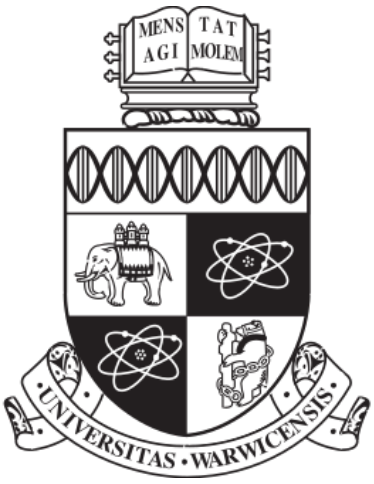
WARWICK
THE UNIVERSITY OF WARWICK



Customising Structure of Graphical Models

by

Rachel Lynne Wilkerson

Thesis

Submitted to the University of Warwick

in partial fulfilment of the requirements

for admission to the degree of

Doctor of Philosophy

Department of Statistics

March 2020

Contents

List of Tables

List of Figures

Acknowledgments

Many thanks to Jim Q. Smith for his enthusiasm, patience, and support. Thank you for sharing your delight in the research process with me.

This research has been strengthened by collaborative effort with several people. Martine Barons originally prompted me to tackle questions facing complex systems like food insecurity from the perspective of decision analysis. Thank you for always lending a listening ear. Manuelli Leonelli, Christianae Gorgen, Eva Ricommagno, Jane Hutton, and particularly Peter Thwaites cultivated a rich conversation on Chain Event Graphs. Ariella Helfgott taught me so much about the importance of storytelling to structural elicitation. Thanks to Aditi Shenvi for troubleshooting diagnostics code and offering a sounding board for new CEG theorems. Thanks to Jack Jewson for collaborative work on beta divergence and CEG model selection.

Thanks are due also to mentors who taught me causal inference, graph theory, and coding hacks: Robert MacKay, Sach Mukherjee, Chris Oates, Leon Danon, Colm Connaughton, and Matthew Shetrone.

This thesis was funded by the Bridges Leverhulme grant and greatly enhanced by the opportunities to host interdisciplinary conversations on causation and reproducibility. Particular thanks to program coordinators Thomas Hills and David Firth, and also to my second supervisor, Emma Uprichard.

Prior work at Baylor University's Texas Hunger Initiative motivated questions about decision analysis and food insecurity. Thanks also to the Baylor Mathematics department for hosting me during the writing up process.

My peers at Warwick taught me perseverance and fostered a real sense of departmental camaraderie. Thanks to my family and friends, particularly my parents for continued support through many math escapades. Lastly, thanks to Calvin for

setting a very compelling thesis deadline, and to Jon for cheering me through to the end.

Declarations

I declare this thesis is based on my own research except where explicitly mentioned. Parts of this thesis have been published as follows.

Most of the material in Chapter Three is to appear in: Expert Judgement in Risk and Decision Analysis to be published in Springer's International Series in Operations Research and Management Science (Editors: Bedford, French, Hanea, and Nane). The content in this chapter was developed in close collaboration with my supervisor Jim Smith, but the paper was written by me, and the applications are mine.

Chapter 4 was developed with Jim Smith and Peter Thwaites. Lemma 54 and the resultant proof is the work of Peter Thwaites. Lemma 36 is the work of Jim Smith. The ancestral graph construction and full separation theorem were developed in close collaboration with Jim Smith, but the methodology and the writing developed in this thesis are mine.

The material in Chapter 5 was similarly developed in close fashion with Jim Smith, but the methodology and applications in this thesis are mine. The material is currently being revised for submission. The R code was developed for both the BN and CEG diagnostics in Chapter 5 are mine. Manuele Leonelli and Ramsiya Ramanathan are coauthors of the R package `bnsens` as they contributed the functions dealing with the sensitivity functions. I am the sole author and maintainer of the CEG package `cegmonitoR`. All figures in this thesis have been generated in RStudio using the `DiagrammeR` package with the exception of Figure 3.9 which was generated with `ggplot2`.

This thesis has not been submitted for examination at another university.

Abstract

Graphical models have proven useful in a wide variety of applications. However, too often the structure of the graphical model is secondary consideration selected for convenience. This thesis makes the case that the chosen structure of the graphical model is fundamental to the resultant analysis. The motivation for this thesis stems from a desire to translate the dynamics described by domain experts into customised statistical models. In this thesis I propose a toolkit for systematically considering other model classes.

The domain of food insecurity motivates the development of models beyond the BN. The examples are illustrated with four graphical model classes: Bayesian Networks, Chain Event Graphs, Multi-regression Dynamic Models, and Flow Graphs. We argue that the problem dynamics should be considered before selecting the model class.

The tree-based Chain Event Graph class of models has proven to be particularly useful for applications in which experts describe a series of events. For this class of models, full checks on the structure are developed, both in the form of theoretical advances in a d-separation theorem and in technical model diagnostics. The full d-separation criteria can be used to verify that the conditional independence relationships implied by the graphs are consistent with the information expressed only through its topology and colouring. The theorem also confirms that using CEG d-separation, conditional independence relationships that cannot be represented by the Bayesian network are expressible in the CEG. The suite of diagnostic monitors check the accuracy of the forecasts that flow from the model. Examining increasingly fine elements of the CEG structure offers checks to see how well the model is

consistent with observations.

Finally, we conclude by considering alternative graphical structures offers nuanced expressions of causation most suitable to certain statistical models. We examine again the four classes of models to illustrate how causal concepts like instrumental variables and intervention become richer in alternative classes of models.

Acronyms

AHC Agglomerative Hierarchical Clustering Algorithm.

BN Bayesian Network.

CEG Chain Event Graph.

DAG Directed Acyclic Graph.

DBN Dynamic Bayesian Network.

DCEG Dynamic Chain Event Graph.

FG Flow Graph.

MDM Multi-regression Dynamic Model.

NSLP National School Lunch Program.

RDCEG Reduced Dynamic Chain Event Graph.

SBP School Breakfast Program.

SMP Summer Meals Program.

SNAP Supplemental Nutrition Assistance Program.

Symbols

G	Graph
$V(G)$	Vertex set
$E(G)$	Edge set
Pa	Parent
Ch	Children
An	Ancestors
De	Descendants
Nd	Non-descendants
Pd	Predecessors
$p(\boldsymbol{x})$	probability mass function
$\mathcal{G}$	DAG
$\mathcal{B}$	BN
$\mathcal{C}$	CEG
$\mathcal{T}$	Tree
s	Situations
l	Leaves
u_i	Stages
v_j	Ancestral positions
w_k	Positions
F	Edge set of $\mathcal{C}$
$\mathcal{F}$	Floret
$\boldsymbol{\theta}_{\mathcal{T}}$	Probability tree
$\mathbb{U}(\mathcal{C})$	Stages of $\mathcal{C}$
$\mathbb{W}(\mathcal{C})$	Positions of $\mathcal{C}$
W	Set of positions $W \subseteq \mathbb{W}$
$\mathcal{C}_\Lambda$	Subgraph

Λ	Set of root to sink paths in $\mathcal{C}$
λ	Path in $\mathcal{C}$
$w(v)$	Set of positions that compose the ancestral positions
W_0	Set of cut-vertices in CEG
$X(w)$	Floret variable
$\boldsymbol{Y}_t$	MDM
$Y_t(i)$	Series of MDM
$\boldsymbol{F}_t$	F of MDM
$\boldsymbol{\theta}_t$	Core states of MDM
$\boldsymbol{G}_t$	Design matrix for MDM
C_0	Variance-covariance matrix
m_0	Mean measurements for MDM
$\mathcal{E}$	Evidence
U	Partition of situations into stages in $\mathcal{C}$
$\boldsymbol{U}$	All possible partition of situations into stages in $\mathcal{C}$
$\mathscr{E}$	Essential graph
$\mathscr{F}$	Filtration

Chapter 1

Introduction

> On their backs were vermiculate patterns that were maps of the world in its becoming. Maps and mazes.
>
> *The Road*, Cormac McCarthy

1.1 Motivation

Graphical models have proven to be an invaluable tool for decision analysis and causal inference. They have been used in a wide range of applications due to their ease of interpretability and flexibility. The structure of graphical models can be elicited directly from domain experts or found using structural learning algorithms. The integrity of this structure ultimately affects the success of the model. Specifically, deriving the model structure from the dynamics as the domain expert describes them creates germane models. The motivation for this thesis is to develop new theorems, methodology, and applications that support drawing the structure of graphical models directly from domain experts' description of a problem.

Of these graphical models, the most well known and widely used is the Bayesian Network (BN). The BN is a collection of conditional independence relationships among a set of random variables. However, sometimes the BN structure is inadequate for the sort of problem dynamics exhibited by complex problems. The problem may exhibit particular dynamics ill-suited to a BN, or asymmetries that render the conditional probability tables of a BN nonsensical. Often, the BN may need to be supplemented by context-specific conditional probabilities. These challenges have prompted development of other models with customised semantics.

Rather than coercing natural language descriptions of a problem to fit the BN structure, I propose that graphical structures can be drawn from a domain experts'

natural language description of a problem. These alternative structures prompt the development of methodology analogous to that of the BN that is customised to the described problem dynamics. The process by of deciding on the structure of a graphical model is not straightforward (Korb and Nicholson, 2010). Describing the process by which decision makers choose a custom structure is a the main contribution of Chapter 3.

Two main contributions of this thesis focus on Chain Event Graphs (CEG)–a class of models that expands BN machinery to a more flexible class of models. CEGs use a tree-based structure to describe unfolding processes–a natural fit to descriptions often given by domain experts. Additionally, the CEGs incorporate context-specific probabilities into a single graphical representation. Discrete BNs represent a subclass of CEGs. Many of the developments for CEGs have drawn from BN methodology. Model selection (Freeman and Smith, 2011a), equivalence class (Görgen and Smith, 2018), evidence propogation w, causal inference (Thwaites and Smith, 2010) have all been extended to the CEG. A number of applications have demonstrated the efficacy off the CEG over other graphical models (Barclay et al., 2013; Barclay, 2014).

This thesis adds two contributions to this methodology. The first is a complete d-separation theorem for the CEG in Chapter 4. This builds from the theorem for simple CEGs from Thwaites and Smith (2015). D-separation theorems proved foundational for BNs. It also contributes a new ancestral CEG construction as well as a construction that shows the dependence between the random variables defined directly from a CEG. The full d-separation theorem admits querying any elicited CEG structure to verify that it represents the domain experts' beliefs as a valid CEG.

The second CEG contribution in Chapter 5 consists of a suite of diagnostic monitors to evaluate the accuracy of the forecasts flowing from the CEG monitor. These prequential monitors, so named for their verification of the predictive, sequential forecasts generated by the model, can be used to detect discrepancies in the structure from the existing fit to the data (Dawid, 1984, 1992; Dawid and Studenỳ, 1999). These build on the work for diagnostic monitors for BNs defined and demonstrated in Cowell et al. (1999) and rely on the message passing algorithm from Collazo et al. (2018). The diagnostics offer a way to check different elements of the structure of a CEG in an online learning environment.

After establishing these additional checks specifically for the CEG, Chapter 6 examines how customising structure affects the nuances of causal inference. The final contribution of this thesis examines examples of customised models and examines how alternative structures prompt a new understanding of concepts like naive cause and instrumental variables. Building on the work of causal inference in the Multi-

regression dynamic model, I prove that every edge in the graph can be thought of as an instrumental variable (Wright, 1921, 1925; Bowden and Turkington, 1990). The full d-separation theorem from Chapter 4 also allows a new definition of instrumental variables for the CEG.

In summary, this work can be summarised by the following research questions:

1. How can customised graphical models be elicited from domain experts?
2. How can d-separation verify the conditional independences of the CEG?
3. How do model diagnostics check the elements of structure of a CEG?
4. How does using customised graphical models offer nuanced definitions of causation?

1.2 Thesis Outline

Chapter 2 outlines basic notation in graphical models. After describing the semantics of the Bayesian Network in Section 2.3.2, Section 2.4.1 describes the development of a CEG from a tree-based model. The chapters concerning the importance of customising models to domain expertise also draw on two other alternative models, the Multi-regression dynamic model (MDM) and the Flow Graph. These models are useful examples of specific natural language descriptions of a problem. Chapter 6 derives new results that show how these new model assumptions add new meaning to cause in graphical systems. The chapter concludes with a discussion of various graphical models.

Chapter 3 exhibits how the custom structure of a problem can be translated from experts' natural language description. General guidelines for selecting an alternative model are proposed, and specific examples are given in the subsequent sections. Each of the examples is drawn from natural language descriptions of a program-specific challenge in food insecurity. The conditional independence statements implied by the graph must concur with experts' description. Checks for each of the custom graphical models is given in each section of Chapter 3. While the d-separation theorem makes these checks possible for the BN, full theorems are not always available for the alternatives to the BN. Useful forecast checks for the MDM are given in Section 3.3.3. The chapter concludes with guidelines for choosing the most relevant model.

Verifying that the conditional independence statements implied by the CEG concur with the intended statements has routinely been checked through several theorems. These queries were first framed as logical and-or gates in Smith and

Anderson (2008), and a theorem for the simple subclass of CEGs was defined in Thwaites and Smith (2015). However, a full d-separation theorem analogous to that of the BNs has not been known. In Chapter 4, I derive such a theorem. This can be shown to confirm all of the conditional independence statements that can be read from a BN in the CEG in addition to the queries that are unique to the CEG. The theorem relies on the construction of an ancestral CEG whose definition appears for the first time here.

After confirming the structure of the CEG, Chapter 5 discusses checks for the forecasts from the model. We begin by reviewing the diagnostic monitors for the BN in Section 5.3 and then define the CEG monitors in Section 5.4. Examples from healthcare and radicalisation illustrate and elucidate these tools.

Chapter 6, explores how the custom models give rise to different notions of causation. Section 6.2.1 discusses the importance of directional invariance across an equivalence class of graphical models and the implications this has on causal reasoning. Section 6.2.2 compares how notions of instrumental variables can be interpreted in custom models. These two definitions are examined in the context of the MDM and CEG in Sections 6.4 and 6.5 respectively.

In the final chapter, I review the contributions made by this thesis and discuss areas of future work.

Chapter 2

Different Semantics, Different Models

...'embrace the monsters' and explore alternative approaches to representation.

David Gooding, "Thinking Through Computing"

2.1 Customised Model Semantics

Incorporating domain expertise is a fundamental tenet of statistical methodology. Complex problems have different dependencies and functional relationships that may not fit in an off-the-shelf model. Graphical models have been developed to depict various different types of dependence relationships in systems. Probabilistic graphical models have since emerged as a particularly powerful methodology for investigating causal relationships.

The accessibility and interpretability of graphical models facilitates a useful exchange between domain experts and statisticians. The small set of variables and visual aid of the graph enables statisticians to translate the relationships to domain experts easily (Pearl, 1986; Smith, 2010). Additionally, framing questions in terms of proposed interventions offers a useful focal point for domain experts because it relates to what they might do in practice. In addition to their accessibility, graphs are also useful for inference and learning. They represent an efficient, compact version of the joint distribution allows for efficient posterior computation and model selection.

2.2 Graphical Models

Among these classes of manipulated causal models, the framework of BNs and their dynamic counterparts are the most widely used to explore causal hypotheses.

BNs represent the dependence structure between sets of random variables. The methodology of BNs and their wide applications has been reviewed in Pearl (2009); Lauritzen and Richardson (2002); Cowell et al. (2007); Korb and Nicholson (2010). The random variables of a BN admits any distribution, allowing them to incorporate either discrete or continuous variables. Many different variations of a BN have been developed. Context-specific BNs take different probability distributions for certain settings of the random variables (Boutilier et al., 1997). Time-varying BNs alter their structure in a regular pattern to capture the dynamics of different structures. The Dynamic Bayesian Network (DBN) extends the BN structure through either discrete or continuous time. DBNs in particular have been used to model food insecurity (Barons and Smith, 2014), an application I will examine in subsequent chapters.

Despite the prevalence and success of BN methods, the BN is not always an appropriate modelling choice. Expressing a process as a series of unfolding events is often more natural to domain experts than thinking in terms of random variables (Shafer, 1996). Rather than expressing a process as a series of random variables, modelling it as events lends itself to the structure of a tree. Decision trees proved foundational to decision and utility theory (Raiffa, 1968). From this development, influence diagrams emerged as a more succinct way to represent the possibilities in the decision tree (Howard and Matheson, 2005). However, what the representation gained in compactness it lost in expressiveness. Building on the advancements from probability trees and the accessibility of influence diagrams, Chain Event Graphs (CEGs) proposed a compact representation of the relationships expressed in a large event tree (Smith and Anderson, 2008). Built from an event tree, the CEG models conditional independence relationships through a colouring known as stages. Dynamic analogues of the CEG have been developed (Barclay and Nicholson, 2015; Collazo and Smith, 2018), as well as the Reduced Dynamic Chain Event Graph (RDCEG) (Shenvi and Smith, 2018).

In addition to the dynamic analogues of the BN and the CEG, there are often situations in which conditional independence structures are preserved across time. Towards that end, Smith (1993) developed the Multi-regression Dynamic Model–a class of multivariate state space time series models. The MDM borrows directly from the linear dynamic model literature, and thus can incorporate seasonal trends and interventions easily (Harvey, 1986; Quintana and West, 1987; West and Harrison, 1997; Durbin and Koopman, 2012). In their simplest form, MDMs link together

univariate Bayesian dynamic linear regression models (DLMs) (Harrison and Stevens, 1976).

Recent advancements include model selection methods using integer programming and a set of diagnostic monitors in Costa et al. (2015). Costa et al. (2017) demonstrates the quick model selection algorithm for both cyclic and acyclic variants of the MDM. Costa et al. (2019) defines a new class of MDM models that shows the dependence structure across groups of individuals while maintaining individual structures.

Another application of the MDM derives a corresponding undirected graph that can be interpreted as an influence diagram (Queen, 1992). The MDM expresses contemporaneous causal relationships for applications ranging from brain connectivity (Costa et al., 2015), traffic networks (Queen and Albers, 2009), to brand forecasting (Queen, 1992). Zhao et al. (2016) defined a variant of the MDM, the dynamic dependence network, that allows the connectivity to change over time.

Another example of a graphical model with semantics different to that of the BN is the Flow Graph (FG). Motivated by applications where physical goods flow through a network of actor interactions, the Flow Graph describes the state of path flows through a network. The Flow Graph offers an example of a structure where model assumptions have been added that break the restrictions set by the BN (Figueroa and Smith, 2007).

These four graphical models are the focus of this thesis and will be used to illustrate food insecurity applications in Chapters 3 and 6. Additional types of graphical models have emerged. Controlled regulatory graphs, composed of hyperclusters, represent new dynamics that describe biological applications including circadian regulation (Liverani and Smith, 2015). Chain graphs incorporate a series of undirected and directed edges that also describe dynamics different to that of the BN (Studenỳ, 2006).

2.2.1 Semi-graphoid axioms

The semi-graphoid axioms formalize the notion of irrelevance that underpins graphical structures (Dawid, 2001). Pearl (1988) linked these axioms to a graphical representation. These were first characterized in the context of probabilistic expert systems (Studenỳ, 1989) and more generally in Smith (2010). Then, the axioms were connected to dependency models (Studenỳ, 1993). Intuitively, these properties maintain that extraneous information remains irrelevant (Smith, 2010). These axioms hold for all probabilities, which prompts looking at graphical models with other meanings. i A semi-graphoid conditional independence model satisfies the four axioms below.

Definition 1 (Symmetry) *The **symmetry** property requires that for three disjoint measurements $X, Y,$ and Z:*

$$X \perp\!\!\!\perp Y|Z \Leftrightarrow Y \perp\!\!\!\perp X|Z$$

Definition 2 *The **decomposition** property requires that for disjoint measurements X, Y, W and Z:*

$$X \perp\!\!\!\perp (Y, W)|Z \Rightarrow X \perp\!\!\!\perp Y|Z \text{ and } X \perp\!\!\!\perp W|Z$$

Definition 3 *The **weak union** property requires that for three disjoint measurements X, Y, W and Z:*

$$X \perp\!\!\!\perp (Y \cup W)|Z \Rightarrow X \perp\!\!\!\perp Y|(Z \cup W)$$

Definition 4 *The **contraction** property requires that for three disjoint measurements X, Y, W and Z:*

$$X \perp\!\!\!\perp Y|Z \text{ and } X \perp\!\!\!\perp W|(Z \cup Y) \Rightarrow X \perp\!\!\!\perp (Y \cup W)|Z$$

These axioms can be used to derive additional rules about independence (Dawid, 1979). The system can also be used to compare the expressiveness of different graph forms (Pearl, 1988). The semi-graphoid axioms frame questions of irrelevance in Chapter 3, offering a way to confirm that the structure given by domain experts is consistent with the dependence structure. In Chapter 4, the functional form of the semi-graphoid axioms is used to prove results affiliated with the d-separation theorem.

2.3 Probabilistic Graphical Models

2.3.1 Basic Graph Theory Definitions

A graph $G = (V, E)$ is a finite set of vertices V and edges E. E is a set of ordered pairs of distinct vertices, a subset of set $V \times V$. Edges can be either directed or undirected, but this thesis largely deals with directed graphs. An edge between two vertices is **directed** when $(v_i, v_j) \in E$ but $(v_j, v_i) \notin E$ A complete graph contains all pairs of vertices in the edge set. A **path** in a graph is a sequence of vertices $v_1, \ldots, v_n$ such that there is an edge between $(v_i, v_{i+1}) \in E$ for $i \in 1, \ldots, n-1$.

If there is a directed edge from v_i pointing at v_j then v_j is the **child** and v_i is the **parent**. The parent and child sets of a set of vertices $V_i \in V(G)$ are denoted

$\mathrm{Pa}(V_i)$ and $\mathrm{Ch}(V_i)$, respectively. We say parents are unmarried when there is not an edge between two parents that share a common child. A graph is **moralised** undirected edges are added between unmarried parents. If there is a directed path from v_i to v_j then v_i is the **ancestor** and v_j is the **descendant**. The ancestor and descendant sets of vertices $V_i, V_j \in V(G)$ are denoted $\mathrm{An}(V_i)$ and $\mathrm{De}(V_j)$. These sets are the vertices $v \in V$ such that there is a directed path from $v_i \in V_i$ to $v_j \in V_j$. It is also convenient to denote the set of **non-descendants** $\mathrm{Nd}(v) = V \setminus (\mathrm{De}(v) \cup \{v\})$.

A cycle is a path that begins and ends at the same vertex. A cycle is directed if there is a directed path between each of the edges.

Definition 5 *A **directed acyclic graph** (DAG) is a directed graph with no directed cycles.*

A graph is **directed** if all edges $e \in E$ are directed. A graph with both directed and undirected edges is called a **hybrid** graph. A **tree** is a graph without cycles that has a unique path between any two vertices. A **directed tree** is a directed acyclic graph that has a tree as its underlying structure. Directed trees are the basis of the CEG structure. The directed tree has one **root** vertex with no parents and the remaining vertices have exactly one parent. The edges are directed away from the root node.

DAGs form the basis for inference on probabilistic graphical models. One question this thesis addresses is what the edges mean in a causal DAG. Domain experts usually mean something by the edges in a DAG that may or may not be compatible with the conditional independence interpretation. Customising graphical models finesses the meaning of the edges in graphical models. For example, in the CEG the edges represent the possible events emanating from a particular vertex. Similarly, with models based on the Dynamic Linear Model (DLM), an edge means that v_i is an input of the function of v_j. In the subsequent sections, these edge definitions alter the underlying framework for causation.

2.3.2 Bayesian Networks

Bayesian Networks have proven to be a powerful method for representing conditional independence relationships between random variables (Pearl, 1986; Cowell et al., 2007).

The nodes of the graph indicate the random variables $\boldsymbol{X} = \{X_1, X_2, \ldots, X_n\}$ and the edges depict possible dependence structures. They are sometimes called Bayesian belief networks or causal networks, although the latter terminology implies additional model assumptions. Object oriented Bayesian Networks expand complex

BNs and became widely used through the HUGIN software (Koller and Pfeffer, 1997; Jensen, 2014).

The BN can be thought of as sets of conditional independence statements. Dawid and Studenỳ (1999) defines conditional independence in terms of factorisations. BNs have a joint probability mass function $p(\boldsymbol{x})$ on set of random variables that can always be written as:

$$p(\boldsymbol{x}) = p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p(x_i | x_1, \ldots, x_{i-1}). \tag{2.1}$$

The **ordered directed Markov property** states that a random variable is independent of its predecessors $\mathrm{Pd}(X_i)$ given its parents.

$$X_i \perp\!\!\!\perp \mathrm{Pd}(X_i) | \mathrm{Pa}(X_i)$$

This allows Equation 2.1 to be written as a **recursive factorization** of conditional probabilities

$$p(\boldsymbol{x}) = p(x_1) \prod_{i=2}^{n} p(x_i | \mathrm{Pa}(x_i)) \tag{2.2}$$

where $\mathrm{Pa}(x_i)$ indicates the parent set of x_i.

Conditional independence is defined as follows

Definition 6 *Random variables X, Y, and Z are **conditionally independent** if and only if $p(x, y) = p(x)p(y)$. The variables are conditionally independent given Z if and only if*

$$p(x, y|z) = p(x|z)p(y|z)$$

for all points other than those on a set of measure zero when $p(z \geq 0)$.

The BNs and CEGs considered in this thesis are discrete models. The MDM and FG rely on continuous joint probability mass function given in Section 2.4.2.

Definition 7 *A Bayesian Network with probability distribution $p(\boldsymbol{x})$ satisfies the **local Markov property** with respect to the directed acyclic graph $\mathcal{G}$ if x_i is independent of its nondescendants $\mathrm{Nd}(x_i)$ given its parents:*

$$x_i \perp\!\!\!\perp \mathrm{Nd}(x_i) | (\mathrm{Pa}(x_i))$$

The local Markov property articulates the conditional independence relationships from the missing edges of the graph G. Additional conditional independence relationships can be deduced from the structure of a graph using the d-separation

theorem, or the global directed Markov property, another formulation of the d-separation theorem (Hammersley and Clifford, 1971; Pearl, 1986; Geiger et al., 1990b; Frydenberg, 1990a,b). This thesis uses the ancestral formulation of the global directed Markov property from Lauritzen et al. (1990). The construction of the ancestral graph requires edges to be moralised, that is, adding an undirected edge between two parent that have a common child.

Theorem 8 (d-separation for BNs) *Given a DAG $\mathcal{G}$ with three disjoint sets of vertices $B_1, B_2,$ and $C \in V(\mathcal{G})$, B_1 d-separates B_2 given C, written as $B_1 \perp_d B_2$ if there is no path from a vertex in B_1 to a vertex in B_2 when the set of vertices in C is removed from the moralised ancestral graph $\left(\mathcal{G}_{An(B_1 \cup B_2 \cup C)}\right)^m$.*

The d-separation theorem for Bayesian Networks respects a graphical representation for a given set of conditional independence relationships (Geiger et al., 1990a; Lauritzen and Richardson, 2002; Cowell et al., 1999; Smith, 2010). Geiger et al. (1990b) first proved that d-separation holds when the sample space of each random variable in the set meets certain assumptions. Then, Pearl proved that given a BN of Gaussian variables, for a query that failed d-separation, a probability distribution could be constructed that showed dependence. This result was then confirmed for binary variables, which also provided a counterexample for a BN with discrete variables. These proofs all rely on the fact that conditional independence relationships can be represented by a single, faithful BN. If the d-separation holds for a particular query, then the variables are conditionally independent. The converse is typically stated as a corollary: if the d-separation query is violated, then there may be a setting of the probabilities that violate the conditional independence relationship. The converse of d-separation does not necessarily hold for any BN with context-specific conditional independence relationships.

Pearl (1986) derived an alternative form of the global directed Markov property in terms of active pathways. Geiger and Pearl (1990) showed that this cannot be improved upon, and thus proved the d-separation theorem. This theorem can be used to check that the conditional independence relationships within the graph are compatible with the domain expert's description of the problem dynamics.

Definition 9 *A Bayesian Network $\mathcal{B}$ on a set of random variables $\boldsymbol{X} = X_1, X_2, \ldots, X_n$ is a directed, acyclic graph that admits a recursive factorization or equivalently, a set of $n-1$ conditional independence statements of the form*

$$X_i \perp\!\!\!\perp Pd(X_i) | \mathrm{Pa}(X_i).$$

Bayesian Networks can also be thought of as set of conditional probability

vectors (CPV) of the form $p(x_i|\text{Pa}(x_i))$. These CPV quantities can be elicited after the structural conditional probability relationships are elicited from experts.

In addition to the Markov assumption, valid BNs also meet the faithfulness assumption. Faithfulness means that there are no lurking dependence structures among the random variables of the graph.

Definition 10 *A Bayesian Network G is* ***faithful*** *if the distribution contains all and only the conditional independence relationships implied by the Markov condition.*

There is a set of DAGs that describes the same conditional independence relationships. There are Markov equivalence classes of graphs that represent the same conditional independence relationships. This means that the edges of a BN cannot be interpreted as causal, as there may be an equivalent graph with the edge reversed.

The Markov and faithfulness assumptions define a BN. The literature on causation strongly emphasizes the importance of manipulation (Holland, 1986; Dawid, 2002; Pearl, 2009). For a BN to be considered causal, Pearl argued that any intervention on a random variable $X_i = \hat{x}_i$ should have the same effect as conditioning on a random variable according to the following intervention formula. External manipulation of setting a value $X_i = \hat{x}_i$ is denoted using the **do operator**, $X_i = \text{do}(x_i)$ or sometimes using $X_i||\hat{x}_i$. The atomic intervention formula is given as

$$p(\boldsymbol{x}_{-i}||\hat{x}_i) = \begin{cases} \frac{p(x_1,\ldots,x_i,\ldots,x_n)}{p(x_i|\text{Pa}(x_i))} & \text{if } x_i = \hat{x}_i \\ 0 & \text{otherwise.} \end{cases} \tag{2.3}$$

This formula shows that the variables that are downstream of X_i inherit the values they would have taken if $X_i = \hat{x}_i$ had occurred naturally. The upstream variables are unaffected. This assumption is quite strong, and often not applicable to particular dynamics experts describe. For instance, interventions in controlled regulatory networks affects upstream variables as well as downstream ones (Liverani and Smith, 2015).

Definition 11 *A causal Bayesian Network $\mathcal{B}$ is causal if it admits the atomic intervention formula given in Equation 2.3 for all variables $X_i \in V(\mathcal{B})$.*

BNs have been adapted to several different variations. These can be used to describe different dynamics, but they are also subject to the same assumptions about faithfulness and intervention. Bayesian Networks have been adapted to Dynamic Bayesian Networks. DBNs show the relationships between variables over time. The BN structure repeats for each time step, and an additional set of edges shows the

dependence structure between time steps. DBNs have been used to model food insecurity for decision analysis (Barons and Smith, 2014).

Object oriented Bayesian Networks (OOBN) use the logic of circuits to model complex, hierarchical problems with different classes (Koller and Pfeffer, 1997; Korb and Nicholson, 2010). The OOBNs can also incorporate temporal and spatial features, as they have for an application studying ecological systems (Chee et al., 2016) .

The numerous variations of BNs are helpful for a range of applications. However, restricting causal questions to the framework of conditional independence of random variables is fundamentally limiting. Alternative graphical models allow us to retain the interpretability and accessibility of graphical structures while allowing for more nuanced representations of problem structure.

2.4 Alternative Graphical Models

The strength of the BN approach is well established in literature, but there have been calls for alternative representations beyond the BN (Spirtes and Zhang, 2016). Much of this thesis will look at expanding the toolkit of BN methodology to these alternative representations. Doing so expands the variety of model classes to describe dynamics identified by domain experts. Several such models have been developed, and a few are highlighted in this Section. This is by no means exhaustive, but rather to be taken as a sample of the alternative structures that could be devised and their ramifications on causal inference.

2.4.1 Chain Event Graphs

Chain Event Graphs (CEGs) are one method for incorporating problems with asymmetries and context-specific conditional independences (Smith and Anderson, 2008). Context-specific independence arises when a model exhibits independence relationships for a particular setting of the parent values (Boutilier et al., 1997).

Definition 12 *For disjoint sets X, Y, and C, X is **conditionally independent** of Y given the particular **context** $C = c$ if*

$$p(X|c, Y) = p(X|c).$$

In the BN setting, the conditional probability vectors must be given different values for different contexts. The CEG offers a cohesive way to encode all the possible contexts in a single, coloured, tree-based model. In this thesis, trees refer to directed trees. I review this construction in this section, first introducing event trees, then a colouring, then a class of staged trees, then a much simpler graph derived from the

staged tree–a CEG. An example of the CEG elicitation and construction is shown in Chapter 3. Chapters 4 and 5 explore theoretical and methodological advancements for the CEG.

Definition 13 *A **tree** $\mathcal{T} = (V, E)$ is a connected, directed graph with no cycles.*

In this definition, V and E denote the node and edge set respectively. The set of vertices $\text{Pa}(v) = \{v' \,|\, \text{there is } (v', v) \in E\}$ represents the parents of $v \in V$ and $\text{Ch}(v) = \{v' \,|\, \text{there is } (v, v') \in E\}$ denotes the children of $v \in V$. It is often helpful to distinguish between the vertices which are **situations** $s \in S$ and the leaf nodes $l \in V \setminus S$. Situations nodes are non-leaf nodes in the event tree. We can denote the set of root-to-leaf paths in an event tree by $\Lambda(\mathcal{T})$. $\Lambda(v)$ and $\Lambda(e)$ refer to **vertex-centred** and **edge-centred events**, the subset of all root-to-leaf paths that pass through either the vertex v or edge e. For a particular situation $v \in V$ and its emanating edges $E(v) = \{(v, v') \in E | v' \in \text{Ch}(v)\}$, define a **floret** as the pair $\mathcal{F}(v) = (v, E(v))$.

We next assign a probability distribution to this event tree with parameters $\theta(e) = \theta(v, v')$ corresponding to the edge $e = (v, v') \in E$. The components of all floret parameter vectors sum to unity $\sum_{e \in E(v)} \theta(e) = 1$ for all $e \in E$ and $v \in V$. Each parameter $\theta(e), e \in E$ is a primitive probability. These primitive probabilities serve a similar role to potentials in BNs. The pair $(\mathcal{T}, \boldsymbol{\theta}_\mathcal{T})$ of a graph $\mathcal{T}$ and all labels $\boldsymbol{\theta}_\mathcal{T} = (\theta(e) | e \in E)$ is called a **probability tree**. Building on the definition of a probability tree as the pair $(\mathcal{T}, \boldsymbol{\theta}_\mathcal{T})$ with graph $\mathcal{T} = (V, E)$ and labels $\boldsymbol{\theta}_\mathcal{T} = (\theta(e) | e \in E)$, now define a **staged tree**. The stagings represent context-specific conditional independence in the CEG.

Definition 14 *Two vertices representing situations $v, v' \in S$ are in the same **stage** u if and only if their floret distributions are equal up to a permutation of their components $\boldsymbol{\theta}_v = \boldsymbol{\theta}_{v'}$.*

Each stage is assigned a unique colour. An event tree can be transformed to a staged tree by colouring the vertices according to their stage memberships. If all vertices are either in the same stage or have pairwise different labels, then $(\mathcal{T}, \boldsymbol{\theta}_\mathcal{T})$ is a staged tree. The set of vertices of the staged tree is partitioned into equivalence classes of vertices in the same stage, denoted as

$$\mathbb{U} = \{u \subseteq V \,|\, v \text{ and } v' \text{ are in the same stage for all } v, v' \in u\}. \tag{2.4}$$

There is a finer partition of events called positions $w \in \mathbb{W}$. Let $\mathcal{T}(v) \subseteq \mathcal{T}$ be the event tree rooted at $v \in V$ and whose root-to-leaf paths are inherited from $\mathcal{T}$. Then the pair $(\mathcal{T}(v), \boldsymbol{\theta}_{\mathcal{T}(v)})$ is a probability subtree of $(\mathcal{T}(v), \boldsymbol{\theta}_{\mathcal{T}(v)})$.

Definition 15 *Two situations $v, v' \in u$ which are in the same stage $u \in U_{\mathcal{T}}$ are also in the same* ***position*** *if their subtrees $(\mathcal{T}(v), \boldsymbol{\theta}_{\mathcal{T}(v)})$ and $(\mathcal{T}(v'), \boldsymbol{\theta}_{\mathcal{T}(v')})$ have the same graph and the same set of edge labels.*

Building on the concepts of stages and positions, a CEG can be constructed from a staged event tree by merging situations that lie in the same position. The leaves of the tree are subsumed into a sink node.

Definition 16 *A* ***Chain Event Graph*** *$\mathcal{C}(\mathcal{T}) = (\mathbb{W}, F)$ is the pair of positions $\mathbb{W}$ and accompanying edge set F. The vertex set $\mathbb{W} = \mathbb{W}_{\mathcal{T}}$ is the set of positions in the underlying tree $\mathcal{T}$.Each position w inherits its colour u from the staged tree. If all edges $e = (v_1, v_1')$, $e' = (v_2, v_2') \in E$ and the vertices v_1, v_2 are in the same position, then there is a corresponding edge $\{f, f'\} \in F$. The labels $\theta(f)$ of edges $f \in F$ are inherited from the corresponding edges in the staged tree. The labelled graph $(\mathcal{C}(\mathcal{T}), \boldsymbol{\theta}_{\mathcal{T}})$ is a Chain Event Graph.*

Within the model class of CEGs, there are subclasses of models with particular properties. CEGs that are equivalent to BNs are stratified, as in Definition 17.

Definition 17 *A CEG $\mathcal{C}$ is* ***stratified*** *if the $\Lambda(\mathcal{C})$ are identified with elements in the product state space of the ordered set of random variables $\boldsymbol{X} = (X_1, \ldots, X_i, \ldots, X_n)$ where every component X_i has a set number of levels, K_i, such that each of the levels is the same distance from the root node.*

This thesis also concerns a second subclass of CEGs, the square-free CEGs. The results derived in Chapter 4 apply to square-free CEGs.

Definition 18 *A CEG $\mathcal{C}$ is* ***square-free*** *if it contains only graphs for which no two situations lying on the same root-to-sink paths also lie in the same stage.*

Model search algorithms

Various statistical methodologies for model selection, estimation, and message passing methods have now been developed (Freeman and Smith, 2011a,b; Barclay et al., 2013, 2014; Thwaites et al., 2008; Cowell and Smith, 2014; Collazo and Smith, 2015a; Collazo et al., 2018; Thwaites and Smith, 2015). Current search algorithms have been developed for stratified CEGs that search the space of trees. These include dynamic programming methods (Collazo et al., 2018) and an Agglomerative Hierarchical Clustering (AHC) algorithm (Freeman and Smith, 2011a). A greedy search algorithm may miss the optimal model, further reason to check the model using our diagnostics in Chapter 5. The AHC algorithm often merges sparsely

populated situations which may return a local rather than global optimum solution. Further adaptations of these search methods have been developed including a search method based on Bayesian Information Criterion (BIC) (Schwarz, 1987). These algorithms have been implemented in Varando et al. (2020). Search for asymmetric structures is currently being developed, as are extensions to search over a range of variable orderings (Collazo et al., 2018).

Model selection algorithms may be used to find a Markov equivalent class of models, and within this, a candidate model may be selected. Methods to identify when two trees encode identical beliefs about the data have now been determined by Görgen and Smith (2018). However, one omission within this technological toolbox are routine diagnostics to apply to this class. The purpose of Chapter 5 is to fill the gap.

Helpfully, the class of stratified CEGs encompasses the class of discrete, context-specific BNs. From this stratified CEG, adaptations such as pruning edges may be made at the suggestion of domain experts. Dynamic programming methods can be used to find the maximum a posteriori CEG (Collazo et al., 2018). For a faster, more scalable method, the Agglomerative Hierarchical Clustering (AHC) algorithm may be used to search the possible colourings of the stages (Barclay et al., 2013). The models returned by the ACH algorithm have been shown to be sufficiently close to the generating model for a surprising number of examples (Barclay et al., 2013).

2.4.2 Multi-regression Dynamic Model

The Multi-regression Dynamic Model is a collection of time series that can be used to describe the dynamics between processes (Smith, 1993; Costa, 2014; Costa et al., 2015). The edges in a Dynamic Bayesian Network (DBN) model the dependence relationships between time steps as in a BN. In contrast, the edges in a MDM represents the effective connectivity between the parent and child time series. This means that, unlike the DBN, the MDM represents contemporaneous causal relationships.

The MDM exemplifies a graph where adding additional model assumptions to the graphical representation offers a custom version of the BN assumptions.

Definition 19 *A collection of time series* $\mathbf{Y}_t = \{Y_t(1), \ldots, Y_t(i), \ldots, Y_t(n)\}$ *can be considered a* ***Multi-regression Dynamic Model (MDM)*** *if the observation equations, a system equation, and initial information as given respectively in Equations 2.5, 2.6, and 2.7 adequately describe the system.*

Each series in an MDM can be represented by an **observation equation** of

the form:

$$Y_t(r) = \boldsymbol{F}_t(r)'\boldsymbol{\theta}_t(r) + v_t(r) \qquad v_t(r) \sim (0, V_t(r)),\ 1 \leq r \leq n \tag{2.5}$$

where $\theta_t = \{\theta_t(1), \ldots, \theta_t(n)\}$ are the state vectors determining the distribution of $Y_t(r)$. $\boldsymbol{F}_t(r)$ is a known function of $\boldsymbol{y}_t(r)$ for $1 \leq r$. That is, each observation equation only depends on the past and current observations rather than the future ones. $V_t(r)$ are known scalar variance observations. These can be estimated from available data or else elicited from experts. The indexing over r encodes the strict ordering of the nodes that is so key for this problem.

The **system equation** is given by:

$$\boldsymbol{\theta}_t = G_t\boldsymbol{\theta}_{t-1} + \boldsymbol{w}_t \qquad \boldsymbol{w}_t \sim (\boldsymbol{0}, \boldsymbol{W}_t); \tag{2.6}$$

where $\boldsymbol{G}_t = \text{blockdiag}\{G_t(1), \ldots, G_t(n)\}$ and $\boldsymbol{w}_t$ has a general distribution. Each $G_t(r)$ represents a $p_r \times p_r$ matrix. For a linear MDM, let $\boldsymbol{G}_t$ be the identity matrix. The term $\boldsymbol{w}_t$ represents the innovations of the latent regression coefficients, that is the difference between the observed and forecasted values. $\boldsymbol{W}_t = \text{blockdiag}\{W_t(1), \ldots, W_t(n)\}$, where each $W_t(r)$ has dimensions $p_r \times p_r$, where p_r is the number of parent of $Y_t(r)$.

Lastly, the **initial information** is expressed as:

$$(\boldsymbol{\theta}_0|y_0) \sim (\boldsymbol{m}_0, C_0). \tag{2.7}$$

where $\boldsymbol{m}_0$ is a vector of mean measurements of the observation and C_0 is the variance-covariance matrix where $C_0 = \text{blockdiag}\{C_0(1), \ldots, C_0(n)\}$.

Chapter 3 contains an example of how the MDM can be elicited. It also explores the conditional independence relationships between different elements of the model. Equivalence classes for the MDM are discussed in Chapter 6 along with the different possible interventions the MDM may admit.

Existing applications of the MDM include brand forecasting (Queen, 1992), brain connectivity (Costa, 2014; Costa et al., 2015), and traffic flows (Queen and Albers, 2009). The MDM has also been used in the context of decision analysis for nuclear emergency response (Leonelli and Smith, 2013). A variant of the MDM, the dynamic chain graph model (Anacleto and Queen, 2016) allows for more complex dependence structures.

2.4.3 Flow Graph

The Flow Graph offers a graphical representation of the flow of goods through a supply chain. A hierarchical flow network can be used to model a supply chain with different levels. This flow of products through a network can be modelled by a network of actors $z(l, j_l)$ where l specifies the hierarchy level and j_l indicates the number of actors in hierarchy level l. The edges between actors in the Flow Graphs represents the transfer of mass from one actor to another in a subsequent level. This graph cannot be construed as a BN because conservation of mass in the network is assumed. This constraint induces other dependencies in a naive BN interpretation. between actors in a given level. Intervening on the level of goods at one actor affects the levels of goods available to the remaining actors in the same level.

In order to translate the causal machinery of BNs to the Flow Graph, Figueroa and Smith (2007) composed a graphical model where the elements of the system are the possible path flows through the system rather than the individual actors. This allows the system to be transferred instead as a set of multivariate multilevel Dynamic Linear Models. Figueroa and Smith (2007) describes a new calculus for intervention in the system.

Chapter 3 again includes an example of how a FG might be elicited from experts. I briefly consider the causal ramifications of the two time slice DBN that represents the path flows through the FG in Chapter 6. While intervention and a new do calculus has been defined for the FG, translating elements like equivalence classes, d-separation, and model selection to this this model remains an open question.

The aforementioned graphical models are the main ones considered in this thesis, but they are only a small sample of possible graphs that could be customised to particular dynamics. The controlled Regulatory Graph represents another instance of a successful translation of described expert dynamics to a bespoke structure. Liverani and Smith (2015) describes this new class of models with respect to biological regulation mechanisms.

Chapter 3

Structural Elicitation for Customised Graphical Representations

> It's just that occasionally the math makes its own rules. The math gets to do that if it wants to.
>
> *Middlegame*, Seanan McGuire

Established methods for structural elicitation typically rely on code modelling standard graphical models classes, most often Bayesian Networks. However, more appropriate models may arise from asking the expert questions in common language about what might relate to what and exploring the logical implications of the statements. Only after identifying the best matching structure should this be embellished into a fully quantified probability model. Examples of the efficacy and potential of this more flexible approach are shown below for four classes of graphical models: Bayesian Networks, Chain Event Graphs, Multi-regression Dynamic Models, and Flow Graphs. To be fully effective any structural elicitation phase must first be customised to an application and if necessary new types of structure with their own bespoke semantics elicited.

3.1 Structural Elicitation

Expert elicitation is a powerful tool when modelling complex problems especially in the common scenario when current probabilities are unknown and data is unavailable for certain regions of the probability space. Such methods are now widely developed,

well understood, and have been used to model systems in a variety of domains including climate change, food insecurity, and nuclear risk assessment (Barons et al.; Rougier and Crucifix, 2018; Hanea et al., 2006). Other methods for deriving the structure of a problem via the gamification of a system have been developed. Scenario testing has been shown to help refine the scope and context of a given problem (Vervoort et al., 2014; Lord et al., 2016). However, eliciting expert probabilities faithfully has proved to be a sensitive task, particularly in multivariate settings. First eliciting structure is critical to the accuracy of the model, particularly as conducting a probability elicitation is time and resource-intensive.

While there are several protocols for eliciting probability distributions such as the Cooke method, SHELF, and IDEA protocols (Cooke, 1991; O' Hagan and Oakley, 2014; Hanea et al., 2018; O'Hagan et al., 2006; Olaf, 2014), the process of determining the appropriate underlying structure has not received the same attention. Protocols for eliciting structural relationships between variables in the continuous range have been developed (Bedford and Cooke, 2001) and basic guidelines for eliciting a discrete Bayesian Network structure are available and well documented (Korb and Nicholson, 2010; Smith, 2010).

Borsuk and Reckhow (2001) describes a process for group elicitation that includes a section on structure and decomposition. One early example of elicited networks is the ALARM model (Beinlich et al., 1989). Other alternatives to the laborious process of expert elicitation include: automating the process by drawing from the literature (Nicholson et al., 2008). Causal machine learning algorithm CaMML has been used to incorporate diverse expert information (Flores et al., 2011).

These methods are widely applicable, but are rarely customised to structural elicitation of models other than the BN. However, it is possible to develop customising protocols to elicit structure, as illustrated through the case studies in this chapter.

3.1.1 Properties of Appropriate Structures

An appropriate model structure fulfils two criteria. Firstly, it should be compatible with how experts naturally describe a process. Ideally, modellers should agree on a structure using natural language. Assuming the domain experts and modeller have an agreed upon natural language description, the most appropriate model class may then be selected.

A compatible model should obey the temporal precedence established by the expert for the given context. The conditional probabilities of the problem should represent real possibilities in the given problem context. The elicited probabilities should represent the actual mechanism. That is, given two candidate models, and

one requires an additional layer of complexity to express the dynamics while another describes it outright, the second model should be selected.

Secondly, any structure should ideally have the potential to eventually be embellished through probabilistic elicitation into a full probability model. Sums and products should obey the axioms of probability. The probability of any of the events happening should sum to 1 for all elements of the graph. This should be rigorously checked using natural language questions posed to the experts.

It is often essential to determine that the structure of a problem as desired by a domain expert is actually consistent with the class of structural models considered. For a full structural elicitation, the domain experts must be shown the essential graph of the model class and confirm that the directionally ambiguous edges are appropriate. If they cannot confirm this, this may prompt a discussion of determining appropriate instrumental variables for each class. I will define instrumental variables for the CEG in Chapter 6.

The following sections demonstrate how particular model classes confirm or violate the properties described above. The logic and dynamics of Bayesian Networks (BN) often do not match with an experts' description of a problem. When this happens, the customising approach illustrated below generates flexible models that are a more accurate representation of the process described by the domain expert. We show that these alternative graphical models often admit a supporting formal framework and subsequent probabilistic model similar to a BN while more faithfully representing the beliefs of the experts.

Towards this end, this chapter explores examples of real case studies that are better-suited to eliciting bespoke structure. We illustrate how experts' natural language description of a problem can determine the structure of a model. Programs to alleviate food insecurity in the United States serve as a running example. Even within this domain, different problem dynamics are naturally more suited to particular structures, and eliciting these custom structures creates more compelling models. These bespoke structures can subsequently be embellished into customised probabilistic graphical models that support a full probabilistic description.

3.2 Eliciting Custom Structure

Structured expert elicitation begins with a natural language description of the problem from domain experts. An expert describes the components of a system and how they are related, and a structure often emerges organically. This process may be aided by the use of informal graphs, a widespread practice. However, the methods and diagrams used by the facilitators may not translate to full probability models.

Nevertheless, there are certain well developed classes of graphical models that do support this translation. The most popular and best supported by available software is the Bayesian Network. However, other graphical frameworks have emerged, each with its own representative advantages. These include event trees, chain event graphs, and dynamic analogues of these (Collazo et al., 2018; Barclay and Nicholson, 2015). This chapter describes some of the competing frameworks and suggest how one can be selected over another.

3.2.1 Choosing an Appropriate Structure

Choosing between candidate structures may not be straightforward. Some domain problems may be compatible with existing structures, while others might require creating new classes of probabilistic graphical models. The task of developing a bespoke graphical framework that supports a translation into a choice of probability models is usually a labour-intensive one requiring some mathematical skills. While some domain problems will require the modeller to undertake developing a customised model class, there are also several such frameworks already built, forming a tool-kit of different frameworks (Collazo et al., 2018; Smith, 1993; Figueroa and Smith, 2007; Liverani and Smith, 2015; Lauritzen and Richardson, 2002). This chapter gives guidelines below to help the modeller decide which of these methods most closely match the problem explanation given by the domain experts.

As a running example, the drivers of food insecurity will be considered. The illustrations used throughout the chapter are based on meetings with actual domain experts. I have simplified these case studies so that I can illustrate the elicitation process as clearly as possible. A meeting of advocates discusses the effect of food insecurity on long-term health outcomes. One advocate voices that food insecurity stems from insufficient resources to purchase food. The experts collectively attest that the two main sources of food are personal funds like disposable income or government benefit programs. The government benefit programs available to eligible citizen include child nutrition programs that provide free school breakfast, lunch, and after school snacks, the Supplemental Nutrition Assistance Program (SNAP), and Temporary Assistance for Needy Families (TANF). From this discussion among experts, modellers need to resolve the discussion into several key elements of the system. One potential set of elements drawn from the expert discussion is shown below:

- Government benefits, B: the rate at which a particular neighbourhood is participating in all available government programs
- Disposable Income, I: the average amount of income available for purchasing

food in the neighbourhood

- Food insecurity, F: the rate at which families and individuals in a neighbourhood experience insufficient access to food
- Long-term health outcomes, H: measured by an overall health index defined at the neighbourhood level.

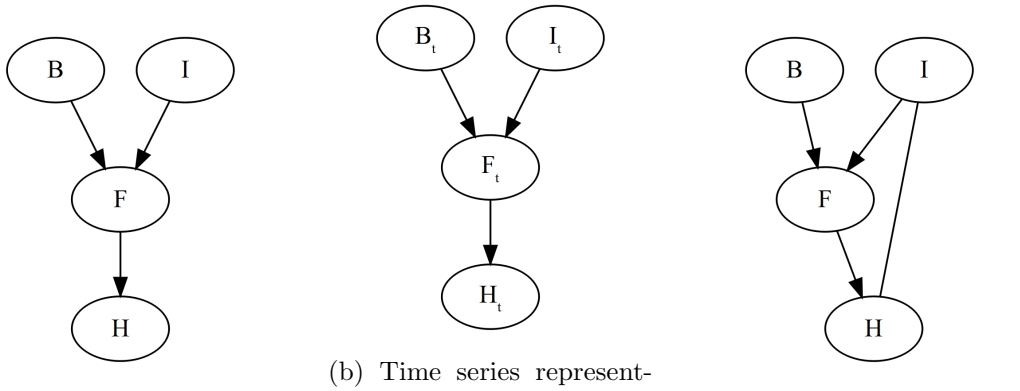
There are several guiding principles to help modellers create a structure that is faithful to the experts' description as shown in Figure 3.2.

Scope One common difficulty that appears in many structural elicitation exercises is the tendency of expert groups to think only in terms of measured quantities, rather than underlying drivers. Food insecurity and poverty researchers often consider elements of the system as documented for policy-makers, whereas those with a first hand knowledge of food insecurity may consider a different set of drivers, like personal trauma (Dowler and O'Connor, 2012; Chilton and Rose, 2009). Anecdotes of food insecurity may often draw out key, overlooked features of the system, but a well-defined problem scope is critical to prevent a drifting purpose. The responsibility of guiding the conversation continually toward general representations instead of off-the-shelf models falls to the facilitator.

Granularity Elicitations typically begin with a coarse description before refining the system. Considering refinements and aggregations can help the experts' opinions of the key elements of the system to coalesce. For instance, rather than modelling all the government benefits together in B, this variable could be removed and instead encompassed by two variables: child nutrition programs, C, and financial support for individuals S.

Because the experts are interested in the well being of the neighbourhood as a whole, it is sensible to model the problem with aggregate rather than individual benefits. The granularity of key elements depends on the modeller's focus. Thinking of the problem at different spatial levels may help to choose the appropriate granularity.

Potential interventions Another guiding principle during the structural elicitation is ensuring that possible interventions are represented by the system components. For instance, if the policy experts wanted to know what would happen after increasing all benefit programs simultaneously, modelling benefits collectively as B would be appropriate. But if they want to study what happens by intervening on child nutrition programs, then separating this node into C, child nutrition programs and



(a) BN of food insecurity at the neighbourhood level

(b) Time series representation of food insecurity drivers over time

(c) Hybrid representation of food insecurity drivers

Figure 3.1: Three different representations of the dependence structure between government benefits (B), disposable income (I), food insecurity (F), and long-term health outcomes (H), customised to the experts' beliefs.

allowing B to represent additional benefit programs would compose a more suitable model.

Context Dependence As the key elements of the system emerge, testing the structure by imagining these key elements in a different structure may either restrict or elucidate additional model features. The drivers that cause food insecurity at the neighbourhood level may vary greatly from those that provoke food insecurity at the individual household level.

For this running example, the experts focus on the neighbourhood level. They speak about each of the variables as the particular incidence rates for a neighbourhood. The modeller could then draw a dependence structure for random variables from their discussion about the dependence between these measurements. This structure would be most conducive to a Bayesian Network. An example of one tentative BN structure that has tried to accommodate these points in Figure 3.1a.

Importance of temporal processes Another key modelling decision is whether or not to use a dynamic network model. Are the experts speaking about potential interventions that are time-dependent or not? Do the key elements of the process change drastically over time? Few elements of a system are ever truly static, but dynamic models should only be chosen when the temporal element is crucial to the experts' description of the system as they are often more computationally intensive.

In contrast to the static example of measurements given above, suppose that the experts believe that yearly fluctuations in disposable income I directly affect the rates of food insecurity F. This is a dynamic process. Another expert might draw

on literature that shows the linear relationship between I and F. Using a standard Bayesian Network for this problem description would not capture the temporal information or the strength between each of the pairs of nodes. The quantities of the graph here are not static random variables, but rather its nodes appear to be representing processes. In this case, a more appropriate choice for the graphical elements would be to represent them as time series B_t, I_t, F_t, H_t. This graph is shown at a single time point in Figure 3.1b. The probabilistic model can be embellished into a number of different stochastic descriptions as will be discussed later in this chapter.

The meaning of the graph begins to change as the modeller learns more about the structure of a problem. This chapter suggests ways in which modellers could begin to frame different models for a desired context in terms of nodes and edges. Nodes for general graphical models can be any mathematical objects suitable to the given domain, provided that the system can be actually represented in terms of a probabilistic distribution which is consistent with the meaning ascribed to the model edges.

After establishing the nodes, the relationships between variables must be represented. These are usually expressed in terms of oriented edges or colourings in the vertices. Continuing with our toy example, the advocates promptly recognize that government benefits and disposable income directly impact the state of food insecurity. It also appears natural, as another expert attests, to associate the long-term health as dependent on food insecurity. These three relationships provide the graph in Figure 3.1a.

The experts comment that the available money for food purchasing directly affects how much food a family can buy, making directed edges a natural fit for B to F and I to F. However, the relationship between long-term health outcomes and disposable income is less clear. One advocate mentions that individuals and families who are battling chronic illness or faced with an outstanding medical bill are less likely to have disposable income, and thus more likely to be food insecure. However, using the typical BN machinery, adding an edge between long-term health outcomes and disposable income would induce a cycle in the graph and thus render the BN inadmissible.

One common solution would be to simply ignore this information and proceed only with the BN given previously. A second solution would be to embellish the model into a dynamic representation that could formally associate this aspect of the process by expressing instantaneous relationships in a single time slice of effects between nodes on different time slices. A time slice simply denotes the observations of the variables at a given time point. Another method might be to incorporate an

undirected edge that could be used to represent the ambiguous relationship between I and H. The result is a hybrid graph with undirected and directed edges with its own logic shown in Figure 3.1c.

Whatever semantic is chosen, edges should represent the experts' natural language description of the relationships. Returning to the instance in which the experts speak about food insecurity as a time series, the edges represent regression coefficients as the system unfolds. As shown below, directed acyclic graphs (DAGs) are particularly convenient for modelling. However, there are graphical representations that permit cycles, should the modeller wish to focus on the cyclic nature of F and H. The choice between the type and orientation of edge affects the semantics of the model as shown below.

3.2.2 Stating Irrelevancies and Checking Conditional Independence Statements

Suppose the domain experts' problem may be represented with a BN. Often, it is more natural for experts to impart meaning to the edges present in a graphical model. Unfortunately, it is the absence of edges that represent the conditional independences. To facilitate a transparent elicitation process, these conditional independence relationships can be expressed in a more accessible way as questions about which variables are irrelevant to the other.

Domain experts who are not statistically trained do not naturally read irrelevance statements from a BN. So it is often important to explicitly unpick each compact irrelevance statement written in the graph and check its plausibility with the domain expert.

Generally, suppose the domain expert believes that X is irrelevant for predicting Y given the measurement Z. That is, knowing the value of X provides no additional information about Y given information about Z. These beliefs can be written as $X \perp\!\!\!\perp Y \,|\, Z$, read as X is independent of Y conditional on Z.

For our example, the missing edges indicate three conditional independence relationships $H \perp\!\!\!\perp B \,|\, F$, $H \perp\!\!\!\perp I \,|\, F$ and $B \perp\!\!\!\perp I$. To check these, the modeller would ask the following questions to the domain expert:

- If I know what the food insecurity status is, does knowing what the disposable income is provide any additional information about long-term health?
- Assuming I know the food insecurity level, does the government benefit level offer any more insight into the long-term health of a neighbourhood?
- Does knowing disposable income levels of a neighbourhood provide further information about the government benefit level?

This last question might prompt the expert to realize that indeed, disposable income influences eligibility for government benefits, so an edge would be added between B and I.

These questions can also be rephrased according to the semigraphoid axioms, a simplified set of rules that hold for a given set of conditional independence statements. It is helpful to include these as they provide a template for different rule-based styles for other frameworks that capture types of natural language. More details can be found in Smith (2010).

The symmetry axiom is given in Definition 1 This axiom asserts that assuming Z is known, if X tells nothing new about Y, then knowing Y also provides no information about X. The second, stronger semi-graphoid axiom is called perfect composition (Pearl, 2014). This semi-graphoid axiom is equivalent to the decomposition, weak union, and contradiction axioms given in Definitions 2, 3, and 4, respectively.

Thus, for any four measurements X, Y, Z, and W:

Definition 20 ***Perfect composition*** *requires that for any four measurements X, Y, Z, and W:*

$$X \perp\!\!\!\perp (Y, Z) \,|\, W \Leftrightarrow X \perp\!\!\!\perp Y \,|\, (W, Z) \text{ and } X \perp\!\!\!\perp Z \,|\, W$$

Colloquially, assuming W is known, then if neither Y nor Z provides additional information about X, then two statements are equivalent. Firstly, if two pieces of information, Y and Z do not offer information about X, then each one on its own also does not help model X. Secondly, if one of the two is given initially alongside W, the remaining piece of information still does not provide any additional information about X. Further axioms are recorded and proved in Pearl (2009). For the purposes of elicitation, these axioms prompt common language questions which can be posed to a domain expert to validate a graphical structure. Given the values of the vector of variables in Z, learning the values of Y would not help the prediction of X. Note that translating this statement into a predictive model implies that $p(x \,|\, y, z) = p(x \,|\, z)$.

BNs encode collections of irrelevance statements that translate into a collection of conditional independence relationships. This can be thought of as what variable measurements are irrelevant to another. Relationships of the form $X \perp\!\!\!\perp Y \,|\, Z$ can be read straight off the graph as missing edges indicate conditional independence relationships. BNs obey the global Markov property, that each node is independent of its non-descendants given its parents (Pearl, 2009). By identifying the non-descendants and parents of each node, the entire collection of independence relationships is readily apparent. To see this in our example, consider the node representing long-term

health, H. In the BN in Figure 3.1a, $\{B, I\}$ are its non-descendants, and F is its parent, so then $H \perp\!\!\!\perp B \,|\, F$ and $H \perp\!\!\!\perp I \,|\, F$.

The independences can be read from the graph using the d-separation criteria. The conditional independence between three sets of variables A, B, and S is determined using d-separation[1]. Investigating d-separation from the graph requires inspecting the moralised ancestral graph of all variables of interest, denoted as $(\mathcal{G}_{\mathrm{An}(A\cup B\cup S)})^m$ (Pearl, 2009; Smith, 2010). This includes the nodes and edges of the variables of interest and all their ancestors. Then, the graph is moralised by drawing an undirected edge between all pairs of variables with common children in the ancestral graph. After disorienting the graph (replacing directed edges on the graph with undirected ones) and deleting the given node and its edges, the conditional independence between variables of interest can be checked. If there is a path between the variables, then they are dependent in the BN; otherwise they are independent.

Pearl and Verma (1995) proved the d-separation theorem for BNs, definitively stating the conditional independence queries that can be answered from the topology of the BN in Figure 3.1a. Lauritzen (1996) provided an alternative formulation of the d-separation criteria using the construction of an ancestral graph. This formulation of the ancestral construction is somewhat more intuitive as it highlights dependence structures due to shared ancestors.

The d-separation criteria and associated theorems formalize this process of reading off conditional independence relationships from a graph and is given in Theorem 8.

As an example, consider the BN of the drivers of food insecurity shown in Figure 3.1a. The d-separation theorem demonstrates that H is d-separated from B and I given the separating set F. In the moralized graph, F d-separates every path from the node H to a node in the set $\{B, I\}$. Thus, d-separation holds for any three disjoint subsets of variables in the DAG.

Separation theorems have been found for more general classes of graphs including chain graphs, ancestral graphs, and chain event graphs (Bouckaert and Studenỳ., 1995; Andersson, 2001; Richardson and Spirtes, 2002). Another class of graphical model, vines, weakens the notion of conditional independence to allow for additional forms of dependence structure (Bedford and Cooke, 2002). Another example of the use of these structures for Bayesian reference is given in Bedford et al. (2016). The results of the separation theorem for BNs can also be used to explore independence relationships in classes of graphs that are BNs with additional restrictions such as those imposed by the Multi-regression Dynamic Models (Smith,

[1]The d in d-separation stands for dependence-separation.

1993) and Flow Graphs (Figueroa and Smith, 2007).

When the structure is verified, it can then be embellished to a full probability model, provided it meets the original assumptions of our model. Understanding the relationship between the elicited conditional independence statements implied by the graph ensures equivalent statements are not elicited, thereby reducing the number of elicitation tasks. Even more importantly, the probabilities will respect the expert's structural hypotheses–hypotheses that are typically much more securely held than their numerical probability assessment.

In a discrete BN, this process involves populating the conditional probability tables with probabilities either elicited from experts or estimated from data. Alternatively, our food insecurity drivers example could be embellished to a full probability representation of a continuous BN. Discrete BNs will be populated by conditional independence tables that assign probabilities to all possible combinations of the values of each term in the factorised joint probability density. New computational approaches for continuous BNs allow for scalable inference and updating of the BN in a high-dimensional, multivariate setting (Hanea et al., 2006). The probabilities underpinning this model can be elicited using additional protocols and procedures from other chapters of (Bedford et al., 2020).

3.3 Examples from Food Insecurity Policy

3.3.1 Bayesian Network

Structural elicitation for a Bayesian Network is well studied (Smith, 2010; Korb and Nicholson, 2010). To see this process in action, consider a food insecurity example. The United States Department of Agriculture (USDA) administers the national School Breakfast Program (SBP), serving free or reduced price meals to eligible students.

A key element of the system is understanding the programmatic operations. Participation in SBP is not as high as it is for the school lunch program (Nolen and Krey, 2015). The traditional model of breakfast service involves students eating in the cafeteria before the beginning of school. Advocates began promoting alternative models of service to increase school breakfast participation. These include: Grab n Go, in which carts are placed through the school hallways and students select a breakfast item en route to class, or Breakfast in the Classroom, where all students eat together during the first period of the day. Only schools which have 80% of students eligible for free or reduced lunch are eligible for universal school breakfast. This means that breakfast is offered to every child in the school, regardless of their free or reduced status. This policy was implemented to reduce stigma of receiving a

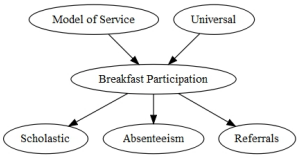
free meal.

The experts would also like to understand the effects of not eating breakfast. Advocates, principals, and teachers have hypothesized that eating a school breakfast impacts scholastic achievement. Food-insecure children struggle to focus on their studies. Experts posit that breakfast reduces absenteeism, as children and parents have the added incentive of breakfast to arrive at school. Some evidence suggests eating breakfast may also reduce disciplinary referrals, as hungry children are more likely to misbehave.

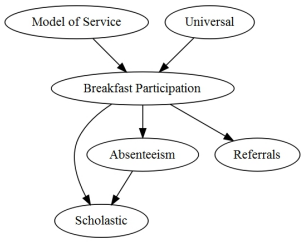
The data for this problem comes from a set of schools who are all eligible for universal breakfast, but some have chosen not to implement the program while others have. As universal breakfast status can be used as a proxy for socio-economic background of students attending a school, the population is narrowed to schools with low socio-economic status. The group of experts do not describe a temporal process here. They do not mention changes in breakfast participation throughout the school year, yearly fluctuations, or a time series of participation rates. Thus, it is natural for the modeller to begin with a BN approach. Given this information about breakfast, led by a facilitator, the modeller could consolidate the discussion into the following nodes:

- X_{m} Model of Service (Yes, No): indicates whether or not an alternative model of service as been implemented
- X_{u} Universal (Yes, No): indicates whether or not an eligible school has opted into universal service, as opposed to checking the economic status of the student at each meal
- X_{b} Breakfast Participation (High, Medium, Low): the binned participation rates at each school
- X_{s} Scholastic Achievement (High, Medium, Low): the standardized test score for each school
- X_{a} Absenteeism (High, Low): the binned absenteeism rate for the year
- X_{r} Disciplinary Referrals (High, Low): absolute number of disciplinary referrals

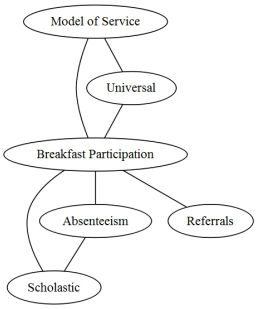
This list of nodes is focused on understanding the effects of school breakfast participation and specific type of breakfast service model. Certainly there are other reasons for absenteeism and disciplinary referrals besides whether or not a student had a good breakfast, but these are beyond the scope of this model. How can modellers determine the structure of this model from these measurable random variables? From



(a) The original BN representing the effects of model of service on breakfast participation and academic outcomes



(b) The BN represents the original BN with an edge added through the described verification process.



(c) The ancestral, moralized DAG of the central BN.

Figure 3.2: Exploring the conditional independence relationships expressed by the directed BN and its moralised analogue.

this set of nodes, the expert is queried about the possible relationship between all possible sets of edges. For instance, the modeller could ask, does knowing whether or not the school has opted into universal breakfast give any other information about whether or not the school has implemented an alternative breakfast model? In this case, the experts believe X_{m} does not give any additional information about X_{u}, because the program model is subject to approval from the cafeteria managers and teachers, whereas the decision to implement universal breakfast is primarily the decision of the principal. Thus no edge is placed between X_{m} and X_{u}. Both X_{m} and X_{u} are helpful in predicting X_{b}, so an arrow is drawn between each of these pairs. X_{s} is affected by X_{b}. These relationships can be seen in Figure 3.2a.

It is important to note that if the population of schools considered had included all schools rather than those with a low socio-economic status, then X_{u} would affect X_{s}, X_{a}, and X_{r} because universal school lunch would then be a proxy for low socio-economic status.

Suppose the domain experts know a school has a low breakfast rate, and they want information about their absenteeism. Will knowing anything about scholastic achievement provide any additional information about absenteeism? In order to check this with d-separation, the modeller may examine the ancestral graph $\mathcal{G}_{\mathrm{An}(X_{\mathrm{s}},X_{\mathrm{a}},X_{\mathrm{r}})}$, the moralised graph $(\mathcal{G}_{\mathrm{An}(X_{\mathrm{s}},X_{\mathrm{a}},X_{\mathrm{r}})})^m$ shown in Figure 3.2c. If there is not a path between X_{s} and X_{a}, then X_{s} is irrelevant to X_{a}. However, if there is a path between X_{s} and X_{a} that does not pass through our given X_{b}, then the two variables are likely to be dependent. Thus, the d-separation theorem checks the validity of the BN. The

symmetry property also imparts a set of equivalent questions. For instance, suppose the domain experts know a school has a low breakfast rate and they want to know information about their scholastic achievement. Will additional information about absenteeism be relevant to scholastic achievement? Asking such a question may prompt our group of experts to consider that students who miss classes often perform worse on exams. Revising the BN is in order, so the modeller add an additional edge from X_{a} to X_{s}. The BN in Figure 3.2a represents the beliefs of the domain experts. This encodes the following irrelevance statements:

- Knowing the model of service provides no additional information about whether or not the school district has implemented universal breakfast.
- The model of service provides no additional information about scholastic achievement, absenteeism, or referrals given information about the percentage of students who eat breakfast.
- Knowing absenteeism rates provides no additional information about disciplinary referrals given information about the breakfast participation rate.
- Knowing scholastic achievement rates provides no additional information about disciplinary referrals given that information about the breakfast participation and absentee rates.

When these irrelevance statements are checked, the domain experts realize that there is an additional link in that absenteeism affects scholastic achievements. Thus the modeller draws an additional arrow between X_{s} and X_{a} as shown in Figure 3.2b. The relationship between referrals and absenteeism is disputed in the literature and among experts, so, at least in this first instance, the modeller omits this edge.

Once the experts agree on the structure and verify it using the irrelevance statements, then the modeller may elicit the conditional distributions. Taken together, the BN represents a series of local judgements.

The joint probability mass function of the BN on the variables $\boldsymbol{X} = \{X_m, X_u, X_b, X_a\}$ given by Definition 2.1 is

$$p(\boldsymbol{x}) = p(x_{\mathrm{m}})p(x_{\mathrm{u}})p(x_{\mathrm{b}}|x_{\mathrm{m}}, x_{\mathrm{u}})p(x_{\mathrm{s}}|x_{\mathrm{b}}, x_{\mathrm{a}})p(x_{\mathrm{a}}|x_{\mathrm{b}})p(x_{\mathrm{r}}|x_{\mathrm{b}})$$

for this example.

Many of these distributions may be estimated by data, and unknown quantities may be supplied through structured expert elicitation. For instance, consider the sample question: what is the probability that scholastic achievement is high given

that breakfast participation rate is medium and the absentee rate is low? When the conditional probability tables are completed, the BN can be used to estimate effects of intervention in the system according to Pearl (2009).

3.3.2 Chain Event Graph

To illustrate an instance when a bespoke representation is more appropriate than the BN, consider the example of obtaining public benefits to address food insecurity. The USDA's Supplemental Nutrition Assistance Program (SNAP) provides funds for food to qualifying families and individuals through Electronic Benefit Transfer (EBT). Although 10.3% of Americans qualify for the program, Loveless (2010) estimates that many more citizens are eligible for benefits than actually receive them. Policy makers and advocates want to understand what systemic barriers might prevent eligible people from accessing SNAP. The application process requires deciding to apply, having sufficient documentation to apply (proof of citizenship, a permanent address), a face to face interview, and correct processing of the application to receive funds.

The structural elicitation phase includes speaking with domain experts to gather a reasonably comprehensive list of steps in the process. Domain experts include case workers, advocates, and individuals applying through the system. For our example, Kaye et al. (2013) collected this information through interviews at 73 community based organizations in New York State and categorized it according to access, eligibility, and benefit barriers. This qualitative information collection is crucial to developing an accurate model. From the qualitative studies, the key barriers were identified as:

- Face-to-face interviews not waived
- Same-day application not accepted
- Excessive documentation required
- Expedited benefit (available to households in emergency situations) not issued
- Failed to receive assistance with application documents
- Barriers experienced by special population: elderly and immigrant
- Ongoing food stamp not issued within 30 days
- EBT card functionality issues

The events selected should be granular enough to encompass the key points at which an applicant would drop out of the process, but coarse enough to minimize model complexity. An important part of the qualitative analysis process includes combining anecdotal evidence into similar groupings. For instance, the benefits office refused to waive the in-office interview for an applicant who did not have transportation to the application centre. In a separate instance, an interview was not waived for a working single mother with four children who could not attend because she was at work. While there are different contexts to each example, the central problem is the failure to waive the face-to-face interview. This type of node consolidation aids in reducing model complexity.

Discretising events can be a convenient way to clarify the model structure. Checking that the discretisation covers all possible outcomes from that event ensures that the model is an accurate representation of the problem. For our example, one possible discretisation with four variables of the problem is:

- X_{r}: At-risk population? (Regular, Elderly, Immigrant)
  - Regular: Households not part of an at-risk population
  - Elderly: Household head is over 65
  - Immigrant: Household head is a citizen, but immigration status of members of the household is uncertain
- X_{a}: Decision to apply (Expedited, Regular application, Decides not to apply)
  - Expedited: Same day applications, used in cases of emergency food insecurity
  - Regular application: The standard procedure
  - Decides not to apply: Eligible households who elect not to apply for a variety of reasons
- X_{v}: Application Verdict (Rejected, Accepted, Revision Required)
  - Rejected: Failed application, no possibility of resubmission
  - Accepted: Successful application
  - Revision required: Application must be resubmitted because of missing documentation, missed interview, or other reasons
- X_{e}: Utilizing an EBT card (Card successfully used for transactions, transaction errors)

- Card used for transactions: EBT arrives within the 30 day deadline and is successfully used at a grocery store
- Transaction errors: Card either does not arrive or returns an error at the grocery store

Figure 3.3 shows a simple BN approach to the natural language problem. Assume that the conditional independence relationships have been checked and that the modeller can now supply the conditional probabilities. Throughout this process, note that some of the probabilities are nonsensical. For example, the modeller must supply a probability for quantities like: the probability of having an accepted application given that the eligible citizen decided not to apply, and the probability of successfully utilizing EBT given that the application was rejected. This probability setting sounds absurd to elicit structurally, and will be distracting during the probability elicitation.

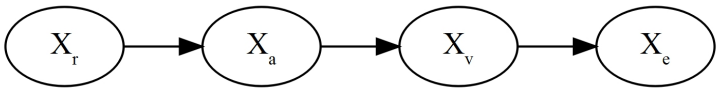


Figure 3.3: An inadmissible BN for the public benefits application process example.

The application process is difficult to coerce into a BN because the problem is highly asymmetrical. For instance, applicants with insufficient documentation will not have the chance to interview, and will not progress through the system. Now, if the natural language of the experts describes this process as a series of events, then these events have a natural ordering. Applicants must first decide to apply, then receive a verdict, and finally use their EBT card. The notion of being a member of an at-risk population does not have an explicit ordering, but the modeller can reasonably order it before the other events as it may affect how downstream events unfold.

Collazo et al. (2018) show that ordering demographic information at the beginning often coincides with higher scoring models during model selection for this class of graphs. Shafer (1996) has argued that event trees are a more natural way to express probabilistic quantities, so this problem may instead be expressed as an event tree in Figure 3.4 according to the framework given in Section 2.4.1. In this instance, there is an alternative graphical framework that provides a better way of accommodating the information provided by the expert.

As defined in Section 2.4.1, the nodes of our event tree are called situations $s_i \in S$ indexed according to temporal precedence; they represent different outcomes faced

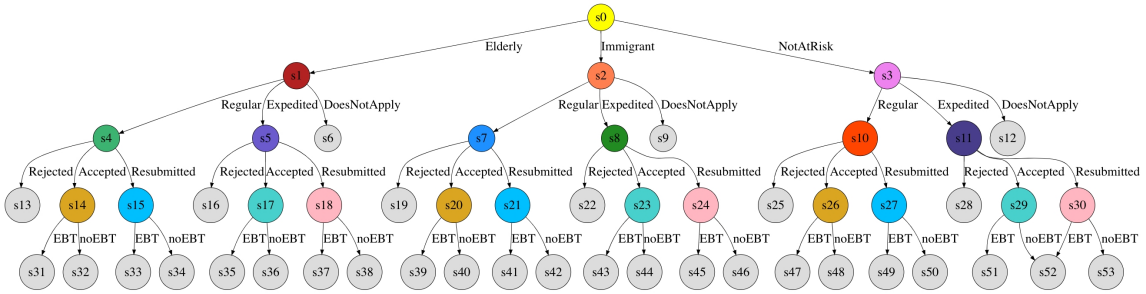


Figure 3.4: Event tree depicting the outcomes of the benefit application process.

by applicants travelling through the system. The edges represent the probabilities of different outcomes of each possible event occurring. The modeller can elicit the probability of observing a unit travelling down each edge of the tree, $\theta(e)$. The probability of a unit travelling down each of those edges should sum to one for each situation. The root-to-sink paths on the tree can be thought of as all possible outcomes of the application procedure. Situations with the same colour on the tree represent events whose outcomes have the same probabilities; these are in the same stage as defined in Definition14. In Figure 3.5, leaf nodes showing terminating outcomes are depicted in light grey.

The tree structure is naturally flexible just like the BN and can easily be modified to accommodate natural language suggestions. For instance, suppose the expert would like to add in a variable: the outcome of an interview process for regular applicants (the expedited process is waived.) Adapting the model simply requires adding two edges representing the outcome of the interview being successful or rejected to the set of situations in which an applicant applies through the regular route $\{s_4, s_7, s_{10}\}$. This simple adjustment in the tree structure would require adding a node to the BN as well as updating the conditional probability tables for the child nodes.

Another feature of the staged event tree structure is that the context specific independences are expressed directly in the tree structure. In this example, elderly applicants are often less likely to apply for benefits because the dollar amount is often too small a motivation for the perceived difficulty of the application. Immigrants are also less likely to apply because, although citizenship is required to apply for benefits, citizens with undocumented family members may fear citizenship repercussions of applying for assistance.

These context-specific probabilities are modelled through the colourings of the positions of the Chain Event Graph (CEG), rather than requiring separate BN

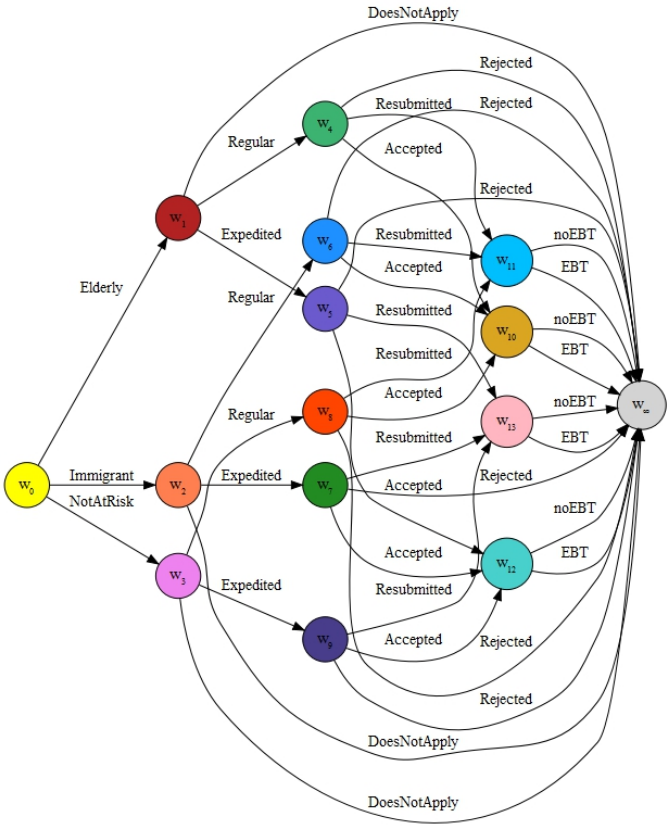


Figure 3.5: Chain Event Graph representation of the benefits application process.

models with context-specific conditional independence relationships. Conditional independence relationships can be read from the graph through the stage structure. Two positions are in the same stage if they are the same colour. In order to draw a condensed representation of the graph, define positions $w_k \in \mathbb{W}$ as in Definition 15. This allows merging the stages for a more compact chain event graph representation, called the Chain Event Graph (CEG), depicted in Figure 3.5.

In the same spirit as the Markov condition for BNs, a result for the CEG can read statements of the form 'the immediate future is independent of the past given the present.' Given that a unit reaches a position, what happens afterwards is independent not only of all developments through which it was reached, but also of the positions that logically cannot happen. These conditional independence statements can be read off the graph just as they can for BNs. Illustrating this process requires new definitions about certain sets of positions in the CEG.

Definition 21 *A set of positions $W' \subseteq \mathbb{W}$ is a **fine cut** if disjoint union of events centred on these vertices is the whole set of root-to-leaf paths.*

That is, none of the positions $w \in W'$ are up- or downstream of another and all of the root-to-sink paths on $\mathcal{C}$ must pass through one of the positions in W'.

Definition 22 *A set of stages $u \in \mathbb{U}$ denoted $W' \subseteq \mathbb{U}$ is a **cut** if the set of positions in the colouring $w \in u | u \in \mathbb{U}$ is a fine cut.*

The definitions of fine cut and cut help to differentiate the 'past' from the 'future' in the graph.

A cut-variable denoted X_W can be thought of as an indicator variable used to define the edges a unit passes through in the present.

Definition 23 *The **cut-variable** X_W is the corresponding set of positions W in a cut or a fine cut and X_W is measurable with respect to the probability space defined by the CEG.*

The past and future can be defined as a vector of random variables whose vertices are located upstream or downstream. Denote the 'past' random variables as

$$Y_{\prec W} = (Y_w | w \text{ upstream of } W)$$

and the 'future' by

$$Y_{W \preceq} = (Y_{w'} | w' \text{ downstream of } W).$$

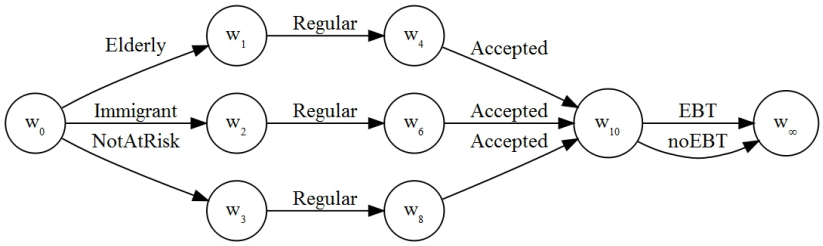
Defining the random variables in a CEG sets up this formal definition of conditional independences in a CEG:

Theorem 24 *Let $\mathcal{C} = (\mathbb{W}, F)$ be a CEG and let $W' \subseteq \mathbb{W}$ be a set of positions then for any cut-variable $X_{W'}$, we find:*

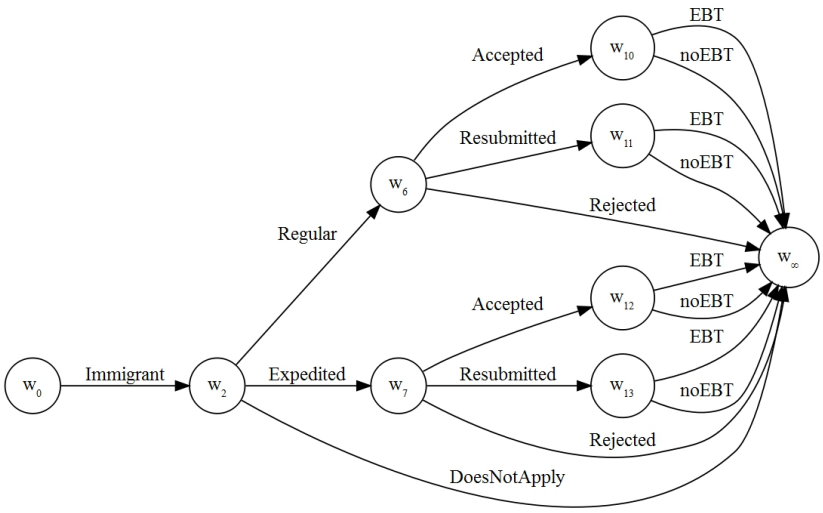
1. *If W' is a fine cut then $Y_{\prec W'} \perp\!\!\!\perp Y_{W' \preceq} | X_{W'}$.*
2. *If W' is a cut then $Y_{\prec W'} \perp\!\!\!\perp Y_{W'} | X_{W'}$.*

Proof can be found in Smith and Anderson (2008).

Theorem 24 explains how to read conditional independence from the CEG structure. The next step is to validate the structure. Just as for the BN, natural language questions from the semigraphoid axioms elucidate the conditional independence relationships. At each cut, consider the conditional independence between each pair of upstream and downstream variables. For instance, given that eligible applicants apply for benefits, does knowing whether or not they are part of an at-risk population provide any additional information about whether or not they apply for expedited benefits? By perfect decomposition, does knowing that the candidate received application assistance provide any information about whether or not they



(a) A pseudo-ancestral CEG representing an independence between the query.



(b) A pseudo-ancestral CEG representing a dependence between the query.

Figure 3.6: Two uncoloured pseudo-ancestral CEGs

will receive the electronic benefits given that they had the correct documentation and passed the interview? Does knowing that they had application assistance provide any additional information about whether or not they passed the interview given that they had the correct documentation? These queries validate the model and may prompt further adaptations.

In the BN, Theorem 8 provides a systematic way to check all of the conditional independence relationships. Thwaites and Smith (2015) proposed a new *d*-separation theorem for simple, uncoloured CEGs. The full d-separation theorem for CEGs will be discussed in Chapter 4. In a BN, the ancestral graph helps to address these queries. The analogue of the ancestral graph for the CEG is given in Chapter 4. Thwaites and Smith (2015) derived the precursor to the full ancestral graph of the CEG– the pseudo-ancestral representation. Pseudo-ancestral graphs depict the nodes

of interest and all the upstream variables, consolidating the downstream variables. Moralizing the graph in a BN corresponds to removing the colourings of the CEG. The examples in this chapter focus on the pseudo-ancestral representation, and the development of the full ancestral graph will be shown in Chapter 4.

Is the ability to complete a transaction on the EBT card independent of whether the applicant is a member of an at-risk population given that they completed a successful regular application? The pseudo-ancestral graph as seen in Figure 3.6a, shows the probability that $\Lambda = \{\text{Regular, Accepted}\}$. Being a part of the at-risk population is independent of being able to utilize an EBT card because all the possible pathways must pass through w_{10}, identifying it as a single vertex composing a fine cut.

On the other hand, testing the independence of the application verdict from the selected method of application for at-risk immigrant population can be done with the ancestral graph in Figure 3.6b. These are not independent because there is no single vertex composing a fine cut.

One of the strengths of the CEG model is that it does not require any algebra, but instead can be elicited entirely using coloured pictures. CEGs are of particular use for problems that exhibit some asymmetry. After validating the structure, populating the model with data or elicited probabilities provides a full statistical model that can be used for inference, details can be found in Collazo et al. (2018). The CEG offers a class of models that is more general than BNs, enabling modellers to represent context-specific independences. The model can also incorporate asymmetries as seen in our non-stratified example.

The CEG is a powerful model particularly well-suited to expert elicitation, as experts often convey information in a story, which naturally expands to an event tree.

3.3.3 Multi-regression Dynamic Model

Our next two examples of customised classes of graphical models consider the problem of assessing participation in the Summer Meals Program (SMP). SMP meal sites are designated as either open or closed. Open sites do not have a set population like in a school or particular program, but rather are open to the public and thus dependent on walk-ins for the bulk of participation.

Although the need in the summer is severe, participation in the program remains relatively low. Advocates generally agree that the two biggest obstacles to program participation are a lack of awareness about the program, and unavailable transportation to the site. These factors affect meal participation which fluctuates throughout the three months of summer holidays. Available data for meal parti-

(a) The correct summary MDM graph (b) An inadmissible MDM summary graph

Figure 3.7: Two DAGs with equivalent BN representations, but unique Multi-regression Dynamic Model representations

cipation records how many meals were served through the program at each day for about three months in the summer. Transportation data records the number of available buses. Awareness can be measured through texting data that records when participants queried a government information line to receive information about where the closest sites serving meals are. The Texas Department of Agriculture collected text records for all phone calls made to the number in the summer of 2013. Figure 3.9 incorporates this data into a dynamic linear model.

Advocates would most like to capture the effect that awareness of SMP has on available transportation, and that transportation in turn has on meal participation. To simplify the elicitation, additional obstacles like low summer school enrolment, poor food quality, and insufficient recreational actives are not considered as primary drivers of meal participation levels. The relationship between awareness and available transportation is well documented, as is the relationship between transportation and meal participation (Wilkerson and Krey, 2015).

The advocates emphasize drastic shifts in awareness, transportation, and meal participation throughout the summer months. On public holidays and weekends, there is a lack of public transportation and a corresponding sharp decline in meals. This temporal aspect of the problem prompts the modeller to consider a time series representation as the most natural class of graphical model.

To emphasize the importance of selecting a time series representation over a BN, consider the limitations of the standard BN model. Suppose the advocates agree on the general structure shown in the DAG in Figure 3.7a, as children and parents must know about the meal before they take transportation to the meal. Then in turn, they must travel to the meal before receiving the meal. However, if the graph is interpreted as a BN, then Figure 3.7a only encodes the conditional independence relationship $M \perp\!\!\!\perp A \,|\, T$, which does not capture the ordering expressed by the advocates. To further stress this point, Figure 3.7b shows a DAG with the reverse ordering that encodes equivalent conditional independence relationships when interpreted as a BN. As shown below, if these are summary graphs of MDMs whose edges represent the strengths given in the model definition in Definition 19, then the models are distinguishable.

The experts remark that a media campaign and corresponding surge in awareness prompts a corresponding increase in the number of people travelling to meal sites. These aspects of the problem, taken with those discussed above prompt a consideration of each of the elements as time series. In order to capture the linear relationship between variables that the experts have expressed, the edges of the graph correspond to regression coefficients between each parent and child node.

Assuming linear relationships exist between awareness and transportation and transportation to the meal site and actual participation, the system can be described as regressions in a time series vector $\boldsymbol{Y}_t = \{Y_t(1), Y_t(2), Y_t(3)\}$. Let the time series of the key measurements denote awareness by $Y_t(1)$, available transportation by $Y_t(2)$, and summer meals participation by $Y_t(3)$. This model corresponds to another example from our toolbox of alternative representations: the Multi-regression Dynamic Model, the general definition of which is given in Definition 19.

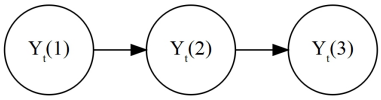
This means that $(\boldsymbol{Y}_t(r)|\boldsymbol{Y}^{t-1}, \boldsymbol{F}_t(r), \boldsymbol{\theta}_t(r))$ follows some distribution with mean $\boldsymbol{F}_t(r)^{t\prime}\boldsymbol{\theta}_t(r)$ and variance $V_t(r)$.

Modelling this behaviour requires dynamic linear models in which the parents are the regression coefficients for each series. For our example in Figure 3.7a, the system and observation model equations are:

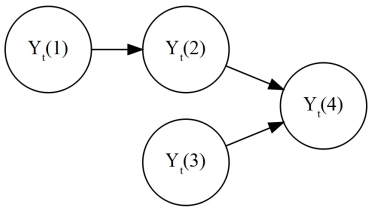
$$\begin{aligned}
\boldsymbol{\theta}_t(1) &= \boldsymbol{\theta}_{t-1}(1) + w_t(1) & Y_t(1) &= \theta_t^{(1)}(1) + v_t(1) \\
\boldsymbol{\theta}_t(2) &= \boldsymbol{\theta}_{t-1}(2) + w_t(2) & Y_t(2) &= \theta_t^{(1)}(2) + \theta_t^{(2)}(2)Y_t(1) + v_t(2) \\
\boldsymbol{\theta}_t(3) &= \boldsymbol{\theta}_{t-1}(3) + w_t(3) & Y_t(3) &= \theta_t^{(1)}(3) + \theta_t^{(2)}(3)Y_t(2) + v_t(3)
\end{aligned}$$

The strengths of the parents are given by the regression coefficients $\theta_t^{(2)}(2)$ for $Y_t(2)$ and $\theta_t^{(2)}(3)$ for $Y_t(3)$. The initial information $\{\boldsymbol{\theta}_0\}$ can be elicited from the domain experts or taken from previous data observations.

Suppose after the experts agree on the structure, the modeller examines the one step ahead forecasts, and notices errors on some days. Examining these days might prompt the experts to recognize that the days of interest correspond to days with a heat advisory. They suggest that the heat index throughout the summer also affects meal participation. This structural change can be quickly integrated into the system by adding observation, system equations, and initial information to represent the new model feature and updating the system for all downstream nodes. The system equations and the initial information is given in Equation 19. Because the ordering in the MDM is strict, and the heat index is a parent of meal participation, meal participation is relabelled as $Y_t(4)$ and the heat index as its parent $Y_t(3)$.



(a) A summary MDM graph with series representing awareness $Y_t(1)$, transportation $Y_t(2)$, and meal participation $Y_t(3)$ respectively.



(b) A MDM summary graph with series representing awareness $Y_t(1)$, transportation $Y_t(2)$ meal participation $Y_t(4)$ and the new heat index variable $Y_t(3)$

Figure 3.8: A summary MDM graph after refining elicitation with experts including the original variables plus a new series with the heat index.

$$\begin{aligned}
\boldsymbol{\theta}_t(4) &= \boldsymbol{\theta}_{t-1}(4) + w_t(4) \\
Y_t(3) &= \theta_t^{(1)}(3) + v_t(3) \\
Y_t(4) &= \theta_t^{(1)}(4) + \theta_t^{(3)}(4)Y_t(2) + \theta_t^{(2)}(4)Y_t(3) + v_t(3)
\end{aligned}$$

In this new model, the regression coefficients $\theta_t^{(3)}(4)$ and $\theta_t^{(2)}(4)$ for meal participation $Y_t(4)$ indicate the strengths of the edges in the summary graph in Figure 3.8.

In this way, the natural language expressions of the domain experts can be used to adjust the model.

Generally, particular observations of $Y_t(r)$ are denoted as $y_t(r)$. The MDM ensures two critical conditional independence relationships. The first holds that if

$$\perp\!\!\!\perp_{r=1}^{n} \boldsymbol{\theta}_{t-1}(r)|\boldsymbol{y}^{t-1} \tag{3.1}$$

then

$$\perp\!\!\!\perp_{r=1}^{n} \boldsymbol{\theta}_{t}(r)|\boldsymbol{y}^{t} \tag{3.2}$$

where $\boldsymbol{y}^{t-1}(i) = \{y_1(i), \ldots, y_{t-1}(i)\}$ and

$$\boldsymbol{\theta}_t(r) \perp\!\!\!\perp Y^t(r+1), \ldots, Y^t(n)|Y^t(1), \ldots, Y^t(r) \tag{3.3}$$

Equation 3.2 demonstrates that the parameters$\{\boldsymbol{\theta}_{t-1}(r)\}$ are independent of each other given the past data $\{\boldsymbol{y}^{t-1}\}$ then $\{\boldsymbol{\theta}_t(r)\}$ is also independent of $\{\boldsymbol{y}^t\}$. Given the initial parameters $\{\boldsymbol{\theta}_0(r)\}$ are independent, then they remain independent as the series unfolds by induction according to Smith (1993).

In the summer meals example, the experts confirm that the independence

between awareness, transportation, and meal participation is independent. That is, $\boldsymbol{\theta}_0(1) \perp\!\!\!\perp \boldsymbol{\theta}_0(2) \perp\!\!\!\perp \boldsymbol{\theta}_0(3)$. Awareness is measured by the amount of public media generated, transportation is a measure of public transportation available, and the participation rate is the number of meals served every day in the summer. The domain experts agree that these can be independent of each other. Additionally, Equation 3.3 ensures the following conditional independence relationships:

$$\boldsymbol{\theta}_t(1) \perp\!\!\!\perp \{y^{t-1}(2), y^{t-1}(3)\} | y^{t-1}(1)$$
$$\boldsymbol{\theta}_t(2) \perp\!\!\!\perp y^{t-1}(3) | \{y^{t-1}(1), y^{t-1}(2)\}$$

An analogue of the d-separation theorem for MDMs identifies part of the topology of the graph that ensures that these conditional independence statements hold.

Theorem 25 *For MDM $\{\boldsymbol{Y}_t\}$ if the ancestral set $\boldsymbol{x}_t(r) = \{y^t(1), \ldots, y^t(r)\}$ d-separates $\boldsymbol{\theta}_t(r)$ from subsequent observations $\boldsymbol{z}_t(r) = \{y^t(r+1), \ldots, y^t(n)\}$ for all $t \in T$, then the one-step ahead forecast holds :*

$$p(\boldsymbol{y}_t | \boldsymbol{y}^{t-1}) = \prod_r \int_{\boldsymbol{\theta}_t(r)} p\{\boldsymbol{y}_t(r) | \boldsymbol{x}^t(r), \boldsymbol{y}^{t-1}(r), \boldsymbol{\theta}_t(r)\}\, p\{\boldsymbol{\theta}_t(r) | \boldsymbol{x}^{t-1}(r), \boldsymbol{y}^{t-1}(r)\} d\boldsymbol{\theta}_t \tag{3.4}$$

Proof. Consider the contrapositive: if the one-step ahead forecast does not hold, then the ancestral set $\boldsymbol{x}_t(r)$ must not d-separate $\boldsymbol{\theta}_t(r)$ from $\boldsymbol{z}_t(r)$. If the form of Equation 3.4 does not hold, then either the first term $p\{\boldsymbol{y}_t(r) | \boldsymbol{x}^t(r), \boldsymbol{y}^{t-1}(r), \boldsymbol{\theta}_t(r)\}$ or the second term $p\{\boldsymbol{\theta}_t(r) | \boldsymbol{x}^{t-1}(r), \boldsymbol{y}^{t-1}(r)\}$ must depend on $\boldsymbol{z}_t(r)$. This would violate the structure of the MDM, inducing arrows between $\boldsymbol{\theta}_t(r)$ and $\boldsymbol{z}_t(r)$. These new arrows violate the d-separation condition. ■

This one step ahead forecast factorises according to the topology of the graph, allowing an examination of the plots of each of the series. For this example, the one step ahead forecast factorises:

$$\begin{aligned} p(\boldsymbol{y}_t | \boldsymbol{y}) = & \int_{\boldsymbol{\theta}_t(1)} p\{\boldsymbol{y}_t(1) | \boldsymbol{y}^{t-1}(1) \boldsymbol{\theta}_t(1)\}\, p\{\boldsymbol{\theta}_t(1)\} d\boldsymbol{\theta}_t(1) \\ & \times \int_{\boldsymbol{\theta}_t(2)} p\{\boldsymbol{y}_t(2) | \boldsymbol{y}(1)^t, \boldsymbol{y}^{t-1}(2), \boldsymbol{\theta}_t(2)\}\, p\{\boldsymbol{\theta}_t(2) | \boldsymbol{y}^{t-1}(1), \boldsymbol{y}^{t-1}(2)\} d\boldsymbol{\theta}_t(2) \\ \times \int_{\boldsymbol{\theta}_t(3)} & p\{\boldsymbol{y}_t(3) | \boldsymbol{y}(1)^t, \boldsymbol{y}(2)^t, \boldsymbol{y}^{t-1}(3), \boldsymbol{\theta}_t(3)\}\, p\{\boldsymbol{\theta}_t(3) | \boldsymbol{y}^{t-1}(1), \boldsymbol{y}^{t-1}(2), \boldsymbol{y}^{t-1}(3)\} d\boldsymbol{\theta}_t(3) \end{aligned}$$

Examining plots of the errors of each forecast can help determine what further structural adjustments should be made. For instance, in the Figure 3.9, awareness

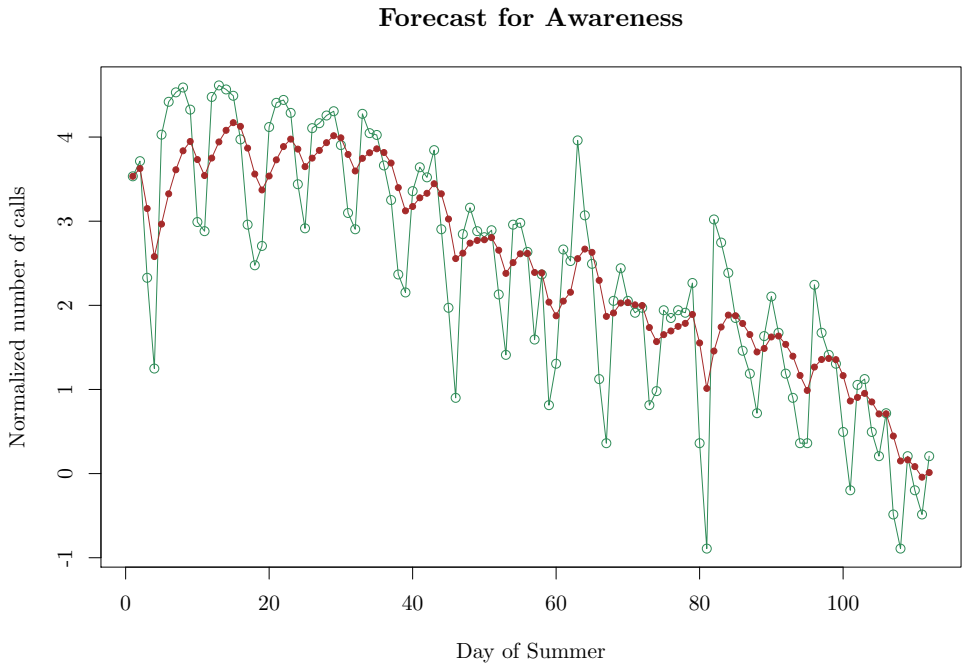


Figure 3.9: The logarithmic plot of awareness (as measured by calls to ask for meal site locations) throughout the summer months. The open green dots are actual observations; the filled brown dots are the one step ahead forecast.

has a cyclical nature, as people are less likely to text for an address of a meal site on weekends and holidays. This model can be adapted to include seasonal shifts using the equations from West and Harrison (1997).

The implementation of this problem as an MDM rather than a BN maintains the strength of the relationships between each series and its regressors, respecting the natural language expression of the system by the domain experts. An additional feature of the MDM is that this representation renders the edges causal in the sense carefully argued in Queen and Albers (2009). For our model, note that while the two DAGs in Figure 3.7 both represent $A_t \perp\!\!\!\perp M_t | T_t$, and are thus indistinguishable, the arrows in the MDM representation are unambiguous. The causal implications of this are developed in Chapter 6. The MDM offers a dynamic representation of a system in which the regressors influence a node contemporaneously.

3.3.4 Flow Graph

Structures can be adapted to meet additional constraints, such as conservation of a homogeneous mass transported in a system. However, these constraints motivate employing yet another graph with different semantics to transparently express the expert structural judgements. To illustrate how to derive this from a natural language expression of a problem, consider the following example from the Summer Meals Program (SMP).

SMP provides no-cost meals to children under 18 at schools and community-based organisations during the summer months. SMP relies on food being procured from vendors, prepared by sponsors, and served at sites. Participation in the program is low, nationally 15% percent of eligible children use the program (Gundersen et al., 2011). Sponsors, entities who provide and deliver meals, are reimbursed at a set rate per participant, but sponsors often struggle to break even. One of the key possible areas for cost cutting is the supply chain of the meals. Community organisers hypothesize different interventions on each of these actors might help make the program more sustainable such as:

- A school district serving as a sponsor (Austin ISD) is having trouble breaking even. What happens when they partner with an external, more financially robust sponsor (City Square) to provide meals to the school. What is the effect on the supply chain of meals to the Elementary and Intermediate schools?
- Several smaller sponsors (among them the Boys and Girls Club) are having trouble breaking even and decide to create a collective to jointly purchase meals from a vendor (Revolution Foods). How does the presence of the new collective alter the flow of meals to the two Boys and Girls Club sites?

- Two sites, say apartment complexes A and B are low-performing, and the management decides to consolidate them. What is the long-term effect on a system?
- What happens when a sponsor, City Square, changes vendors from Revolution Foods to Aramark?
- What happens when one sponsor, Austin ISD, no longer administers the program and another sponsor, Boys and Girls Club takes responsibility for delivering food to the Intermediate and High Schools?

Hearing the domain expert describe what types of intervention they would like to be able to model can elucidate the critical elements of the structure. In this example, the effect of the supply and transportation of meals through the network is key to the types of behaviour the modeller hopes to capture. This problem can be framed as a set quantity of meals moving through the system. Key model assumptions must always be checked with the domain expert. In this case, one of the key assumptions is that the number of children who are in need of meals and are likely to attend the program is relatively stable throughout the summer. This is a reasonable assumption, particularly when modelling a set population such as students in summer school or extracurricular programming. Community advocates verify that the assumption is reasonable because all of these sites and sponsors need a relatively set population in order to break even on the program.

Additionally, to estimate the effect of the addition or removal of actors in the system, it is important to assume that the number of meals for children in need is conserved. Thus, if a sponsor and subsequent sites leave the program, then those children will access food at another sponsor's meal sites, provided transportation is available. This assumption permits modelling particular interventions of interest, where combining, removing, or adding actors to the system is of particular interest. The dynamics of this particular problem involve the switching of ownership–what happens when the path flow of meals through the system changes–either a sponsor buys a meal from a different vendor, or a site turns to a different sponsor to supply their meals. This is a key component of the problem, but unfortunately it renders the problem intractable for the BN as shown below. However, Figueroa and Smith (2007) discovered a methodology for re-framing this problem as a tractable variant of a BN that simultaneously remains faithful to the dynamics of the problem described above (Figueroa and Smith, 2007).

Modelling the process as a BN begins with identifying the actors involved. A scenario for the key players in the city of Austin, Texas may consist of the following players at the vendor, sponsor, and site level. Levels are denoted by $z(i, j)$ where i

indicate the level (vendor, sponsor, or site), and j differentiates between actors on a particular level. In this example the players are:

$z(1,1)$ Revolution Foods

$z(1,2)$ Aramark

$z(2,1)$ City Square

$z(2,2)$ Austin Independent School District

$z(2,3)$ Boys and Girls Club

$z(3,1)$ Apartment complex A

$z(3,2)$ Apartment complex B

$z(3,3)$ Elementary School

$z(3,4)$ Intermediate School

$z(3,5)$ High School

$z(3,6)$ Boys and Girls Club site A

$z(3,7)$ Boys and Girls Club site B

These actors compose the nodes of the network; the edges represent the flow of meals between entities. For instance, vendor Aramark $z(1,2)$ prepares meals for sponsors at Austin ISD, $z(2,2)$, who in turn dispenses them at the Intermediate School, $z(3,4)$. Domain experts assume that each day, a set number of meals runs through the system. This list of actors can be readily obtained from natural language descriptions of the problem. Eliciting this information would simply require the modeller to ask the domain experts to describe the flow of meals through each of the actors in the system. This structural elicitation and resultant graph in Figure 3.10 are transparent to the expert, an advantage of customised modelling.

As the modeller begins to check the relationships encoded in the graphical model elicited in Figure 3.10, the missing edges between actors in a given level means that each of the sponsors is unaffected by the meals being transported to and from the other sponsors. However, this is not realistic for closed sites because knowing the number of meals served at all but one sponsor gives perfect information about the remaining sponsor, as the number of meals served by sponsors remains constant. For instance, knowing how many meals are prepared by Aramark, $z(1,1)$, provides perfect information about how many are prepared by Revolution Foods, $z(1,2)$, because meals are conserved at each level, implying a directed line from $z(1,1)$ to $z(1,2)$. Modelling this process graphically, as in Figure 3.10, induces severe dependencies in the network when the process is modelled as a BN. Thus, the problem as the experts have expressed it cannot be represented as a BN.

Decomposing the information in Figure 3.10 to into paths as shown in Figueroa and Smith (2007), admits a representation as a Dynamic Bayesian Networks. Denote

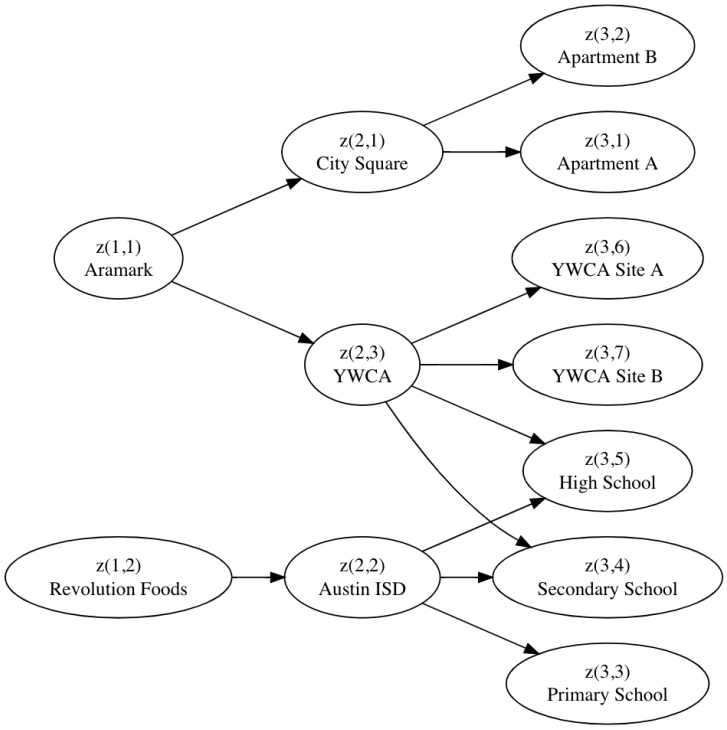


Figure 3.10: Flow Graph showing transfer of meals from vendors $z(1, j)$, to sponsors $z(2, j)$, to sites $z(3, j)$.

$\boldsymbol{\phi}'_t[l] = (\phi_t(l, 1), \phi_t(l, 2), \ldots, \phi_t(l, n_l))$, where $l = \{1, 2, 3\}$ as the node states vector for each of the three levels, where $\phi_t(l, j_l)$ represents the mass owned by player $z(l, j_l)$ during time t. This probabilistic representation allows the modeller to retain the advantages of the clear representation in Figure 3.10 to draw information about the system from the experts as well as the computational convenience of the BN machinery.

The full methodology for translating the hierarchical Flow Graph to the dynamic Bayesian Network (DBN) representation is given in Figueroa and Smith (2007), this chapter simply states the elements of the model that would need to be a part of the probability elicitation. Information about the numbers of meals held by each entity at each day during the summer can be represented by a time series vector $\boldsymbol{X}'_t = (\boldsymbol{X}'_t[1], \boldsymbol{X}'_t[2], \boldsymbol{X}'_t[3])$, representing the number of meals at the vendor, sponsor, and site levels respectively. Next, the paths of meals travelling from vendor to meal site are represented as aggregates of the product amounts. The paths in this

diagram are:

$$\begin{aligned}
\pi(1) &= \{z(1,1), z(2,1), z(3,2)\} & \pi(2) &= \{z(1,1), z(2,1), z(3,1)\} \\
\pi(3) &= \{z(1,1), z(2,3), z(3,6)\} & \pi(4) &= \{z(1,1), z(2,3), z(3,7)\} \\
\pi(5) &= \{z(1,1), z(2,3), z(3,5)\} & \pi(6) &= \{z(1,1), z(2,3), z(3,4)\} \\
\pi(7) &= \{z(1,2), z(2,2), z(3,5)\} & \pi(8) &= \{z(1,2), z(2,2), z(3,4)\} \\
\pi(9) &= \{z(1,2), z(2,2), z(3,3)\}
\end{aligned} \tag{3.5}$$

Fully embellishing this model involves eliciting the core states, the underlying drivers of the number of meals passing through each of the actors. These can be readily adapted to reflect the beliefs of different domain experts. For instance, different school districts often follow different summer school schedules, so if the advocates were interested in applying the model to a different region, it would simply require updating the core state parameters. The information about the path flows is most readily supplied through available data about the number of meals prepared, transported, and served throughout the summer.

As with the MDM, the conditional independence relationships can be read from the model. The dynamic linear model is essentially a Markov chain, so checking the flow of items in the network only depends on the previous iteration. If not, then the model must be adapted to express a Markov chain with memory. Furthermore, validating the structure requires checking that the past observations of how much stuff is in the model at each level are independent of future amounts given all of the governing state parameters for that particular time-step. The one-step ahead forecast allows a structural check similar to that of the MDM.

3.4 Discussion

The case studies in Section 3.3 show how drawing the structure from the experts' natural language description motivates the development of more flexible models that can highlight key features of a domain problem. The SBP example shows that a BN is appropriate when the expert describes a problem as a set of elements that depend on each other. The SNAP application example highlights the advantages of a tree-based approach when the experts describe a series of events and outcomes. The open SMP example shows how additional restrictions on the BN structure can draw out the contemporaneous strengths between elements of the model that is crucial to the experts' description. Lastly, the flow of meals in a system shows how working with the accessible representation of meal flow in a system can be translated into a valid structure while remaining faithful to the assumptions expressed by the expert.

A summary table is shown in Table 3.1 citing additional examples of applications of these bespoke graphical models examined in this chapter. References are given for two classes of models, chain graphs and regulatory graphs that are not explored in this chapter. This is of course a small subset of all the formal graphical frameworks now available. These case studies and applications in the table are examples from possible customised models.

Generally, allowing these representations to capture dynamics uniquely to a given application cultivates more suitable representations. Just as the d-separation theorem articulates the conditional independence relationships in the BN, analogous theorems elucidate the dependence structure of custom representations. Each of these examples of elicited structure has its own logic which can be verified by examining the conditional independence statements and confirming with the expert that the model accurately conveys the expert's beliefs.

Carefully drawing structure from an expert's natural language description is not an exact science. This chapter offers a few guidelines for when to use particular models summarised in the flow chart in Figure 3.11. The examples discussed here are far from exhaustive and Figure 3.11 also highlights areas of open research. Spirtes and Zhang (2016) confirms that determining what new classes of models might be more appropriate than a BN for a given domain. A full protocol for choosing one customising model over another remains to be formalised. While software for BN elicitation is ubiquitous, robust software for these alternative models is under development.

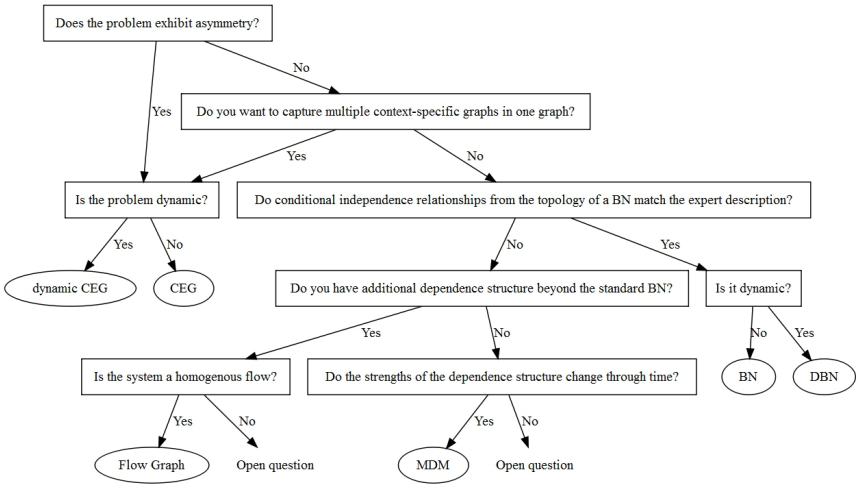


Figure 3.11: Flow chart to guide picking an appropriate structure.

The premise of drawing the structure from a natural language description rather than tweaking a model to fit an existing structure represents a substantial

Name	Description	When to use	Applications
(Dynamic) Bayesian Network	Directed acyclic graph of random variables	Systems naturally expressed as dependence structure between random variables	Biological networks (Smith, 2010), ecological conservation (Korb and Nicholson, 2010)
(Dynamic) Chain Event Graph	Derived from event tree coloured to represent conditional independence	Asymmetric problems, problem description is told as a series of unfolding events	Healthcare outcomes (Barclay et al., 2014), forensic evidence (Collazo et al., 2018)
Chain Graphs	Hybrid graph with directed and undirected edges	Problem description has both directional and ambiguous relationships	Mental health (Cox and Wermuth, 1993), social processes (Cox and Wermuth, 2014)
Flow Graph	Hierarchical flow network	Supply and demand problems, homogeneous flows	Commodity supply (Figueroa and Smith, 2007)
Multi-regression Dynamic model	Collection of regressions where the parents are the regressors	Contemporaneous effects between time series	Marketing(Smith, 1993), traffic flows(Queen and Albers, 2009), neural fMRI activity(Costa et al., 2015)
Regulatory Graph	Graph customised to regulatory hypotheses	Need to test a regulatory hypothesis	Biological control mechanisms (Liverani and Smith, 2015)

Table 3.1: Examples of customised graphical models.

shift in how modellers elicit structure. Furthermore, inference on each of these novel representations engenders customised notions of causation, as each of the full probability representations of customised models admits its own causal algebras. The causal effects following intervention in a BN are well studied, and these methods can be extended to custom classes of models discussed here. A thorough investigation of causal algebras is beyond the scope of this chapter, but it offer further motivation for careful attention to structure in the elicitation process. Chapter 6 shows how each structural class admits a different interpretation of cause. Future work will demonstrate how each structural class has its own causal algebra and that for causation to be meaningful the underlying structure on which it is based needs to properly reflect domain knowledge. Many open questions remain as to how to customise structure elicitation.

Chapter 4

Checking CEG Structure with d-Separation Theorem

> If you cud even jus see 1 thing clear the woal of whats in it you cud see every thing clear. But you never wil get to see the woal of any thing youre all ways in the middl of it living it or moving through it
>
> *Riddley Walker*, Russell Hoban

4.1 Background

The d-separation theorem for BNs has been used to systematically list and verify the irrelevance statements implied by the graph. This has enabled advancements in causal inference and decision modelling. However, the semantics of a BN are not always suited to a given application. Deriving an ancestral construction and accompanying d-separation theorem for CEGs permits a consistent querying of context-specific conditional independence relationships within a single graphical representation.

Separation theorems for CEGs were first formulated as configurations of noisy-and gates (Smith and Anderson, 2008). The equivalence classes for CEGs can be traversed via a polynomial equivalence class. One of the operators involved in determining the polynomial equivalence class, the swap operator is crucial to constructing the ancestral CEG class (Görgen and Smith, 2018). A separation theorem for simple, uncoloured CEGs proved that the existence of a cut vertex created conditions for d-separation (Thwaites and Smith, 2015). This chapter

contributes a much more general d-separation theorem which can be applied to any coloured, square-free CEG. It uses a novel ancestral graph construction for the CEG analogous to ancestral constructions used in the querying of BNs.

The ancestral CEG construction is a DAG that can be used to verify the collection of conditional independence statements implied by the chain event graph constructed from the staged event tree, $\mathcal{C}$. In that sense, it can be used to ease the transition to the full model with graphical and probabilistic components, $(\Omega(\mathcal{C}), P(\mathcal{C}, \mathbb{U}, \mathbb{W}))$, where $\Omega(\mathcal{C})$ denotes the sample space of atomic events on $\mathcal{C}$, $\mathbb{U}$ indicates the stages of the graph, and $\mathbb{W}$ denotes the position structure. Only once this topology and colouring is discovered will the full model $(\Omega(\mathcal{C}), P(\mathcal{C}, \mathbb{U}, \mathbb{W}))$ be estimated. This is extremely important when eliciting a CEG. Many dependence queries can be examined, confirmed, or disputed by domain experts before the putative framework is quantified. Adding this stage to the elicitation process helps us make sure, with minimal effort, that the actual, broad framework can be embellished into a full probability model is faithful to the expert's structural beliefs as discussed in Chapter 3. In this way, the modeller does not waste time eliciting probabilities on models that are ultimately inappropriate.

In the Section 4.2, I present the technical prerequisites necessary for the ancestral graph including intrinsic events, random variables, and ancestors within the CEG. This includes preliminary results about dependence between the random variables of a CEG. The full construction algorithm for the ancestral CEG graph as a function of the query is given in Section 4.3. This novel construction of a graph of a valid BN represents the conditional independence structures of a special class of random variables measurable with respect to the event space generated by $\mathcal{C}$. Section 4.4, proves the sufficiency and necessity of CEG d-separation. Section 4.4.3 proves that under certain regularity conditions the method described in the ancestral construct gives the full list of such statements. This new construction gives a complete list of all irrelevance statements we can check to validate a CEG model before we proceed to embellish it into a full probability specification with the necessary addition of vectors of quantified conditional probabilities. Section 4.5 proves that the d-separation queries in the BN can be addressed with an equivalent CEG. Thus, the d-separation theorem for CEGs encompasses a much broader class of models. A brief discussion of these results follows.

4.2 Technical Prerequisites

4.2.1 Semi-graphoid Axioms and Properties of Conditional Independence

The following properties are central to querying dependence relationships between $\boldsymbol{Y}_{B_1}, \boldsymbol{Y}_{B_2}.\boldsymbol{Y}_{C_1}.\boldsymbol{Y}_{C_2}$ given subcomponents of the vector $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_m)$ of random variables where $\boldsymbol{Y}_{B_1}$ denotes the components of $\boldsymbol{Y}$ whose indices lie in the set $B_1 \subset M = \{1, 2, \ldots, m\}$. These properties hold for any random variable defined on a discrete space. In particular, all the properties below hold irrespective of any positivity conditions on the corresponding mass functions. This is important in our context because there are often functional relationships between the sets of variables of interest.

If a random variable $\boldsymbol{Y}_{B_1}$ is degenerate given a conditioning set $\boldsymbol{Y}_{C_1}$– that is it takes a single value with probability 1– then the vector of the other random variables $\boldsymbol{Y}_{M\setminus B_1}$ are irrelevant to it. That is,

$$\boldsymbol{Y}_{B_1} \perp\!\!\!\perp \boldsymbol{Y}_{M\setminus\{B_1\cup C_1\}} | \boldsymbol{Y}_{C_1} \tag{4.1}$$

For any function $f(\boldsymbol{Y}_{B_1})$ of $\boldsymbol{Y}_{B_1}$, then for all sub-vectors $\boldsymbol{Y}_{B_2}$ and conditioning set $\boldsymbol{Y}_{C_1}$,

Property 26

$$\boldsymbol{Y}_{B_1} \perp\!\!\!\perp \boldsymbol{Y}_{B_2} | \boldsymbol{Y}_{C_1} \Leftrightarrow (\boldsymbol{Y}_{B_1}, f(\boldsymbol{Y}_{B_1})) \perp\!\!\!\perp \boldsymbol{Y}_{B_2} | \boldsymbol{Y}_{C_1}$$

Definition 27 *The* ***strong decomposition property for random vectors*** $\boldsymbol{Y}_{B_1}, \boldsymbol{Y}_{B_2}, \boldsymbol{Y}_{C_1}, \boldsymbol{Y}_{C_2}$ *states that*

$$\boldsymbol{Y}_{B_1} \perp\!\!\!\perp (\boldsymbol{Y}_{B_2}, \boldsymbol{Y}_{C_1}) | \boldsymbol{Y}_{C_2} \Leftrightarrow \boldsymbol{Y}_{B_1} \perp\!\!\!\perp \boldsymbol{Y}_{B_2} | \boldsymbol{Y}_{C_1}, \boldsymbol{Y}_{C_2} \text{ and } \boldsymbol{Y}_{B_1} \perp\!\!\!\perp \boldsymbol{Y}_{C_1} | \boldsymbol{Y}_{C_2}.$$

Property 28 (Symmetry property for random vectors) *The symmetry property states that*

$$\boldsymbol{Y}_{B_1} \perp\!\!\!\perp \boldsymbol{Y}_{B_2} | \boldsymbol{Y}_{C_1} \Leftrightarrow \boldsymbol{Y}_{B_2} \perp\!\!\!\perp \boldsymbol{Y}_{B_1} | \boldsymbol{Y}_{C_1}.$$

Here we plan to derive necessary and sufficient conditions for two random vectors given an amenable event to be independent of each other for all probability models in $\mathcal{C}$.

4.2.2 Relevant Class of CEGs and their Probability Models

The results for ancestral graphs derived here pertain to minimal, square-free CEGs. Here we show the probability model described by each graph. Recall from Chapter 2, that the CEG $\mathcal{C}(\mathcal{T}) = (\mathbb{W}, F)$ is built from a staged tree $\mathcal{T}$ on a set of positions $\mathbb{W}$ and corresponding edge set F. In this chapter, we will simplify $\mathcal{C}(\mathcal{T})$ as $\mathcal{C}$.

Recall a CEG $\mathcal{C}$ has an associated directed acyclic graph where $V(\mathcal{C})$ is the vertex set and $F(\mathcal{C})$ is the edge set. As the CEG is formed from a directed tree, the edge set includes a single root vertex w_0 and a single sink vertex w_∞.

Positions in the same stage and edges with the same probability are coloured– the remaining edges and positions are shown in the following diagrams as remaining black and unfilled, although they each have a unique colour. Colouring the edges allows us to express the conditional independence relationships implicit in $\mathcal{C}$ without specifying a particular probability model.

In this chapter, W denotes the set of positions unique to a particular query about conditional independence. Recall from Chapter 2 that the general notation for CEGs is as follows. The non-sink vertices $V(\mathcal{C}) \setminus w_\infty$ form the set of *positions* $w \in \mathbb{W}(\mathcal{C})$. The *stages*, $u \in \mathbb{U}(\mathcal{C})$, are subsets of the positions, $\mathbb{W}(\mathcal{C})$, where each subset referred to as a *stage* corresponds to a unique vertex colour. If $w \in \mathbb{W}(\mathcal{C})$ is uncoloured then it lies in a stage $u \in \mathbb{U}(\mathcal{C})$ such that $u = \{w\}$. On the other hand, if $w \in \mathbb{W}(\mathcal{C})$ is coloured and $n(u)$ other positions share that colour then $u \in \mathbb{U}(\mathcal{C})$ consists of all those positions in $\mathbb{W}(\mathcal{C})$ sharing that colour.

Each CEG now acts as an index of a family of probability models $\mathbb{P}(\mathcal{C})$ uniquely determined by the coloured graph $\mathcal{C}$ as follows. The set of atoms $\omega \in \Omega(\mathcal{C})$ of the finite event space of $\mathbb{P}(\mathcal{C})$ is constructed to be in one-to-one correspondence to the set $\lambda \in \Lambda(\mathcal{C})$ of root to sink paths. Then each $u \in \mathbb{U}(\mathcal{C})$ is assigned a strictly positive probability vector $\boldsymbol{\pi}(u)$ that has a length of the number of outgoing edges from that stage, u. Each edge $f \in F(\mathcal{C})$ is associated to one of the components π_f of $\{\boldsymbol{\pi}(u) : u \in \mathbb{U}(\mathcal{C})\}$. For two edges f, f', the corresponding probabilities $\pi_f = \pi_{f'}$ if and only if the edges f, f' are coloured the same in $\mathcal{C}$. The probability mass function $\{p_{\mathcal{C}}(\omega) : \omega \in \Omega(\mathcal{C})\}$ of $\mathbb{P}(\mathcal{C})$ is then defined by the formula:

$$p_{\mathcal{C}}(\omega(\lambda)) = \prod_{f \in \lambda} \pi_f. \tag{4.2}$$

In this way, once we specify the values of the strictly positive probability vectors $\{\boldsymbol{\pi}(u) : u \in \mathbb{U}\}$, the CEG $\mathcal{C}$ indexes a single probability model $(\Omega(\mathcal{C}), \mathbb{P}(\mathcal{C}))$. On the other hand, until we specify $\{\boldsymbol{\pi}(u) : u \in \mathbb{U}\}$, the DAG $\mathcal{C}$ represents a class of probability models just as the graph of a BN does.

Each CEG $\mathcal{C}$ has a unique, minimal representation that maintains the order

of edges in a CEG. Two different CEGs can represent the same class of probability model. However there is a unique *minimal* CEG $\mathcal{C}$ that has the smallest number of positions. We can construct a minimal CEG from one that is not minimal by simply repeatedly merging two positions into a single position whenever the coloured subtrees rooted at those positions are isomorphic, merging the subsequent edges and vertices in these subtrees in the obvious way. We continue to do this until no such pairs of positions exist.

Definition 29 *A **minimal** CEG has no two positions whose subtrees are isomorphic.*

It is easy to check that such an operation leaves the set of root-to-sink paths and the sequence of colours on their edges the same in the unmerged and merged CEG, ensuring that both the probability space and the probabilities assigned to its atoms in Equation 4.2 are the same. All the CEGs we consider here will henceforth be assumed to be minimal. We will later see that minimal CEGs are especially important to prove the sufficiency of results given here that are based solely on the topology of a graph. The results we derive here hold for square-free CEGs.

4.2.3 Random Variables of a CEG

Producing an analogue of the d-separation theorem requires first specifying subsets of random variables related to the queries. When querying the DAG of a BN the pre-specified set of variables represented in the vertex set of the DAG may be queried. However, for a CEG this selection is not quite so straightforward because it is originally specified in terms of an event tree and so only indirectly in terms of random variables. Despite this, there are three types of variables which are central to describing how a unit might traverse a CEG. These variables are good candidates for useful dependence queries. The first set of random variables, measurable with respect to $\Omega(\mathcal{C})$ are the position *incident* variables.

Definition 30 *For any position $w \in \mathbb{W}(\mathcal{C})$ and the given path a unit traverses $\lambda \in \Lambda$, define an incident variable to be*

$$I(w) = \begin{cases} 1, \{\lambda : w \in \lambda\} \\ 0, \{\lambda : w \notin \lambda\} \end{cases} \tag{4.3}$$

Note that the incidence variable of the root node $I(w_0) \equiv 1$ is degenerate because all root-to-leaf paths λ go through the root.

Definition 31 *The position $w \in \mathbb{W}(\mathcal{C})$ is a* ***cut vertex*** *when all root-to-sink paths $\lambda \in \Lambda$ pass through w. The set of all cut vertices is written as W_0.*

The second variable is associated with indicators on positions in the ancestral graph, referred to as an *ancestral incident* variable. The ancestral graph is composed of the ancestors of the query set $\text{An}(W) \subseteq \mathbb{W}$. This can result in a graph with positions that must be merged to be a minimal representation. This results in an additional partition $\mathbb{V}$ of the positions that is finer than the stages but coarser than the positions $\mathbb{W}(\mathcal{C}) \preceq \mathbb{V}(\mathcal{C}) \preceq \mathbb{U}(\mathcal{C})$. This additional partition $\mathbb{V}(\mathcal{C})$ accounts for the positions in the same stage that cannot be merged in the full CEG as they have different subsequent subtrees, but they can be merged when we take the ancestral CEG on a subset of the original nodes in Section 4.4. Let

$$\mathbb{V}(\mathcal{C}) = \{v_1, \ldots, v_j, \ldots, v_{\#(v)}\}$$

be partition of the set of positions we get from merging the finer set of positions

$$\text{An}(W) \subseteq \mathbb{W}(\mathcal{C}) = \{w_1, \ldots, w_i, \ldots, w_{\#(w)}\}.$$

The ancestral position partition is no coarser than the stage partition

$$\mathbb{U}(\mathcal{C}) = \{u_1, \ldots, u_k, \ldots, u_{\#(u)}\}.$$

Each position corresponds to an ancestral position $w_i \in v_j$ and stage $v_j \in u_k$. For each ancestral position v, $w(v)$ denotes the set of positions that have been merged into the new ancestral position.

Note that when ancestral position v merges two positions $w_1, w_2 \in v_j$ then these vertices are coloured the same $w_1, w_2 \in u_k$ and $v_j \in u_k$. The need for the set of ancestral positions will become clear as we construct the ancestral graph on a smaller set of variables. Essentially, ancestral positions arise in our ancestral construction when we have two positions in the same stage with the same subtrees that have had some descendants removed.

Definition 32 *For any ancestral position $v \in \mathbb{V}$, define the* ***ancestral position incident variable*** *as*

$$I(v|\mathbb{V}) = \begin{cases} 1, \{\lambda : v \in \lambda\} \\ 0, \{\lambda : v \notin \lambda\} \end{cases} \tag{4.4}$$

The incident variable and ancestral incident variable indicate whether or not a unit has passed through the (ancestral) position. The actual edge the unit traverses

is given by the floret variables, defined below.

Definition 33 *For each $w \in \mathbb{W}$ a **floret variable** is defined as*

$$X(w) = \begin{cases} f(w), \{\lambda : w \in \lambda\} \\ 0, \{\lambda : w \notin \lambda\}. \end{cases}$$

For any directed path λ, an $X(w)$ is only initiated when its corresponding $I(w) = 1$. Additionally, $X(w) = w'$ instantiates $I(w')$ and this process carries on recursively from the root to the sink until all the $w \in \lambda$ are instantiated. For all other $w \in \mathbb{W}(\mathcal{C})$, $I(w) = 0$ and hence, $X(w)$ will not be instantiated when considering path λ. Each random variable in $\mathcal{C}$ is defined for any position w and is measurable with respect to $\Omega(\mathcal{C})$. Note that for each w, the random variables $I(w)$ and $X(w)$ are measurable with respect to the sigma field $\Omega(\mathcal{C})$ whose atoms are the different root to sink paths of $\mathcal{C}$.

The colour of the stages and positions indicates that there are many dependences between these random variables. However, all three types of random variable with $\Omega(\mathcal{C})$ are functions of the set of random variables $\{X(w) : w \in \mathbb{W}(\mathcal{C})\}$, which in this sense gives a complete picture of the underlying processes. Henceforth, this work considers only dependence queries associated with these variables.

4.2.4 Ancestors, Descendants, and Conditional Independence

This section translates the terminology for ancestors and descendants to the CEG. In this section, the existence of a directed path in $\mathcal{C}$ from w to w' is denoted $w \prec w'$. For a set of positions $W \subseteq \mathbb{W}(\mathcal{C})$ write

$$\begin{aligned} \mathrm{An}(W) &\triangleq \left\{w : w \in W \text{ or } \exists\ w' \in W \text{ such that } w \prec w'\right\} \\ \mathrm{An}^{-}(W) &\triangleq \left\{w : \exists\ w' \in W \text{ such that } w \prec w'\right\} \\ \mathrm{De}(W) &\triangleq \left\{w : w \in W \text{ or } \exists\ w' \in W \text{ such that } w' \prec w\right\} \end{aligned}$$

to denote the *ancestral set*, *non-ancestral set*, and *descendent set* respectively of W in $\mathcal{C}$. Then, $\mathrm{Nd}(W)$ is the non-descendent set of W and note that:

$$W \subseteq \mathrm{An}(W) \subseteq \mathrm{Nd}(W) \subseteq \mathbb{W}(\mathcal{C})$$

The set of positions in question is contained in the ancestral set which is contained in the non-descendant set, which is contained in the set of all positions. Notice that directly from the topology of $\mathcal{C}$, we can determine whether or not w is in any of the sets.

Definition 34 *For a a given position $w \in \mathbb{W}(\mathcal{C})$, the **ancestral floret** and **ancestral incident vectors**, denoted $\boldsymbol{X}_{\mathrm{An}^-(w)}$ and $\boldsymbol{I}_{\mathrm{An}^-(w)}$ respectively, are the random vectors whose components consists of*

$$\left\{X(w') : w' \prec w\right\} \text{ and } \left\{I(w') : w' \prec w\right\}.$$

Definition 35 *Similarly, for a a given position $w \in \mathbb{W}(\mathcal{C})$, the **non-descendant floret** and **non-descendant incident vectors**, denoted as $\boldsymbol{X}_{\mathrm{Nd}(w)}$ and $\boldsymbol{I}_{\mathrm{Nd}(w)}$ respectively, are the random vector whose components consist of*

$$\left\{X(w') : w' \in \mathrm{Nd}(w)\right\} \text{ and}$$

$$\left\{I(w') : w' \in \mathrm{Nd}(w)\right\}.$$

Knowing incidence at a position in $\mathcal{C}$ renders the incidence and floret of non-descendant positions irrelevant to the floret variable in question in the sense of Lemma 36.

Lemma 36 *For all $w \in \mathbb{W}(\mathcal{C})$*

$$X(w) \perp\!\!\!\perp \left(\boldsymbol{I}_{\mathrm{Nd}(w)}, \boldsymbol{X}_{\mathrm{Nd}(w)}\right) | I(w) \tag{4.5}$$

Proof. Note that since $\boldsymbol{I}_{\mathrm{Nd}(w)}$ is by definition a function of $\boldsymbol{X}_{\mathrm{Nd}(w)}$ from Properties 28 and 26 it is sufficient to prove that

$$X(w) \perp\!\!\!\perp \boldsymbol{X}_{\mathrm{Nd}(w)} | I(w) \tag{4.6}$$

which is equivalent to requiring both

$$X(w) \perp\!\!\!\perp \boldsymbol{X}_{\mathrm{Nd}(w)} | I(w) = 0 \tag{4.7}$$

and

$$X(w) \perp\!\!\!\perp \boldsymbol{X}_{\mathrm{Nd}(w)} | I(w) = 1. \tag{4.8}$$

Since by definition the event $\{X(w) = 0\}$ implies that $\{I(w) = 0\}$, $X(w)$ is degenerate when $I(w) = 0$. Thus, Equation 4.7 is a direct consequence of Equation 4.1. Next note that directly from the definition of a CEG,

$$X(w) \perp\!\!\!\perp \boldsymbol{X}_{\mathrm{An}^-(w)} | I(w) = 1. \tag{4.9}$$

Furthermore given a unit passes along a path λ that reaches w so that $I(w) = 1$,

again by definition, it cannot pass through or have passed through any positions in $\mathrm{Nd}(w)\backslash\mathrm{An}^{-}(w)$. So the event

$$\{I(w)=1\}=\{I(w)=1\}\cap\left\{I(w')=0:w'\in\mathrm{Nd}(w)\backslash\mathrm{An}^{-}(w)\right\},$$

and consequently $\boldsymbol{X}_{\mathrm{Nd}(w)\backslash\mathrm{An}^{-}(w)}=\mathbf{0}$. Therefore,

$$\begin{aligned}
X(w)\perp\!\!\!\perp\boldsymbol{X}_{\mathrm{Nd}(w)}|\{I(w)=1\},\boldsymbol{X}_{\mathrm{An}^{-}(w)}\\
X(w)\perp\!\!\!\perp\boldsymbol{X}_{\mathrm{Nd}(w)\backslash\mathrm{An}^{-}(w)},\boldsymbol{X}_{\mathrm{An}^{-}(w)}|\{I(w)=1\}\\
X(w)\perp\!\!\!\perp\boldsymbol{X}_{\mathrm{Nd}(w)\backslash\mathrm{An}^{-}(w)}|\boldsymbol{X}_{\mathrm{An}^{-}(w)},\{I(w)=1\}\cap\left\{I(w')=0:w'\in\mathrm{Nd}(w)\backslash\mathrm{An}^{-}(w)\right\}\\
X(w)\perp\!\!\!\perp\boldsymbol{X}_{\mathrm{Nd}(w)\backslash\mathrm{An}^{-}(w)}|\boldsymbol{X}_{\mathrm{An}^{-}(w)}|\{I(w)=1\}
\end{aligned}\tag{4.10}$$

which is true trivially by Definition 1 and Equation 4.1 because $\boldsymbol{X}_{\mathrm{Nd}(w)\backslash\mathrm{An}^{-}(w)}$ under the conditioning event above is degenerate. Thus Equations 4.9 and 4.10 prove the result by Property 26. ■

Lemma 36 expresses the independence of non-descendants from a floret variable of a position given the incident variable. Additional results can be proved by defining the parent set of positions.

Definition 37 *For a a given position $w'\in\mathbb{W}(\mathcal{C})$, the **parent floret** and **parent incident vectors**, denoted $\boldsymbol{X}_{\mathrm{Pa}(w')}$ and $\boldsymbol{I}_{\mathrm{Pa}(w')}$, are the random vectors whose components are given by $\{w:(w,w')\in F(\mathcal{C})\}$.*

Denote $\boldsymbol{X}_{\overline{\mathrm{Pa}}(w')}\triangleq\boldsymbol{X}_{\mathbb{W}(\mathcal{C})\backslash\mathrm{Pa}(w')\cup\{w'\}}$, the set of all incident variables not w' and not in this set.

Lemma 38 *For any cut vertex $w_c\in W_0$*

$$I(w_c)\perp\!\!\!\perp\boldsymbol{X}(w_c)\tag{4.11}$$

whilst for all $w'\in\mathbb{W}(\mathcal{C})\backslash W_0$

$$I(w')\perp\!\!\!\perp\boldsymbol{X}_{\overline{\mathrm{Pa}}(w')}|\boldsymbol{X}_{\mathrm{Pa}(w')}\tag{4.12}$$

Proof. If $w'\in W_0$ then $I(w')$ is degenerate so Equation 4.11 is a direct consequence of Equation 4.1. The conditional independence in Equation 4.12 comes from noting that by definition:

$$I(w)=\sup_{w\in\mathrm{Pa}(w')}\chi\left\{X(w)=(f,f')\right\}$$

where χ is the indicator variable. χ is a function of the parent floret variables

leading into the cut vertex position $w' \in \mathbb{W}(\mathcal{C})$. So this property is also a result of Equation 4.1 ■

4.2.5 Intrinsic Events and Conditional Independence

There are certain events in $\Lambda \in \Omega(\mathcal{C})$ that are of particular interest. First, note that the event $\Lambda \in \Omega(\mathcal{C})$ induces a subgraph $\mathcal{C}_\Lambda \subseteq \mathcal{C}$. Events that can be represented by the set of all the root to sink paths in the subgraph are called intrinsic. It is easy to check that these events form a pi-system, that is they are closed under intersection. All conditioned queries in a DAG are sets of queries conditioning on an intrinsic event. Although a CEG has many non-intrinsic events, these correspond to events that have no direct relationship to the graph defining the process.

Definition 39 *A set of root-to-sink paths Λ in $\Omega(\mathcal{C})$ defines an **intrinsic event** if there is a subgraph $\mathcal{C}_\Lambda$ such that Λ consists of all the root to leaf paths in $\mathcal{C}$ that pass through $\mathcal{C}_\Lambda$.*

The ancestral CEG uses a smaller subset of ancestral positions $v_j \in \mathbb{V}(\mathcal{C}) \subseteq \mathbb{W}(\mathcal{C})$, so we define amenable events as an analogue of intrinsic events for ancestral CEGs. The ancestral graph will be defined in Section 4.3, but for the moment, but the analogue of intrinsic events is introduced here.

Definition 40 *A set of paths Λ in the ancestral graph $\mathcal{C}^{m\prime}_{\Lambda_{\mathrm{An}(W)}}$ defines an **amenable event** if $\mathcal{C}^{m\prime}_{\Lambda_{\mathrm{An}(W)}}$, which is itself a subgraph containing the root-to-sink paths contains exactly the root-to-sink paths specified in Λ.*

Conditioning on an intrinsic event preserves the conditional independence relationships on the subgraph. As d-separation is a graphical criterion, the results of the d-separation theorem will be restricted to queries that induce amenable events in the ancestral construction.

If $\mathcal{C}$ represents a semi-Markov process where units pass along one of its paths then intrinsic and amenable events are natural to discuss. They represent the typical conditioning events that might arise when there is only partial information about a particular unit's path. For example, if the positions in $\mathcal{C}$ represent certain previous conditions of a patient, a doctor may learn from the patient's records and conversations with her a collection of some of the states she had passed through or at least the colour of position. The doctor would then need to infer both the gaps in this record and the possible future unfolding of the patient's pathway.

Not all events are of this form. A simple example of an event that is non-intrinsic can be seen in the graph of Figure 4.1. The event $F = \lambda_{(0,0),(1,1)}$ induces the

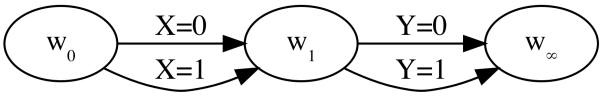


Figure 4.1: The event $\lambda_{(0,0),(1,1)}$ where $\{X = 0, Y = 0\} \cup \{X = 1, Y = 1\}$ represents a non-intrinsic event as the subgraph admits events not in the set such as $\{X = 0, Y = 1\}$.

subgraph that contains root-to-sink paths like $\lambda_{(0,1),(1,0)}$ that are not in the original event. There is no subgraph such that the event corresponds to all the root to sink paths.

If the CEG represents a faithful BN, then by addressing the queries associated with a CEG concerning amenable conditioning events we can query at least as many implied conditional independent statements as we would if we were to use the d-separation theorem on the BN directly.

The functional relationships between the random variables in a CEG renders conditioning on intrinsic events quite subtle. The more specific the conditioning event the more it will tend to force independences. One extremely important issue here is that, directly from the definition of the incident variables, knowing incidence at one position automatically imparts knowledge about incidence at a set of positions the unit logically could not have passed through. (Note that definitions in this section are defined for positions but also apply to ancestral positions.) Formally, define the conditioning set C corresponding to the intrinsic event Λ in the following way. $\overline{T}$ can be thought of as the trash set that no longer applies once we observe the incidence of a particular position. $T(C)$ represents the root-to-leaf paths of the conditioning event.

$$\{I(w) = 1 : w \in C\} \Rightarrow \left\{I(w') = 0 : w' \in \overline{T}(C)\right\}$$

where

$$\overline{T}(C) \triangleq \bigcap_{w \in C} \left\{\mathrm{Nd}(w) \backslash \mathrm{An}^{-}(w) : w \in C\right\}$$

is the set of all positions $w \in \mathbb{W}(\mathcal{C})$ which are neither in the ancestor set nor in the descendant set of all $w \in C$. Note that

$$\overline{T}(C)^c \triangleq T(C) \triangleq \bigcup_{w \in C} \left\{\mathrm{An}(w) \cup \mathrm{De}(w)\right\}.$$

Conditioning on $T(C)$ conditions on the root-to-leaf paths containing all $w \in C$. Unless the cardinality of C is small, $T(C)$ defines a small event and $\overline{T}(C)$ contains

most of the positions $w \in \mathbb{W}(\mathcal{C})$.

$$\{I(w) = 1 : w \in C\} \Leftrightarrow \left\{I(w) = 1, I(w') = 0, w' \in \overline{T}(C)\right\}$$

$\overline{T}(C)$ may be the empty set and if not

$$\coprod_{w' \in \overline{T}(C)} X(w') | \{I(w) = 1 : w \in C\}$$

and

$$\boldsymbol{X}_{\overline{T}(C)} \perp\!\!\!\perp \boldsymbol{X}_{T(C)} | \{I(w) = 1 : w \in C\}.$$

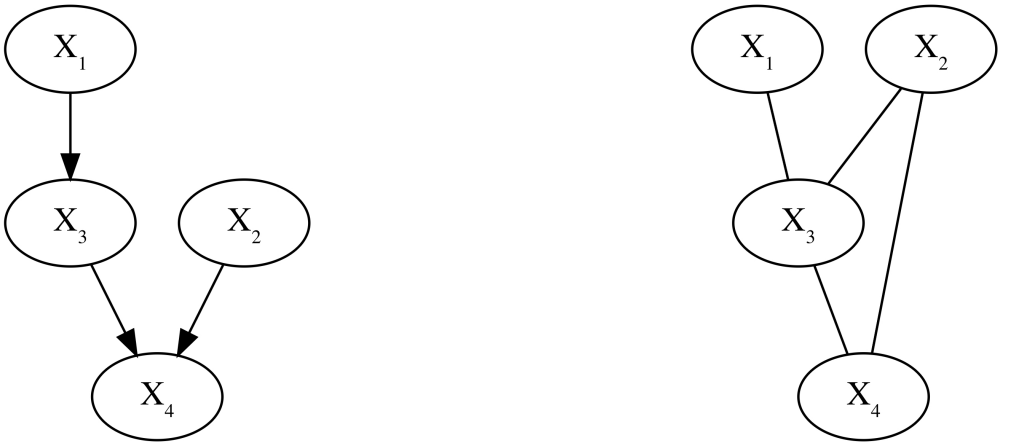
The implied conditional independences concerning $\boldsymbol{X}_{\overline{T}(C)}$ are trivial. Conditional on $\{I(w) = 1 : w \in C\}$ the components of $\boldsymbol{X}_{\overline{T}(C)}$ are all mutually independent of each other and also all independent of $\boldsymbol{X}_{T(C)}$ because the values of all these variables are known once this event happens. So the only conditional independences given $\{I(w) = 1 : w \in C\}$ are those that concern the variables in $\boldsymbol{X}_{T(C)}$. Henceforth we will consider relationships only between these variables when conditioning on such events.

4.3 Ancestral Graphs for the CEG

The ancestral construction enables us to answer dependence queries for sets of variables conditioned on an intrinsic event. Proving the necessity of the d-separation condition to show independence requires an additional construct that extends the compact representation of the CEG to a valid BN whose vertices are random variables. Framing dependence queries in terms of the random variables defined in Section 4.4 answers additional lemmas about the dependence structure based on ancestrality. Proving d-separation for the BN will require the extended ancestral graph although the CEG d-separation criteria requires only the CEG ancestral construction.

4.3.1 Ancestral CEGs

One novel construction we use in this thesis is the *ancestral* CEG $\mathcal{C}_{A(W)}$. Thwaites and Smith (2015) determined that the existence of a cut vertex is a sufficient condition to read conditional independence from a simple (uncoloured) CEG. The novel ancestral construction in this section addresses queries about conditional independences that arise from the colouring of the positions. This requires several steps, notably incorporating the results associated with the swap operator from Görgen and Smith (2018). The work in this chapter exploits these relatively new



(a) A sample BN

(b) Moralized BN

Figure 4.2: An example of an ancestral BN construction on binary variables $\{X_1, X_2, X_3.X_4\}$ corresponding to the ancestral CEG shown below

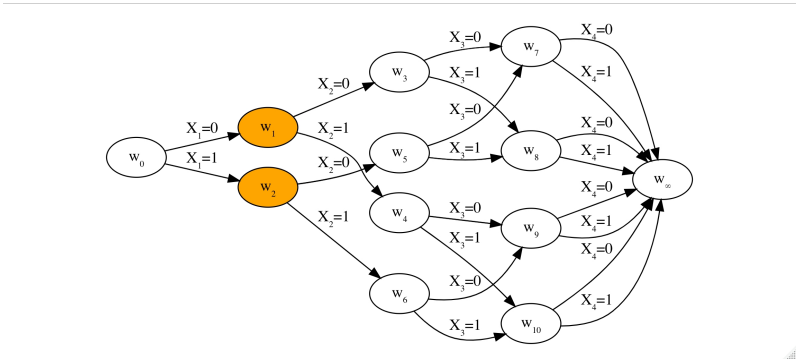


Figure 4.3: The CEG corresponding to Figure 4.2.

results and enables us to give a new definition of context-specific ancestrality. The ancestral construction requires some preliminary definitions.

Definition 41 *For a CEG, $\mathcal{C}$, the **ancestral vertex** and **edge sets** for the base ancestral CEG, $\mathcal{C}^0_{\mathrm{An}(W)}$ are defined as*

$$V\left(\mathcal{C}^0_{\mathrm{An}(W)}\right) = \{V(\mathcal{C}) \cap \mathrm{An}(W)\} \cup \{w_\infty\}$$

$$E\left(\mathcal{C}^0_{\mathrm{An}(W)}\right) = \left\{(w, w') \,|w, w' \in V\left(\mathcal{C}^0_{\mathrm{An}(W)}\right) \wedge (w, w') \in F(\mathcal{C})\right\} \cup \left\{(w, w_\infty) \,|w \in V\left(\mathcal{C}^0_{\mathrm{An}(W)}\right) \nexists\, w' \notin V\left(\mathcal{C}^0_{\mathrm{An}(W)}\right)\right\}$$

That is, the vertex set is the set of ancestors of all the positions in the query plus a new sink node. The edge set is given as in $\mathcal{C}$, with the exception that if $w \notin V(\mathcal{C}^0_{\mathrm{An}(W)})$ then $e = (w, w')$ is mapped to enter the new sink vertex w_∞ of $\mathcal{C}^0_{\mathrm{An}(W)}$.

Part of the construction of the ancestral CEG requires choosing the equivalent graph with a particular order. Defining this order necessitates formalizing distance between sets of positions on the graph.

Definition 42 *For two non-overlapping sets of positions in the CEG, B_1 and B_2, the **distance** $d(B_1, B_2)$ is the sum of the lengths of the directed paths between each position in each set divided by the total number of pathways between the two sets given by*

$$d(B_1, B_2) = \frac{\sum_{w_1 \in B_1} \sum_{w_2 \in B_2} |\lambda(w_1, w_2)|}{\#\lambda(w_1, w_2)}.$$

Finding the optimal ordering requires the swap operator, one of the functions necessary to traverse the equivalence class. The swap operator is analogous to arc reversals in the BN. Algebraically, the swap operator changes the order of summation in the interpolating polynomial of a CEG (Görgen and Smith, 2018). Unique to the ancestral construction, positions can also be swapped when we look at the subgraphs implied by the conditioning set. These graphs have singleton edges that may also be swapped. We define a swap and a twin as in Görgen and Smith (2018).

Definition 43 *A probability subtree $(\mathcal{T}, \Theta_\mathcal{T})_u \subseteq (\mathcal{T}, \Theta_\mathcal{T})$ is a **twin** if all root-to-leaf paths consist of exactly two edges and all children of its root are in the same stage u.*

Definition 44 *Let $(\mathcal{T}, \Theta_\mathcal{T})$ be a staged tree with $(\mathcal{T}, \Theta_\mathcal{T})_u \subseteq (\mathcal{T}, \Theta_\mathcal{T})$ a twin around stage u. Denote a tree polynomially equivalent to $(\mathcal{S}, \Theta_\mathcal{S})_u \subseteq (\mathcal{S}, \Theta_\mathcal{S})$ The map $(\mathcal{T}, \Theta_\mathcal{T}) \mapsto (\mathcal{S}, \Theta_\mathcal{S})$ is a **swap** if $(\mathcal{S}, \Theta_\mathcal{S})$ is a staged tree.*

The motivation for the ancestral ordering stems from attempting to force cut-vertices in the ancestral graph by juxtaposing B_1 and B_2 where possible.

Definition 45 *The **ancestral ordering** $I_{An(B_1 \cup B_2 \cup C)}$ for non-overlapping sets of positions in a query B_1, B_2 and C can be found by applying swaps in order to:*

1. *minimize* $d(B_1, B_2)$
2. *minimize* $\min\{d(B_1, C), d(B_2, C)\}$

The ancestral CEG requires examining the subgraphs induced by the conditioning set. The set of vertices $W = B_1 \cup B_2 \cup C$ is composed of two query sets B_1 and B_2 and the conditioning set C. The conditioning set C consists of positions $w_C^1, w_C^2, \ldots w_C^{\#(C)}$. Each position has a number of emanating edges. The product of these numbers of emanating edges across all valid pathways gives us the number of contexts we consider for the isomorphic subgraphs. Each conditioning context gives rise to an induced subgraph.

Definition 46 *For a given CEG $\mathcal{C}$ and query conditioning on the intrinsic event Λ_F involving the set of positions $w \in C$ with context $X(w) = c$, the **conditioned subgraph** for a given context c is the set of pathways through the given context c, denoted Λ_c.*

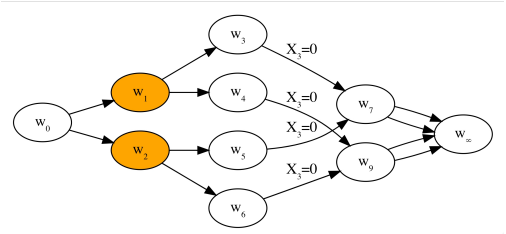
Note that conditioned subgraphs are still valid CEGs because we are assuming that each individual context happens with probability 1. In the ancestral CEG construction, we will examine the subgraphs both upstream and downstream of each conditioning set. If for all values of $X(w) = c$, Λ_c are isomorphic up to a relabelling of the colours, we can merge the corresponding positions to a new node w_c.

In the base ancestral CEG $C^0_{\mathrm{An}(W)}$, new ancestral positions may have arisen from the consolidation of the conditioning context and the removal of the non-descendants of the query set. Because it is a coloured subtree of $\mathcal{C}$ it may now exhibit two positions $w_1, w_2 \in \mathbb{W}\left(\mathcal{C}^0_{\mathrm{An}(W)}\right)$ where the sub-graphs rooted at w_1 and w_2 respectively are colour isomorphic and have identical subtrees. If this is the case then the same family could be represented by a graph with a new position merging w_1 and w_2 into w_{12}, for example. The final set of nodes in the ancestral CEG is the minimal, conditioned subgraph version of $\mathcal{C}^{m\prime}_{\Lambda_{\mathrm{An}(W)}}$, is obtained by repeatedly merging positions until this is no longer possible.

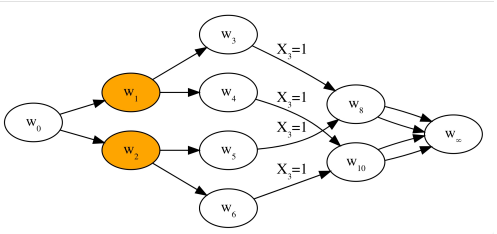
These preliminaries aside, the construction algorithm for the ancestral CEG is as follows:

Definition 47 *For disjoint query on sets of positions $W = \{B_1 \cup B_2 \cup C\}$ an ancestral CEG is constructed according to Algorithm 48*

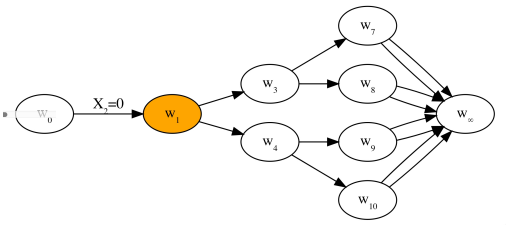
Algorithm 48 *Given sets of positions $W = \{B_1 \cup B_2 \cup C\}$, construct the following graph:*



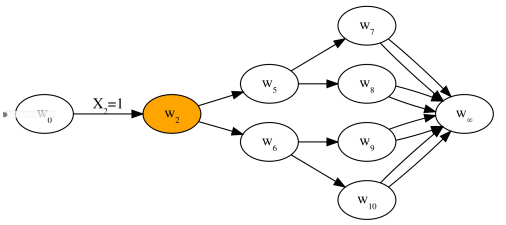
(a) Conditioned subgraph for query $X_1 \perp\!\!\!\perp X_4 | X_3 = 0$



(b) Conditioned subgraph for query $X_1 \perp\!\!\!\perp X_4 | X_3 = 1$



(c) Conditioned subgraph for query $X_1 \perp\!\!\!\perp X_4 | X_2 = 0$



(d) Conditioned subgraph for query $X_1 \perp\!\!\!\perp X_4 | X_2 = 1$

Figure 4.4: Subgraphs shown for each conditioning context for Example 49 that reveal isomorphic trees up to a relabelling of the colours.

1. *Take the ancestral vertex and edge sets* $V\left(\mathcal{C}^0_{\mathrm{An}(W)}\right)$ *and* $E\left(\mathcal{C}^0_{\mathrm{An}(W)}\right)$.
2. *Import the colouring for* $\mathcal{C}^0_{\mathrm{An}(W)}$ *from* $\mathcal{C}$.
3. *Examine the conditioned subgraphs* $\mathcal{C}^c_{\Lambda_{\mathrm{An}(W)}}$ *for each context* $c \in \mathbb{X}(w)$ *for all* $w \in C$ *and merge* $w \in C$ *to a new node* w^*_c *if all subtrees upstream and downstream of the set* C *are isomorphic. Denote this graph as* $\mathcal{C}_{\Lambda_{\mathrm{An}(W)}}$.
4. *Select the graph from the equivalence class of* $\mathcal{C}_{\Lambda_{\mathrm{An}(W)}}$ *with the ancestral ordering* $I_{\mathrm{An}(B_1 \cup B_2 \cup C)}$ *for the query.*
5. *Take the minimal graph by merging all ancestral positions, denoted* $\mathcal{C}^m_{\Lambda_{\mathrm{An}(W)}}$.
6. *Separate the ancestral graph into components with the cut vertices representing the sink node of one component and the root of the subsequent component, denoted* $\mathcal{C}^{m\prime}_{\Lambda_{\mathrm{An}(W)}}$.

Dependence in a BN relies on having pathways that do not travel through the conditioning set. In the CEG, dependent pathways must not *need* to traverse a cut vertex in the ancestral CEG construction. The CEG d-separation process mirrors the process for a BN, and permits a construction of a condensed ancestral graph. These conditions give a non-trivial definition of ancestrality.

(a) The merged graph for $\mathcal{C}_{\mathcal{G}}$. (b) The merged and minimal graph for $\mathcal{C}_{\mathcal{G}}$.

Figure 4.5: Construction of the ancestral CEG for the corresponding CEG of the BN in Figure 4.2

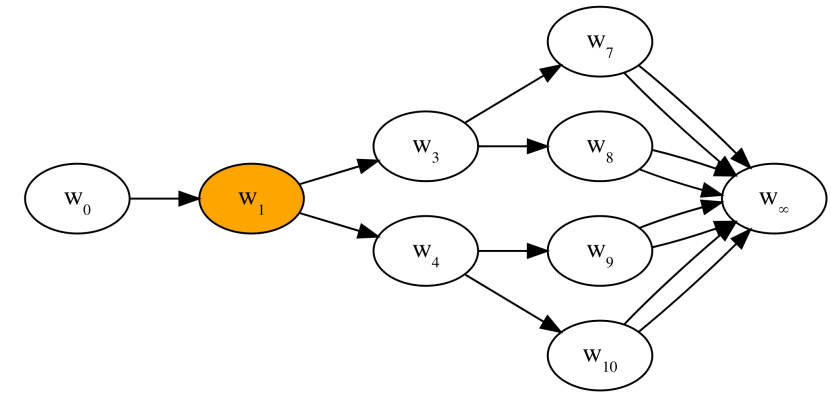


Figure 4.6: The merged and minimal ancestral graph for query two in Example 49. The ancestral graph was adapted from the conditioned subgraphs in Figure 4.4. No new positions required merging and the resultant graph is minimal.

We will first show how the ancestral construction for a CEG equivalent with a BN can be used to address the same queries.

Example 49 *(Construction of an Ancestral CEG) Figure 4.2a shows an example of a BN, $\mathcal{G}$. $X_i \in \{0,1\}$ for $X_i \in \mathcal{G}$. The equivalent CEG, $\mathcal{C}_{\mathcal{G}}$, is shown in Figure 4.3. This has the ancestral ordering $X_2 \prec X_1 \prec X_3 \prec X_4$ according to the algorithm in Definition 45. From the moralized graph of the BN in Figure 4.2b we can read two queries that we will test in $\mathcal{C}_{\mathcal{G}}$: $X_1 \perp\!\!\!\perp X_4|\{X_3\}$ and $X_1 \not\perp\!\!\!\perp X_4|\{X_2\}$. For the both queries, $X_1 \perp\!\!\!\perp X_4|\{X_3\}$, the initial ancestral vertex and edge sets are given by $V(\mathcal{C}_{\mathcal{G}})$ and $E(\mathcal{C}_{\mathcal{G}})$ respectively. The conditioned subgraphs for the two queries are given by Figure 4.4. Figures 4.4a and 4.4b show the subgraph for query $X_1 \perp\!\!\!\perp X_4|\{X_3\}$ where we condition $X_3 = 0$ and $X_3 = 1$ respectively. The upstream and downstream subtrees are isomorphic, so we merge positions $w_c = \{w_3, w_4, w_5, w_6\}$ shown in Figure 4.5a. This merge creates ancestral positions that are in the same position, so the new minimal graph for query $X_1 \perp\!\!\!\perp X_4|\{X_3\}$ is shown in Figure 4.5b. When this graph is separated into components at each of the cut vertices $\{w_0, w_{12}, w_c\}$ in Figure 4.5b, this confirms that $X_1 \perp\!\!\!\perp X_4|\{X_3\}$, $w_{12} \perp\!\!\!\perp \{w_8, w_{10}\}|w_c$.*

For the second query, $X_1 \not\perp\!\!\!\perp X_4|\{X_2\}$, the merged and minimal graph is shown in Figure 4.6. The ancestral positions again have the ancestral ordering $X_2 \prec X_1 \prec X_3 \prec X_4$. When the graph in Figure 4.6 is separated into components at the cut vertices w_0 and w_1, representing X_2 and X_1 respectively, there is still a pathway from w_1 to $\{w_7, w_8, w_9, w_{10}\}$ representing X_4.

4.3.2 A Valid BN for the CEG

The relationship between the incidence and floret variables of a CEG can form a valid BN, denoted as $\mathcal{B}(\mathcal{C})$. To ensure that the dependence relationships between the variables are valid, the extended ancestral graph is a function of the vertex and edge sets derived in the ancestral construction.

Construct $\mathcal{B}(\mathcal{C})$ by indexing the ancestral positions $v_i \in \mathbb{V}(\mathcal{C})$ for $i = 0, 1, 2, \ldots, n$ in the ancestral ordering $I_{\mathrm{An}(B_1 \cup B_2 \cup C)}$. Then, order the $2n+1$ variables so that we introduce any non-root incident variable before its floret variable:

$$X(v_0), I(v_1), X(v_1), I(v_2), X(v_2), \ldots, I(v_i), X(v_i), \ldots, I(v_n), X(v_n).$$

Definition 50 *For disjoint sets $B_1, B_2, C \in \mathcal{C}$ let the **extended ancestral graph** $\mathcal{B}\left(\mathcal{C}^{m\prime}_{\Lambda_{\mathrm{An}(W)}}\right)$ be the directed acyclic graph with vertex set inherited from the ancestral CEG $\Lambda^{m\prime}_{\mathrm{An}(B_1 \cup B_2 \cup C)}$ in the ancestral ordering $I_{\mathrm{An}(B_1 \cup B_2 \cup C)}$ with vertex set:*

$$\{X(v_0), I(v_1), X(v_1), I(v_2), X(v_2), \ldots, I(v_i), X(v_i), \ldots, I(v_n), X(v_n)\}.$$

$X(v_0)$ has no parents; $X(v_i)$ has as its single parent $I(v_i)$; $I(v_i)$ has no parents if the set of positions is a cut vertex $v_i^* \in V_0$ and parents $\boldsymbol{X}_{\mathrm{Pa}(v_i)}$ if otherwise. $\mathrm{Pa}(v_i)$ *is the parent set of v_i' in $\Lambda^{m\prime}_{\mathrm{An}(B_1 \cup B_2 \cup C)}$, $v_i \in \mathbb{V}\left(\Lambda^{m\prime}_{\mathrm{An}(B_1 \cup B_2 \cup C)}\right)$.*

Note here that, because $\Lambda^{m\prime}_{\mathrm{An}(B_1 \cup B_2 \cup C)}$ is minimal, the vertex set may be of smaller cardinality than $\mathrm{An}(W)$ because the colouring may enable us to merge some of the positions of $\mathcal{C}$. This is because the colouring of the ancestral CEG encodes more information than the extended ancestral graph, but the latter enables us to leverage the d-separation theorem for BNs in the proof of d-separation for CEGs.

Lemma 51 *$\mathcal{B}\left(\Lambda^{m\prime}_{\mathrm{An}(B_1 \cup B_2 \cup C)}\right)$ for any set $W = \{B_1 \cup B_2 \cup C\} \subseteq \mathbb{W}(\mathcal{C})$ are valid BNs.*

Proof. This derives immediately from the definition of the DAG of a BN and the Equations 4.10, 4.11, 4.12. ■

Lemma 51 confirms that d-separation for BNs applies to extended ancestral graphs.

4.4 Querying Conditional Independences on a CEG

4.4.1 A Theorem for D-separation in the CEG

Suppose disjoint subsets B_1 and B_2 are subsets of the positions $\mathbb{W}(\mathcal{C})$ and C represent the set of positions with singleton edges emerging in the conditioned subgraphs of $\Lambda_{A(W)}$. The original intrinsic event becomes an amenable event in the ancestral CEG and C represents the set of ancestral positions that are merged from the set of original positions $A(W)$.

$$F(C|\mathbb{V}) = \bigcap_{v \in C} \{I(v|\mathbb{V}(\mathcal{C})) = 1\}$$

We are interested in discovering whether or not on the basis of $\mathcal{C}$ we can assert

$$\boldsymbol{X}_{B_1} \perp\!\!\!\perp \boldsymbol{X}_{B_2} | F(C|\mathbb{V}) \tag{4.13}$$

To do this we consider the ancestral graph of the subsets and the incidence variable of the set of positions in question.

Definition 52 *For non-overlapping sets of positions in $B_1, B_2 \in W(\mathcal{C})$ and intrinsic event $F(C|\mathbb{V})$, B_1 is d-separated from B_2 written as $B_1 \perp_d B_2 | F(C|\mathbb{V})$ if there is no directed pathway from the edge of floret $\mathcal{F}_{B_1}$ to $\mathcal{F}_{B_2}$ in $\mathcal{C}_{A(B_1 \cup B_2 \cup C)}$.*

In CEG d-separation, the construction of the ancestral graph attempts to force a cut vertex to appear in the conditioned, ancestral subgraph. Choosing the correct order via permissible swaps is equivalent to a query-specific moralization process for a DAG of a BN.

4.4.2 Sufficient Conditions

Lemma 53 shows that d-separation in the CEG is a sufficient condition for $\boldsymbol{X}_{B_1} \perp\!\!\!\perp \boldsymbol{X}_{B_2} | F(C|\mathbb{V}(\mathcal{C}))$.

Lemma 53 *Given a query on a set of positions, $W = \{B_1 \cup B_2 \cup C\}$, if $B_1 \perp_d B_2 | F(C|\mathbb{V}(\mathcal{C}))$ in the ancestral subgraph $\mathcal{C}^{m\prime}_{\Lambda_{\mathrm{An}(W)}}$, then*

$$\boldsymbol{X}_{B_1} \perp\!\!\!\perp \boldsymbol{X}_{B_2} | F(C|\mathbb{V}(\mathcal{C})).$$

Proof. If $B_1 \perp_d B_2 | F(C|\mathbb{V}(\mathcal{C}))$ in $\mathcal{C}^{m\prime}_{\Lambda_{\mathrm{An}(W)}}$ by the definition of CEG d-separation, we know that there is no path from the edge floret $\mathcal{F}(w_1)$ to $\mathcal{F}(w_2)$ for any $w_1 \in B_1$ and $w_2 \in B_2$. The positions w_1 and w_2 are not in the same two-connected component, because otherwise, this would violate the CEG d-separation definition.

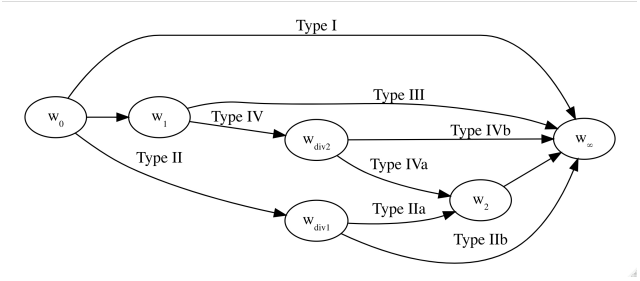


Figure 4.7: Types of paths in a CEG.

Consequently, $X(w_1)$ and $X(w_2)$ for any $w_1 \in B_1$ and $w_2 \in B_2$ belong to different two-connected components by the construction of the extended ancestral graph, $\mathcal{B}\left(\mathcal{C}^{m\prime}_{\Lambda_{\text{An}(W)}}\right)$. $\boldsymbol{X}_{B_1} \perp\!\!\!\perp \boldsymbol{X}_{B_2} | F(C|\mathbb{V}(\mathcal{C}))$ is confirmed by d-separation for BNs. There are no edges between two-connected components in the extended ancestral graph. No vertices in different two-connected components have a common child, so moralization does not introduce any additional pathways. ■

4.4.3 Necessary Conditions

Demonstrating d-separation in the ancestral graph is linked to the existence of certain types of pathways in $\mathcal{C}$. We can define a set of pathways on $\mathcal{C}$ according to whether they include subpaths involving either $w_1 \in B_1$ or $w_2 \in B_2$. These paths are shown in Figure 4.7. Type IIa paths, denoted Λ_{01} represent all the paths passing through $w_2 \in B_2$ but not $w_1 \in B_1$. Type IIb paths, formally denoted Λ_{00} represent all the paths passing through neither $w_1 \in B_1$ nor $w_2 \in B_2$ that share edges with paths that do go to $w_1 \in B_1$. Type IVa paths, denoted Λ_{11} pass through both $w_1 \in B_1$ and $w_2 \in B_2$. Type IVb paths, denoted Λ_{10} represent all the paths passing through $w_1 \in B_1$ but not $w_2 \in B_2$ that share edges with paths that go from w_1 to w_2.

Type I subpaths pass through neither $w_1 \in B_1$ nor $w_2 \in B_2$ and do not share any edges with subpaths going to w_1. Type III subpaths pass through $w_1 \in B_1$, but do not share any edges with subpaths from w_1 to w_2. These edge types will be important for the proof of necessity. Figure 4.7 shows all path types in a CEG.

Formal definitions of the pathways are as follows for $w_1 \in B_1, w_2 \in B_2$:

$$\lambda_{00} = \{\lambda \in \Lambda(\mathcal{C}) : w_1, w_2, \notin \lambda \wedge \exists e \in \lambda : e \in \lambda(w_0, w_1) \cup \lambda(w_0, w_2)\}$$

$$\Lambda_{00} = \bigcup \lambda_{00} \in \Lambda(\mathcal{C}) : w_1, w_2 \notin \lambda_{00} \text{ and } e(\lambda_{00}) \in \lambda(w_0, w_2) \text{ for } e \in \lambda_{00}.$$

λ_{00} may also contain edges such that

$$e(\lambda_{00}) \in \lambda(w_0, w_1).$$

$$\Lambda_{01} = \bigcup \lambda_{01} \in \Lambda(\mathcal{C}) : w_1 \notin \lambda_{01} \text{ and } w_2 \in \lambda_{01}$$

and

$$e(\lambda_{01}) \in \lambda(w_0, w_2) \text{ for } e \in \lambda_{01}.$$

Again, λ_{01} may also contain edges such that

$$e(\lambda_{01}) \in \lambda(w_0, w_1).$$

$$\Lambda_{10} = \bigcup \lambda_{10} \in \Lambda(\mathcal{C}) : w_1 \in \lambda_{10} \text{ and } w_2 \notin \lambda_{10}$$

and some edges

$$e(\lambda_{10}) \in \lambda(w_1, w_2) \text{ for } e \in \lambda_{10}.$$

λ_{10} may also contain edges such that

$$e(\lambda_{10}) \in \lambda(w_1, w_2).$$

$$\Lambda_{11} = \bigcup \lambda_{11} \in \Lambda(\mathcal{C}) : w_1 \in \lambda_{11} \text{ and } w_2 \in \lambda_{11}.$$

The union of these pathways represents all possible root-to-sink paths in the CEG in the event that $\boldsymbol{X}_{B_1} \perp\!\!\!\perp \boldsymbol{X}_{B_2}$. Lemma 54 means that we have independence between two positions if we do not have pathways of Type I or III from Figure 4.7.

Lemma 54 *If* $\boldsymbol{X}_{B_1} \perp\!\!\!\perp \boldsymbol{X}_{B_2}$*, then* $\Lambda(\mathcal{C}_{A(B_1 \cup B_2 \cup C)}) = \Lambda_{00} \cup \Lambda_{01} \cup \Lambda_{10} \cup \Lambda_{11}$

Proof. It suffices to show that there are no paths of two types.

1. $\lambda'_{00} \in \Lambda(\mathcal{C})$ such that $w_1, w_2, \notin \lambda'_{00}$ and $\nexists\, e \in \lambda'_{00}$ such that $e \in \lambda(w_0, w_2)$.
2. $\lambda'_{10} \in \Lambda(\mathcal{C})$ such that $w_1 \in \lambda'_{10}$ and $w_2 \notin \lambda'_{10}$ and $\nexists\, e \in \lambda'_{10}$ such that $e \in \lambda(w_1, w_2)$.

For the first type of the first path, let

$$\sum_{\lambda'_{00} \in \Lambda'_{00}} \boldsymbol{\pi}(\lambda'_{00}) = \alpha$$

and

$$\sum_{\lambda_{00}\in\Lambda_{00}} \boldsymbol{\pi}(\lambda_{00}) + \sum_{\lambda_{01}\in\Lambda_{01}} \boldsymbol{\pi}(\lambda_{01}) = \beta$$

where $\alpha, \beta < 1$. We can write

$$p\left(\sum_{\lambda_{01}\in\Lambda_{01}} \boldsymbol{\pi}(\lambda_{01})\right) = \beta\gamma$$

for some $\gamma \leq 1$ and probability p. Let $P(I(w_2) = 1|X(w_1) = 0) = p$.

$$\frac{P\left(\sum_{\lambda_{01}\in\Lambda_{01}} \boldsymbol{\pi}(\lambda_{01})\right)}{\sum_{\lambda'_{00}\in\Lambda'_{00}} \boldsymbol{\pi}(\lambda'_{00}) + \sum_{\lambda_{00}\in\Lambda_{00}} \boldsymbol{\pi}(\lambda_{00}) + \sum_{\lambda_{01}\in\Lambda_{01}} \boldsymbol{\pi}(\lambda_{01})} = \frac{\beta\gamma}{\alpha+\beta}$$

$$\gamma = \left(\frac{\alpha+\beta}{\beta}\right) p \tag{4.14}$$

We could find values for α, β, γ, p so this could hold, but for true independence in the graph, this must hold for all possible values of α, β, p. We require that $p > 0$, but in theory it is possible for $\alpha = 0$ or $\beta = 0$. If $\alpha = 0$, Equation 4.14 is satisfied by $\gamma = p$. In the case where $\beta = 0$, $\alpha > 0$, Equation 4.14 must hold $\forall\, \alpha, \beta, p \in (0, 1)$ such that $\alpha + \beta = 1$. Consider the case when $\alpha = 0.8, \beta = 0.1, p = 0.9$. Then $\gamma = 8.1$. This is not possible, thus, there are no paths $\lambda'_{00} \in \Lambda'_{00}$.

For the second type of path,

$$P(I(w_2) = 1|X(w_1) = i) = p \qquad \forall\, i \in 0, 1, 2, \ldots$$

But if $\exists \lambda'_{10}$, then $\exists$ an edge e_{i1} emanating from w_1 such that $e_{i1} \notin \lambda(w_1, w_2)$. Let the edge label of $X(w_1)$ corresponding to this edge be m. Then

$$P(I(w_2) = 1|X(w_1) = m) = 0.$$

Hence there are no paths $\lambda'_{01} \in \Lambda'_{01}$. ■

This Lemma exposes two pathways that induce dependence in the CEG, Type I and Type III paths.

Theorem 55 *For a query on sets of disjoint positions $W = \{B_1, B_2, C\} \in \mathbb{W}(\mathcal{C})$, if*

$$\boldsymbol{X}_{B_1} \perp\!\!\!\perp \boldsymbol{X}_{B_2}|F(C|\mathbb{V}(\mathcal{C}) \Rightarrow$$

$$B_1 \perp_d B_2|F(C|\mathbb{V}(\mathcal{C})) \text{ in } \mathcal{C}^{m\prime}_{\Lambda_{\mathrm{An}(W)}}$$

Proof.

Consider the contrapositive:

$$B_1 \not\perp_d B_2 | F(C|\mathbb{V}(\mathcal{C}))$$

in $\mathcal{C}^{m\prime}_{\Lambda_{\mathrm{An}(W)}}$ then

$$\boldsymbol{X}_{B_1} \not\perp\!\!\!\perp \boldsymbol{X}_{B_2} | F(C|\mathbb{V}(\mathcal{C})))$$

in $\mathcal{C}$.

Note that for any $v_1 \in B_1$ and $v_2 \in B_2$, there is a path $\lambda(v_1, v_2)$. The ancestral positions v_1 and v_2 are in the same two-connected component of $\mathcal{C}^{m\prime}_{\Lambda_{\mathrm{An}(W)}}$.

Now we want to show that if this is the case,

$$X(v_1) \perp\!\!\!\perp X(v_2)$$

is certainly false. As

$$I(v_1) \not\perp\!\!\!\perp I(v_2) \Rightarrow X(v_1) \not\perp\!\!\!\perp X(v_2), \tag{4.15}$$

it suffices to show that

$$I(v_1) \not\perp\!\!\!\perp I(v_2).$$

For each setting of values in the conditioning set $c \in C$, it is sufficient to verify that:

$$p(I(v_2)|I(v_1), \boldsymbol{c}) \neq p(I(v_2)|\boldsymbol{c}).$$

If the intrinsic event $F(C|\mathcal{V}(\mathcal{C}))$ consolidates a cut to a single vertex in $\mathcal{C}^{m\prime}_{\Lambda_{\mathrm{An}(W)}}$, then it must be either upstream or downstream of both B_1 and B_2. Otherwise, it would have forced B_1 and B_2 in to different two connected components. Each cut between B_1 and B_2 has at least two ancestral positions.

For each setting of values $C = c$, we can assign these probabilities in the following way:

$I(v_1)\backslash I(v_2)$	0	1	
0	p_{00}	p_{01}	$1 - p_1$
1	p_{10}	p_{11}	p_1
	$1 - p_2$	p_2	

By definition $I(v_1) \perp\!\!\!\perp I(v_2)$ always in $\mathcal{C}$ when

$$p_{00}p_{11} = p_{01}p_{10}. \tag{4.16}$$

All atomic probabilities have to be strictly positive $p_1, p_2 > 0$, so that

$$p_{11} = p_1 p_2 > 0.$$

Otherwise, this violates the assumption that we have one root-to-sink path in the ancestral CEG $\lambda : v_1, v_2 \in \lambda$. There are now three cases to consider.

The first case is when $p_1 = 1$ –when by definition all root-to-sink paths must pass through v_1. Then

$$I(w_1) \perp\!\!\!\perp I(w_2)$$

because

$$p_{00} + p_{01} = 1 - p_1 = 0.$$

However, the inverse of Equation 4.15 is not necessarily true. The independence of the incident ancestral position vectors does not necessarily imply the independence of the edge ancestral position vectors.

In this case, assume without loss of generality that v_1 is the root node of $\mathcal{C}^{m\prime}_{\Lambda_{\text{An}(W)}}$because any situation where everything passes through it must be the root node, and we do not have any squares. If the conditioning set of ancestral positions C is consolidated to a single ancestral position in $\mathcal{C}^{m\prime}_{\Lambda_{\text{An}(W)}}$, then it must be upstream of both sets B_1 and B_2. Otherwise, this would create a cut vertex between the sets, violating the assumption that there is a path between sets. If the set of ancestral positions C is not consolidated to a single ancestral position in $\mathcal{C}^{m\prime}_{\Lambda_{\text{An}(W)}}$, then, this tells us that the minimal subgraphs either upstream or downstream of the conditioning set with the ancestral ordering are not isomorphic. In either case, there are no permissible swaps. For each path $\lambda(v_1, v_2)$, there must be a path in $\mathcal{C}^{m\prime}_{\Lambda_{\text{An}(W)}}$that passes through v_1, but does not share any subpaths with the path $\lambda(v_1, v_2)$, denoted $\lambda(v_1, \overline{v}_2)$. There are no cut-vertices in the ancestral positions between B_1 and B_2, so each set of ancestral positions at each length away from v_0 has at least two ancestral positions. Thus, $\lambda(v_1, \overline{v}_2)$ can be constructed from the set of alternate ancestral positions. Thus, there is a setting of the probabilities in $\mathcal{C}^{m\prime}_{\Lambda_{\text{An}(W)}}$such that $X(v_1) \not\perp\!\!\!\perp X(v_2)$.

The second case is when $p_2 = 1$ – when by definition all root-to-sink paths must pass through v_2. Then

$$I(v_1) \perp\!\!\!\perp I(v_2)$$

because

$$p_{00} + p_{10} = 1 - p_2 = 0,$$

Without loss of generality, v_2 is the sink node of a two-connected component in $\mathcal{C}^{m\prime}_{\Lambda_{\text{An}(W)}}$. By our default assumption, we know there exists a path $\lambda(v_1, v_2)$ for

$v_1 \in B_1$ and $v_2 \in B_2$. Mirroring the argument above, every set of ancestral positions at lengths away from the sink node has at least two ancestral positions in it. Thus, we can construct an alternative path $\lambda(v_1, \overline{v}_2)$ for $v_1 \in B_1$ and $\overline{v}_2 \in B_2$ that does not share any subpaths with $\lambda(v_1, v_2)$. There is a setting of the probabilities in $\mathcal{C}^{m\prime}_{\Lambda_{\mathrm{An}(W)}}$ such that $X(v_1) \not\perp\!\!\!\perp X(v_2)$.

The final case occurs when

$$p_1, p_2 < 1.$$

This is the most interesting since now for

$$I(v_1) \perp\!\!\!\perp I(v_2)$$

to hold we would need all the subsets of paths corresponding to

$$\{I(v_1) = i, I(v_2) = j\}$$

$\forall\, i, j = \{0, 1\}$ must be non empty. If they are empty, then to not violate

$$I(v_1) \perp\!\!\!\perp I(v_2)$$

we would need a probability on the opposite side to be zero as well, which would degenerate to one of the above cases.

In this case, the pathways in $\mathcal{C}^{m\prime}_{\Lambda_{\mathrm{An}(W)}}$ still have two vertices in each level, and thus we can always construct a pathway of Type I or III. From the contrapositive of Lemma 54, we know that this implies

$$\boldsymbol{X}_{B_1} \not\perp\!\!\!\perp \boldsymbol{X}_{B_2} | F(C|\mathbb{V}(\mathcal{C})))$$

in $\mathcal{C}$.

■

This construction allows us to identify topological configurations conducive to reading off the conditional independence relationships. Theorem 55 subsumes the pre-existing theorems we have proved for these special cases. The following established results now all follow as a special case:

Corollary 56 *For $B_1, B_2, w_c \in V(\mathcal{C})$ such that $B_1 \cap B_2 = \emptyset$ and w_c is a cut vertex such that $B_1 \preceq w_c \preceq B_2$ then*

$$\boldsymbol{X}(B_1) \perp\!\!\!\perp \boldsymbol{X}(B_2) | \boldsymbol{I}(w_c)$$

Proof. We begin by constructing $\mathcal{C}^{m\prime}_{\Lambda_{\text{An}(W)}}$ from the given assumptions. $A(B_1 \cup B_2 \cup w_c)$ is the ancestral vertex set. Conditioning on $\boldsymbol{I}(w_c)$ induces subgraphs that are isomorphic downstream. If the trees were not isomorphic, there would not be a cut vertex, w_c. To determine the ancestral ordering $I_{B_1 \cup B_2 \cup w_c}$, there is always a series of swaps that changes the ordering from $B_1 \prec w_c \preceq B_2$ to $B_1 \prec B_2 \prec w_c$. Then, the positions $w \in B_2$ inherit the colouring from w_c and thus, $w \in B_2$ is merged to a new cut vertex in the ancestral positions of the minimal $\mathcal{C}^{m\prime}_{\Lambda_{\text{An}(W)}}$. This imposes separation between all $w_1 \in B_1$ and $w_2 \in B_2$, and by the d-separation theorem for CEGs, we can conclude that

$$\boldsymbol{X}(B_1) \perp\!\!\!\perp \boldsymbol{X}(B_2) | \boldsymbol{I}(w_c)$$

■

Theorem 24 is a corollary of our ancestral construction.

Corollary 57 *If W' is a fine cut, then*

$$\boldsymbol{X}_{\prec W'} \perp\!\!\!\perp \boldsymbol{X}_{W' \preceq} | \boldsymbol{I}_{W'}$$

If W' is a cut, then

$$\boldsymbol{X}_{\prec W'} \perp\!\!\!\perp \boldsymbol{X}_{W'} | \boldsymbol{I}_{W'}$$

Proof. The ancestral vertex set is given by $A(W') \cup w_\infty$. In $\mathcal{C}^0_A$, all edges from the cut W' map to w_∞. This means that they all the downstream trees are isomorphic. The upstream trees are isomorphic because W' is a fine cut. Thus W' is consolidated to a single cut vertex w_c which d-separates $\boldsymbol{X}_{\prec W'}$ from $\boldsymbol{X}_{W' \preceq}$. ■

4.5 CEG D-separation Extends the BN D-separation

Every BN can be equivalently written as a CEG. This section demonstrates that for this case, the d-separation theorem for CEGs simply replicates the results of d-separation for a BN as expected.

Any dependence query of a BN answered by the d-separation theorem can also be answered by constructing an equivalent CEG and performing CEG d-separation. Any CEG equivalent to a BN will be stratified and coloured according to the existing conditional independence relationships.

Theorem 58 *For disjoint sets of random variables $\boldsymbol{X}_{B_1}, \boldsymbol{X}_{B_2}$ and $\boldsymbol{X}_C$ in the BN $\mathcal{G}$, if $\boldsymbol{X}_{B_1} \perp_d \boldsymbol{X}_{B_2} | \boldsymbol{X}_C$ in $\mathcal{G}^{m\prime}_{A(\boldsymbol{X}_{B_1} \cup \boldsymbol{X}_{B_2} \cup \boldsymbol{X}_C)}$, then for an equivalent CEG $\mathcal{C}_\mathcal{G}$ and corresponding sets of positions $W = \{B_1 \cup B_2 \cup C\} \in \mathbb{W}(\mathcal{C}_\mathcal{G})$ and intrinsic event*

$F(C|\mathbb{V}(\mathcal{C}_\mathcal{G}))$ corresponding to the original conditioning set of random variables in $\mathcal{C}_\mathcal{G}$, $B_1 \perp_d B_2 | F(C|\mathbb{V}(\mathcal{C}_\mathcal{G}))$.

Proof.

First, we demonstrate that a query between two sets of variables can be reduced to a collection of queries on individual random variables. That is, a query on the set:

$$\boldsymbol{X}_{B_1} \perp\!\!\!\perp \boldsymbol{X}_{B_2} | \boldsymbol{X}_C = x_c.$$

can be reduced to a collection of queries of the form

$$X_i \perp\!\!\!\perp X_j | \boldsymbol{X}_C = x_c$$

for $w_1 \in B_1$ and $w_2 \in B_2$.

$$X_i, \boldsymbol{X}_{B_1 \setminus i} \perp\!\!\!\perp X_j, \boldsymbol{X}_{B_2 \setminus j} | \boldsymbol{X}_C = x_c$$

By contraction

$$X_i, \boldsymbol{X}_{B_1 \setminus i} \perp\!\!\!\perp \boldsymbol{X}_{B_2 \setminus j} | X_j, \boldsymbol{X}_C = x_c \text{ and } X_i, \boldsymbol{X}_{B_1 \setminus i} \perp\!\!\!\perp X_j | \boldsymbol{X}_C = x_c$$

By symmetry

$$\boldsymbol{X}_{B_2 \setminus j} \perp\!\!\!\perp X_i, \boldsymbol{X}_{B_1 \setminus i} | X_j, \boldsymbol{X}_C = x_c \text{ and } X_j \perp\!\!\!\perp X_i, \boldsymbol{X}_{B_1 \setminus i} | \boldsymbol{X}_C = x_c$$

By strong decomposition

$$\boldsymbol{X}_{B_2 \setminus j} \perp\!\!\!\perp \boldsymbol{X}_{B_1 \setminus i} | X_i, X_j, \boldsymbol{X}_C = x_c \text{ and } \boldsymbol{X}_{B_2 \setminus j} \perp\!\!\!\perp X_i | \boldsymbol{X}_{B_1 \setminus i}, X_j, \boldsymbol{X}_C = x_c \text{ and}$$

$$X_j \perp\!\!\!\perp \boldsymbol{X}_{B_1 \setminus i} | \boldsymbol{X}_{B_2 \setminus j}, \boldsymbol{X}_C = x_c \text{ and } X_j \perp\!\!\!\perp X_i | \boldsymbol{X}_C = x_c$$

By symmetry

$$X_i \perp\!\!\!\perp X_j | \boldsymbol{X}_C = x_c \text{ and } X_i \perp\!\!\!\perp \boldsymbol{X}_{B_2 \setminus j} | \boldsymbol{X}_{B_1 \setminus i}, X_j, \boldsymbol{X}_C = x_c \text{ and}$$

$$\boldsymbol{X}_{B_1 \setminus i} \perp\!\!\!\perp \boldsymbol{X}_{B_2 \setminus j} | X_i, X_j, \boldsymbol{X}_C = x_c \text{ and } \boldsymbol{X}_{B_1 \setminus i} \perp\!\!\!\perp X_j | X_i, \boldsymbol{X}_C = x_c$$

Using this argument

$$\boldsymbol{X}_{B_1} \perp\!\!\!\perp \boldsymbol{X}_{B_2} | \boldsymbol{X}_C$$

is equivalent to statements of the form

$$X_i \perp\!\!\!\perp X_j | \boldsymbol{X}_C = x_c$$

By the BN d-separation theorem,

$$X_i \perp_d X_j | \boldsymbol{X}_C = x_c$$

$\mathcal{C}_\mathcal{G}$ is a stratified CEG, so each set of positions corresponding to individual random variables X_i, X_j, and $\boldsymbol{X}_C$, denoted $W = \{w_{X_i} \cup w_{X_j} \cup w_{\boldsymbol{X}_C}\} \in \mathbb{W}(\mathcal{C}_\mathcal{G})$ is at the same level from the root node.

To construct $\mathcal{C}^{m\prime}_{\Lambda_{\text{An}(W)}}$, we must determine the ancestral ordering $I_{A(w_{X_i}, w_{X_j}, w_{\boldsymbol{X}_C})}$. Assuming, without loss of generality that $w_{X_i} \prec w_{X_j}$, the three relevant orderings are:

i) $w_{X_i} \preceq w_{X_j} \preceq w_{\boldsymbol{X}_C}$

ii) $w_{X_i} \preceq w_{\boldsymbol{X}_C} \preceq w_{X_j}$

iii) $w_{\boldsymbol{X}_C} \preceq w_{X_i} \preceq w_{X_j}$

For case i), $w_{X_i} \preceq w_{X_j} \preceq w_{\boldsymbol{X}_C}$, we want to show that there must be a cut vertex in $\mathcal{C}^{m\prime}_{\Lambda_{\text{An}(W)}}$between w_{X_i} and w_{X_j}. Because $\mathcal{C}_\mathcal{G}$ is stratified, the subtrees conditioning on the event $F(C|\mathbb{V}(\mathcal{C}_\mathcal{G}))$ will consolidate to a single node. If not, this would violate the assumption that $\mathcal{C}_\mathcal{G}$ is stratified. This forces the positions in $w_{\boldsymbol{X}_C}$ to merge to a single cut vertex. The ancestral ordering must juxtapose w_{X_i} and w_{X_j}. If not, then there is some other variable in $\mathcal{G}$, say X_k, there would be result in a path X_i, X_k, X_j. However, this chain would violate the initial assumption that $X_i \perp_d X_j | \boldsymbol{X}_C = x_c$ Each colour in the set of positions w_{X_j} is the same. Otherwise, this would induce a dependence in the graph and violate the initial assumption. The subtrees from each position in w_{X_j} are identical. If they were not, this would induce a moralised edge in the original ancestral BN. Thus, the positions in w_{X_j} can be merged to a single vertex in the minimal ancestral graph, and when this graph is separated into components, this confirms that $w_{X_i} \perp_d w_{X_j} | \boldsymbol{X}_C = x_c$.

Again in case ii), when $w_{X_i} \preceq w_{\boldsymbol{X}_C} \preceq w_{X_j}$, it is sufficient to show that there is a cut vertex in the ancestral graph between w_{X_i} and w_{X_j}. Because the ancestral ordering first seeks to minimize the difference between w_{X_i} and w_{X_j}, we know that there must be a chain in $\mathcal{G}$, (X_i, X_C, X_j). There must be no additional pathways between X_i and X_j in $\mathcal{G}$. Again, the subgraphs induced by conditioning on $\boldsymbol{X}_C = x_c$ will necessarily induce isomorphic subtrees upstream and downstream. The positions in $\boldsymbol{X}_C$ merge to a single vertex, inducing a cut vertex between w_{X_i} and w_{X_j}. When

the ancestral graph is separated into two-connected components, then this confirms that $w_{X_i} \perp_d w_{X_j} | \boldsymbol{X}_C = x_c$.

Finally, consider case iii) when $w_{\boldsymbol{X}_C} \preceq w_{X_i} \preceq w_{X_j}$. The subtrees induced by the conditioning set remove the branches where $\boldsymbol{X}_C \neq x_c$. The remaining sets of positions in w_{X_j} all have the same colour. Otherwise, there would be a chain or a moralised edge into X_j. Because there is nothing downstream of w_{X_j}, the positions in that set can be merged to a single vertex in the minimal ancestral graph. This will again separate w_{X_i} and w_{X_j} into different two-connected components in the ancestral graph and confirms that $w_{X_i} \perp_d w_{X_j} | \boldsymbol{X}_C = x_c$.

■

4.6 Discussion

The results proved in the previous section provide a d-separation theorem directly analogous to that of Pearl or Lauritzen, extending their work to a much larger class of models. The d-separation theorem for CEGs can also be used for dynamic CEGs and RDCEGs. The construct of the extended BN, while not strictly necessary for the CEG, is a useful articulation of the random variables in a CEG. This has further ramifications for explaining genuine cause and other mediation formulas in the CEG.

Chapter 5

CEG Diagnostics

> It's not the note you play that's the wrong note—it's the note you play afterward that makes it right or wrong.
>
> Miles Davis

5.1 Background

Bayesian Networks are useful, widely implemented, and one of the best structural tools to use when a set of predetermined measurement variables are available. However, even when elaborated into Object Oriented Bayesian Networks, these structural frameworks are not always ideal: see for example Koller and Pfeffer (1997); Korb and Nicholson (2010). These representations don't always encode all of the symmetries in a problem. Context-specific BNs emerged as one way to address this problem (McAllester et al., 2004; Boutilier et al., 1997; Geiger and Heckerman, 1996), but these approaches often abandon graphical representation of the symmetry.

Event trees respect the symmetries in a problem. However, one problem with event trees is that they do not convey the information about conditional independences encoded in a BN. The class of Chain Event Graphs was therefore designed to express conditional independence relationships encoded in the colouring (Smith and Anderson, 2008). This class of tree-based models is more general than the BN; it includes the context-specific BNs, albeit depicted in a different but equivalent way.

Chain Event Graphs (CEGs) are a useful graphical model representation. They generalise the class of Bayesian Networks (BNs), representing context-specific independence and graphical asymmetry. Furthermore it can be argued that because

they are drawn from a tree-based structure, CEGs allow a more natural way to express a series of unfolding events (Shafer, 1996).

As with other graphical models, CEGs are then populated with distributions, often inferred by data. Typically, these parameters of the distribution can be updated sequentially as more data becomes available. In this setting the routine use of diagnostics is essential. They reveal problematic structural elements, expose when changes in the data are no longer compatible with the model, or alternatively demonstrate its plausibility.

Within the Bayesian paradigm the prequential diagnostics of Dawid (1984) have proved particularly useful and simple to apply. These examine the one-step ahead forecasts of each subsequent observation in a dataset to determine the compatibility of the model with the data. In particular, prequential diagnostics determine how well the model predicts future data based on past performance (Dawid, 1992). These have been used successfully to provide diagnostics for the Bayesian Network class (Cowell et al., 1999).

Prequential diagnostics have since been extended to other graphical models including the Multi-regression Dynamic Model (Costa et al., 2015). Here I extend them to CEGs. The prequential approach is especially attractive for use with this class since its focus is on a model's ability to forecast the future development of a unit in the population given the past. This harmonises beautifully with the type of modelling structure expressed by a CEG which encodes possible future pathways for each unit.

In this chapter I describe the suite of diagnostic monitors developed for detecting ill-fitting CEGs. Section 5.2 explains the meaning and estimation of the Chain Event Graphs and their derivation from the staged trees. In Section 5.3, I review the prequential diagnostics for the Bayesian Network (BN) and define analogous diagnostics for the CEG in Section 5.4. Section 5.5 shows the diagnostics applied to two different examples. First, the Christchurch Health and Development Study (CHDS) example shows the process of households' circumstances that may result in a child being admitted to the hospital. This example demonstrates the ability of the diagnostic monitors to differentiate between candidate models including a BN and two CEGs. The example in Section5.5.2 describes radicalisation data that shows how individuals in a prison may choose to engage in radical activity. Our second example shows how these diagnostics improve model interpretability as I begin to scale the CEG. Together, these examples demonstrate how the diagnostics highlight misspecifications in the structure.

5.2 The Meaning and Estimation of CEGs

5.2.1 Christchurch Data Set

In this chapter I consider two examples to illustrate our methodology. The first has the advantage that it has been subject to various different CEG models and so is already well studied (Barclay and Nicholson, 2015; Cowell and Smith, 2014). The study was conducted at the University of Otago, New Zealand (Fergusson et al., 1986). It encompassed a five year longitudinal study of several explanatory variables including:

- X_{s}: Family social background, a categorical variable differentiating between high and low levels according to educational, socio-economic, ethnic measures, and information about the children's birth.
- X_{e}: Family economic status, a categorical variable distinguishing between high and low status with regard to standard of living.
- X_{l}: Family life events, a categorical variable signalising the existence of low (0 to 5 events), average (6 to 9 events) or high (10 or more events) number of stressful events faced by a family over the five years.
- X_{h}: Hospital admissions, a binary variable indicating whether or not a child in the household was hospitalised.

The aim of the CHDS study was to better understand how the different variables above might relate to one another. Previous studies of the CHDS data demonstrated the flexibility and expressiveness of the CEG model over the BN (Barclay et al., 2013). We will demonstrate below how the diagnostics I develop here pinpoint exactly how the CEG structure can model the processes better than a BN.

The partition specifying the stages a CEG is analogous to specifying conditional independence asserted through the graph of a BN (Dawid, 1979; Studenỳ, 2002). Situations in the same stage are independent conditional on their respective histories and the proofs of can be found in Smith and Anderson (2008); Thwaites and Smith (2010).

For this chapter, we consider the class of stratified CEGs because they offer the most direct comparison to a standard BN.

The $\mathrm{CEG}_{\mathrm{BN}}$ in Figure 5.1 on 86 encodes the same conditional independence relationships as the BN in Figure 5.3 on page 89. The BN in Figure 5.3 models that X_{h} is independent of X_{e} given X_{l} and X_{s}. $\mathrm{CEG}_{\mathrm{BN}}$ in Figure 5.1 encodes this through the colouring in the set of stages representing X_{h}. For $X_{\mathrm{s}} =$ High (or Low),

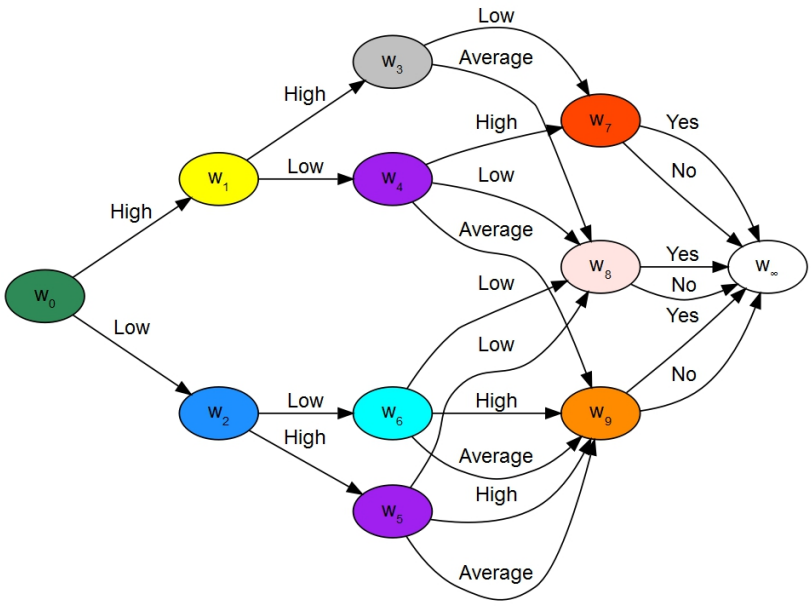


Figure 5.1: CEG_{BN}, a CEG adapted from the BN used in previous CHDS study. X_s corresponds to $\{w_0\}$; X_e to $\{w_1, w_2\}$; X_l to $\{w_3, w_4, w_5, w_6\}$; and X_h to $\{w_7, w_8, w_9\}$

the future development of X_{l} is not dependent on X_{e} The edges for both levels of X_{e} go into the same stages. CEG_{AHC} in Figure 5.2 represents the CEG found by the Agglomerative Hierarchical Clustering (AHC) algorithm.

5.3 BN Prequential Diagnostics

5.3.1 BN Conjugate Dirichlet Analysis

A Bayesian Network G is given by a set of random variables X_i for $\{i \in 1, \ldots, n\}$, each taking different values x_k for $\{k \in 1, \ldots, K_i\}$. The possible configurations of the parents of X_i are denoted $\rho_i = j$ are $\{1, \ldots, q_i\}$. Although the methodology presented here is generic, I will illustrate the use of the prequential methods using the simplest and most widely used sort of prior for the CEG, the product Dirichlet, where for each set of parents of node and values of X_i governed by parameter $\boldsymbol{\theta}_{ijk}$.

Thus suppose we observe $\boldsymbol{y}_i = \{y_1, \ldots, y_m, \ldots, y_M\}$, a series of observations for the variable X_i, where each possible value of each random variable is assigned a Dirichlet prior $\mathcal{D}(\alpha_1, \ldots, \alpha_{K_i})$. In a discrete BN, the entries in the conditional probability tables for a particular parent setting sum to one over all possible levels of the node. That is, the parameter for the ith node with the jth setting of the parents for the kth value, $\theta_{ij} = \sum_{k=1}^{K_i} \boldsymbol{\theta}_{ijk} = 1$.

Setting a Dirichlet prior for each θ_{ij}, permits the conjugate posterior analysis.

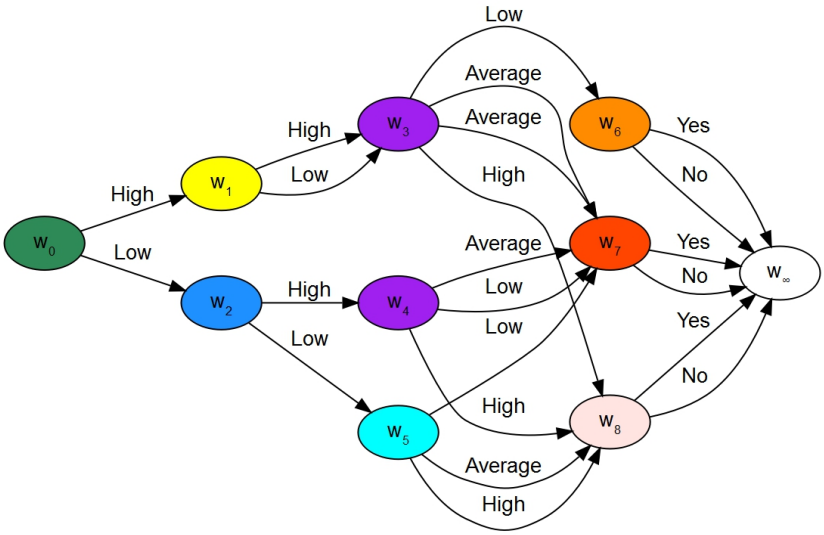


Figure 5.2: CEG_{AHC}, The CEG for the CHDS data found using the AHC algorithm. X_s corresponds to $\{w_0\}$; X_e to $\{w_1, w_2\}$; X_l to $\{w_3, w_4, w_5\}$; and X_h to $\{w_6, w_7, w_8\}$

As data is accumulated about the system, the Dirichlet prior can be updated by adding the counts of the observation to the prior. We can compute a reference Dirichlet prior by taking the highest number of levels of a given variable (X_1 gives an effective sample size of $\alpha = 3$ for the CHDS example) and dividing it by the number of levels outgoing from each situation.

The prequential diagnostics compute the surprise of seeing each subsequent observation given the past observations. Towards that end, our monitors use the likelihood of observing the complete data $\boldsymbol{y}$ as given by Heckerman et al. (1995). Assuming it was randomly sampled, the likelihood of the probability vectors is:

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = c \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{K_i} \theta_{ijk}^{y_{ijk}}$$

where $c = \frac{Y!}{\prod_{k=1}^{K_i} y_{ijk}!}$, where $Y = \sum_{k=1}^{K_i} y_{ijk},$. The parameter for each value and parent pair for each node θ_{ijk} is governed by a Dirichlet distribution. Thus the prior is given by:

$$p(\boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\sum_{k=1}^{K_i} \alpha_{ijk})}{\prod_{k=1}^{K_i} \Gamma(\alpha_{ijk})} \prod_{k=1}^{K_i} \theta_{ijk}^{\alpha_{ijk}-1}$$

Following the conjugate analysis, we obtain the following form of the posterior distribution:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = c\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(\sum_{k=1}^{K_i}\alpha_{ijk}+y_{ijk})}{\prod_{k=1}^{K_i}\Gamma(\alpha_{ijk}+y_{ijk})}\prod_{k=1}^{K_i}\theta_{ijk}^{y_{ijk}+\alpha_{ijk}-1}$$

Which gives us the closed form:

$$p(\boldsymbol{y}) = c\int_{\Theta}\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(\sum_{k=1}^{K_i}\alpha_{ijk}+y_{ijk})}{\prod_{k=1}^{K_i}\Gamma(\alpha_{ijk}+y_{ijk})}\prod_{k=1}^{K_i}\theta_{ijk}^{\alpha_{ijk}+y_{ijk}-1}\,d\boldsymbol{\theta}$$

$$p(\boldsymbol{y}) = c\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(\alpha_+)}{\prod_{k=1}^{K_i}\Gamma(\alpha_{ijk})}\frac{\prod_{k=1}^{K_i}\Gamma(\alpha_{ijk}+y_{ijk}))}{\Gamma(\alpha_++Y)} \tag{5.1}$$

where $\alpha_+ = \sum_{k=1}^{K_i}\alpha_{ijk}$.

5.3.2 Scoring Rules

In this chapter, in order to check the accuracy of the forecasts, we can use the logarithmic scoring rule because of its close links to Bayesian inference through the Bayes factor score.

The temporal ordering, denoted $m = (1, \ldots, M)$, is taken in this case to be the ordering in the dataset. The subsequent prequential methods we derive below all rely on this ordering. Note that there may be scenarios in which we would like to reorder the data by a covariate, or according to external information available about how the sample was collected.

Let y_m denote the mth observation of the data for which y_m is observed at a specific level of the random variable y_k. Let p_m denote the predictive density of observing y_k after learning from the first $m-1$ cases. The logarithmic score of the mth observation of Y taking the value y_k is denoted:

$$S_m = -\log p_m(y_k)$$

There are two methods of standardisation. Relative standardisation examines the logarithmic difference between the penalties under two different models. The absolute difference does not require an alternative model. Instead, we compute a standardised test statistics Z_m using the expectation E_m and variance V_m following Cowell et al. (1999):

$$E_m = -\sum_{k=1}^{K} p_m(y_k) \log p_m(y_k) \tag{5.2}$$

$$V_m = \sum_{k=1}^{K} p_m(y_k) \log^2 p_m(y_k) - E_m^2 \tag{5.3}$$

$$Z_M = \frac{\sum_{m=1}^{M} S_m - \sum_{m=1}^{M} E_m}{\sqrt{\sum_{m=1}^{M} V_m}}. \tag{5.4}$$

For sufficiently large sample sizes under the model assumptions, for all but small indices m, Z_M will have an approximate standard Normal distribution if the model could have plausibly generated the data.

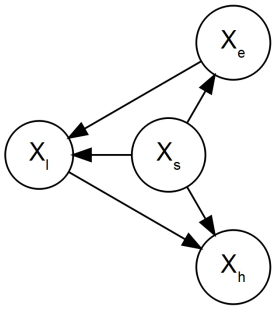


Figure 5.3: BN CHDS: A BN obtained from previous studies of the CHDS data (Barclay et al., 2013).

For the global monitors, we can now examine alternative models under the relative standardisation technique. Our candidate models include the baseline BN shown in Figure 5.3, a CEG based on the BN that includes additional information in Figure 5.1, and another CEG found from the AHC algorithm in Figure 5.2. This enables us to identify structural improvements with an increasingly fine set of monitors.

5.3.3 Diagnostic Monitors for BNs

The prequential methods are similar to cross-validation, with the key difference being that they rely on information from the previous iterations, rather than predicting on the variables excepting the one of interest.

Within a Bayesian framework these diagnostics are especially attractive, because if the estimated conditionals are treated as one-step ahead predictives, then the log marginal likelihood is simply the sum of these scores. So the prequential methods then decompose an aggregate score into scores associated with different

subsets of the contributions to the data. Each such subset can then be scrutinized for its fidelity to the fitted model as it applies to that subset within the context of a full Bayesian analysis.

When most effective, the prequential approach is able to adopt an interpretable and natural ordering of the observational data. When a temporal component is not immediately obvious, it may be helpful to order the data according to some covariate of the observables. For instance, modelling healthcare outcomes might benefit from ordering the data according to the length of time each patient spent in the hospital. The prequential approach is well suited to detect where the model is no longer a good fit to the data.

The monitors discussed in Cowell et al. (1999) that we reproduce for the BN include the global monitor for overall model fit, the node monitor to check the probability distributions, and the parent-child monitor to assess the contribution of individual parent settings. Cowell et al. (1999) also used a batch monitor, essentially a chi-square test to detect significant differences between observed and expected counts of each variable. In the same spirit, we develop a situation monitor based on expected and observed counts, but elsewhere we focus on the Bayesian monitors.

The monitors discussed in this section review the well-established BN monitors. In the context of this thesis, the interpretation of the parent-child monitor for BNs is a novel contribution as it assesses when a BN is not an adequate model. The CHDS example illustrates this when the BN parent child monitor suggests that the data has additional context-specific conditional independence information. The implementation of these monitors in R is also a new contribution.

Global monitors The global monitor for BNs is defined as the logarithmic probability of the mth observation : $-\log p_m(y_m)$ after $m-1$ cases are processed.

Definition 59 *The overall **global monitor** for all M cases is:*

$$G_{BN} = -\log \prod_{m=1}^{M} p_m(y_m) = -\log \prod_{m=1}^{M} p_m(y_m|y_1,\ldots,y_{m-1}) \tag{5.5}$$

$$= -\log p(y_1 \ldots, y_m) = -\log p(\boldsymbol{y}). \tag{5.6}$$

Calculating the global monitor for two different systems provides an immediately interpretable comparison between models. These monitors have been shown to provide quick checks of BN structure against data. To illustrate, the log marginal likelihood, equivalent to the global monitor, for BN CHDS is $G_{\text{BN}} = -2495.01$. In Section 5.5, we will see how this compares to the global monitor of competing models.

X_M	$Z_{\text{marg node}}$	$Z_{\text{cond node}}$
X_{s}	1.708	0.1737
X_{e}	0.582	-1.560
X_{l}	2.953	2.454
X_{h}	0.340	-0.450

Table 5.1: Final BN node monitors for the CHDS example where $|Z| > 1.96$ suggests an ill fit.

Node monitors The node monitor assesses the adequacy of the marginal and conditional probability distributions for each node in the model.

Definition 60 *The **marginal node monitor** is given by*

$$N_{marg} = -\log p_m(x_k)$$

after $m-1$ cases are processed.

This is calculated by ignoring the other evidence in the mth case after X_i is observed. The unconditional node monitor checks the suitability of the probability distribution of the node.

The conditional node monitor uses probabilities that are conditioned on evidence in the mth case. To compute the conditional node monitor, all of the evidence in $\mathcal{E}$ is propagated except for $X_i = x_i$.

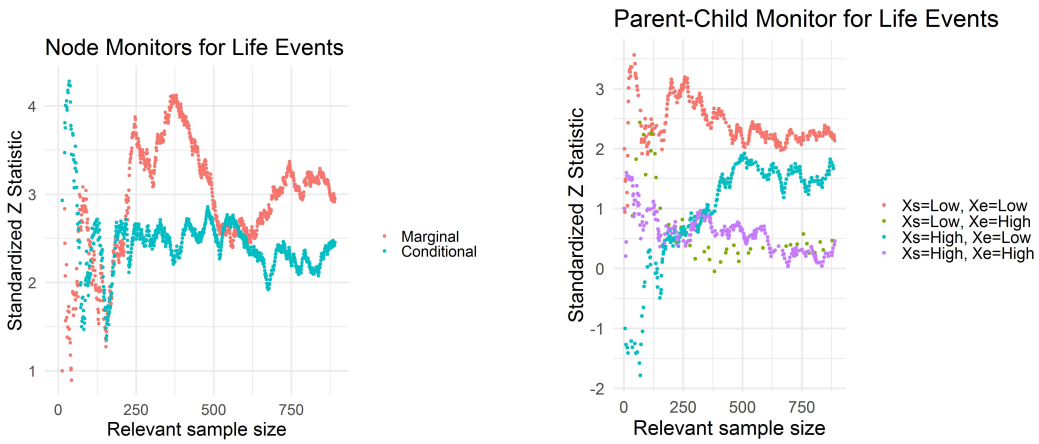
Definition 61 *The **conditional node monitor** can be represented as:*

$$N_{cond} = -\log p_m(x_i|\mathcal{E}_m \setminus X_i).$$

The conditional node monitor checks how well the model predicts each node given the other evidence in the observation. First, we specified the conditional probability tables, with θ_i after learning from the first $m-1$ cases. For the conditional node monitors, we propagated the evidence from the other variables omitting the node under consideration, and then queried the BN with the functions in the R package `gRain` (Højsgaard, 2012).

For instance, to compute the conditional node monitor for X_{h}, we propagated the evidence $\mathcal{E} = \{X_{\text{s}} = \text{High}, X_{\text{e}} = \text{High}, X_{\text{l}} = \text{High}\}$ and queried X_{h} according to the structure in Figure 5.4a. The node monitors are then standardised according to Equation 5.6.

Computing the final node monitors offers a quick check to see which node probability distributions might be incorrectly specified. The final node monitors for



(a) The marginal and conditional node monitor for X_{l}.

(b) CHDS BN: the parent-child monitor for positions all possible parent settings of X_{l}

Figure 5.4: Node monitors detect ill-fitting distribution for X_{l}

the CHDS BN are shown in Table 5.1. The marginal and conditional node monitors for X_{s}, X_{e}, and X_{h} are properly calibrated. However, we notice that the predictive probability distribution appears to be misspecified for X_{l}. The plot in Figure 5.4a confirms that both the marginal and conditional monitors indicate that we should not trust the modelling of X_{l}. There are context specific conditional probability distributions for X_{l} that should be adjusted.

As we will see in Section 5.2, the nodes of the BN are not exactly analogous to the positions of a CEG. Additional checks on the stages and the situations composing the stages will be required.

Parent-child monitors After identifying the problematic node, the parent-child monitor can be used to pinpoint the configurations of the parent values which might be associated with the misspecification.

Definition 62 *For any node X_i in a BN (noting that this is distinct from the situations and vertices v in a CEG), the* ***parent-child monitor*** *is defined as the predictive posterior of the mth observation with parents ρ after learning from the first $m-1$ cases with parents ρ:*

$$R = p_m(x_i | X_{pa(i)}^{m-1} = \rho).$$

Historically, the parent-child monitor has been used to confirm the effects of learning and the selected priors on the model (Cowell et al., 1999). The parent-child monitor can also be used to assess the appropriateness of different priors on individual nodes. We use it here to identify BNs that have context-specific probability

distributions that could be remedied by re-expressing the problem as a CEG. A good heuristic is that any predictive model with $|Z_m| > 1.96$ should be viewed with suspicion (Cowell et al., 2007). Following the example of Cowell et al. (2007), the parent-child monitor is computed without a formal hypothesis test here, as that would require the asymptotic theory. Instead, indicate we should be cautious, or even reject the model.

In Figure 5.4b, we check the parent-child monitor for X_l given all possible parent settings. This indicates that the household with $X_\text{s} = \text{Low}$ and $X_\text{e} = \text{Low}$ are a particularly poor fit to the data. Because the parent-child monitor assesses how sensitive a model is to particular setting, we use it here to indicate when a BN should be adapted to a CEG model.

This section has reviewed the existing prequential diagnostics for a BN. While these diagnostics are well established in the literature, they have been surprisingly little used in practice. However, coding these monitors in the R package `bnmonitoR` should elevate the profile of these diagnostics. The code for these diagnostics along with an example that uses a dataset on lung cancer used in Cowell et al. (1999). The BN diagnostics can be found at https://github.com/rachwhatsit/BNdiagnostics.

5.4 CEG Diagnostics

The monitors below explain what we might expect to see from the model in a predictive space. Prequential monitors can pinpoint where and how forecasts from candidate models deviate. The model fit might deviate because there can be two different data generating processes, and in this situation we might want to use the diagnostics to help explain why one model is a better fit than another. Additionally, data exchangeability might not hold, or the data might have some other built up dependence that the current structure does not capture. The diagnostics might reveal that the Dirichlet is an inappropriate choice for the model and other distributions might be needed. Certain copula families may be more appropriate for particular structures as demonstrated by Elfadaly and Garthwaite (2013); Zapata-Vázquez et al. (2014); Wilson et al. (2018); Elfadaly and Garthwaite (2017).

5.4.1 CEG Conjugate Dirichlet Analysis

Within a conjugate analysis, product Dirichlet-Multinomial distributions describe the posterior and more importantly the predictive distributions we use in our specific prequential analysis. Suppose we have either elicited or used model selection techniques to acquire the CEG, $\mathcal{C}$ with K stages denoted $u_1, \ldots, u_K$. Each stage u_i in $\mathcal{C}$ has floret parameters $\boldsymbol{\theta}_i$ for $i \in 1, \ldots, K$. Edges in a stage are $E(u_i) =$

$\{f_{i1}, \ldots, f_{iK_i}\}$ with labels $\theta_{ij} = \theta(f_{ij})$ for $j = 1, \ldots, K_i$ and $i = 1, \ldots, K$. Then suppose we observe a sample $\boldsymbol{Y} = \boldsymbol{y}$. From this we know in part how many observed counts arrive at each of the K stages. We denote the counts at each individual stage as $\boldsymbol{y} = (\boldsymbol{y}_0, \ldots, \boldsymbol{y}_i, \ldots \boldsymbol{y}_K)$ where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ij}, \ldots, y_{iK_i})$.

Assuming that the experiment was randomly sampled, then the floret parameter vector $\boldsymbol{\theta}_i$ has a Multinomial distribution Multi$(N_i, \boldsymbol{\theta}_i)$ where $N_i = \sum_{j=1}^{K_i} y_{ij}$ whose mass function we denote as $p_i(\boldsymbol{y}_i|\boldsymbol{\theta}_i)$. The separable form of the likelihood of the probability vectors for stages $u_1, \ldots, u_K$ is given by:

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{K} p_i(\boldsymbol{y}_i|\boldsymbol{\theta}_i) = \prod_{i=1}^{K} \prod_{j=1}^{K_i} \theta_{ij}^{y_{ij}}.$$

The Dirichlet prior distribution for each of the stages is denoted as $\boldsymbol{\alpha_i} = (\alpha_{i1}, \ldots, \alpha_{iK_i})$. Thus the prior is given by:

$$p(\boldsymbol{\theta}) = \prod_{i=1}^{K} \frac{\Gamma(\sum_{j=1}^{K_i} \alpha_{ij})}{\prod_{j=1}^{K_i} \Gamma(\alpha_{ij})} \prod_{j=1}^{K_i} \theta_{ij}^{\alpha_{ij}-1}.$$

Following the conjugate analysis in Collazo et al. (2018), under closed sampling we obtain the following form for the posterior distribution:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = \prod_{i=1}^{K} \frac{\Gamma(\sum_{j=1}^{K_i} \alpha_{ij})}{\prod_{j=1}^{K_i} \Gamma(\alpha_{ij})} \prod_{j=1}^{K_i} \theta_{ij}^{y_{ij}+\alpha_{ij}-1}$$

$$= \prod_{i=1}^{K} p(\boldsymbol{\theta}_i|\boldsymbol{y}_i) = \prod_{i=1}^{K} \frac{\Gamma(\sum_{j=1}^{K_i} \alpha_{ij+})}{\prod_{j=1}^{K_i} \Gamma(\alpha_{ij+})} \prod_{j=1}^{K_i} \theta_{ij}^{\alpha_{ij+}-1}.$$

where $\alpha_{i+} = \alpha_i + y_i$. Under closed sampling, then we can write the marginal likelihood in closed form:

$$p(\boldsymbol{y}) = \int_{\boldsymbol{\Theta}} \prod_{i=1}^{K} \frac{\Gamma(\sum_{j=1}^{K_i} \alpha_{ij})}{\prod_{j=1}^{K_i} \Gamma(\alpha_{ij})} \prod_{j=1}^{K_i} \theta_{ij}^{y_{ij}+\alpha_{ij}-1} d\boldsymbol{\theta}$$

$$= \prod_{i=1}^{K} \frac{\Gamma(\sum_{j=1}^{K_i} \alpha_{ij})}{\Gamma(\sum_{j=1}^{K_i} \alpha_{ij+})} \prod_{j=1}^{K_i} \frac{\Gamma(\alpha_{ij+})}{\Gamma(\alpha_{ij})}.$$

$$\log p(\boldsymbol{y}) = \sum_{i=1}^{K} \left[\log \Gamma(\bar{\alpha}_i) - \log \Gamma(\bar{\alpha}_{i+}) - \left(\sum_{j=1}^{K_i} \log \Gamma(\alpha_{ij}) - \log \Gamma(\alpha_{ij+}) \right) \right] \tag{5.7}$$

where $\bar{\alpha}_i = \sum_{j=1}^{K} \alpha_{ij}$ for all $i \in 1, \ldots, K$ and $\alpha_{i+} = \alpha_i + y_i$.

X_s	X_e	s_i	w_k	U_{AHC}	U_{AHC_1}	U_{AHC_2}	U_{AHC_3}	U_{AHC_4}
High	High	s_l^1	w_3					
High	Low	s_l^2	w_3					
Low	High	s_l^3	w_4					
Low	Low	s_l^4	w_5					

Table 5.2: Possible stagings for the cut X_1.

5.4.2 Global Monitor

As shown in Section 5.3.3, the global monitor is the probability of observing all of the evidence for a particular case m after processing $m-1$ cases, $P_m(\mathcal{E}_m)$. Evidence for the CEG is defined as the root-to-leaf path containing the observation.

Definition 63 *The overall **CEG global monitor** then is defined as the product of observing each of the m cases:*

$$G_{CEG} = -\log p(y_1, \ldots, y_m) = -\log p(\boldsymbol{y})$$

For a CEG, this is given by the marginal likelihood $p(\boldsymbol{y})$ shown in Equation 5.7. The global monitor offers an immediately interpretable comparison of candidate models. It also defines a way to directly compare a CEG equivalent to a BN with a CEG found using another method, as we see for the CHDS example in Section 5.5. After making changes to finer aspects of the structure, the global monitor may be computed to show improvements in the overall model.

5.4.3 Staging Monitors

Staging monitors are designed to identify problems with the staging of the colourings for a given cut as defined in Definition 22. For the comparison to the BN diagnostics, note that a cut in a stratified CEG is equivalent to a random variable in the BN. This does not have an analogy to the BN monitors because it is designed to detect discrepancies within the context-specific conditional independence relationships and ordinary BNs do not accommodate such structure. However, it can be used on a CEG representation equivalent to a BN to detect particular context-specific independences within this class. The relevant sample size here simply refers to the index in the dataset.

The set of situations $\{s_i \in V(\mathcal{T}) | s_i \in w_j \text{ for } w_j \in \mathbb{W}(\mathcal{C})\}$ associated with the positions of a cut are partitioned to compose the staging. For instance, in the CHDS example, the cut representing X_1 has four associated situations $\{s_l^1, s_l^2, s_l^3, s_l^4\}$ that correspond to the contexts X_s = High and X_e = High, X_s = High and X_e =

Low, $X_{\rm s}$ = Low and $X_{\rm e}$ = High, $X_{\rm s}$ = Low and $X_{\rm e}$ = Low, respectively. The stage structure shown in Figure 5.2 has the partition

$$U_{\rm AHC} = \{\{s_l^1, s_l^2, s_l^3\}, \{s_l^4\}\}$$

This staging may change for different points in the dataset. Consider the set of alternative stagings, denoted $U' \in \boldsymbol{U}$ to be the stagings that are one move on a Hasse diagram away from the given staging. For the example CEG, the alternative stagings we consider are shown in Table 5.2. The first alternative staging represents all situations merged together, and the last three indicate the situations with another stage removed from $\{s_l^1, s_l^2, s_l^3\}$. The staging may vary as forecasts flow from the data, so the stagings and alternative stagings are indexed as U_m and $\boldsymbol{U}_m$ respectively.

Definition 64 *For a given CEG with staging U_m and data observations $\{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_m\}$, the **CEG staging monitor** is the staging predictive distributions, $p(U_m|\boldsymbol{y}^{m-1})$.*

The form of the one step ahead predictive allowing for first-order Markov transitions between stages is given in Freeman and Smith (2011b). Because our primary aim is to see if the model staging is an appropriate fit given the data, we do not allow for transitions between stagings. However, for some applications, the computation of the one step ahead predictive could be adjusted to account for known transitions in the stage structure.

To assess the appropriateness of the staging to the data, we need the quantity:

$$p(U_m = U'|\boldsymbol{y}^{m-1}) \propto p(\boldsymbol{y}_{m-1}|U_{m-1} = U')p(U_{m-1} = U'|\boldsymbol{y}^{m-2})$$

$$= \frac{p(\boldsymbol{y}_{m-1}|U_{m-1} = U')p(U_{m-1} = U'|\boldsymbol{y}^{m-2})}{\sum_{U' \in \boldsymbol{U}} p(\boldsymbol{y}_{m-1}|U_{m-1} = U')p(U_{m-1} = U'|\boldsymbol{y}^{m-2})}.$$

As shown in Freeman and Smith (2011b), $P(U_{m-1} = U'|\boldsymbol{y}^{m-2})$ is available at time $t-1$ and

$$p(\boldsymbol{y}_{m-1}|U_{m-1} = U') = \int_{\boldsymbol{\Theta}_{m-1}} p(\boldsymbol{y}_{m-1}|\boldsymbol{\theta}_{m-1}, U_{m-1} = U')p(\boldsymbol{\theta}_{m-1}|U_{m-1} = U')d\boldsymbol{\theta}_{m-1}$$

$$= \prod_{i=1}^{K} \frac{\Gamma(\sum_{j=1}^{K_i} \alpha_{m-1}^{ij})}{\Gamma(\sum_{j=1}^{K_i} \alpha_{m-1}^{ij+})} \prod_{j=1}^{K_i} \frac{\Gamma(\alpha_{m-1}^{ij+})}{\Gamma(\alpha_{m-1}^{ij})}.$$

Here we have embedded the time index so that α_{m-1}^{ij} denotes α_{ij} at observation

$m-1$ and α_{m-1}^{ij+} denotes $\alpha_{ij}+y_{ij}$ at observation $m-1$. The staging monitor identifies places where the data is no longer a good fit for the existing stage structure.

The plots of the staging monitor depict $p(U_m = U'|\boldsymbol{y}^{m-1})$ for the assumed stage U and each alternative staging U' over the number of observations in the dataset. This allows us to see how the suitability of the model changes over time. If one of the alternative stages emerges as the highest probability forecast, then this indicates that the alternative staging in the model class should be used instead. If no clear staging emerges, this indicates that the appropriate staging may be outside the model class. This could indicate that the data-generating process draws from different stagings at different times. This might necessitate the use of different dependence structures. One example of such a dependence structure can be found in Wilson et al. (2018).

We will see how this enables us to differentiate between possible stagings in the CHDS data in Section 5.5.

5.4.4 Position Monitors

The position monitors, as the nodes of the CEG, rely on the message passing algorithm. Collazo and Smith (2015a) derived this algorithm, and the tree-based nature of the CEG makes it much faster than the BN propagation algorithm.

The node monitors for a BN detect discrepancies in the probability distribution specified for each node. For the CEG, we want to check the probability distributions specified for each position. Mirroring the BN methodology, we will compute a marginal and conditional probability.

To compute the marginal position monitor, N_{marg} for the mth observation in our dataset, we first compute the probability florets for each of the positions based on the previous $m-1$ observations in the dataset. Because positions only apply to data that matches the appropriate upstream pathways of w_i, the position monitors are only computed for those observations. Then, we compute the marginal probability of observing the mth observation for each of the values f_k emerging from the position w_i. We compute these by summing the probability of each of the root to sink paths that goes through the edge of interest f_k. The relevant sample size refers to the index in the subset of the data that arrives at each position.

Definition 65 *The **CEG marginal monitor** is given by*

$$N_{marg} = -\log p_m(\Lambda(\theta(w_i) = f_k))$$

The marginal monitor is then standardized against the actual observed value of w_i in the mth observation according to the Equation 5.2, 5.3, and 5.4.

The conditional node monitor computes the probability of observing evidence for the mth case after propagating evidence from the observations in the mth observation, excluding the outcome in the position of interest w_i. The conditional node monitor was designed for BNs to check the appropriateness of a distribution for a node conditional on the evidence for all the other nodes in the BN. As the CEG is automatically conditioning on all of the upstream variables, the conditional monitors for the positions of a CEG only provides information additional to the marginal node monitor for certain structures defined below. Like the marginal position monitor above, these are functions of observations within a given position of interest.

Whereas with the marginal monitor, we can compute the marginals from the probability florets directly, we need to use message passing to pass the evidence to update the probability florets for the conditional monitors. The propagation algorithm for the CEG is given in Thwaites et al. (2008) with additional details in Collazo et al. (2018). The propagation algorithm relies on evidence, which is the full root-to-sink path in the CEG. The evidence for the mth observation is some subset of the settings of random variables at the mth observation.

Evidence is propagated through a sub-graph of the CEG called the transporter. The transporter inherits the probabilities $\theta(w_i)$ for the set of positions and edges in the transporter. In the conditional monitor, we compute the p_m from the probabilities from the previous $m-1$ cases. The probabilities are back-propagated, i.e. summed at each position to compute the potential, $\phi(w_i)$ and then updated by dividing each $\theta(w_i)$ by $\phi(w_i)$. Thus, if the potential for w_i sums to one, then the updated probabilities are the same as the original.

Definition 66 *The **CEG conditional position monitor** is given by:*

$$N_{cond} = -\log p_m(\theta(w_i) = f_k | \mathcal{E}_m \setminus \theta(w_i)).$$

The conditional monitors are then standardized according to Equations 5.2, 5.3, and 5.4. For our examples in Section 5.5, $\phi(w_i) = 1$, so for our example, we need show only the marginal monitors.

The position monitors can be compared to a BN node monitor to confirm the suitability of the CEG structure. Within the CEG model class, it can detect discrepancies within the specified probability distribution. If the marginal position monitor indicates a poor fit, but the conditional position monitor indicates an appropriate fit, then we may continue cautiously using the selected model. However, if both the marginal and conditional position monitors indicate a poor fit, then we may want to consider alternative models. The monitors are designed to be used from the coarsest to finest, so we would only detect an issue with the position after

confirming that the staging is appropriate. Thus the position monitor detects issues that may be at or downstream of the position. Perhaps certain situations that are in the same stage should not be in the same position. The position monitor can also be used to detect when data has been generated from a model with additional positions or information available.

5.4.5 Situation Monitors

At the finest level, the CEG is composed of situations defined in Section 5.2. A stage u_i in a CEG is composed of situations $\{s_1, \ldots, s_k, \ldots, s_M\}$ that are by definition exchangeable. A situation monitor highlights situations when this exchangeability assumption might be violated.

The prequential methods check the validity of the forecasts. To check the forecasts from each of the stages in the structure, we need to compare the forecasts coming from each of the different situations. The stage order monitor imposes a new order to retain the prequential methodology. The leave one out stage monitor offers a quick check and additional aid to model transparency.

Leave one out stage monitor Using a method similar to the leave one out cross validation, we can examine the Bayes factor contribution from the stage u_i with a particular situation s'_k removed, denoted $f(\boldsymbol{y}'_{i,-k})$, and compare it to the Bayes factor contribution from the stage as a whole, $f(\boldsymbol{y}_i)$ as above. We expect that the stage with all contributing situations to be preferable to the one with the situation removed. Thus, this offers a quick check if any removing any situations leads to a higher Bayes factor score. We refer to this as the leave one out monitor, given by

$$Q(u_k, s_k) = \log f(\boldsymbol{y}_{i,-k'}) - \log f(\boldsymbol{y}_i)$$

where the contribution from the stage with situation s_k left out is

$$\log f(\boldsymbol{y}_{i,-k'}) = \log \Gamma(\bar{\alpha}_i) - \log \Gamma(\bar{\alpha}_{i+,-k'}) - \left(\sum_{j=1}^{K_i} \log \Gamma(\alpha_{ij}) - \log \Gamma(\alpha_{ij+,-k}) \right)$$

where $\alpha_{i+,-k'} = \alpha_i + y_{i,-k}$. A quick visual check can plot the actual observed proportions in each situation against the proportion we expect to see from the predictive posterior with data from the stage of interest missing. We examine the proportions of a particular level $l = l'$ for each of the stages. The stages associated with the variables that take extreme values are often of particular interest. For instance, for X_{h} in the CHDS data, we consider the proportion of households for which $X_{\mathrm{h}} = \text{Yes}$.

We could use this for more than two levels, but it would be more difficult to picture the discrepancy, and thus more difficult to display and interpret the output. Reducing the problem to a binary question allows us to leverage the properties of the Dirichlet distribution closure to marginalisation (Collazo et al., 2018). We can compute the conjugate posterior $\text{Beta}(\alpha', \beta') = (\alpha^+_{-k'}, \beta^+_{-k'})$ with the situation s_{-k} removed and take the expectation $\frac{n\alpha'}{(\alpha'+\beta')}$ where α' corresponds to the level of interest $l = l'$. We can compare this to the observed proportion of units where $y_i = l'$.

Situation order monitor To use a prequential check on the stage structure, we can impose an ordering on the relevant situations $I_{\tilde{M}} = \{s_1, \ldots, s_m, \ldots, s_M\}$. This ordering could correspond to some notion of severity of the situations. For instance, in the CHDS data, we might order the situations in cut X_1 according to increasing adversity $I(X_1) = \text{Low, Average, High}$. Imposing this ordering ensures that the corresponding residuals are independent.

For interpretability, reframe the data as Beta-Binomial distributed, where we are interested in the number of counts of the 'worst' level. The one step ahead predictives for each subsequent situation takes the counts of the data from the preceding situations as its parameters. Let $\alpha_{\prec m} = \alpha_i + \sum_{m=1}^{m-1} y_{im}$ and $\beta_{\prec m} = \beta_i + \sum_{m=1}^{m-1} y_{im}$ represent the count data from only the preceding situations. The surprise of observing the number of counts y_{ml} of the 'worst' level in the subsequent situation s_m is given by:

$$p(y_{m,l}) = \frac{\Gamma(\alpha_{\prec m} + \beta_{\prec m})}{\Gamma(\alpha_{\prec m})\Gamma(\beta_{\prec m})} \binom{y_m}{y_{ml}} \frac{\Gamma(\alpha_{\prec m} + y_{ml})\Gamma(\beta_{\prec m} + y_m - y_{ml})}{\Gamma(\alpha_{\prec m} + \beta_{\prec m} + y_{ml})}$$

Computing this quantity for each situation in turn allows us to determine when and if there is a certain point where the stage is a poor forecast for the subsequent data.

5.5 Examples

To investigate the diagnostic applications, we use two additional examples. The first example with the Christchurch Health and Development Study (CHDS) data is used to extend the BN diagnostics to the CEG. This enables us to see precisely where and how the CEG outperforms the BN, besides giving us a suite of diagnostics for general CEGs. The second example on radicalisation shows how the diagnostics scale.

5.5.1 CHDS

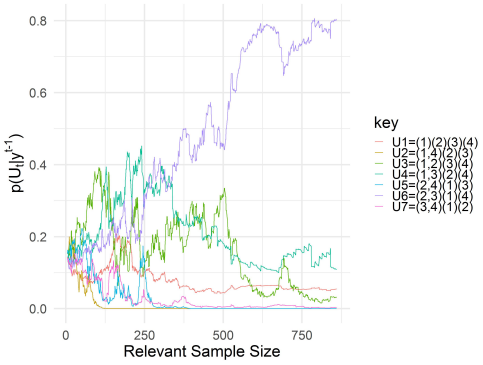
I have chosen this example because it has been subject to various different CEG models and so is already well studied, see Barclay and Nicholson (2015); Cowell and Smith (2014). The cohort study was conducted at the University of Otago, New Zealand (Fergusson et al., 1986). The sample used in this thesis used the categories derived from a latent class model (Barclay, 2014). Complete data was available for 890 households, and the analysis was performed on this sample. Description of the relevant variables are given below:

- X_{s}: Family social background, a categorical variable differentiating between high and low levels according to educational, socio-economic, ethnic measures, and information about the children's birth.
- X_{e}: Family economic status, a categorical variable distinguishing between high and low status with regard to standard of living.
- X_{l}: Family life events, a categorical variable signalising the existence of low (0 to 5 events), average (6 to 9 events) or high (10 or more events) number of stressful events faced by a family over the five years.
- X_{h}: Hospital admissions, a binary variable indicating whether or not a child in the household was hospitalised.

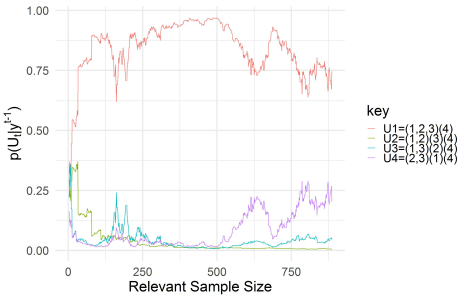
Other studies of the CHDS example have shown that the CEG give a much higher MAP score than the BN model. In this chapter, we focus on the diagnostics for stratified CEG models and show how the diagnostics can be used to explain why the fit of the CEG is better. More explicitly, our diagnostics can be used to show where predictions from the CEG model outperform those of the BN. To enable this comparison, we will compare two CEGs and the original BN. Figure 5.1 shows a $\mathrm{CEG}_{\mathrm{BN}}$ that encodes additional context-specific information from previous studies (Collazo et al., 2018).

The log marginal likelihood of this model is $Q(M_{\mathrm{CEG}_{\mathrm{BN}}}) = -2,495.01$. Under the relative standardization method, we obtain a Bayes Factor of 2,421,748. This is a tremendous improvement over the existing BN model already. With the assumed variable ordering $(X_{\mathrm{s}}, X_{\mathrm{e}}, X_{\mathrm{l}}, X_{\mathrm{h}})$, the AHC algorithm returns the structure $\mathrm{CEG}_{\mathrm{AHC}}$ in Figure 5.2. The marginal log likelihood for $\mathrm{CEG}_{\mathrm{AHC}}$ is -2478.49. This model is an even more sizeable improvement over the original BN with a Bayes Factor of 14,946,684. Comparing the two CEG models, the model generated by the AHC algorithm is six times as likely to have been data generating model, with a Bayes Factor of 6.172. This offers strong evidence that $\mathrm{CEG}_{\mathrm{AHC}}$ is a more suitable model

for the CHDS data than the equivalent BN representation in CEG_{BN}. We will nevertheless consider both as candidate models in order to demonstrate how our monitors identify the differences in the structure.



(a) CEG_{BN}: the staging monitor for variable X_{l}. This eventually recovers the optimal staging found in CEG_{BN}.



(b) CEG_{AHC}: the staging monitor for variable X_{l} depicting the selected staging as the most likely one.

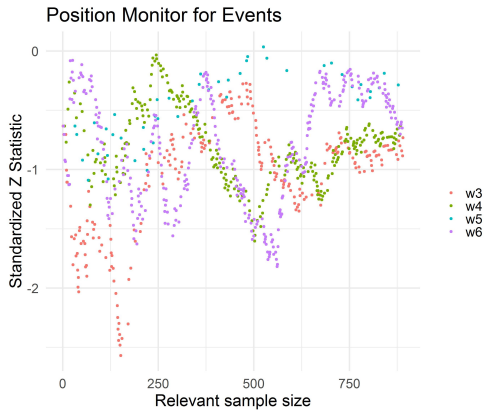
Figure 5.5: Staging monitors for two candidate CEG models. The staging with the highest probabilitiy indicates the best fit.

The staging monitor examines the possible partitions of the stages, called stagings at each cut in the tree. The staging monitor for CEG_{BN} is shown in Figure 5.5a. It confirms that $\{X_{\text{s}} = \text{High}\ X_{\text{e}} = \text{Low},\ X_{\text{s}} = \text{Low}\ X_{\text{e}} = \text{High}\}$, $\{X_{\text{s}} = \text{High}\ X_{\text{e}} = \text{High}\}$, $\{X_{\text{s}} = \text{Low}\ X_{\text{e}} = \text{Low}\}$ (denoted (1)(23)(4) emerges as the clear preference for the staging.

We see that the model struggles to distinguish between $\{X_{\text{s}} = \text{High}\ X_{\text{e}} = \text{High},\ X_{\text{s}} = \text{High}\ X_{\text{e}} = \text{Low}\}$, $\{X_{\text{s}} = \text{Low}\ X_{\text{e}} = \text{High}\}$, $\{X_{\text{s}} = \text{Low}\ X_{\text{e}} = \text{Low}\}$ (denoted (12)(3)(4)) and $\{X_{\text{s}} = \text{High}\ X_{\text{e}} = \text{High},\ X_{\text{s}} = \text{Low}\ X_{\text{e}} = \text{High}\}$, $\{X_{\text{s}} = \text{High}\ X_{\text{e}} = \text{High}\}$, $\{X_{\text{s}} = \text{Low}\ X_{\text{e}} = \text{Low}\}$ (denoted (13)(2)(4)) in the early observations. This suggests that an alternative model with a different stage structure might be more suitable for the data.

However, the monitor for CEG_{AHC} in Figure 5.5b, indicates a better fit to the data. The current staging is given by $\{X_{\text{s}} = \text{High}\ X_{\text{e}} = \text{Low},\ X_{\text{s}} = \text{Low}\ X_{\text{e}} = \text{High},\ X_{\text{s}} = \text{High}\ X_{\text{e}} = \text{High}\ \}$, $\{X_{\text{s}} = \text{Low}\ X_{\text{e}} = \text{Low}\}$, (denoted (123)(4) in Figure 5.5b). This remains the most likely staging throughout the data. The probability of the subsequent staging remaining the same based on the previous observations consistently stays around 0.75 as each subsequent observation in the dataset is realised.

To confirm the more accurate modelling of positions modelling context-specific probability distributions of X_l in the candidate CEGs, we can check the position monitors applied to CEG_{BN} and CEG_{AHC} in Figures 5.6a and 5.6b respectively. Both



(a) CEG_{BN}: the position monitor for positions w_3, w_4, w_5, and w_6 modelling X_{l}

(b) CEG_{AHC}: the position monitor for positions w_3, w_4, and w_5 modelling X_{l}

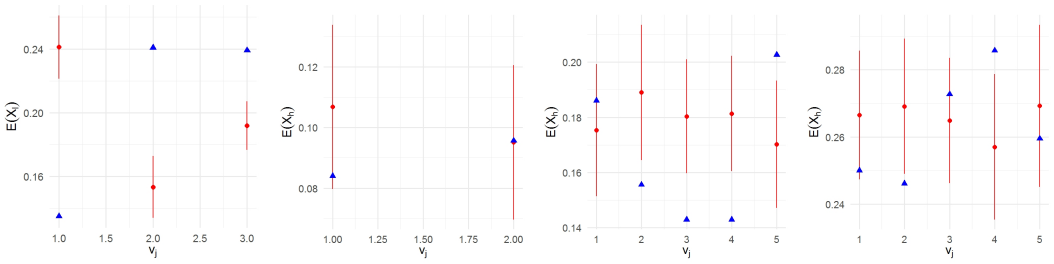
Figure 5.6: Position monitors for two candidate CEG models.

models are acceptable, and a substantive improvement over the position monitor of the original BN in Figure 5.4a.

After checking the staging, we turn our attention to the composition of the stages themselves. We consider the situations for the best-fitting CEG, CEG_{AHC}. In this CEG, stages u_0, u_1, u_2, and u_4 only have one contributing situation, so we examine u_3, u_5, u_6, and u_7. (Recall that stages are not labelled in Figure 5.2, but can be identified by assigning sequential labels to the unique colours.) The leave one out monitors return Bayes factor scores very close to zero, so we examine the plots of the expected and observed proportions of the levels of interest. We consider the proportion of $X_{\text{l}} = \text{High}$ for u_3 and $X_{\text{h}} = \text{Yes}$ for stages u_5, u_6, and u_7 in Figure 5.7.

While the staging and position monitors for u_3 and w_3 and w_4 respectively suggest that the probability distribution is a good fit for the data overall, the situation monitor in Figure 5.7a suggest that we should be cautious about the forecasts CEG_{AHC} for families experiencing a high level of adverse events. If we estimate the proportion of high adverse life events from households with either high social and low economic or low social and high economic capital, we will overestimate for households with high economic and social capital. Conversely, we underestimate the proportion of high adverse life events when we examine the leave one out proportions for s_2 and s_3.

Examining the prequential monitors here with the ordering of decreasing capital $I(X_{\text{l}}) = \{s_1, s_2, s_3\} = \{X_{\text{s}} = \text{High}\ X_{\text{e}} = \text{High},\ X_{\text{s}} = \text{High}\ X_{\text{e}} = \text{Low},\ X_{\text{s}} = \text{Low}\ X_{\text{e}} = \text{High}\ \}$ gives $p(y_{2,\text{High}}) = 0.028$ and $p(y_{2,\text{High}}) = 0.102$. This further confirms that situations s_1 and s_2 are not exchangeable. To adjust the model, we might consider the process by which families experience a number of life events. The



(a) u_3 LOO monitor (b) u_5 LOO monitor (c) u_6 LOO monitor (d) u_7 LOO monitor

Figure 5.7: The observed (blue triangles) and expected (red dots) proportions of households with high adverse life events (a) and children admitted to the hospital (b,c,d) with the respective situations in Table 5.3 on page 104 removed.

u_5				u_6				u_7			
s_i	X_s	X_e	X_l	s_i	X_s	X_e	X_l	s_i	X_s	X_e	X_l
s_1	High	High	Low	s_1	High	High	Average	s_1	High	High	High
s_2	High	Low	Low	s_2	High	Low	Average	s_2	High	Low	High
				s_3	Low	High	Average	s_3	Low	High	High
				s_4	Low	High	Low	s_4	Low	Low	Average
				s_5	Low	Low	Low	s_5	Low	Low	High

Table 5.3: Situations composing stages modelling X_h in stages u_5, u_6, and u_7

leave one out monitor for u_3 in particular suggests that something fundamentally different might be contributing to adverse life events for families with high social and high economic standing, one plausible explanation.

Stage u_5 is composed of the situations listed in Table 5.3. This is the moderately fortunate group. They are characterized by low life events and high social standing. The prequential monitor is given by: $p(y_{2,\mathrm{No}}) = 0.082$, again with no evidence of a structural issue.

Stage u_6 represents people who have access to either social or economic capital who experience an average number of life events, and families of individuals with low socio-economic standing who experience a low number of life events. This group has an average level of vulnerability. Examining the prequential stage monitors does not reveal any particular poor fits to the data: $p(y_{2,\mathrm{No}}) = 0.072$ $p(y_{3,\mathrm{No}}) = 0.272$ $p(y_{4,\mathrm{No}}) = 0.220$ $p(y_{5,\mathrm{No}}) = 0.075$.

Finally, u_7 represents the group with particularly unfortunate circumstances, regardless of their socio-economic stressors. All of the families of individuals reporting a high frequency of adverse life events contribute to this stage except for the group with no access to social or economic credit. Again, the prequential monitors do not indicate any situations of ill-fitting structure: $p(y_{2,\mathrm{No}}) = 0.065$ $p(y_{3,\mathrm{No}}) = 0.243$

$p(y_{4,\text{No}}) = 0.050$ $p(y_{5,\text{No}}) = 0.054$.

For stages u_5, u_6, and u_7, the leave one out monitors suggest that we should be cautious about forecasts for situations where the observed proportion of hospitalisations falls outside the bounds of our expected posterior. The monitors in Figures 5.7b, 5.7c, and 5.7d tell us which situations are over and underestimating hospitalisations, respectively.

5.5.2 Radicalisation Example

The radicalisation dataset examines the process by which individuals in a prison population are likely to be radicalised. Because of the sensitive nature of this domain, the data was constructed from a simulated model based on expert judgements which were then calibrated to publicly available statistics within the UK. Detailed information on the coding and simulation of the variables is available in Collazo and Smith (2015b). The dataset was previously used to describe the effect of non-local priors (Collazo and Smith, 2015a). The dataset has 85,000 simulated observations. The variables of interest are as follows:

- X_{g} Gender: Binary variable with values Male, and Female
- X_{r} Religion: Ternary variable with values Religious, Non-religious, and Non recorded
- X_{a} Age: Ternary variable with values Old, Medium, Young
- X_{o} Offence: Values include i) Violence against another person ii) RBT Robbery Burglary or Theft iii) Drug offence iv) Sexual offence v) other offence
- X_{n} Nationality: Binary variable indicating if an individual is a British citizen or a foreigner
- X_{w} Network: Indicates whether the individual has intense, frequent, or sporadic engagement with known members of target criminal organisation
- X_{e} Engagement: Binary variable that indicates whether or not the individual engages in radical activities.

In this second example, we illustrate how our diagnostics can be applied to a much larger study.

The model was built to better explain the pathways that lead to criminal engagement. So in this context, diagnostics are best used to examine how well the situations are predicting engagement in radical activities, X_{e}. Due to the complexity

stage	n
u_{33}	24
u_{34}	350
u_{35}	232
u_{36}	72
u_{37}	112
u_{38}	46
u_{39}	54
u_{null}	190

Table 5.4: Number of situations in each of the stages modelling X_{e}

and number of variables, the CEG model of the radicalisation data encodes a much richer space of causal hypotheses than the previous example. A Bayes factor model selection with the AHC algorithm using the ordering assumed in the dataset returns a CEG structure with a log marginal likelihood of -400007.3, which we use here as a baseline to determine better fitting adjustments to the structure.

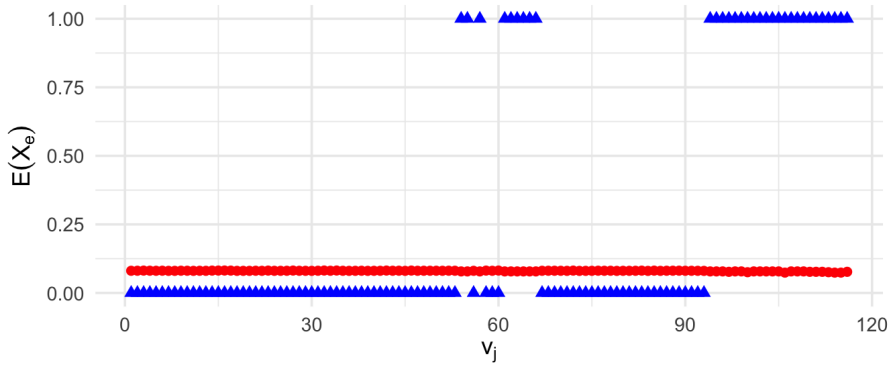
The stage partitioning for engagement X_{e} has six stages

$$U = \{u_{33}, u_{34}, u_{35}, u_{36}, u_{37}, u_{38}, u_{39}, u_{\text{null}}\}$$

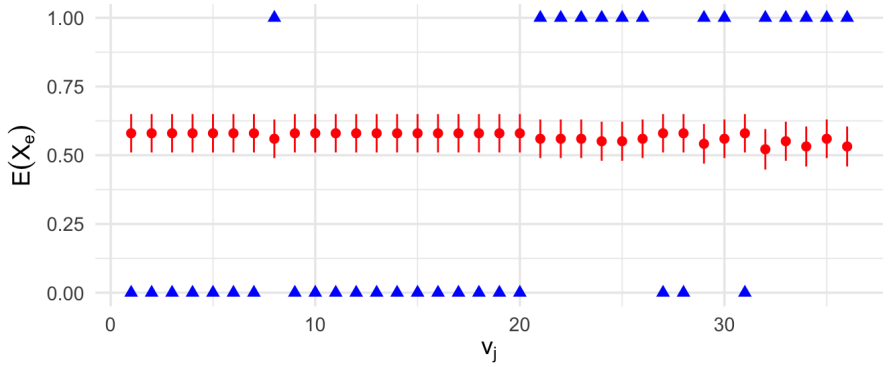
and 1080 unique positions, a much richer model. u_{null} represents the stage encompassing all situations that are unpopulated. This is a convenient and methodologically sound way of processing the empty stages. A large number of situations is difficult to inspect for cohesion, so our diagnostics are particularly important here. The size of each stage modelling engagement is shown in Table 5.4. Due to the high number of situations in each stage, situations will be indexed according to their particular stage. (That is, a situation s_1 in u_{31} is a distinct vertex from s_1 in u_{35}.)

One of the key questions concerning the radicalisation dataset is whether or not it is accurately captures radical engagement. The stages u_{33} and u_{38} contains sparse situations where all engage in radical activities. Stage u_{34} contains situations where no one engages in radical activities. Stage u_{35} contains several situations that do engage in radical activities alongside a large number of more sparsely populated situations that do not. Stages u_{36}, u_{37} and u_{39} reflect the same pattern. The plots of expected versus observed proportions when we leave a stage out are plotted in Figure 5.8. Plots are only shown for three stages, u_{35}, u_{36}, and u_{37}, as these are the stages that exhibit situations that exhibit both levels of X_{e}.

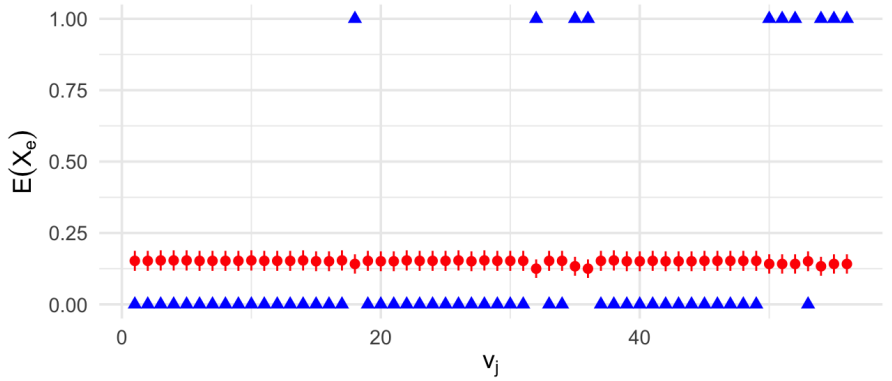
In stage u_{35} each of the situations is sparsely populated ($n < 15$ observations). Upon inspection of the dataset, the situations with all observed individuals engaging



(a) Leave one out monitor for u_{35}



(b) Leave one out monitor for u_{36}



(c) Leave one out monitor for u_{37}

Figure 5.8: Leave one out monitors for engagement, X_{e}. Large differences between the expected (red) and observed values (blue) indicate poor fit.

in radical activities are all religious, British males, traits not shared by the situations in which individuals do not engage. Because the counts of situations in stage u_{35} are quite sparse and radical activity is not abundant, it is difficult to tell if the situations are exchangeable. However, the common traits seem to suggest that it would be sensible to separate out the situations representing religious British males.

In stage u_{36} each of the situations is very sparsely populated ($n < 5$ observations). All of the situations that have no observations that engage in radical activities have $n = 1$. Thus, the expected posterior for the leave one out method has been heavily weighted by the situations containing individuals that do engage in radical activity. Again, sparsity obscures the model fit here, but inspection reveals that in u_{35}, the situations with observations that do engage in radical activity are all male.

This pattern holds in the last stage we consider, u_{37}. Again, the situations in the stage that do engage in radical activities are all male. This suggests that there is perhaps some additional information about the differences between male and female prisoners that is determining levels of radical engagement.

This second example shows that the diagnostics are particularly useful as our model accommodates larger data sets. The changes to the situations and staging structure can be adjusted and a new global monitor computed to show that the diagnostics suggest genuine model improvements.

Notice here that engagement with these diagnostics helps us discover currently articulated structure and help the domain experts to develop new hypotheses and models to check.

5.6 Discussion

Our extension of the prequential diagnostics from Bayesian Networks to the more general class of Chain Event Graphs has enables us to highlight places in which the selected structure is a poor fit to the given data. We have demonstrated how earlier analyses would have been much richer by employing these diagnostics and drawing out the reasons for certain variables failing or why one model is preferred to another. These monitors shown here are derived for stratified staged trees to build on the existing diagnostics available for a BN, but these methodologies also work for asymmetrical trees, a powerful example of CEG models.

These can also be applied to new classes of CEG like the dynamic CEG Barclay and Nicholson (2015), where the ordering is explicit and need not be assumed from the ordering of the data. We have only considered models from the AHC model selection algorithm here, although we can apply these diagnostics to additional advancements in model selection criteria. This work can also be extended to incorporate different

score functions.

The code for these CEG diagnostics as well as the subsequent two examples is available for download at https://github.com/rachwhatsit/cegmonitor. With the addition of the `stagedtrees` packages, we have a convenient implementation of the CEG software for practitioners.

Diagnostic monitors can be used to show how subsequent data performs when configured with the initial model. They may also be used to highlight an underlying dependence structure not captured by the existing CEG. As we have seen in the second example, the diagnostics pick out particular places for refinement as well as where situations in the model can be consolidated. The prequential diagnostics shown here are a critical performance check and a useful addition to the suite of CEG methods currently available.

Chapter 6

Customised Causal Inference

> "Focusing on what was done & what followed should clarify cause & effects. Alas..."
>
> *Railsea*, China Miéville

6.1 Background

As in Chapter 3, causal explanations of processes are often embedded in an expert's structural descriptions of a problem. These are often expressed in terms of an intervention on the system, either externally by nature or internally by people associated with the expert. There is now a wide literature about how such causal hypotheses relate to the structural hypotheses associated with the BN. However there is surprisingly little written about how these sorts of embeddings – critical if a the model is going to be used to guide policy – apply when the underlying structure is not a BN.

Statistical causation focuses largely on Bayesian Networks and frames causation as queries about independence among random variables. This rarely captures the sort of mechanisms domain experts speak about when they talk about causation (Cox and Wermuth, 2014). While the proliferation of structure discovery algorithms has enabled domain experts to apply new classes of models to a variety of problems, causal explanations of these models have lagged behind. Paying attention to the dynamics of the problem is particularly important to understand the causal mechanisms of a process.

There are several clear frameworks for expressing causal inference, including Granger causation from econometrics, the potential outcome framework of Rubin, and Pearl's interventional do calculus (Granger, 1969; Holland, 1986; **?**). Each of

these frameworks has a rich history of theory and applications. In particular, Pearl's BN approach is widely used. However, BNs represent a very small part of possible models available for causal analysis. Additionally, the de facto assumption that all modelling aspects are captured by a BN is rarely true. The faithfulness assumption and requirements for a BN to be causal are particularly stringent. Across different problem domains, certain features of a problem may render it intractable for a BN. These features limit the notion of cause that can be expressed by the BN.

Instead, I argue that causal analysis should begin by mathematically describing the processes described by domain experts, as illustrated in Chapter 3 For interventional causation, modelling the intervention mechanism should then be customised to the problem domain via this bespoke structural description (Aalen et al., 2016). It can then be expressed as the mapping between an idle and controlled system (Aoki, 1976; Materassi and Salapaka, 2016). The natural process as described in the idle system is not necessarily expressible as a faithful BN, although the BN can also be situated in this framework. This process allows controlled extensions to other model classes that are genuinely faithful to a given domain to be developed.

Expressing the wide variety of causal relationships observed in nature requires first respecting the underlying structure of the domain problem. The process of determining a bespoke structure enhances transparency with the domain experts and prompts more nuanced versions of causation in a system.

As stated in Chapter 3, these bespoke structures have now been developed for a number of different structures beyond the common BN approach. There are of course a large number of possible model classes, but for the purposes of this thesis it is sufficient to consider four alternative models. Thwaites and Smith (2010) developed a structure for tree-based Chain Event Graphs that can be used to describe an unfolding process. The Multi-regression Dynamic Model describes the semantics of a problem in which one time series affects another (Smith, 1993). The Flow Graph describes the movement of goods in a network where conservation of products destroys the dependence structure of a Bayesian Network (Figueroa and Smith, 2007). The elicitation of these models from domain experts has been described in Chapter 3.

The dynamics of problems suited to these graphical models are fundamentally different from the underlying BN Pearl's do calculus is based on. These systems give rise to new definitions of causation in a graphical model. To date, there is not a general catalogue of causal definitions in graphical models beyond that for BNs. This chapter aims to fill that gap by examining how standardized, general definitions of causation in a graphical model extend across new classes of graphs customised to problem semantics.

In this chapter after reviewing how the BN can be embedded in the more general context of bespoke causation, I review methods in which models discussed in the thesis have been similarly embedded. This discussion is restricted mainly to two ideas: intervention and genuine cause. Interventional causation is key in economic and epidemiological models and was developed independently by Pearl (2009) and Spirtes et al. (2000) to deduce causal hypotheses based only on observations of the idle systems. The procedure for interventional cause first determines which variables can be causes of others and the formulae for how we might hypothesise the strength of this effect were it to be true.

My novel contribution demonstrates how the concept of genuine cause identification and its measurement applies to the CEG and the MDM in Sections 6.4.2 and 6.5.2, respectively. From the CEG, the definition of genuine cause is more nuanced and flexible than that of the BN. Within the context of the MDM, the embedded dynamics enable us to match the concept of a genuine causal hypothesis to the initial elicited structure of the idle system.

The Section 6.2 describes general definitions of different levels of causation in bespoke systems. Definitions for temporal precedence, instrumental variables, and genuine cause are given in addition to customised interventions. The subsequent sections consider how each of these ideas might be customised to different models. Sections 6.3, 6.4, 6.5, and 6.6 explore BNs, CEGs, MDMs, and FGs respectively. Many of these definitions in alternative representations represent open questions and new areas of research, explored in the discussion in Section 6.7.

6.2 General Approaches to Custom Cause

Several sources have articulated different gradations of causation. Working solely in the context of BNs, Pearl identifies a three rung ladder of causation: association, intervention, and couterfactuals. Features are associated if they are related to some extent; intervention enables us to measure an effect if someone performs an action; and counterfactuals enable hypothetical interventions in additional contexts.

This corresponds to the levels of causation articulated in Cox and Wermuth (2014). In this framework, zero-level causation denotes association. First-level causation is again concerned with intervention, including case control studies. Second-level causation seeks to explain the dependencies observed in zero and first-level causation. Second-level causation explains mechanisms, and ties in with broader conceptions of causation from sociology and economics (Goldthorpe, 1998; Demiralp and Hoover, 2003).

The Bradford Hill criteria describes generic criteria for causation formulated

in an epidemiological setting (Hill, 1965). As many of the criteria are set in a clinical trial setting, it is likely that Hill was interested in interventional causation. This chapter focuses only on causal hypotheses that are motivated by the concept of intervention in the modelled system. While his criteria are not sufficient on their own to register a causal relationship, expressing them in statistical models offers a powerful way to incorporate domain expertise.

In the following sections, the proposed general definitions of cause will be defined for a general DAG and in subsequent sections applied to particular models, beginning with the BN.

6.2.1 Essential Graphs and Temporal Precedence

When speaking about causation, domain experts expect any potential causal candidates to occur before the effect is observed. Temporal precedence is perhaps the most basic of Hill's criteria. Rather than assuming temporal precedence from the directionality of a particular directed acyclic graph (DAG), $\mathcal{G}$, a causal candidate must be among the edges that are directionally invariant in a model. To articulate this in a graphical model, begin by examining the partial order implied by the essential graph of a DAG. The essential graph is a hybrid graph in which the only directed edges are common across all graphs in the equivalence class. The undirected edges change orientation within the equivalence class of DAGs.

Essential graphs, first named by Andersson et al. (1997), are chain graphs that characterize all graphs in the Markov equivalence class. Essential graphs are also sometimes known as completed patterns or complete pdags. Flesch and Lucas (2007) examined the meaning of the essential arcs in the essential graph.

To find the essential graph $\mathscr{E}(\mathcal{G})$, suppose the given a structure $\mathcal{G}$ has been found from a structural discovery algorithm or an expert elicitation. For general definitions require $\mathcal{G}$ to be a DAG. No causal directionality between X_i and X_j could be discriminated from a dataset of the process, however big if there are two equivalent Markov graphs with $X_i \prec X_j$ in one and $X_j \prec X_i$ in the other. Traversing the equivalence class of graphs and identifying the directionally invariant edges determines the edges that are causal candidates. The methods for finding the equivalence class are specific to the class of graphical models. For some classes of models, the equivalence class and subsequent essential graph remains an open question. Given the class of equivalent graphical models $\mathcal{G}$, a partial order I can be associated with the set of graphs.

Definition 67 *For $X_i, X_j \in \mathscr{E}(\mathcal{G})$, X_i **precedes** Y, written $X_i \prec X_j$, in a DAG $\mathcal{G}$ if there is a directed path from X_i to X_j in the essential graph $\mathscr{E}(\mathcal{G})$ where the*

undirected edges are equivalent to both $\preceq$ *and* $\succeq$.

The associated partial ordering can then be used to determine which nodes in the graph are causal candidates. This then motivates the naive causal definition

Definition 68 *With respect to the class of graphical models* $\mathcal{G}$ *with partial order* $I_{\mathcal{G}}$ *and essential graph* $\mathscr{E}(\mathcal{G})$, X_i ***naively causes*** X_j *if there is a directed path from* X_i *to* X_j *in the essential graph for* $X_i, X_j \in E(\mathcal{G})$.

Naive cause can be used to determine what the causal candidates are in a given structure $\mathcal{G}$. Methods of finding the equivalence class and corresponding essential graph for different graphs represents an open area of research. The link between temporal precedence and essential graphs, while implicit in their construction, has not been explicitly explored. Note that if X_i is not a naive cause of X_j it doesn't logically mean that X_i is not a cause of X_j. Rather, it means that there is no evidence for a causal relationship in the given graph (Wermuth, 2017).

6.2.2 Instrumental Variables

Identifying naive cause and genuinely embedded cause helps distinguish spurious associations from ones that could be causal from a relatively data rich environment. Another technique for determining spurious association from genuine cause is using instrumental variables.

Instrumental variables are a widely used technique in economics. Wright (1928) first proposed the concept of instrumental variables, and Wright (1921) showed that path analysis and instrumental variables were equivalent. An economics perspective on instrumental variables can be found in Goldberger (1972) and Morgan et al. (1990), while Bowden and Turkington (1990) offers a technical look at their development. Pearl (2009) reframed instrumental variables as genuine cause, differentiating it from spurious association. Instrumental variables can be thought of as approximations for randomised controlled trials. A valid instrumental variable is exogenous and relevant.

Definition 69 *For disjoint random variables* X_i, X_j, X_k, *and context* X_U, *say* X_k *is an* ***instrumental variable*** *if*

i) $X_i \not\perp\!\!\!\perp X_k$

ii) $X_U \perp\!\!\!\perp X_k$

iii) $X_j \perp\!\!\!\perp X_k | X_U, X_i$

Intuitively, we can think of an instrumental variable X_k as being independent of all variables with an influence on X_j that are not mediated by X_i where X_k is not independent of X_i.

Definition 70 *X_i is a genuine cause of X_j if there is a random variable X_k and context X_U in the model $\mathcal{G}$ that satisfies the definition of an instrumental variable.*

The definitions above are given in terms of random variables. Sections 6.4.2 and 6.5.2 will examine how graphical models whose vertices are not necessarily random variables can expand the definition of instrumental variables and genuine cause. For instance, random variables are implicit in the event tree, but examining the edge floret variables enables us to define genuine cause in the CEG.

6.2.3 Intervention

The previous sets of general definitions aim to distinguish between spurious association and associations that may be causal. The second level of causation requires certain sets of definitions to hold following an intervention in a system. Intervention also allows us to estimate effects in the controlled system. Total effects assess the impact of the intervention on the entire system. The average causal effect takes the difference between the idle and controlled systems. Expanding these definitions of intervention to models that are not BNs allows us to model interventions that are closer to the actual underlying mechanisms.

Implicitly, Pearl maintains that possible interventions occur on a factorizable joint distribution. For simplicity, this chapter restricts plausible interventions to ones that are enacted on factorizations of the system. This restriction can be expanded in future work to include other graphical forms.

Definition 71 *A joint probability distribution is **factorisable** when it can be written as a product of factors $f(\boldsymbol{x}) = f(x_1) \dots f(x_i) \dots f(x_n)$*

Intervening on the factorization can be denoted several different ways. Pearl uses do notation, $\text{do}(X_i = x_i)$, to describe when a random variable takes a particular setting (**?**). Alternatively, this is sometimes written as $X_i = \hat{x_i}$ or $X_i||x_i$ (Lauritzen, 1996) or P_{man} (Spirtes et al., 2000). The intervention notation in a BN represents an external setting of the random variables. As we will see in subsequent sections, sometimes domain experts want to model interventions that are not necessarily settings of random variables. Adapting these semantics to new models allows us to model interventions that are close to the mechanism described by domain experts.

Control theory offers a general way of framing interventional cause as a map between the idle and controlled systems. We define idle and controlled systems in

terms of the filtrations. This partial (or in other cases, total) order I can be used to construct a filtration, $\mathscr{F}_t$ which can be generally thought of as the observable information at time t.

Definition 72 *Graphical model G has an idle joint probability distribution represented by a filtered probability space $(\Omega, \mathcal{A}, \mathscr{F}_t, P)$.*

Definition 73 *The controlled system can be represented by a filtered probability space, $(\Omega, \mathcal{A}, \mathscr{F}_t, \hat{P})$ in which $\hat{P}$ represents the new controlled probability distribution.*

It is well documented if a BN is called causal, then after an intervention on a variable in the system, upstream variables are unaffected, downstream variables are affected as if the intervention had taken that value naturally, and variables not downstream are unaffected (Dawid, 2002; Dawid and Didelez, 2010; Eichler and Didelez, 2007). The control mechanism affects the downstream variables, that is variables that happen after the intervention at time t. We formalize this for general graphical models and their probability distributions.

Definition 74 *A model G with sample space Ω exhibits **idle determinism** if following intervention at time t_i, the following holds:*

i) All atoms in previous filtrations, $\{\mathscr{F}_{t<t_i}\}$, inherit the probability distribution from the idle distribution, P. That is, $\hat{P}(\omega_i) = P(\omega_i) \, \forall \omega_i \in \{\mathscr{F}_{t<t_i}\}$

ii) In $\mathscr{F}_{t_i}$, the probability of the atomic intervention $\hat{\omega}$ is one, and the probability of all other atoms is set to zero in the controlled distribution. That is, $\hat{P}(\hat{\omega}) = 1$ and $\hat{P}(\omega_i : \{\omega_i \in \mathscr{F}_{t_i}\} \setminus \hat{\omega}) = 0$.

iii) In subsequent filtrations, $\mathscr{F}_{t>t_i}$ containing the intervention $\hat{\omega}$, there is a map Γ between the idle and controlled distributions.

This property is not shared by all graphs in which we want to speak about causation. For instance, the controlled regulatory graphs do not exhibit this property (Liverani and Smith, 2015). However, defining idle determinism allows us to determine when the domain expert is speaking about types of causation that meet this property. There may be more general properties about intervention in classes of graphical models, but focusing on the above definitions demonstrates then nuances of causation in different model classes.

6.3 Bayesian Networks

BNs are a common modelling choice for causal inference. This section confirms that these general definitions developed in Section 6.2 also apply to the BN.

6.3.1 Naive cause

The essential graph of a BN is given by the skeleton of the BN with the only directed edges as the ones invariant across the equivalence class. This basic requirement is not trivial, even when interpreting a BN. In a BN, two graphs are equivalent if they have the same v-structures. The v-structure, also called a collider occurs when two unmarried parents share a common child. Changing the orientation of any arrow that is undirected in the essential graph must not result in any new v-structures.

Figure 3.2 shows the original graph $\mathcal{B}$ elicited from domain experts in Section 3.3.1. In this particular example, the essential graph $\mathscr{E}(\mathcal{B})$ is the same as $\mathcal{B}$ as reversing the arrows would result in additional v-structures. This graph gives us the partial order I:

$$X_{\mathrm{m}} \preceq X_{\mathrm{u}} \prec X_{\mathrm{b}} \prec X_{\mathrm{a}} \preceq X_{\mathrm{r}} \prec X_{\mathrm{s}}.$$

Then, for example, we can say that for any directed paths, breakfast model X_{m} naively causes breakfast participation rates X_{b}. Directed edges in the essential graph of the BN are naively causal.

6.3.2 Intervention in the BN

Interventions in the BN correspond to an external intervention that manipulates a given variable to a particular value. While this operation may only crudely capture the causal mechanism, it is a widely used approach to causal questions. For a causal BN in which x_j is set to $\hat{x}_j$, the post-intervention joint probability mass function of the remaining variables in the system is given by the **total effect** formula in Equation 6.1.

$$p(\boldsymbol{x}_{-j}||\hat{x}_j) = \begin{cases} \dfrac{p(x_1,\ldots,x_j,\ldots,x_n)}{p(x_j|\boldsymbol{x}_{Pa(j)})} & \text{if } x_j = \hat{x}_j \\ 0 & \text{otherwise} \end{cases} \tag{6.1}$$

Joint interventions in a BN may also be defined as an external manipulation of a set of variables $\boldsymbol{X}_J$ to a particular set of values $\hat{X}_J$. The compound post-intervention distribution is given by Equation 6.2.

$$p(\boldsymbol{x}_{-J}||\hat{x}_J) = \begin{cases} \dfrac{p(\boldsymbol{x})}{\prod_{j\in J} p(x_j|\boldsymbol{x}_{Pa(j)})} & \text{if } x_j = \boldsymbol{x}_J \\ 0 & \text{otherwise} \end{cases} \tag{6.2}$$

These post-intervention distributions represent interventions enacted on single and compound elements of the original factorisation, meeting the general requirement of considering interventions on factorisations.

To confirm that the BN admits idle determinism, define a filtration compatible with the partial ordering of a BN. A discrete BN has sample space

$$\Omega = \prod_{i=1}^{n} \mathbb{X}_i$$

where each of the atoms corresponds to a configuration of the values of the BN,

$$\omega_i = \{X_1 = x_1^i, \ldots, X_n = x_n^k\}.$$

The probability space of each of these atoms is inherited from the conditional probability tables. The sigma algebra is given by the power set of each of the possible values of each $\mathbb{X}_i$ where N_i denotes the number of values each ith variable may take:

$$\mathscr{F}_1 = \mathcal{P}(X = x_1^i, \boldsymbol{X}_{-\{1\}}), i \in \{1, N_1\}$$

$$\mathscr{F}_2 = \mathcal{P}(X = x_1^i, X_2 = x_2^j, \boldsymbol{X}_{-\{1,2\}}), i \in \{1, N_1\}, j \in \{1, N_2\}, \ldots,$$

$$\mathscr{F}_n = \mathcal{P}(X = x_1^i, X_2 = x_2^j, \ldots, X_n = x_n^k), i \in \{1, N_1\}, j \in \{1, N_2\}, k \in \{1, N_n\}.$$

The BN meets the requirements of idle determinism outlined above. That is, for an intervention at time $t_i \in \{1, n\}$,

i) $\forall \omega_i \in \{\mathscr{F}_{t<t_i}\}$, $p(\omega_i) = \hat{p}(\omega_i)$

ii) $\forall \omega_i \in \{\mathscr{F}_{t_i}\} : X_j = \hat{x}_j \in \omega_i$, $\hat{p}(\omega_i) = 1$ and
$\forall \omega_i \in \{\mathscr{F}_{t_i}\} : X_j \neq \hat{x}_j \in \omega_i$, $\hat{p}(\omega_i) = 0$

iii) $\forall \omega_i \in \{\mathscr{F}_{t>t_i}\} : X_j = \hat{x}_j \in \omega_i$, $\hat{p}(\omega_i) = \Gamma(p(\omega_i))$ and
$\forall \omega_i \in \{\mathscr{F}_{t>t_i}\} : X_j \neq \hat{x}_j \in \omega_i$, $\hat{p}(\omega_i) = 0$

where $\Gamma(p(w_i))$ is the map given by the total effect formula in Equation 6.1. Then the total effect of the manipulation alternatively $\text{do}(X_j = \hat{x}_j)$ is the marginal probability mass function of $Y(\boldsymbol{X}_{-j})$ using the probability mass function of $\boldsymbol{X}_j$ given by Equation 6.1.

6.4 Chain Event Graphs

CEGs encompass a broader class of models than the BN by admitting context-specific conditional independence and asymmetries. The vertices of the graphical structure represent positions, and the edges represent the occurrence of events.

Shafer (1996) posited that causality is more naturally expressed through trees than by measurements of random variables.

6.4.1 Naive Cause

The essential graph of the CEG does not have a convenient graphical representation. However, it does have a convenient polynomial representation (Görgen and Smith, 2018).

Definition 75 *In a chain event graph, $\mathcal{C}$, w_i **precedes** w_j, written $w_i \prec w_j$ if across all graphs in the equivalence class w_i is closer to the root w_0 than w_j.*

The essential graph for a CEG does not have a direct analogue. The results from Görgen and Smith (2018) characterized the statistical equivalence class. Transformations known as swaps and resizes allow us to algebraically traverse the equivalence class. Each CEG $\mathcal{C}$ has a total ordering, but true temporal precedence must be considered across the Markov equivalence class.

Definition 76 *For two positions w_i and w_j, w_i **naively causes** w_j if w_i precedes w_j in all compatible partial orderings of the given CEG.*

Future work could entail writing an algorithm that lists all of the CEGs in the same equivalence class. The number of swaps that correspond to equivalent interpolated polynomials can be determined by considering the number of levels of nested brackets in the interpolating polynomial defined in Görgen and Smith (2018). Within the partial ordering of a class of CEGs, positions that represent naively causal relationships are invariant to swaps. An algorithm can articulate all of the swaps from an interpolating polynomial of corresponding staged tree $\mathcal{T}$. This partial ordering, in turn, produces a list of naively causal edges.

6.4.2 Genuine Cause

Within the more generalized framework described above, I will next examine how we might determine evidence of a genuine cause when instead of the underlying structure is a CEG.

Definition 77 *For sets of positions $w_A, w_B, w_C, w_D \in \mathcal{C}$, and intrinsic event there is a **genuine cause** between sets of positions w_C and w_D if there exists a set of **instrumental positions** w_B and a context w_A such that:*

i) $\boldsymbol{X}_C \not\perp\!\!\!\perp \boldsymbol{X}_B$

ii) $\boldsymbol{X}_A \perp\!\!\!\perp \boldsymbol{X}_B$

iii) $\boldsymbol{X}_D \perp\!\!\!\perp \boldsymbol{X}_B | F(\{A \cup C\} | \mathbb{V}(\mathcal{C}))$

These dependence queries can be checked with the d-separation theorem from Chapter 4. Genuine cause in the CEG context highlights the correspondence to a RCT.

Example 78 *Example 79 in Chapter 3 on page 121 does not have any instrumental variables implicit in the graph. Let* $\boldsymbol{X}_{\mathrm{p}}$ *be a new variable that suggests whether or not clients are randomly sorted into groups that test a new application method. When we let the new application method denote the set of instrumental positions for the context in which we only focus on elderly immigrants, there is a genuine cause between the decision to apply and receiving EBT. Let* $\boldsymbol{X}_{\mathrm{A}} = X_{\mathrm{r}}$, $\boldsymbol{X}_{\mathrm{B}} = X_{\mathrm{p}}$, $\boldsymbol{X}_{\mathrm{C}} = X_{\mathrm{a}}$, *and* $\boldsymbol{X}_{\mathrm{D}} = X_{\mathrm{e}}$.

This example illustrates the use of instrumental positions in an asymmetric CEG. This cannot be represented by instrumental variables in a BN because of the structural zeroes in the conditional probability tables.

6.4.3 Intervention in the CEG

Intervention in the CEG corresponds to the occurrence of an event rather than the external setting of a variable as in the BN. The flexibility of the CEG to incorporate context-specific information permits interventions with new experimental designs or trials on a specific sub-population. Cowell and Smith (2014) argue that the process of causal discovery applied to a BN can be mirrored for the CEG.

As in Collazo et al. (2018), the intervention formulae for a staged tree are analogous to that of the BN. As in Section 2.4.1, let $(\mathcal{T}, \boldsymbol{\theta}_{\mathcal{T}})$ be a probability tree. Let $\mathcal{T} = (V, E)$ be the graph of that tree and let $\hat{e} = (s, s') \in E$ be an edge between situations of the graph. The effect of the tree-atomic manipulation of forcing any unit arriving at the situation s along edge $\hat{e}$ produces a new (degenerate) probability tree $(\mathcal{T}, \boldsymbol{\theta}_{\mathcal{T}})_{\hat{e}}$. This new tree has the same graph $\mathcal{T}_{\hat{e}} = \mathcal{T}$, but the edge probability of the enforced edge is set to one, $\theta(\hat{e}) = 1$, and the probabilities of all other edges $e' \in E(s) \setminus \{\hat{e}\}$ emanating from v are set to zero, $\theta(e') = 0$. Otherwise, the new tree inherits all edge probabilities from $(\mathcal{T}, \boldsymbol{\theta}_{\mathcal{T}})$.

Every probability tree is a graphical representation of atoms $\omega \in \Omega$, the set of all root-to-leaf paths. A tree-atomic manipulation changes the atomic probabilities $p_{\boldsymbol{\theta}}(\omega)$ of every atom $\omega \in \Omega$ in that space which are associated with a path containing the edge $\hat{e}$. The new probability of such an event is:

$$p_{\hat{e}}(\omega \parallel \hat{e}) = \frac{p_{\boldsymbol{\theta}(\omega)}}{\theta(\hat{e})} \tag{6.3}$$

and zero for all atoms ω associated with root-to-leaf paths passing through s but not through the edge $\hat{e}$. This formula given in Equation 6.3 for the probability mass function $p_{\hat{e}}$ can be expressed using the new map $\Gamma_{\hat{e}}$ which enacts the tree-atomic manipulation described above. This map meets the requirements for idle determinism in Definition 74.

The effects of this intervention can be seen in the example from Section 3.3.2.

Suppose we propose two different interventions at situation s_i and s_j forcing the units arriving at s_i and s_j at the head of edges $\hat{e}_i(s_i, s_i')$ and $\hat{e}_j(s_j, s_j')$ forcing the units along each edge. Then the effect of the joint intervention on the controlled mass function is the composition of the two effects:

$$p_{\hat{e}_i\hat{e}_j} = \Gamma_{\hat{e}_j}(p_{\hat{e}_i}) = \Gamma_{\hat{e}_j}(\Gamma_{\hat{e}_i}(p))$$

The total effect of a manipulation of an idle CEG as in Equation 6.3 is the probability mass function of the manipulated CEG. The new formula for the probability mass function $p_{\hat{e}}$ can be expressed using a map $\Gamma_{\hat{e}}$ which enacts the tree-atomic manipulation described in Equation 6.3. This map represents a new way of conceiving of cause in a CEG.

Example 79 *During a humanitarian crisis, emergency SNAP funds may be administered as expedited SNAP. Given the original tree in Figure 3.4, the emergency SNAP intervention corresponds to forcing units along the edges indicating expedited results:* $(s_1, s_5), (s_2, s_8)$, *and* (s_3, s_{11}). *The resultant subtree showing the possible paths that unfold from the expedited applications is shown in Figure 6.1.*

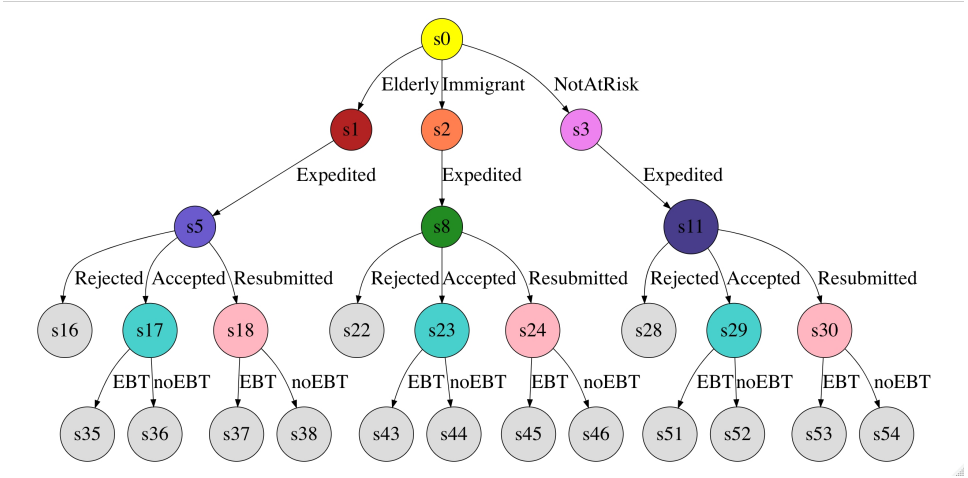


Figure 6.1: The staged tree shown for the emergency SNAP example.

The CEG is causal when the probabilities downstream remain the same after an intervention. In Example 79, the modeller can confirm with domain experts

that the three demographic groups considered still have different probabilities of having their applications rejected, accepted, or resubmitted. The CEG in Figure 6.1 shows that the probabilities of expedited accepted applicants successfully using Electronic Benefit Transfer (EBT) cards is the same regardless of demographic group. Applicants who must resubmit their application report a different probability of successful use, possibly due to inherited errors from the process upstream. If the domain experts did not confirm this stage structure, then the original CEG in Figure 3.4 could not be considered causal.

Idle determinism in the CEG Idle determinism holds for interventions in the CEG. The sample space of the CEG is all of the possible outcomes, each atom w_i is represented by a root-to-sink path $\lambda \in \Lambda(\mathcal{C})$. The sigma algebra of the CEG $\mathcal{C}$ is the power set of all of the root-to-sink paths, $\mathcal{P}(\Lambda(\mathcal{C}))$. As the filtration for the BN partitions the set adding a variable at each successive partition, the CEG adds an additional set of positions at a particular depth from the root node. The filtration of the CEG is given by the power set of the set of vertex centred events at a given depth $l = \{1, \ldots, N\}$ where N is the length of the longest path in $\Lambda(\mathcal{C})$. We will denote the set of positions W_l as the set of positions at depth l from the root.

$$\mathscr{F}_1 = \mathcal{P}(\Lambda(w_0)), \mathscr{F}_2 = \mathcal{P}(\{\Lambda(W_2)\}), \ldots, \mathscr{F}_N = \mathcal{P}(\{\Lambda(W_N)\}).$$

When an intervention occurs, say for some set of positions in $\hat{W} \subseteq W_{\hat{l}}$ at depth $\hat{l}$, then idle determinism is satisfied:

i) $\omega_i \in \{\mathscr{F}_{l<\hat{l}}\}$, $p(\omega_i) = \hat{p}(\omega_i)$

ii) $\forall \omega_i \in \mathscr{F}_{\hat{l}}$, if $\omega_i \in \Lambda(W_{\hat{l}}), \hat{p}(\omega_i) = 1$
$\forall \omega_i \in \mathscr{F}_{\hat{l}}$, if $\omega_i \in (\Lambda(W_{\hat{l}}))^c, \hat{p}(\omega_i) = 0$

iii) $\forall \omega_i \in \{\mathscr{F}_{l>\hat{l}}\}$: for some $w \in \hat{W}_{\hat{l}} \in \omega_i, \hat{p}(\omega_i) = \Gamma(\omega_i)$
$\forall \omega_i \in \{\mathscr{F}_{l>\hat{l}}\}$: for some $w \in \hat{W}_{\hat{l}} \notin \omega_i, \hat{p}(\omega_i) = 0$

For the subsequent filtrations, $\mathscr{F}_{l>\hat{l}}$ the downstream events are also zero. Upstream filtrations $\mathscr{F}_{l<\hat{l}}$ are unaffected. $\Gamma(\hat{\omega}_i)$ is again given by the total effect formula in Equation 6.3.

Alternative interventions in the CEG Returning to the example of the SNAP applications from Chapter 3 illustrates the flexibility of this new model class. The standard CEG distribution described above corresponds to forcing all of the units arriving at a vertex to travel down a particular edge. Possible interventions must

be interpreted within the context of the graph. For instance, suppose we want to compute the total effect of forcing all applications to be expedited following a natural disaster. This could not be modelled with the BN due to the structural zeroes from the population that decides not to apply.

Other interventional dynamics arise from either adding a refinement to or consolidating the sequence of events. For instance, suppose after speaking with domain experts, lack of sufficient documentation is voiced as a serious barrier to the application process. Then, we could add an additional position to the graph that demonstrates whether or not an applicant has sufficient documentation. This could not be modelled by the BN because of the asymmetries. Furthermore, it is easier to add a position to the CEG than to add a node to the BN. This is because updating the conditional probability tables of a BN requires editing not only the new node, but all downstream nodes. Confirming with domain experts that this is the case indicates whether a CEG is causal or not. If the CEG is not causal, then when used for policy interventions it will often mislead. The CEG only requires adding the probability on the new edges added to the graph. This allows for quick inference that can be adjusted quickly with a group of experts.

6.5 Multi-regression Dynamic Models

6.5.1 Naive Cause

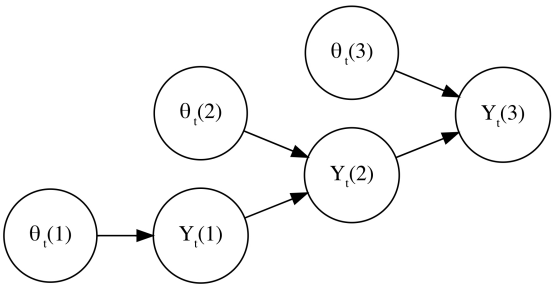


Figure 6.2: The full essential graph of the MDM shown for a single time slice at t.

In a MDM, there is an observational partial order and a full partial order. The essential graph for the sample MDM shown in Figure 6.2 is equivalent to the existing summary graph. The summary graph refers to the graph drawn for a single time slice in which the nodes are instantiations of only the series vectors.

The full essential graph includes arrows from the core state values $\theta_t(m)$ to the observation $Y_t(m)$ as shown in Figure 6.2. The full essential graph introduces a

finer partition on the ordering of the series observational essential graph. While the core state values precede the instantiation of the series, they are contemporaneous with the series observation. Consequently, it makes sense to only consider the core state values as causal candidates for the downstream observation series and not the reverse. The core state values play a key role in developing a dynamic notion of genuine cause as seen in Section 6.5.2. Each series observation depends on a set of parents chosen from the preceding set of series variables, and this strict ordering means that a unique MDM is its own essential graph as shown in Lemma 82.

Definition 80 *In the essential graph drawn between the series,* $Y_t(i) \prec Y_t(j)$ *there is an observational partial order in a MDM if there is a directed path from* $Y_t(i)$ *to* $Y_t(j)$ *in the series observational essential graph.*

This partial order allows us to have contemporaneous naive cause, suitable as the core state and series observation occur simultaneously. This partial ordering on the series can again be used to define a naive cause.

Definition 81 $Y_t(i)$ *naively causes* $Y_t(j)$ *if there is a directed path in the essential graph of the MDM.*

Lemma 82 *The equivalence class of a MDM is a singleton.*

Proof. By construction, as each series $Y_t(i)$ has its own corresponding core state $\boldsymbol{\theta}_t(i)$, this creates a v-structure between this edge and $\text{Pa}(Y_t(i))$. Because the MDM is a valid BN, the set of v-structures determines the equivalence class. ■

Thus, each MDM is also its own essential graph. The strict ordering of the MDM is more restrictive than the BN, but these stricter assumptions create more powerful causal relationships.

In the Summer Meals Program example elicited in Section 3.3.3, the MDM elicited from experts shown in Figure 6.2 is also the essential graph. The strict temporal precedence agrees with the description given by domain experts.

6.5.2 Genuine Cause

The model assumptions of the MDM engender new notions of genuine cause in a dynamic setting in addition to the naive cause. In the MDM setting, the core state values of the parents of a series can be thought of as instrumental variables. The core state vector of the parent series acts as a randomizing agent in the hypothetical RCT. For the Summer Meals example in Section 3.3.3, the number of children transported to the meal site directly affects the number of meals eaten at the site. Suppose the modeller set two versions of the MDM with different core state values for

transportation to the meal site, as availability of buses changes depending on school district. Then, observing proportionate changes in the number of children eating at a meal site in both models would indicate that the model accurately captured the causal mechanism between transporting children to the meal site and the number of meals eaten.

Formally, Definition 83 establishes the core state of a parent series as an instrumental variable. For the MDM, the associated context of the instrumental variable can be thought of as the preceding core state variables. Setting the core state value of the instrumental variable establishes the strength of the causal mechanism between the two series in the MDM.

Definition 83 *For two series vectors $Y_t(i)$ and $Y_t(j)$ where $Y_t(i) \in \text{pa}(Y_t(j))$, $\boldsymbol{\theta}_t(i)$ is an instrumental variable of the genuine cause $Y_t(i)$ on $Y_t(j)$ for a set context U if:*

i) $Y_t(i) \not\perp\!\!\!\perp \boldsymbol{\theta}_t(i)$

ii) $U \perp\!\!\!\perp \boldsymbol{\theta}_t(i)$

iii) $Y_t(j) \perp\!\!\!\perp \boldsymbol{\theta}_t(i) | Y_t(i), U$

Lemma 84 *Given an MDM, a cause between a series $Y_t(i)$ and its parents $\text{pa}(Y_t(i))$ is genuinely causal.*

Proof. By the model specifications in Definition 19, each series $Y_t(i)$ that is not a root node has a set of parents $\text{pa}(Y_t(i))$. The context can be defined as the preceding observations of the core state values:

$$U = \{\boldsymbol{\theta}^{t-1}(1), \ldots, \boldsymbol{\theta}^{t-1}(i-1)\} \setminus \boldsymbol{\theta}_t(\text{pa}(Y_t(i)).$$

Then, $\boldsymbol{\theta}_t(\text{pa}(Y_t(i))$ meets the requirements of an instrumental variable as

i) $\text{pa}(Y_t(i)) \not\perp\!\!\!\perp \boldsymbol{\theta}_t(\text{pa}(Y_t(i))$ by Equation 2.5,

ii) $U \perp\!\!\!\perp \boldsymbol{\theta}_t(\text{pa}(Y_t(i))$ by Result 1 of Smith (1993), and

iii) $Y_t(i) \perp\!\!\!\perp \boldsymbol{\theta}_t(\text{pa}(Y_t(i)) | \text{pa}(Y_t(i)), U$.

Criteria iii) is true because a single time slice of the full MDM is a valid BN, and $\text{pa}(Y_t(i))$ d-separates $Y_t(i)$ from $\boldsymbol{\theta}_t(\text{pa}(Y_t(i))$. Thus, an instrumental variable $\boldsymbol{\theta}_t(\text{pa}(Y_t(i))$ can be constructed for every series with parents in the MDM. ■

6.5.3 Intervention

In dynamic systems there is a vast array of ways in which we can intervene and each of these can have a different consequence. Here for the MDM–because it can be unfolded as a DBN with latent states, we can use standard BN intervention calculus to read new causal algebras for the MDM corresponding to different interventions. Intervention in the MDM can be customised to the sort of manipulation each time series is subjected to. Interventions on the series $Y_t(i)$ and the underlying state vectors $\boldsymbol{\theta}_t(i)$ have been defined in Queen and Albers (2009).

The idle probability distribution for the series is given by

$$(Y_t(i) \,|\, \boldsymbol{\theta}_t(i)) \sim (\boldsymbol{F}_t(i)'\boldsymbol{\theta}_t(i), V_t(i));$$

Definition 85 *A **series intervention** on $Y_t(i)$ occurs in an atomic manipulation on the series. It can occur on any number of and combination of the items in the series $Y_t(1), \ldots, Y_t(i), \ldots, Y_t(i)$. Given intervention C, the post-intervention distribution is given by*

$$(Y_t(i) \,|\, \boldsymbol{\theta}_t(i), C) \sim (\boldsymbol{F}_t(i)'\boldsymbol{\theta}_t(i) + h_t(i), V_t(i) + H_t(i)).$$

The ability to intervene on different combinations of the series observations enables us to customise the interventions to the dynamics described by domain experts. The Summer Meals Program offers an example of a series intervention. Suppose that public transportation is cancelled at time $\hat{t}$, then $T_{\hat{t}}||t_{\hat{t}}$. Generally, $Y_{\hat{t}}(j)||\hat{Y}_{\hat{t}}(j)$. This could occur for a one off change. In our example, this could be a cancelling of bus services for a public holiday.

Intervention for an ongoing series of observations represents another type of intervention. For this after some time of intervention $\hat{t}$, we set a value of a series for ongoing $t \geq \hat{t}$. This could occur when there is a time point where there is an ongoing disruption to the service. For instance, the bus services could be cancelled for school children following the end of summer school.

At some time of intervention $\hat{t}$, we can propose an intervention to $Y_{\hat{t}}(j)$ that lasts for a period of time $\hat{t}_1 < t < \hat{t}_2$. For instance, a mid-summer awareness campaign would increase the number of radio ads and text messages for the first week after summer school, but then the awareness would resume to normal levels.

Another intervention corresponds to altering the variance of the observed series. For instance, after summer school changes, the variability in the number of children using public transportation and eating meals drastically increases.

Interventions may also occur on the core state values. The idle probability distribution for the state vectors is given by:

$$(\boldsymbol{\theta}_t \,|\, \boldsymbol{\theta}_{t-1}) \sim (\boldsymbol{G}_t\boldsymbol{\theta}_{t-1}, \boldsymbol{W}_t);$$

Definition 86 *A **state vector intervention** externally manipulates $\boldsymbol{\theta}_t(i)$. As with compound interventions in the BN, any number or combination of interventions on $\boldsymbol{\theta}_t(1), \ldots, \boldsymbol{\theta}_t(n)$ is possible. The post-state vector intervention distribution is given by*

$$(\boldsymbol{\theta}_t \,|\, \boldsymbol{\theta}_{t-1}, C) \sim (\boldsymbol{G}_t^*\boldsymbol{\theta}_{t-1}, \boldsymbol{W}_t^*).$$

Here, $h_t(i)$ and $\boldsymbol{G}_t^*(i)$ represent the change in $Y_t(i)$ and $H_t(i)$ and $\boldsymbol{W}_t^*(i)$ represent the change in uncertainty (Queen and Albers, 2009).

The MDM can be customised to the types of intervention in dynamic linear models. Queen and Albers (2009) explored both series and core state interventions, but only for interventions on a single observation. Dynamic linear models accommodate periodic shifts. This can be used to show the effect of serving weekend meals, or the effect of holidays on children eating summer meals.

6.6 Flow Graph

Where the MDM offers a set of stricter model assumptions, the Flow Graph relaxes those assumptions. The Flow Graph is not compatible with a Bayesian Network as the additional mass conservation constraint induces severe dependencies. However, relaxing this assumption enables us to model flows of goods through a network.

The details of the Flow Graph construction are given in Chapter 2. The flow is first described as an Hierarchical Flow Network (HFN). This is then transformed to a two time slice dynamic Bayesian Network (2TS-DBN) representation that expresses the flows in terms of measurable random variables. As the 2TS-DBN representation of the HFN is a valid BN, we can define an essential graph and naive cause.

6.6.1 Naive Cause

The additional constraints of the decomposed HFN to the 2TS-DBN has meaningful ramifications for the essential graph and the extension of naive cause.

Lemma 87 *The essential graph of the 2TS-DBN of the HFN is the undirected skeleton of the 2TS-DBN.*

Proof. The construction of the 2TS-DBN requires linked chains. There are no v-structures. Thus, the essential graph is entirely undirected. ■

This undirected essential graph confirms that modelling intervention in a flow is entirely reversible. Having an undirected essential graph is by design, as it

allows for reversible flow depending on what actors in the network are subject to intervention. This means that there are no naively causal candidates in the Flow Graph.

6.6.2 Genuine Cause

The Flow Graph represents several open questions, one of which is what genuine cause means in the context of the Flow Graph. Similarly to the MDM, I posit that the structure of the Flow Graph is a valid representation of the conditional independence relationships specified in Figueroa and Smith (2007) when the one-step ahead forecasts hold. Instrumental variables in the Flow Graph are an open question.

The type of intervention in the FG alters the dependence structure and specifies a directionality. Thus, different interventions have different genuine causes. Interventions in dynamic systems take many forms and each of these types of intervention may result in a unique genuine cause.

6.6.3 Intervention

The intervention breaks the Pearlean definition by manipulating the error variance. The calculus of intervention for direct manipulation of the state random vector nodes of the true process has been found in Figueroa and Smith (2007). The Flow Graph admits a factorization of the path flows. This calculus works for changing path flows as well as for interventions that remove nodes from the system.

Alternative intervention in the Flow Graph The unconventional dynamics of the Flow Graph admit customised interventions. The dynamics of the example of the transfer of meals from vendors to sponsors to sites from Chapter 3 prompt different types of intervention. For instance, suppose two meal sites (perhaps a local school and a nearby community centre) wanted to merge sites. This intervention could not be modelled in the standard BN frame, but in the flow graph framework, the post-interventional distribution could be computed after merging two actors on the same level.

A second alternative intervention consists of removing or adding mass at a particular level in the system. In the example from the Summer Meals Program, this might correspond to a policy change to only reimburse meals for children aged 0-12 rather than the existing restrictions to supply meals for ages 0-18. Adding mass to the system could correspond to a vendor receiving a donation of meals at a particular time point. This intervention on the path flows would again lead to a different post-interventional distribution.

Directly intervening on the structure of the path flows would again yield a custom intervention. If a sponsor changed vendors, this would directly alter the structure of the Hierarchical Flow Network and the consequent path flows. The richness of the Flow Graph admits interventions that align with the interventions expressed by domain experts.

6.7 Discussion

This chapter broadens the understanding of causation in probabilistic graphical models with respect to temporal precedence, instrumental variables, and interventional cause.

In causal inference with the BN, causal relationships are framed around a known set of background variables. Particular settings of the random variables correspond to different contexts, effects, and causes. While this has proven useful for causal inference, it is very restrictive with respect to the sort of dynamics it can describe. This chapter demonstrates that describing causal relationships within filtrations and defining a mapping between idle and controlled systems is a more flexible way of describing this. Describing intervention in the BN and the CEG demonstrates the suitability of filtrations to a discrete, tree-based structure. Generalizations with the MDM and the Flow Graph show two examples of an alternative setting for causation that hold for continuous domains. This chapter describes the flexibility of interventions in each different class of model.

Chapter 7

Discussion

> The universe works on a math
> equation that never even ever really
> ends in the end
>
> "Never Ending Math Equation,"
> Modest Mouse

The new applications, methods, and theory addressed in this thesis opens additional questions about causation and structure in customised graphical models. Section 7.1 outlines the main contributions of this thesis. Section 7.2 explores a particular area of inquiry related to CEG model selection. Section 7.3 describes several areas of further inquiry that build from the work outlined in this thesis.

7.1 Summary

This thesis has sought to demonstrate the importance of selecting an appropriate structure to probabilistic graphical modelling.

Chapter 3 provides a framework for conducting a qualitative structural elicitation that accurately represents experts' natural language description of a problem. My main contribution in this chapter is merging domain expertise of issues in the realm of food insecurity with appropriate model classes. Exploring the different model classes within the domain of food insecurity exemplifies this process. Checking the structure of the MDM with the one step ahead forecast represents a novel contribution. Checking the CEG structure with the preliminary separation theorems motivates the work of the full d-separation theorem in Chapter 4.

The full d-separation theorem for the CEG represents a substantial contribution to CEG theory and methodology. The construction of the ancestral graph provided relies on a new understanding of ancestrality in systems with context-specific

independence. The full d-separation theorem allows for a much more flexible class of models that admit asymmetries and context-specificity. D-separation for BNs only holds for faithful BNs, making the CEG model class much less restrictive.

Chapter 5 offers a practical advancement to CEG methodology. Prequential diagnostics check the consistency of forecasts that flow from the model with structural elements of the CEG. The contribution of two software packages `bnmonitoR` and `cegmonitoR` form a useful addition to the toolkit available for modellers. The CEG diagnostics can be used to check consistency between different cohorts, tying into some of the wider ideas about causation.

Finally, Chapter 6 explores elements of causation that are necessarily widened by model classes that are different from that of the BN. The concepts of essential graphs, instrumental variables, and intervention offer different nuanced definitions of causation across different model classes. In particular, I demonstrate that the equivalence class of the MDM is a singleton, rendering all edges in the MDM as instrumental variables. The full d-separation theorem for CEGs also allows us to define instrumental variables for the CEG.

This chapter reviews one particular area of work underway on Beta Divergence in Section 7.2, and then concludes with a look at areas of future work in Section 7.3.

7.2 Beta divergence

The traditional Bayes factor search for the CEGs uses the logarithmic score. The logarithmic score is very sensitive to outliers.

The logarithmic score is associated with the Kullback-Leibler divergence. Model search for the CEG can be conducted with beta-divergence instead of the Kullback-Leibler divergence. The Agglomerative Hierarchical Search algorithm uses the logarithmic score to determine the stage structure of the CEG. By varying the parameter beta, we can determine how sensitive to outliers the model should be. Preliminary work confirms that as beta goes to zero, we recover the same stage structure as given by the logarithmic score. As beta increases, the AHC algorithm returns a stage structure with an increasingly coarse staging. The Bayes factor search tends to lump sparsely populated situations in the biggest stage to cut losses, an issue that this remedies. In an online learning context, this means that the stage structure is more resilient to outliers.

Simulations shows that increasing beta results in increasingly coarse CEG stagings. Beta can be used to make a staging more robust to sparsely populated situations.

7.3 Future Work

The custom classes of models explored in Chapter 3 could each be developed into their own elicitation protocols. Additional work could be done to translate natural language into customised graphical models.

The CEG diagnostics form a practical addition to the CEG methodology. These diangostics have already been extended to the DCEG and the RDCEG (Shenvi and Smith, 2018). Beyond applications to the CEG, the diagnostics may be used to address open questions around Bayesian model criticism for causal inference. Prequential diagnostics have recently resurfaced in the literature as a way to assess causal discrepancies in an online learning setting (Tran et al., 2016). Applying these diagnostics across different cohorts or populations offers a way to evaluate causal relationships in a graph. The CEG diagnostics are particularly useful as they admit context-specific conditional independence.

The full separation theorem for CEGs imparts a powerful representation of context-specific conditional independence relationships. The new separation theorem confirms that we can develop new models that relax the assumptions of the BN to encompass broader models. Additional software development to traverse the equivalence class of CEGs and identify both naive cause and genuine cause is in progress.

Chapter 6 outlined examples of customised graphical models, a first step towards developing a general theory of causal modelling. Future work could continue to define general definitions of causation. New classes of graphical models present further extensions of concepts like the essential graph. For example, the equivalence class of the Flow Graph remains an open question. The Controlled Regulatory Graph also prompts open questions about naive cause, instrumental variables, and the features of custom intervention. These may prompt additional general definitions of causation for custom models. As more custom models are developed, the framework for causation can be continually revised to include new forms of control.

Bibliography

O.O. Aalen, K. Røysland, J.M. Gran, R. Kouyos, and T. Lange. Can we believe the dags? a comment on the relationship between causal dags and mechanisms. *Statistical Methods in Medical Research*, 25(5):2294–2314, 2016.

O. Anacleto and C. Queen. Dynamic Chain Graph Models for Time Series Network Data. *Bayesian Analysis*, 12(2):1–19, 2016. doi: 10.1214/16-BA1010.

S.A. Andersson. An Alternative Markov Property for Chain Graphs. *Scandinavian Journal of Statistics*, 28(1):33–85, 2001. URL `https://arxiv.org/ftp/arxiv/papers/1302/1302.3553.pdf`.

S.A. Andersson, D. Madigan, and M.D. Perlman. A characterization of markov equivalence classes for acyclic digraphs. *The Annals of Statistics*, 25(2):505–541, 1997.

M. Aoki. *Optimal control and system theory in dynamic economic analysis*. North Holland publishing company, Amsterdam, 1976.

Collazo R.A. Smith J.Q. Thwaites P.A. Barclay, L.M. and A.E. Nicholson. The dynamic chain event graph. *Electronic Journal of Statistics*, 9(2):2130–2169, 2015. ISSN 19357524. doi: 10.1214/15-EJS1068.

L.M. Barclay. *Modelling and Reasoning with Chain Event Graphs in Health Studies*. PhD thesis, Univesity of Warwick, 2014. URL `http://go.warwick.ac.uk/wrap`.

L.M. Barclay, J.L. Hutton, and J.Q. Smith. Refining a Bayesian Network using a Chain Event Graph. *International Journal of Approximate Reasoning*, 54(9): 1300–1309, 2013. doi: 10.1016/j.ijar.2013.05.006.

L.M. Barclay, J.L. Hutton, and J.Q.Smith. Chain event graphs for informed missingness. *Bayesian Analysis*, 9(1):53–76, 2014. doi: 10.1214/13-BA843.

M. J. Barons, S.K. Wright, and J.Q. Smith. Eliciting probabilistic judgements for integrating decision support systems. In *Elicitation: The science and art of structuring judgement*.

Zhong X. Barons, M.J. and J.Q. Smith. Dynamic bayesian networks for decision support and sugar food security. *CRiSM Rep.(submitted)*, 2014.

T. Bedford and R.M. Cooke. *Probabilistic risk analysis: foundations and methods*. University of Cambridge Press, 2001.

T. Bedford and R.M. Cooke. Vines - A new graphical model for dependent random variables. *Annals of Statistics*, 30(4):1031–1068, 2002. ISSN 00905364. doi: 10.1214/aos/1031689016.

T. Bedford, A. Daneshkhah, and K.J. Wilson. Approximate uncertainty modeling in risk analysis with vine copulas. *Risk Analysis*, 36(4):792–815, 2016.

T. Bedford, S. French, A.M. Hanea, and G.F. Nane, editors. *Expert Judgement in Risk and Decision Analysis*. Springer, 2020.

I.A. Beinlich, H.J. Suermondt, R.M. Chavez, and G.F. Cooper. The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. In *AIME 89*, pages 247–256. Springer, 1989.

Clemen R. Maguire L. Borsuk, M. and K. Reckhow. Stakeholder values and scientific modeling in the Neuse River watershed. *Group Decision and Negotiation*, 10(4): 355–373, 2001.

R.R. Bouckaert and M. Studenỳ. Chain graphs: Semantics and expressiveness. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 69–76. Springer, 1995.

C. Boutilier, M. Goldszmidt, M. Park, and D. Koller. Context-Specific Independence in Bayesian Networks. *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence*, pages 115–123, 1997.

R.J. Bowden and D.A. Turkington. *Instrumental variables*, volume 8. Cambridge University Press, 1990.

Y.E. Chee, L. Wilkinson, A.E. Nicholson, P.F. Quintana-Ascencio, J.E. Fauth, D. Hall, K. J. Ponzio, and L. Rumpff. Modelling spatial and temporal changes with GIS and Spatial and Dynamic Bayesian Networks. *Environmental Modelling & Software*, 82:108–120, 2016. doi: 10.1016/j.envsoft.2016.04.012.

M. Chilton and D. Rose. A rights-based approach to food insecurity in the United States. *American Journal of Public Health*, 99(7):1203–11, 2009. doi: 10.2105/AJPH.2007.130229.

R.A. Collazo and J.Q. Smith. A New Family of Non-Local Priors for Chain Event Graph Model Selection. *Bayesian Analysis*, (4):1–37, 2015a. doi: 10.1214/15-BA981.

R.A Collazo and J.Q. Smith. A New Family of Non-Local Priors for Chain Event Graph Model Selection. *Bayesian Analysis*, (4):1–37, 2015b. ISSN 1936-0975. doi: 10.1214/15-BA981. URL `http://projecteuclid.org/euclid.ba/1448852254`.

R.A. Collazo, C. Görgen, and J.Q. Smith. *Chain Event Graphs*. CRC Press, 2018.

Rodrigo A Collazo and Jim Q Smith. An n time-slice dynamic chain event graph. *arXiv preprint arXiv:1808.05726*, 2018.

R.M. Cooke. *Experts in Uncertainty: Opinion and Subjective Probability in Science*. Oxford University Press, New York, 1991.

L. Costa. Studying effective brain connectivity using multiregression dynamic models. 2014. URL `http://wrap.warwick.ac.uk/65774`.

L. Costa, J.Q. Smith, T. Nichols, J. Cussens, E.P. Duff, and T.R. Makin. Searching multiregression dynamic models of resting-state fMRI networks using Integer programming. *Bayesian Analysis*, 10(2):441–478, 2015. doi: 10.1214/14-BA913.

L. Costa, T. Nichols, J.Q. Smith, et al. Studying the effective brain connectivity using multiregression dynamic models. *Brazilian Journal of Probability and Statistics*, 31(4):765–800, 2017.

L. Costa, J.Q. Smith, and T. Nichols. A group analysis using the multiregression dynamic models for fmri networked time series. *Journal of statistical planning and inference*, 198:43–61, 2019.

R. G. Cowell, R. J. Verrall, and Y. K. Yoon. Modeling operational risk with Bayesian networks. *Journal of Risk and Insurance*, 74(4):795–827, 2007. doi: 10.1111/j.1539-6975.2007.00235.x.

R.G. Cowell and J. Q. Smith. Causal discovery through MAP selection of stratified chain event graphs. *Electronic Journal of Statistics*, 8(1):965–997, 2014. doi: 10.1214/14-EJS917.

R.G. Cowell, A.P. Dawid, S.L. Lauritzen, and D.J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer- Verlag, New York, US, 1999.

D. R. Cox and N. Wermuth. *Multivariate dependencies: Models, analysis and interpretation*. Chapman and Hall/CRC, 2014.

D.R. Cox and N. Wermuth. Linear Dependencies represented by Chain Graphs. *Statistical Science*, 8(3):204–283, 1993.

A. P. Dawid. Prequential Data Analysis. *Lecture Notes-Monograph Series*, 17: 113–126, 1992. doi: doi:10.2307/4355629.

A. P. Dawid. Separoids: A mathematical framework for conditional independence and irrelevance. *Annals of Mathematics and Artificial Intelligence*, 32(1-4):335–372, 2001. doi: 10.1023/A:1016734104787.

A.P. Dawid. Conditional Independence in Statistical Theory. *Journal of the Royal Statistical Society, Series B*, 41(1):1–31, 1979.

A.P. Dawid. Statistical Theory: The Prequential Approach, 1984. ISSN 00359238. URL `http://scholar.google.com/scholar?hl=en{&}btnG=Search{&}q=intitle:Statistical+Theory:+The+Prequential+Approach{#}1`.

A.P. Dawid. Influence diagrams for causal modelling and inference. *International Statistical Review*, 70(2):161–189, 2002.

A.P. Dawid and V. Didelez. Identifying the consequences of dynamic treatment strategies: A decision-theoretic overview. *Statistics Surveys*, 4:184–231, 2010.

A.P. Dawid and M. Studenỳ. Conditional products: An alternative approach to conditional independence. In *AISTATS*, 1999.

S. Demiralp and K.D. Hoover. Searching for the causal structure of a vector autoregression. *Oxford Bulletin of Economics and statistics*, 65:745–767, 2003.

E.A. Dowler and D. O'Connor. Rights-based approaches to addressing food poverty and food insecurity in Ireland and UK. *Social Science & Medicine*, 74(1):44–51, 2012. ISSN 02779536. doi: 10.1016/j.socscimed.2011.08.036. URL `http://linkinghub.elsevier.com/retrieve/pii/S0277953611005545`.

J. Durbin and S.J. Koopman. *Time series analysis by state space methods*. Oxford university press, 2012.

M. Eichler and V. Didelez. Causal reasoning in graphical time series models. *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence, UAI 2007*, 1: 109–116, 2007.

Fadlalla G Elfadaly and Paul H Garthwaite. Eliciting dirichlet and connor–mosimann prior distributions for multinomial models. *Test*, 22(4):628–646, 2013.

Fadlalla G Elfadaly and Paul H Garthwaite. Eliciting dirichlet and gaussian copula prior distributions for multinomial models. *Statistics and Computing*, 27(2): 449–467, 2017.

D.M. Fergusson, L. J. Horwood, and F. T. Shannon. Social and family factors in the childhood hospital admission. *Journal of Epidemiology and Community Health*, (40):50–58, 1986.

L.J. Figueroa and J.Q. Smith. A causal algebra for dynamic flow networks. *Advances in Probabilistic Graphical Models*, 213:39–54, 2007. ISSN 14349922. doi: 10.1007/978-3-540-68996-6_2.

I. Flesch and P.J.F. Lucas. Markov equivalence in Bayesian networks. *Studies in Fuzziness and Soft Computing*, 213:3–38, 2007. ISSN 14349922. doi: 10.1007/978-3-540-68996-6_1.

M.J. Flores, A.E. Nicholson, A. Brunskill, K.B. Korb, and S. Mascaro. Incorporating expert knowledge when learning Bayesian network structure: a medical case study. *Artificial intelligence in medicine*, 53(3):181–204, 2011.

G. Freeman and J. Q. Smith. Bayesian MAP model selection of chain event graphs. *Journal of Multivariate Analysis*, 102(7):1152–1165, 2011a. ISSN 0047259X. doi: 10.1016/j.jmva.2011.03.008. URL `http://dx.doi.org/10.1016/j.jmva.2011.03.008`.

G Freeman and J.Q. Smith. Dynamic staged trees for discrete multivariate time series: Forecasting, model selection and causal analysis. *Bayesian Analysis*, 6(2): 279–306, 2011b. ISSN 19360975. doi: 10.1214/11-BA610.

M. Frydenberg. The Chain Graph Markov Property. *Scandinavian Journal of Statistic*, 17(4):333–353, 1990a.

M. Frydenberg. Marginalization and Collapsibility in Graphical Interaction Models. *The Annals of Statistics*, 18(2):790–805, 1990b.

D. Geiger and D. Heckerman. Knowledge representation and inference in similarity networks and Bayesian multinets. *Artificial Intelligence*, 82(1-2):45–74, apr 1996. ISSN 00043702. doi: 10.1016/0004-3702(95)00014-3.

D. Geiger, T. Verma, and J. Pearl. d-Separation: From Theorems to Algorithms. In *5th Workshop on Uncertainty in AI, Windsor, Ontario, Canada*, number August 1989, pages 118–124, 1990a.

D. Geiger, T.S. Verma, and J. Pearl. Identifying independence in Bayesian networks. Networks 20 507–534. *Mathematical Reviews (MathSciNet): MR1064736 Digital Object Identifier: doi*, 10, 1990b.

A.S. Goldberger. Structural equation methods in the social sciences. *Econometrica: Journal of the Econometric Society*, pages 979–1001, 1972.

J.H. Goldthorpe. Causation, statistics and sociology. Twenty-ninth Geary Lecture, 1998.

C. Görgen and J.Q. Smith. Equivalence Classes of Staged Trees. *Bernoulli*, 24(4A): 2676–2692, 2018. URL `http://arxiv.org/abs/1512.00209`.

C.W.J. Granger. Investigating Causal Relations By Econometric Models. *Econometrica*, 37(3):424–438, 1969.

C. Gundersen, B. Kreider, and J. Pepper. The Economics of Food Insecurity in the United States. *Applied Economic Perspectives and Policy*, 33(3):281–303, 2011. doi: 10.1093/aepp/ppr022.

J.M. Hammersley and P. Clifford. Markov fields on finite graphs and lattices. *Unpublished manuscript*, 46, 1971.

A. M. Hanea, M. F. McBride, M. A. Burgman, and B. C. Wintle. Classical meets modern in the IDEA protocol for structured expert judgement. *Journal of Risk Research*, 21(4):417–433, 2018. doi: 10.1080/13669877.2016.1215346.

A.M. Hanea, D. Kurowicka, and R.M. Cooke. Hybrid Method for Quantifying and Analyzing Bayesian Belief Nets. *Quality and Reliability Engineering International*, 22:709–729, 2006. doi: 10.1002/qre.

P. J. Harrison and C. F. Stevens. Bayesian forecasting. *Journal of the Royal Statistical Society: Series B (Methodological)*, 38(3):205–228, 1976.

A. C. Harvey. Analysis and generalisation of a multivariate exponential smoothing model. *Management Science*, 32(3):374–380, 1986.

D. Heckerman, D. Geiger, and D.M. Chickering. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Machine Learning*, 20(3):197–243, 1995. ISSN 15730565. doi: 10.1023/A:1022623210503.

A.B. Hill. President' s Address The Environment and Disease: Association or causation? *Proc R Soc Med*, (58):295–300, 1965.

S. Højsgaard. Graphical independence networks with the gRain package for R. *Journal of Statistical Software*, 46(10):1–26, 2012. URL `http://www.jstatsoft.org/v46/i10/`.

P. W. Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

R. A. Howard and J. E. Matheson. Influence diagrams. *Decision Analysis*, 2(3): 127–143, 2005.

F. Jensen. *sHUGIN API Refernece Manual*, 2014. URL `http://www.hugin.com`.

L. Kaye, E. Lee, and Y.Y.i Chen. Barriers to Food Stamps in New York State: A Perspective from the Field. *Journal of Poverty*, 17(1):13–28, 2013. doi: 10.1080/10875549.2012.747995.

D. Koller and A. Pfeffer. Object-Oriented Bayesian Networks. *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence.*, page 302, 1997.

K.B. Korb and A.E. Nicholson. *Bayesian Artificial Intelligence*. CRC Press, London, 2010.

S. L. Lauritzen. *Graphical Models*, volume 17. Clarendon Press, 1996.

S.L. Lauritzen and T.S. Richardson. Chain Graph Models and Their Causal Interpretations. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 64(3):321–361, 2002.

S.L. Lauritzen, A.P. Dawid, B.N. Larsen, and H.G. Leimer. Independence properties of directed Markov fields. *Networks*, 20(5):491–505, 1990.

M. Leonelli and J. Q. Smith. Dynamic Uncertainty Handling for Coherent Decision Making in Nuclear Emergency Response. *Risk Management for Complex Socio-Technical Systems*, 109(November), 2013. ISSN 0003018X.

Silvia Liverani and Jim Q. Smith. Bayesian selection of graphical regulatory models. *International Journal of Approximate Reasoning*, 77:87–104, 2015. ISSN 0888613X. doi: 10.1016/j.ijar.2016.05.007. URL `http://dx.doi.org/10.1016/j.ijar.2016.05.007`.

S. Lord, A. Helfgott, and J.M. Vervoort. Choosing diverse sets of plausible scenarios in multidimensional exploratory futures techniques. *Futures*, 77:11–27, 2016. doi: 10.1016/j.futures.2015.12.003.

T. A. Loveless. Food Stamp/ Supplemental Nutrition Assistance Program (SNAP) Receipt in the Past 12 Months for Households by State: 2010 and 2011 American Community Survey Briefs, 2010.

D. Materassi and M. V. Salapaka. Graphoid-based methodologies in modeling , analysis , identification and control of networks of dynamic systems. pages 4661–4675, 2016.

D. McAllester, M. Collins, and F. Pereira. Departmental Papers (CIS) Case-Factor Diagrams for Structured Probabilistic Modeling Case-Factor Diagrams for Structured Probabilistic Modeling. *Intelligence*, pages 382–391, 2004.

Mary S Morgan et al. *The history of econometric ideas*. Cambridge University Press, 1990.

A.E. Nicholson, C. R. Twardy, K. B. Korb, and L. R. Hope. Decision support for clinical cardiovascular risk assessment. *Bayesian networks: A practical guide to applications bayesian networks*, pages 33–52, 2008.

E. Nolen and K. Krey. The Effect of Universal-Free School Breakfast on Milk Consumption and Nutrient Intake. *Food Studies: An Interdisciplinary Journal*, 5 (4):23–33, 2015.

A. O' Hagan and J. Oakley. SHELF: the Sheffield elicitation framework, 2014. URL `http://www.tonyohagan.co.uk/shelf/`.

A O'Hagan, C. E. Buck, A. Daneshkhah, J. R. Eiser, P. H. Garthwaite, D. J. Jenkinson, J. E. Oakley, and T. Rakow. *Uncertain Judgements: Eliciting Experts' Probabilities.* John Wiley & Sons, 2006.

M. Olaf. Guidance on Expert Knowledge Elicitation in Food and Feed Safety Risk Assessment 1. 12(6):1–278, 2014. doi: 10.2903/j.efsa.2014.3734.

J. Pearl. Fusion, Propagation, and Structuring in Belief Networks. Technical report, 1986.

J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 1988.

J. Pearl. *Causality*. Cambridge University Press, New York, US, 2 edition, 2009. URL `http://site.ebrary.com/lib/warwick/reader.action?docID=10697730{#}`.

J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

J. Pearl and T.S. Verma. A Theory of Inferred Causation. *Studies in Logic and the Foundations of Mathematics*, 134:789–811, 1995.

C. M. Queen. Using the Multiregression Dynamic Model to Forecast Brand Sales in a Competitive Product Market. *Journal of the Royal Statistical Society . Series D*, 43(1):87–98, 1992.

C. M. Queen and C. J. Albers. Intervention and causality: forecasting traffic flows using a dynamic Bayesian network. *Journal of the American Statistical Association*, 104(486):669–681, 2009. doi: 10.1198/jasa.2009.0042.

J. M. Quintana and M. West. An analysis of international exchange rates using multivariate DLM's. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 36(2-3):275–281, 1987.

H. Raiffa. Decision analysis: Introductory lectures on choices under uncertainty. 1968.

T. Richardson and P. Spirtes. Ancestral Graph Markov Models. *The Annals of Statistics*, 30(4):962–1030, 2002.

J. Rougier and M. Crucifix. Uncertainty in climate science and climate policy. *Climate Modelling*, pages 361—-380, 2018. doi: 10.1007/978-3-319-65058-6_12.

G. Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 15(4): 461–464, 1987.

G. Shafer. The Art of Causal Conjecture. 1996.

A. Shenvi and J. Q. Smith. The reduced dynamic chain event graph. *arXiv preprint arXiv:1811.08872*, 2018.

J. Q. Smith. *Bayesian Decision Analysis: Principles and Practice*. Cambridge University Press, Cambridge, 2010.

J.Q. Smith. Multiregression Dynamic Models. *Journal of the Royal Statistical Society: Series B*, 55(4):849–870, 1993. doi: 6/96/58267.

J.Q. Smith and P.E. Anderson. Conditional independence and chain event graphs. *Artificial Intelligence*, 172(1):42–68, 2008. ISSN 00043702. doi: 10.1016/j.artint.2007.05.004.

P. Spirtes and K. Zhang. Causal discovery and inference: concepts and recent methodological advances. *Applied Informatics*, 3(1):3, 2016. doi: 10.1186/s40535-016-0018-x.

P. Spirtes, C.N. Glymour, and R. Scheines. *Causation, prediction, and search*. MIT Press, 2000.

M. Studenỳ. Attempts at axiomatic description of conditional independence. *Kybernetika*, 25(7):72–79, 1989.

M. Studenỳ. Formal properties of conditional independence in different calculi of ai. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 341–348. Springer, 1993.

M. Studenỳ. Characterization of essential graphs by means of an operation of legal component merging. *Graphical Models*, 2002. ISSN 02184885. doi: 10.1142/S0218488504002576.

M. Studenỳ. *Probabilistic conditional independence structures*. Springer Science & Business Media, 2006.

P. A. Thwaites and J. Q. Smith. A New Method for tackling Asymmetric Decision Problems. (Id), 2015.

P.A. Thwaites and E. Smith, J.Q.and Riccomagno. Causal analysis with chain event graphs. *Artificial Intelligence*, 174(12):889–90, 2010.

P.A. Thwaites, J.Q. Smith, and R.G. Cowell. Propagation using Chain Event Graphs. *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence, UAI 2008*, 2008.

Dustin Tran, Francisco JR Ruiz, Susan Athey, and David M Blei. Model criticism for bayesian causal inference. *arXiv preprint arXiv:1610.09037*, 2016.

G. Varando, F. Carli, M. Leonelli, and E. Riccomagno. *stagedtrees: Staged Event Trees*, 2020. URL `https://cran.r-project.org/package=stagedtrees`.

J. M. Vervoort, P. K. Thornton, P. Kristjanson, W. Förch, P. J. Ericksen, K. Kok, J. S.I. Ingram, M. Herrero, A. Palazzo, A.. Helfgott, A. Wilkinson, P. Havlík, D. Mason-D'Croz, and C. Jost. Challenges to scenario-guided adaptive action on food security under climate change. *Global Environmental Change*, 28:383–394, sep 2014. ISSN 09593780. doi: 10.1016/j.gloenvcha.2014.03.001.

N. Wermuth. Personal correspondence, October 2017.

M. West and J. Harrison. *Bayesian forecasting and dynamic models*. New York, 1997. ISBN 9780387947259.

R. L. Wilkerson and K. Krey. Associations Between Neighborhoods and Summer Meals Sites : Measuring Access to Federal Summer Meals Programs. *Journal of Applied Research on Children: Informing Policy for Children at Risk*, 6(2), 2015.

Kevin James Wilson et al. Specification of informative prior distributions for multinomial models using vine copulas. *Bayesian Analysis*, 13(3):749–766, 2018.

P.G. Wright. *Tariff on animal and vegetable oils*. Macmillan Company, New York, 1928.

S. Wright. Correlation and causation. *Journal of Agricultural Research*, 7(7):557–585, 1921.

S. Wright. *Corn and hog correlations*. Number 1300. US Department of Agriculture, 1925.

Rita Esther Zapata-Vázquez, Anthony O'Hagan, and Leonardo Soares Bastos. Eliciting expert judgements about a set of proportions. *Journal of Applied Statistics*, 41(9):1919–1933, 2014.

Z.Y. Zhao, M. Xie, and M. West. Dynamic dependence networks: Financial time series forecasting and portfolio decisions. *Applied Stochastic Models in Business and Industry*, 32(3):311–332, 2016.