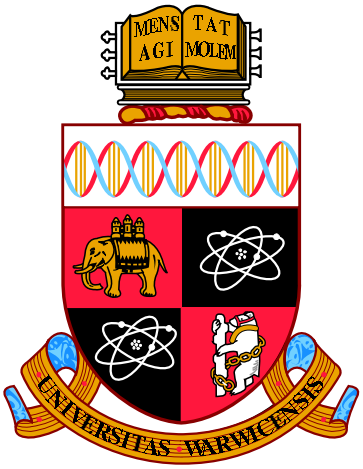


Modelling and Reasoning with Chain Event Graphs in Health Studies

by

Lorna Marion Barclay

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Department of Statistics

February 2014

THE UNIVERSITY OF WARWICK

Contents

List of Tables **iii**

List of Figures **v**

Acknowledgments **ix**

Declarations **x**

Abstract **xi**

Chapter 1 Introduction **1**

1.1 Motivation and Thesis Outline 1

1.2 Applications to Health Data 4

1.2.1 The Christchurch Health and Development Study 5

1.2.2 The UK Cerebral Palsy Birth Cohort 8

1.2.3 A Fictional Example on Influenza 12

Chapter 2 Introduction to Graphical Models and Chain Event Graphs **13**

2.1 Graphical Models 14

2.1.1 An Overview of Graphical Models 14

2.1.2 Graph Theory 16

2.2 Review of Bayesian Networks 18

2.3 Chain Event Graphs and Their Semantics 23

2.3.1 Review of CEGs 24

2.3.2 Conditional Independency in CEGs 32

2.3.3 Causality in CEGs 34

2.4 Writing a BN as a CEG 36

2.5 The Ordinal CEG 40

2.6 Alternative Graphical Models 42

Chapter 3 Bayesian Learning of Bayesian Networks and Chain Event Graphs **45**
3.1 Learning the Parameters of BNs and CEGs 46
3.2 Model selection for BNs and CEGs 50
3.3 Application to the CHDS 56
3.3.1 The Maximum a Posteriori BN for the CHDS Example . . . 56
3.3.2 The Maximum a Posteriori CEG for the CHDS Example . . 59
3.3.3 The Ordinal MAP CEG and a Causal Intervention on the MAP CEG 63

Chapter 4 Modelling Missingness using Chain Event Graphs **66**
4.1 Introduction to Analysing Missing Data Structures 66
4.2 CEGs for Informed Missingness 69
4.2.1 CEG Representation of MCAR, MAR and MNAR 71
4.2.2 Application to the UKCP 76
4.3 CEGs for Informed Missingness for more than One Variable 80
4.4 CEGs for Informed Variable Construction 87

Chapter 5 The Dynamic Chain Event Graph **94**
5.1 Dynamic Chain Event Graphs and Their Semantics 95
5.1.1 Infinite Trees and the DCEG 95
5.1.2 The DCEG with Holding Times 99
5.2 Writing a Dynamic BN as a DCEG 103
5.3 The DCEG and Markov Processes 112
5.4 Bayesian Learning of DCEGs 119

Chapter 6 Discussion **128**
6.1 Summary 128
6.2 CEGs with Informative Priors 129
6.3 CEGs with Sparse Cell Counts 132
6.4 Future Work 136

Appendix A Latent-Class Model **138**

Appendix B R Code for the AHC Algorithm **140**

List of Tables

1.1 Summary statistics on the variables social background, economic situation, life events and hospital admission of the CHDS example (% of individuals admitted to hospital for each category is given in brackets) 7

1.2 Number of individuals in the UK cerebral palsy cohort with non-severe, severe or missing visual, ambulatory and manual impairment (% of individuals with survival up to or above the age of 5 is given in brackets) 10

1.3 Number of individuals in the UK cerebral palsy cohort with non-severe, severe or missing visual or ambulatory impairment and very low, low or normal birth weight (% of individuals with survival up to or above the age of 5 is given in brackets) 11

3.1 Scale of evidence for Bayes Factors according to Kass and Raftery [1995] 54

3.2 The associated table of CPVs associated with the MAP BN from Figure 3.1 58

3.3 The iterations of the AHC algorithm when finding the MAP CEG structure for the CHDS example given by the event tree in Figure 2.3 60

3.4 Table of CPVs associated with the MAP CEG structure given in Figure 3.3 61

3.5 Table of CPVs associated with the manipulated CEG of the CHDS example given in Figure 3.4 64

4.1 Table of CPVs associated with the MAP CEG for the UKCP example on birth weight, visual impairment and survival given in Figure 4.8 . 77

4.2 Table of CPVs associated with the MAP CEG for the UKCP example on birth weight, visual impairment, ambulatory impairment and survival given in Figure 4.10 83

4.3 Plausibility of the MAR assumption in the UKCP example describing the effect of birth weight, visual impairment, ambulatory impairment and missing impairments on survival 86
4.4 Table of CPVs associated with the MAP CEG for the UKCP example on visual, ambulatory, manual impairment and missingness of impairments on survival given in Figure 4.12 88

5.1 Prior distributions on CPVs and conditional holding times associated with the DCEG from Figure 5.14 for the flu example 125

6.1 Number of individuals in the UK cerebral palsy cohort with non-severe, severe or missing visual, ambulatory and manual impairment 133

List of Figures

2.1 The BN of the CHDS example on social background, economic situation, life events and hospital admission derived from Fergusson et al. [1986] 21
2.2 An alternative ordering of the variables in the BN of the CHDS example 21
2.3 An event tree $\mathcal{T}$ for the CHDS example, with the variables taking the ordering: social background, economic situation, life events and hospital admission 25
2.4 A staged tree version of the event tree $\mathcal{T}$ for the CHDS example, with the variables taking the ordering: social background, economic situation, life events and hospital admission 30
2.5 The CEG of the staged tree version of $\mathcal{T}$ on the variables social background, economic situation, life events and hospital admissions of the CHDS example 31
2.6 The CEG for the CHDS example from Figure 2.5 manipulated such that individuals from a low social background are given financial aid 35
2.7 A possible BN structure for the social background, economic situation, life events and hospital admission of the CHDS example 36
2.8 The staged tree version of the BN of the CHDS example from Figure 2.7 38
2.9 The CEG of the staged tree version of the BN given in Figure 2.7 . . 39

3.1 The Maximum a Posteriori BN of the CHDS example on social background, economic situation, life events and hospital admission. BN score (logarithm of the marginal likelihood) $\log L(\mathcal{B}|\boldsymbol{N}) = -2489.776$ 57
3.2 High scoring BN structures for the CHDS example on social background, economic situation, life events and hospital admission 58
3.3 The MAP CEG structure found through the AHC algorithm from the event tree in Figure 2.3. CEG score $L(\mathcal{C}|\boldsymbol{N}) = -2478.490$ 61

3.4 The MAP CEG for the CHDS example from Figure 3.3 manipulated such that individuals from a low social background are given financial aid 64

4.1 Event Tree for the UKCP example describing the effect of birth weight, visual impairment and missingness on survival 70
4.2 Ordinal CEG for the UKCP example describing the effect of birth weight, visual impairment and missingness on survival when data are MAR 72
4.3 Ordinal CEG for the UKCP example describing the effect of birth weight, visual impairment and missingness on survival when data are MCAR 74
4.4 Ordinal CEG for the UKCP example describing the effect of birth weight, visual impairment and missingness on survival when data are MNAR 74
4.5 Ordinal CEG for the UKCP example describing the effect of birth weight, visual impairment and missingness on survival when data are MNAR conditional on birth weight 75
4.6 Ordinal CEG for the UKCP example describing the effect of birth weight, visual impairment and missingness on survival when data are MNAR, with missingness having a similar effect as a severe impairment 75
4.7 Ordinal CEG for the UKCP example describing the effect of birth weight, visual impairment and missingness on survival when data are MNAR, but missingness has a positive effect on survival 76
4.8 Ordinal MAP CEG structure for the UKCP example describing the effect of birth weight, visual impairment and missingness on survival 77
4.9 Event tree for the UKCP example describing the effect of birth weight, visual impairment, ambulatory impairment and missingness on survival 81
4.10 Ordinal MAP CEG structure for the UKCP example describing the effect of birth weight, visual impairment, ambulatory impairment and missingness on survival 82
4.11 Reduced Ordinal MAP CEG structure for the UKCP example describing the effect of birth weight, visual impairment, ambulatory impairment and missingness on survival 83
4.12 Ordinal MAP CEG structure for the UKCP example describing the effect of visual, ambulatory, manual impairment and missingness of impairments on survival 89

4.13 Reduced Ordinal MAP CEG structure for the UKCP example describing the effect of visual, ambulatory, manual impairment and missingness of impairments on survival 90

5.1 The beginning of the infinite tree, $\mathcal{T}$, for the flu example describing the development of catching flu, recovering with or without treatment and getting a flu vaccine 96
5.2 The beginning of the infinite staged tree, $\mathcal{T}$, for the flu example with two assumptions: 1. the probability of catching flu does not depend on whether the flu was caught in the past and 2. the probability of the individual returning to his normal life is independent of whether treatment was taken or not 98
5.3 The DCEG of the infinite staged tree from Figure 5.2 for the flu example 99
5.4 The beginning of the infinite tree, $\mathcal{T}^*$, for the flu example where catching flu is represented by the time spent at the root vertex . . . 102
5.5 The DCEG with holding times for the flu example, where catching flu is described by the time spent at the root vertex 103
5.6 A two time-slice DBN structure, $\mathcal{B}$, on two binary variables X_1 and X_2, where at each time point X_1 is only affected by the previous value of X_1 and X_2 is affected by both the previous value of X_2 and the current value of X_1 106
5.7 The illustration of the trees $\mathcal{T}_1$ and $\mathcal{T}_2$ of the two time-slice DBN from Figure 5.6 107
5.8 The infinite tree representation of the two time-slice DBN from Figure 5.6 108
5.9 The infinite staged tree representation of the two time-slice DBN from Figure 5.6 110
5.10 The DCEG of the infinite staged tree representation of the two time-slice DBN from Figure 5.6 111
5.11 The state-transition diagram of a Markov process on the state space (a, b, c) with initial distribution α and transition matrix P 114
5.12 DCEG representation of the Markov process with state space (a, b, c), initial distribution α and transition matrix P 114
5.13 DCEG representation of a Markov process describing the game of tossing a coin until heads has appeared three times 115
5.14 The DCEG with holding times for the flu example, where catching flu is described by the time spent at the root vertex 122

6.1 Ordinal MAP CEG structure for the UKCP example describing the effect of visual, ambulatory, manual impairment and missingness of impairments on survival, with sparse cell counts marked as dotted (counts: 0-6) or dashed (counts: 7-12 individuals) edges 134
6.2 Reduced Ordinal MAP CEG structure for the UKCP example describing the effect of visual, ambulatory, manual impairment and missingness of impairments on survival with sparse cell counts marked as dotted (counts: 0-6) or dashed (counts: 7-12 individuals) edges . . 135

Acknowledgments

First of all I would like to thank my supervisors, Prof. Jim Smith and Prof. Jane Hutton, for their guidance and support throughout the past years. Their enthusiasm and encouragement has been invaluable to this thesis and has especially encouraged me to present my work during the PhD.

Furthermore I would also like to thank Prof. J. Horwood and the CHDS group as well as the UKCP collaboration for providing the data sets discussed in this thesis. I am also grateful to Prof. Ann Nicholson for welcoming me at Monash University in November 2012.

Special thanks also goes to the whole of the Statistics Department for providing a stimulating and friendly research environment. Finally, thank you very much to my family for supporting me and believing in me as well as for proof-reading my work!

I am grateful to the EPSRC for funding my research.

Declarations

I hereby declare that this thesis is based on my own research, except when stated otherwise. Some of this work has been published or accepted for publication as follows: Some of the material in Chapter 3.3 together with the preliminary analysis in Chapter 1.2.1 and parts of Chapter 2.3 and 2.4 has now appeared in the *International Journal of Approximate Reasoning* under the title "Refining a Bayesian Network using a Chain Event Graph" and can be found in the bibliography and throughout the thesis as Barclay et al. [2013a]. Also, the first three sections of Chapter 4, as well as part of the introduction to the UKCP study in Section 1.2.2 and the Ordinal CEG in Section 2.5 has been accepted for publication in *Bayesian Analysis* under the title "Chain Event Graphs for Informed Missingness" and is cited in the thesis as Barclay et al. [2013b]. Early drafts of these papers are also available as CRiSM Research Reports 12-06 and 12-17. Both papers are co-authored with my two supervisors Prof. Jim Q. Smith and Prof. Jane L. Hutton. However, both papers were entirely written by me and the methodology and applications developed in this thesis are mine.

The material of Chapter 5 forms the basis for a third paper which is currently being revised for submission. The paper is co-authored with Prof. Jim Q. Smith and was further written in collaboration with Peter A. Thwaites and Ann E. Nicholson. However, I have been the lead author of the paper and what appears in this thesis is based on my own contribution to the paper and is written by me. An early version of this paper is available as the CRiSM Research Report 14-04.

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

Abstract

The Chain Event Graph (CEG) is a new class of graphical model, first introduced in Smith and Anderson [2008], which is derived from a probability tree by merging vertices whose associated conditional probabilities are the same. It is proving to be a useful framework for modelling asymmetric problems and further generalises the Bayesian Network (BN), by allowing for context-specific dependence structures between the variables of the problem. This thesis provides a first demonstration of the value of using the CEG in real-world applications and the new techniques developed here are motivated by problems that arise from two health studies; the Christchurch Health and Development Study (CHDS) and the UK Cerebral Palsy (UKCP) Cohort Study.

A direct comparison of the BN and CEG on the CHDS demonstrates that the CEG can lead to significantly higher scoring models than the BN and further that it enables additional conclusions to be drawn on the health study directly from the topology of its graph. An extension of the CEG, the Ordinal CEG, is developed in this thesis, which further enhances the graphical representation of the CEGs for studies with a binary outcome. Motivated by the UKCP this thesis further investigates how missing data structures can be explicitly represented by a CEG and how its graph can consequently provide a precise understanding of the influence of missingness. Finally, a dynamic version of the CEG is developed and it is demonstrated how this new class of models generalises the Dynamic BN and is further closely linked to (semi-) Markov processes. The expressiveness of this model is illustrated through a fictional example.

Chapter 1

Introduction

1.1 Motivation and Thesis Outline

Chain Event Graphs (CEGs) are a new class of graphical models which were first introduced in Smith and Anderson [2008]. To date, the most widely used graphical model is the Bayesian Network (BN). Particularly medicine and health care are two of the most popular application areas of the BN. Two well-known examples are the CHILD network [Cowell et al., 2007], which represents possible diseases that may lead to cyanosis (blue baby), or the ALARM network [Beinlich et al., 1989] for monitoring patients in intensive care units. In these applications the BN has proven to be extremely valuable for modelling complex relationships between variables and provides a useful framework for medical diagnosis, monitoring and prediction. However, a well-known short-coming of the BN is that it does not accommodate context-specific dependencies between the variables. In particular, in a type of cohort study where we are interested in a single outcome variable, such as survival or the onset of a disease, the BN may be restrictive in terms of the conclusions that can be made about the combined effect of risk factors on the outcome.

The CEG provides a richer class of models which incorporates these types of dependence structures, as well as retaining the property that conclusions can be easily read back to the client. It is derived from a probability tree by merging vertices whose associated conditional probabilities are the same and whose emanating edges explain the same unfolding events. In contrast to related models, such as the context-specific BN [Boutilier et al., 1996] and the Probabilistic Decision Graph [Jaeger et al., 2006], the CEG gives a single graphical representation of the entire problem and includes the BN as a particular subclass.

Over the past years a substantial amount of research has been carried out on CEGs, including evidence propagation [Thwaites et al., 2008], causal inference [Thwaites et al., 2010; Thwaites, 2013] and a model selection algorithm [Freeman and Smith, 2011a]. These developments are often adapted from the BN methodology, as any discrete BN can be represented by a CEG. The developments of the CEG have so far been predominantly theoretical. One of the contributions of this thesis is to motivate the applicability of the CEG to real-world problems. This is supported by a number of small applications throughout the thesis to subsets of two cohort studies, the Christchurch Health and Development Study (CHDS) and the UK Cerebral Palsy Cohort Study (UKCP).

In Chapter 3 of the thesis I will demonstrate, using an example from the CHDS, how the CEG leads to higher scoring models and more refined conclusions than can be made from a BN. This analysis gives a first explicit comparison of the CEG and the BN. Another important feature of the construction of the CEG is that it retains the paths of its associated probability tree within its graph. In an event tree, the root-to-leaf paths explain the unfolding of a sequence of events, describing different possible stories over time, and this description is retained within the structure of the CEG. In a cohort study, the use of CEGs therefore seems more intuitive than the use of a standard BN. A BN represents a set of conditional independence statements through its graph, where a directed edge from one variable to another expresses a possible dependency between the variables. However, it is demonstrated in Chapter 2.2 that directed edges can sometimes be reversed to give a BN with the same set of conditional independence statements. The CEG, in contrast, can directly represent that risk factors occur before the outcome variable of interest by ordering the variables in its associated tree. This has motivated two further new developments of the CEG in this thesis: Firstly, I have developed the Ordinal CEG to further enhance the graphical representation of the CEG for binary outcome variables, where the final vertices in the graph classify the cohort according to the outcome variable of interest. Secondly, it is common for substantial amounts of missing data to be present in such studies due to, for example, retrospective collection of data from routine health records, loss of contact with participants or participants declining to answer certain questions. Chapter 4 discusses how missing data on risk factors can be incorporated into the CEG framework and how informative conclusions on the influence of missingness can be read from the graph. In particular, it will be shown that the CEG proves to be useful for preliminary analyses on the effect of missing risk factors on survival and consequently that new informative categories of variables can be defined through the final positions in the CEG, which can be

used in a later survival analysis.

Finally, building on the substantive literature and applications of the Dynamic BN (DBN), the development of a formal dynamic version of the CEG is the next step in developing the CEG to provide a useful graphical framework for modelling longitudinal processes. In many larger cohort studies, the life history of the cohort is recorded by taking measurements repeatedly over time. This thesis gives a first representation of a new class of dynamic graphical model, the Dynamic CEG (DCEG), which formally extends the CEG to infinite trees to model the occurrence of repeated events over time as well as the time spent at each vertex in the graph.

The above can then be summarised into the following research questions discussed in this thesis:

1. What advantages does the CEG have over the commonly used BN and how can the CEG's graphical representation be improved?

2. How can missing data structures be represented in a CEG and how can this representation aid the analysis of processes where missingness is influential?

3. How can a dynamic version of the CEG be defined and how does it compare to other dynamic graphical models?

Below I will outline the specific chapters in which these research questions will be approached.

Thesis Outline

In the remainder of this chapter I will look at the two datasets used throughout the thesis to demonstrate the applicability of the CEG and illustrate the newly developed methodology. First, I will briefly introduce the Christchurch and Development Study (CHDS) and describe in detail the subset of the study considered in Chapters 2 and 3 as well as previous analyses carried out by Fergusson et al. [1986]. I will then introduce the second study of this thesis, which is the UK Cerebral Palsy Cohort Study (UKCP study) and provide summary statistics of the relevant variables considered in Chapters 4 and 6. Finally, I will provide a third, fictional example which will illustrate the methodology developed in Chapter 5.

In Chapter 2 I will first give an overview of graphical models and introduce some standard notation in graph theory. In Section 2.2 I will review BNs in more detail and then move to the CEG in Section 2.3. This section defines the CEG and explains its semantics as well as the conclusions that can be drawn from a CEG.

Section 2.4 then demonstrates how any discrete BN can be expressed as a CEG and Section 2.5 introduces the newly developed Ordinal CEG. The chapter concludes with a discussion on alternative graphical models.

In Chapter 3 I look at conjugate learning of the parameters in a BN and, in parallel, the learning of parameters in a CEG. Section 3.2 then reviews the standard model selection using Bayes Factors for BNs and CEGs. In the final Section I will apply the methodology introduced on model selection to the example of the CHDS, comparing the BN and CEG using this real dataset. I will also briefly return to the Ordinal CEG as well as to causal interventions and conditional independence statements described in Chapter 2, and apply these concepts to the CHDS.

In Chapter 4 I will explore how missing data can be incorporated into a CEG. I will first review the three well-known types of missing data; Missing Completely At Random (MCAR), Missing At Random (MAR) and Missing Not At Random (MNAR) and show how these can be explicitly represented within a CEG. I will then apply the methodology to the UKCP study in Sections 4.2.2 and 4.3 and show how we can read off informative conclusions about the missingness structure from the CEG. In particular, I will introduce a reduced version of the Ordinal CEG. In the final Section I will show how we can further construct new informative categories of variables with the CEG, which could then be used in a later analysis.

In Chapter 5 I will illustrate how observational studies with repeated measurements could be represented well by a dynamic version of the CEG, the DCEG. In Section 5.1.1 I will define a DCEG and in Section 5.1.2 I will extend this framework to add holding-time distributions to the DCEG. I will then compare the DCEG first to the DBN in Section 5.2, showing that any DBN can be written as a DCEG, and then to Markov and semi-Markov processes in Section 5.3. Finally, I will discuss how we could learn the parameters in a DCEG and give an outlook on model selection.

In the final Chapter I summarise the contributions made by this thesis. Finally, I will discuss some issues associated with the complexity of the CEG and further points of development that would extend the work of this thesis.

1.2 Applications to Health Data

The two datasets considered in this thesis are based on two birth cohort studies, the Christchurch Health and Development Study and the UK Cerebral Palsy Cohort Study. The term cohort is used in epidemiology to 'refer to a group of individuals who share a common characteristic' [Salkind, 2010]. In the two examples considered below, the first cohort consists of people born in mid-1977 in Christchurch, New

Zealand, and the second cohort looks at people in the UK born between 1966 and 1999 who are diagnosed to have cerebral palsy. A cohort study then describes a type of observational study which follows a cohort of people over time, comparing the occurrence of a particular outcome (i.e. a disease or survival), to risk factors or exposure variables. In the CHDS, data are collected over several years looking at the effect of social and family factors on later physical health as an outcome, while in the UKCP study, early measurements of risk factors such as various impairments or birth weight, are analysed to determine their impact on survival. The final example, discussed in Chapter 5 of the thesis, is fictional, where a group of individuals is followed-up on the development of influenza and their recovery with or without antiviral treatment. As an individual may develop influenza several times in his life, it is assumed that repeated measurements are taken. These measurements may either be taken at regular time intervals (e.g. daily or monthly), or data is collected as event histories, recording when an event occurs and the duration between events.

1.2.1 The Christchurch Health and Development Study

The Christchurch Health and Development Study (CHDS) has been carried out by a research group, led by Professor David Fergusson, in the Department of Psychological Medicine at Otago University. It is a cohort study, in which children born in Christchurch, New Zealand, in mid-1977 have been followed up for over 30 years. Out of 1310 children, born between 15th April and 5th August 1977, 1265 children were included in the study. Up to the present day data is still collected on around 71% of these individuals.

The study started off looking mostly at infant health during the first five years of the child's life and at possible factors affecting it, for example smoking during pregnancy, breastfeeding or the social and economic background of the family [Fergusson et al., 1980, 1981, 1986]. Later attention was drawn more towards behavioural problems such as conduct disorder and child-rearing problems, while during the child's teenage years further analysis was carried out on early alcohol and drug use, as well as mental health issues [Fergusson et al., 1994a,b]. These were related to family problems and social disadvantage, with a particular focus on various measures taken on parenting [Boden et al., 2007]. The CHDS research group has continued to follow-up the individuals' lives through adolescence up to the present.

The example used throughout Chapters 2 and 3 considers an early subset of the CHDS discussed in Fergusson et al. [1986], which studies the first five years of the Christchurch cohort. The study looks at the effect of the family's social

background, the economic status and the number of family life events on the child's health, measuring whether at least one hospital admission occurs during the first five years of the child's life. Based on previous studies of the dataset, Fergusson et al. [1986] considered only admissions due to illness and accidents as these were the only reasons for admissions that were sensitive to the social and family situation [Beautrais et al., 1982]. To describe the family's social background the CHDS group collected information about the mother's education and age at birth, the family's socioeconomic status, the child's ethnic origin, and whether the child grew up in a single or two parent family, with all variables consisting of two or three categories. These variables were then combined using factor analysis to give a single measure of the social background (see Fergusson et al. [1984] for details). Similarly, the economic status was measured as a function of the family income, possible financial difficulties, the standard of living and the quality of the accommodation of the child, as rated by an interviewer. Again these were simplified into a single measure of the overall economic situation. Of particular interest in this study was whether the effect of adverse life events in a child's lives might be associated with increased health problems. Twenty events were classed as life events, based on a variation of the Holmes and Rahe Social Readjustment Rating Scale [Holmes and Rahe, 1967], in which the mother of the child was interviewed and reported on the events that occurred. These included the experience of moving house, the husband changing job, the death of a close friend or relative, serious financial problems within the family, divorce, or a serious illness or accident within the family [Beautrais et al., 1982]. The number of live events were then grouped into four categories. Complete data was available for 890 children and so the analysis was carried out on this dataset.

Rates of hospital admissions per 100 children age $0-5$ were first compared for the various levels of each covariate separately using one-way analysis of variance, concluding that the more socially or economically disadvantaged the child's background (without adjusting for the other factors) the higher the hospital admission rate tended to be. In addition to this the probability of an admission increased significantly with the number of family life events. For further analysis a Cox proportional hazards model estimating the risk of at least one hospital admission during the five years was fitted with the three covariates 'social background', 'economic situation' and 'number of life events'. Results from Fergusson et al. [1986] showed that according to this model the family's economic status did not influence the risk of admission significantly after adjusting for the other covariates, suggesting that in this type of population financial problems were not the main reason for health problems. In contrast to this, family life events and social background both appeared to

have a significant impact on the admission rates, even after having adjusted for the other covariates. In particular, the most significant association was found between the hospital admissions rate and the number of family life events.

In Chapters 2 and 3 I will use the above example to compare BNs and CEGs in terms of the conclusions we can draw from their graphs and their model selection process. To construct the four variables of interest I aimed to follow as far as possible the methodology of Fergusson et al. [1986]. However, as the variables describing the social and economic background are discrete, predominantly with few categories, I slightly adapted the methods of Fergusson et al. [1986]. I instead fitted a latent class model using the package 'poLCA' in R [Linzer and Lewis, 2011], which relates the set of observed categorical variables to a latent categorical variables describing the social background and the economic situation respectively. For simplicity I assumed a binary latent variable. I then predicted the latent class for the social background and the economic situation that each individual is in by determining the modal probability of the classes given a particular configuration of the observed variables. A more detailed explanation of the construction of the latent-class model and the prediction of the latent classes using the Expectation-Maximisation algorithm, following Linzer and Lewis [2011], is given in the appendix A. I further split the number of life events into three approximately equal categories: $0-5$ events, $6-9$ events and ≥ 10 events. Finally, similar to Fergusson et al. [1986], the variable on hospital admission distinguishes between 'no hospital admission' and 'at least one hospital admission' Table 1.1 shows summary statistics of the four variables with the probability of at least one hospital admission added in brackets to each category of the three covariates.

Admissions	No admission 721	≥ 1 Admission 169	
Social background	High 507 (14.8%)	Low 383 (24.5%)	
Economic situation	High 283 (14.8%)	Low 607 (20.9%)	
Number of life events	$0-5$ Events 329 (11.9%)	$6-9$ Events 295 (21.0%)	≥ 10 Events 266 (25.6%)

Table 1.1: Summary statistics on the variables social background, economic situation, life events and hospital admission of the CHDS example (% of individuals admitted to hospital for each category is given in brackets)

There are 169 (19.0%) children overall with at least one admission, varying from 11.9% to 25.6% per category. For a high social background and a high economic

situation the admissions probability is around 14.8% increasing to 24.5% and 20.9% respectively for a low social background and low economic situation. The admissions probability according to the number of life events is as low as 11.9% for $0-5$ life events, 21.0% for $6-9$ life events and 25.6% for ≥ 10 life events. Hence, similarly to Fergusson et al. [1986] the life events appear to have the strongest effect on the hospital admissions probability.

I will use this example throughout Chapter 2 to introduce the semantics of the BN and the CEG. I will first demonstrate how the dependence structure between the variables of this problem can be depicted by a BN. I will further illustrate the limitations of the BN for this problem and the more refined conclusions that can be drawn from the CEG. In Chapter 3 I will then find the most probable BN and CEG structure for this data, demonstrating the added value of using a CEG in such an analysis. The Ordinal CEG, which will be introduced in Chapter 3.3.3, adds further value to the graphical representation of the CEG when a particular binary outcome variable is of interest. To avoid confusing with the terminology used in graph theory I will henceforth refer to the children in the CHDS as 'offspring'.

1.2.2 The UK Cerebral Palsy Birth Cohort

The second cohort study used to illustrate the usefulness of the CEG, specifically when we have substantial amounts of missing data, is the UK Cerebral Palsy Cohort Study (UKCP study) [Surman et al., 2006]. The Surveillance of Cerebral Palsy in Europe (SCPE) estimates cerebral palsy to occur in around 2 individuals per 1000 live births and states that it is the most common cause of significant physical disability in children [Cans, 2000]. The most cited definition of cerebral palsy is given by Bax [1964] as 'a disorder of posture and movement due to a defect or lesion in the immature brain'. The UKCP cohort defines cerebral palsy according to the SCPE, who give an extended version of this definition, and includes and postnatal cerebral palsy in their study. Apart from mobility impairments due to brain lesions, hearing, visual and mental impairments may also arise. Consequently, cerebral palsy affects the individual's life significantly and has a large social and financial impact on affected families.

The UKCP cohort is a combination of five cerebral palsy registers (Merseyside and Cheshire, Scotland, Oxford, Northern Ireland, North of England) of people born between 1960 and 1999, comprising of 6294 subjects in total, followed up until censoring in May 2012, by which time 863 deaths have occurred. Data are collected on demographics, the individual's background and on clinical information, such as severity of impairments and type of cerebral palsy [Surman et al., 2006].

It is known that cerebral palsy is associated with increased premature mortality and that people who have severe impairments are more likely to die prematurely than the general population (see for example Strauss et al. [1998]; Blair et al. [2001]; Hemming et al. [2005]). I consider the following three impairments as risk factors for survival:

- Visual impairment: binary variable distinguishing between severe and not severe (severe: acuity less than 6/60)
- Ambulatory impairment: binary variable distinguishing between severe and not severe (severe: unable to walk even with aids, in a wheelchair or bedridden)
- Manual impairment: binary variable distinguishing between severe and not severe (severe: unable to feed or dress)

Survival to early childhood is considered as a binary variable distinguishing between survival up to or above the age of 5. When an individual dies very young, before measurements on some impairments can be taken, then these data are often recorded as missing, leading to an overall substantial amount of missing data. In other situations the impairments may be so severe that taking measurements is not possible. It can be seen from Table 1.2 that all three impairments have substantial amounts of missing data. Due to follow-up using hospital records and national death and emigration records survival, on the other hand, is practically fully observed. I will discuss in Chapter 4 how the CEG enables us to determine explicitly the effect of missingness on survival and that, in this study, missingness is consequently associated with poorer survival and severity of impairments.

Four individuals are omitted from the analysis that emigrated before the age of five and one individual is omitted due to missing age giving a total of 6289 individuals which are included in the analysis. Summary statistics concerning the cohort are given in Table 1.2. The percentage of survival up to or above the age of 5 is given in brackets next to the total number of individuals in each category.

Visual impairment appears to have a strong effect on survival with the survival probability ranging from 98.7% to 85.4%. In particular, survival is lowest when visual impairment data is missing. Ambulatory impairment appears to have an even more significant effect on survival. Here a missing impairment gives a survival of only 78.4% in comparison to survival of 99.8% and 90.1% for a non-severe or severe impairment, respectively. Manual impairment has a similarly strong influence on survival as the ambulatory impairment, with a survival probability of 99.6% for non-severe, 88.6% for severe and 78.9% for missing impairment. As expected, three missing impairments lead to a comparatively low probability of survival of 78.9%

(a) Visual, ambulatory and manual impairment

Visual impairment	Ambulatory impairment	Manual impairment		
		Not severe	Severe	Missing
Not severe	Not severe	3470 (99.8)	69 (100.0)	44 (97.7)
	Severe	343 (99.1)	585 (93.0)	23 (91.3)
	Missing	19 (94.7)	0	35 (91.4)
Severe	Not severe	72 (100.0)	4 (100.0)	1 (100.0)
	Severe	37 (89.2)	413 (87.2)	8 (62.5)
	Missing	0	1 (100.0)	7 (57.1)
Missing	Not severe	325 (100.0)	9 (88.9)	16 (100.0)
	Severe	92 (97.8)	290 (79.0)	16 (37.5)
	Missing	2 (100.0)	3 (100.0)	405 (76.5)

(b) Total

	Not severe	Severe	Missing
Visual impairment	4588 (98.7)	543 (88.4)	1158 (85.4)
Ambulatory impairment	4010 (99.8)	1807 (90.1)	472 (78.4)
Manual impairment	4360 (99.6)	1374 (88.6)	555 (78.9)
Total	6289 (95.4)		

Table 1.2: Number of individuals in the UK cerebral palsy cohort with non-severe, severe or missing visual, ambulatory and manual impairment (% of individuals with survival up to or above the age of 5 is given in brackets)

and three non-severe impairments lead to a high survival probability of 99.8%. However, the combination of missing and severe impairments in some cases lead to even lower probabilities of survival of 37.5% or 57.1% than three missing impairments. It is also important to note that some cell counts are sparse, possibly zero, and consequently conclusions of a survival probability of 100.0% need to be treated with care. I will discuss the sparse cell counts in Table 3.1 further at the end of the thesis in Section 6.3 of the Discussion.

I also consider two examples where birth weight is included as a further covariate influencing survival, where the first looks at the effect of birth weight and visual impairment on survival and the second adds ambulatory impairment into the model. Birth weight is split into three categories: very low ($\leq$ 1.5kg), low ($1.5 - 2.5$kg) and normal ($>$ 2.5kg) birth weight. There are 55 individuals in the study which have missing birth weight. As this is less than 1% of the individuals in the study, it seems reasonable to omit these individuals to avoid sparseness of the cell counts and I will hence assume throughout that birth weight is fully observed. As mentioned above, sparsity of cell counts will be discussed in Chapter 6.3. Summary

statistics on birth weight, visual and ambulatory impairment and their effect on survival is given in Table 1.3.

(a) Birth weight, ambulatory and manual impairment

Birth weight	Visual impairment	Ambulatory impairment			
		Not severe	Severe	Missing	Total
Very low	Not severe	650(99.5)	191(96.3)	11(81.8)	1100 (96.5)
	Severe	18 (100.0)	61 (85.2)	4 (75.0)	
	Missing	66 (100.0)	53 (81.1)	46 (84.8)	
Low	Not severe	857 (100.0)	235 (94.5)	9 (100.0)	1521 (96.6)
	Severe	15 (100.0)	123 (89.4)	2 (50.0)	
	Missing	105 (100.0)	83 (89.2)	92 (83.7)	
Normal	Not severe	2066 (99.76)	524 (95.0)	31 (93.5)	3613 (95.2)
	Severe	44 (100.0)	271 (86.3)	2 (50.0)	
	Missing	176 (99.4)	261(79.3)	238 (79.4)	

(b) Total

	Not severe	Severe	Missing
Visual impairment	4574 (98.7)	540 (88.5)	1120 (87.1)
Ambulatory impairment	3997 (99.8)	1802 (90.1)	435 (82.1)
Total	6234 (95.7)		

Table 1.3: Number of individuals in the UK cerebral palsy cohort with non-severe, severe or missing visual or ambulatory impairment and very low, low or normal birth weight (% of individuals with survival up to or above the age of 5 is given in brackets)

From Table 1.3 it can be concluded that birth weight does not appear to have as strong an effect on the survival as either of the disabilities. However, there is a slight tendency that a normal birth weight reduces survival slightly. Also note that Table 1.3 has 55 individuals less than Table 1.2 and this appears to affect mostly the number of individuals with missing visual or missing ambulatory impairment, improving the survival for these categories to 87.1% and 82.1%. This suggests that missing birth weight is associated with missing impairments and poorer survival. Again two missing or one missing and one severe impairment lead to the lowest probablities of survival across all birth weights, especially when birth weight is normal. This is followed by two severe impairments across all birth weights. However, given a very low birth weight, non severe visual impairment and missing ambulatory impairment also leads to a survival probability of only 81.8%. The survival probability is overall extremley high for a non-severe ambulatory impairment.

1.2.3 A Fictional Example on Influenza

The final example used in this thesis is a fictional example on a cohort of people who may develop influenza. Influenza is an acute viral disease caused by RNA (ribonucleic acid) viruses. We distinguish between three types: influenza virus A, B and C, where virus C occurs much less frequently than the other two. Typical symptoms such as fever, headache or nausea last for about one week. Antiviral treatments can reduce the length and severity of the infection. The impact of flu can vary from year to year and usually peaks in winter. The infection rate is commonly measured in terms of the number of infected people per week. For example, the Department of Health (UK) reported that at the end of 2010 the illness rate peaked at around 120 people per week, and in the winter of 2000 this was as many as 250 people per week. During these periods the excess-death rate was estimated to be around 16.8 per $100,000$. To prevent an infection it is possible to take an annual influenza vaccine which, according to the World Health Organisation, can prevent $70 - 90\%$ of influenza illnesses.

To illustrate the new methodology developed on Dynamic CEGs, I will slightly simplify the problem by assuming that when an individual catches flu and takes an antiviral, then this always leads to full recovery. Similarly, I will assume that the vaccine is 100% effective. I consider first a study where measurements record monthly whether the individual catches influenza and then whether he takes antiviral treatment or not, his potential recovery and his decision to either take an influenza vaccine or risk catching influenza again. I will then also consider the possibility of a more detailed study where also the time until an event occurs is recorded. This would monitor explicitly the duration of the individual being healthy and the duration of recovery after illness, as well as the time until a decision to take treatment or a vaccine is made.

Chapter 2

Introduction to Graphical Models and Chain Event Graphs

In this Chapter I will introduce the CEG [Smith and Anderson, 2008; Thwaites et al., 2010; Freeman and Smith, 2011a; Barclay et al., 2013a,b] describing its derivation as well as explaining its main features as a graphical model. I will first review the most commonly used graphical model, the BN, in Section 2.2, which has been studied for example in Lauritzen [1996]; Korb and Nicholson [2004] or Cowell et al. [2007]. However, as mentioned in Chapter 1.2.3, in certain cases the BN does not provide a rich enough structure to incorporate all information obtainable from the data set. This is the case, for example, when the conditional independence statements of the problem are asymmetric or only certain combinations of variables affect another variable which cannot be represented simply by the directed edges between variables in the BN [Poole and Zhang, 2003].

The CEG has therefore been proposed as a new flexible class of graphical models which can represent asymmetric structures directly in its topology and which I will introduce in Section 2.3. Throughout this and the subsequent chapter I will illustrate the advantages of modelling a problem using CEGs over the BN and will show in Section 2.4 that the BN is a subclass of the CEG, supporting the claim that the CEG can give a more detailed representation than a BN. The semantics of the CEG can be further extended to the 'Ordinal CEG', which I will define in Section 2.5. This orders the positions of the CEG according to an outcome variable, hence adding to the expressiveness of the graph. Extensions of the BN to allow for asymmetric dependencies, such as the context-specific BN [Friedman

and Goldszmidt, 1998; Boutilier et al., 1996] and Bayesian Multinets [Geiger and Heckerman, 1996], as well as alternatives to the CEG will be briefly discussed at the end of this chapter in Section 2.6 and contrasts the CEG with alternative graphical models. I will begin the chapter by giving a brief review of the developments within graphical models and by defining some basic graph theory.

2.1 Graphical Models

2.1.1 An Overview of Graphical Models

Graphical models are statistical models for a set of random variables whose joint probability density function (pdf) or probability mass function (pmf) is described in terms of a graph and which hence possesses many advantages in terms of 'representation, inference and learning' [Kollar and Friedman, 2009]. One substantial benefit of graphical models is that the statistical model is accessible through the graphical representation of a given problem. Pearl [1986] reasons that in real world problems it is often easier for a client or a group of experts to identify conditional probabilities between a small number of variables instead of the full probability model and that therefore judgements are usually made only on a subset of variables, which the graphical models take into account. This, in particular, enables better interaction between a statistician and a client or domain, who can see for himself how the statements made are encoded in the graphical model and can amend these on the graph as necessary.

However, graphical models have not only become popular as a representational tool, but have also proven to be extremely useful for inference and learning within a complex problem. The joint distribution on a set of variables is described more compactly in the graph, commonly through conditional independence statements between the variables. This allows for efficient computation of posterior distributions and the propagation of evidence within a high-dimensional setting from which inference can be drawn. Also, when learning a statistical model the computational benefits of graphical models allow for efficient model selection techniques.

Graphical models have now been widely studied for example in Lauritzen [1996]; Studenỳ [2005]; Kollar and Friedman [2009]; Smith [2010] and many more. We commonly distinguish between three classes of graphical models: directed acyclic graphs, undirected graphs and chain graphs [Studenỳ, 2005]. Directed acyclic graphs (see Definition 7) have only directed edges between the vertices in the graph and do not exhibit cycles. They form the underlying graph of the Bayesian Network or Belief Network (BN) [Pearl, 1986] which may have discrete variables, variables with

a Gaussian distribution or a combination of the two. The BN is one of the most commonly used graphical models and will be described in detail in Section 2.2 of this chapter. In contrast to this, undirected graphs or Markov Networks describe conditional independence statements between the variables in the graph only by undirected edges. Examples of these are the Gaussian graphical models with continuous variables or log-linear graphical models with discrete variables [Lauritzen, 1996]. Finally, we may have a mixture of directed and undirected edges in the graph, which are called chain graphs and were developed in Lauritzen and Wermuth [1989]. As this thesis focuses on CEGs, which are directed graphical models describing a set of discrete random variables, I will only consider in detail the discrete Bayesian Network and its extensions such as the context-specific BN [Boutilier et al., 1996] and Bayesian Multinets [Geiger and Heckerman, 1996] which will briefly be discussed in Section 2.6.

The above graphs all explain the distribution of a set of random variables where the variables of the problem are represented by the nodes in the graph and the edges explain possible dependencies between the variables. A different approach to representing a complex problem graphically was taken in the area of decision analysis, where probability trees and decision trees were employed. Although these are formally graphical models, they are historically not included within the graphical models literature described above. In Raiffa [1968] a decision tree describes the unfolding of a sequence of events and decisions then can be made at different points in the tree where the leaf nodes represent the final outcome, often given in terms of a utility or monetary value. Decision problems can then be analysed and an optimal decision rule found by working backwards through the tree. However, a decision tree can quickly become extremely large and so in 1976 Miller et al. [1976] attempted to tackle a large decision problem by more compact representations of the decision tree using ideas of coalescence. This led to the first formulation of the Influence Diagram [Howard and Matheson, 1981], which quickly became extremely popular as a tool for decision analysis and communication. Today it is known that the Influence Diagram is in fact an extension of the BN by adding decision nodes and a utility node to its graph (see for example Smith [2010]). Although Influence Diagrams were, at first, translated back into decision trees to be solved, it was soon shown in Shachter [1986] that Influence Diagrams could be analysed directly to find the optimal decision policy in a similarly efficient way. Decision trees consequently lost attention and Influence Diagrams and BNs became the main graphical framework for decision analysis and reasoning under uncertainty.

However, in practice domain experts still describe many problems in terms

of tree structures rather than directly as a BN. In particular, when the dependence structure between variables in the problem is asymmetric, trees often give a more intuitive representation. Therefore, part of the statistician's task is then to translate these presented problems into BNs and Influence Diagrams (see Section 2.2 on the elicitation of a model). As a consequence, Anderson and Smith [2006] developed the idea of reverting back to carrying out analyses directly on a coalesced probability tree as these appeared to be more effective in communicating with a client, leading to the Chain Event Graph [Smith and Anderson, 2008]. However, in contrast to the developments in the early 1980s, the CEG admits not only coalescence of two vertices and all future developments, but also coalescence of two vertices where only the immediate events are the same. This is represented by colouring of the graph, called the 'stage' partition of the model. Unlike the Influence Diagram and the BN the CEG still retains all the paths of the tree within its graph, giving therefore a very different, yet still expressive representation of a problem. Research has so far centred around problems that are represented by probability trees. However several research reports on the development of a 'Decision Event Graph' [Cowell et al., 2013] exist which define a CEG for a decision tree and how it contrasts to the Influence Diagram.

2.1.2 Graph Theory

In this section I will review some standard notation within graph theory according to Lauritzen [1996].

Definition 1. *A **graph** $\mathcal{G}$ consists of a set of vertices or nodes $V(\mathcal{G})$ and a set of edges $E(\mathcal{G})$, where each edge in $E(\mathcal{G})$ connects a pair of vertices in $V(\mathcal{G})$. A **directed edge** from a vertex v_i to a vertex v_j can be written as $e(v_i, v_j)$ and is drawn as an arrow from v_i into v_j.*

Definition 2. *A **cut** is a partition of the vertices $V(\mathcal{G})$ of a graph $\mathcal{G}$ into two disjoint subsets and the associated **cut-set** is an edge set in $E(\mathcal{G})$, where the vertices connected by these edges are in different subsets of the cut.*

Definition 3. *A graph is said to be a **complete graph** when there exists an edge between all pairs of vertices.*

Definition 4. *A graph is **infinite** when either the set $V(\mathcal{G})$ or the set $E(\mathcal{G})$ is infinite or both.*

Definition 5. *A set of vertices $v_1, ..., v_n$, such that there exists an edge between v_i and $v_{i+1}, \forall i = 1, ..., n-1$ is called a **path**, ϵ, of length n between v_1 and v_n.*

Let $\mathcal{E}$ be the set of all paths in $\mathcal{G}$. Similarly, if the edges are directed such that $e(v_i, v_{i+1}) \in E(\mathcal{G})$ for $i = 1, ..., n-1$, then we say we have a ***directed path*** *from v_1 to v_n. Rather than defining the path in terms of its vertices a path can also be defined by its edges $e(v_i, v_{i+1}) \in E(\mathcal{G}), i = 1, ..., n-1$.*

Definition 6. *Call $v_j \in V(\mathcal{G})$ a* ***child*** *of $v_i \in V(\mathcal{G})$ and $v_i \in V(\mathcal{G})$ a* ***parent*** *of $v_j \in V(\mathcal{G})$, if there exists a directed edge from v_i to v_j. Write the parent set of a vertex v_j as $pa(v_j)$ and the set of children of v_i as $ch(v_i)$. Similarly, call the vertices with a directed path into v_j the* ***ancestors*** *of v_j, written as $an(v_j)$, and the set of vertices with a directed path from v_i the* ***descendants*** *of v_i, written as $de(v_i)$. Finally, we have that the set $nd(v_i) = V(\mathcal{G}) \backslash de(v_i)$ is the set of* ***non-descendants*** *of v_i.*

Definition 7. *A graph is said to be* ***directed*** *if its edges are directed. A* ***cycle*** *is a special type of directed path which starts and ends with the same vertex. Hence, a* ***directed acyclic graph (DAG)*** *is a directed graph with no directed cycles.*

We can well-order the vertices in a DAG in the sense that every parent vertex comes before its children in the ordering.

Definition 8. *Call the vertices that come before a vertex v_j the* ***predecessors*** *of v_j, written as $pr(v_j)$.*

In order to define a CEG some further terminology as introduced in Shafer [1996] and Smith [2010] is needed:

Definition 9. *A* ***tree*** *$\mathcal{T} = (V(\mathcal{T}), E(\mathcal{T}))$ is a connected directed graph with no cycles. It has one vertex, called the root vertex v_0, with no parents, while all other vertices have exactly one parent.*

Definition 10. *A* ***leaf vertex*** *$l \in V(\mathcal{T})$ is a vertex with no children. A non-leaf vertex of a tree $\mathcal{T}$ is called a* ***situation****, s_i, and the set of situations is denoted by $S(\mathcal{T}) \subseteq V(\mathcal{T})$ and the set of leaf nodes by $\mathcal{L}(\mathcal{T}) = V(\mathcal{T}) \backslash S(\mathcal{T})$.*

In this thesis I will only consider **event trees** [Smith, 2010], which describe the way a process develops and the different sequences of events that may occur. Therefore, all situations in the tree are chance nodes and the edges of the tree label the possible events that can occur. This stands in contrast to the decision tree, which also includes decision nodes, describing the decisions made by a decision maker throughout the process. When the edges of the tree are labelled with the conditional probabilities of observing the following event given we have reached a particular node, the tree is called a **probability tree**.

Definition 11. *A **subtree** $\mathcal{T}(s_i)$ is a tree with vertex set $V(\mathcal{T}(s_i)) = \{s_i\} \cup de(s_i)$, and the edge set $E(\mathcal{T}(s_i))$ consisting of all the edges between s_i and its descendants in $\mathcal{T}$.*

Definition 12. *A **floret** is a subtree $\mathcal{F}(s_i) = (V(\mathcal{F}(s_i)), E(\mathcal{F}(s_i)))$ of $\mathcal{T}$, $s_i \in S(\mathcal{T})$ with vertex set $V(\mathcal{F}(s_i)) = \{s_i\} \cup ch(s_i)$, and the edge set $E(\mathcal{F}(s_i))$ consisting of all the edges between s_i and its children in $\mathcal{T}$.*

I will assume throughout this thesis that every situation, s_i, has a finite number of edges, m_i, emanating from it. Write $e_{s_i k}$ for the kth edge associated with the floret $F(s_i)$ emanating from s_i. When describing a unique path within a CEG this is essential, as multiple edges between the same two vertices may exist. Consequently, a path ϵ is therefore defined by its edge set $(e_{s_{i_1}k_1}, e_{s_{i_2}k_2}, \ldots, e_{s_{i_n}k_n})$.

2.2 Review of Bayesian Networks

The Bayesian Network [Pearl, 1986; Korb and Nicholson, 2004; Cowell et al., 2007; Jensen and Nielsen, 2007] is the most widely used graphical model which expresses the relationship between the variables of the system in terms of conditional independence statements. More explicitly, the vertices of the graph represent the variables of the problem and the directed edges between the nodes indicate possible dependencies between the variables. The BN was first defined in Pearl [1986] and its simple graphical structure has proved to be a particularly useful tool for feeding conclusions back to a client, as well as being an efficient framework for evidence propagation and model selection procedures that exploit its graphical structure. The BN has therefore been employed in many real-world applications as a framework for reasoning under uncertainty. Korb and Nicholson [2004] give an overview of real-world problems in which BNs have been employed for prediction, monitoring or diagnosis. The most common areas of applications include medicine and health care, environment, engineering, education, business and computing.

Assume that we have a joint probability mass function (pmf), $p(\boldsymbol{x})$, on a vector of random variables $\boldsymbol{X} = (X_1, X_2, ..., X_p)$. The definitions below hold simultaneously for probability density functions. However, as only discrete graphical models are being considered within this thesis, the definitions are restricted to pmfs.

We can then write the pmf p as a product of conditional probabilities:

$$p(\boldsymbol{x}) = p(x_1) \prod_{i=2}^{p} p(x_i | x_1, ..., x_{i-1}),$$

which can then be easily simplified by realising that in a given problem not all variables are dependent on each other, but that some of the variables may be conditionally independent.

Definition 13. *Let X, Y and Z be random variables. We say X and Y are **independent**, written as $X \perp\!\!\!\perp Y$, if and only if $p(x, y) = p(x)p(y)$. Also, X and Y are **conditionally independent given Z**, written as $X \perp\!\!\!\perp Y \mid Z$, if and only if*

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

when $p(z) > 0$.

It is these conditional independence statements which are represented within the topology of the BN and which form the basis for efficient model selection and propagation. In order to represent the joint probability of the variables in terms of a graph, $\mathcal{G}$, we assume a one-to-one correspondence between the set of vertices and the vector of random variables $\boldsymbol{X}$. Then by Cowell et al. [2007]:

Definition 14. *A probability distribution P on a set of random variables obeys the **ordered directed Markov property** relative to a DAG $\mathcal{G}$ if any variable is conditionally independent of its predecessors, given its parents*

$$X_i \perp\!\!\!\perp pr(X_i) \backslash pa(X_i) \mid pa(X_i). \tag{2.1}$$

The graph therefore describes the relationship between the variables via conditional independence statements between the variables with a missing edge representing conditional independence. This further leads to a simpler form of the pmf on $\boldsymbol{X}$ as given in equation 2.2

$$p(\boldsymbol{x}) = p(x_1) \prod_{i=2}^{p} p(x_i|pa(x_i)), \tag{2.2}$$

and we say that P admits a **recursive factorisation relative to** $\mathcal{G}$. A BN can now be formally defined as follows:

Definition 15. *A **Bayesian Network** (BN) on a vector of variables $\boldsymbol{X} = (X_1, X_2, ..., X_p)$ is made up of three components:*

1. *A DAG $\mathcal{B}$, with vertex set $V(\mathcal{B}) = \{X_1, X_2, ..., X_p\}$ and with a directed edge from X_j to X_i if and only if $X_j \in pa(X_i)$.*

2. *A set of* $p-1$ *conditional independence statements of the form*

$$X_i \perp\!\!\!\perp pr(X_i) \backslash pa(X_i) \mid pa(X_i).$$

3. *A set of Conditional Probability Vectors (CPVs),* $p(x_i|pa(x_i))$*, associated with each of the vertices* X_i*.*

When $pr(X_i)\backslash pa(X_i)$ is the empty set, then the corresponding conditional independence statement is trivial. When defining a BN commonly only the non-trivial conditional independence statements are given. The elicitation of the first two components of the BN is often referred to as qualitative modelling [Cowell et al., 2007]. Here, the expert can build a model focusing at first only on the qualitative structure and discuss the dependence structure between the variables by looking at the edges in the graph, even without specifying the CPVs yet. The final quantitative component of the BN can then be elicited later, once the structure has been determined. Kjaerulff and Madsen [2007] or Smith [2010] explain how to proceed when eliciting a BN qualitatively from an expert, while details on how to elicit the probabilities can be found in O'Hagan et al. [2006]. The elicited probabilities of the BN can further be updated with available data and Bayesian model selection techniques can be used to determine the maximum a posteriori (MAP) BN structure given data. This will be discussed in detail in Chapter 3.1.

Below I give an example of a BN on four variables using the example of the Christchurch Health and Development Study introduced in Chapter 1.2.1.

Example 1. *Consider the following four variables of the CHDS.*

- X_1 = *family social position: binary variable: 'low', 'high'*
- X_2 = *family economic situation: binary variable: 'low', 'high'*
- X_3 = *number of family life events: variable with three categories : 'low', 'average', 'high'*
- X_4 = *hospital admission: binary variable: 'yes', 'no'.*

Recall that the conclusion drawn in Fergusson et al. [1986] states that the economic situation has no effect on the hospital admissions probability once adjusting for the social background and the family life events. This can be directly translated into the conditional independence statement:

$$X_4 \perp\!\!\!\perp X_2 \mid X_1, X_3,$$

with the corresponding DAG given in Figure 2.1.

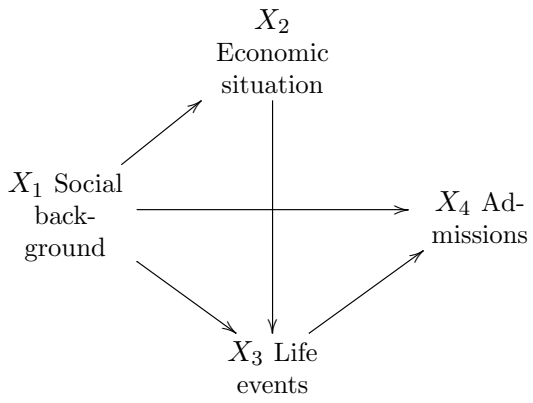


Figure 2.1: The BN of the CHDS example on social background, economic situation, life events and hospital admission derived from Fergusson et al. [1986]

Note that the definition of a BN requires an ordering of the variables such that the parent variables come before its children. This well-ordering of variables in a DAG is however not always unique. Consider, for example, the DAG given in Figure 2.2:

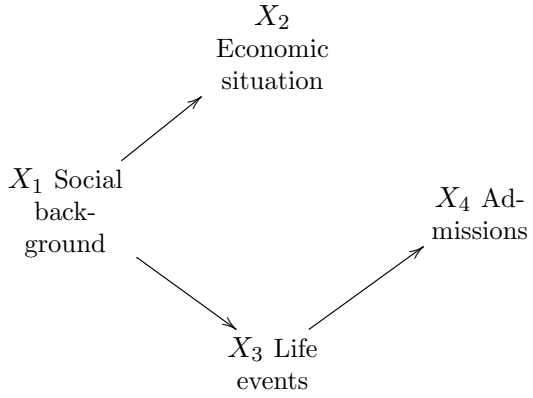


Figure 2.2: An alternative ordering of the variables in the BN of the CHDS example

We could have either the immediate ordering $\boldsymbol{X} = (X_1, X_2, X_3, X_4)$ with conditional independence statements $X_3 \perp\!\!\!\perp X_2 \mid X_1$ and $X_4 \perp\!\!\!\perp X_1, X_2 \mid X_3$, or alternatively $\boldsymbol{X} = (X_1, X_3, X_4, X_2)$ with conditional independence statements $X_4 \perp\!\!\!\perp X_1 | X_3$ and $X_2 \perp\!\!\!\perp X_3, X_4 \mid X_1$. In both cases parents come before children. However, it is proven, for example in Smith [1989], that the conditional independence statements derived from the DAG are deducible from each other. A stronger statement than the ordered directed Markov property is the **local directed Markov property**, which takes this into account by stating that X_i is independent of all non-descendants given its parents, i.e. $X_i \perp\!\!\!\perp nd(X_i) \backslash pa(X_i) \mid pa(X_i)$.

Apart from the conditional independencies that are apparent from the missing edges, further conditional independencies on subsets of variables can be deduced directly from the graph. These independencies are given by the **d-separation theorem** defined by Pearl [1986] or, by the **global directed Markov property** [Lauritzen et al., 1990], a different formulation of the d-separation theorem. During elicitation the statistician may deduce conditional independencies from the d-separation theorem and use these to check the validity of the model with the expert.

From the above definitions it becomes evident that a set of conditional independence statements could be represented by several DAGs. Hence, for example, the conditional independence statement $X_4 \perp\!\!\!\perp X_2 \,|\, X_1, X_3$ given in Figure 2.1 could also be represented by a DAG with the edge from X_2 to X_3 reversed. Therefore, two BNs may encode exactly the same conditional independence statements even though their associated DAGs are different and we then say that the two BNs are **equivalent**. It was proved in Verma and Pearl [1990] that two BNs are equivalent if and only if 1. their associated DAGs have the same topology, and 2. whenever two parents of a variable X are unconnected in one of the DAGs then this structure is also present in the other DAG. Hence, the edges within a standard BN are not to be interpreted causally. An edge from a variable X_1 to a variable X_2 does not mean that X_1 causes X_2 but simply that there exists a possible association between the two variables. Pearl's book Causality (2000) is a major contribution to the literature on causality and claims that while associations between variables can be described by joint probability distributions, causal relationships always underlie some judgments or assumptions made that cannot be derived from an observational study. Based on Holland's [Holland, 1986] slogan 'No causation without manipulation', Pearl [2000] and many others (e.g. Lauritzen [2001] and Dawid [2002]) describe a cause in terms of an intervention or manipulation. So, when 'X_1 causes X_2', this means that manipulating X_1 (i.e. forcing it to a particular value) changes the probability distribution of X_2 for at least some value x_1^*. To distinguish between 'conditioning by observation' and 'conditioning by intervention', Lauritzen [2001] uses the notation $p(\boldsymbol{x}||x_A^*)$ to describe the joint pmf given that a subset of variables X_A has been manipulated to x_A^*. This then gives the so-called intervention formula:

$$\begin{aligned} p(\boldsymbol{x}||x_A^*) &= \prod_{x_i \notin x_A} p(x_i|pa(x_i))\Big|_{x_A=x_A^*} \\ &= \frac{\prod_{i=1}^{p} p(x_i|pa(x_i))}{\prod_{x_i \in x_A} p(x_i^*|pa(x_i))}\Big|_{x_A=x_A^*}. \end{aligned}$$

The above can be understood by thinking of the joint probability in the

form of its factorisation according to the associated DAG, evaluated at $X_A = x_A^*$. By fixing a subset of variables to a particular value, $X_A = x_A^*$, the conditional probabilities for $x_i \in x_A$, are simply equal to one, as the x_i have been fixed to x_i^* by intervention. In general, conditioning by observation is not the same as conditioning by intervention. A Causal BN is then defined as follows:

Definition 16. *A BN is a **Causal BN** if the system, under manipulation, admits the intervention formula.*

Pearl's approach is often criticised for several reasons [Smith, 2010; Shafer, 1996]. The obvious criticism is that it is assumed that an intervention, e.g. fixing $X_k = x_k^*$, say, has the same effect on the child variables as if x_k^* had simply been observed. This assumption however, often does not appear to be valid in applications when dealing with rational individuals with respect to problems in social sciences and medical applications. Consider, for example, a policy intervention programme for the CHDS, where financially disadvantaged families receive financial aid. It is likely that the effect of a family receiving financial help is not the same as when they have earned the money themselves.

A further criticism is that Pearl does not incorporate a time element into his model. A cause must always occur before the affected variable, but this is only implicitly assumed through the ordering of the variables. Shafer [1996] therefore suggests inferring causal assumptions from probability trees, as a tree gives a natural description of the way a process unfolds. Similarly, I will show in the subsequent section that the CEG retains the paths of its associated probability tree and consequently that the CEG may be advantageous for problems, such as cohort studies, where the variables take a particular ordering over time.

2.3 Chain Event Graphs and Their Semantics

In 2008, Smith and Anderson [2008] proposed the CEG as an alternative graphical model to the BN, which has the significant advantage of allowing for asymmetric dependence structures. For example, a dependence structure between two variables may only occur for a particular parent configuration, which cannot be captured by a BN. Boutilier et al. [1996] call this a context-specific conditional independency and define it as follows:

Definition 17. *Let X, Y C be random variables in a given problem. Then X and Y are **contextually independent** given the context $C = c$ if*

$$P(X|c, Y) = P(X|c), \text{ whenever } P(c, Y) > 0. \tag{2.3}$$

This means that two variables are independent given the value of a third variable.

In this section I will first introduce the CEG and its semantics. I will then show in Section 2.3.2 how the CEG lets us read off these types of conditional independencies as well as additional conclusions directly from its topology, and I will also briefly discuss the causal extension of the CEG.

2.3.1 Review of CEGs

The CEG is a graphical model which is derived from a finite probability tree by merging the vertices in the tree together whose associated conditional probabilities are the same. As described in Section 2.1.1 its derivation from a tree is particularly advantageous when we want to describe the way a process unfolds and how the combination of different subsequent events leads to different conclusions. Starting at the root vertex and traversing a path, the tree describes the different events in a process or story. Each situation, s_i, in the tree displays a situation or state the individual may be in and its children vertices are the possible events that may follow from this situation. In cohort studies, such as those described in Chapter 1.2, a tree can hence naturally take us through a part of the individual's life or represent explicitly the way in which different factors affect an outcome. The CEG retains these features of the tree of describing a sequence of events within in its graph.

I again use the example of the CHDS to illustrate the derivation of the CEG.

Example 2. *Recall from Chapter 1.2.1 Table 1.1 that we have binary variables describing the social background, X_1, and the economic situation, X_2, as well as a variable describing the number of family life events, X_3, occurring over the five years of the offsprings's lives. As before the interest lies in the effect of these variables on the probability of a hospital admission, X_4 occurring over the five years. A possible event tree of this problem is given in Figure 2.3.*

The edges $e_{s_i k} \in E(F(s_i)), k = 1, \ldots, m_{s_i}$ of a floret in the tree are labelled by the particular events that occur after s_i is reached. Hence, each floret $F(s_i)$ can be associated with a random variable $X(s_i)$ taking values $\{x_1, x_2, \ldots, x_{m_{s_i}}\}$, where s_i has m_{s_i} children. For example, s_3 in Figure 2.3 is associated with the random variable $X(s_3) = X_3$ describing the number of life events given a high social background and a high economic situation, taking the values 'high' and 'low'. Consequently, the paths in the tree $\mathcal{T}$ of Figure 2.3 correspond to the set of all configurations of values that the set of variables $\mathbf{X} = (X_1, X_2, X_3, X_4)$ can take.

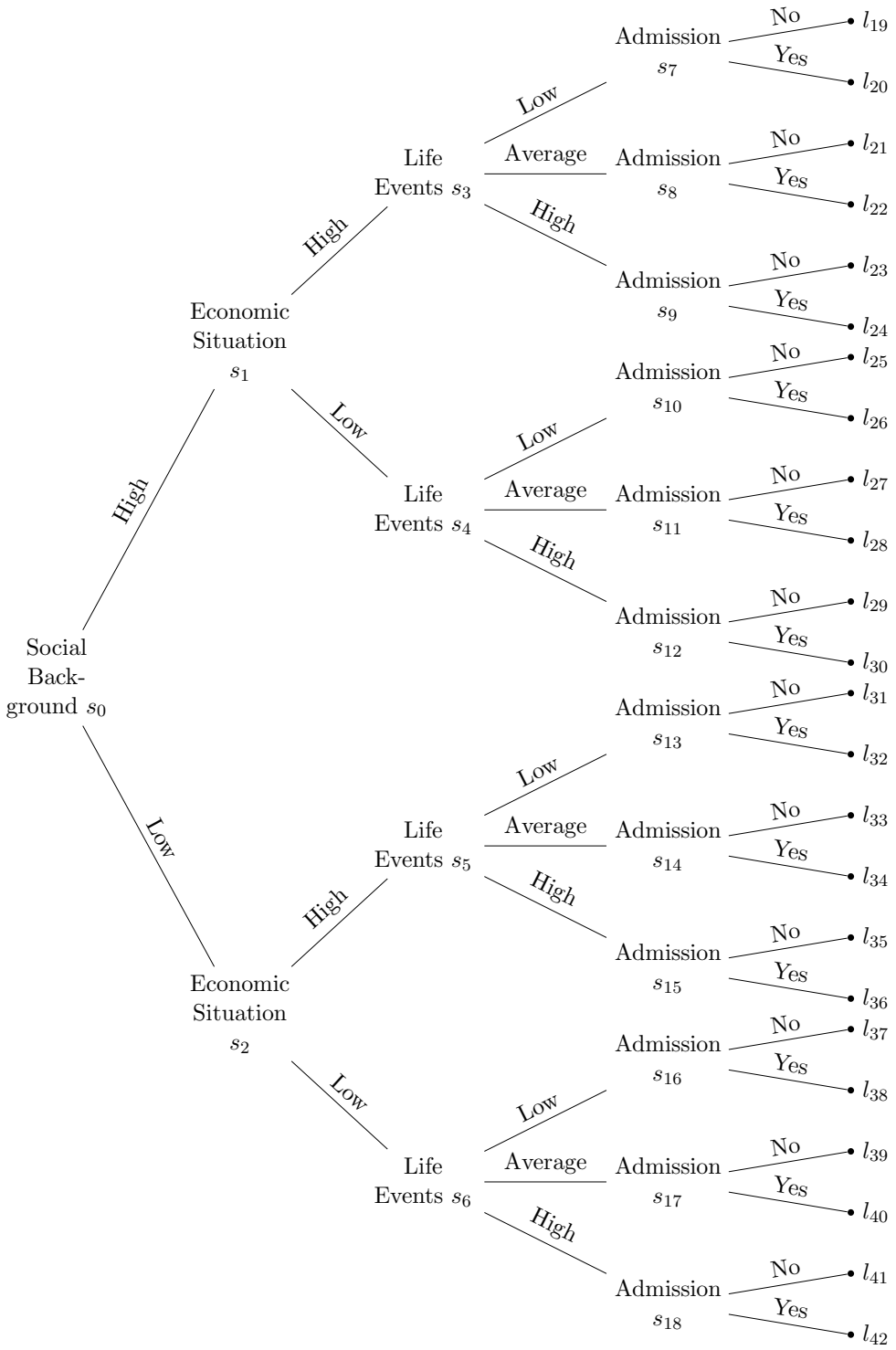


Figure 2.3: An event tree $\mathcal{T}$ for the CHDS example, with the variables taking the ordering: social background, economic situation, life events and hospital admission

Like the BN, an event tree is commonly elicited from a description of a problem by a client or domain expert. This can often be simpler than the elicitation of a BN, as the tree can explicitly express the described sequence of events from its root-to-leaf paths and depict the order in which these occur. Details are given for example in Smith [2010] who discusses the elicitation process and the advantages of using tree structures. This set-up further requires that an ordering of the variables first needs to be chosen when representing a problem as a specific tree. Often, a plausible order can be determined which is compatible with the temporal development of each individual within the study. In a cohort study this will often be the case, where a set of variables are measured throughout the individuals' lives, which will affect the outcome variable measured last. In some scenarios several orderings of the variables may be plausible. For example in the CHDS, putting the social background as our first variable is an obvious choice, as it is measured only at birth and hence cannot be affected by variables measured after birth. The hospital admissions variable is placed as the final variable in the tree, as we are interested in the effect that the other three variables have on it. In this example, I have further placed the economic situation before the life events, which suggests that the economic situation may affect the number of life events (e.g. change of job, financial problems). However, it is also plausible that the life events affect the economic situation. Therefore, in this case, we have two plausible orderings and I will discuss the effect of switching the economic situation and the life events in the probability tree in Chapter 3.3. I will further discuss in the same chapter the possibility of relaxing the restriction of an ordering by searching over several plausible tree structures. For now, however, I assume we have a single tree representation of a given problem from which we will later derive the CEG.

To define a CEG I will start with a finite probability tree and then introduce the concepts of **stages** and **positions**, which group the vertices of the tree together when the probabilities on their florets are the same. A probability tree is, by definition, an event tree with a vector of conditional probabilities (CPV) associated with each floret, $F(s_i)$, which is written as

$$\pi_{s_i} = (\pi_{s_i 1}, \pi_{s_i 2}, .., \pi_{s_i m_{s_i}}), \tag{2.4}$$

where $\pi_{s_i k} = P(e_{s_i k}|s_i)$ is the probability that an individual transitions from s_i along the kth edge $e_{s_i k}$ and $\sum_{k=1}^{m_i} \pi_{s_i k} = 1$. The CPVs then describe the distribution of the random variable, $X(s_i)$, associated with floret $F(s_i)$. For example, floret $F(s_3)$ is associated with the random variable describing the number of family life events

conditional on a high social background and economic situation, taking the three values 'low', 'average', 'high' with distribution $\pi_{s_3} = (\pi_{s_3 1}, \pi_{s_3 2}, \pi_{s_3 3})$.

Definition 18. *We say two situations s_i and s_j are in the same **stage**, u, if and only if*

1. *the topology of the florets $F(s_i)$ and $F(s_j)$ is the same, i.e. there exists a bijection Φ_{ij} between $E(\mathcal{F}(s_i))$ and $E(\mathcal{F}(s_j))$, where $\Phi_{ij}(e_{s_i k}) = e_{s_j k}$ and $e_{s_i k}$ and $e_{s_j k}$ describe the same unfolding event, and*

2. *$\pi_{s_i} = \pi_{s_j}$, i.e. the probability distributions associated with the florets is the same.*

If s_i and s_j are in the same stage then we assign their pairs of edges $\left(e_{s_i k}, e_{s_j k}\right)$ the same colour. (See for example Figure 2.4.)

When there is only a single situation in a stage, then this stage is called **trivial**. The situations of the tree can hence be partitioned into stages, associated with a set of bijections $\{\Phi_{ij} : s_i, s_j \in S(\mathcal{T})\}$ and we denote the set of stages by $U(\mathcal{T})$. Also, given a stage $u \in U(\mathcal{T})$, $\pi_u = (\pi_{u1}, \pi_{u2}, .., \pi_{um_u})$ is the CPV of stage u. Note that by definition all situations in u have the same number of emanating edges, m_u. Given a stage partition $U(\mathcal{T})$ of the situations in $\mathcal{T}$ we can then define a **staged tree** version of $\mathcal{T}$ as follows:

Definition 19. *A **staged tree** version of $\mathcal{T}$ is a tree with coloured edges, where for every non-trivial stage $u \in U(\mathcal{T})$ and $s_i, s_j \in u$, corresponding edges under Φ_{ij} are assigned the same colour.*

In previous publications not only the edges but also the vertices of the staged tree are coloured. In this thesis I will colour only the edges of the situations that are in the same stage, as in small examples this gives the clearest presentation. However, in Chapters 4 and 5, I will instead only colour the situations that are in the same stage itself, where corresponding edges are consequently identified only by their labelling.

If $U(\mathcal{T})$ is the trivial partition, such that every situation is in a different stage, then the staged tree is uncoloured and contains no additional information about the process that is not already contained in $\mathcal{T}$. However, Smith and Anderson [2008] give numerous examples of trees where the stage partition of a proposed model is non-trivial. For example, any discrete BN has an equivalent representation in terms of a stage partition, which is only trivial when we have a complete BN structure, as will be shown in Chapter 2.4.

A finer partition of the vertices in the tree can be given by the position partition, where a position is defined as follows:

Definition 20. *Two situations s_i, s_j in the same stage are also in the same* ***position*** *w if and only if*

1. *the topology of the subtrees $T(s_i)$ and $T(s_j)$ is the same, i.e. there exists a bijection Ψ_{ij} between $E(\mathcal{T}(s_i))$ and $E(\mathcal{T}(s_j))$ and all corresponding edges describe the same unfolding events, and*

2. *the probability distributions on corresponding florets in the subtrees are the same.*

The definition requires that for two situations to be in the same position there must not only be a map between the edge sets $E(\mathcal{T}(s_i))$ and $E(\mathcal{T}(s_j))$ of the two subtrees but also the colours of any edges under this map must correspond in the associated staged tree. When s_i, s_j are a distance of one edge from a leaf node then $\mathcal{T}(s_i) = \mathcal{F}(s_i)$ and $\mathcal{T}(s_j) = \mathcal{F}(s_j)$ and so they will be in the same position if and only if they are in the same stage. But if these situations are further from a leaf, not only do these two situations need to be in the same stage but also all their children must have a parallel child in the same stage, and so on. Therefore a potentially finer partition can be obtained through the set of positions, denoted by $W(\mathcal{T})$. As above each position w made up of several situations has an associated CPV given by

$$\pi_w = (\pi_{w1}, \pi_{w2}, .., \pi_{wm_w}),$$

where all situations in w have the same number m_w of emanating edges. The CEG $\mathcal{C}$ of a finite staged tree $\mathcal{T}$ is then the staged tree collapsed over its positions, where the positions form the vertices of the graph and the set of leaf nodes are collected in a single position called w_∞. The formal definition of the CEG derived from a finite tree similar to Smith and Anderson [2008] is given below. I will extend this definition to infinite trees in Chapter 5.

Definition 21. *A* ***Chain Event Graph*** *(CEG) $\mathcal{C} = (V(\mathcal{C}), E(\mathcal{C}))$ of a staged tree $\mathcal{T}$ has vertex set $V(\mathcal{C}) = W(\mathcal{T}) \cup w_\infty$, the set of all positions of the tree $\mathcal{T}$ and the position of leaf nodes. Emanating edges from a position $w_i \in W(\mathcal{T})$ are constructed as follows: Choose a single representative situation $s(w_i) \in S(\mathcal{T})$. Then there is an edge from w_i to a position $w_j \in V(\mathcal{C})$ for each child $v_j \in ch(s(w_i)), v_j \in w_j$ in the tree $\mathcal{T}$. When two positions are in the same stage then they are connected by an undirected dashed line and their edges are coloured according to their colouring in the staged tree.*

I will henceforth denote the stage partition of a CEG $\mathcal{C}$ by $U(\mathcal{C})$ and the position partition by $W(\mathcal{C})$.

Definition 22. *Say the CEG is* ***simple*** *whenever $U(\mathcal{C}) = W(\mathcal{C})$, i.e. the stages and positions coincide and hence the CEG is uncoloured.*

Similarly to eliciting the dependence structure of the variables in a BN, the stage structure of a CEG can be elicited from a client or domain expert. Starting with an event tree, the client is queried on whether future developments in the story are likely to be the same given that certain events have occurred. Because of this the qualitative structure of the CEG can be determined prior to eliciting specific probabilities (see Smith and Anderson [2008] for more details).

I will now illustrate the above definitions and the conclusions we can draw from the CEG on the example for the CHDS.

Example 3. *Assume the following stage partition on the tree in Figure 2.3 of the CHDS example:*

$$u_0 = \{s_0\}, u_1 = \{s_1\}, u_2 = \{s_2\}, u_3 = \{s_3, s_4, s_5\}, u_4 = \{s_6\}, u_5 = \{s_7, s_{10}\},$$
$$u_6 = \{s_8, s_{11}, s_{13}, s_{14}, s_{16}\}, u_7 = \{s_9, s_{12}, s_{15}, s_{17}, s_{18}\},$$

where the resulting staged tree with coloured edges is given in Figure 2.4. In this example the stage and position partition do not coincide completely, as s_5 is in the same stage as s_3 and s_4, however, the probability distributions on the subtree $\mathcal{T}(s_5)$ do not coincide with $\mathcal{T}(s_3)$ and $\mathcal{T}(s_4)$. This is evident from the colouring of the edges emanating from situation s_{13} which does not match the colouring of s_7 and s_{10}. Therefore s_5 is not in the same position as s_3 and s_4 and we hence have the slightly finer position partition:

$$w_0 = \{s_0\}, w_1 = \{s_1\}, w_2 = \{s_2\}, w_3 = \{s_3, s_4\}, w_4 = \{s_5\}, w_5 = \{s_6\}, w_6 = \{s_7, s_{10}\},$$
$$w_7 = \{s_8, s_{11}, s_{13}, s_{14}, s_{16}\}, w_8 = \{s_9, s_{12}, s_{15}, s_{17}, s_{18}\}, w_\infty = \{l_{19}, l_{20}, \ldots, l_{42}\}.$$

Collapsing the staged tree over its positions then gives the CEG in Figure 2.5.

We can then give a detailed account of the effect a combination of covariates has on the hospital admissions directly from the topology of the CEG, where the conclusions drawn from the graph go beyond the usual conditional independence statements obtainable from a BN:

- *The social background appears to have an effect on the economic situation ($w_0 \to w_1, w_0 \to w_2$).*

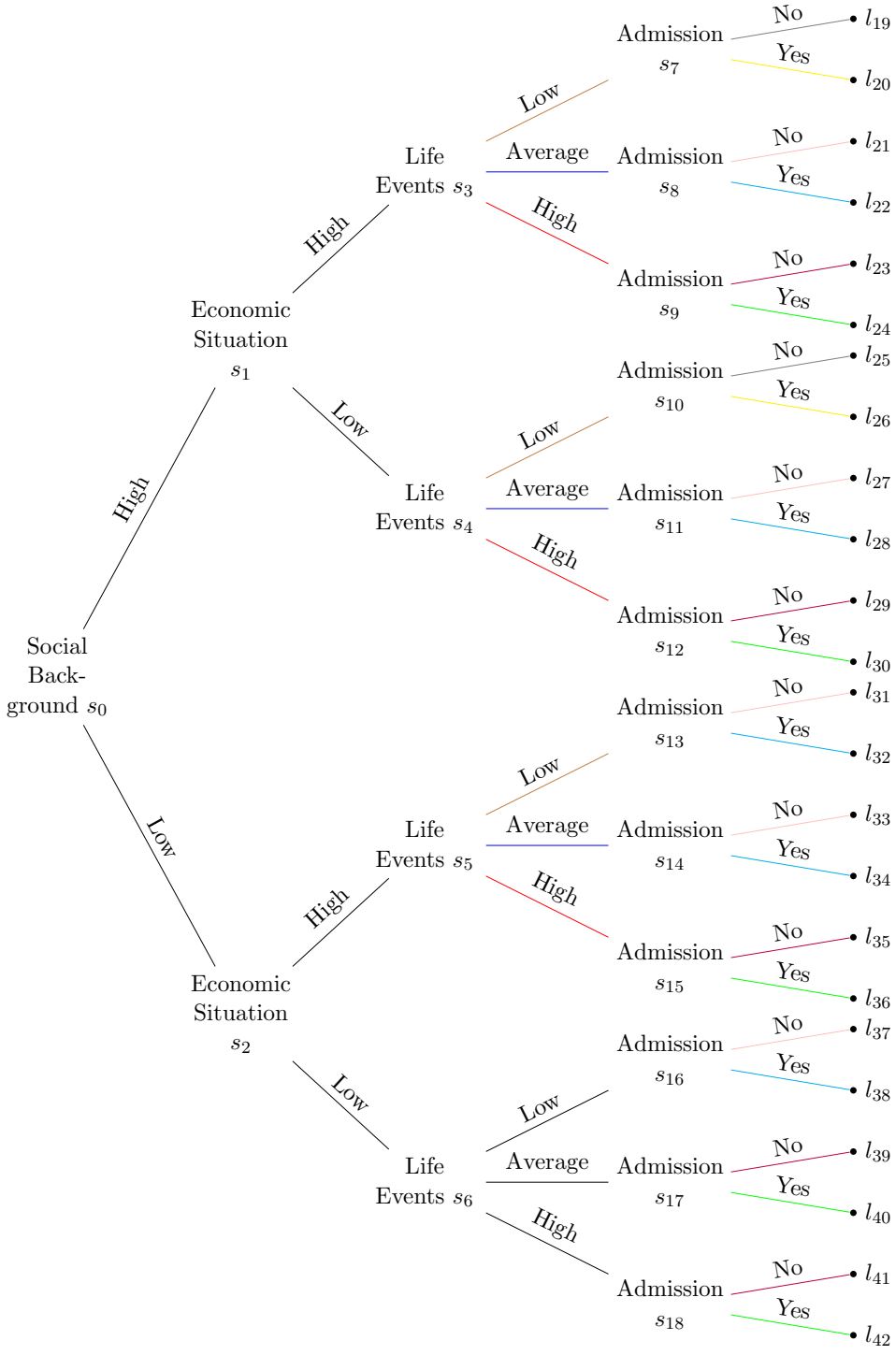


Figure 2.4: A staged tree version of the event tree $\mathcal{T}$ for the CHDS example, with the variables taking the ordering: social background, economic situation, life events and hospital admission

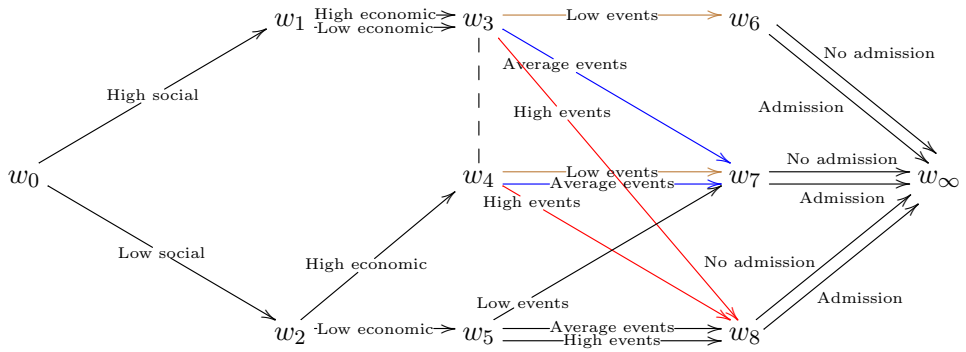


Figure 2.5: The CEG of the staged tree version of $\mathcal{T}$ on the variables social background, economic situation, life events and hospital admissions of the CHDS example

- *The economic situation seems to have no effect on the number of life events for families from a higher social background ($w_1 \to w_3$). However, in a family from a lower social background the economic situation seems to affect the number of life events that occur ($w_2 \to w_4, w_2 \to w_5$).*
- *Offspring from a family of high social background and a low number of life events are in one position, independent of the economic situation ($w_3 \to w_6$).*
- *Offspring from socially advantaged families with an average number of life events are in the same position as offspring from socially disadvantaged families with a high economic situation and a low or average number of life events, as are offspring from a low economic situation with a low number of life events ($w_3 \to w_7, w_4 \to w_7, w_5 \to w_7$).*
- *All individuals with a high number of life events are in the same position irrespective of their social or economic background. Further, an individual from a low social and economic background with only an average number of life events is also in this position ($w_3 \to w_8, w_4 \to w_8, w_5 \to w_8$).*

It is now also possible to observe that the CEG does not lose any information over the tree, as any path in the tree can be identified in the CEG. For example, an individual with a high social background, a high economic situation, a high number of life events and no hospital admission goes along the edges $e(w_0, w_1)$, $e_{w_1 1}$ labelled 'high economic', $e(w_3, w_8)$, $e_{w_3 1}$ labelled 'no admission'. However, unlike a tree, the CEG can have 'double edges' in the graph, meaning that we have two edges (or more) going from a single parent position into the same child position. (Compare, for example, Figure 2.5 $e(w_1, w_3)$.) Because of this we define a path through the CEG by its edges, rather than its positions. Further, by the definition of a stage the CEG satisfies the Markov property in the sense that the probability of going along a particular edge depends only on the current stage we are at and is independent of the

path through which the stage was reached. Let $\epsilon = (e_{w_0k_0}, e_{w_{i_1}k_1}, e_{w_{i_2}k_2}, \ldots, e_{w_{i_n}k_n})$ be a path starting at w_0 and going along $n+1$ edges ending at w_∞, where $e_{w_{i_n}k_n}$ leads to w_∞. Then,

$$\begin{aligned}
p(\epsilon|\mathcal{C}) &= p(e_{w_0k_0}, e_{w_{i_1}k_1}, e_{w_{i_2}k_2}, \ldots, e_{w_{i_n}k_n}) \\
&= p(e_{w_0k_0})p(e_{w_{i_1}k_1}|e_{w_0k_0}) \ldots p(e_{w_{i_n}k_n}|e_{w_0k_0}, e_{w_{i_1}k_1}, \ldots, e_{w_{i_{n-1}}k_{n-1}}) \\
&= p(e_{w_0k_0}|w_0)p(e_{w_{i_1}k_1}|w_{i_1}) \ldots p(e_{w_{i_n}k_n}|w_{i_n}) \text{ by the definition of a position} \\
&= p(e_{w_0k_0}) \prod_{a=1}^{n} p(e_{w_{i_a}k_a}|w_{i_a}) \\
&= \pi_{w_0k_0} \prod_{a=1}^{n} \pi_{w_{i_a}k_a}. \qquad (2.5)
\end{aligned}$$

Hence, P admits a **factorisation relative to** $\mathcal{C}$, similar to the factorisation for BNs in Equation 2.2. Note, that $\pi_{w_{i_a}k_a} = \pi_{u_{i_a}k_a}$, where $w_{i_a} \in u_{i_a}$ and so each path can be written as a product of conditional probabilities associated with the stages in $U(\mathcal{C})$. Hence,

$$p(\epsilon|\mathcal{C}) = \pi_{u_0k_0} \prod_{a=1}^{n} \pi_{u_{i_a}k_a}, \qquad (2.6)$$

where u_{i_a}, $a = 1, \ldots, n$ are not necessarily distinct as a stage may consist of several positions.

2.3.2 Conditional Independency in CEGs

As for BNs, conditional independence statements can be read off directly from the topology of the CEG. Thwaites and Smith [2011] attempt to characterise all conditional independence statements that can be read from the CEG, including an analogue of the d-separation theorem for BNs. For this thesis, however, only the dependence structures discussed in Smith and Anderson [2008] are needed. In contrast to the BN, the conditional independence statements that can be read from the CEG are context-specific conditional independencies (Definition 2.3) or event-specific conditional independencies [Smith and Anderson, 2008], where we condition on a variable taking a particular value or a sequence of possible events having happened.

Smith and Anderson [2008] define these types of conditional independencies on a CEG as follows: By Thwaites et al. [2010], let $Y(w)$ be the variable identified with the set of paths from w_0 to w. Similarly, $Z(w)$ is the variable associated with the set of paths from w to w_∞. Finally, let $\mathcal{E}(w)$ represent the event that the

individual passes through the position w, which is associated with all paths from w_0 to w_∞ which go through w. Smith and Anderson [2008] then prove that we can read off conditional independence statements of the form

$$Y(w) \perp\!\!\!\perp Z(w) \,|\, \mathcal{E}(w). \tag{2.7}$$

So, given that an individual reaches a position w, the path through which w was reached is independent of all future developments from this position. Similarly, let $Y(u)$ be the variable whose state space is the set of all paths from w_0 to $w, w \in u$, $\mathcal{E}(u)$ is the event of passing through $w \in u$ and finally, as defined in Section 2.3, $X(u)$ is the random variable associated with the floret $F(u)$. By Smith and Anderson [2008] we can then also read off conditional independencies of the form

$$Y(u) \perp\!\!\!\perp X(u) \,|\, \mathcal{E}(u). \tag{2.8}$$

So, given an individual reaches a position $w \in u$, the path taken to reach u is independent of the immediate future. I will show on the CHDS example below that, apart from simple conditional independencies as defined in Definition 2.3, more complex deductions can be drawn from the CEG through these two statements, where different sequences of events lead into the same position.

Example 4. *In the CHDS example w_1 and w_2 are reached through a high or low social background respectively. Therefore, by (2.7), given w_1 and w_2, the paths through which these positions are reached is independent of the future developments. This is, however, a trivial conditional independency as w_1 and w_2 are reached by two unique paths associated with the two values taken by the social background. Similarly, w_4 and w_5 are reached by two unique paths describing a low social background and a high or a low economic situation respectively, leading again to a trivial context-specific conditional independence statement. However, w_3 is reached by two paths, namely by an individual from a high social background and with either a high or a low economic situation. Then by (2.7), given w_3 the paths through which w_3 is reached, namely by either a low or a high economic situation, are independent of all future developments, namely the life events and the hospital admission. This is equivalent to the context-specific conditional independency that the life events are independent of the economic situation given a high social background. Similarly, w_6 is reached through a high social background and a low number of life events independent of the economic situation. Hence, given a high social background and a low number of live events, hospital admission is independent of the economic situation. Further to these context-specific independencies, by inserting w_7 into (2.7), we can conclude*

that, given w_7 is reached, the hospital admission is independent of the five paths through which an individual can get to w_7. These five paths are made up of particular configurations of the three covariates or, equivalently, describe the different events that happen leading to w_7. Finally, w_8 can be reached via a high number of life events or by a low social background, low economic situation and an average number of life events, and this is independent of a hospital admission given w_8 has been reached. Also, note that, as w_3 and w_4 are in the same stage, (2.8) can be applied in this case. This leads to the deduction that the life events are in fact not only independent of the economic situation given a high social background but independent of the paths through which w_3 or w_4 are reached.

Like the d-separation theorem for BNs, these types of conditional independence statements derived from the CEG can be fed back to the client to check that the elicited CEG structure correctly formalises his beliefs.

2.3.3 Causality in CEGs

Just as the BN, the CEG admits a causal extension [Thwaites et al., 2010; Thwaites, 2013] by linking it to causal hypotheses about the likely effect of an intervention. As pointed out by Shafer [1996], causal assumptions are most easily inferred from tree-like structures as these naturally respect not only an ordering according to time but also an implicit causal ordering. Further, in contrast to the BN we are not restricted to intervening on a particular variable. Instead it is possible to intervene at a specific position and hence the CEG allows for context-specific interventions. For example, in the CHDS study, a policy maker may be interested in the effect of an intervention which only gives families from a low social background financial aid and this type of intervention can be simply represented within a CEG. The simplest form of intervention, then forces an individual that reaches a position w^* along a particular edge, $e_{w^*k^*}$, say. Under such an intervention $p(e^*_{w^*k}|w^*)$ is set equal to 1 and hence $p(e_{w^*k}|w) = 0, \forall k \neq k^*$. Consequently, the CPV of w^*, π_{w^*}, has one entry equal to 1 and zeros everywhere else. All other edge probabilities are assumed to remain unchanged. As in Pearl [2000] and related work, an intervention which forces individuals along a particular edge is therefore assumed to have the same affect on its children as if we had observed an individual going along this edge.

A CEG on which an intervention is imposed is called a **manipulated CEG** [Thwaites et al., 2010] and it is drawn by deleting all paths emanating from w^* that do not go along the required edge, $e_{w^*k^*}$.

Definition 23. *We say a CEG C is a **Causal CEG** when, under manipulation at a*

position w^*, the CEG is a manipulated CEG C^* with $\pi_{w^*k^*} = 1$, $\pi_{w^*k^*} = 0, \forall k \neq k^*$ and all other $\pi_w, w \in W$ are as in C.

I will illustrate a simple intervention on the CHDS example below:

Example 5 (CHDS)**.** *Consider again the CHDS with the CEG given in Figure 2.5. Assume that a policy maker wants to know the effect of giving individuals from a low social background financial aid on hospital admission. Then, if we are prepared to read the graph causally, the effect of giving only families from a low social background financial aid corresponds to forcing all individuals that reach w_2 along the edge $e(w_2, w_4)$. The manipulated CEG is given in Figure 2.6, where the edges $e(w_2, w_5)$, as well as w_5 and its emanating edges, have been removed. The effect in terms of reducing the probability of a hospital admission is calculated in Chapter 3.3.3.*

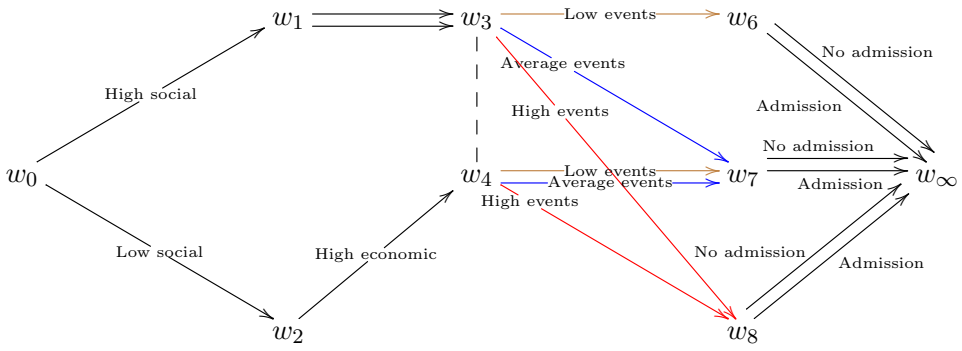


Figure 2.6: The CEG for the CHDS example from Figure 2.5 manipulated such that individuals from a low social background are given financial aid

I have shown in this section that, because of its graphical derivation, the CEG inherits many of the benefits of a BN: I have demonstrated here that a CEG can give a more detailed representation of a problem than a BN, such that detailed inference on the way in which a combination of variables affects another variable can be drawn from the graph and also that, similar to the BN, conditional independencies can be read off and a causal CEG can be defined. In Chapter 3.1 I will further review how we can carry out estimation and model selection on CEGs according to Freeman and Smith [2011a], again showing the analogy to the BN. Further developments of the CEG over the past years include a propagation algorithm [Thwaites et al., 2008] based on the junction tree algorithm for BNs, separation theorems, developed in Thwaites and Smith [2011] and more detailed causal developments, including analogues to Pearl's backdoor and frontdoor theorems [Thwaites, 2013]. In this

thesis, however, I will focus primarily on model selection and inference that can be drawn from a selected model, as described in this section.

2.4 Writing a BN as a CEG

In this section I will demonstrate, using the CHDS example, that any discrete BN can be written as a staged tree and therefore as a CEG. I will show that the BN imposes certain restrictions on the stage structure of the associated tree, leading to an entirely symmetric CEG. This supports the previous suggestions that one of the CEG's main advantages lies in the possibility of representing asymmetries within the dependence structure of the variables of a problem that the BN cannot capture. A technical proof that every discrete BN can be written as a CEG is given in Smith and Anderson [2008].

To write a BN as a CEG the graph is first translated into a tree. To do so the variables in the BN are ordered such that parent variables appear before their children. There is always such an ordering due to the acyclicity of the corresponding graph. However, the ordering is not necessarily unique, as explained in Section 2.2.

Example 6. *Consider a possible BN of the CHDS Example to be the graph given in Figure 2.7 with conditional independence statements* $X_3 \perp\!\!\!\perp X_2 \mid X_1$ *and* $X_4 \perp\!\!\!\perp X_1, X_2 | X_3$. *Here, we choose the straightforward ordering* $\boldsymbol{X} = (X_1, X_2, X_3, X_4)$ *with the corresponding tree given already in Figure 2.3.*

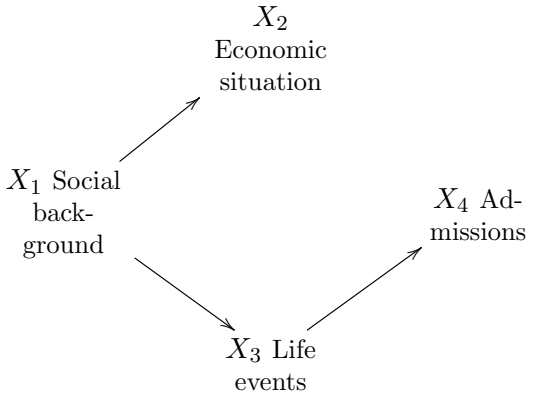


Figure 2.7: A possible BN structure for the social background, economic situation, life events and hospital admission of the CHDS example

The conditional independence statements of a BN can then be uniquely represented by defining stages on the tree. Recall that a conditional independency in

a BN is of the form

$$X_i \perp\!\!\!\perp pr(X_i) \backslash pa(X_i) \mid pa(X_i).$$

To represent this in terms of stages in the tree, all the vertices whose floret describes X_i and whose previous events differ only in $pr(X_i)\backslash pa(X_i)$ are put into the same stage. So, two situations, s_i and s_j, whose florets $F(s_i)$ and $F(s_j)$ are associated with the variable X_i, are in the same stage whenever their set of parent variables $pa(X_i)$ takes the same values on the corresponding paths leading to s_i and s_j in the tree. Hence, to move from a BN to a staged tree we can go through each of the conditional independence statements $X_i \perp\!\!\!\perp pr(X_i)\backslash pa(X_i) \mid pa(X_i)$ for $i = 2, 3, ..$ determining at each step the florets describing X_i and colouring them accordingly in the tree to display the different stages. Consequently, a BN with no conditional independencies corresponds to a CEG where each situation is in a separate stage.

Example 7. *In the CHDS example the corresponding staged tree of the BN $\mathcal{B}$ therefore has stages*

$$u_0 = \{s_0\}, u_1 = \{s_1\}, u_2 = \{s_2\}, u_3 = \{s_3, s_4\}, u_4 = \{s_5, s_6\},$$
$$u_5 = \{s_7, s_{10}, s_{13}, s_{16}\}, u_6 = \{s_8, s_{11}, s_{14}, s_{17}\}, u_7 = \{s_9, s_{12}, s_{15}, s_{18}\},$$

and is given in Figure 2.8. As an edge from X_1 to X_2 exists, the two situations s_1 and s_2 are in separate stages, u_1 and u_2. The conditional independence statement $X_3 \perp\!\!\!\perp X_2 | X_1$ puts s_3 into a stage with s_4, called u_3, and s_5 into a different stage with s_6, called u_4. Similarly, the conditional independence statement $X_4 \perp\!\!\!\perp X_1, X_2 \mid X_3$ is described by the stages u_5, u_6 and u_7, where all situations reached by the edge describing a low number of life events, namely s_7, s_{10}, s_{13} and s_{16} are in the same stage u_5. The same applies to situations reached through an edge describing an average number of live events and a high number of life events respectively, giving stages u_6 and u_7.

To summarise, the restrictions imposed by the BN can be defined within the associated staged tree by the two following rules which lead to a symmetric CEG:

1. Let two situations, s_i and s_j whose florets describe a variable X_i be in the same stage. Their paths will differ in some of the values taken by the variables in $pr(X_i)\backslash pa(X_i)$. Then all situations whose florets describe the same variable X_i are also in this stage whenever their associated paths differ by the same variables, $pr(X_i)\backslash pa(X_i)$.

2. Let two situations whose florets describe a variable X_i be in different stages. Their paths will differ by some of the values taken by the variables in $pa(X_i)$.

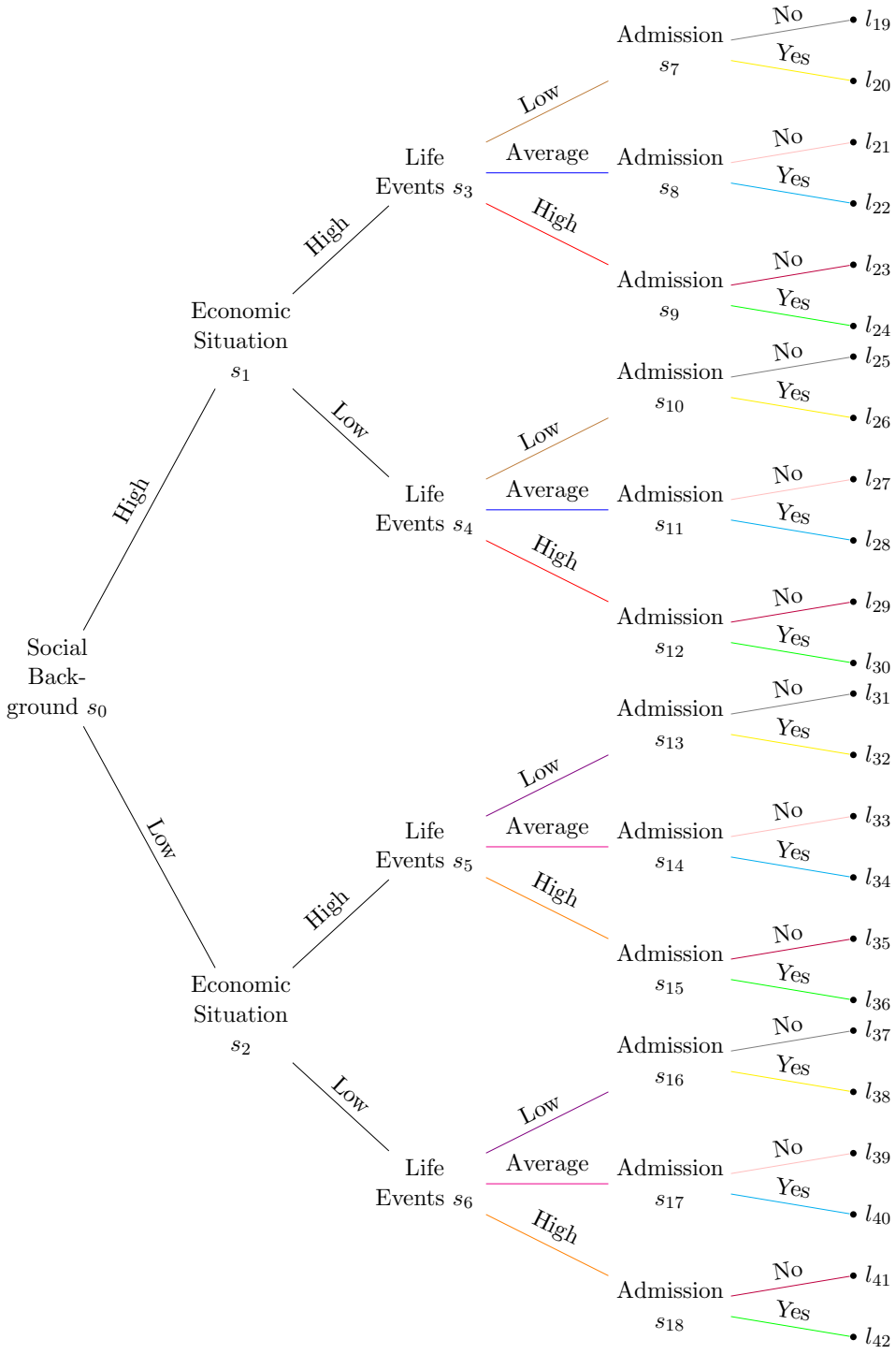


Figure 2.8: The staged tree version of the BN of the CHDS example from Figure 2.7

Then all other situations whose paths differ by the same variables are also in different stages.

From the staged tree representation of a BN the corresponding CEG can then be constructed, as defined in Section 2.3.

Example 8. *From the staged tree (Figure 2.8) we obtain the CEG given in Figure 2.9. Note that this CEG is uncoloured as the stages and positions coincide. From*

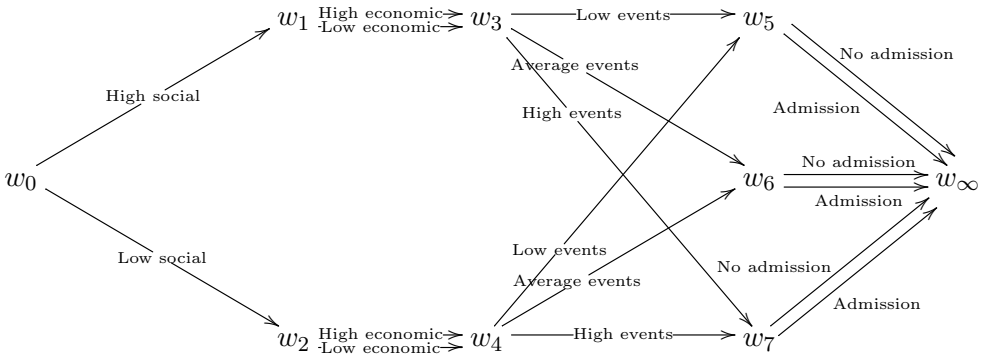


Figure 2.9: The CEG of the staged tree version of the BN given in Figure 2.7

Equations 2.7 and 2.8 the conditional independence statements of the original BN can be read from the CEG. For example, inserting w_3 and w_4 into 2.7 gives

$$Y(w_3) \perp\!\!\!\perp Z(w_3) \,|\, \mathcal{E}(w_3) \text{ and } Y(w_4) \perp\!\!\!\perp Z(w_4) \,|\, \mathcal{E}(w_4),$$

which says that the paths through which w_3 and w_4 have been reached is independent of the future developments, i.e. independent of the life events and the hospital admissions. Both positions can be reached via two paths, one describing a low economic situation and the other a high economic situation. Combining these two conditional independence statements of the CEG, we can hence conclude from the CEG that the economic situation is independent of the life events and the hospital admission. Similarly, the conditional independence statements on w_5, w_6 and w_7 can be combined to deduce that, given the life events, the hospital admissions are independent of the social background and the economic situation.

Figure 2.9 shows that the CEG of a BN is entirely symmetric. Consider also again the CEG given in Figure 2.5. This is not symmetric as, for example, the edges $e(w_1, w_3)$ in comparison with the edge $e(w_2, w_4)$ and $e(w_2, w_5)$ do not follow the same pattern and therefore, by the rules given above, this CEG structure is not a BN.

2.5 The Ordinal CEG

I have shown in the previous sections that CEGs are a particularly useful tool to identify and visualise the different ways in which certain combinations of covariates affect a variable of interest. In many medical applications this variable is binary, such as survival of a patient or the onset of a disease. Of course, many examples in other application areas, such as social sciences, also look at binary outcome variables. However, in this thesis the examples are restricted to medical applications. In this section I will introduce a new graphical development of the CEG, the Ordinal CEG, which provides an enhanced graphical representation of the standard CEG for problems with a binary outcome variable, by imposing an ordering on the positions of the graph. A possible extension of the Ordinal CEG to problems with an outcome variable with more than two categories is briefly discussed at the end of the section.

Assume we have a tree describing a problem on p variables $X_1, X_2, \ldots, X_p$. The situations in the tree can then be partitioned into vertex subsets, such that each subset consists of those vertices whose emanating edges describe the same succeeding events or, equivalently, whose associated florets describe the same variable. Denote these vertex subsets by V_{X_i}, where X_i, $i = 1, \ldots, p$ is the variable described by the floret $F(s_j), s_j \in V_{X_i}$. Then, by the definition of a stage (Definition 18), all situations in a vertex subset may be merged into stages from which a CEG structure can be derived.

Example 9. *Consider again the tree in Figure 2.3 of the CHDS example with variables X_1 = social background, X_2 = economic situation, X_3 = life events and X_4 = hospital admissions. We then have the following vertex subsets:*

$$V_{X_1} = \{s_0\}, V_{X_2} = \{s_1, s_2\}, V_{X_3} = \{s_3, s_4, s_5, s_6\},$$
$$V_{X_4} = \{s_7, s_8, s_9, s_{10}, s_{11}, s_{12}, s_{13}, s_{14}, s_{15}, s_{16}, s_{17}, s_{18}\}.$$

Given a CEG structure the vertex subsets $V_{X_i}, i = 1, \ldots, p$ can alternatively be defined in terms of the positions in the CEG. Hence, given a problem on p variables, $X_1, \ldots, X_p$, and an associated CEG structure, $\mathcal{C}$, the set of positions of $\mathcal{C}$ can be partitioned into p vertex subsets $V_{X_1}, \ldots V_{X_p}$. Note that the vertex subset of the first variable in the ordering of the tree will always only consist of w_0.

Example 10. *For the CHDS example assume again the CEG given in Figure 2.5. The positions can then be partitioned into the four vertex subsets as follows.*

$$V_{X_1} = \{w_0\}, V_{X_2} = \{w_1, w_2\}, V_{X_3} = \{w_3, w_4, w_5\}, V_{X_4} = \{w_6, w_7, w_8\}.$$

The graph of the Ordinal CEG then vertically aligns the positions within a vertex subset to enhance the graphical representation of the problem according to X_p. I will give the formal definition of this below:

Definition 24. *[Barclay et al., 2013b] Let $\mathcal{C}$ represent a problem on p variables with a binary outcome variable, X_p, described by the final florets in $\mathcal{C}$. We say that the CEG, $\mathcal{C}$, is an* ***Ordinal CEG with respect to*** *X_p when the positions in each vertex subset V_{X_i}, associated with variable X_i, are vertically aligned in descending order with respect to the probability $P(X_p = 0|\mathcal{C})$.*

In the CHDS example the variable of interest, X_p, describes whether the individual is admitted to hospital or not. So, if X_p describes hospital admission, with $X_p = 0$ meaning that the individual is not admitted to hospital, then the ordering occurs such that the position with the highest probability of no admission is at the top of the graph. This allows us to read off directly from the graph how the different combinations of covariates affect the admissions probability: the higher up the graph a combination takes us the better the effect on the outcome variable. The Ordinal CEG further retains the natural time ordering of its associated tree by listing the vertex subsets from V_{X_1} to V_{X_p} from left to right in the graph. In the examples considered so far, the root-to-leaf paths in the tree are always of the same length, where each path describes a particular configuration of the variables of the problem. In this case, each vertex subset, V_{X_i}, of the Ordinal CEG defines a cut in the graph and we can look at each cut-set associated with V_{X_i} to discuss the different dependence structures and the effect on the outcome variable, X_p, at each point in time.

Example 11. *Let us assume that the CEG in Figure 2.5 is an Ordinal CEG. We would then deduce from the graph that the position w_6 is associated with the lowest admission probability, followed by w_7 and then w_8. Similarly, we would read from the topology of the Ordinal CEG that the hospital admission for an individual from a high social background (w_1) is lower than that for an individual with a low social background (w_2). Further, we believe that an admission is less probable when the individual is from a low social background but has a high economic situation than when he has a low economic situation (Compare w_4 and w_5). I will return to the Ordinal CEG of the CHDS example in Chapter 3.3.3 after discussing Bayesian model selection techniques for CEGs and the calculation of the posterior CPVs given available data.*

In this thesis all outcome variables of a given problem are binary and so all CEGs can be represented as Ordinal CEGs. It may, however, also be possible to use

the Ordinal CEGs when we have an outcome variable with more than two categories. One option would be to choose simply one outcome that is of greatest interest and order the positions according to the probability of this outcome. Alternatively, we could draw several CEG structures, one for each possible category.

In higher dimensional problems, the full CEG structure can become complicated when we have a large number of positions. In an Ordinal CEG this means that we have a large number of positions in each vertex subset. To improve the visual aspect of the Ordinal CEG, I have further developed the concept of a **Reduced Ordinal CEG**. Here, interest is restricted entirely to the combined effect of the covariates on the outcome variable by considering only the positions in the final subset, V_{X_p}, and re-expressing the paths leading to these positions in terms of new variables. Hence, in the Reduced Ordinal CEG, only the final subset of positions remains, while V_{X_1} up to $V_{X_{p-1}}$ are redefined to describe simply the intermediate steps leading to the final subset of positions. These intermediate steps are represented as intermediate positions which are denoted by w^I. In the CHDS example considered so far the CEG structures are simple and accessible and hence a reduction is not necessary. In Chapter 4 two slightly larger examples based on the UKCP study are presented and here the simplification to the Reduced Ordinal CEG is extremely useful. I will therefore postpone a more detailed description of the Reduced Ordinal CEG until that Chapter.

2.6 Alternative Graphical Models

I have emphasised in the previous sections that one of the drawbacks of the BN is that it does not allow for context-specific dependence structures. To take these features into account extensions to the BN have been proposed, mostly in the form of tables or tree-like structures which are added to the graph, leading to the **context-specific Bayesian Network** [Boutilier et al., 1996; Poole and Zhang, 2003; Friedman and Goldszmidt, 1998]. Recall again the CHDS example with the CEG structure given in Figure 2.5. Then, by the definition of a contextual independency (Definition 2.3), the conclusion that the economic situation is independent of the life events and the hospital admission given a high social background is a contextual independency and this type of dependence structure can be represented by a context-specific BN. However, these types of models focus primarily on efficient propagation and learning, where contextual independencies are represented either through a table of CPVs or by a separate tree structure, one for each variable. Therefore, the benefit of the BN's expressiveness for the client is quickly lost. Simi-

lar extensions have also been proposed in Smith et al. [1993] for influence diagrams, who use coalesced decision trees to represent the conditional probability distribution of a variable given its parents.

Apart from its graphical benefits, the CEG further enables us to draw additional conclusions from its graph. For example, a context-specific BN can only demonstrate that, given a high number of life events, the hospital admission's variable is independent of the social background and the economic situation. However, it fails to represent that a low social background, low economic situation and average number of life events lead to the same conclusion. Hence the CEG not only allows for the types of context-specific independencies of Definition 2.3 but can also illustrate further conditional independencies based on the paths leading into and from a position, as described in detail in Section 2.3.2.

Two further ways to represent asymmetric dependence relationships are discussed in Geiger and Heckerman [1996], where inference is carried out on **Bayesian Multinets** or **Similarity Networks**. In a Bayesian Multinet, a variable, called 'hypothesis variable', is selected and a separate BN is drawn for each value of the variable, i.e. for each hypothesis. Each of the BNs can then represent different independence structures for different hypotheses. A Similarity Network takes a slightly different approach and attempts to assess the similarity between two values of the hypothesis variable through a 'similarity graph' and describes each comparison in terms of a separate directed graph on the covariates of the hypothesis variable. In both cases the graphical representation is hence across multiple networks rather than within a single graphical model and becomes complicated when several hypothesis variables are of interest.

Another class of model which benefits from admitting asymmetric dependence structures is the **Probabilistic Decision Graph** (PDG) [Jaeger, 2004; Jaeger et al., 2006]. Although the CEG's structural syntax is closely linked to PDGs, it is a more general class of models due to the additional colouring of the CEG when two positions are in the same stage. Jaeger [2004] showed that PDGs and BNs are incomparable regarding the conditional independence statements they encode and hence that the BN is not a subclass of the PDG. In contrast to this, I have demonstrated in Section 2.4 that any BN can be written as a CEG. Also, the model given in Figure 2.5 could not be represented as a PDG, due to the colouring of the CEG with respect to w_4 and w_5, which gives additional information on the effect of the social background and economic situation on the life events. Like context-specific BNs, the PDG concentrates primarily on computational efficiency and less on an accessible representation.

Recently, Edwards and Ankinakatte [2013] discussed in a research report how **Acyclic Probabilistic Finite Automata** (APFA) [Ron et al., 1995] relate to more commonly used graphical models, including CEGs, and suggest that the APFA is a special case of the CEG. An APFA is a class of graphical model which generates strings of symbols and is commonly used within speech recognition and natural language processing. However, it has otherwise not been employed within other applications of statistics. Similar to the CEG, the APFA has a single root vertex from which every path in the graph emanates and a single sink vertex which forms the end point of every path. It can also have multiple edges between its vertices, where each edge has an associated symbol and a conditional probability attached to it. Edwards and Ankinakatte [2013] then show that an APFA can be constructed from a probability tree by 'contracting' two situations and their entire subtrees. However, this is determined by non-Bayesian methods and once a situation is merged with another, all subsequent situations in their corresponding subtrees are automatically also merged. Although, the APFA can be shown to encode conditional independencies and certain BNs, the BN is not a subclass of the APFA. As before, it is the additional stage structure and colouring of the CEG, which is not present within an APFA, which ensures that every BN can be written as a CEG. Further, an APFA does not allow for asymmetric trees with paths of different lengths, so that the CEGs developed in Chapter 4 could not be represented as an APFA.

I have demonstrated in this Chapter that the CEG is unique in its graphical representation of problems based on probability trees and is particularly expressive for highly asymmetric dependence structures. Unlike other models it captures the BN within its model class, while still giving a stand-alone representation of a problem. Together with the development of the Ordinal CEG and the Reduced Ordinal CEG, detailed conclusions can be drawn directly from its graph which go beyond the conclusions that can be read from a BN or a context-specific BN. In the following Chapter I will discuss Bayesian learning for CEGs and BNs as well as applying these to the CHDS example. The inference made on the resulting graphs will support the findings in this chapter.

Chapter 3

Bayesian Learning of Bayesian Networks and Chain Event Graphs

The usual Bayesian techniques for making inference and learning about parameters and structure have been widely applied to Bayesian Networks [Korb and Nicholson, 2004; Cowell et al., 2007; Jensen and Nielsen, 2007; Heckerman, 2008] and more recently to CEGs [Freeman and Smith, 2011a; Barclay et al., 2013a]. In this chapter I will review the learning of the parameters and model selection in a CEG. I will first demonstrate in Section 3.1 how we learn the parameters in a BN according to Heckerman et al. [1995] and Heckerman [2008] and then move on to the analogous learning of the parameters in CEGs, showing that Dirichlet priors on the stages are essential [Freeman and Smith, 2011a]. The Bayesian Dirichlet (BD) metric can then be derived directly for both classes of models in Section 3.2. I will further introduce a greedy search algorithm for CEGs developed by Freeman and Smith [2011a], which searches quickly over the CEG model space using Bayes Factors, and briefly conclude with possible alternatives and extensions to this algorithm. In Section 3.3 the introduced methodology is applied to the example of the CHDS study discussed in Chapter 2. I will demonstrate that the CEG not only enables more refined conclusions of the problem but also results in a significantly higher scoring model. Finally, I will review the Ordinal CEG and causal interventions in a CEG for the CHDS example on the learnt Maximum a Posteriori (MAP) model.

3.1 Learning the Parameters of BNs and CEGs

In this section I will describe how we learn parameters in a BN according to Heckerman [2008] and will then extend this to the CEG based on Freeman and Smith [2011a].

Assume that we have a BN with graph $\mathcal{B}$ on p variables $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$. Also assume that each variable X_i can have q_i different parent configurations associated with it. Following the notation of Heckerman [2008] define

$$\boldsymbol{\theta}_{ij} = p(x_i|pa(x_i) = j; \boldsymbol{\theta}), \tag{3.1}$$

for all $i = 1, \ldots, p$ and $j = 1, \ldots, q_i$. Assuming that each variable X_i can take m_i different values, $\boldsymbol{\theta}_{ij} = (\theta_{ij1}, \ldots, \theta_{ijm_i})$, where

$$\theta_{ijk} = p(x_i = k|pa(x_i) = j; \boldsymbol{\theta}),$$

for $k = 1, .., m_i$. Finally, let $\boldsymbol{\theta}_i = \{\boldsymbol{\theta}_{ij} : j = 1, \ldots, q_i\}$ and $\boldsymbol{\theta} = \{\boldsymbol{\theta}_{ij} : i = 1, \ldots, p, j = 1, \ldots, q_i\}$. By the recursive factorisation (Chapter 2.2, Equation 2.2) the joint pmf of observing $\boldsymbol{X} = \boldsymbol{x}$ can be written as

$$p(\boldsymbol{x}|\boldsymbol{\theta}, \mathcal{B}) = \prod_{i=1}^{p} p(x_i|pa(x_i); \boldsymbol{\theta}_i, \mathcal{B}).$$

Then given a complete random sample $\mathcal{S} = \{\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^n\}$ of n realisations, the likelihood of $\boldsymbol{\theta}$ takes the form

$$L(\boldsymbol{\theta}|\mathcal{S}, \mathcal{B}) = \prod_{s=1}^{n} \prod_{i=1}^{p} p(x_i^s|pa(x_i^s), \boldsymbol{\theta}_i, \mathcal{B}).$$

From Equation 3.1 it is straightforward that this is a multinomial likelihood with parameters θ_{ijk}, $i = 1, \ldots p$, $j = 1, \ldots, q_i$, $k = 1, \ldots, m_i$ which can be rewritten by counting the number of times, N_{ijk}, that each $x_i = k$ with parent configuration $pa(x_i) = j$ is observed. We then obtain

$$L(\boldsymbol{\theta}|\boldsymbol{N}, \mathcal{B}) = \prod_{i=1}^{p} \prod_{j=1}^{q_i} \prod_{k=1}^{m_i} \theta_{ijk}^{N_{ijk}},$$

with $\boldsymbol{N} = \{\boldsymbol{N}_{ij}, i = 1, \ldots p, j = 1, \ldots, q_i\}$ and $\boldsymbol{N}_{ij} = \{N_{ij1}, \ldots, N_{ijm_i}\}$. The likelihood then immediately separates such that

$$L(\boldsymbol{\theta}|\boldsymbol{N},\mathcal{B}) = \prod_{i=1}^{p}\prod_{j=1}^{q_i} L_{ij}(\boldsymbol{\theta}_{ij}|\boldsymbol{N}_{ij},\mathcal{B}),$$

and

$$L_{ij}(\boldsymbol{\theta}_{ij}|\boldsymbol{N}_{ij},\mathcal{B}) = \prod_{k=1}^{m_i} \theta_{ijk}^{N_{ijk}}. \tag{3.2}$$

Let $p(\boldsymbol{\theta}|\mathcal{B})$ be the joint pdf of $\boldsymbol{\theta}$. When further making the assumptions of **global and local parameter independence** [Spiegelhalter and Lauritzen, 1990], which assume that all $\boldsymbol{\theta}_{ij}$ are a priori mutually independent, then the prior also factorises as

$$p(\boldsymbol{\theta}|\mathcal{B}) = \prod_{i}^{p}\prod_{j}^{q_i} p_{ij}(\boldsymbol{\theta}_{ij}|\mathcal{B}). \tag{3.3}$$

As a consequence each of the $\boldsymbol{\theta}_{ij}$ can be learnt separately. It was then shown in Heckerman et al. [1995] that under the two additional assumptions of **parameter modularity** and **likelihood equivalence** a Dirichlet distribution on the $\boldsymbol{\theta}_{ij}$ is inevitable. Parameter modularity requires that, when a variable X_i has the same parent configuration j in two BN structures $\mathcal{B}_1$ and $\mathcal{B}_2$, then the corresponding prior on $\boldsymbol{\theta}_{ij}$ is the same ($p(\boldsymbol{\theta}_{ij}|\mathcal{B}_1) = p(\boldsymbol{\theta}_{ij}|\mathcal{B}_2)$). Likelihood equivalence says that when two BN structures are equivalent (see Chapter 2.2), then they will have the same marginal likelihood ($L(\mathcal{B}_1|\boldsymbol{N}) = L(\mathcal{B}_2|\boldsymbol{N})$). This further assumes that all structures have a prior probability greater than zero (**structure possibility**). Therefore, each $\boldsymbol{\theta}_{ij}$ is given a prior Dirichlet distribution, $\boldsymbol{\theta}_{ij} \sim Dir(\boldsymbol{\alpha}_{ij}), \boldsymbol{\alpha}_{ij} = (\alpha_{ij1},\ldots,\alpha_{ijm_i})$, which takes the form

$$p_{ij}(\boldsymbol{\theta}_{ij}|\mathcal{B}) = \frac{\Gamma(\sum_{k=1}^{m_i}\alpha_{ijk})}{\prod_{k=1}^{m_i}\Gamma(\alpha_{ijk})}\prod_{k=1}^{m_i}\theta_{ijk}^{\alpha_{ijk}-1}. \tag{3.4}$$

The Dirichlet priors can then be updated in closed form given a multinomial likelihood as in Equation 3.2 to obtain the posterior distribution

$$p_{ij}(\boldsymbol{\theta}_{ij}|\boldsymbol{N}_{ij},\mathcal{B}) = \frac{\Gamma(\sum_{k=1}^{m_i}(\alpha_{ijk}+N_{ijk}))}{\prod_{k=1}^{m_i}\Gamma(\alpha_{ijk}+N_{ijk})}\prod_{k=1}^{m_i}\theta_{ijk}^{\alpha_{ijk}+N_{ijk}-1}, \tag{3.5}$$

i.e. $\boldsymbol{\theta}_{ij}|\boldsymbol{N}_{ij} \sim Dir(\boldsymbol{\alpha}_{ij}+\boldsymbol{N}_{ij})$, and the full posterior of $\boldsymbol{\theta}$ on a BN $\mathcal{B}$ therefore takes the form

$$p(\boldsymbol{\theta}|\boldsymbol{N},\mathcal{B}) = \prod_{i=1}^{p}\prod_{j=1}^{q_i} p_{ij}(\boldsymbol{\theta}_{ij}|\boldsymbol{N}_{ij},\mathcal{B}).$$

I will now demonstrate an analogous prior to posterior analysis for CEGs.

Assume a CEG structure $\mathcal{C}$ with position partition $U(\mathcal{C})$. Then, analogous to the above, let

$$\pi_{uk} = p(e_{uk}|u;\boldsymbol{\pi}),$$

for $k = 1,\ldots,m_u$ and let $\boldsymbol{\pi}_u = (\pi_{u1},\ldots,\pi_{um_u})$ and $\boldsymbol{\pi} = \{\boldsymbol{\pi}_u : u \in U(\mathcal{C})\}$. Recall that by the factorisation in Chapter 2.3, Equation 2.6, the joint pmf of an individual going along a particular path $\epsilon = (e_{w_0k_0}, e_{w_{i_1}k_1},\ldots,e_{w_{i_n}k_n})$ in $\mathcal{C}$, which starts at w_0 and end at w_∞, can be written as

$$p(\epsilon|\mathcal{C}) = \pi_{u_0k_0}\prod_{a=1}^{n}\pi_{u_ak_a},$$

where u_a, $a = 1,\ldots,n$ are not necessarily distinct as a stage may consist of several positions. Given a complete random sample $\mathcal{S}$ the likelihood of $\boldsymbol{\pi}$ can be written by counting the number of times, N_{uk}, an individual reaches a stage u and passes along the edge e_{uk}. This also gives a multinomial likelihood with parameters π_{uk}, with $u \in U$ and $k = 1,\ldots,m_u$, of the form

$$L(\boldsymbol{\pi}|\boldsymbol{N},\mathcal{C}) = \prod_{u\in U(\mathcal{C})}\prod_{k=1}^{m_u}\pi_{uk}^{N_{uk}},$$

where $\boldsymbol{N}_u = (N_{u1}, N_{u2},\ldots,N_{um_u})$ and $\boldsymbol{N} = \{\boldsymbol{N}_u, u \in U(\mathcal{C})\}$. This likelihood then separates according to the stage partition U of $\mathcal{C}$ such that

$$L(\boldsymbol{\pi}|\boldsymbol{N},\mathcal{C}) = \prod_{u\in U(\mathcal{C})} L_u(\boldsymbol{\pi}_u|\boldsymbol{N}_u,\mathcal{C}), \text{ and}$$

$$L_u(\boldsymbol{\pi}_u|\boldsymbol{N}_u,\mathcal{C}) = \prod_{k=1}^{m_u}\pi_{uk}^{N_{uk}}. \tag{3.6}$$

Assuming prior independence of $\boldsymbol{\pi}_u$, associated with each stage u, is the equivalent condition to the local and global parameter independence for BNs. Under this assumption the prior pdf of $\boldsymbol{\pi}$ can be written as

$$p(\boldsymbol{\pi}|\mathcal{C}) = \prod_{u\in U} p_u(\boldsymbol{\pi}_u|\mathcal{C}). \tag{3.7}$$

Given a tree structure, let the CEG where every situation is in a separate stage be called $\mathcal{C}_0$. Under two assumptions, namely that 1. the rates at which individuals go along the tree are independent and 2. the probability at which individuals take an edge after reaching a situation is independent of the rate at

which they arrived at this situation, Freeman and Smith [2011a] show that each parameter vector π_u associated with a floret in $\mathcal{C}_0$ has an independent Dirichlet prior. Using ideas similar to parameter modularity, they follow an argument exactly analogous to that of Heckerman et al. [1995] and prove that, under the additional assumption that the stage priors are identical for equivalent stages in different CEG structures ($p_u(\boldsymbol{\pi}_u|\mathcal{C}_1) = p_u(\boldsymbol{\pi}_u|\mathcal{C}_2)$), Dirichlet priors on the $\pi_u, u \in U(\mathcal{C})$ are inevitable. We hence put a Dirichlet prior on each CPV $\pi_u \sim Dir(\boldsymbol{\alpha}_u)$ with parameters $\boldsymbol{\alpha}_u = (\alpha_{u1}, \alpha_{u2}, \ldots, \alpha_{um_u})$, which takes the form

$$p_u(\boldsymbol{\pi}_u|\mathcal{C}) = \frac{\Gamma(\sum_{k=1}^{m_u} \alpha_{uk})}{\prod_{k=1}^{m_u} \Gamma(\alpha_{uk})} \prod_{k=1}^{m_u} \pi_{uk}^{\alpha_{uk}-1}. \tag{3.8}$$

Then, as before, given a complete random sample, the parameters $\boldsymbol{\pi}_u$ can be updated separately and in closed form using Equation 3.6 resulting in the posterior distribution

$$p_u(\boldsymbol{\pi}_u|\boldsymbol{N}_u, \mathcal{C}) = \frac{\Gamma(\sum_{k=1}^{m_u} (\alpha_{uk} + N_{uk}))}{\prod_{k=1}^{m_u} \Gamma(\alpha_{uk} + N_{uk})} \prod_{k=1}^{m_u} \pi_{uk}^{\alpha_{uk}+N_{uk}-1}. \tag{3.9}$$

A useful consequence of this closed form prior to posterior analysis is that the posterior predictive probabilities of observing $x_i = k$ given a parent configuration j in a BN or going along an edge e_{uk} after reaching u in a CEG are also easy to calculate. These are simply given by the expectations of the Dirichlet posterior for $\boldsymbol{\theta}_{ij}$ and $\boldsymbol{\pi}_u$ respectively. Hence for BNs,

$$\begin{aligned} p(x_i = k|pa(x_i) = j; \boldsymbol{N}_{ij}, \mathcal{B}) &= \int \theta_{ijk} \times p_{ij}(\boldsymbol{\theta}_{ij}|\boldsymbol{N}_{ij}, \mathcal{B}) d\boldsymbol{\theta}_{ij} \\ &= E_{p_{ij}(\boldsymbol{\theta}_{ij}|\boldsymbol{N}_{ij}, \mathcal{B})}(\theta_{ijk}) \\ &= \frac{\alpha_{ijk} + N_{ijk}}{\sum_{k=1}^{m_i} (\alpha_{ijk} + N_{ijk})}, \end{aligned}$$

and for CEGs,

$$\begin{aligned} p(e_{uk}|u; \boldsymbol{N}_u, \mathcal{C}) &= \int \pi_{uk} \times p(\boldsymbol{\pi}_u|\boldsymbol{N}_u, \mathcal{C}) d\boldsymbol{\pi}_u \\ &= E_{p(\boldsymbol{\pi}_u|\boldsymbol{N}_u, \mathcal{C})}(\boldsymbol{\pi}_u) \\ &= \frac{\alpha_{uk} + N_{uk}}{\sum_{k=1}^{m_u} (\alpha_{uk} + N_{uk})}. \end{aligned}$$

Hence, the posterior CPVs for BNs and CEGs can be easily calculated by the above. In the following applications of the thesis I will attach the posterior predictive

probability of the binary outcome variable to each of the final positions of the Ordinal CEGs, as it is these probabilities that are of most interest and through which the ordering of the positions in an Ordinal CEG are determined.

Given a plausible CEG structure it is further important to validate the model in order to determine whether it is an accurate representation of the process being modelled. In real-world example such as the ones considered in this thesis, real data on the history of the process is desirable. One common approach is to split the data into a training set and a test set, for example with a 90% to 10% split, and to learn the model only using the training data set. The test data set can then determine the predictive accuracy of the CEG by comparing the predictive outcome, for example whether the individual is admitted to hospital or not, to the actual value of the outcome. However, this approach does not take into account the probabilistic nature of the predictions [Korb and Nicholson, 2004]. Further, when only a small data set or data with highly asymmetric cell counts, then this may not be a desirable approach, as a subset of the data is no longer used to determine the CEG structure and the predictive probabilities. Alternatively, a jackknife analysis could be used which calculates the parameter estimates of the model of each subsample obtained by leaving out one observation at a time. The jackknife estimator of the parameters is then given by the average of these estimates. Given a CEG structure a jackknife analysis may be useful for variance and bias estimation of the predictive probabilities associated with the outcome variable. A different approach is further suggested in Cowell et al. [2007] for BNs. Here a penalty, given by the negative logarithm of the overall probability that the data is observed, is calculated and the hypothesis that the observed events are occurring with the probabilities stated by the model is tested. This approach checks for an overall misfit between the model and the data.

3.2 Model selection for BNs and CEGs

Rather than assuming a BN or CEG structure and updating its conditional probabilities given the data provided, we may instead want to learn the structure of the BN or CEG that best fits the given data set. The standard Bayesian methodology then assigns prior distributions to the structures and updates these to its posterior distribution using the marginal likelihood of the structure given the data. The two approaches that are commonly used when we are uncertain about the structure are model selection, in which we aim to select a 'good' model from the set of possible models, or selective model averaging, where we determine a set of possible models with a high posterior probability and average over these. In this thesis I will con-

sider only model selection as the interest lies in determining a particular model and its graphical representation from which inference can be drawn.

Common model selection algorithms are either constraint-based learning algorithms, such as the PC algorithm [Spirtes et al., 1993], or metric learning algorithms, where each graph is scored according to a certain metric and the structure with the highest score selected. Throughout this thesis I will use the Bayesian Dirichlet (BD) metric to score BNs and analogously CEGs. Other common metrics include the Minimum Message Length (MML) metric [Korb and Nicholson, 2004], the Bayesian Information Criterion (BIC) [Schwarz, 1978] and Akaike's Information Criterion (AIC) [Akaike, 1974]. All these metrics score the model according to its fit to the data, while penalising for model complexity and are useful alternatives to the BD metric.

The BD metric for scoring BNs was first derived in Cooper and Herskovits [1992] and was further developed in Heckerman et al. [1995]. It is given by the joint density $p(\mathcal{G}, \boldsymbol{N})$ of a graph $\mathcal{G}$, here a BN or a CEG, and a complete random sample. This can be factorised into $p(\mathcal{G}, \boldsymbol{N}) = p(\mathcal{G})L(\mathcal{G}|\boldsymbol{N})$, the prior of the graph $\mathcal{G}$ multiplied by the marginal likelihood of the graph given the data. By scoring the models in this way, the structure which obtains the highest score, called the Maximum a Posteriori (MAP) model structure, can be determined. Under the assumption of Dirichlet priors the **marginal likelihood** of a BN, $L(\mathcal{B}|\boldsymbol{N})$, can then be calculated in closed form directly from Equations 3.4 and 3.5. Hence the BD metric for a BN $\mathcal{B}$ is given by:

$$p(\mathcal{B}) \prod_{i=1}^{p} \prod_{j=1}^{q_i} \frac{\Gamma(\sum_{k=1}^{m_i} \alpha_{ijk})}{\Gamma(\sum_{k=1}^{m_i} (\alpha_{ijk} + N_{ijk}))} \prod_{k=1}^{m_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}. \tag{3.10}$$

To find the BD metric of a given structure it is hence only necessary to determine the hyperparameters a_{ijk} of the priors $p(\boldsymbol{\theta}_{ij})$, with $i = 1, ..., p, j = 1, ..., q_i, k = 1, ..., m_i$.

Call $\alpha = \sum_{i=1}^{p} \sum_{j=1}^{q_i} \sum_{k=1}^{m_i} \alpha_{ijk}$ the **equivalent sample size** of the prior $p(\boldsymbol{\theta})$, which specifies the number of data points the prior is worth and hence gives a measure of confidence about the prior. Heckerman et al. [1995] then prove that under likelihood equivalence, parameter modularity and local and global parameter independence, the hyperparameters α_{ijk} of any BN $\mathcal{B}$ on a set of variables is determined by:

$$\alpha_{ijk} = \alpha p(x_i = k, pa(x_i) = j|\mathcal{B}_c),$$

where $\mathcal{B}_c$ is the complete BN of the variables in the problem. Heckerman et al. [1995] suggest finding $p(\boldsymbol{x}|\mathcal{B}_c)$ by specifying a prior network. The simplest form is a

prior network such that $p(\boldsymbol{x}|\mathcal{B}_c)$ is uniform and Heckerman's approach then simply corresponds to letting

$$\alpha_{ijk} = \frac{\alpha}{q_i m_i}, \tag{3.11}$$

where q_i is the number of parent configurations associated with variable x_i and m_i is the number of values x_i can take. The BD metric defined in 3.10 together with 3.11 was previously developed by Buntine [1991] and is called the BDeu-metric in Heckerman et al. [1995] due to the assumption of likelihood equivalence and the uniform prior on $p(\boldsymbol{x}|\mathcal{B})$. To determine the equivalent sample size, Neapolitan [2004] suggests setting $\alpha = \max m_i$, the largest number of possible values a variable in the problem can take, to ensure that the α_{ijk} are simple fractions when a weak uniform prior is appropriate. However, Silander et al. [2012] demonstrate that model selection is highly sensitive towards the equivalent sample size. Therefore, when carrying out model selection it is important to carry out a sensitivity analysis on the equivalent sample size to determine how it affects the model.

Analogously, for CEGs, the BD metric for a CEG structure $\mathcal{C}$ can be immediately derived from Equations 3.8 and 3.9 as:

$$p(\mathcal{C}) \prod_{u \in U} \frac{\Gamma(\sum_{k=1}^{m_u} \alpha_{uk})}{\Gamma(\sum_{k=1}^{m_u} (\alpha_{uk} + N_{uk}))} \prod_{k=1}^{m_u} \frac{\Gamma(\alpha_{uk} + N_{uk})}{\Gamma(\alpha_{uk})}. \tag{3.12}$$

Throughout the thesis the logarithm of the BD metric is often used for more efficient calculations, which is given by

$$\log p(\mathcal{C}) + \sum_{u \in U(\mathcal{C})} \left(\log \Gamma \left(\sum_{k=1}^{m_u} \alpha_{uk} \right) - \log \Gamma \left(\sum_{k=1}^{m_u} (\alpha_{uk} + N_{uk}) \right) + \sum_{k=1}^{m_u} \left(\log \Gamma(\alpha_{uk} + N_{uk}) - \log \Gamma(\alpha_{uk}) \right) \right). \tag{3.13}$$

Freeman and Smith [2011a] then prove that, under the assumptions that the stage priors are identical for equivalent stages in different CEG structures and assuming independent Dirichlet priors on $\mathcal{C}_0$, the hyperparameters $\alpha_{uk}, k = 1, \dots m_u$, associated with each stage $u \in U(\mathcal{C})$ in a CEG $\mathcal{C}$ can be deduced from the equation

$$\alpha_{uk} = \alpha \sum_{s_i \in u} p(e_{s_i k}, s_i | \mathcal{C}_0),$$

where $p(e_{s_i k}|s_i; \mathcal{C}_0)$ corresponds to the probability of going along the edge $e_{s_i k}$ in $\mathcal{C}_0$, given s_i has been reached, which is equal to the probability of having reached this situation in the associated tree and going along the kth edge. So, priors across models are deduced by summing corresponding hyperparameters when two stages are merged. Similarly to 3.11, in the simplest case, a default uniform prior is given to

the paths in $\mathcal{C}_0$, i.e. we put a uniform prior on the root-to-leaf paths in the associated tree, such that all possible combinations of developments are a priori equally likely. The equivalent sample size of the prior is determined by $\alpha = \sum_{u \in U} \sum_{k=1}^{m_u} \alpha_{uk}$. The approach of Neapolitan [2004] can be adapted to specify α to be equal to $\alpha = \max m_u$ such that the hyperparameters of α_{uk} result in simple fractions given a weak uniform prior. When the marginal likelihood cannot be calculated in closed form, is is possible to resort to MCMC methods to sample from the posterior distribution.

When determining the priors on the structures, $p(\mathcal{B})$ for BNs or $p(\mathcal{C})$ for CEGs, the default approach has been to either assign equal probabilities to all possible structures or to select a set of plausible structures and assign equal probabilities to these. In this case the marginal likelihood suffices as a selection criterion, as the contribution of $p(\mathcal{B})$ and $p(\mathcal{C})$ in 3.10 and 3.12 respectively, is the same for all structures. An improved approach would be to determine prior probabilities on different stage partitions from an expert and use these to assign CEG structures different prior probabilities according to the expert's beliefs.

To compare two competing graphical structures $\mathcal{G}_1$ and $\mathcal{G}_2$ we can calculate

$$\frac{p(\mathcal{G}_1|\boldsymbol{N})}{p(\mathcal{G}_2|\boldsymbol{N})} = \frac{p(\mathcal{G}_1)L(\mathcal{G}_1|\boldsymbol{N})}{p(\mathcal{G}_2)L(\mathcal{G}_2|\boldsymbol{N})}, \tag{3.14}$$

the **posterior odds** in favour of $\mathcal{G}_1$, or equivalently the ratio of the BD metric. Under the assumptions that the structure priors are equal this reduces to

$$\frac{L(\mathcal{G}_1|\boldsymbol{N})}{L(\mathcal{G}_2|\boldsymbol{N})},$$

the **Bayes Factor** of the two models $\mathcal{G}_1$ and $\mathcal{G}_2$ or equivalently the ratio of the marginal likelihood of the two models. By the BD metric given in (3.10) and (3.12) the posterior odds or the Bayes Factor of two models is very easily calculated in closed form. One possible interpretation of the Bayes Factor is given by Kass and Raftery [1995] who divide the range of values the Bayes Factor can take into intervals that determine the strength of evidence in favour of one of the models. The suggested intervals and interpretations are repeated in Table 3.1 which will be used as a guideline for comparing model structures throughout the thesis. To find the MAP BN or MAP CEG the models are therefore scored according to the BD metric in 3.10 and 3.12 and compared using Bayes Factors or posterior odds. For BNs, routine model selection procedures can be carried out using the 'deal' package in R [Bøttcher and Dethlefsen, 2003]. Here all prior structures are set to be equally likely and, when carrying out a full model search, the package returns the logarithm

$log(BF_{\mathcal{G}_1,\mathcal{G}_2})$	$BF_{\mathcal{G}_1,\mathcal{G}_2}$	Evidence against $\mathcal{G}_2$
$0 - 1.10$	$1 - 3$	Not worth more than a bare mention
$1.10 - 3$	$3 - 20$	Substantial
$3 - 5$	$20 - 150$	Strong
> 5	> 150	Decisive

Table 3.1: Scale of evidence for Bayes Factors according to Kass and Raftery [1995]

of the marginal likelihood as a network score for each BN and the Bayes Factor of each BN with respect to the most probable network. However, when the number of variables in the BN is large, there may be too many possible structures to calculate each network score and perform an exhaustive search, so that heuristic search algorithms are used. In R a greedy search algorithm with random restarts according to Heckerman et al. [1995] has been implemented: This starts with a randomly chosen BN and then finds the MAP BN by adding, deleting or reversing an edge at every step of the algorithm until no improvement in score can be obtained.

For CEGs, it is also possible to score all CEGs given a tree representation of a given problem. However, the model space of CEGs is far larger than the space of possible BN structures. For example, assume we have a tree structure with only binary variables, so that the first variable in the tree is described by the single root floret, the second variable by two florets, the third by four florets and so on. Then there will be $B_2 = 2$ possible ways to partition the situations whose emanating edges describe the second variable, $B_4 = 15$ partitions for the situations of the third variable, $B_8 = 4140$ for the fourth set of situations etc., where B_i is the ith Bell number. Therefore, an exhaustive search is infeasible in all but the simplest case. Consequently, Freeman and Smith [2011a] developed a greedy search algorithm for CEGs called the Bayesian Agglomerative Hierarchical Clustering (AHC) algorithm. The algorithm starts at the finest partition of the CEG, $\mathcal{C}_0$. It then quickly searches over the model space by finding at every step the two stages, which, when merged, provide the highest CEG score. The algorithm stops once the coarsest partition of the CEG has been reached and the CEG with the highest overall score is selected. The CEG with the coarsest partition is called $\mathcal{C}_\infty$ and corresponds to a CEG where all situations whose florets have the same topology have been merged into a single stage (compare Definition 18). I repeat the exact steps of the algorithm below:

1. From a given tree structure first construct the CEG $\mathcal{C}_0$, where all leaf nodes are in the terminal position w_∞ and all situations are in separate stages and calculate the logarithm of the BD metric using 3.13

2. For each pair of stages, u_i and u_j, whose florets $F(u_i)$ and $F(u_j)$ have the

same topology, calculate the posterior log-odds given in 3.14

$$\log \frac{p(\mathcal{C}_1^*, \boldsymbol{N})}{p(\mathcal{C}_0, \boldsymbol{N})},$$

where $\mathcal{C}_1^*$ is the CEG constructed by merging the two stages u_i and u_j.

3. Let $\mathcal{C}_1 = \arg\max_{\mathcal{C}_1^*} \left(\frac{p(\mathcal{C}_1^*, \boldsymbol{N})}{p(\mathcal{C}_0, \boldsymbol{N})} \right)$.

4. Calculate $\mathcal{C}_2^*$ by merging two stages in $\mathcal{C}_1$ and hence find $\mathcal{C}_2$.

5. Continue until the coarsest partition, $\mathcal{C}_\infty$, is reached and select $\mathcal{C} = max(\mathcal{C}_0, \mathcal{C}_1, \ldots, \mathcal{C}_\infty)$ as the MAP CEG.

Recall that it is assumed that stage priors are identical for equivalent stages in different CEG structures. Therefore, as we are summing over the stages in the logarithm of the BD metric (compare 3.13), in order to calculate the difference in score we only need to compare the contributions to the score made by the stages in which the two structures differ. A further important aspect of the description of the algorithm is that a particular tree structure is assumed from which the MAP CEG is found and hence a fixed ordering of the variables in the tree is assumed. However, this can be simply extended by searching separately across all plausible tree structures, i.e. all possible permutations of the variables in the tree, and choosing the CEG which gives the overall highest score.

A potential weakness of using a greedy search algorithm, like the AHC, is that it might find a local maxima of the model space rather than the MAP CEG. In contrast to the heuristic algorithm suggested for BNs, the AHC algorithm allow us only to merge stages but not to split these again. An alternative to this algorithm could therefore be to implement a greedy search algorithm similar to the one described for BNs, where at each step two stages can either be merged or split to improve the model score. An initial CEG structure could be, for example, $\mathcal{C}_0$ as before, or alternatively, the MAP BN structure of the given problem translated into a CEG. Further, although the AHC algorithm is a greedy search algorithm, the number of calculations necessary in step 2 of the algorithm still increases exponentially with the number of variables. To overcome this, Freeman [2011] suggests restricting the model space to allow only certain situations in the tree to be merged and hence restrict the model space a priori.

Two further search algorithms have been suggested: Freeman [2011] proposes to reformulate the search as a weighted Maximum Satisfiability (MAX-SAT) problem, which was previously developed for BNs in Cussens [2008]. However, when

doing so, the score of each possible stage structure needs to be calculated a priori and therefore an exact search quickly becomes infeasible. Again restricting the weighted MAX-SAT algorithm by the number of situations that can be in a stage is suggested, however this does not seem to be optimal. Freeman [2011] gives an example on a tree with 170 situations and demonstrates the weighted MAX-SAT algorithm needed to be restricted according to the number of situations that can be in a stage, while the AHC algorithm still performed efficiently and found a higher scoring model.

Silander and Leong [2013] recently developed a dynamic programming algorithm for exact learning of CEGs, which not only searches across a single tree but across all possible permutations of the variables in the tree. However, also here an exact search quickly becomes infeasible, as the complexity of the algorithm grows super-exponentially with the number of variables. Silander and Leong [2013] suggest combining the algorithm with faster heuristic clustering algorithms, using K-means clustering. They claim that, when using their dynamic programming approach using K-means clustering, the search is feasible for up to around 30 variables.

For this thesis I have implemented the AHC algorithm in R based on previous work in Freeman [2011] to find the MAP CEG structure of a given problem (see B. The CEG space of the examples considered in the thesis is sufficiently small, so that the model space does not need to be restricted a priori.

3.3 Application to the CHDS

In this section I will apply the methodology introduced in this chapter to the CHDS example. I will first find the MAP BN structure by scoring each possible structure according to the BD metric and selecting the BN with the highest score. I will then similarly find the MAP CEG of the CHDS example using the AHC algorithm. I will then compare the two models and show that the MAP CEG has a significantly higher score than the MAP BN and hence that the CEG of the CHDS example gives a better fitting model. Having found the MAP CEG structure, I will briefly return to the Ordinal CEG and the Causal CEG from Chapter 2.5 and 2.3.3.

3.3.1 The Maximum a Posteriori BN for the CHDS Example

Recall from Example 1 in Chapter 2.2 that the CHDS example consists of the following four variables:

- X_1 = family social position: binary variable: 'low', 'high'

- X_2 = family economic situation: binary variable: 'low', 'high'
- X_3 = number of family life events: variable with three categories : 'low', 'average', 'high'
- X_4 = hospital admission: binary variable: 'yes', 'no',

and that this can be represented by a BN on four variables. Based on the conclusions of Fergusson et al. [1986] I originally deduced the BN in Figure 2.1. This stated that the economic situation has no effect on hospital admission once adjusting for the social background and the family life events, and further that this is the only non-trivial conditional independence statement. In this section I will now instead use the BD metric to find the best fitting BN structure given the data set of the CHDS example discussed in Chapter 1.2.1.

To set up the Dirichlet prior distributions on the $\boldsymbol{\theta}_{ij} = p(x_i|pa(x_i) = j; \boldsymbol{\theta})$, I assume a uniform prior on $p(\boldsymbol{x}|\mathcal{B}_c)$ such that the distribution over all possible configurations is uniform and hence the hyperparameters α_{ijk} of the $p(\boldsymbol{\theta}_{ij})$ are given by equation 3.11. I further specify an equivalent sample size of $\alpha = 3$, the maximum number of categories taken by a variable in the CHDS problem, As recommended in Neapolitan [2004]. Finally, I assume that structures are a priori equally likely and hence Bayes Factors are used throughout for the comparison of different models. An exhaustive search using the 'deal' package in R [Bøttcher and Dethlefsen, 2003] over all possible BNs on the four variables scores each BN according to the logarithm of the marginal likelihood of the structure given the data and finds the MAP model to be the DAG given in Figure 3.1 with associated CPVs given in Table 3.2.

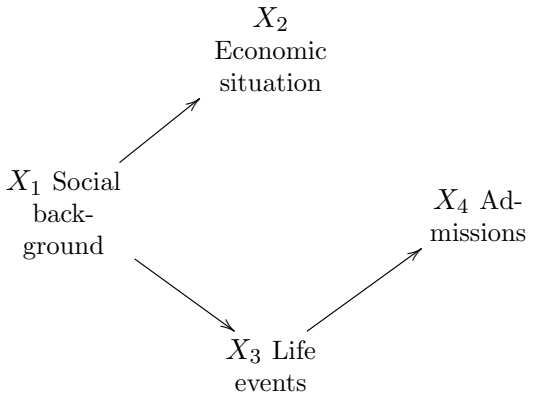


Figure 3.1: The Maximum a Posteriori BN of the CHDS example on social background, economic situation, life events and hospital admission. BN score (logarithm of the marginal likelihood) $\log L(\mathcal{B}|\boldsymbol{N}) = -2489.776$

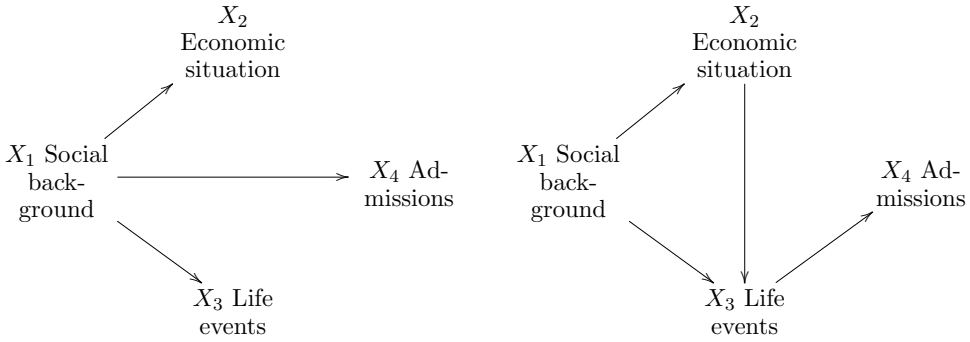
Similar to the network structure derived from Fergusson et al. [1986] (Figure 2.1) the MAP model suggests that hospital admission is independent of the

Conditional Probability Vector	
$(P(X_1 = \text{High}), P(X_1 = \text{Low}))$	$(0.569, 0.431)$
$(P(X_2 = \text{High} \mid X_1 = \text{High}), P(X_2 = \text{Low} \mid X_1 = \text{High}))$	$(0.468, 0.532)$
$(P(X_2 = \text{High} \mid X_1 = \text{Low}) P(X_2 = \text{Low} \mid X_1 = \text{Low}))$	$(0.122, 0.878)$
$(P(X_3 = \text{Low} \mid X_1 = \text{High}), P(X_3 = \text{Average} \mid X_1 = \text{High}), P(X_3 = \text{High} \mid X_1 = \text{High}))$	$(0.461, 0.347, 0.192)$
$(P(X_3 = \text{Low} \mid X_1 = \text{Low}), P(X_3 = \text{Average} \mid X_1 = \text{Low}), P(X_3 = \text{High} \mid X_1 = \text{Low}))$	$(0.248, 0.311, 0.441)$
$(P(X_4 = \text{No admission} \mid X_3 = \text{Low}), P(X_4 = \text{Admission} \mid X_3 = \text{Low}))$	$(0.880, 0.120)$
$(P(X_4 = \text{No admission} \mid X_3 = \text{Average}), P(X_4 = \text{Admission} \mid X_3 = \text{Average}))$	$(0.789, 0.211)$
$(P(X_4 = \text{No admission} \mid X_3 = \text{High}), P(X_4 = \text{Admission} \mid X_3 = \text{High}))$	$(0.743, 0.257)$

Table 3.2: The associated table of CPVs associated with the MAP BN from Figure 3.1

economic situation given the social background and the number of life events. However, exhibits several additional conditional independencies between the variables: It suggests that the economic situation and the family life events are independent given the social background ($X_3 \perp\!\!\!\perp X_2 \,|\, X_1$) and expresses that a direct dependency occurs only between the life events and the hospital admissions and not between social background and admissions ($X_4 \perp\!\!\!\perp X_1, X_2 \mid X_3$). Table 3.2 shows that the hospital admissions vary between 12% and 25.7% depending on the number of life events.

Nevertheless, the exhaustive search over all possible structures reveals two further BN structures scoring only slightly less than the MAP model which are given in Figure 3.2. Network structure (a) swaps the directed edge from family life



(a) 2nd BN, $\log L(\mathcal{B}|\boldsymbol{N}) = -2490.073$ (b) 3rd BN, $\log L(\mathcal{B}|\boldsymbol{N}) = -2490.751$

Figure 3.2: High scoring BN structures for the CHDS example on social background, economic situation, life events and hospital admission

events to admissions with an edge from the social background to the admissions. Structure (b) introduces an extra edge between the economic situation and the family life events. In comparison to the MAP model the log Bayes Factors are 0.297 and 0.975 favouring the MAP model. By Table 3.1, giving the scale of evidence for Bayes Factors, these differences in scores are negligible and hence, given the data

set provided, all three structures are believed to be similarly plausible. As noted in Section 3.1 model selection may be sensitive to the selected equivalent sample size. Nevertheless, in this case, increasing the equivalent sample size leads to the same three highest scoring models, together with the BN derived from Fergusson's results.

Although all three models suggested in Figures 3.1 and 3.2 have similar scores, the conclusions drawn from these three BN structures differ. While all three structures suggest that the social background affects the economic situation and the life events and that the economic situation does not influence hospital admissions, it is not clear in what way the social background and the life events affect the hospital admissions and whether the life events depend on the economic situation. This suggests that a model, which combines features of different competing BNs may be closer to the underlying true model. When searching the CEG space it will always be possible to find the CEG corresponding to the MAP BN structure, as the BN is a subclass of the class of CEGs. However, when considering the BN structures in Figure 3.1, it seems likely that we will be able to find a CEG which combines vertices into stages and positions in an asymmetric and hence result in a higher model score.

3.3.2 The Maximum a Posteriori CEG for the CHDS Example

In order to find the MAP CEG structure of the CHDS example the AHC algorithm is used to search across the space of CEGs. As discussed in Chapter 2.3 a plausible ordering of the variables is given by the ordering: social background first, economic situation, number of life events and finally hospital admission, with the associated tree representation given in Figure 2.3 of Chapter 2.3. In this section I will search across the CEG space with respect to this ordering and will briefly discuss at the end the effect of switching the economic situation and the number of life events.

To allow for a direct comparison with the BN I specify the same equivalent sample size of $\alpha = 3$ and assume that all paths in C_0 are a priori equally likely. I further assume that all CEG structures are a priori equally likely such that BNs and CEGs can be compared directly by their marginal likelihoods, $L(\mathcal{G}|\boldsymbol{N})$.

As described in the previous section the AHC algorithm starts at $\mathcal{C}_0$, calculates its score, and then at each step of the algorithm finds the two stages, which when merged, give the highest improvement in score. Table 3.3 gives the iterations of the algorithm until C_∞ is reached and shows at each step which stages (described in terms of their situations) have been merged (column 2), the improvement in the logarithm of the Bayes Factor (column 3) and the score of the new CEG (column

4).

CEG	Stages merged	Log-Bayes Factor	CEG score
C_0			-2512.708
C_1	$\{s_4, s_5\}$	5.528	-2507.180
C_2	$\{s_{17}, s_{18}\}$	3.731	-2503.449
C_3	$\{s_8, s_{16}\}$	3.453	-2499.996
C_4	$\{s_{12}, s_{17}, s_{18}\}$	3.377	-2496.619
C_5	$\{s_7, s_{10}\}$	3.305	-2493.314
C_6	$\{s_8, s_{11}, s_{16}\}$	3.060	-2490.254
C_7	$\{s_9, s_{12}, s_{17}, s_{18}\}$	3.041	-2487.213
C_8	$\{s_{13}, s_{14}\}$	2.565	-2484.648
C_9	$\{s_9, s_{12}, s_{15}, s_{17}, s_{18}\}$	2.514	-2482.134
C_{10}	$\{s_8, s_{11}, s_{13}, s_{14}, s_{16}\}$	2.342	-2479.792
C_{11}	$\{s_3, s_4, s_5\}$	1.302	-2478.490
C_{12}	$\{s_8, s_9, s_{11}, s_{12}, s_{13}, s_{14}, s_{15}, s_{16}, s_{17}, s_{18}\}$	-0.812	-2479.302
C_{13}	$\{s_7, s_8, s_9, s_{10}, s_{11}, s_{12}, s_{13}, s_{14}, s_{15}, s_{16}, s_{17}, s_{18}\}$	-8.764	-2488.066
C_{14}	$\{s_3, s_4, s_5, s_6\}$	-36.638	-2524.704
C_∞	$\{s_1, s_2\}$	-62.440	-2587.144

Table 3.3: The iterations of the AHC algorithm when finding the MAP CEG structure for the CHDS example given by the event tree in Figure 2.3

The highest scoring CEG structure found is the CEG $\mathcal{C}_{11}$ with a total score of -2478.49. It corresponds to the CEG structure that has already been described in Chapter 2.3 and which is repeated in Figure 4.12 together with its CPVs in Table 3.4. Here, the posterior predictive admission probabilities have been added to the final positions w_6, w_7 and w_8, which are, together with their 95% credible intervals given by: $9.1(5.7, 13.0)\%$, $17.7(13.5, 22.3)\%$ and $26.5(22.1, 31.1)\%$. In fact, an exhaustive search across the associated probability tree, carried out in Cowell and Smith [2011], shows that the resulting CEG structure of the AHC algorithm is the MAP model. Further, the model appears to be robust with respect to the chosen equivalent sample size. Increasing the equivalent sample size continuously up to $\alpha = 60$ keeps the same final positions and only splits w_3 and w_4 into separate stages.

As indicated in Chapter 2.3, these final positions are of particular interest, as they give an interpretation of the effect of a combination of variables on the hospital admissions and can be seen as describing three different 'health states' that the individual can reach, where the CEG lets us trace the different paths the individuals can take before ending up in one of these three 'states'. As expected, the lowest predictive probability of a hospital admission (9.1%) is present when the individual reaches position w_6, while the highest admission probability of 26.5% occurs when the individual has a high number of life events or a low social background, low

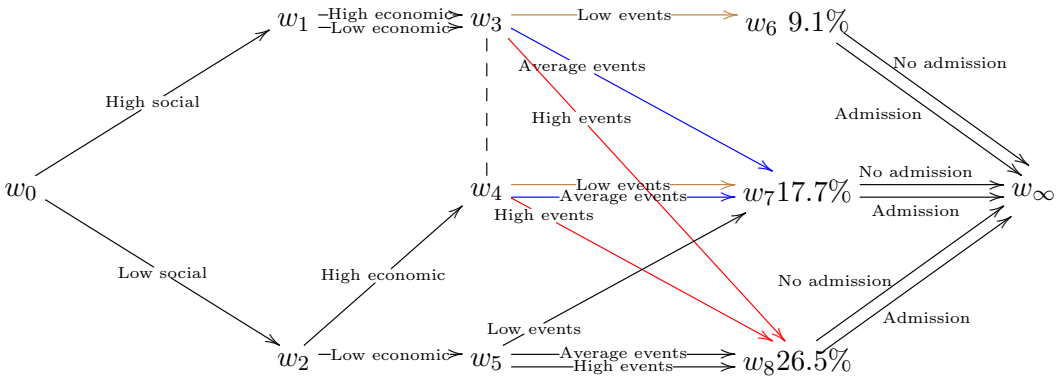


Figure 3.3: The MAP CEG structure found through the AHC algorithm from the event tree in Figure 2.3. CEG score $L(\mathcal{C}|\boldsymbol{N}) = -2478.490$

Stage/Position	Conditional Probability Vector	
$u_0 = w_0$	$(P(X_1 = \text{High}), P(X_1 = \text{Low}))$	$(0.569, 0.431)$
$u_1 = w_1$	$(P(X_2 = \text{High}\|u_1), P(X_2 = \text{Low}\|u_1))$	$(0.468, 0.532)$
$u_2 = w_2$	$(P(X_2 = \text{High}\|u_2), P(X_2 = \text{Low}\|u_2))$	$(0.122, 0.878)$
$u_3 = \{w_3, w_4\}$	$(P(X_3 = \text{Low}\|u_3), P(X_3 = \text{Average}\|u_3), P(X_3 = \text{High}\|u_3))$	$(0.460, 0.344, 0.196)$
$u_4 = w_5$	$(P(X_3 = \text{Low}\|u_4), P(X_3 = \text{Average}\|u_4), P(X_3 = \text{High}\|u_4))$	$(0.219, 0.312, 0.469)$
$u_5 = w_6$	$(P(X_4 = \text{No admission}\|u_5), P(X_4 = \text{Admission}\|u_5))$	$(0.909, 0.091)$
$u_6 = w_7$	$(P(X_4 = \text{No admission}\|u_6), P(X_4 = \text{Admission}\|u_6))$	$(0.823, 0.177)$
$u_7 = w_8$	$(P(X_4 = \text{No admission}\|u_7), P(X_4 = \text{Admission}\|u_7))$	$(0.735, 0.265)$

Table 3.4: Table of CPVs associated with the MAP CEG structure given in Figure 3.3

economic situation and an average number of life events. Finally, individuals that reach w_7 are predicted to have an admission probability of 17.7%. It is noticeable that, while a high number of life events forces the individuals into position w_8 with the highest admission probability, an individual from a low social background will never reach position w_6 even with a low number of life events. Table 3.4 further illustrates that an individual from a low social background is more likely to also have a low economic background (53.2% versus 87.8%). Similarly, an individual from a low social and economic background has a predictive probability of 46.9% of having a high number of life events and 21.9% for a low number of life events, while for the remaining offspring these probabilities are 19.6% and 46.0%, respectively.

It is also interesting to compare these probabilities with the CPVs of the MAP BN in Table 3.2. The predictive probabilities of a high and low social background and of the economic situation given the social background are naturally identical, as the CEG does not merge situations s_1 and s_2 and the BN, similarly, has a directed edge from the social background to the economic situation. However, the probabilities of the life events differ slightly as the CEG here represents the first

asymmetry, as three edges lead into $u_3 = \{w_3, w_4\}$ and only one edge into position $u_4 = \{w_5\}$. Finally, the CEG allows us to make slightly greater distinctions between the predictive probabilities of a hospital admission associated with the final three positions, which range from 0.091 to 0.265. In contrast to this, the MAP BN, which only considers the number of life events as influential, predicts probabilities between 0.120 and 0.257. The results of the CEG support the conclusions of Fergusson et al. [1986] that the effect of life events on admissions is strongest. However, the CEG further explains explicitly the way in which the social background and the economic situation may have an additional effect on hospital admissions.

From Table 3.3 the score of the MAP CEG structure is calculated to be -2478.490. In contrast to this, the score of the MAP BN which is -2489.776, giving a log-Bayes Factor or 11.284 or, equivalently, a Bayes Factor of about $80,000$ in favour of the CEG. By Table 3.1 this strongly suggests that the more flexible CEG model is highly preferable. Thus, although only a small number of asymmetries are present in this representation, the effect on the model score is vast and the added detail within the interpretation and the inference drawn is substantial.

I have so far only considered finding the MAP CEG with respect to a particular ordering of the variables and therefore a particular tree structure. A search across all possible permutations of the ordering of the variables in the tree would, be possible in order to find the MAP CEG given any ordering. Nevertheless, it seems plausible to keep the social background at the start of the tree due to the logical time ordering of the variables. The hospital admissions could be switched with the economic situation and the number of life events, as the variables are measured across the same years and an admission could affect, for example, the financial situation of the family. However, this analysis focuses on the effect of the three covariates on the hospital admission. I therefore only consider switching the economic situation and the number of life events in the ordering of the tree. Doing so a MAP CEG structure with the same final three positions, describing the same three health states as w_6, w_7 and w_8 in Figure 3.3, is obtained. As only two variables have been switched, only the effect of the life events on the economic situation is novel, while the overall conclusions on hospital admissions remain the same. The new CEG structure suggests that offspring from a high social background with a low or average number of life events have the same distribution for the economic situation. Also, offspring from a high social background with a high number of life events or offspring from a low social background with a low number of life events are in one stage and, similarly, families from a low social background with an average or high number of life events are in the same stage. In comparison to the MAP CEG of Figure 3.3 this CEG only

scores very slightly less, with a Bayes Factor of 1.27 in favour of the previous CEG.

3.3.3 The Ordinal MAP CEG and a Causal Intervention on the MAP CEG

The predictive probabilities attached to the final three positions in Figure 3.3 suggest that the CEG is written as an Ordinal CEG with respect to the probability of no hospital admission. From Table 3.4 the predictive admission probabilities for the remaining positions can be calculated: The probability of being admitted to hospital given the individual reaches w_1 is 15.5% and 24.0% given the individual reaches w_2. Similarly, the probabilities of hospital admission are 15.5% for w_3, 19.4% for w_4 and 24.6% for w_5. Even, without these calcuations, the Ordinal CEG lets us read directly from its graph that a hospital admission is less likely for a high social background than for a low social background, and similarly, that a hospital admission is less likely for an individual from a high economic situation and a low social background than for an individual with a low economic situation and a low social background.

The introduction of the Ordinal CEG in Chapter 2.5 suggests a brief comparison to logistic regression, the standard modelling technique for binary outcome variables. In the CHDS I have defined the hospital admissions variable to be the outcome variable by which the positions in the Ordinal CEG are ordered. In a corresponding regression analysis I would therefore let the hospital admission variable be the response variable and the social background, the economic situation and the life events the covariates. When carrying out a logistic regression all possible two-way and three-way interaction terms would need to be included to be able to make inference on the combined effect of the covariates on the outcome. Given the parameter estimates of the regression model the estimated effect of each combination of covariates could then be calculated and compared. The CEG, however, illustrates directly which combination of covariates have similar effects on the admission probability from the topology of its graph. The use of the BD metric to score models automatically determines when the effects of two different combinations of covariates can be interpreted as being the same (the vertices are in the same stage) taking into account the complexity of the model and the number of counts in each category, while in a regression analysis, the decision to merge different combinations would be decided by the investigation of the regression coefficients. Finally, the CEG also allows conclusions about the effect of the social background on the economic situation and the effect of social background and economic situation on life events and combines all this information within a single graph.

As introduced in Chapter 2 in applications like these we are often interested

in the causal effect of the covariates on the hospital admission. For example, if it were possible to intervene on the number of life events and enact a policy ensuring that the life events on a particular unit would always lie in the 'low' category, then it could be concluded from the CEG in Figure 3.3 that the consequent probability of hospital admission would be reduced to 9.1% for a high social background and 17.7% for a low social background. However, these types of interventions may also be asymmetric. Let us consider the intervention discussed in Chapter 2.3.3, where an intervention at position w_2 in Figure 3.3 forces all individuals from a low social background along the edge $e(w_2, w_4)$, indicating that individuals from a low social background and with a low economic situation are to be given financial aid. The resulting manipulated CEG was given in Figure 2.6 and is repeated together with its CPVs (Table 3.5) below.

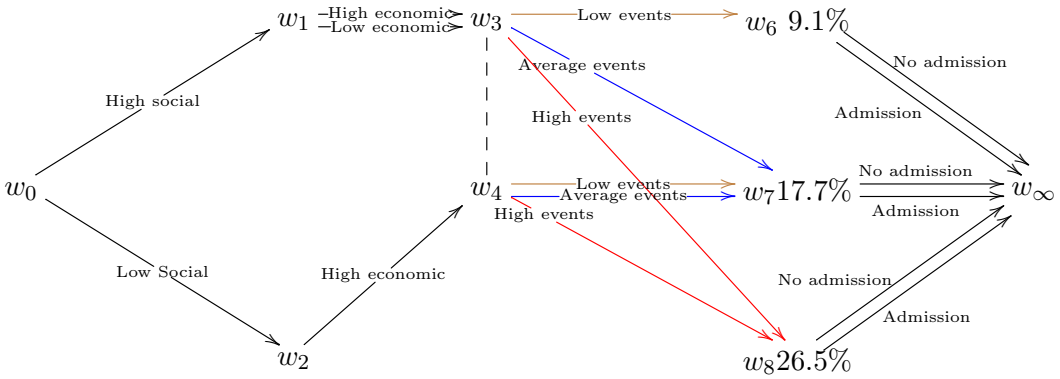


Figure 3.4: The MAP CEG for the CHDS example from Figure 3.3 manipulated such that individuals from a low social background are given financial aid

Stage/Position	Conditional Probability Vector	
$u_0 = w_0$	$(P(X_1 = \text{High}), P(X_1 = \text{Low}))$	$(0.569, 0.431)$
$u_1 = w_1$	$(P(X_2 = \text{High}\|u_1), P(X_2 = \text{Low}\|u_1))$	$(0.468, 0.532)$
$u_2 = w_2$	$P(X_2 = \text{High}\|u_2)$	1
$u_3 = \{w_3, w_4\}$	$(P(X_3 = \text{Low}\|u_3), P(X_3 = \text{Average}\|u_3), P(X_3 = \text{High}\|u_3))$	$(0.460, 0.344, 0.196)$
$u_5 = w_6$	$(P(X_4 = \text{No admission}\|u_6), P(X_4 = \text{Admission}\|u_6))$	$(0.909, 0.091)$
$u_6 = w_7$	$(P(X_4 = \text{No admission}\|u_7), P(X_4 = \text{Admission}\|u_7))$	$(0.823, 0.177)$
$u_7 = w_8$	$(P(X_4 = \text{No admission}\|u_8), P(X_4 = \text{Admission}\|u_8))$	$(0.735, 0.265)$

Table 3.5: Table of CPVs associated with the manipulated CEG of the CHDS example given in Figure 3.4

It can be read directly from the manipulated CEG that the predictive probability of hospital admissions for families from a low social background and an average number of life events given this intervention is improved to 17.7%. Further, Table 3.2 shows that this intervention reduces the predictive probability of a high number of life events in socially disadvantaged families from 46.9% to 19.6%, giving an improvement in the probability of admissions mediated through the life events. The probability of an individual reaching w_8 can then be calculated to reduce from 41.7% to 19.6% and therefore the overall probability of admission reduces from 19.1% to 17.2%.

In this Chapter I have reviewed the learning of the parameters and model selection using the Bayesian Dirichlet scoring function for BNs and CEGs. I have shown that through an application to the CHDS example that, whilst a BN search can be useful in providing a graphical framework for feeding back the analysis to a client, the CEG provides useful additional conclusions and refinements to the BN. This is not only apparent in the significantly high Bayes Factor of the derived CEG in comparison to the MAP BN, but also in its expressiveness to the client.

Chapter 4

Modelling Missingness using Chain Event Graphs

In this Chapter I will discuss how the CEG provides a new way of systematically exploring the effect of missing covariate data within a study and hence enables us to draw informative conclusions about the type of missingness and its influence. I will first review the three well-known types of missing data: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) [Rubin, 1976] and will demonstrate on a subset of the UKCP study, introduced in Chapter 1.2.2, how these different types of missing data can be represented explicitly within the topology of the CEG. In particular, the graph of the Ordinal CEG enables us to obtain a precise understanding of the subtleties associated with the three common types of missingness and differentiate further between more refined MNAR structures. I will then apply the model selection techniques developed in Chapter 3.2 to several examples of the UKCP and consequently show how the graph of the CEG helps us understand the influence of missingness, as well us allowing us to investigate the plausibility of the MAR assumption within these models. In the final section I will demonstrate how we can use the CEG to define new informative categories of variables in the UKCP study. As some of the examples in this chapter have slightly larger CEG structures, I will further demonstrate the use of the Reduced Ordinal CEG in these examples.

4.1 Introduction to Analysing Missing Data Structures

In many situations the full data set of a given problem may not be observed. Problems caused by missingness can be especially acute in cohort studies when it is

typical for substantial amounts of data to be missing over certain periods of time. Reasons for missing values may be, for example, non-response due to the individual refusing to disclose information, dropout, migration, or simply loss of data. Also, in the UKCP study introduced in Chapter 1.2 measurements for impairments may be missing for more informative reasons: For example, the impairment may be too severe to be measured or the individual may have died before a measurement could be taken. It is common practice to partition missing data mechanisms into three categories, which are Missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) as proposed in Rubin [1976]. I here briefly review these three missingness mechanisms by looking at the simplest case where we have two variables: X_1, which is fully observed and X_2 which has missing values. Also, let R_2 be a missingness indicator, which is equal to 1 when X_2 is observed and equal to 0 when X_2 is missing. Little and Rubin [2002] then define that data are MCAR when missingness does not depend on the observed and unobserved values, and this can be written as

$$p(r_2 \mid x_1, x_2) = p(r_2).$$

In line with the graphical models introduced in the previous two chapters, this can be rewritten as the conditional independence statement:

$$R_2 \perp\!\!\!\perp X_1, X_2. \tag{4.1}$$

In the UKCP study let X_1 describe the birth weight, which is assumed to be fully observed, and let X_2 be the visual impairment. From Table 1.2 it is known that X_2 has large amounts of missing data, so let R_2 be the indicator distinguishing whether the visual impairment is missing or not. Then, assuming that X_1 is fully observed, MCAR would imply that the reasons for missingness are independent of the visual impairment and of the birth weight. This assumption is, however, often not plausible in a study like this. A slightly less restrictive assumption defined in Little and Rubin [2002] states that missingness only depends on the components of the data that are observed and not on the components that are missing and we then say data are MAR. The standard notation introduced in Little and Rubin [2002] and Schafer [1997] makes it difficult to write the MAR assumption in terms of conditional independencies. However, examples of data that are MAR in the above references and in Daniels and Hogan [2008] show that MAR corresponds to

the conditional independency:

$$R_2 \perp\!\!\!\perp X_2 \mid X_1, \tag{4.2}$$

or equivalently to

$$p(r_2 \mid x_1, x_2) = p(r_2 \mid x_1).$$

In our example on the cerebral palsy cohort this would mean that the missingness indicator of the visual impairment is independent of the visual impairment itself but dependent on the birth weight. Again, in studies like these, even the MAR assumption may not be a plausible assumption to make. For example, an impairment may be so severe that it is impossible to measure and therefore it is missing. In such a scenario the missingness indicator will also depend on the impairment itself and hence the MAR assumption would be violated. Little and Rubin [2002] then say that data are MNAR, where the missingness mechanism depends on the observed and unobserved values, which can be written as

$$R_2 \not\perp\!\!\!\perp X_1, X_2, \tag{4.3}$$

or, equivalently,

$$p(r_2 \mid x_1, x_2) = p(r_2 \mid x_1, x_2).$$

Various methods for addressing inference when data are missing have been developed [Schafer, 1997; Little and Rubin, 2002], such as complete-case or available-case analysis and single or multiple imputation. Research has centred around circumstances when it is appropriate to assume that data are MAR. It has been shown that in this case it is possible to use efficient computational methods, for example, the Expectation-Maximisation algorithm or MCMC methods [Little and Rubin, 2002; Heckerman, 2008] to find Maximum Likelihood estimates, or in a Bayesian setting, the posterior distribution of the parameters of interest. However, in many situations the MAR assumptions are not plausible. As such methods then seriously bias inferences, as demonstrated for example in Sterne et al. [2009], the missingness process has to be modelled explicitly to avoid bias.

One method for analysing incomplete data of categorical variables is to treat missingness as an additional category for each variable that has missing values. This is, however, not always appropriate. Winship et al. [2002] discuss that this is not always appropriate. For example, two variables X and Y may be marginally dependent, but independent conditional on a binary variable Z taking values 0 or 1. It is assumed that Z is, however, only partially observed and the missing values of

Z are treated as a separate category. Then the missingness category of Z may not detect the conditional independence of X and Y as it contains cases with $Z = 0$ and cases with $Z = 1$ and estimates of the dependency between X and Y conditional on Z may be biased. However, in other situations this approach seems appropriate when missingness of an observation can be hypothesised as an informative measurement of the development of that individual in an unfolding process. For example, as mentioned earlier, in the UKCP, we know that missingness of an impairment is associated with the impairment being impossible to measure due to severity or early death. This type of hypothesis is represented well using a tree or a CEG, where missingness is expressed either by an extra edge or an extra branch in the graph and represents the missingness indicator explicitly. The issue discussed in Winship et al. [2002] could then be overcome, as context-specific independencies in the CEG would demonstrate that X and Y are independent conditional on the observed values of Z.

4.2 CEGs for Informed Missingness

I will demonstrate in this section how the CEG can be used to classify different types of missingness by applying it to a subset of the UKCP study. Let X_1 describe the birth weight, X_2 the visual impairment and X_3 the variable describing the survival to age 5 or above. Also, recall that X_2 has a large amount of missing values, and let R_2 be the variable indicating whether X_2 is missing or not. The corresponding event tree of this problem is given in Figure 4.1. Here, survival is the variable of interest and represents the final variable in the tree as we are interested in the effect of the other two covariates on the probability of survival. Birth weight is introduced first, while the impairment, which is measured later, is introduced second, giving the ordering of the variables: (X_1, R_2, X_2, X_3). As X_3 is binary the CEG can be written as an Ordinal CEG with respect to the probability of survival and thus enhance the expressiveness of the graph.

I will first illustrate on this example how different missingness structures can be represented by an Ordinal CEG and show that, when data are assumed to be MAR or MCAR, a particular set of CEG structures is observed which describe the randomness of the missingness mechanism. However, when data are MNAR, then the CEG structures can be used to distinguish between hypotheses about different types of MNAR mechanisms. I will then apply the methods to the MAP CEG structure of the given example.

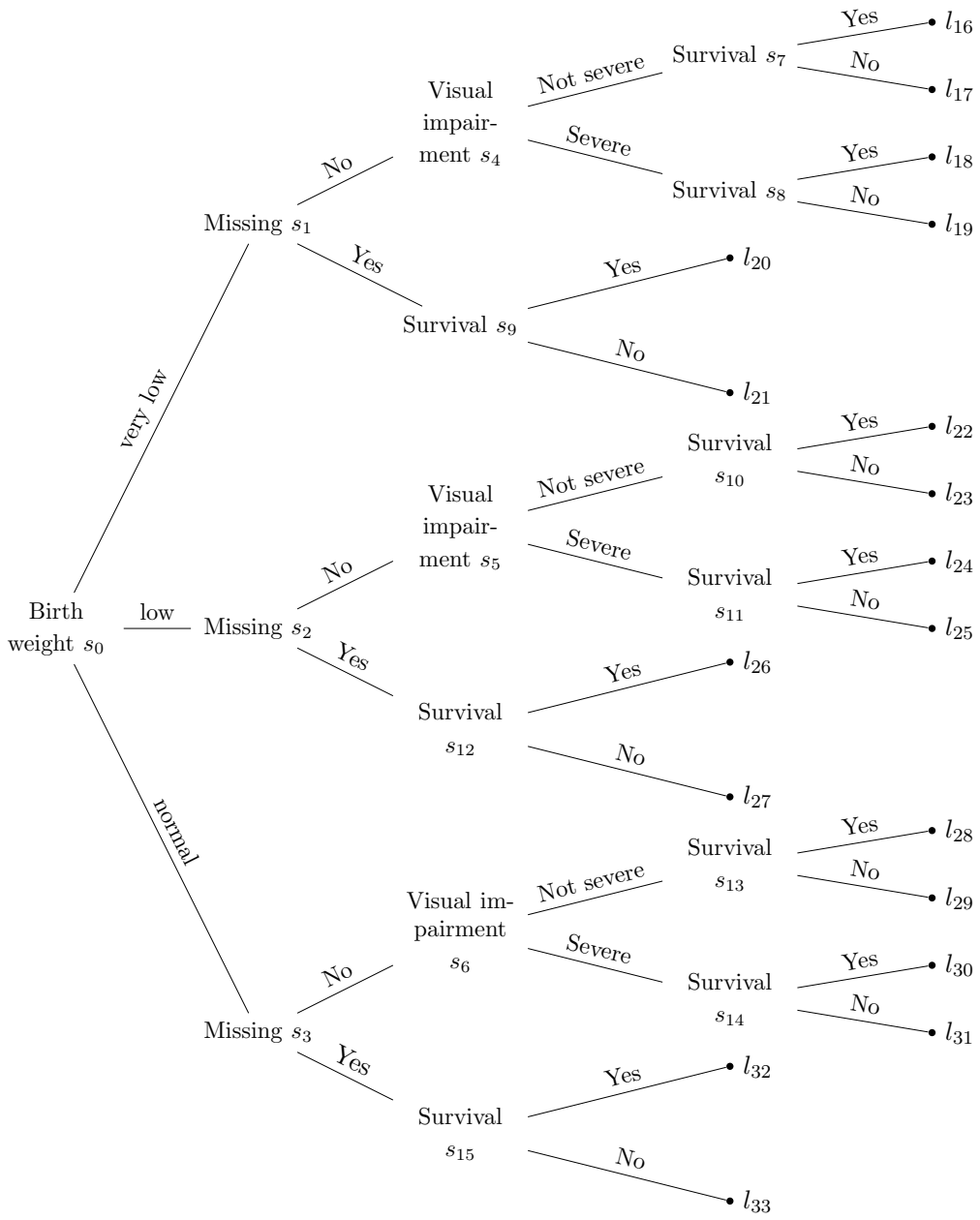


Figure 4.1: Event Tree for the UKCP example describing the effect of birth weight, visual impairment and missingness on survival

4.2.1 CEG Representation of MCAR, MAR and MNAR

In the literature, to model the joint density of X_2 and the missingness process R_2 it is common to distinguish between two model families, selection models and pattern-mixture models. These are based on two different factorisations of the joint density of the variable, X_2 and R_2 [Little and Rubin, 2002]. A selection model factorises the joint density into the marginal density of X_2 and the missingness process R_2 conditional on X_2, while in a pattern-mixture model the opposite factorisation is used, where first the marginal density of the missingness process is modelled and then the density of the variable X_2 given the missingness process. In the above tree structure the ordering where R_2 appears before X_2 has been chosen, which is the ordering implied by the pattern-mixture model. Alternatively, I could have picked the ordering where X_2 appears before R_2 which is implied by the selection model. Which of these two orderings is more convincing is dependent on context: For example, if data had already been collected and then some of the data lost, then the ordering (X_2, R_2) would be more natural. If someone from the cohort left the study early before any outcome variable could be measured then the ordering (R_2, X_2) would be more plausible. The MAR statement $R_2 \perp\!\!\!\perp X_2 \mid X_1$ in 4.2 implicitly takes the variables in the order (X_1, X_2, R_2). In this case the argument would be that the variables X_1 and X_2 exist for each individual a priori, however these variables might not be recorded for X_2 for various reasons. However, the assumption of MAR is equivalent to the assumption $X_2 \perp\!\!\!\perp R_2 | X_1$. This reinterprets MAR in terms of viewing data as if it were consistent with the order (X_1, R_2, X_2), as in the tree in Figure 4.1. Either way, by the semi-graphoid axioms, the MCAR and MAR assumptions are equivalent under either ordering. I hence choose the second ordering which allows models that violate the MAR assumption to still be estimated and where I assume that X_1 and the missingness indicator R_2 can be seen as measurements of events happening that might influence X_2. Hence, unlike standard representations of MCAR, MAR and MNAR, the CEG suggests to first decide explicitly a plausible ordering of the variables within the given context and retains this within the structure of its graph.

In this example, apart from birth weight (X_1), and visual impairment (X_2) a further outcome variable, X_3, describing survival up to or above 5 exists, which is fully observed. Due to follow-up using hospital records and national death and emigration records, it is often reasonable to assume that the outcome variable is fully observed even when the individual has dropped out throughout the study. In

this case the MCAR/MAR assumptions commonly further include that

$$X_3 \perp\!\!\!\perp R_2 \mid X_1, X_2,$$

i.e. that the probability of survival is independent of the missingness process given X_1 and X_2 [Hemming and Hutton, 2012]. Under this assumption the following can be deduced:

$$\begin{aligned} p(x_3|x_1, r_2 = 0) &= \sum_{x_2} p(x_3|x_1, r_2 = 0, x_2) p(x_2|x_1, r_2 = 0) \\ \text{under MAR} &= \sum_{x_2} p(x_3|x_1, r_2 = 1, x_2) p(x_2|x_1, r_2 = 1). \end{aligned} \tag{4.4}$$

So, under MAR, $p(x_3|x_1, r_2 = 0)$ is given by a weighted average of the probability of survival given birth weight and visual impairment, which is observed, weighted according to the probability of visual impairment given it is observed and given a particular birth weight. The probability of survival given missing visual impairment will therefore lie between the probability of survival for a non-severe and a severe impairment. So, with respect to our example and the associated tree in Figure 4.1, under the assumption of MAR, s_9 is expected to be in a position whose posterior predictive probability of survival (posterior CPV) is a weighted average of the predictive probabilities of survival for severe and non-severe visual impairment given a very low birth weight, weighted according to the probability of a severe or non-severe impairment, given the impairment is observed. In an Ordinal CEG representation, the position describing survival for the missing category given a very low birth weight, is expected to lie between the positions for survival of individuals with severe or non-severe visual impairment given a very low birth weight. The same holds for the vertices s_{12} and s_{15}. A possible Ordinal CEG structure under the MAR assumption would therefore be the one given in Figure 4.2 (Note: birth weight = bw).

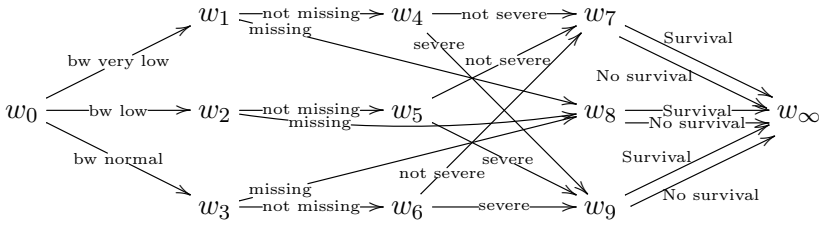


Figure 4.2: Ordinal CEG for the UKCP example describing the effect of birth weight, visual impairment and missingness on survival when data are MAR

For simplicity, I have assumed that survival is hypothesised to be independent of the birth weight given the visual ability and I will assume that this holds throughout this subsection. If this were not the case then the positions w_7, w_8 and w_9 might be split into several positions depending on the birth weight. As the CEG is said to be an Ordinal CEG the graph also states that a very low birth weight leads to the highest survival, followed by a low and normal birth weight [Hutton and Pharoah, 2002].

However, the graphical observation of Figure 4.2 is only a necessary but not sufficient condition. Data may produce a CEG structure as in Figure 4.2 and hence appear to be MAR. However, the predictive probability $p(x_3|x_1, r_2 = 0)$ in Equation 4.4 may nevertheless differ significantly from

$$\sum_{x_2} p(x_3|x_1, r_2 = 1, x_2)p(x_2|x_1, r_2 = 1)$$

such that data are unlikely to be MAR. This will be case when the disability categories are very imbalanced. For example, when a large proportion of individuals has a non-severe impairment, i.e. $x_2 = 0$, then

$$p(x_2 = 0|x_1, r_2 = 1) \approx 1 \text{ while } p(x_2 = 1|x_1, r_2 = 1) \approx 0,$$

and hence the right hand side of Equation 4.4

$$\sum_{x_2} p(x_3|x_1, r_2 = 1, x_2)p(x_2|x_1, r_2 = 1) \approx p(x_3|x_1, r_2 = 1, x_2 = 0),$$

So, the above sum may be close to the predictive probability of individuals with a non-severe disability, i.e. close to the predictive probability associate with position w_7, but the predictive probability $p(x_3|x_1, r_2 = 0)$ associated with position w_8 may not be close to this value. We can consequently only deduce from the graph when data are MNAR but we need further calculations to deduce that they are MAR. We can do so by looking at the CPVs of the CEG and simply calculate from these the left hand side and right hand side of Equation 4.4 and compare these. The graph on its own nevertheless gives an indication of the possibility that the MAR assumption holds. I will give an example of this in the next subsection, in which I will carry out model selection on the UKCP example illustrated here.

From equation 4.1 that, if data are MCAR, then $R_2 \perp\!\!\!\perp X_1$ is additionally requires on top of the MAR assumption. This independence statement can be directly deduced from the topology of the CEG, as in this case all positions associated with

the missingness indicator R_2 will be in the same stage, such that the probability of having a missing value is indistinguishable across the birth weight, X_1. In this example this means that w_1, w_2 and w_3 are in the same stage, as represented in Figure 4.3.

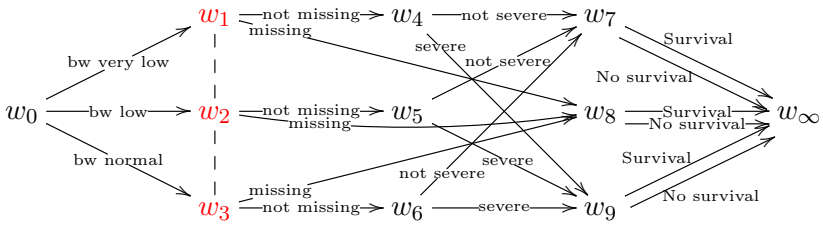


Figure 4.3: Ordinal CEG for the UKCP example describing the effect of birth weight, visual impairment and missingness on survival when data are MCAR

When data are MNAR then the missingness process depends on both the observed and the unobserved values such that R_2 depends on both X_1 and X_2. A simple case for MNAR occurs when all situations describing survival, given that missingness has occurred, are in positions with a lower survival probability than when the visual impairment is observed. In the UKCP example this means that s_9, s_{12} and s_{15} in Figure 4.1 are in lower positions than $s_7, s_8, s_{10}, s_{11}, s_{13}$ and s_{14}. Hence a missing value predicts that the visual impairment is likely to be even worse than the usual visual impairment which is classed as 'severe' and hence is associated with poorer survival. This is represented by the Ordinal CEG structure given in Figure 4.4. In this scenario it is possible deduce directly from the graph alone that missingness is unlikely to be MAR, as all edges labelled 'missing' lead to a lower position of survival than the edges labelled 'severe' or 'non-severe'.

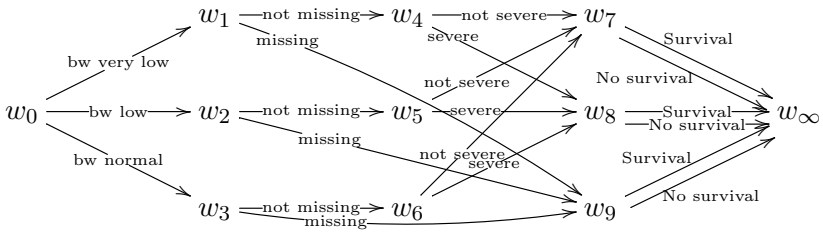


Figure 4.4: Ordinal CEG for the UKCP example describing the effect of birth weight, visual impairment and missingness on survival when data are MNAR

However, the Ordinal CEG can also distinguish between different types of MNAR data. Hence, alternatively, data may be MNAR conditional only on certain values of another variable. For example, data may be MAR given that the birth weight is very low or low but MNAR when birth weight is normal. This hypothesis

is represented by the Ordinal CEG with the structure given in Figure 4.5. The MNAR conclusion can be deduced directly from the graph as in Figure 4.4, while the MAR conclusions given very low and low birth weight would need to be checked as described above.

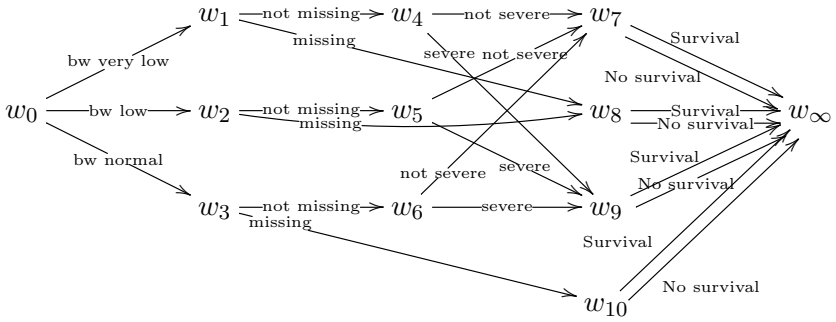


Figure 4.5: Ordinal CEG for the UKCP example describing the effect of birth weight, visual impairment and missingness on survival when data are MNAR conditional on birth weight

However, the topology of the Ordinal CEG is able to provide information on the strength of the influence due to missingness. Consider, the CEG in Figure 4.6. Here an individual with missing visual disability is in the same position as an individual whose visual impairment is classed as 'severe'. Unless the proportion of individuals with severe impairments is very large, such that $p(x_2 = 1|x_1, r_2 = 1) \approx 1$, so that under MAR $p(x_3|x_1, r_2 = 0) \approx p(x_3|x_1, r_2 = 1, x_2 = 1)$ by Equation 4.4, data are likely to be MNAR. When comparing this to Figure 4.4, it is apparent that the missing category has a stronger effect on survival in Figure 4.4 than in Figure 4.6.

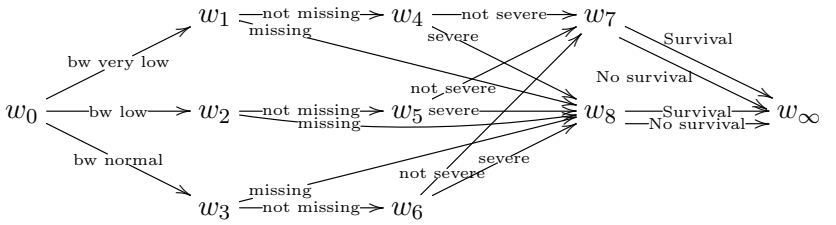


Figure 4.6: Ordinal CEG for the UKCP example describing the effect of birth weight, visual impairment and missingness on survival when data are MNAR, with missingness having a similar effect as a severe impairment

Finally, the opposite effect of missingness could also be hypothesised from the graph, where the survival probability given that visual impairment is missing is in the position with the highest probability of survival. An example of this is illustrated

in Figure 4.7. In this situation data are again MNAR, but now the conclusion made would be that missingness occurs only when the visual impairment is non-severe. Of course, an expert may deem such a CEG and associated hypotheses implausible. This is however simple to address within the Bayesian methodology: As discussed in Chapter 3.1 models considered implausible by the expert could simply be excluded from the search space, or alternatively assigned small prior probabilities.

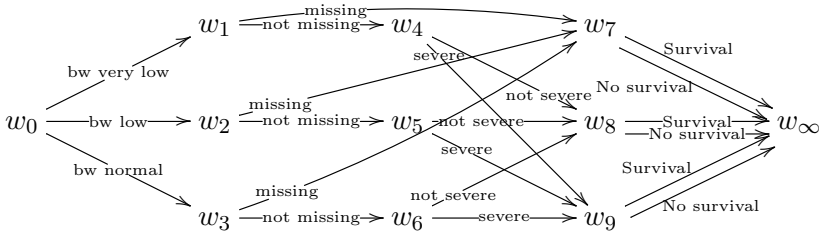


Figure 4.7: Ordinal CEG for the UKCP example describing the effect of birth weight, visual impairment and missingness on survival when data are MNAR, but missingness has a positive effect on survival

Note that in this example missingness does not occur for the outcome variable. Nevertheless, this could also be incorporated into the CEG structure. In this case the outcome variable would have three categories, 'survival', 'no survival' and 'missing', and the predictive probabilities of survival and of missing survival could be added to the graph. If we were interested, not only in survival up to or above 5, but in several survival categories, such as age $0-5, 5-10, 10-15$ and >15, the situations describing survival in the associated tree would have four categories. The MAP CEG structure could then be found on this tree and a Kaplan-Meyer plot to each final position could be added.

I have shown that the graph of the CEG allows for a direct analysis of the reasons of missingness and gives an explicit representation of the different types of missingness mechanisms. I have also illustrated that the Ordinal CEG can distinguish between different types of MNAR and the way in which this is made explicit in the graph.

4.2.2 Application to the UKCP

In this section I will find the MAP CEG structure for the running example, used in the previous section, given the available data from the UKCP study. The resulting model can then be used to draw inference on the effect of the birth weight and visual impairment on survival and gives an understanding of the missingness structures beyond the three established mechanisms. As I have ensured that the missingness

indicator appears before the variable with missing values within the ordering of the variables in the tree, the model selection techniques of Chapter 3.2 can be directly applied to this example and the scores of the CEGs in the model space can be calculated in closed form as before.

As in Chapter 3.2 I assume a uniform prior on the root-to-leaf paths and an equivalent sample size of 3, equal to the number of categories the birth weight variable takes. A discussion of possible informative priors for the examples of the UKCP study are discussed in Chapter 6.2. Running the AHC algorithm finds the MAP CEG to be the CEG given in Figure 4.8 with the CPVs given in Table 4.1. The predictive probabilities of survival up to or above the age of 5 are attached to the final positions in the CEG, which, together with the 95% credible intervals of the posterior distribution of survival, are: 98.7 $(98.4, 99.0)$% for position w_5, 89.5 $(87.5, 91.3)$% for w_6 and 84.6 $(81.8, 87.2)$% for w_7. The CEG is again drawn as an Ordinal CEG such that the positions describing the same succeeding event are vertically aligned in descending order with respect to the predictive probability of survival. To calculate the predictive probability of survival for positions w_1, w_2 and w_3 Table 4.1 can then be used to obtain a survival probability of 96.5% for w_1, 96.2% for w_2 and 95.3% for w_3. So, from the topology of the Ordinal CEG a low birth weight is predicted to give the highest probability of survival and a slightly lower probability for a very low and a normal birth weight.

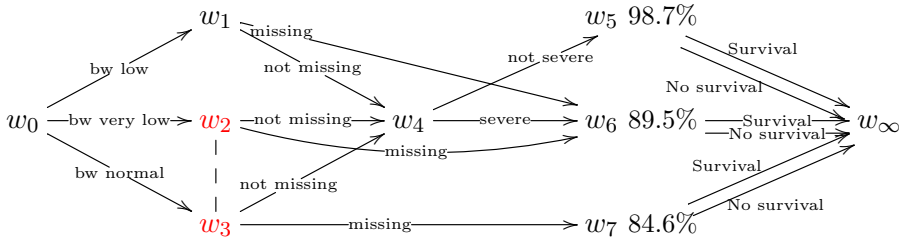


Figure 4.8: Ordinal MAP CEG structure for the UKCP example describing the effect of birth weight, visual impairment and missingness on survival

Stage/Position	Conditional Probability Vector	
$u_0 = w_0$	$(P(X_1 =\text{Low}), P(X_1 =\text{Very low}), P(X_1 =\text{Normal}))$	$(0.244, 0.177, 0.579)$
$u_1 = w_1$	$(P(X_2 =\text{Not missing}\|u_1), P(X_2 =\text{Missing}\|u_1))$	$(0.850, 0.150)$
$u_2 = \{w_2, w_3\}$	$(P(X_2 =\text{Not missing}\|u_2), P(X_2 =\text{Missing}\|u_2))$	$(0.814, 0.186)$
$u_3 = w_4$	$(P(X_3 =\text{Not severe}\|u_3), P(X_3 =\text{Severe}\|u_3))$	$(0.894, 0.106)$
$u_4 = w_5$	$(P(X_3 =\text{Survival}\|u_4), P(X_3 =\text{No survival}\|u_4))$	$(0.987, 0.013)$
$u_5 = w_6$	$(P(X_4 =\text{Survival}\|u_5), P(X_4 =\text{No survival}\|u_5))$	$(0.895, 0.105)$
$u_6 = w_7$	$(P(X_4 =\text{Survival}\|u_6), P(X_4 =\text{No survival}\|u_6))$	$(0.846, 0.154)$

Table 4.1: Table of CPVs associated with the MAP CEG for the UKCP example on birth weight, visual impairment and survival given in Figure 4.8

As illustrated in Chapter 2.3.2 on the CHDS example, a number of conclusions can be drawn from the CEG about the likely dependence structure of the three variables considered. The distribution of the missingness is indistinguishable for a very low and normal birth weight as w_2 and w_3 are in the same stage. Further, recall the conditional independence statement of Chapter 2.3 Equation 2.8, which states that $Y(w) \perp\!\!\!\perp X(w) \,|\, \mathcal{E}(w)$, where $Y(w)$ is a variable identified with the set of paths from w_0 to w, $X(w)$ is the variable associated with the edges emanating from w and $\mathcal{E}(w)$ represents the event that the individual passes through the position w. This can be used as before to read off conditional independencies from the CEG associated with the variables in the graph by looking at the cut-sets of the graph associated with each of the edges emanating from the vertex subsets V_{R_2}, V_{X_2} and V_{X_3}. The first cut-set in the graph consists of the edges emanating from positions w_1, w_2 and w_3. These are reached by three unique paths and hence $Y(w) \perp\!\!\!\perp X(w) \,|\, \mathcal{E}(w)$ applied to w_1, w_2 and w_3 gives the trivial conditional independence statement that the birth weight affects the missingness process. Moving further along the graph we can deduce the conditional independency $Y(w_4) \perp\!\!\!\perp Z(w_4) \,|\, \mathcal{E}(w_4)$, where by Equation 2.7 $Z(w_4)$ is the variable associated with the paths from w_4 to w_∞. Here the event, $\mathcal{E}(w_4)$, i.e. going through w_4 corresponds to observing visual ability. $Z(w_4)$ describes visual impairment and survival, while $Y(w_4)$ represents the birth weight. We then have that, given visual impairment is observed, the visual impairment and survival are independent of birth weight, such that, when visual disability is observed, we have that the distribution of visual disability and survival is the same ($w_1 \to w_4$, $w_2 \to w_4, w_3 \to w_4$). Finally, consider the three final positions, w_5, w_6 and w_7, which can be interpreted as describing the 'health state' of the individual. We have that $X(w_5)$, $X(w_6)$ and $X(w_7)$ describe the variable survival and then, from $Y(w) \perp\!\!\!\perp X(w) \,|\, \mathcal{E}(w)$ applied to w_5, w_6 and w_7 we conclude that survival depends only on these three positions and not on the paths through which they have been reached.

As expected, the highest probability of survival is obtained when visual impairment is observed to be non-severe. In this case survival up to or above 5 is predicted to be 98.7%. When visual impairment is observed to be severe, the individual is forced into the final position w_6 with survival of 89.5%, which is significantly lower than survival with a non-severe disability. The poorest survival is found to be for individuals whose visual impairment is not observed. Here a very low and low birth weight leads to a survival probability equal to the predictive survival probability for severe impairment, while for a normal birth weight survival is predicted to be only 84.6%. This is significantly lower than survival when visual disability is observed.

We hence deduce directly from the Ordinal CEG structure that the data are unlikely to be MAR. As explained in the previous section the expected survival probabilities, under MAR, for individuals for whom visual disability is missing, can be calculated from the right hand side of Equation 4.4, where the survival probability conditional on a particular birth weight is expected to be the weighted average of the survival probability for individuals of that birth weight with a severe or non-severe disability, weighted according to the probability of observing a severe or non-severe visual impairment. In Figure 4.8 all individuals go through position w_4 and therefore the calculated expected probability of survival under MAR will be the same for all individuals. This is given by

$$98.7 \times 0.894 + 89.5 \times 0.106 = 97.7\%,$$

with 95% credible interval $(96.8\%, 98.5\%)$. This is compared to the predictive survival probabilities when visual impairment is missing, which correspond to the predictive survival probabiltiies associated with the positions w_6 and w_7, of 89.5% and 84.6%. We see that the predictive survival for a missing impairment is much lower and, in either case, does not lie within the calculated 95% credible interval. The conclusion is therefore that the data are unlikely to be MAR. In the situation where the individual has a normal birth weight this can be read off directly from the Ordinal CEG. For a very low or low birth weight, the missing edge leads to the same position as severe visual disability with survival probability 89.5%. Figure 4.8 suggests that the data are not MAR, however this needed to be calculated explicitly to make reliable conclusions. (Compare Figure 4.6).

Having found the MAP CEG structure for the tree given in Figure 4.1 the hypothesis that data are MCAR can also be examined. The first requirement for this is that w_1, w_2 and w_3 are in the same stage, which suggests that there is no evidence that missingness is dependent on the birth weight of the individual, such that $R_2 \perp\!\!\!\perp X_1$. However, this is only the case for w_2 and w_3 but not for w_1. The second requirement, that missingness is independent of visual disability ($R_2 \perp\!\!\!\perp X_2$) has also shown to be implausible by the above.

Note that again the equivalent sample size can be varied to check whether the selected model is sensitive towards the strength of the uniform prior. Doing so shows that the only observed change is that an individual with a very low birth weight and severe visual impairment also moves into position w_7 as the sample size increases. This happens as the uniform prior implies a survival probability of 50% a priori and this affects a reduction in the predictive probability of survival for

the relatively small number of individuals with a very low birth weight and severe impairment. The Ordinal CEG then proposes that data may be MAR for a very low birth weight, as now the position reached when the individual has a very low birth weight and a missing impairment lies between the other two position. However, calculations as the ones performed above and the close predictive probabilities of w_7 and w_6 in comparison to w_5 show that this is not the case.

4.3 CEGs for Informed Missingness for more than One Variable

When only one variable has missing values and all other variables are fully observed, then the MAR assumption, that the missingness indicator depends only on the observed variables, is defined as in Equation 4.2. However, when more than one variable has missing values and we have more than one missingness indicator, then the MAR definition according to Rubin [1976] is not well defined. Given that two variables are partially observed, one suggestion (see for example Lu and Copas [2004]) is to distinguish three different cases, according to which of the two variables or both are missing, and determine whether data are MAR separately for each case. In this section I will adopt this approach when looking at CEG structures where more than one variable has missing values. I will demonstrate that the CEG enables us to obtain a good understanding and a useful visual representation of the influence of missingness, even when more than one variable has missing values.

To illustrate this I extend the model space by including a further variable into the model describing ambulatory impairment, for which we distinguish between a severe or a non-severe impairment. I again choose an ordering where birth weight occurs first, followed by visual impairment and then ambulatory impairment, and finally survival up to or above the age of 5 appears last in the ordering as the variable of interest. The corresponding tree structure of this extended problem is given in Figure 4.9. I have now no longer explicitly included a missingness indicator in the tree but have instead simply added missingness as a separate category to the two variables describing visual and ambulatory impairment. Comparing this tree to the tree in Figure 4.1 the florets describing the visual impairment, s_5, s_6, and s_7, can be thought of as being merged with s_2, s_3 and s_4, which describe whether the visual impairment is missing or not, by removing the edges between s_2 and s_5, and similarly between s_3 and s_6, and between s_4 and s_7. Consequently, the resulting CEG can now no longer distinguish between MAR and MCAR. However, the same informative conclusions on the impact of missingness on survival can still

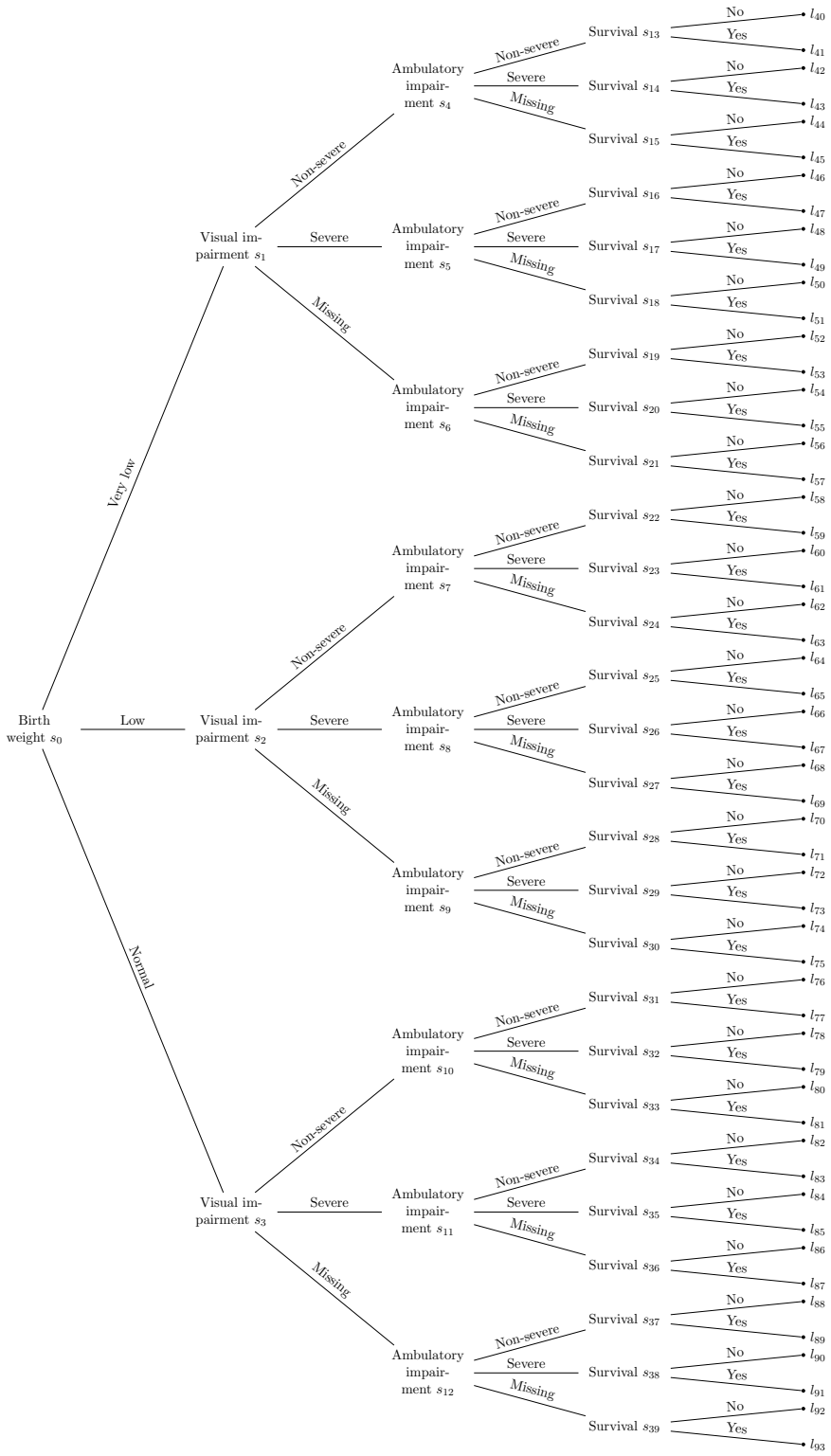


Figure 4.9: Event tree for the UKCP example describing the effect of birth weight, visual impairment, ambulatory impairment and missingness on survival

be deduced.

The most probable CEG structure for this problem can be found using the AHC algorithm described, where the priors are set up as in the previous section where $\alpha = 3$ and the paths of the three are assumed to be a priori equally likely. The resulting CEG found through the algorithm is given in Figure 4.10 with the associated table of CPVs given in Table 4.2. As before it is drawn as an Ordinal CEG, where here positions that are in the same stage are given the same colour, rather than their emanating edges. (Note Vis. = visual impairment, Amb. = ambulatory impairment, non-sev. = non-severe, sev. = severe and miss. = missing.)

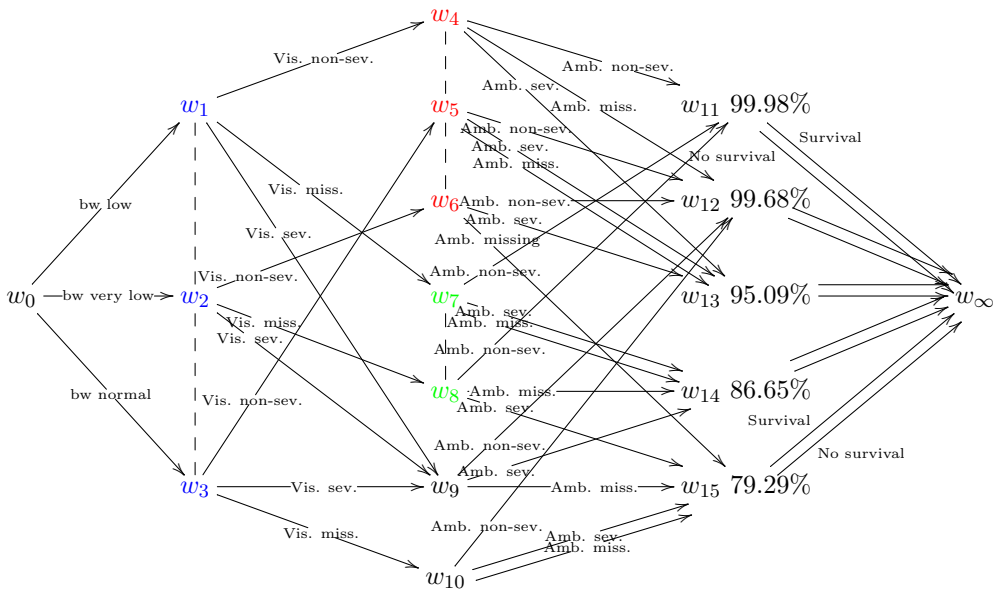


Figure 4.10: Ordinal MAP CEG structure for the UKCP example describing the effect of birth weight, visual impairment, ambulatory impairment and missingness on survival

Due to the relatively large tree structure the MAP CEG is complicated such that it cannot be easily read by a client. More explicitly, the situations s_1 to s_{12} are often only merged into stages but not positions. In this example the situations $s_4, s_5, \ldots, s_{12}$ are merged only into seven positions with 21 edges emerging from them leading into five final positions. It may be possible to simplify the structure slightly by reordering the variables: As birth weight appears to have less impact on survival than the two impairments, putting birth weight as the third variable in the tree may make the presentation slightly clearer. However, there would still be a large number of edges emanating from different positions. A better simplification of Figure 4.10 can be found by transforming the CEG into a Reduced Ordinal CEG, as introduced

Stage/Position	Conditional Probability Vector	
$u_0 = w_0$	$(P(X_1 = \text{Low}), P(X_1 = \text{Very low}), P(X_1 = \text{Normal}))$	$(0.2440, 0.1766, 0.5794)$
$u_1 = \{w_1, w_2, w_3\}$	$(P(X_2 = \text{Non-severe}\|u_1), P(X_2 = \text{Severe}\|u_1), P(X_2 = \text{Missing}\|u_1))$	$(0.7335, 0.0868, 0.1797)$
$u_2 = \{w_4, w_5, w_6\}$	$(P(X_3 = \text{Non-severe}\|u_2), P(X_3 = \text{Severe}\|u_2), P(X_3 = \text{Missing}\|u_2))$	$(0.7811, 0.2077, 0.0112)$
$u_3 = \{w_7, w_8\}$	$(P(X_3 = \text{Non-severe}\|u_3), P(X_3 = \text{Severe}\|u_3), P(X_3 = \text{Missing}\|u_3))$	$(0.3842, 0.3057, 0.3101)$
$u_4 = w_9$	$(P(X_3 = \text{Non-severe}\|u_4), P(X_3 = \text{Severe}\|u_4), P(X_3 = \text{Missing}\|u_4))$	$(0.1429, 0.8417, 0.0154)$
$u_5 = w_{10}$	$(P(X_3 = \text{Non-severe}\|u_5), P(X_3 = \text{Severe}\|u_5), P(X_3 = \text{Missing}\|u_5))$	$(0.2608, 0.3866, 0.3526)$
$u_6 = w_{11}$	$(P(X_4 = \text{Survival}\|u_6), P(X_4 = \text{No survival}\|u_6))$	$(0.9998, 0.0002)$
$u_7 = w_{12}$	$(P(X_4 = \text{Survival}\|u_7), P(X_4 = \text{No survival}\|u_7))$	$(0.9968, 0.0032)$
$u_8 = w_{13}$	$(P(X_4 = \text{Survival}\|u_8), P(X_4 = \text{No survival}\|u_8))$	$(0.9509, 0.0491)$
$u_9 = w_{14}$	$(P(X_4 = \text{Survival}\|u_9), P(X_4 = \text{No survival}\|u_9))$	$(0.8665, 0.1335)$
$u_{10} = w_{15}$	$(P(X_4 = \text{Survival}\|u_{10}), P(X_4 = \text{No survival}\|u_{10}))$	$(0.7929, 0.2071)$

Table 4.2: Table of CPVs associated with the MAP CEG for the UKCP example on birth weight, visual impairment, ambulatory impairment and survival given in Figure 4.10

briefly in Chapter 2.5. This is done by defining a new variable on 'number of severe disabilities' with the six categories: no impairment, one non-severe and one missing impairment, exactly one severe impairment, one severe and one missing impairment, two severe impairments and both impairments missing [Hutton et al., 1994]. The corresponding new illustration is given in Figure 4.11.

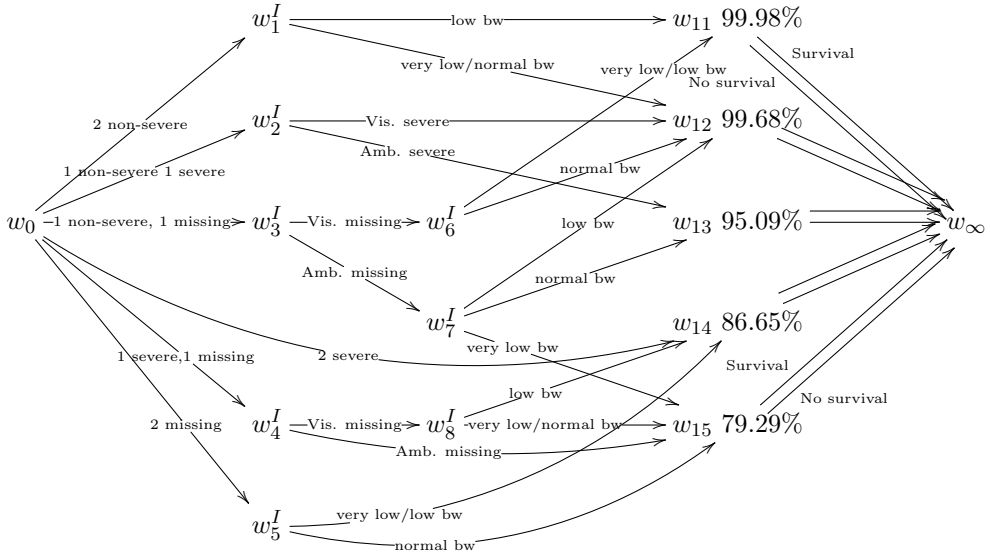


Figure 4.11: Reduced Ordinal MAP CEG structure for the UKCP example describing the effect of birth weight, visual impairment, ambulatory impairment and missingness on survival

The paths of the full Ordinal CEG up to the final positions are then redefined using this variable, which is described by the edges emanating from w_0, and birth weight and the type of impairments are only added into the graph as necessary. For example, all individuals with both impairments severe lead directly to a final

position in the graph, while in the other cases it is necessary to distinguish further between the type of impairment, the birth weight, or both. The intermediate steps between the new variable and the final positions are represented as positions denoted by w^I. As mentioned in Chapter 2.5 these positions are marked by an 'I', as they are not derived directly from a given tree structure and it may be possible that some of these positions are also in the same stage. Further, there are no 'double edges' emanating from these positions and their edges sometimes describe two different levels of a covariate (see, for example, $e(w_1^I, w_{12})$). The graphical representation up to the final positions has therefore so far only been developed to serve as a visual improvement, while further analysis would be carried out on the full CEG. However, the CPVs of the Reduced Ordinal CEG can be calculated from Table 4.2. Also, it would be possible to construct a new tree structure on the number of non-severe, severe or missing impairments with six categories, followed by the type of impairment and the birth weight and ending with survival as before. Then the full MAP CEG structure of this tree could be found to determine a possible coarser stage structure and a full CEG representation.

The percentages attached to the five final positions in Figures 4.10 and 4.11 give the posterior predictive percentages of survival given an individual reaches that position. We have the following percentages and 95% credible intervals: 99.98 $(99.87, 1.00)$% for position w_{11}, 99.68 $(99.45, 99.85)$% for w_{12}, 95.09 $(93.65, 96.35)$% for w_{13}, 86.65 $(83.99, 89.11)$% for w_{14} and 79.29 $(75.88, 82.51)$% for w_{15}. When both disabilities are not severe, we distinguish between a low birth weight and a very low or normal birth weight. However, the difference in the predictive probability of survival for the two positions reached (w_{11} and w_{12}) are extremely close, both being near 100%. Due to the large sample size of these two positions, namely 1028 and 2969 respectively, the two positions are however not merged when scoring the CEGs according to the BD metric. The difference in score is, nevertheless, only 1.44 favouring the CEG in Figure 4.11, which only gives slight evidence for preferring this CEG. When one disability is severe and one is non-severe, then again the birth weight influences whether the individual reaches position w_{12} or position w_{13}. Again, the probabilities of survival associated with the two positions are fairly close, namely 99.68% and 95.09%. If one disability is non-severe and one is missing we first distinguish between the disability that is missing and then by birth weight. The associated predictive probabilities of survival of the positions range from 79.29% to 99.98%. Overall, it appears that missing ambulatory impairment predicts poorer survival than missing visual impairment: When visual ability is missing, but ambulation is non-severe, then survival appears to be approximately

as equally high as when both impairments are non-severe. However, when the ambulatory ability is missing, then, depending on the birth weight, the predictive probability of survival ranges from 79.29%, the lowest position, to 99.68%. This wide variation could be explained by the relatively small cell counts associated with a very low or low birth weight in this category. Further, Hutton et al. [1994] show that a low birth weight is often associated with lower rates of severe disability. Hence, as ambulation is missing, the low birth weight represents the likely severity of the impairment. Moving further down the graph, we observe that two severe impairments have a predictive survival probability of 86.65%, which, as expected, is significantly worse than two non-severe impairments. When one disability is severe and the other is missing then survival splits according to which disability is observed and, when visual impairment is missing, also according to the birth weight. The graph suggests that survival is slightly lower when ambulatory impairment is missing than when visual impairment is missing. In the latter case the probability of survival is indistinguishable from the case where both disabilities are severe for a low birth weight, as both paths lead to position w_{14}. In all other cases position w_{15} is reached with a probability of survival of 79.29%. In either case the graph suggests that it is unlikely for data to be MAR, as in neither case do we reach a position between 'one impairment severe and one non-severe' and 'both impairments severe', which is what we would expect under MAR. Finally, when both impairments are missing, then again the lowest probability of survival, namely 79.29%, is predicted for a normal birth weight and the second lowest, 86.65%, is predicted for a very low or low birth weight.

The CEG in Figure 4.11 therefore suggests that, when one impairment is missing and the other is severe, or when both impairments are missing, missingness is highly influential by leading to poorer survival and data are unlikely to be MAR. In contrast to this, when one impairment is observed to be non-severe and the other is missing, then the predictive survival probabilities are close to 100% when ambulation is observed and vary strongly between 99.68% and 79.29% when visual impairment is observed. As in the previous section the expected probabilities of survival under the MAR assumption can be calculated and are given in the left column of Table 4.3. These can then be compared to the predictive probabilities of survival suggested by the CEG, given in the right column of the same table.

When only one variable is missing then we condition on the other variable as if it were fully observed and deduce whether data are MAR or not as in Equation 4.4. When both variables are missing the MAR assumption is assumed to require that the missingness process is independent of both impairments given the birth weight. For

	Probability of survival in %					
	Expected under MAR			Predictive		
Birth weight	Very low	Low	Normal	Very low	Low	Normal
1 non-severe, 1 missing						
Visual ability observed	98.73(93.88,1.00)	98.96(89.95,1.00)	98.73(93.88,1.00)	79.29	99.68	95.09
Ambulatory ability observed	99.68(97.12,1.00)	99.89(98.95,1.00)	99.68(97.12,1.00)	99.98	99.98	99.68
1 severe, 1 missing						
Visual ability observed		88.54(61.29, 99.71)			79.29	
Ambulatory ability observed		92.32(89.51, 94.73)		79.29	86.65	79.29
2 missing	97.62(95.56,99.05)	97.83(94.04,99.73)	97.62(95.56,99.05)	86.65	86.65	79.29

Table 4.3: Plausibility of the MAR assumption in the UKCP example describing the effect of birth weight, visual impairment, ambulatory impairment and missing impairments on survival

example, the expected probability of survival, under MAR, for a low birth weight and given that the visual impairment is observed to be non-severe, can be calculated as follows from Table 4.2: The probability of survival given a low birth weight and given both impairments are non-severe is 99.98%. Similarly, given only visual impairment is non-severe and ambulatory impairment is severe, the predictive probability of survival is 95.09%. Also, the probability of a non-severe ambulatory impairment, given a low birth weight and observed non-severe visual impairment, is 0.7899, and, similarly, 0.2101 for a severe ambulatory impairment. Then the expected probability of survival under MAR is $99.98 \times 0.7899 + 95.09 \times 0.2101 = 98.96\%$. The remaining probabilities of Table 4.3 are calculated in the same way.

As already deduced from the graph missingness is strongly influential when both impairments are missing. Similarly, when the ambulatory impairment is observed to be severe, then data are unlikely to be MAR, as the predictive probabilities of survival are lower than the expected probability under MAR and do not lie in their respective 95% credible intervals. It was further deduced from the graph that, when the visual impairment is observed to be severe, the predictive probability of survival is 79.29%, which is again substantially lower than the expected probability under MAR, which is 88.54%. However, due to the small number of cell counts within this category (compare Table 1.3), the credible interval is very wide and in fact includes the predictive probability of survival. Therefore, although data appears to be MNAR, we need to be cautious with our conclusions due to the sparse cell counts. When ambulatory impairment is observed to be non-severe and visual impairment is missing, then the table suggests that data is likely to be MAR. This was also suggested by the CEG structure in Figure 4.11. Finally, when the visual impairment is observed to be non-severe, then data appears to be MNAR for a very low birth weight, but MAR for a very low or low birth weight.

4.4 CEGs for Informed Variable Construction

I have demonstrated in the previous chapters that in cohort studies interest usually lies particularly in the final set of positions of the CEG. In the examples discussed in this thesis the final positions describe the 'health state' the individual is in with an associated probability of hospital admission or survival by which the positions in the Ordinal CEG are vertically aligned. In the first cerebral palsy example I deduced from the graph that survival depends on whether visual impairment is missing and, when it is missing, on birth weight, while, when it is observed, on the severity of the impairment. Consequently it was possible to distinguish between three health states with an associated probability of survival. In the second example, discussed in the previous section, I distinguished between five different final positions. The paths leading up to these positions were reduced by defining a new variable which counts the number of non-severe, severe and missing impairments, leading to the Reduced Ordinal CEG. Hence the five 'health states' depended on this new variable as well as on a combination of birth weight and type of impairment.

The above examples motivate the idea that in a setting like this the CEG may also be useful in defining informative categories of variables to predict the probability of the outcome variable, for example survival. In Chapter 3 it was shown that the final situations in the probability tree are merged into the same position when the associated probability distribution on their emanating edges is similar. The final positions in the MAP CEG structure give an informative description of the way in which a combination of variables affects the outcome and this segmentation of the paths in the CEG can then be used to define a new variable with categories equal to the final positions in the graph. The resulting new variable could then be employed in a subsequent analysis. This concept of course becomes more interesting the more variables are included in the problem.

One of the aims of the full UKCP study is to investigate the range of causes of death in the cohort and its association with impairments (Maudsley et al. [1999] and personal communication with J.L. Hutton). The effect of various risk factors on the causes of death are to be analysed using logistic regression models and competing risk models. Three impairments, namely visual, ambulatory and manual impairment, all with substantial amounts of missing data, are classed as risk factors of survival for people with cerebral palsy. However, including all three impairments as well as their two-way and three-way interaction terms would make the model extremely complex. An alternative, proposed in Hutton [2006], is to count the number of severe and missing impairments and define a new variable with states 0, 1, 2, and

3 severe or missing impairments, which is then used in a later analysis. However, this assumes that a missing impairment has the same effect on survival as a severe impairment and the previous examples show that this is not the case. Instead, I suggest here constructing a CEG on the three impairments and survival and use the final positions in the resulting graph to define a new variable with informed categories. I illustrate this approach below and show that the final positions of the MAP CEG give a significant improvement to the other two suggested approaches which propose to either include all interaction terms or count the number of severe and missing impairments.

In order to draw informative conclusions about survival, the variable describing survival occurs last in the probability tree from which the CEG can be derived. It would then be possible to search across all six possible trees (i.e. all permutations of the three impairments). However, as before, I will reduce the CEG to a Reduced Ordinal CEG for clarity, focusing only on the final positions describing survival, in which case the ordering of the three covariates is not relevant. I here choose again a chronological ordering of the variables according to the time at which they are likely to be observed. I again put uniform priors on the root-to-leaf paths of the tree and specify an equivalent sample size of $\alpha = 3$. (A comparison of the MAP CEG when specifying a larger equivalent sample size is given in 6.2.) The resulting MAP CEG structure found using the AHC algorithm is given in Figure 4.12 with its CPVs given in Table 4.4. (Note that Man. = manual impairment.)

Stage/Position	Conditional Probability Vector	
$u_0 = w_0$	$(P(X_1 = \text{Non-severe}), P(X_1 = \text{Severe}), P(X_1 = \text{Missing}))$	$(0.7293, 0.0863, 0.1842)$
$u_1 = w_1$	$(P(X_2 = \text{Non-severe}\|u_1), P(X_2 = \text{Severe}\|u_1), P(X_2 = \text{Missing}\|u_1))$	$(0.7809, 0.2027, 0.0118)$
$u_2 = w_2$	$(P(X_2 = \text{Non-severe}\|u_2), P(X_2 = \text{Severe}\|u_2), P(X_2 = \text{Missing}\|u_2))$	$(0.1422, 0.8425, 0.0153)$
$u_3 = w_3$	$(P(X_2 = \text{Non-severe}\|u_3), P(X_2 = \text{Severe}\|u_3), P(X_2 = \text{Missing}\|u_3))$	$(0.3023, 0.3437, 0.3540)$
$u_4 = \{w_4, w_5\}$	$(P(X_3 = \text{Non-severe}\|u_4), P(X_3 = \text{Severe}\|u_4), P(X_3 = \text{Missing}\|u_4))$	$(0.9676, 0.0200, 0.0124)$
$u_5 = w_6$	$(P(X_3 = \text{Non-severe}\|u_5), P(X_3 = \text{Severe}\|u_5), P(X_3 = \text{Missing}\|u_5))$	$(0.9280, 0.0260, 0.0460)$
$u_6 = w_7$	$(P(X_3 = \text{Non-severe}\|u_6), P(X_3 = \text{Severe}\|u_6), P(X_3 = \text{Missing}\|u_6))$	$(0.3607, 0.6150, 0.0243)$
$u_7 = w_8$	$(P(X_3 = \text{Non-severe}\|u_7), P(X_3 = \text{Severe}\|u_7), P(X_3 = \text{Missing}\|u_7))$	$(0.3517, 0.0021, 0.6462)$
$u_8 = w_9$	$(P(X_3 = \text{Non-severe}\|u_8), P(X_3 = \text{Severe}\|u_8), P(X_3 = \text{Missing}\|u_8))$	$(0.0810, 0.9013, 0.0177)$
$u_9 = w_{10}$	$(P(X_3 = \text{Non-severe}\|u_9), P(X_3 = \text{Severe}\|u_9), P(X_3 = \text{Missing}\|u_9))$	$(0.2312, 0.7284, 0.0404)$
$u_{10} = \{w_{11}, w_{12}\}$	$(P(X_3 = \text{Non-severe}\|u_{10}), P(X_3 = \text{Severe}\|u_{10}), P(X_3 = \text{Missing}\|u_{10}))$	$(0.0053, 0.0101, 0.9846)$
$u_{11} = w_{13}$	$(P(X_4 = \text{Survival}\|u_{11}), P(X_4 = \text{No survival}\|u_{11}))$	$(0.9979, 0.0021)$
$u_{12} = w_{14}$	$(P(X_4 = \text{Survival}\|u_{12}), P(X_4 = \text{No survival}\|u_{12}))$	$(0.9871, 0.0129)$
$u_{13} = w_{15}$	$(P(X_4 = \text{Survival}\|u_{13}), P(X_4 = \text{No survival}\|u_{13}))$	$(0.9289, 0.0711)$
$u_{14} = w_{16}$	$(P(X_4 = \text{Survival}\|u_{14}), P(X_4 = \text{No survival}\|u_{14}))$	$(0.8736, 0.1264)$
$u_{15} = w_{17}$	$(P(X_4 = \text{Survival}\|u_{15}), P(X_4 = \text{No survival}\|u_{15}))$	$(0.7757, 0.2243)$
$u_{16} = w_{18}$	$(P(X_4 = \text{Survival}\|u_{16}), P(X_4 = \text{No survival}\|u_{16}))$	$(0.4841, 0.5159)$

Table 4.4: Table of CPVs associated with the MAP CEG for the UKCP example on visual, ambulatory, manual impairment and missingness of impairments on survival given in Figure 4.12

As in the previous examples the CEG is written as an Ordinal CEG and the

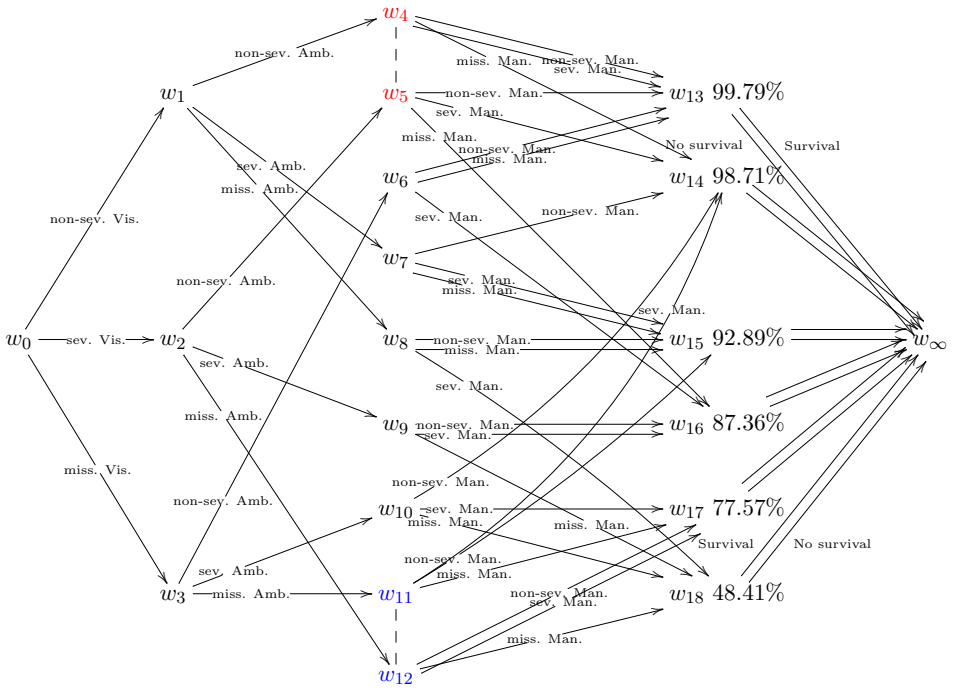


Figure 4.12: Ordinal MAP CEG structure for the UKCP example describing the effect of visual, ambulatory, manual impairment and missingness of impairments on survival

predictive probabilities of survival are attached to its final positions. These together with their 95% credible intervals are: $99.79(99.63, 99.91)\%$, $98.71(97.53, 99.51)\%$, $92.89(90.81, 94.71)\%$, $87.36(84.17, 90.23)\%$,$77.57(74.40, 80.59)\%$ and $48.41(31.43, 65.58)\%$. It can be seen from Table 1.2 in the introductory chapter that for several combinations of impairments the cell counts are sparse as only a very small number of individuals have a particular combination of non-severe, severe or missing impairments. When drawing conclusions from the graph the sparse cell counts need to be treated with caution due to the lack of data available. Further, there are two paths up to the final positions with no individuals. These are the paths describing 'severe visual impairment', 'missing ambulatory impairment', 'non-severe manual impairment' and 'non-severe visual impairment', 'missing ambulatory impairment', 'severe manual impairment'. As I have so far assumed a uniform prior. i.e. a prior survival probability of 50%, these paths will lead to positions with a low probability of survival. As no data is available it is not possible to deduce plausible probabilities of survival for these two combinations of covariates from the data set. In this case incorporating expert knowledge through informative priors into the model would be desirable. Sparseness of cells for Table 1.2 and the use of informative priors for this

example will be discussed in Chapter 6. I again draw the Reduced Ordinal CEG to enhance the graphical representation of the problem, which is given in Figure 4.13.

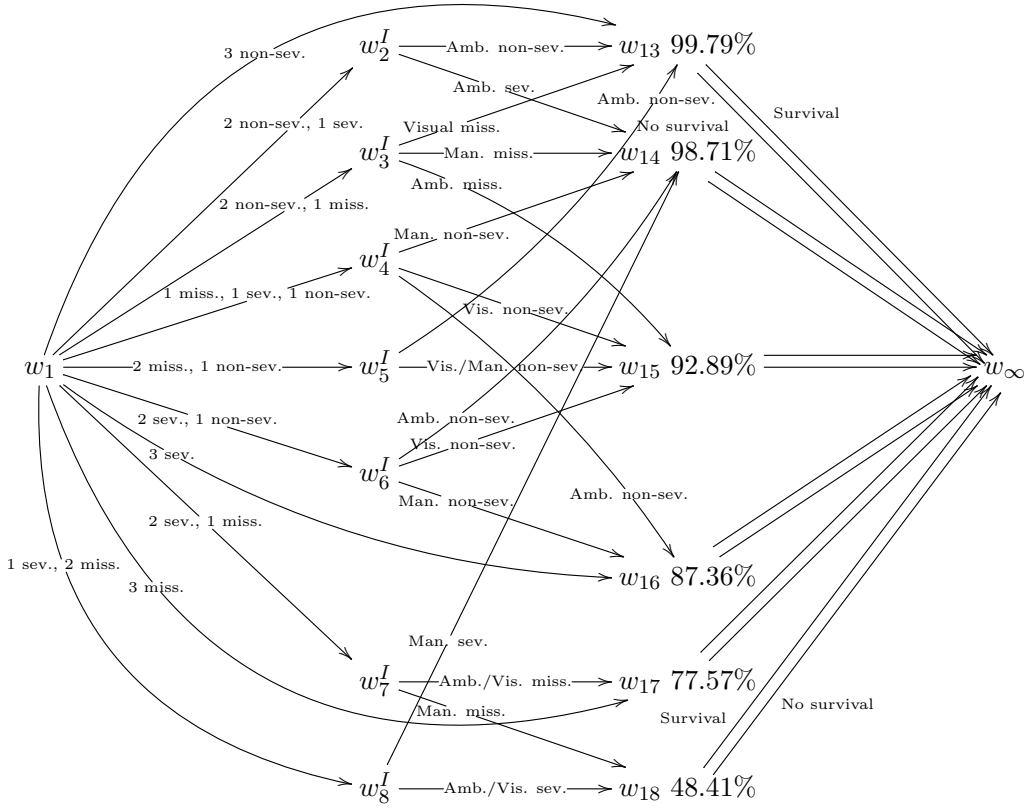


Figure 4.13: Reduced Ordinal MAP CEG structure for the UKCP example describing the effect of visual, ambulatory, manual impairment and missingness of impairments on survival

As before, it is plausible to define a variable which counts the number of non-severe, severe and missing impairments and which consequently has ten levels. Note that the two paths with zero individuals have been moved from positions w_{17} and w_{18} to position w_{14} and w_{15} respectively, to allow for the reduced representation in Figure 4.13. The logarithm of the Bayes Factor between the Reduced CEG and the original CEG is 0.41 (Bayes Factor: 1.51) favouring the MAP CEG only insignificantly. The following conclusions can then be drawn from the graph: As expected missingness appears to be highly influential when all three impairments are missing. Further missingness is associated with poorer survival when one impairment is missing and two are severe. In both cases w_{17} is reached with a predictive probability of survival of 77.57% and even w_{18} with an expected probability of survival of 48.41%, when manual impairment is missing. Further, missingness shows to be especially influential when two impairments are missing and one is observed to be severe. When

either ambulatory impairment or visual impairment is observed to be severe then the survival is predicted to be only 48.41%. When manual impairment is observed then the survival probability is surprisingly high. However, as mentioned above, this is most likely due to only three individuals being observed in this category. Cases like this will be further discussed in Chapter 6.3.

In contrast to this, three severe impairments lead to position w_{16} with a probability of survival of 87.36%. Hence, the CEG suggests that missingness has a strong influence on survival, especially when position w_{18} is reached, for which survival is below 50%. When two impairments are missing and the third is observed to be non-severe, then survival is predicted to be at 92.89%, and at 99.79% when the ambulatory impairment is observed to be non-severe. As in the previous examples, non-severe ambulatory impairment is influential for predicting a high probability of survival. Further, when two impairments are observed to be non-severe and only one impairment is missing, then the type of impairment which is missing influences the probability of survival, where a missing ambulatory impairment reduces the survival to 92.89%. Finally, when one impairment is missing, one is severe and the third non-severe, then the impairment that is not severe distinguishes between different probabilities of survival. Surprisingly, here a non-severe ambulatory impairment has the lowest probability of survival. However, Table 1.2 shows that only 10 individuals go along this path, in comparison to 23 and 92 individuals going along the other two paths with non-severe visual or ambulatory impairment. Hence again sparse data for this category give misleading results.

Comparing this to the CEG in Figure 4.11 on birth weight, visual and ambulatory impairment, we observe that the five final positions in this graph are similar in survival probability to the top five final positions in Figure 4.13. However, now an extra final position in Figure 4.13 exists with a significantly lower probability of survival of 48.41%. This position is reached by individuals with one severe and two missing impairments or two severe and one missing impairment. A number of further distinctions can be detected: For example, two severe impairments in Figure 4.11 lead to a predictive probability of survival of 86.65%. When the manual impairment is observed to be non-severe or severe in Figure 4.13 then we obtain approximately the same probability of survival. However, when manual impairment is missing the survival probability reduces to 48.41%. Also, when one impairment is observed to be severe and the other non-severe in Figure 4.11 then, comparing this to Figure 4.13, shows that manual impairment improves survival when it is non-severe and reduces survival when it is missing. Overall, manual impairment appears to be influential for survival. Nevertheless including only two impairments

already allows us to draw similar substantial conclusions.

I will now compare the deductions made from this CEG to the two approaches suggested at the beginning of this section. Recall from Chapter 2.5 that a regression analysis, which includes all two-way and three-way interaction terms, is comparable to the CEG, $\mathcal{C}_0$, where all situations whose emanating edges describe survival, are in separate positions. Note that C_0 hence has 27 final positions, one for each combination of impairments, while the MAP CEG only has six. To compare the approach of including all interaction terms to the approach of using the six final positions of the MAP CEG to define new informative categories, the difference in score between C_0 and the MAP CEG is calculated. (Note that for an accurate comparison, which focuses on the final positions in the tree, C_0 is compared to the CEG with the final positions as in the MAP CEG and all other positions left separate.) The difference in score between these two CEGs is 40.22 (Bayes Factor: $e^{40.22}$) favouring the CEG with six final positions. Therefore, by Table 3.1, this CEG is strongly preferred over the CEG C_0.

A further approach suggested at the beginning of this section was to count the number of severe or missing impairments and to include only a single variable into the model, taking the values 0, 1, 2, or 3 missing or severe impairments. It is simple to construct a CEG which describes this variable: This would be a CEG with four final positions, each position representing one value of the variable. Consequently, the path describing three non-severe impairments would be in one separate position, the paths describing two non-severe and one severe or missing impairment would lead to another position and so on. Again the difference in score between this CEG with four final positions and the CEG with six final positions, obtained from the MAP CEG, can be calculated to give a score of 17.92 (Bayes Factor: $e^{17.92}$) favouring the CEG with six positions. Although counting the number of severe or missing impairments is strongly preferable to treating every combination separately, the score is still significantly worse than the score obtained from the MAP CEG. I therefore propose using the final positions of the MAP CEG to define a new variable, with six informative categories, which gives a description of the severity of the three impairments. In particular, in this setting, it is further possible to also incorporate missingness in a structured way across the different impairments. Instead of treating it simply the same as a severe impairment (as suggested in the second approach), missingness is treated differently dependent on whether other impairments are observed to be severe or non-severe. The CEG therefore not only provides further insight into the way in which the severity and the number of impairments (including missing values) influence survival, but also allows us to make informed decisions

about the construction of the covariates, which can then be used in a later survival analysis.

This Chapter suggests that the CEG provides a useful new way of exploring systematically the influence of missing data within cohort studies. The (Reduced) Ordinal CEG, in particular, allows us to obtain a precise understanding of the subtleties associated with the three common types of missingness, as well as letting us derive new informative categories of variables, which can be used in a later analysis.

Chapter 5

The Dynamic Chain Event Graph

I have shown in the previous chapters that a CEG is a discrete graphical model which can capture many important features of a process, in particular how a combination of events in an individual's life affects an outcome. In Chapter 2.4 the topology of the CEG has been exploited to represent and generalise models such as the BN. However, the CEG cannot be used to generalise discrete dynamic processes like the Dynamic BN (DBN), as its semantics have, up to this point, only been developed to describe processes whose underlying probability tree is finite and whose final situations describe the variable of interest such as survival.

It has been shown that an event tree provides a natural framework to describe the various possible sequences of events an individual can experience. In this chapter the model space is extended to infinite trees to describe potentially infinite discrete longitudinal processes. Hence, I now assume that the events encountered by an individual could be infinite and, above all, that the events an individual experiences may be repeated at later points in time. These could be measured either at regular time points, for example when yearly measurements are taken, or at irregular time intervals, i.e. at the time at which an event occurs. A different extension of the CEG to a dynamic graphical model was developed in Freeman and Smith [2011b]. Here the underlying probability tree is finite but the stage structure of the possible CEGs is allowed to change across discrete time-steps. This model, however, addresses an entirely different problem to the one considered here, as it looks at different cohorts entering the tree at discrete time-points rather than assuming that repeated measurements are taken over time.

I will develop the Dynamic CEG (DCEG) [Barclay et al., 2013c], which is

derived from an infinite probability tree, and will demonstrate that it gives a powerful representation for modelling discrete dynamic processes. I will show that the DCEG can not only model the effect of a combination of variables on an outcome but also the temporal relationships between these variables, again allowing for asymmetries within the dependence structure. I will first define the DCEG, derived from an infinite tree. Analogously to the CEG, stages and positions on the infinite tree can be defined and hence the DCEG is able to represent the originally elicited tree in a much more compact and easily interpretable form. I will illustrate its representation on the fictional example introduced in Chapter 1.2. In the subsequent subsection I will then demonstrate that this framework can be extended to allow for time to be modelled explicitly by attaching holding time distributions to the vertices in the graph. Analogous to Chapter 2.4 I will show in Section 5.2 that any general DBN can be written as a DCEG. In Section 5.3 I will further show that the DCEG is closely linked to discrete-time Markov processes, while the DCEG with holding time distributions is related to semi-Markov processes. In the final section I will discuss the learning of the parameters in a DCEG, which suggests that, like its CEG and BN analogues, the DCEG not only provides an expressive representation of a process but also supports conjugate learning and closed form model selection.

5.1 Dynamic Chain Event Graphs and Their Semantics

In this Section I will describe the derivation of the DCEG from an infinite tree. I will first define the DCEG and then extend this further to the DCEG with conditional holding times attached to the vertices in the graph to allow for time to be modelled explicitly.

5.1.1 Infinite Trees and the DCEG

Recall from Chapter 2.1.2, Definition 4 that an infinite graph is a graph with an infinite number of vertices and/or edges. I will continue to assume in this thesis that each situation s_i has a finite number of edges m_i emanating from it. An infinite tree therefore will have an infinite number of situations. It may have no leaf vertices when all paths are infinite. However, there may be examples, such as the ones demonstrated in this chapter, where some paths in the tree are finite and hence a set of leaf vertices exists.

I consider the following fictional example, introduced in Chapter 1.2, to illustrate an infinite tree, $\mathcal{T}$.

Example 12. *An individual is every month at risk of catching flu. When he catches the flu he either decides to take antiviral treatment or not. If he takes antiviral treatment he will always recover, but if he does not take antiviral treatment he either manages to recover or he dies from the virus. After a full recovery the individual either decides to go back to his normal life where he is at risk of catching flu again or he decides to receive an influenza vaccine to prevent him from being at risk again. As the tree is infinite, only an informal depiction of the corresponding tree can be given (Figure 5.1), where implicit continuations of the tree are given by the notation '...'.*

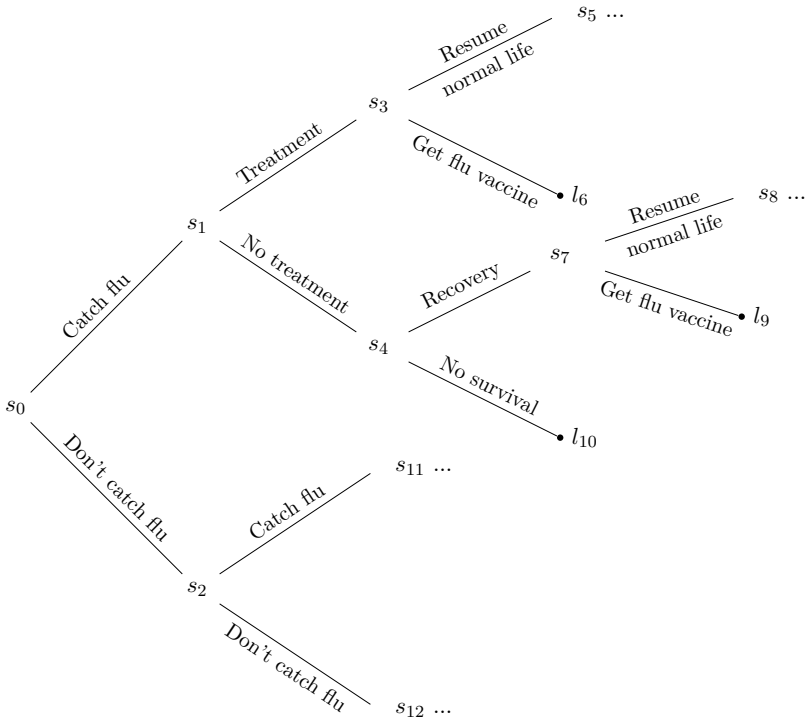


Figure 5.1: The beginning of the infinite tree, $\mathcal{T}$, for the flu example describing the development of catching flu, recovering with or without treatment and getting a flu vaccine

Like in the finite case, the set of situations of the infinite tree, $\mathcal{T}$, can be partitioned into a set of stages, $U(\mathcal{T})$, or positions $W(\mathcal{T})$. The definitions of stages and positions are as given in Chapter 2.3 in Definitions 18 and 20. However, now the number of situations in a stage or position may be infinite. As the set of positions is now defined on an infinite tree, two situations, s_i, s_j, lying on the same directed path from the root, can be in the same position. This is impossible for two situations in a finite tree, where the subtree rooted at a situation further along the path must

necessarily have fewer vertices than the one closer to the root of the tree. So, in particular, in a finite tree no bijection between $\mathcal{T}(s_i)$ and $\mathcal{T}(s_j)$ can exist, when s_i and s_j lie on the same path. Consequently, the set of positions of an infinite tree $\mathcal{T}$ can sometimes be associated with a coarser partition of its situations than a finite subtree of $\mathcal{T}$ with the same root. I continue the flu example below to demonstrate how an infinite staged tree can be constructed and a DCEG derived from this.

Example 13. *Assume that the probability of catching flu does not change over the months and does not depend on whether flu has been caught before. This implies that s_0, s_2, s_5, s_8 and s_{12} are in the same stage, as well as all subsequent situations describing this event, which are not represented in Figure 5.1. Similarly, s_1 and s_{11} are in the same stage, such that whether the antiviral medication is taken or not is also independent of the number of months until the individual catches flu and independent of flu having been caught before. I further assume that the probability of the individual returning to his normal life after recovery is the same when he recovers after treatment as when he successfully recovers without treatment. This means that s_3 and s_7, as well as all other situations representing the probability of returning to a normal life after recovery, are in the same stage. The corresponding staged tree representation is given in Figure 5.2.*

It can be seen from the staged tree that, in this example, whenever two situations are in the same stage, they are also in the same position as their subtrees have the same topology and the same colouring of corresponding edges. Not all paths in the tree are infinite and hence a set of leaf vertices, $\{l_6, l_9, l_{10}, \ldots\}$, exists.

The DCEG then represents the infinite staged tree in a way that is analogous to the way the CEG represents structural and probabilistic symmetries in a finite tree. It is defined in exactly the same way as the CEG, however on an infinite staged tree with possibly infinite stages and positions. I give a formal definition below:

Definition 25. *A **Dynamic CEG (DCEG)**, $\mathcal{D} = (V(\mathcal{D}), E(\mathcal{D}))$, of an infinite staged tree $\mathcal{T}$, has vertex set $V(\mathcal{D}) = W(\mathcal{T}) \cup w_\infty$, the set of positions of the infinite staged tree $\mathcal{T}$, together with a single sink vertex, w_∞, comprising the leaf nodes of $\mathcal{T}$, if these exist. Emanating edges from a position $w_i \in W(\mathcal{T})$ are constructed as follows: Choose a single representative situation $s(w_i) \in S(\mathcal{T})$. Then there is an edge from w_i to a position $w_j \in V(\mathcal{D})$ for each child $v_j \in ch(s(w_i)), v_j \in w_j$ in the tree $\mathcal{T}$. When two positions are also in the same stage then they are connected by an undirected dashed line and their edges are coloured according to their colouring in the staged tree.*

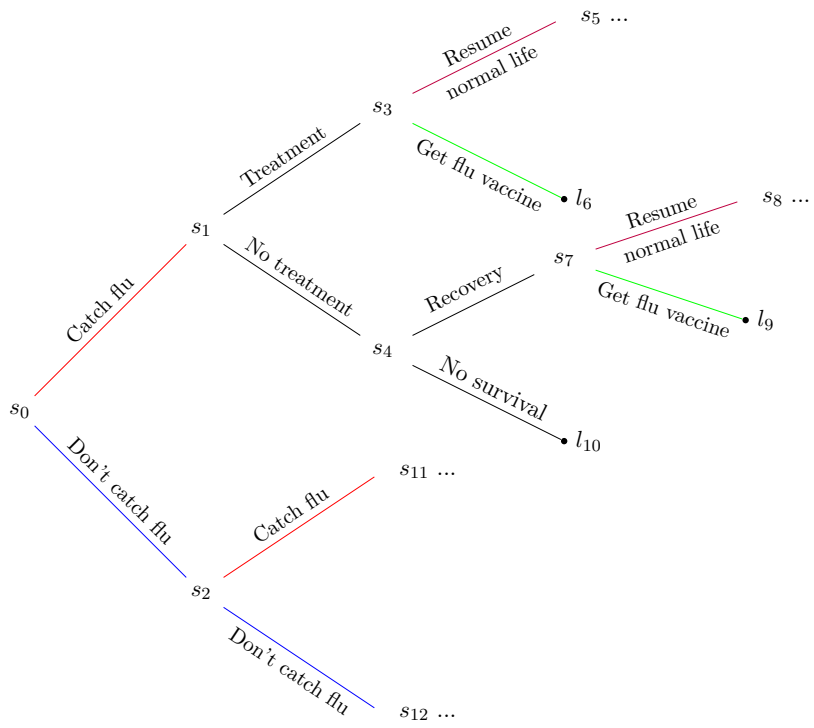


Figure 5.2: The beginning of the infinite staged tree, $\mathcal{T}$, for the flu example with two assumptions: 1. the probability of catching flu does not depend on whether the flu was caught in the past and 2. the probability of the individual returning to his normal life is independent of whether treatment was taken or not

Given a DCEG $\mathcal{D}$ I will henceforth denote the stage partition of the DCEG by $U(\mathcal{D})$ and its position partition by $W(\mathcal{D})$. From Definition 25 the DCEG associated with the staged tree in Figure 5.2 can then directly be derived.

Example 14. *The DCEG of the flu example is given in Figure 5.3 with stage and position partition given as follows:*

$$w_0 = u_0 = \{s_0, s_2, s_5, s_8, s_{12} \ldots\}, w_1 = u_1 = \{s_1, s_{11}, \ldots\},$$
$$w_2 = u_2 = \{s_3, s_7, \ldots\}, w_3 = u_3 = \{s_4, \ldots\}, w_\infty = \{l_6, l_9, l_{10}, \ldots\}.$$

Again the notation '...' implies that we have an infinite number of situations in each stage or position, as well as an infinite number of leaf vertices. The loop from w_0 into itself illustrates that every month the individual could remain well and not catch flu. Alternatively, the individual may move to w_1 at some point, meaning that he has caught flu. In this case he can recover either by getting treatment ($w_1 \to w_2$) or recover on his own ($w_1 \to w_3 \to w_2$). Having recovered the individual either

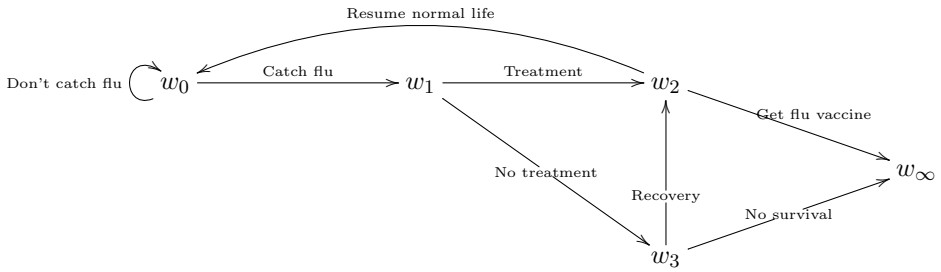


Figure 5.3: The DCEG of the infinite staged tree from Figure 5.2 for the flu example

decides to take a flu vaccine to avoid getting flu again ($w_2 \to w_\infty$) or to simply resume his normal life and risk getting flu again ($w_2 \to w_0$). Finally, when not taking treatment, the individual may not recover, and hence move from w_3 to w_∞.

The example illustrates that in many applications the number of positions and stages of a staged tree may be finite even though the tree's vertex set is infinite due to repeated measurements. When this is the case the DCEG is a finite graph, as above, and therefore provides a succinct picture of the structural and probabilistic relationships in the process, which would be difficult to capture by representing the problem simply by an infinite tree. In contrast to the CEG, which is always acyclic, the DCEG exhibits cycles when it has an infinite number of paths but a finite graph. As for a CEG it would be possible to elicit the tree structure of a given problem from a client or domain expert and similarly elicit possible stage and position partitions.

5.1.2 The DCEG with Holding Times

So far I have assumed that the DCEG models a dynamic process where measurements are taken at regular intervals, such as daily or monthly. For example in the DCEG of the flu example in Figure 5.3, the individual is, every month, at risk of catching flu: If he catches flu, he traverses through the rest of the DCEG ending up either at w_∞ or back at w_0; if not he loops back directly to w_0. In this section I will extend the above methodology so that time spent until an event occurs can be modelled directly.

Going back to the tree representation of a problem call the time an individual stays in a situation s_i the **holding time** H_{s_i} associated with this situation. Further, let the **conditional holding time** associated with each edge $e_{s_i k}, k = 1, \ldots, m_i$ in the tree be denoted by $H_{s_i k}$. This describes the time an individual stays at a situation s_i given that he moves along the edge $e_{s_i k}$ next. Analogously to this, holding times on the positions of the associated DCEG $\mathcal{D}$ can be defined as follows: Let H_w be the random variable describing the holding time at position $w \in W(\mathcal{D})$

in the DCEG and $H_{wk}, k = 1, .., m_w$ the random variable describing the conditional holding time on w given the individual moves along the edge e_{wk} next.

In a DCEG the time an individual stays in a particular position w, with a loop into itself, simply follows a geometric distribution. So, if we assume that the kth edge of w loops back into w, then the probability that an individual stays in position w for t months is equal to $\pi_{wk}^t \times (1-\pi_{wk})$, where $e_{wk} = e(w, w)$. Further, it has been assumed that once an individual catches flu, only the events of taking treatment, recovering, and receiving a vaccine are recorded and not the time until these events occur. These could, for example, be recorded retrospectively when measurements are taken a month later. The holding time distributions on a position without a loop into itself are therefore degenerate.

As in the flu example, the processes to be modelled are often event driven and these are well represented within a tree and hence a DCEG: When moving from one position to another the individual transitions away from a particular state into a different state associated with a new probability distribution of what will happen next. In these scenarios, interest commonly lies not only in the transition probabilities through the graph but also in the amount of time spent at each position. Hence, rather than measurements being taken at regular time-steps it is more natural to think of the measurements being taken when an event happens, where the time until the event happens is recorded. For example, the individual may not record whether he catches flu or not every month but instead monitor the time spent at w_0 not catching flu, until one day he falls ill. Similarly, the time until seeing the doctor for treatment or the time until recovery may be of different lengths and so he spends different amounts of time at each position in the DCEG. In order to incorporate this into the graph conditional holding time distributions can be attached to each edge in the DCEG.

By the definition of a DCEG, two situations are in the same stage whenever their emanating edges have the same probability. Similarly, it is assumed that, the conditional holding time depends only on the current stage and the next edge the individual moves along but not on the previous path up to reaching the current stage.

Definition 26. *A DCEG is* ***time-homogeneous*** *whenever two situations that are in the same stage also have the same conditional holding time distributions on their edges, i.e. the holding times are independent of the path taken through which the stage is reached. Denote the random variable of the conditional holding time associated with each stage by* $H_{uk}, k = 1, \ldots, m_u$.

I will assume throughout that the DCEG is time-homogeneous, which further

implies that when two positions are in the same stage u then their conditional holding time distributions are also the same. Note that an individual may spend a certain amount of time in position $w_i \in u$ before moving along the kth edge to a position w_j which is in the same stage. So an individual may make a transition into a different position but arrive at the same stage.

I further assume throughout that the conditional probabilities of going along a particular edge after reaching a stage, do not vary with previous holding times. In the flu example this would mean that the time until catching flu does not effect the probability of taking treatment and the probability of recovery without treatment. Similarly, the holding times are assumed to be independent of previous holding times. So, for example, the time until recovery is independent of the time to catching flu. Contexts where the holding time distribution may affect the transition probabilities and future holding times can provide an interesting extension to the DCEG, which, however, will not be covered in this thesis. Under these assumptions a time-homogeneous DCEG with holding times can therefore be defined as follows:

Definition 27. *A **DCEG with holding times**, $\mathcal{D} = (V(\mathcal{D}), E(\mathcal{D}))$ is a DCEG with no loops from a position into itself and with conditional holding time distributions conditioned on the current stage, u, and the next edge, e_{uk}, to be passed through:*

$$F_{uk}(h) = P(H_{uk} \leq h | u, e_{uk}), h \geq 0, u \in U, k = 1, \dots m_u.$$

Hence $F_{uk}(h)$ describes the time an individual stays in a position $w \in u$ before moving along the kth edge, e_{wk}. A frame around a position in $\mathcal{D}$ indicates that holding time distributions have been attached to its associated edges.

Consequently, given a position $w \in W(\mathcal{D})$ is reached, the joint probability of staying at this position for a time less than or equal to h and then moving along the kth edge is

$$P(H_{wk} \leq h, e_{wk} | w) = P(H_{wk} \leq h | w, e_{wk}) P(e_{wk} | w) = F_{uk}(h)\pi_{uk}, w \in u. \quad (5.1)$$

Finally, the joint density of e_{wk} and h is

$$p(e_{wk}, h | w) = \pi_{uk} f_{uk}(h),$$

where f_{uk} is the pdf or pmf of the holding time at stage u going along edge $e_{wk}, w \in u$ next. A time-homogeneous DCEG $\mathcal{D}$ with stage partition $U(\mathcal{D})$ is therefore fully specified by its set of conditional holding time distributions
$\{F_{uk} : u \in U(\mathcal{D})\}$ and its collection of CPVs $\{\pi_u : u \in U(\mathcal{D})\}$ and the elicitation

process of a DCEG could be extended to include the elicitation of holding time distributions for each position and its associated edges. The conditional holding times could in general take any distribution. For example, an exponential holding time distribution may be plausible if it is assumed that the event will occur at a constant rate. Other plausible distributions would be alternative survival distributions such as the Weibull distribution, when the occurrence of the event is expected to increase or decrease with time, or a log-normal distribution or a log-logistic distribution, when a unimodal event rate is appropriate. I will postpone further discussion of the holding time distribution for this example to the end of this chapter when looking at the learning of the parameters of the DCEG.

Example 15. *Consider the following variant of the flu example represented by the infinite tree, $\mathcal{T}^*$, in Figure 5.4. Instead of measuring every month whether the individual catches flu, the individual will spend a certain amount of time at s_0 before moving along the tree. Hence the second edge emanating from s_0 in Figure 5.2 and its entire subtree have been removed. As before, it is assumed that the probability of*

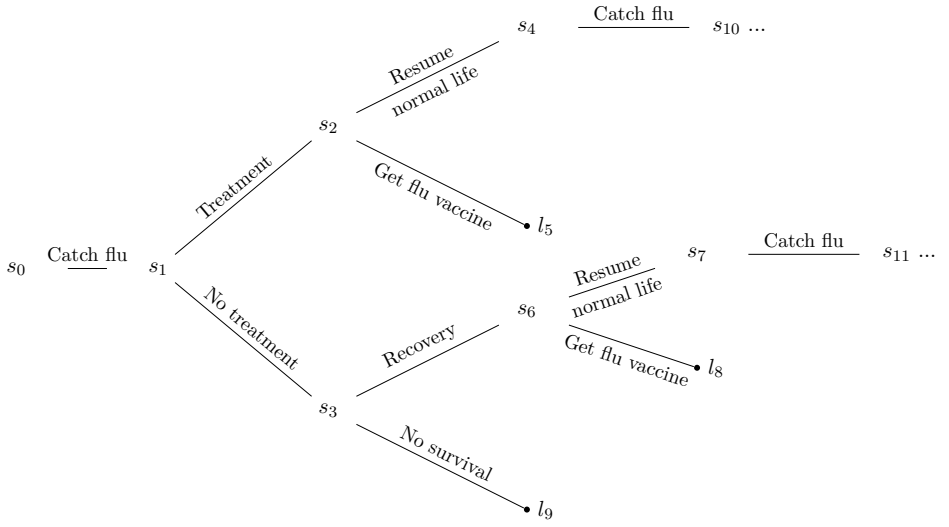


Figure 5.4: The beginning of the infinite tree, $\mathcal{T}^*$, for the flu example where catching flu is represented by the time spent at the root vertex

catching flu and the decision to take treatment does not depend on whether the flu has been caught before. Also, recovery with or without treatment is assumed not to affect the probability of receiving a vaccine. The corresponding DCEG is given in

Figure 5.5 with the stages and positions given by

$$w_0 = u_0 = \{s_0, s_4, s_7, \ldots\}, w_1 = u_1 = \{s_1, s_{10}, s_{11}, \ldots\},$$
$$w_2 = u_2 = \{s_2, s_6, \ldots\}, w_3 = u_3 = \{s_3, \ldots\}, w_\infty = \{l_5, l_8, l_9, \ldots\}. \quad (5.2)$$

In comparison to Figure 5.3 the loop from w_0 into itself has been removed. Instead

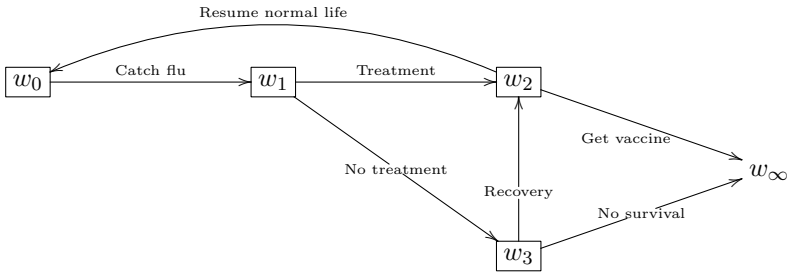


Figure 5.5: The DCEG with holding times for the flu example, where catching flu is described by the time spent at the root vertex

the time spent at w_0 is described by the holding time at position w_0. Similarly, the time until treatment is taken or not, the time until recovery or death and the time to receiving the flu vaccine or not are of interest and holding time distributions can be defined on these. Hence, visually the only difference between Figures 5.3 and 5.5 is that the positions have a frame around them to illustrate that the conditional holding times are of interest and w_0 no longer contains a loop into itself.

5.2 Writing a Dynamic BN as a DCEG

Various dynamic graphical models to model longitudinal data are already well studied. The most widely used dynamic graphical model is the Dynamic BN (DBN) [Korb and Nicholson, 2004; Koller and Lerner, 2001; Murphy, 2002], a BN which is repeated across discrete time-steps. In fact also the two other classes of graphical models mentioned in 2.1.1, namely Markov Networks and Chain Graphs, have a dynamic counterpart which expands the model over discrete time-steps. In this section I will demonstrate how a discrete DBN can be represented by a DCEG and hence that, as in the comparison between CEGs and BNs, the DCEG is a more general class of models than the DBN by allowing for asymmetric dependence structures between the variables.

It has been shown in Chapter 2.4 how a BN can be written as a staged tree and hence as a CEG. This can be simply extended to a dynamic setting where a discrete DBN can be represented as an infinite staged tree and therefore as a DCEG. A

DBN is an extension of the BN for discrete-time stochastic processes, which models not only the relationship of the variables at some point in time but also models their temporal relationships. Assume, as for the BN, that a given problem is defined by a vector of p random variables, $\mathbf{X} = (X_1, X_2, \ldots, X_p)$. To construct a DBN, which also models change over time, a vector of p variables $\mathbf{X}_t = (X_{1,t}, X_{2,t}, \ldots, X_{p,t})$ for each time point t is assumed, where $t \in T$ and $T = \{t_0 < t_1 < t_2, \ldots\}$, represents the discrete time points of the process. The variables $\mathbf{X}_t$ then form a time-slice or time-step of the DBN for each time point t and the graph represents the conditional independence structure between the set of variables $\{X_{i,t}, i = 1, \ldots p, t \in T\}$. As in previous chapters, I assume that every $X_{i,t}$ takes a finite number of values, m_i. The DBN was first defined in Dean and Kanazawa [1989] under the name of a probabilistic temporal network. Most applications assume regular time-steps, however this is not strictly required. In the most general case, see for example Murphy [2012], the DBN can be defined as follows:

Definition 28. *A **Dynamic BN** on $\{\mathbf{X}_t : t \in T\}$ is made up of*

1. *an infinite DAG $\mathcal{B}$ with vertex set $V(\mathcal{B}) = \{\mathbf{X}_t : t \in T\}$ and with a directed edge from $X_{j,s}$ into $X_{i,t}$ if and only if $X_{j,s} \in pa(X_{i,t})$, where*

$$pa(X_{i,t}) = \{X_{j,s} : t_0 \leq s \leq t, j \in \{1, 2, \ldots, p\}\}.$$

2. *A set of conditional independence statements of the form*

$$X_{i,t} \perp\!\!\!\perp pr(X_{i,t}) \backslash pa(X_{i,t}) \mid pa(X_{i,t}). \tag{5.3}$$

3. *A set of CPVs associated with $p(x_{i,t}|pa(x_{i,t}))$.*

So there is a directed edge into $X_{i,t}$ from the variables indexed by time t or before it. In practice it is often assumed that the DBN is first-order Markov such that a variable is only affected by variables of the previous time-step and the current time-step, and hence $s = \{t-1, t\}$. Further, the structure of the time-slices is assumed to be time-invariant, such that the dependence structure and associated CPVs of the variables $\mathbf{X}_t$ in time-slice t, given parents from time-slice $t-1$ and t, is the same independent of t. This DBN is known as the two time-slice DBN, which can be simply defined by a BN structure on two time-slices, $t-1$ and t, with an associated set of CPVs for time-slice t with parents from time-slice t or $t-1$ and an initial set of CPVs for time-slice t_0. The two time-slice DBN has gained similar popularity to the BN and has been used in a variety of applications, such

as robotics, speech recognition or in environmental and medical applications. For a better illustration the two time-slice DBN is often 'rolled out' over several time-slices and I will do so for Example 16 when demonstrating how to rewrite a DBN as a DCEG.

To show how a DBN can be written as a DCEG, I will first show how to write the variables of the DBN as an infinite tree. I will then define the conditional independence statements of the DBN by colouring the florets in the tree to form a stage partition of the situations. In order to write a DBN $\mathcal{B}$ as an infinite tree the variables are first ordered so that parents come before children and time-slices come before each other according to their time index. From Equation 5.3 it can be seen that there is always such an ordering due to the acyclicity and the time element of the DBN. The variables of the DBN $\{X_{i,t} : i = 1, \ldots, p, t \in T\}$ can then be re-indexed according to this ordering as X_r, $r = 1, 2, 3, \ldots$ so that whenever $q < r$ then $X_q = X_{j,s} \in pa(X_{i,t})$. As already explained for the non-dynamic case in Chapter 2.4 several orderings of the variables in $\{\mathbf{X}_t : t \in I\}$ may exist from which one is chosen. Given this ordering a corresponding infinite tree of the DBN $\mathcal{B}$ can be constructed, where the variables up to index r can be represented by a finite tree, which is denoted by $\mathcal{T}_r = (V(\mathcal{T}_r), E(\mathcal{T}_r))$, and each path in the tree $\mathcal{T}_r$ represents a particular combination of values that the variables $X_q, q \leq r$ can take. This allows us to recursively define a set of trees $\{\mathcal{T}_r : r \geq 1\}$, where $\mathcal{T}_r$ is a subtree of $\mathcal{T}_{r+1}$, as follows:

Recall that $L(\mathcal{T}_r)$ is the set of leaf vertices of $\mathcal{T}_r$ and denote $l_{rn}, n = 1, 2, ..., N_r$ as a single leaf vertex in $L(\mathcal{T}_r)$, where $\mathcal{T}_r$ has N_r leaf vertices.

1. For $r = 1$, let $\mathcal{T}_1$ be the floret, $F(s_0)$, associated with X_1 which can take m_1 values. Therefore $V(\mathcal{T}_1) = \{s_0, l_{11}, l_{12}, \ldots, l_{1m_1}\}$ and $E(\mathcal{T}_1) = \{e_{s_0k} : k = 1, \ldots, m_1\}$.

2. Given $\mathcal{T}_r = (V(\mathcal{T}_r), E(\mathcal{T}_r))$, define $V(\mathcal{T}_{r+1})$ and $E(\mathcal{T}_{r+1})$ as follows: Let

$$E\left(\mathcal{T}_{r+1}\right) = E\left(\mathcal{T}_r\right) \cup E^+_{r+1},$$

where

$$E^+_{r+1} = \{e_{l_{rn}k} : l_{rn} \in L(\mathcal{T}_r), k = 1, 2, ..., m_{r+1}\} \tag{5.4}$$

is a set of $N_r \times m_{r+1}$ new edges, where m_{r+1} edges emanate from each vertex l_{rn}, each describing the values the random variable X_{r+1} can take. Now attach

a new leaf vertex to each of the edges in E^+_{r+1} and let

$$V^+_{r+1} = \{ch(l_{rn}) : l_{rn} \in L(\mathcal{T}_r)\}. \tag{5.5}$$

Then

$$V(\mathcal{T}_{r+1}) = V(\mathcal{T}_r) \cup V^+_{r+1}.$$

Finally, the infinite tree $\mathcal{T}$ of the DBN $\mathcal{B}$ is defined as

$$\mathcal{T} = \lim_{r\to\infty} \mathcal{T}_r,$$

where the vertex set is

$$V(\mathcal{T}) = V(\lim_{r\to\infty} \mathcal{T}_r)$$

and the edge set is given by

$$E(\mathcal{T}) = E(\lim_{r\to\infty} \mathcal{T}_r).$$

I demonstrate this recursive construction of the infinite tree for an example of a two time-slice DBN below:

Example 16. *Consider two binary variables X_1 and X_2 taking values 0 and 1, which are measured at discrete time points $t \in T$. Further assume that at each time point, X_1 is only affected by the previous value of X_1 and X_2 is affected by both the previous value of X_2 and the current value of X_1. This can be represented by a two time-slice DBN, whose graph is given in Figure 5.6, 'rolled-out' over three time-slices. Its variables can be re-indexed as*

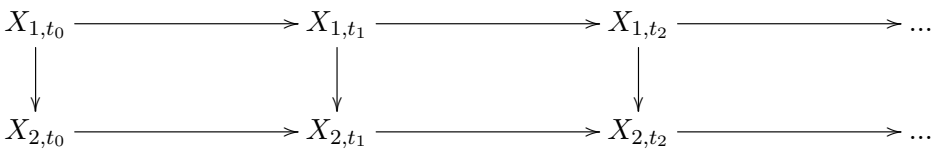
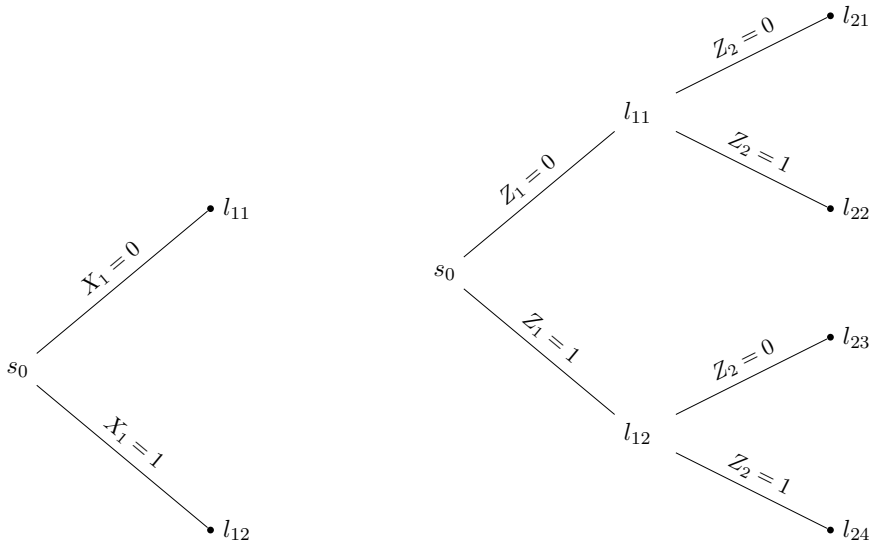


Figure 5.6: A two time-slice DBN structure, $\mathcal{B}$, on two binary variables X_1 and X_2, where at each time point X_1 is only affected by the previous value of X_1 and X_2 is affected by both the previous value of X_2 and the current value of X_1

$$X_1 = X_{1,t_0}, X_2 = X_{2,t_0}, X_3 = X_{1,t_1}, X_4 = X_{2,t_1}, X_5 = X_{1,t_2}, X_6 = X_{2,t_2}. \tag{5.6}$$

As $X_1 = X_{1,t_0}$, $\mathcal{T}_1$ hence corresponds to the tree given in Figure 5.7 (a) with root vertex s_0 and two emanating edges labelled $X_1 = 0$ and $X_1 = 1$. To obtain $\mathcal{T}_2$



(a) Illustration of $\mathcal{T}_1$

(b) Illustration of $\mathcal{T}_2$

Figure 5.7: The illustration of the trees $\mathcal{T}_1$ and $\mathcal{T}_2$ of the two time-slice DBN from Figure 5.6

(Figure 5.7 (b)) from $\mathcal{T}_1$ attach $m_2 = 2$ edges to each leaf vertex of $\mathcal{T}_1$ as defined by Equation 5.4 and attach a child to each new edge as defined in Equation 5.5. Similarly, to obtain $\mathcal{T}_3$ from $\mathcal{T}_2$ attach $m_3 = 2$ edges describing $X_3 = 0$ and $X_3 = 1$ to each leaf of $\mathcal{T}_2$ and attach a new leaf to each new edge. Continuing in this way the infinite tree $\mathcal{T}$ of the DBN $\mathcal{B}$ is obtained as given in Figure 5.8, where again the notation '...' describes the continuation of the process.

Given an infinite tree, the conditional independencies of the DBN can then be represented by a staged tree. (Note that in this example the situations that are in the same stage are coloured rather than their associated edges.) The reasoning is entirely analogous to the non-dynamic case of writing a BN as a CEG: From the construction of the infinite tree we know that the edges emanating from a vertex $l_{rn} \in V(\mathcal{T}_r)$, $(e_{l_{rn}1}, e_{l_{rn}2}, \ldots, e_{l_{rn}m_{r+1}})$, describe the values X_{r+1} can take with the associated CPV being of the form $(\pi_{l_{rn}1}, \pi_{l_{rn}2}, \ldots, \pi_{l_{rn}m_{r+1}})$. Also, the path up to l_{rn} describes a particular history, i.e. a particular combination of values taken by $pr(X_r) \cup X_r$. Hence, l_{rn_1},$l_{rn_2} \in V(\mathcal{T}_r), n_1, n_2 \in N_r$, are in the same stage when

$$P(X_{r+1} = x_{r+1}|l_{rn_1}) = P(X_{r+1} = x_{r+1}|l_{rn_2}),$$

for all values of x_{r+1}, or equivalently,

$$\pi_{l_{rn_1}} = \pi_{l_{rn_2}}. \tag{5.7}$$

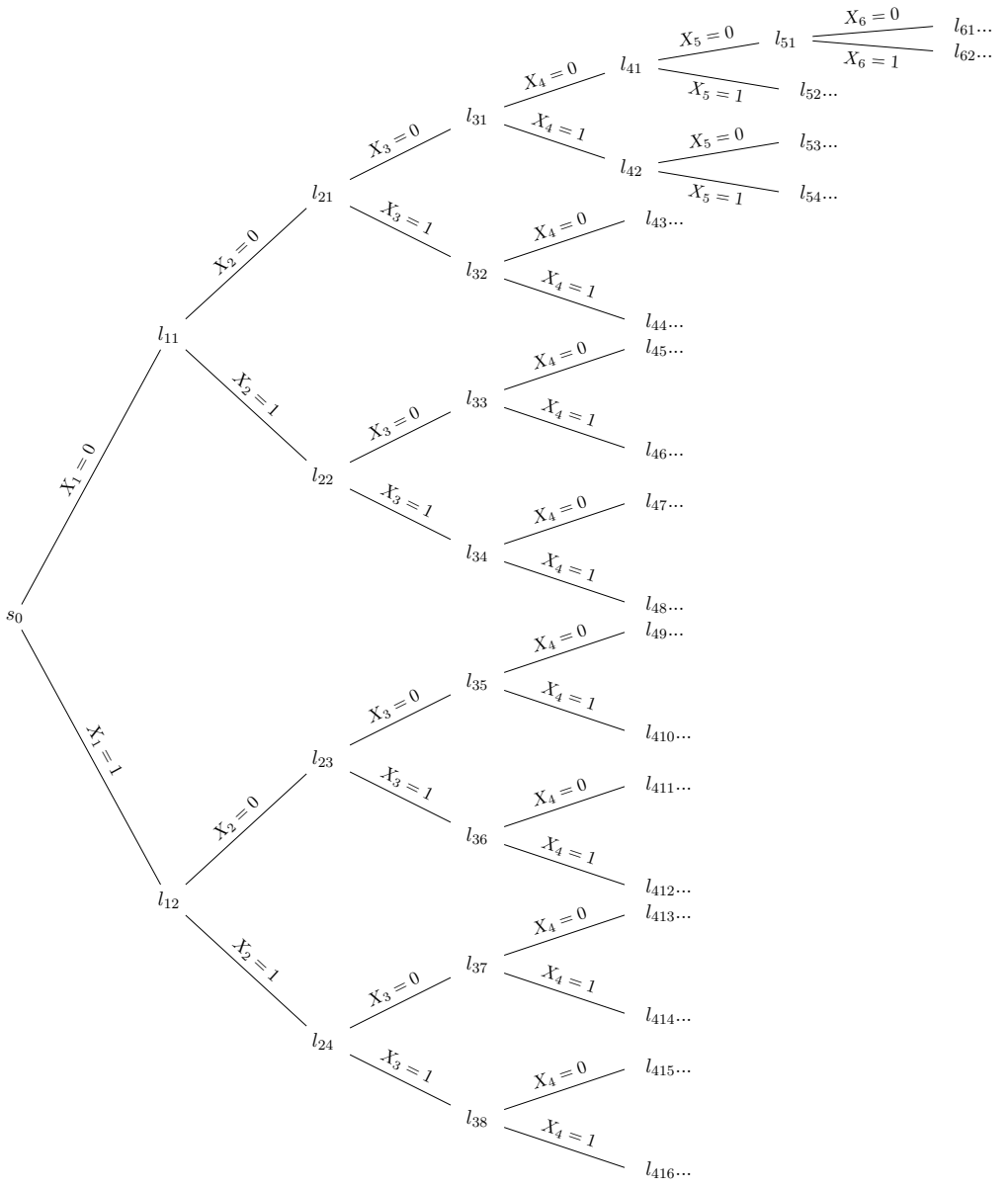


Figure 5.8: The infinite tree representation of the two time-slice DBN from Figure 5.6

By the definition of a DBN (Definition 28)

$$X_{r+1} \perp\!\!\!\perp pr(X_{r+1}) \backslash pa(X_{r+1}) \mid pa(X_{r+1}),$$

and so l_{rn_1},$l_{rn_2} \in V(\mathcal{T})$ are in the same stage whenever the paths leading to l_{rn_1} and l_{rn_2} differ only by the values taken by the variables in $pr(X_{r+1}) \backslash pa(X_{r+1})$. The vertices l_{rn_1} and l_{rn_2} are then assigned the same colour. From the infinite staged tree the corresponding DCEG can then be directly derived as described in Definition 25.

Example 17. *In the example of the two time-slice DBN the following colouring on $\mathcal{T}$ is obtained, as given in Figure 5.9. Observe that the DBN in Figure 5.6 represents the conditional independence statements*

$$X_{1,t} \perp\!\!\!\perp pr(X_{1,t}) \mid X_{1,t-1}, \ \forall t > t_0 \tag{5.8}$$

and

$$X_{2,t} \perp\!\!\!\perp pr(X_{2,t}) \mid X_{2,t-1}, X_{1,t}, \ \forall t > t_0. \tag{5.9}$$

Then by the previous re-indexing (see 5.6), $X_3 \perp\!\!\!\perp X_2 \mid X_1$ and hence $P(X_3 = 0|l_{21}) = P(X_3 = 0|l_{22})$ and $P(X_3 = 1|l_{21}) = P(X_3 = 1|l_{22})$ and, similarly, $P(X_3 = 0|l_{23}) = P(X_3 = 0|l_{24})$ and $P(X_3 = 1|l_{23}) = P(X_3 = 1|l_{24})$. Therefore, l_{21} and l_{22} are in the same stage as well as l_{23} and l_{24}. Similarly, by the same conditional independence given in 5.8, $X_5 \perp\!\!\!\perp X_1, X_2, X_4 \mid X_3$, which is depicted by the colouring of l_{41} and l_{42} as well as l_{43} and l_{44} and so on. Further, 5.9 requires that $X_4 \perp\!\!\!\perp X_1 \mid X_2, X_3$, such that l_{31} and l_{35} are in the same stage, as are l_{32} and l_{36}, l_{33} and l_{37}, and l_{34} and l_{38}. As this is a two time-slice DBN the CPVs of $\mathbf{X}_t$ given its parents from time-slice $t-1$ and t are the same for all $t > t_0$. Consequently, the colouring of l_{41} and l_{42} is identical to the colouring of l_{21} and l_{22} and so on. The DCEG of the infinite staged tree of Figure 5.9 is given in Figure 5.10. The two time-slice structure of the DBN is depicted in the DCEG as follows: The initial variables $X_1 = X_{1,t_0}$ and $X_2 = X_{2,t_0}$ are represented by the paths up to w_3, w_4, w_5 and w_6. After this the paths continue to positions w_7, w_8, w_9 and w_{10} and then loop back to positions w_3, w_4, w_5 and w_6.

In this example, which only depicts standard conditional independencies, the graph of the DCEG is much more complicated and the DBN is topologically much simpler. However, when many of the configurations of the variables are impossible and consequently a large number of zeros within the table of CPVs exist, then the DCEG can be simplified and can be more expressive than the DBN. Consider, for

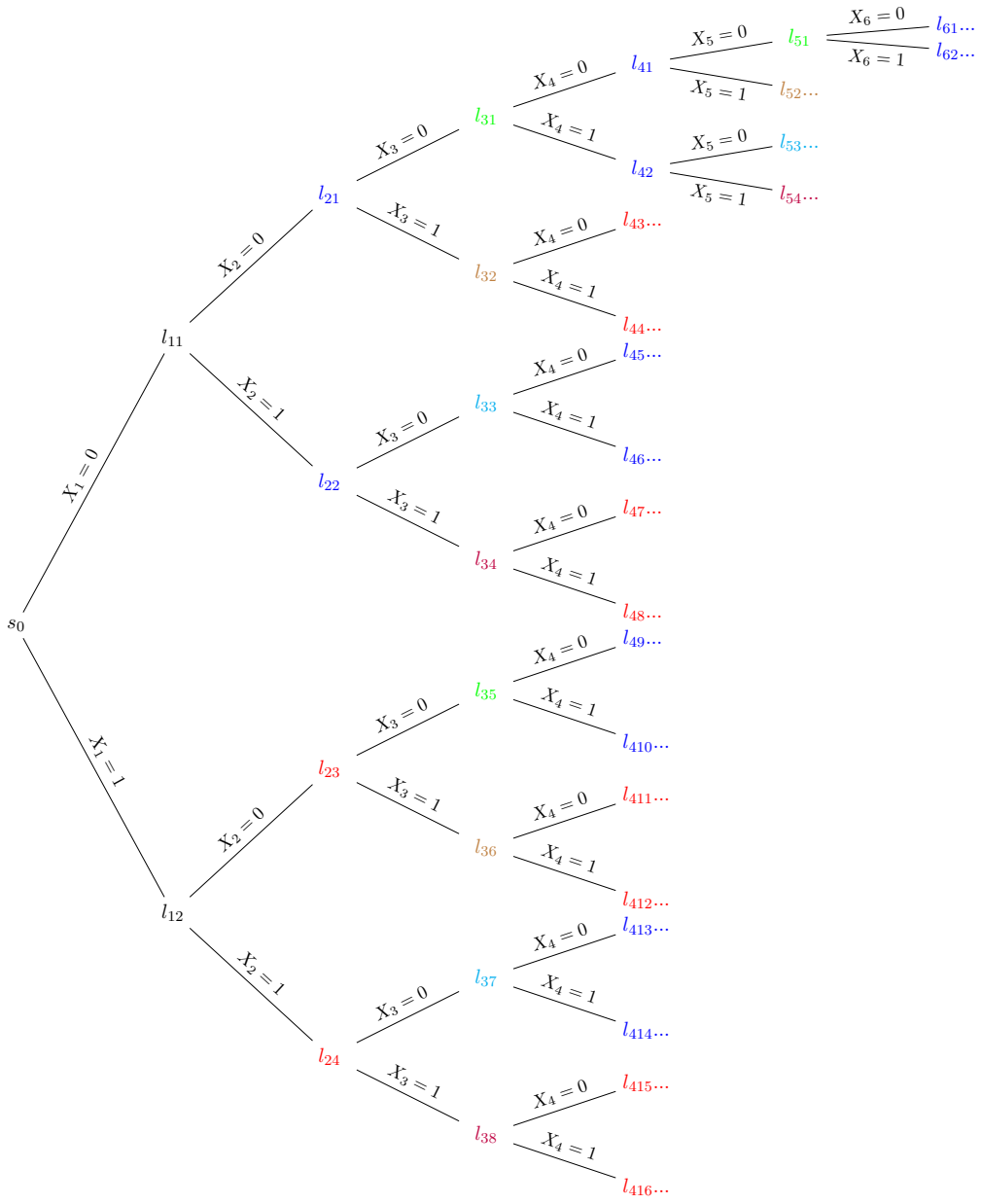


Figure 5.9: The infinite staged tree representation of the two time-slice DBN from Figure 5.6

example Figure 5.9: If the CPVs of the BN state that $P(X_{2,t} = 0|X_{1,t} = 0) = 0$ for $t \in T$ then the edge describing this probability and the entire subtree, $\mathcal{T}(l_{31})$, can be omitted from $\mathcal{T}$ as well as all subtrees whose root is reached by the events $X_{1,t} = 0, X_{2,t} = 0$. Hence, unlike the BN and its dynamic analogue, the DCEG

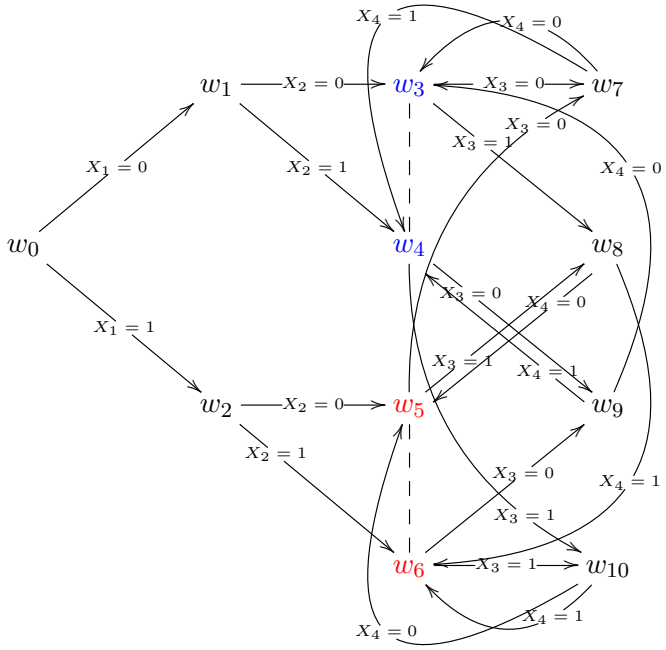


Figure 5.10: The DCEG of the infinite staged tree representation of the two time-slice DBN from Figure 5.6

also allows zeros in the corresponding table of CPVs to be represented by missing edges in its graph. This is particularly useful when representing processes which have many logical constraints.

Recall further from Chapter 2.4 that the BN imposes certain restrictions or symmetries on the stage structure of the CEG. When representing a DBN as an infinite tree these restrictions are directly extended across time-slices by the re-indexing of the variables, where the conditional independence statements of the form $X_{i,t} \perp\!\!\!\perp pr(X_{i,t}) \backslash pa(X_{i,t}) \mid pa(X_{i,t})$ force situations that are reached by the same parent configuration into the same stage. The usual DBN therefore only admits certain very specific stage partitions. The additional restrictions imposed by the two time-slice DBN are also represented within the topology of the graph of the DCEG: Due to the set of CPVs being the same across time-slices, the DCEG simply loops round the positions w_3 to w_{10} for each time-slice t. In contrast to the DBN, the DCEG can further allow asymmetric dependence structures between the variables of a time-slice and also across time-slices. For example, the DCEG could distinguish in the flu example between different conditional independence structures for different influenza viruses. It could depict whether the probability of taking treatment depends not only on whether a virus is caught but also on which virus.

To allow for irregular time-steps in a DBN, Nodelman et al. [2002] suggested

the development of a Continuous-Time BN whose variables evolve continuously over time. The model combines the standard BN with continuous-time Markov processes, where a DAG describes the local dependence structure between the variables and the evolution of each variable is given by a set of conditional Markov processes on $X_i|Pa(x_i) = j$, such that the development of X_i over time depends on the state of its parents. One problem of these models stated in Nodelman et al. [2002] is, however, that exact inference is intractable and approximate techniques need to be used.

Another interesting class of graphical model, related to the DBN, is the local independence graph [Didelez, 2008] or the graphical duration model [Gottard, 2007]. Here it is assumed that data is available on the event history of a group of people, which includes particular events that occur and the time until an event occurs. Data are then modelled by a marked point process, which is given as a pair of random variables $(T_s, E_s), s = 1, 2, \ldots$, where $T_s \in \mathcal{T}, 0 < T_1 < T_2, \ldots$ are the time of the occurrences of the different types of event $E_s \in \mathcal{E}$. The dependence structure between the number of occurrences of each event is then depicted by a local independence graph, where the events are the vertices in the graph and missing edges represent conditional independencies stating that the intensity of a future event is independent of certain past events given other past events. Although the problems modelled by a local independence graph are similar to those that can be expressed by a DCEG with holding time distributions, the dependence structures depicted by a local independence graph are more closely related to the DBN. Considering again the flu example, the DCEG could express the asymmetric dependence structure describing that the probability of taking treatment is different when the individual catches flu for the first time from when he has had flu before. The local independence graph, however, assumes that the conditional independencies do not change with time.

5.3 The DCEG and Markov Processes

In this section I will compare the DCEG to Markov processes. I will show that the Markov process is a particular subclass of the DCEG and illustrate the structural and conceptual differences in their construction. Similarly, I will further prove that a certain subclass of the DCEG with holding time distributions corresponds to a semi-Markov process.

Recall that a discrete-time Markov process [Norris, 1998; Suhov, 2008] can be defined as follows:

Definition 29. *A **discrete-time Markov process** is a discrete-time stochastic process $\{X_n : n \in \mathbb{N}\}$ with discrete state space which satisfies the Markov property, such that*

$$P(X_{n+1} = i_{n+1} | X_n = i_n, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = P(X_{n+1} = i_{n+1} | X_n = i_n).$$

It is assumed throughout that the Markov process is time-homogeneous such that $P(X_{n+1} = j | X_n = i) = p_{ij}, \forall n$, and the Markov process can therefore be defined through an initial distribution α and a transition matrix P with ijth entry p_{ij}.

A state-transition diagram of a Markov process has vertices describing the states of the process and an edge from state i to j labelled with the probability p_{ij}. When the DCEG, as defined in Definition 25, describing a dynamic process, has a finite number of positions, then its topology resembles the state-transition diagram of a discrete-time Markov process, where the positions of the DCEG form the states of the Markov process. Further, by the definition of the DCEG, the CPV associated with a particular stage depends only on the stage reached, i.e. the Markov property, that the transition probability to the next position depends only on the current position, is satisfied.

However, there are several structural differences between the Markov process and the DCEG which demonstrate that the DCEG can provide additional information about a given problem: Firstly, the graph of the DCEG preserves the paths of the infinite tree it is derived from, and, secondly, the DCEG may be coloured, which provides additional information about the stage structure of its corresponding tree. I consider the following two simple Markov processes to demonstrate these differences:

Example 18. *Let $\{X_n : n \in \mathbb{N}\}$ be a discrete-time Markov process on the state space $\{a, b, c\}$ with initial distribution $\alpha = (0.4, 0.4, 0.2)$ and with transition matrix P given by*

$$P = \begin{pmatrix} 0.2 & 0.3 & 0.5 \\ 0.5 & 0.3 & 0.2 \\ 0.5 & 0.3 & 0.2 \end{pmatrix}$$

Its state-transition diagram is given in Figure 5.11. Observe further that the transition probabilities from states b and c are the same. Due to this the DCEG representation gives a different structure, which becomes apparent when looking first at the tree representation of the problem. As the process is infinite, the number of situations of the tree is also infinite. The initial situation s_0, the root of the tree,

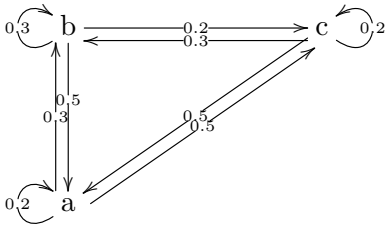


Figure 5.11: The state-transition diagram of a Markov process on the state space (a, b, c) with initial distribution α and transition matrix P

has emanating edges which represent the choice of initial state with associated CPV $\pi_{s_0} = (0.4, 0.4, 0.2)$. *The other situations could be indexed as* $\{s_{i,n}, i = a, b, c, n \in \mathbb{N}\}$ *with CPVs* $\pi_{s_a,n} = (0.2, 0.3, 0.5)$ *and* $\pi_{s_b,n} = \pi_{s_c,n} = (0.5, 0.3, 0.2)$. *It is then immediate that the corresponding DCEG only has three stages and positions with the stage and position partition given by*

$$u_0 = w_0 = \{s_0\}, u_a = w_a = \{s_{a,n}, n \in \mathbb{N}\}, u_{bc} = w_{bc} = \{s_{b,n}, s_{c,n}, n \in \mathbb{N}\}.$$

There is no w_∞ *as all paths are infinite and hence no leaf vertices exist in the tree. The DCEG can then be drawn as given in Figure 5.12 and the associated CPVs are* $\pi_{w_0} = (0.4, 0.4, 0.2)$, $\pi_{w_a} = (0.2, 0.3, 0.5)$ *and* $\pi_{w_{bc}} = (0.5, 0.3, 0.2)$. *For a better comparison the CPVs have here also been attached to the edges of the DCEG.*

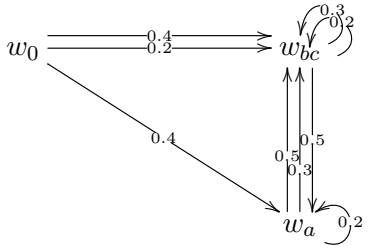


Figure 5.12: DCEG representation of the Markov process with state space (a, b, c), initial distribution α and transition matrix P

Even here, where the process is initially defined through a transition matrix, the graph of the DCEG automatically identifies states which have equivalent roles, here state b being identified with state c, and illustrates the identical conditional probabilities associated with the two states by putting $s_{b,n}$ and $s_{c,n}$, for $n \in \mathbb{N}$ in the same position. The DCEG also depicts explicitly the initial distribution of the process given by the edges emanating from w_0 and acknowledges the initially elicited distinctions of the states b and c through the double edge from w_0 to w_{bc}. This may have important interpretive value, as the DCEG can discover a different partition

of the states of a variable through this or even help to construct new informative variables to represent a problem, as described in Chapter 4. The DCEG further retains the distinction between the two states by the double edge from w_a to w_{bc} and the double loop from w_{bc} into itself.

The next example demonstrates the added expressiveness of the DCEG through the colouring of the graph whenever two or more positions are in the same stage.

Example 19. *A coin is tossed and has probability* $P(H) = \lambda$ *of landing heads and probability* $P(T) = 1 - \lambda = \bar{\lambda}$ *of landing tails. The coin is tossed until three heads have appeared when the game terminates. The DCEG of this example has four positions describing whether* 0, 1, 2 *or* 3 *heads have been tossed and is given in Figure 5.13. As each toss has the same probability* λ *of returning heads, the positions*

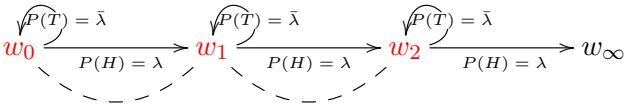


Figure 5.13: DCEG representation of a Markov process describing the game of tossing a coin until heads has appeared three times

w_0, w_1 *and* w_2 *are here also all in the same stage and so they are connected by an undirected dashed line and their edges are coloured.*

The example shows that the additional colouring allows us to identify further symmetries within the transition probabilities between states in a consistent way. In particular, it is the stage structure of the tree that supports the learning of the parameters and the model selection algorithms of the CEG, hence allowing for more efficient learning procedures than obtainable by the position partition. This is analogous in the dynamic case, as will be described at the end of this chapter. To summarise, a Markov process is a DCEG that is simple and that has no two edges leading from the same parent into the same child. Finally, the initial distribution of the Markov process is directly depicted through w_0 and its emanating edges.

However, above all, the elicitation of the DCEG from a tree distinguishes it from a Markov process. As described in the previous section and in Chapter 2.3 the DCEG and CEG are constructed from the description of a process as a tree. It allows us to identify relevant stages, e.g. 'health states' and transitions between these states, either through expert elicitation or model selection methods. In a Markov process these transitions and states are directly defined at the beginning of the analysis, with the state-transition diagram being a graphical representation of this. Therefore, a further possible use of the DCEG could be a method to elicit the states and transitions of the Markov process.

When the DCEG has holding time distributions attached then it is closely linked to semi-Markov processes [Barbu and Limnios, 2008; Medhi, 1994] in a similar way. Semi-Markov processes are a generalisation of Markov processes by allowing for the holding times, describing the length of time spent at the states of the process, to have any distribution, instead of restricting them to having a geometric distribution (discrete-time Markov processes) or an exponential distribution (continuous-time Markov processes). I give the definition of a semi-Markov process similar to Medhi [1994] below:

Definition 30. *Let $\{Y_t : t \geq 0\}$ be a stochastic process with discrete state space and with transitions occurring at times $t_0, t_1, t_2, \ldots$. Also, let X_n describe the state of the process at time t_n, where $Y_t = X_n$ on $t_n \leq t < t_{n+1}$, and let H_n be the holding time before transition to X_n. If*

$$P(X_{n+1} = j, H_{n+1} \leq t | X_0, X_1, .., X_n, H_1, .., H_n) = P(X_{n+1} = j, H_{n+1} \leq t | X_n), \tag{5.10}$$

*then $\{X_n, H_n, n \in \mathbb{N}\}$ is called a **Markov Renewal process** and $\{Y_t : t \geq 0\}$ a **semi-Markov process**. Also, $\{X_n, n \in \mathbb{N}\}$ is the associated **jump process**, which is a discrete-time Markov process with transition matrix P, where*
$p_{ij} = P(X_{n+1} = j | X_n = i)$.

A semi-Markov process is usually specified by an initial distribution α and by its semi-Markov kernel Q whose ijth entry is given by

$$\begin{aligned} Q_{ij}(t) &= P(X_{n+1} = j, H_{n+1} \leq t | X_n = i) \\ &= p_{ij} F_{ij}(t), \end{aligned} \tag{5.11}$$

where

$$F_{ij}(t) = P(H_{n+1} \leq t | X_{n+1} = j, X_n = i)$$

is the conditional holding time distribution, i.e. the holding time at $X_n = i$ assuming that the next state moved to is $X_{n+1} = j$. As before time-homogeneity is assumed and hence the above equations do not depend on the index n. It can then be shown that a particular subclass of the time-homogeneous DCEG with holding times corresponds to a semi-Markov process.

Theorem 1. *Let a DCEG $\mathcal{D}$ with holding times be simple and let no two children lead from the same parent into the same child. Then this DCEG is a semi-Markov process with state space $\mathcal{S} = \{V(\mathcal{D}) \backslash w_0\}$ and with the entries of its transition matrix*

given by

$$p_{ij} = \begin{cases} \pi_{w_i k} & : \textit{ if } e_{w_i k} = e(w_i, w_j) \textit{ exists} \\ 1 & : \textit{ if } w_i = w_j = w_\infty \\ 0 & : \textit{ otherwise,} \end{cases}$$

and with conditional holding time distributions

$$F_{ij}(t) = \begin{cases} P(H_{w_i k} \leq t | e_{w_i k}, w_i) & : \textit{ if } e_{w_i k} = e(w_i, w_j) \textit{ exists} \\ 1 & : \textit{ if } w_i = w_j = w_\infty \\ 0 & : \textit{ otherwise.} \end{cases}$$

If the position w_0 is a source vertex then the initial distribution is given by $\alpha = \pi_{w_0}$. Otherwise the initial distribution assigns probability 1 to w_0 and w_0 is included in the state space.

Proof. Assume we have a DCEG $\mathcal{D}$ which is simple and which has no double edges from one vertex into another. To show that this can be written as a semi-Markov process the state space needs to be defined and the semi-Markov kernel and initial distribution need to be specified.

Define the state space of the semi-Markov process and its jump process to be $\mathcal{S} = \{V(\mathcal{D}) \backslash w_0\}$, the set of positions not including w_0. As no two edges lead from the same parent into the same child each edge is uniquely determined by the two positions it connects. First consider the case where $w_i \neq w_\infty$ and then the case where $w_i = w_\infty$. Note that not every DCEG will have a final position of leaf vertices, in which case the second case does not apply.

Case 1: $w_i \neq w_\infty$**:** If $e(w_i, w_j)$ exists, then the ijth entry of the transition matrix P of the jump process is given by

$$p_{ij} = P(X_{n+1} = w_j | X_n = w_i) = P(e(w_i, w_j) | w_i).$$

Assuming without loss of generality that the kth edge of w_i leads to w_j, then,

$$\begin{aligned} P(X_{n+1} = w_j | X_n = w_i) &= P(e_{w_i k} | w_i) \\ &= \pi_{w_i k} \\ &= \pi_{u_i k} \text{ where } u_i = w_i \text{ as the DCEG is simple.} \end{aligned}$$

The conditional holding time distributions can be derived in a similar way.

Assuming again that the kth edge of w_i leads to w_j

$$\begin{aligned} F_{ij}(t) &= P(H_{n+1} \leq t | X_{n+1} = w_j, X_n = w_i) \\ &= P(H_{w_i k} \leq t | e_{w_i k}, w_i) \\ &= P(H_{u_i k} \leq t) \text{ where } u_i = w_i \text{ as the DCEG is simple.} \end{aligned}$$

By Equation 5.11 the ijth entry of the semi-Markov kernel is then given by $Q_{ij}(t) = p_{ij}F_{ij}(t)$. If $e(w_i, w_j)$ does not exist then the ijth entry of the semi-Markov kernel is zero as no transition from w_i to w_j occurs.

Case 2: $w_i = w_\infty$: When $w_i = w_\infty$, then the individual stays in w_∞ forever once reaching this state and hence $Q_{ij}(t) = 1$ when $w_j = w_\infty$ and 0 otherwise.

When w_0 in the DCEG is a source node and no edges lead back to w_0, so that it solely serves as a starting point of the process, then the initial distribution of the corresponding semi-Markov process is given by $\alpha = \pi_{w_0} = \pi_{u_0}$. If w_0 can be reached again throughout, then w_0 is included in the state space and the initial distribution of the semi-Markov process assigns w_0 probability 1. □

I have shown in this section that when a DCEG is simple and no two edges lead from the same parent into the same child, then it corresponds to a Markov process and to a semi-Markov process when holding times are added to the DCEG. I have further demonstrated that the construction of the DCEG from an infinite staged tree provides further information about the dynamic process than the state-transition diagram of a Markov process. By eliciting symmetries from the infinite tree, different states within a Markov process with the same transition probabilities may be merged into one state or coloured when only the immediate transition probabilities are the same. For example, in the flu example, we may distinguish between two different types of flu: virus B and virus C. It could then be determined, by eliciting the stage structure of the problem, whether the effect of treatment and recovery is the same for both flu viruses, hence leading to a double edge from w_0 to w_1 labelled with the two different viruses, or whether the viruses react entirely differently such that w_0 has two emanating edges leading to two different subgraphs and hence a more complex DCEG structure. Alternatively the viruses may only react differently with respect to treatment but otherwise develop in the same way, which would result in a coloured DCEG graph.

The correspondence between (semi-)Markov processes and DCEGs can further be very useful as many of the well-developed results on Markov processes could

be extended to the DCEG. From Equation 5.1 the probability of staying at a position w for a time $\leq h$ and then moving along the edge e_{wk} can be calculated. This equation corresponds to the entries of the semi-Markov kernel (Equation 5.11) of a semi-Markov process. Then, for example, Barbu and Limnios [2008] or Kulkarni [1995] have shown how to derive the transition matrix of the semi-Markov process from the semi-Markov kernel, in order to calculate the probability of being in state j at time t given that we are initially in state i. These types of calculations could be directly extended to the DCEG. This would further enable the DCEG to be applicable to the wide-ranging domain of semi-Markov processes, which includes reliability theory, finance and insurance or traffic modelling.

Within Health Economic Decision making it is desirable to measure the cost-effectiveness of health-care interventions and their clinical impact simultaneously. To do so Cooper et al. [2003, 2007] developed a Bayesian approach using MCMC methods such as Gibbs sampling. The graphical representations given in these papers resemble the graphs of the DCEG given for the flu example in this chapter suggesting a further possible application of the DCEG worth investigating.

5.4 Bayesian Learning of DCEGs

In the final section of this chapter I will extend the methodology of learning parameters in a CEG to the DCEG. I assume throughout that the number of stages in the DCEG is finite, as in the example given in the previous sections of this chapter. In a DCEG the stage parameters, $\boldsymbol{\pi}_u$ can be learnt exactly analogously to learning in a CEG, which was described in Chapter 3.1, where each $\boldsymbol{\pi}_u$ is assigned a prior Dirichlet distribution and is updated in closed form by counting the number of individuals going along each edge in the DCEG. As introduced in Section 5.1.2, a vector of conditional holding time distributions $(F_{u1}, F_{u2}, \ldots, F_{um_u})$ may further be attached to each stage u in the DCEG to express the time spent at each position before going along a particular edge in the graph. Denote the parameter of the distribution F_{uk} by λ_{uk} and call the full set of parameters $\boldsymbol{\lambda} = \{\lambda_{uk}, u \in U, k = 1, \ldots m_u\}$. I will first show in this section how the likelihood of $\boldsymbol{\pi}$ and $\boldsymbol{\lambda}$ given a complete random sample separates and that, under the assumption that $\boldsymbol{\pi}$ and $\boldsymbol{\lambda}$ are a priori independent, these can be learnt separately. I will then discuss conjugate learning on $\boldsymbol{\pi}$ and $\boldsymbol{\lambda}$ and will illustrate this on the already familiar flu example. At the end of this section I will briefly show how different DCEG structures could be scored and compared using the marginal likelihood of the DCEG structure given the data.

Given a DCEG $\mathcal{D}$, for each individual that traverses the DCEG, the edges he

passes along can be recorded as well as the holding times at each position. Assume the individual s takes the path $(e_{w_{i_0}k_0}, e_{w_{i_1}k_1}, \ldots, e_{w_{i_{n_s}}k_{n_s}})$ along n_s+1 edges starting at $w_{i_0} = w_0$. Then, following the notation from Chapter 2.3, let $w^s_{i_a}$ describe the ath position reached by individual s and $h^s_{i_a}$ the holding time at position $w^s_{i_a}$ and $e^s_{i_a k_a}$ the ath edge passed along, where $a = 0, 1, \ldots, n_s$. Then, by the definition of a DCEG (see Definition 27) the likelihood, given an individual s, with path ϵ^s and vector of holding times $\mathbf{h}^s = (h^s_{i_0}, h^s_{i_1}, \ldots, h^s_{i_n})$, is given by

$$\begin{aligned} L(\boldsymbol{\pi}, \boldsymbol{\lambda}|\epsilon^s, \mathbf{h}^s, \mathcal{D}) &= \prod_{a=0}^{n_s} p(e^s_{i_a k_a}, h^s_{i_a}|w^s_{i_a}) \\ &= \prod_{a=0}^{n_s} \pi_{w^s_{i_a}k_a} f_{w^s_{i_a}k_a}(h^s_{i_a}). \end{aligned}$$

This can now be generalised to a complete random sample $\mathcal{S}$ of n individuals going through the tree to obtain the likelihood

$$L(\boldsymbol{\pi}, \boldsymbol{\lambda}|\mathcal{S}, \mathcal{D}) = \prod_{s=1}^{n} L(\boldsymbol{\pi}, \boldsymbol{\lambda}|\epsilon^s, \mathcal{D}) = \prod_{s=1}^{n}\prod_{a=0}^{n_s} \pi_{w^s_{i_a}k_a} f_{w^s_{i_a}k_a}(h^s_{i_a}).$$

As for the CEG this likelihood can be rewritten by counting the number of times the individuals pass through a position $w \in u$ and go along the kth edge, $k = 1, .., m_u$, which is denoted by N_{uk}. Let $\mathbf{h}_{uk}$ be the vector of conditional holding times for the individuals who arrive at stage u and move along the kth edge next and let h_{ukl} be the holding time of the lth pass along this edge. Denote the full set of holding times by $\mathbf{h} = \{\mathbf{h}_{uk}, u \in U, k = 1, .., m_u\}$ and the set of the number of times each edge is taken by $\mathbf{N} = \{N_{uk}, u \in U, k = 1, .., m_u\}$. The likelihood of $\boldsymbol{\pi}$ and $\boldsymbol{\lambda}$ given a complete random sample and a DCEG $\mathcal{D}$ is therefore given by

$$L(\boldsymbol{\pi}, \boldsymbol{\lambda}|\mathbf{N}, \mathbf{h}, \mathcal{D}) = \prod_{u\in U}\prod_{k=1}^{m_u} \pi_{uk}^{N_{uk}} \prod_{l=1}^{N_{uk}} f_{uk}(h_{ukl}), \tag{5.12}$$

where the individuals go N_{uk} times along edges $e_{wk}, w \in u$ each time staying for a time h_{ukl} at the previous position. Then, immediately from Equation 5.12 the likelihood $L(\boldsymbol{\pi}, \boldsymbol{\lambda}|\mathbf{N}, \mathbf{h}, \mathcal{D})$ of a complete random sample separates. Explicitly, we have that

$$L(\boldsymbol{\pi}, \boldsymbol{\lambda}|\mathbf{N}, \mathbf{h}, \mathcal{D}) = L_1(\boldsymbol{\pi}|\mathbf{N}, \mathcal{D}) \times L_2(\boldsymbol{\lambda}|\mathbf{h}, \mathbf{N}, \mathcal{D})$$
$$= \prod_{u \in U} \prod_{k=1}^{m_u} \pi_{uk}^{N_{uk}} \times \prod_{u \in U} \prod_{k=1}^{m_u} \prod_{l=1}^{N_{uk}} f_{uk}(h_{ukl}). \tag{5.13}$$

If $\boldsymbol{\lambda}$ and $\boldsymbol{\pi}$ are believed to be a priori independent so that

$$p(\boldsymbol{\pi}, \boldsymbol{\lambda}|\mathcal{D}) = p_1(\boldsymbol{\pi}|\mathcal{D}) p_2(\boldsymbol{\lambda}|\mathcal{D}),$$

then $p_1(\boldsymbol{\pi}|\mathcal{D})$ and $p_2(\boldsymbol{\lambda}|\mathcal{D})$ can be updated independently using $L_1(\boldsymbol{\pi}|\mathbf{N}, \mathcal{D})$ and $L_2(\boldsymbol{\lambda}|\mathbf{h}, \mathbf{N}, \mathcal{D})$ respectively, to obtain the posterior density

$$p(\boldsymbol{\pi}, \boldsymbol{\lambda}|\mathbf{h}, \mathbf{N}, \mathcal{D}) = p_1(\boldsymbol{\pi}|\mathbf{N}, \mathcal{D}) p_2(\boldsymbol{\lambda}|\mathbf{h}, \mathbf{N}, \mathcal{D}),$$

which also separates. Therefore the updating of the stage parameters $\boldsymbol{\pi}$ and the holding time parameters $\boldsymbol{\lambda}$ can be performed without reference to the other.

I consider first the updating of $p_1(\boldsymbol{\pi}|\mathcal{D})$. As mentioned in the introduction of this section, this is completely analogous to learning the stage parameters in a CEG and therefore we put a Dirichlet prior on the stage priors of the form

$$p_1(\boldsymbol{\pi}|\mathcal{D}) = \prod_{u \in U} p_u(\boldsymbol{\pi}_u|\mathcal{D}) = \prod_{u \in U} \frac{\Gamma(\sum_{k=1}^{m_u} \alpha_{uk})}{\prod_{k=1}^{m_u} \Gamma(\alpha_{uk})} \prod_{k=1}^{m_u} \boldsymbol{\pi}_{uk}^{\alpha_{uk}-1},$$

assuming, as for the CEG, that the stage priors are a priori independent (compare Equation 3.7 of Chapter 3). Recall from Chapter 3.1, that in a CEG the hyperparameters of the stage priors α_{uk} are found by specifying the parameters of the CEG, $\mathcal{C}_0$, where every situation is in a separate stage. In the simplest case all paths in the associated tree are assumed to be a priori equally likely and the priors of any other CEG $\mathcal{C}$ is determined by summing the parameters of the situations that are in the same stage. Given that the DCEG $\mathcal{D}$ has a sink node w_∞, which can be eventually reached from any position, the hyperparameters of the DCEG can be found in a similar way: The priors on the finest partition $\mathcal{D}_0$ are determined and the hyperparameters of the situations that are merged in $\mathcal{D}$ are summed to find the hyperparameters of $\mathcal{D}$. When situations on the same path are merged into a position then the summing of the hyperparameters corresponds to finding the limit of a geometric series. I will show below how to find the priors of a DCEG $\mathcal{D}$ for the flu example.

Example 20. *Recall the DCEG of the variant of the flu example, repeated in Figure 5.14, whose stage and position partition is given in 5.14.*

$$w_0 = u_0 = \{s_0, s_4, s_7, \ldots\}, w_1 = u_1 = \{s_1, s_{10}, s_{11}, \ldots\},$$
$$w_2 = u_2 = \{s_2, s_6, \ldots\}, w_3 = u_3 = \{s_3, \ldots\}, w_\infty = \{l_5, l_8, l_9, \ldots\}. \tag{5.14}$$

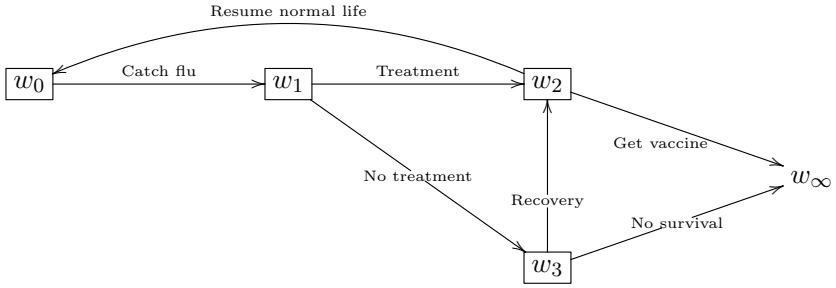


Figure 5.14: The DCEG with holding times for the flu example, where catching flu is described by the time spent at the root vertex

As w_∞ exists the hyperparameters of the stage priors can be found by the standard approach of summing the hyperparameters of the situations in each stage. To specify the prior Dirichlet distribution a slightly larger equivalent sample size of 14 is assumed to ensure later that the holding time distributions have a mean. For example, from Equation 5.14 we have that $u_1 = \{s_1, s_{10}, s_{11}, \ldots\}$. Under the assumption that the paths in the tree (Figure 5.4) are a priori equally likely the situations in u_1 have the distributions: $s_1 \sim Dir(7,7), s_{10} \sim Dir(7 \times \frac{1}{4}, 7 \times \frac{1}{4}), s_{11} \sim Dir(7 \times \frac{1}{8}, 7 \times \frac{1}{8})$. Similarly, the next situations of u_1, which are not explicitly represented in the associated infinite tree, will have the distributions $Dir(7 \times \frac{1}{4}^2, 7 \times \frac{1}{4}^2), Dir(7 \times \frac{1}{8}^2, 7 \times \frac{1}{8}^2)$, etc. Summing the hyperparameters of these situations gives the following two geometric series:

$$7 + 7 \times \frac{1}{4} + 7 \times \left(\frac{1}{4}\right)^2, \ldots = \frac{7}{1 - \frac{1}{4}} = 9\frac{1}{3},$$

$$\frac{7}{8} + \frac{7}{8} \times \frac{1}{8} + \frac{7}{8} \times \left(\frac{1}{8}\right)^2, \ldots = \frac{\frac{7}{8}}{1 - \frac{1}{8}} = 1.$$

Hence, we can deduce that the distribution of π_{u_1} is $Dir(10\frac{1}{3}, 10\frac{1}{3})$. The hyperparameters of the remaining priors on u_2 and u_3 can be found in a similar way to be $\pi_{u_2} \sim Dir(6\frac{2}{3}, 6\frac{2}{3})$ and $\pi_{u_3} \sim Dir(4,4)$. The distribution of u_0 is trivial, assigning probability 1 to the edge $e(w_0, w_1)$.

If we do not have an absorbing position, then summing the hyperparameters

of the stages merged results in infinitely large priors. In Barclay et al. [2013c] it is shown that a possible alternative approach could be taken which makes use of the direct correspondence between the DCEG and Markov processes, where the limiting distribution of the process is found and the hyperparameters are derived from this. Having set up the priors, these can again be updated in closed form, given a complete random sample, to obtain

$$p(\boldsymbol{\pi}|\mathbf{N},\mathcal{D}) = \prod_{u\in U} p_u(\boldsymbol{\pi}_u|\boldsymbol{N}_u,\mathcal{D}) = \prod_{u\in U} \frac{\Gamma(\sum_{k=1}^{m_u}(\alpha_{uk}+N_{uk}))}{\prod_{k=1}^{m_u}\Gamma(\alpha_{uk}+N_{uk})} \prod_{k=1}^{m_u} \pi_{uk}^{\alpha_{uk}+N_{uk}-1}. \tag{5.15}$$

Next the updating of $p(\boldsymbol{\lambda}|\mathcal{D})$ to $p(\boldsymbol{\lambda}|\mathbf{h},\mathbf{N},\mathcal{D})$ using the second component of the likelihood, $L_2(\boldsymbol{\lambda}|\mathbf{h},\mathbf{N},\mathcal{D})$ is considered. Recall from Equation 5.13 that the likelihood for a complete random sample is given by

$$L_2(\boldsymbol{\lambda}|\mathbf{h},\mathbf{N},\mathcal{D}) = \prod_{u\in U}\prod_{k=1}^{m_u}\prod_{l=1}^{N_{uk}} f_{uk}(h_{ukl}),$$

where N_{uk} is the number of times stage u is reached followed by going along edge e_{uk} and $h_{ukl}, l = 1,\ldots,N_{uk}$ are the conditional holding times associated with this edge. Note that, equivalently, this can be written as

$$L_2(\boldsymbol{\lambda}|\mathbf{h},\mathbf{N},\mathcal{D}) = \prod_{u\in U}\prod_{k=1}^{m_u} L_{uk}(\boldsymbol{\lambda}_{uk}|\mathbf{h}_{uk},\mathbf{N}_{uk},\mathcal{D}),$$

where the components $L_{uk}(\boldsymbol{\lambda}_{uk}|\mathbf{h}_{uk},\mathbf{N}_{uk},\mathcal{D})$ of the likelihood can be described by different holding time distributions as mentioned at the end of Section 5.1.2. In the simplest case the likelihood could be a product of exponential distributions, where each component takes the form

$$L_{uk}(\lambda_{uk}|\mathbf{h}_{uk},\mathbf{N}_{uk},\mathcal{D}) = \prod_{l=1}^{N_{uk}} \frac{1}{\lambda_{uk}} exp\left(-\frac{1}{\lambda_{uk}}h_{ukl}\right). \tag{5.16}$$

A more general case would be to assume that the holding times have a Weibull distribution with known shape parameter, K. The likelihood then takes the form

$$L_{uk}(\lambda_{uk}|\mathbf{h}_{uk},\mathbf{N}_{uk},\mathcal{D}) = \prod_{l=1}^{N_{uk}} \frac{K}{\lambda_{uk}^K} h_{ukl}^{K-1} exp\left(-\left(\frac{h_{ukl}}{\lambda_{uk}}\right)^K\right). \tag{5.17}$$

Other possibilities would be to use a log-normal distribution or a log-logistic distribution, when a unimodal density function is appropriate. To learn the parameters

of the conditional holding time distributions, the priors λ_{uk} are assumed to be mutually independent. Of course in certain contexts this assumption may not be appropriate. For example it may be plausible to assume that the time until catching flu has an effect on the time until recovery and in this case the parameters would depend on each other. However, as discussed in Section 5.1.2, here the simplest case is considered, where the CPV and the conditional holding times associated with a stage are assumed to be unaffected by previous holding times, and hence the above assumption is appropriate. Therefore,

$$p_2(\boldsymbol{\lambda}|\mathcal{D}) = \prod_{u \in U} \prod_{k=1}^{m_u} p_{uk}(\lambda_{uk}|\mathcal{D}).$$

Putting an Inverse-Gamma prior, $IG(\alpha_{uk}, \beta_{uk})$ on λ_{uk}, which takes the form

$$p_{uk}(\lambda_{uk}|\mathcal{D}) = \frac{\beta_{uk}^{\alpha_{uk}}}{\Gamma(\alpha_{uk})} \lambda_{uk}^{-\alpha_{uk}-1} exp\left(-\frac{\beta_{uk}}{\lambda_{uk}}\right), \tag{5.18}$$

these parameters of the conditional holding time distributions can then be updated separately and in closed form given a random sample of exponentially distributed holding times as shown in Equation 5.16. This gives Inverse-Gamma posteriors of the form:

$$p_{uk}(\lambda_{uk}|\mathbf{h}_{uk}, \mathcal{D}) = \frac{(\beta_{uk} + \sum_{l=1}^{N_{uk}} h_{ukl})^{\alpha_{uk}+N_{uk}}}{\Gamma(\alpha_{uk} + N_{uk})} \lambda_{uk}^{-\alpha_{uk}-N_{uk}-1} exp\left(-\frac{\beta_{uk} + \sum_{l=1}^{N_{uk}} h_{ukl}}{\lambda_{uk}}\right). \tag{5.19}$$

Similarly, given a Weibull likelihood with known shape parameter K as in Equation 5.17, an inverse-Gamma prior on $\lambda_{uk}^K \sim IG(\alpha_{uk}, \beta_{uk})$ again allows for a conjugate analysis to a posterior $IG(\alpha_{uk}+N_{uk}, \beta_{uk}+\sum_{l=1}^{N_{uk}} h_{ukl}^K)$. Further, if a log-normal likelihood is chosen with known precision, then conjugacy is again obtained by putting a Normal prior on λ_{uk}. When conjugacy cannot be obtained, MCMC methods can be used to find the corresponding posterior distribution.

Example 21. *Recall again the DCEG of the flu example given in Figure 5.14. In this example, it may be plausible to assume an exponential distribution on $H_{u_0 1}$, which describes the time until catching flu, with scale parameter $\lambda_{u_0 1}$, the average time until the individual gets ill. Further it could be assumed that $H_{u_1 1}$ has the more general Weibull distribution, with scale parameter $\lambda_{u_1 1}$ and with known shape parameter $K_1 > 1$, describing the time until taking treatment and recovering. As $K_1 > 1$ it is assumed that the recovery rate increases with time. The time until the individual decides not to take the treatment could again be exponentially distributed with scale parameter $\lambda_{u_1 2}$, i.e. it is assumed to occur at a constant rate. Similarly*

to $H_{u_1 1}$, $H_{u_3 1}$ could also have a Weibull distribution with known shape parameter $K_2 > 1$. In contrast to this, $H_{u_3 2}$ could have a Weibull distribution with scale parameter $\lambda_{u_3 2}$ and known shape parameter $K_3 < 1$ indicating that the death rate decreases with time. The holding times $H_{u_2 1}$ and $H_{u_2 2}$ could again have exponential distributions with parameters $\lambda_{u_2 1}$ and $\lambda_{u_2 2}$ respectively. Here the time until getting the vaccine or resuming a normal life is measured. The holding time distributions and their parameter priors together with the stage priors are given in Table 5.1.

Description	Holding time distribution	Prior
Time until catching flu	$H_{u_0 1} \sim Exp(\lambda_{u_0 1})$	$\lambda_{u_0 1} \sim IG(20\frac{1}{3}, 19\frac{1}{3})$
Take treatment	$N_{u_1} \sim Mult(\pi_{u_1})$	$\pi_{u_1} \sim Dir(10\frac{1}{3}, 10\frac{1}{3})$
Time until recovery with treatment	$H_{u_1 1} \sim Weibull(\lambda_{u_1 1}, K_1)$	$\lambda_{u_1 1}^{K_1} \sim IG(10\frac{1}{3}, 9\frac{1}{3})$
Time until decide against treatment	$H_{u_1 2} \sim Exp(\lambda_{u_1 2})$	$\lambda_{u_1 2} \sim IG(10\frac{1}{3}, 9\frac{1}{3})$
Recovery	$N_{u_3} \sim Mult(\pi_{u_3})$	$\pi_{u_3} \sim Dir(4, 4)$
Time until recovery	$H_{u_3 1} \sim Weibull(\lambda_{u_3 1}, K_2)$	$\lambda_{u_3 1}^{K_2} \sim IG(4, 3)$
Time until death	$H_{u_3 2} \sim Weibull(\lambda_{u_3 2}, K_3)$	$\lambda_{u_3 2}^{K_3} \sim IG(4, 3)$
Get vaccine	$N_{u_2} \sim Mult(\pi_{u_2})$	$\pi_{u_2} \sim Dir(6\frac{2}{3}, 6\frac{2}{3})$
Time until resume normal life	$H_{u_2 1} \sim Exp(\lambda_{u_2 1})$	$\lambda_{u_2 1} \sim IG(6\frac{2}{3}, 5\frac{2}{3})$
Time until vaccine taken	$H_{u_2 2} \sim Exp(\lambda_{u_2 2})$	$\lambda_{u_2 2} \sim IG(6\frac{2}{3}, 5\frac{2}{3})$

Table 5.1: Prior distributions on CPVs and conditional holding times associated with the DCEG from Figure 5.14 for the flu example

If Inverse-Gamma priors on $\lambda_{u_0 1}$, $\lambda_{u_1 1}^{K_1}$, $\lambda_{u_1 2}$, $\lambda_{u_2 1}$, $\lambda_{u_2 2}$, $\lambda_{u_3 1}^{K_2}$ and $\lambda_{u_3 2}^{K_3}$ are assumed, a conjugate analysis as described above can be carried out. The priors can be specified by assuming a prior mean equal to 1 for all prior holding times and an equivalent sample size corresponding to the strength of the prior belief on the edge associated with each conditional holding time distribution (see Table 5.1). Then, given a complete random sample of individuals going through the DCEG for a certain length of time, the number of times, N_{uk}, each edge, e_{uk}, is reached can be recorded, as well as the time spent at each position before moving along a particular edge. By Equations 5.15 and 5.19 the prior distributions on $\boldsymbol{\pi}$ and $\boldsymbol{\lambda}$ could then be updated in closed form and the CPVs and expected time spent at each position, before moving along a certain edge, calculated.

When the learning of the parameters can be carried out in closed form then the Bayesian Dirichlet scoring methods of Chapter 3.2 will also be similarly efficient for the DCEG. Recall from Chapter 3.2 that the BD metric is given by the prior of the CEG structure and the marginal likelihood of the structure given a complete random sample. Similarly, the BD metric for a DCEG structure given a complete random sample is given by

$$p(\mathcal{D})L(\mathcal{D}|\mathbf{h}, \mathbf{N}). \tag{5.20}$$

Assuming that all DCEG structures are a priori equally likely, the marginal likelihood, $L(\mathcal{D}|\mathbf{h},\mathbf{N})$ again suffices as a model selection criteria. When the likelihood separates, as in the above situations, then the marginal likelihood also separates into two parts, one associated with the transitions and another with the holding times:

$$L(\mathcal{D}|\mathbf{h},\mathbf{N}) = L_1(\mathcal{D}|\mathbf{N})L_2(\mathcal{D}|\mathbf{h},\mathbf{N}). \tag{5.21}$$

Then, exactly analogously to the finite CEG, the first component of the marginal likelihood of a DCEG takes the form (compare Equation 3.12):

$$L_1(\mathcal{D}|\mathbf{N}) = \prod_{u\in U} \frac{\Gamma(\sum_{k=1}^{m_u}\alpha_{uk})}{\Gamma(\sum_{k=1}^{m_u}\alpha_{uk}+N_{uk})} \prod_{k=1}^{m_u} \frac{\Gamma(\alpha_{uk}+N_{uk})}{\Gamma(\alpha_{uk})}.$$

When the holding times are either geometric or degenerate, then model selection of the DCEG is analogous to model selection of the CEG, where the models are scored according to (3.12) and the model selection algorithms discussed in Chapter 3.2 can be used. When we have a DCEG with holding times then the second component of the marginal likelihood, $L_2(\mathcal{D}|\mathbf{h},\mathbf{N})$, also needs to be calculated. For example, given only exponential holding times, the second component of the marginal likelihood takes the form

$$L_2(\mathcal{D}|\mathbf{h},\mathbf{N}) = \prod_{u\in U}\prod_{k=1}^{m_u} \frac{\beta_{uk}^{\alpha_{uk}}}{\Gamma(\alpha_{uk})} \frac{\Gamma(\alpha_{uk}+N_{uk})}{\beta_{uk}+\sum_{l=1}^{N_{uk}} h_{ukl}^{\alpha_{uk}+N_{uk}}},$$

and the full marginal likelihood, as given in Equation 5.21, can be used to score and compare possible DCEG structures. Techniques for searching the model space to find the MAP DCEG structure are currently being developed. The most prominent difficulty here is that, without further constraints, the size of the model class of DCEGs is extremely large and so techniques need to be developed that search efficiently over the model space, using for example algorithms such as the dynamic programming algorithm by Silander and Leong [2013] or by restricting the model space a priori (see Chapter 3.2).

I have in this chapter developed a formal representation for a dynamic version of the CEG. I have demonstrated that it usefully generalises the discrete DBN and I believe that it provides a valuable complementary tool to alternative dynamic graphical models. It is particularly suited to domains where the number of categories of the variables is large but the associated transitions are sparse, leading to a large number of zeros in the associated CPVs, or to domains where context-specific symmetries are present. Further, the link of the DCEG to (semi-)Markov processes

suggests that the methodology developed for these processes could be usefully exploited by the DCEG. The learning of the parameters of the DCEG has shown to be a straightforward generalisation of the non-dynamic case as discussed in Chapter 3 and suggests that model selection techniques adapted from the current techniques for CEGs can be developed.

Chapter 6

Discussion

In the concluding chapter I will first summarise the contributions made by this thesis. In Sections 6.2 and 6.3 I will then briefly discuss two areas of research, which have not been covered in the previous chapters, but which need careful consideration when carrying out model selection within a Bayesian framework. Section 6.2 discusses the use of informative priors with respect to the UKCP study and Section 6.3 looks at problems associated with sparse cell counts. Finally, I will conclude with possible areas of future research on CEGs.

6.1 Summary

In this thesis I have developed a number of new aspects of the CEG motivated by the application of the CEG to various health studies. I have shown in Chapter 2 how the CEG provides an improved graphical framework to the BN for the description of an unfolding sequence of events, particularly when asymmetric dependence structures arise. The results in Chapters 2 and 3 demonstrated on the CHDS that the CEG search can lead to a significantly higher scoring model and further that it lets us draw plausible additional conclusions from its graph.

The focus of this thesis has been predominantly concerned with applications, where the effect of risk factors in a health study on a variable of interest, such as physical health or survival, is considered. Hence, apart from the additional conclusions that can be drawn from the CEG, I have also introduced the Ordinal and Reduced Ordinal CEG, which enhance the graphical representation of the CEG for a binary outcome. In Chapter 4 this framework has been exploited to represent processes where missingness is influential and data cannot plausibly be hypothesised to be MAR in all situations. In particular, I have shown that it is often possible to

draw informative conclusions on the missingness structure directly from the CEG, supported by calculations on the CPVs. With respect to the UKCP study the CEG provided further insight into the way in which the severity and the number of impairments, including missing values, influence survival. I have demonstrated that the final positions in the Ordinal CEG can be usefully employed to make informed decisions about the construction of new covariates with informative categories which can be used in a later analysis.

Finally, in Chapter 5 I have defined a new dynamic version of the CEG, the DCEG. This new class of models extends the semantics of the CEG, which had so far only been developed for finite trees, to the infinite tree and is further able to model the time spent at each stage or position in the graph. I have shown that the discrete DBN as well as the state-transition diagrams of Markov and semi-Markov processes are included in this class of models, suggesting that the DCEG can provide a useful complementary tool to alternative graphical models for modelling longitudinal processes with asymmetric dependence structures. The introduction to parameter learning and model selection for DCEGs at the end of that chapter proposes that the methods developed in Chapters 2 and 3 could be directly extended to the DCEG.

6.2 CEGs with Informative Priors

Throughout the thesis I have assumed a uniform prior across the paths in the tree, such that each path is a priori equally likely. Further, I have consistently picked a small equivalent sample size to demonstrate weak prior beliefs. Ideally, in a full analysis, domain knowledge should be brought in, where prior information is obtained through expert elicitation or from the available literature on previous studies. This could then be incorporated into the model through informative priors. Details on how expert elicitation is practised is discussed for example in O'Hagan et al. [2006]. To illustrate how external information can be incorporated into a CEG, I consider the example on three impairments and survival in the UKCP study described in Chapter 4.4.

Throughout the thesis the final situations in the tree, whose associated edges describe survival, have taken Dirichlet distributions with mean 0.5, i.e. survival probability of 50%. However, although cerebral palsy is associated with increased premature mortality, the survival rate up to or above the age of 5 is assumed to be 98% in the UK (Source: Office of National Statistics, 2012 tables). Also, it is known that individuals with severe impairments are more likely to die prematurely

than the general population and that this increases with the number of impairments (see for example Blair et al. [2001] or Hutton [2006]). It is therefore plausible to put informative priors onto the final florets in the tree which describe survival. Two possibilities could be the following:

1. Assume that the prior on the final florets takes the Dirichlet distribution $Dir(1.96, 0.02)$, with mean 98%, which may be considered as the known survival rate above the age of 5 for people with Cerebral Palsy.

2. Assume a Dirichlet distribution on the final florets of the tree, which differentiates between the number of severe or missing impairments based on expert judgment, as indicated for example in Hutton [2006]. Depending on the number of severe and missing impairments assume the following Dirichlet priors:

   - 0 impairments: $Dir(1.98, 0.02)$
   - 1 impairment: $Dir(1.9, 0.1)$
   - 2 impairments: $Dir(1.6, 0.4)$
   - 3 impairments: $Dir(1.4, 0.6)$

   These describe an expected probability of survival of $99\%, 95\%, 80\%$ and 70%.

In both cases all other priors are set such that the paths up to the final situations in the tree are a priori equally likely. However, also here prior information could be obtained. For example, it may be assumed that, given visual and ambulatory impairment are observed to be non-severe, a non-severe manual impairment is more likely than severe or missing manual impairment.

Note also that I have here assumed an equivalent sample size of 2 on the final situations, which leads to an overall equivalent sample size of 54. As already mentioned in Chapter 3.2 it is important to test the sensitivity of the prior with respect to different equivalent sample sizes. Therefore, for an accurate comparison between the MAP CEGs, given the informative priors above, and the MAP CEG with a uniform prior on the root-to-leaf paths, I first investigate the sensitivity of the MAP CEG with uniform priors according to its equivalent sample size. Increasing the equivalent sample size leads to similar conclusions as before. However, instead of six final positions, there are seven final positions, as an extra position with a predictive probability of survival of 96.43% exists which splits position w_{14}. Also, due to the larger equivalent sample size combinations of impairments with sparse cell counts tend to move to positions with lower survival probabilities. Nevertheless, the predictive probabilities of survival on the final positions are extremely close to those of Figure 4.12, differing at most by 0.37%. The model therefore appears to be

fairly robust with respect to the equivalent sample size and similar conclusions to those from the previous model are drawn from the graph. The CEG's score (given by the logarithm of the marginal likelihood) is -11850.56 and I will compare this score to the scores obtained when using the informative priors described above.

The MAP CEG, found when setting priors according to the known survival rate, also has seven final positions with predictive probabilities of survival: 99.98%, 99.77%, 98.74%, 92.98%, 87.50%, 77.42% and 44.22%. Most of the predictive probabilities of survival are extremely close to those of Figure 4.12. However, the predictive probability of the lowest position has reduced to 44.22%, as this position is now made up only of the category of individuals with two missing impairments and the ambulatory impairment observed as severe. Also position w_{13} has been split into two positions with extremely close predictive probabilities of survival, where the combination 'missing visual', 'non-severe ambulatory' and 'non-severe manual impairment' is now in a separate position with a probability of survival of 99.98%. Overall, due to the prior suggesting an overall survival of 98%, combinations with sparse cell counts tend to be moved to higher positions. Nevertheless, the general deductions made on the effect of the three impairments and missingness on survival remain similar, demonstrating a robustness of the model to prior assumptions.

The MAP CEG, found when using the second type of informative priors, similarly, has seven final positions with predictive probabilities of survival: 99.97%, 99.76%, 98.64%, 92.76%, 87.25%, 77.54%, 51.89%. Again the combination of 'missing visual', 'not-severe ambulatory' and 'not-severe manual impairment' has been placed into a separate position with a probability of survival of 99.97%. The next five final positions are again extremely close in their prediction of survival to the predictions from the original CEG. However, the seventh position now has a higher predictive probability of survival of 51.89% in comparison to the original CEG and the CEG with the first type of informative prior. This happens as only the combination 'missing visual', 'severe ambulatory' and 'missing manual impairment' was in this position in the previous CEG. Now, as for the original CEG, in addition the combinations 'severe visual', 'missing ambulatory' and 'missing manual impairment' as well as 'severe visual', 'severe ambulatory' and 'missing manual impairment' are in this position. Hence the prior information that individuals with only missing and severe impairments have a lower probability of survival has affected this position. Overall the position partition of this CEG is extremely close to the partition of the original CEG, with only a few edges with sparse cell counts having been moved into a lower position. Hence, again the general deductions made on survival are similar for this informative prior.

The scores of the CEG with informative priors are -11824.70 for the first approach and -11817.85 for the second approach, giving a Bayes Factor of 943.88 favouring the second approach. In comparison to the MAP CEG structure with uniform priors we have a Bayes Factor of $e^{25.86}$ favouring the first approach and a Bayes Factor of $e^{32.71}$ favouring the second approach. This suggests that our results are consistent with previous findings on the effect of impairments on survival, which were incorporated into the model through informative priors. In particular, the prior information of counting the number of severe and missing impairments with an equivalent sample size of 54 has shown to give nearly identical predictions to the CEG with a weak uniform prior.

Apart from putting informative priors on the CPVs, it may also be possible to include expert knowledge on the plausibility of particular situations being combined, by putting prior information on different CEG structures. For example, it may be thought that the situation describing the survival with three non-severe impairments cannot plausibly be merged with a situation describing survival with three severe or missing impairments. Especially, when looking at DCEGs, where the model space is even larger, restricting the model space according to expert knowledge may be one useful approach to making model selection algorithms feasible (see also Section 6.4).

6.3 CEGs with Sparse Cell Counts

Consider again the example of the UKCP study, where we are interested in the effect of three impairments, visual, ambulatory and manual, on survival. As we distinguish between a non-severe, severe or missing impairment, there are 27 different combinations of impairments that individuals can have. If we represent the problem as a tree, with survival as our final variable, then these 27 combinations are given by the 27 final situations in this tree. Table 6.1 gives the number of individuals associated with the 27 combinations of the three impairment variables.

There are two zero cell counts and three cells with only one or two individuals. Further, there are two cells with three and four individuals and three cells with seven, eight and nine individuals respectively, while all other cells have ≥ 16 individuals. Several possibilities to determine whether a cell is to be classed as 'sparse' could be considered:

1. Let a cell count be sparse when it is less than a specified proportion of the full sample size. As we have 6289 individuals we could say that a sparse cell will have ≤ 6 individuals, which corresponds to $< 1/1000$ of the full sample size.

Visual impairment	Ambulatory impairment	Manual impairment		
		Not severe	Severe	Missing
Not severe	Not severe	3470	69	44
	Severe	343	585	23
	Missing	19	0	35
Severe	Not severe	72	4	1
	Severe	37	413	8
	Missing	0	1	7
Missing	Not severe	325	9	16
	Severe	92	290	16
	Missing	2	3	405

Table 6.1: Number of individuals in the UK cerebral palsy cohort with non-severe, severe or missing visual, ambulatory and manual impairment

2. Let a sparse cell count be determined not only by the total sample size but also by the number of different paths the individual can take in the tree. In our example there are 27 categories and 6289 individuals, so on average each cell should have 233 individuals. A cell is then considered to be sparse when it has ≤ 12 individuals ($\leq 5\%$ of the average number of individuals in each cell).

3. In a $\mathcal{X}^2$-test of independence a cell is considered as sparse when the expected number in the cell is less than a particular value, commonly 1, 3 or 5. The overall death rate before the age of 5 in this data set is 4.6% and therefore a cell with 22 individuals is expected to include one death. A cell could then be considered to be sparse when it has less than 22 individuals.

To mark a sparse cell it is helpful to draw the edges whose associated counts are sparse as dotted or dashed edges. Using approach 1 above the edges with $0-6$ individuals going along them, could are as dotted edges. Similarly, by approach 2 the edges with $7-12$ individuals are drawn as dashed edges and we consider the corresponding cell counts as 'small'. When interpreting the CEG structures these edges, with only a sparse or small number of individuals attached to them, need to be treated with care, as the combinations are so rare that we cannot draw reliable conclusions about the survival rate for individuals with these combinations. Consider again the MAP CEG structure on three impairments from Chapter 4.4, which is repeated in Figure 6.1 with the sparse cell counts marked as described.

I also repeat, in Figure 6.2, the Reduced Ordinal CEG of the example and mark the edges with a sparse number of individuals in the same way. For a better illustration I draw the edges that describe several combinations as double edges, in-

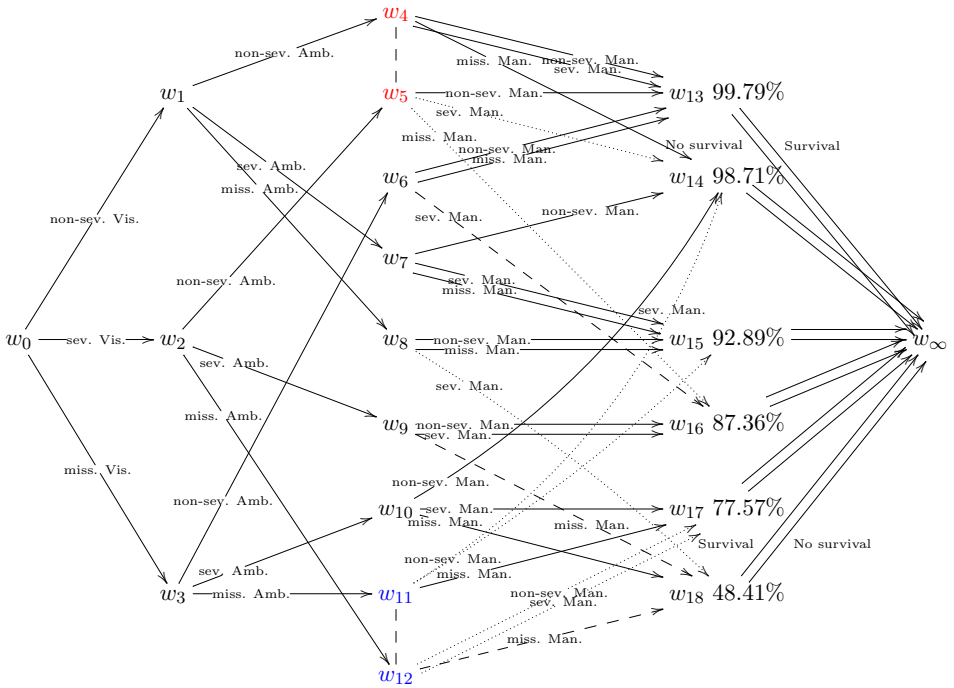


Figure 6.1: Ordinal MAP CEG structure for the UKCP example describing the effect of visual, ambulatory, manual impairment and missingness of impairments on survival, with sparse cell counts marked as dotted (counts: 0-6) or dashed (counts: 7-12 individuals) edges

stead of as single edges as before. It can now be seen directly which interpretations should be made carefully. For example, in Chapter 4.4 I suggested that the conclusions drawn from the combination of 'missing visual', 'missing ambulatory' and 'severe manual impairment' may be implausible and the graph now depicts explicitly that there are less than 6 individuals going along this edge. Also, the surprisingly low probability of survival for individuals with not severe ambulatory impairment but severe visual and missing manual impairment, or vice versa, is now suggested to be due to sparsity of cell counts.

Sparsity could be avoided by pruning the tree prior to running the AHC algorithm. For example, of the individuals with severe visual impairment and non-severe ambulatory impairment, one individual has missing manual impairment, 72 non-severe manual impairment and four severe manual impairment. The tree could then be pruned to only include the first two impairments in the analysis and, in the case of a severe visual and a non-severe ambulatory impairment, move directly to survival without considering manual impairment. Hence the third impairment variable is only included if enough data is available to draw reliable conclusions.

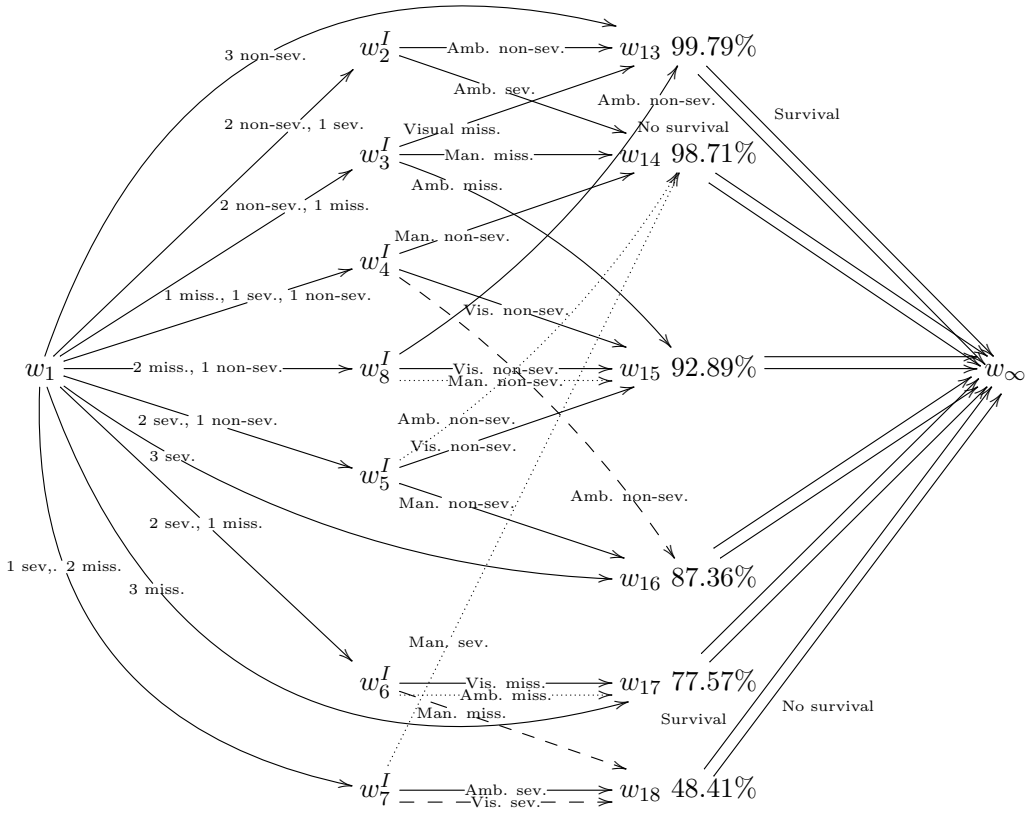


Figure 6.2: Reduced Ordinal MAP CEG structure for the UKCP example describing the effect of visual, ambulatory, manual impairment and missingness of impairments on survival with sparse cell counts marked as dotted (counts: 0-6) or dashed (counts: 7-12 individuals) edges

So, in this example, we would move directly to survival from positions w_5, w_{11} and w_{12} in Figure 6.1. In other cases, such as position w_6, we may distinguish only two categories for manual impairment, namely non-severe and severe or missing. Alternatively, considering position w_8, it may be preferred to simply omit the edge (with zero individuals) in order to keep the information about the other two impairments. Finally, in position w_9 with counts 8, 37 and 413 for missing, non-severe and severe manual impairment respectively, we may want to keep the 'small' cell count of 8 rather than omitting the edge. In every case, sparsity of cell counts should be considered when determining a plausible CEG structure to ensure that reliable conclusions can be obtained.

Finally, it is worth recognising that situations with sparse cell counts tend to be merged into other stages, even when their associated survival probabilities do not appear to be sufficiently close. This is a general problem in statistical analyses: We can distinguish small changes in probability within a large sample, such that with

enough data we can always obtain significant results (compare Figure 4.11 positions w_{11} and w_{12}). However, small cell counts give little information on survival and are likely to be combined. In this example, stronger prior information on the survival probabilities, as shown in the previous section, could prevent this behaviour of sparse cells being combined implausibly.

6.4 Future Work

I believe that the class of CEGs and its dynamic analogue are not only expressive models but also have the promise of many useful future applications. Perhaps one of the biggest challenges for the methods developed in this thesis is that, as the number of variables of the problem increases, the number of situations in the associated probability tree, and with it the model space of the CEG, quickly becomes extremely large. This may cause difficulties when applying the CEG to larger health studies (or other areas of application) than the ones discussed in this thesis. The issue of resulting sparse cell counts, which may result from a large probability tree, has been discussed above in Section 6.3. However, the complexity of the CEG needs to be addressed in future to fully exploit the methods developed in this thesis for large-scale problems. Below I will give three possible areas for further research regarding the application of the CEG to larger studies:

- I have briefly discussed at the end of Chapter 3 that, as the number of variables increases in the CEG, a search across the model space can become complex. Particularly when moving to the DCEG, the size of the model space is vast and model selection techniques that efficiently traverse the model space need to be devised. As mentioned in Chapters 3.2 and 5.4, Freeman [2011] suggests restricting the model space a priori by allowing only certain situations to be combined and hence prevent particular CEG structures. To do so, informative priors would need to be put on the various model structures either to prevent certain structures entirely or to simply give less plausible CEG structures a smaller prior probability. Also, the recent paper by Silander and Leong [2013] suggests a dynamic programming algorithm that can deal with around 30 variables. Both approaches can be investigated further.

- A further suggested way to deal with complexity has been introduced in Chapter 2.5, namely the Ordinal CEG, which allows an improved graphical representation by listing the positions in descending order according to a variable of interest. In addition to this, I have demonstrated in Chapter 4 that it is

possible to reduce the original CEG structure by defining new variables resulting from the dependence structure of the variables depicted in the topology of the CEG. This has led to the Reduced Ordinal CEG, which retains the final positions in the graph and reduces the paths leading to these. Although this has shown to produce a small informative number of final positions, the question of how to determine a Reduced Ordinal CEG from an Ordinal CEG systematically for larger problems still remains. In the UKCP example, counting the number of impairments is a straightforward solution. However, in a different context further domain knowledge may be necessary.

- The methods used in Chapter 3 could be further exploited by applying the CEGs to a subset of a large BN model. This would allow us to refine parts of the BN model for which the dependence structure between the variables is particularly complex. Especially, when there is data missing, this could be extremely useful. Complex BNs are now commonly represented as Objected-Oriented BNs (OOBNs) [Koller and Pfeffer, 1997]. These are defined by a set of classes, which are network fragments that can be used multiple times throughout the construction of the OOBN. Further hierarchical structures are obtained by allowing the attributes in a class to be network fragments in themselves and by letting subclasses with additional attributes be defined from existing classes. It is worth investigating whether the CEG and DCEG could be incorporated into this framework.

Finally, to motivate the use of CEGs within a wider domain a CEG software tool is desirable. With the development of BNs a large number of BN software tools have been developed, such as Netica, GeNIe and Hugin, which has strongly influenced the use of BNs across a wide range of domains. Similar software for CEGs would be desirable but are so far not publicly available. To find the MAP CEG structures for the examples given in the thesis I wrote a simple implementation of the AHC algorithm in R based on a previous implementation by Freeman [2011]. I have further been collaborating with the School of Information Technology at Monash University, Melbourne, Australia, on the development of a CEG software tool with a graphical user interface, which allows the drawing of a tree, elicitation of stages, as well as model selection using the AHC algorithm. So far a prototype exists and a full development of the tool is currently being investigated.

Appendix A

Latent-Class Model

Following Linzer and Lewis [2011] we have J categorical variables $X_1, ..X_J$, which can take K_j, $j = 1, .., J$ possible values. (In the CHDS example $J = 4$ for the economic situation and $J = 5$ for the social background). Further assume a random sample of N individuals. Then define Y_{ijk}, $i = 1, .., N, j = 1, .., J, k = 1, .., K_j$ as follows:

$$\begin{aligned} Y_{ijk} &= 1 \text{ when individual } i \text{ takes value } k \text{ on variable } j \\ &= 0 \text{ otherwise.} \end{aligned} \tag{A.1}$$

Let Z be the latent class variable(describing the overall economic situation or social background) which is assumed to have R classes. Then let π_{jrk} describe the probability that an individual that has latent class r takes value k on variable j. So,

$$\pi_{jrk} = P(X_j = k|Z = r). \tag{A.2}$$

Also let

$$p_r = P(Z = r). \tag{A.3}$$

Assuming local independence, i.e. given the latent class of an individual i is known, the probability of observing $X_j = k$ is conditionally independent of the probabilities of the observations on the other observed variables. So the probability that we observe a particular set of outcomes for the J variables on an individual i given class r is:

$$P(Y_i|Z = r) = \prod_{j=1}^{J} \prod_{k=1}^{K_j} \pi_{jrk}^{Y_{ijk}}, \tag{A.4}$$

where Y_i is the vector of values $(Y_{i11}, .., Y_{i1K_1}, .., Y_{iJ1}, .., Y_{iJK_J})$. Summing over all possible classes, we get the probability of observing a set of of outcomes for the J variables on an individual i:

$$P(Y_i) = \sum_{r=1}^{R} p_r \prod_{j=1}^{J} \prod_{k=1}^{K_j} \pi_{jrk}^{Y_{ijk}}. \tag{A.5}$$

Then given estimates $\widehat{p}_r$ and $\widehat{\pi}_{jrk}$, the estimated probabilities that an individual is in class r given his observed values Y_{ijk}, for $j = 1, .., J$ and $k = 1, .., K_j$, are

$$\widehat{P}(r|Y_i) = \frac{\widehat{p}_r \widehat{P}(Y_i|Z = r)}{\sum_{q=1}^{R} \widehat{p}_q \widehat{P}(Y_i|Z = q)}. \tag{A.6}$$

The command 'poLCA' in R estimates $\widehat{p}_r$ and $\widehat{\pi}_{jrk}$ using the Expectation-Maximisation algorithm: This starts with random starting values for $\widehat{p}_r$ and $\widehat{\pi}_{jrk}$ and then finds $\widehat{P}(r|Y_i)$ using equation A.6. The log-likelihood function is given by

$$log\ L(p_r, \pi_{jrk}|Y) = \sum_{i=1}^{N} log \sum_{r=1}^{R} p_r \prod_{j=1}^{J} \prod_{k=1}^{K_j} \pi_{jrk}^{Y_{ijk}}. \tag{A.7}$$

In the maximisation step the maximum likelihood estimates for p_r and π_{jrk} given $\widehat{P}(r|Y_i)$ are found, which are given by

$$\widehat{p}_r = \frac{1}{N} \sum_{i=1}^{N} \widehat{P}(r|Y_i) \quad \text{and} \quad \widehat{\pi}_{jr} = \frac{\sum_{i=1}^{N} Y_{ij} \widehat{P}(r|Y_i)}{\sum_{i=1}^{N} \widehat{P}(r|Y_i)}, \tag{A.8}$$

where $\widehat{\pi}_{jr} = (\widehat{\pi}_{jr1}, .., \widehat{\pi}_{jrK_j})$ and $Y_{ij} = (Y_{ij1}, .., Y_{ijK_j})$.

These new estimates are then put back into equation A.6 and so on. The algorithm is iterated until convergence (specified by the package to be reached once subsequent estimates differ by $< 1^{-10}$).

Appendix B

R Code for the AHC Algorithm

```
CEG.AHC<-function(exampledata=exampledata,equivsize=3){
exampledata<-exampledata
equivsize<-equivsize
numbvariables<-dim(exampledata)[2]
numbcat<-c()
for(k in 1:numbvariables){
numbcat<-c(numbcat,nlevels(exampledata[,k]))
}
numb<-c(1)
for(i in 2:numbvariables){
numb<-c(numb,prod(numbcat[1:(i-1)]))
}
prior<-c()
for(i in 1:numbvariables){
for(j in 1:numb[i]){
prior<-c(prior,list(rbind(rep(equivsize/(numbcat[i]*numb[i]),numbcat[i]))))
}
}
#Datalist1: list of the number of individuals going from the stage along a
    particular edge in C_{0}
data<-c(list(rbind(table(exampledata[,1]))))
for (i in 2:numbvariables){
for (j in 1:numb[i]){
data<-c(data,list(rbind(ftable(exampledata[,1:i])[j,])))
}
}
#List of the stages that can be merged in the first step
comparisonset<-c()
for (i in 2:numbvariables){
comparisonset<-c(comparisonset,list(c((sum(numb[1:(i-1)])+1):(sum(numb[1:i]))
    )))
}
labelling<-c()
for (k in 1:(numbvariables-1)){
label<-c(1,rep("NA",sum(numb[1:k])-1))
label<-c(label,rep(levels(exampledata[,k]),numb[k]))
```

```
if (k<(numbvariables-1)){
for (i in (k+1):(numbvariables-1)){
label<-c(label,rep(levels(exampledata[,k]),each=numb[i+1]/numb[k+1],numb[k+1]
    /numbcat[k]))
}
}
labelling<-cbind(labelling,label)
}
mergedlist<-c()
for (i in 1:sum(numb)){
mergedlist<-c(mergedlist,list(labelling[i,]))
}
merged1<-c()
lik<-0
for( i in 1: sum(numb)){
  alpha<-unlist(prior[i])
  N<-unlist(data[i])
  lik<-lik+sum(lgamma(alpha+N)-lgamma(alpha))+sum(lgamma(sum(alpha))-lgamma(
      sum(alpha+N)))
}
score<-c(lik)
#At each step we calculate the difference between the current CEG and the CEG
     in which two stages in the current comparison set have been merged.
#We go through every possible combination of stages that can be merged. k is
    an index for the comparisonset we are in,
#and i and j the position of the stages within the comparison set.
diff.end<-1 #to start the algorithm
while(diff.end>0){  #We stop when no positive difference is obtained by
    merging two stages
#while(length(unlist(comparisonset))>3){
difference<-0
for (k in 1:length(comparisonset)){
if(length(comparisonset[[k]])>1){    #can only merge if more than one stage
    in the comparisonset
for (i in 1:(length(comparisonset[[k]])-1)){
for (j in (i+1):length(comparisonset[[k]])){
#to compare
compare1<-comparisonset[[k]][i]
compare2<-comparisonset[[k]][j]
#we calculate the difference between
#the CEG where two stages are merged
result<-lgamma(sum(prior[[compare1]]+prior[[compare2]]))-lgamma(sum(prior[[
    compare1]]+data[[compare1]]+prior[[compare2]]+data[[compare2]]))+
sum(lgamma(prior[[compare1]]+data[[compare1]]+prior[[compare2]]+data[[
    compare2]]))-sum(lgamma(prior[[compare1]]+prior[[compare2]]))-
#and the CEG where the two stages are not merged
(lgamma(sum(prior[[compare1]]))-lgamma(sum(prior[[compare1]]+data[[compare1
    ]]))+sum(lgamma(prior[[compare1]]+data[[compare1]]))-
sum(lgamma(prior[[compare1]]))+lgamma(sum(prior[[compare2]]))-lgamma(sum(
    prior[[compare2]]+data[[compare2]]))+
sum(lgamma(prior[[compare2]]+data[[compare2]]))-sum(lgamma(prior[[compare2]])
    )
```

```
#if the resulting difference is greater than the current difference then we
    replace it
if (result > difference){
difference<-result
merged<-c(compare1,compare2,k)
}
}
}
}
}
diff.end<-difference
#We update our priorlist, datalist and comparisonset to obtain the priorlist,
     datalist and comparisonlist for C_{1}
if(diff.end>0){
prior[[merged[1]]]<-prior[[merged[1]]]+prior[[merged[2]]]
prior[[merged[2]]]<-cbind(NA,NA)
data[[merged[1]]]<-data[[merged[1]]]+data[[merged[2]]]
data[[merged[2]]]<-cbind(NA,NA)
comparisonset[[merged[3]]]<-comparisonset[[merged[3]]][-(which(comparisonset
    [[merged[3]]]==merged[2]))]
mergedlist[[merged[1]]]<-cbind(mergedlist[[merged[1]]],mergedlist[[merged
    [2]]])
mergedlist[[merged[2]]]<-cbind(NA,NA)
lik<-lik+diff.end
score<-c(score,lik)
merged1<-cbind(merged1,merged)
}
}
#Output: stages of the finest partition to be combined to obtain the most
    probable CEG structure
stages<-c(1)
for (i in 2:numbvariables){
stages<-c(stages,comparisonset[[i-1]])
}
result<-mergedlist[stages]
newlist<-list(prior=prior,data=data,stages=stages,result=result,score=score,
    merged=merged1,comparisonset=comparisonset,mergedlist=mergedlist,lik=lik)
return(newlist)
}
```

./AHC.R

Bibliography

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

P.E. Anderson and J.Q. Smith. Bayesian representations using Chain Event Graphs. 2006.

V.S. Barbu and N. Limnios. *Semi-Markov chains and hidden semi-Markov models toward applications: their use in reliability and DNA analysis*, volume 191 of *Lecture Notes in Statistics*. Springer, 2008.

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, 2013a.

L.M. Barclay, J.L. Hutton, and J.Q. Smith. Chain Event Graphs for Informed Missingness. *Bayesian Analysis*, 2013b. DOI:10.1214/13-BA843.

L.M. Barclay, J.Q. Smith, P.A. Thwaites, and A.E. Nicholson. Dynamic Chain Event Graphs. *CRiSM Research Report 14-04, University of Warwick*, 2013c.

M. C. Bax. Terminology and classification of cerebral palsy. *Developmental Medicine & Child Neurology*, 6(3):295–297, 1964.

A.L. Beautrais, D.M. Fergusson, and F.T. Shannon. Life events and childhood morbidity: A prospective study. *Pediatrics*, 70(6):935–940, 1982.

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. London,UK, 1989.

E. Blair, L. Watson, N. Badawi, and F.J. Stanley. Life expectancy among people with cerebral palsy in Western Australia. *Developmental Medicine & Child Neurology*, 43:508–515, 2001.

J.M. Boden, D.M. Fergusson, and L.J. Horwood. Exposure to single parenthood in childhood and later mental health , educational, economic and criminal bahaviour outcomes. *Arch Gen Psychiatry*, 64(9):1089–1095, 2007.

S.G. Bøttcher and C. Dethlefsen. deal: A package for learning Bayesian Networks. *Journal of Statistical Software*, 8(20):1–40, 2003.

C. Boutilier, N. Friedman, M. Goldszmidt, and D. Koller. Context-specific independence in Bayesian Networks. In *Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence*, pages 115–123, Portland, OR, USA, 1996. Morgan Kaufmann Publishers Inc.

W. Buntine. Theory refinement on bayesian networks. In *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence*, volume 91, pages 52–60, Los Angeles, CA, USA, 1991.

C. Cans. Surveillance of cerebral palsy in Europe: a collaboration of cerebral palsy surveys and registers. *Developmental Medicine & Child Neurology*, 42(12):816–824, 2000.

G. F. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine learning*, 9(4):309–347, 1992.

N. J. Cooper, K. R. Abrams, A. J. Sutton, D. Turner, and P. C. Lambert. A Bayesian approach to markov modelling in cost-effectiveness analyses: application to taxane use in advanced breast cancer. *Journal of the Royal Statistical Society: Series A*, 166(3):389–405, 2003.

N. J. Cooper, P. .C Lambert, K. R. Abrams, and A. J. Sutton. Predicting costs over time using Bayesian markov chain monte carlo methods: an application to early inflammatory polyarthritis. *Health economics*, 16(1):37–56, 2007.

R.G. Cowell and J.Q. Smith. Causal discovery through MAP selection of stratified Chain Event Graphs. *CRiSM Research Report 13-14, University of Warwick*, 2011.

R.G. Cowell, A.P. Dawid, S.L. Lauritzen, and D.J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer Verlag, New York, USA, 2007.

R.G. Cowell, P.A. Thwaites, and J.Q. Smith. Decision making with Decision Event Graphs. *CRiSM Research Report 10-15, University of Warwick*, 2013.

James Cussens. Bayesian Network learning by compiling to weighted MAX-SAT. In *UAI*, pages 105–112, Helsinki, 2008. AUAI Press.

M. J. Daniels and J. W Hogan. *Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis*. Chapman and Hall/CRC, Boca Raton, FL, USA, 2008.

A.P. Dawid. Influence diagrams for causal modelling and inference. *International Statistical Review*, 70(2):161–189, 2002.

T. Dean and K. Kanazawa. A model for reasoning about persistence and causation. *Computational Intelligence*, 5(2):142–150, 1989.

V. Didelez. Graphical models for marked point processes based on local independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):245–264, 2008.

D. Edwards and S. Ankinakatte. A class of context-specific graphical models for discrete longitudinal data. *arXiv preprint arXiv:1311.5066*, 2013.

D.M. Fergusson, L.J. Horwood, and F.T. Shannon. Parental smoking and respiratory illness in infancy. *Archives of Disease in Childhood*, 55(5):358, 1980.

D.M. Fergusson, L.J. Horwood, and F.T. Shannon. Birth placement and child health. *The New Zealand medical journal*, 94(688):37, 1981.

D.M. Fergusson, M.E. Dimond, L.J. Horwood, and F.T. Shannon. The utilisation of preschool health and education services. *Social Science & Medicine*, 19(11): 1173–1180, 1984.

D.M. Fergusson, L.J. Horwood, and F.T. Shannon. Social and family factors in childhood hospital admission. *Journal of Epidemiology and Community Health*, 40(1):50, 1986.

D.M. Fergusson, L.J. Horwood, and M.T. Lynskey. A longitudinal study of early childhood education and subsequent academic achievement. *Australian Psychologist*, 29(2):110–115, 1994a.

D.M. Fergusson, M.T. Lynskey, and L.J. Horwood. Alcohol consumption and associated problems in a birth cohort of 15 year olds. *The New Zealand medical journal*, 107(977):167, 1994b.

G. Freeman. *Learning and Predicting with Chain Event Graphs*. PhD thesis, University of Warwick, UK, 2011.

G. Freeman and J.Q. Smith. Bayesian MAP model selection of Chain Event Graphs. *Journal of Multivariate Analysis*, 102(7):1152–1165, 2011a.

G. Freeman and J.Q. Smith Smith. Dynamic staged trees for discrete multivariate time series: forecasting, model selection and causal analysis. *Bayesian Analysis*, 6(2):279–305, 2011b.

N. Friedman and M. Goldszmidt. Learning Bayesian Networks with local structure. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 421–460. MIT Press, 1998.

D. Geiger and D. Heckerman. Knowledge representation and inference in similarity networks and Bayesian Multinets. *Artificial Intelligence*, 82(1):45–74, 1996.

A. Gottard. On the inclusion of bivariate marked point processes in graphical models. *Metrika*, 66(3):269–287, 2007.

D. Heckerman. A tutorial on learning with Bayesian Networks. *Innovations in Bayesian Networks*, pages 33–82, 2008.

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.

K. Hemming and J. L. Hutton. Bayesian sensitivity models for missing covariates in the analysis of survival data. *Journal of Evaluation in Clinical Practice*, 18(2): 238–246, 2012.

K. Hemming, J.L. Hutton, A. Colver, and M.-J. Platt. Regional variation in survival of people with cerebral palsy in the United Kingdom. *Pediatrics*, 116(6):1383–1390, 2005.

P.W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

T.H. Holmes and R.H. Rahe. The social readjustment rating scale. *Journal of psychosomatic research*, 11(2):213–218, 1967.

R.A. Howard and J.E. Matheson. Influence Diagrams. In R. A. Howard and J. E. Matheson, editors, *Readings on the Principles and Applications of Decision Analysis*, volume 2, pages 721–762. Strategic Decisions Group, Menlo Park, CA, 1981.

J.L. Hutton. Cerebral palsy life expectancy. *Clinics in perinatology*, 33(2):545–555, 2006.

J.L. Hutton and P.O.D. Pharoah. Effects of cognitive, motor, and sensory disabilities on survival in cerebral palsy. *Archives of disease in Childhood*, 86(2):84–89, 2002.

J.L. Hutton, T. Cooke, and P.O.D. Pharoah. Life expectancy in children with cerebral palsy. *British Medical Journal*, 309(6952):431–435, 1994.

M. Jaeger. Probabilistic Decision Graphs - combining verification and AI techniques for probabilistic inference. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 12:19–42, 2004.

M. Jaeger, J.D. Nielsen, and T. Silander. Learning Probabilistic Decision Graphs. *International Journal of Approximate Reasoning*, 42(1):84–100, 2006.

F.V. Jensen and T.D. Nielsen. *Bayesian Networks and Decision Graphs*. Springer Verlag, New York, USA, 2nd edition, 2007.

R.E. Kass and A.E. Raftery. Bayes Factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

U.B. Kjaerulff and A.L. Madsen. *Bayesian networks and Influence Diagrams: a guide to construction and analysis*. Springer Verlag, New York, USA, 2007.

D. Kollar and N. Friedman. *Probabilistic graphical models: principles and techniques*. The MIT Press, 2009.

D. Koller and U. Lerner. Sampling in factored dynamic systems. In *Sequential Monte Carlo Methods in Practice*, pages 445–464. Springer, 2001.

D. Koller and A. Pfeffer. Object-oriented bayesian networks. In *Proceedings of the Thirteenth conference on Uncertainty in Artificial Intelligence*, pages 302–313. Morgan Kaufmann Publishers Inc., 1997.

K.B. Korb and A.E. Nicholson. *Bayesian Artificial Intelligence*. cRc Press, Boca Raton Florida, USA, 2004.

V.G. Kulkarni. *Modeling and analysis of stochastic systems*, volume 36. CRC Press, 1995.

S.L. Lauritzen. *Graphical models*. Oxford University Press, Oxford, UK, 1996.

S.L. Lauritzen. Causal inference from graphical models. In O.E. Barndorff-Nielsen, D.R. Cox, and C. Klppelberg, editors, *Complex stochastic systems*, pages 63–107. Chapman and Hall/CRC, London, Boca Raton, 2001.

S.L. Lauritzen and N. Wermuth. Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics*, 7(1):31–57, 1989.

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.

D.A. Linzer and J.B. Lewis. poLCA: An r package for polytomous variable latent class analysis. *Journal of Statistical Software*, 42(10):1–29, 2011.

R.J.A. Little and D.B. Rubin. *Statistical Analysis with Missing Data*. John Wiley and Sons Inc., Hoboken, NJ, USA, 2nd edition, 2002.

G. Lu and J. B Copas. Missing at random, likelihood ignorability and model completeness. *The Annals of Statistics*, 32(2):754–765, 2004.

G. Maudsley, J.L. Hutton, and P.O.D. Pharoah. Cause of death in cerebral palsy: a descriptive study. *Archives of disease in childhood*, 81(5):390–394, 1999.

J. Medhi. *Stochastic processes*. New Age International, 1994.

A.C. Miller, M. W. Merkhofer, R. A. Howard, J.E. Matheson, and T.R. Rice. Development of automated aids for decision analysis. Technical report, DTIC Document, 1976.

Kevin Patrick Murphy. *Dynamic bayesian networks: representation, inference and learning*. PhD thesis, University of California, 2002.

K.P. Murphy. *Machine learning: a probabilistic perspective*. The MIT Press, Cambridge, MA, USA, 2012.

R. E. Neapolitan. *Learning Bayesian Networks*. Pearson Prentice Hall, Upper Saddle River, NJ, USA, 2004.

U. Nodelman, C.R. Shelton, and D. Koller. Continuous time Bayesian Networks. In *Proceedings of the Eighteenth conference on Uncertainty in Artificial Intelligence*, pages 378–387, Edmonton, AB, Canada, 2002.

J. R. Norris. *Markov Chains (Cambridge Series in Statistical and Probabilistic Mathematics)*. Cambridge University Press, Cambridge, UK, 1998.

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 Inc, Chichester, UK, 2006.

J. Pearl. Fusion, propagation, and structuring in belief networks* 1. *Artificial intelligence*, 29(3):241–288, 1986.

J. Pearl. *Causality: models, reasoning, and inference*. Cambridge University Press, Cambridge, MA, USA, 2000.

D. Poole and N.L. Zhang. Exploiting contextual independence in probabilistic inference. *Journal of Artificial Intelligence Research*, 18:263–313, 2003.

H. Raiffa. *Decision analysis: introductory lectures on choices under uncertainty*. Addison-Wesley, 1968.

D. Ron, Y. Singer, and N. Tishby. On the learnability and usage of acyclic probabilistic finite automata. In *Proceedings of the eighth annual conference on Computational learning theory*, pages 31–40, Santa Cruz, Ca, USA, 1995.

D.B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

N.J. Salkind. *Encyclopedia of research design*, volume 1. Sage Publications, Inc., 2010.

J.L. Schafer. *Analysis of incomplete multivariate data*. Chapman & Hall/CRC, Boca Raton, FL, USA, 1997.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461–464, 1978.

R.D. Shachter. Evaluating Influence Diagrams. *Operations research*, 34(6):871–882, 1986.

G. Shafer. *The Art of Causal Conjecture*. The MIT Press, 1996.

T. Silander and T.-Y. Leong. A dynamic programming algorithm for learning Chain Event Graphs. In *16th International Conference on Discovery Science (DS 2013)*, pages 201–216, Singapore, 2013.

T. Silander, P. Kontkanen, and P. Myllymäki. On sensitivity of the map bayesian network structure to the equivalent sample size parameter. *CoRR*, abs/1206.5293, 2012.

J. Q. Smith and P. E. Anderson. Conditional independence and Chain Event Graphs. *Artificial Intelligence*, 172(1):42–68, 2008.

J.E. Smith, S. Holtzman, and J.E. Matheson. Structuring conditional relationships in Influence Diagrams. *Operations research*, 41(2):280–297, 1993.

J.Q. Smith. Influence diagrams for statistical modelling. *The Annals of Statistics*, 17(2):654–672, 1989.

J.Q. Smith. *Bayesian Decision Analysis: Principles and Practice*. Cambridge University Press, Cambridge, UK, 2010.

D.J. Spiegelhalter and S.L. Lauritzen. Sequential updating of conditional probabilities on directed graphical structures. *Networks*, 20(5):579–605, 1990.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. The MIT Press, Cambridge, MA, USA, 1993.

J.A.C. Sterne, I.R. White, J.B. Carlin, M. Spratt, P. Royston, M.G. Kenward, A.M. Wood, and J.R. Carpenter. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *British Medical Journal*, 338: 157–160, 2009.

D.J. Strauss, R.M. Shavelle, and T.W. Anderson. Life expectancy of children with cerebral palsy. *Pediatric neurology*, 18(2):143–149, 1998.

M. Studenỳ. *On Probabilistic Conditional Independence Structures*. Springer, London, UK, 2005.

Y. Suhov. *Markov Chains: A Primer in Random Processes and Their Applications*, volume 2. Cambridge University Press, Cambridge, UK, 2008.

G. Surman, S. Bonellie, J. Chalmers, A. Colver, H. Dolk, K. Hemming, A. King, J.J. Kurinczuk, J. Parkes, and M.-J. Platt. UKCP: a collaborative network of cerebral palsy registers in the United Kingdom. *Journal of Public Health*, 28(2): 148–156, 2006.

P. A. Thwaites, J. Q. Smith, and E. Riccomagno. Causal analysis with Chain Event Graphs. *Artificial Intelligence*, 174(12):889–909, 2010.

P.A. Thwaites. Causal identifiability via Chain Event Graphs. *Artificial Intelligence*, 195:291–315, 2013.

P.A. Thwaites and J.Q. Smith. Separation theorems for Chain Event Graphs. *CRiSM Research Report 11-09, University of Warwick*, 2011.

P.A. Thwaites, J.Q. Smith, and R.G. Cowell. Propagation using Chain Event Graphs. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, pages 546–553, Helsinki, 2008. AUAI Press.

T. S. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, pages 255–270, Cambridge, MA, USA, 1990. Elsevier Science Inc.

C. Winship, R.D. Mare, and J.R. Warren. Latent class models for contingency tables with missing data. In J.A. Hagenaars and A.L. McCutcheon, editors, *Applied Latent Class Analysis*, pages 408–432. Cambridge University Press, 2002.