The Dynamic Chain Event Graph

L. M. Barclay[a,*], J. Q. Smith[a], P. A. Thwaites[b], A. E. Nicholson[c]

[a] *Department of Statistics, University of Warwick, Coventry CV4 7AL, United Kingdom*

[b] *School of Mathematics, University of Leeds, LS2 9JT, United Kingdom*

[c] *Faculty of Information Technology, Monash University, Victoria, Australia*

Abstract

In this paper we develop a formal dynamic version of Chain Event Graphs (CEGs), a particularly expressive family of discrete graphical models. We demonstrate how this class is closely linked to semi-Markov models and provides a convenient generalisation of the Dynamic Bayesian Network (DBN). In particular we develop a two time-slice Dynamic CEG providing a useful analogue to the two time-slice DBN. We demonstrate how the Dynamic CEG's graphical formulation exhibits the essential qualitative features of the hypotheses it embodies and also how each model can be estimated in closed form enabling fast model search over the class. The expressive power of this model class together with its estimation is illustrated throughout by a variety of examples.

Keywords: Chain Event Graphs, Markov processes, Probabilistic Graphical Models, Dynamic Bayesian Networks

1. Introduction

A Chain Event Graph (CEG) [1, 2, 3, 4] is a coloured discrete graphical model constructed from an event tree together with certain elicited symmetries called stages, which describes through its topology many salient features of a process. However its semantics have so far only been developed to describe processes whose underlying event tree is finite. In this paper we extend the model space to infinite trees from which we can derive the dynamic analogue of the CEG, the Dynamic CEG (DCEG), which extends the CEG to describe potentially infinite discrete longitudinal processes. Note that an event tree provides a natural framework through which time sequences can be incorporated, as each path in the tree describes the various possible sequences of events an individual can experience. It is often convenient to assume that the potential events encountered by a unit could be infinite. The corresponding event tree is then an infinite graph and, in particular, the events an individual experiences may be repeated at later time points in the tree. Analogously to the CEG we show

*Corresponding author

Email addresses: L.M.Barclay@warwick.ac.uk (L. M. Barclay), J.Q.Smith@warwick.ac.uk (J. Q. Smith), P.A.Thwaites@leeds.ac.uk (P. A. Thwaites), ann.nicholson@monash.edu (A. E. Nicholson)

in this paper that we can write any infinite tree as a DCEG and hence represent the originally elicited tree in a much more compact and easily interpretable form. Like its CEG and Bayesian Network (BN) analogues the DCEG not only provides an evocative representation of a process but also supports conjugate learning and fast closed form model selection over large numbers of collections of hypotheses encoded within the graph. We further demonstrate that we can extend this framework by attaching holding time distributions to the nodes in the graph, so that under certain assumptions the DCEG corresponds directly to a semi-Markov process.

We revisit the CEG through a simple example, adapted from [5], which we later develop further as an illustration of the DCEG.

Example 1. *A trekker who is in a fit state to trek decides to go trekking. While trekking, he either meets or does not meet a life threatening danger. If he encounters such a danger, he either manages to avoid it or suffers such injury that he is unable or unwilling to trek again. After a safe trek that evening he decides whether to keep trekking in the future or give up trekking. An event tree depicting this story is given in Figure 1.*

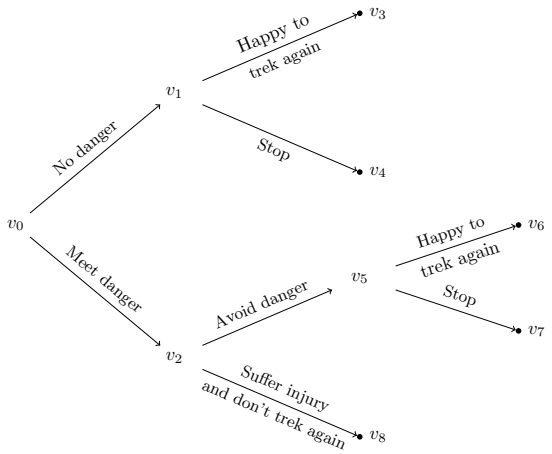

Figure 1: Trekking example

After eliciting this event tree we can hypothesise possible probabilistic symmetries that might plausibly exist in the corresponding probability tree. For example, we might hypothesise that avoiding danger will not affect the trekker's probability to decide to trek again, demanding that the probabilities on the edges emanating from v_1, labelled "happy to trek again" and "stop". are believed to be identical to the probabilities on the edges emanating from v_5 with the same labels. A CEG depicts not only the unfolding of events expressed in a tree like this, but also these types of probabilistic symmetries: Briefly, we say that two non-leaf vertices are in the same stage whenever their associated outgoing edges are described by the same set of labels and two edges with the same label are believed to have the same probability associated with them. Similarly, when the entire probability distribution over the atoms of future events unfolding from two different vertices are the same, we say that these two vertices are in the same position, w. When two vertices are in the same stage but not in the same position this is indicated by additional colouring of the vertices and their emanating edges so that both the identified vertices and their identified outgoing edges are matched in colour. The CEG is then obtained by

collapsing the tree over its positions, such that the vertices of the CEG are given by the set of positions. The leaf nodes of the tree are collected into a single position called w_∞. A full description of this construction can be found in [1, 2, 6]

Example 1 (continued). *The hypothesis that avoiding danger does not affect the probability that the trekker decides to trek again, places v_1 and v_5 in the same position. We then obtain the following CEG given in Figure 2.*

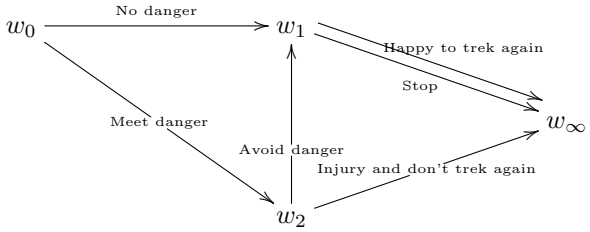

Figure 2: CEG of trekking example

Although the topology of a CEG is usually more complicated than the corresponding discrete BN, it is often much more expressive. Not only the conditional independencies, but also the context-specific symmetries, are directly depicted in the *topology and colouring* of the graph of the CEG. Further, structural zero probabilities in the conditional probability tables are directly depicted by the absence of edges in its graph. It has further been shown that the CEG retains most of the useful properties of a BN like closure to learning under complete sampling [2], causal expressiveness [3, 6, 7, 8] and efficient propagation [9, 10] and hence CEGs provide a natural and expressive framework for various tasks associated with representing, propagating and learning, especially when the tree of the underlying sample space is asymmetric [11].

Another important class of graphical models related to CEGs is the Probabilistic Decision Graph (PDG) [12, 13]. This has been widely studied as an alternative inferential framework to the BN and efficient model selection algorithms [14, 15] have been developed for it. However, unlike the PDG the class of CEGs contains all discrete BNs, as well as the extension of the BN to context-specific BNs [16, 17] and Bayesian multinets [18, 19], in the sense that *all* the conditional independences entailed by these model classes are embodied in the topology of the graph of a single CEG (see for example [1, 20] as well as many others). The topology of the CEG has hence been exploited to fully represent and generalise models such as context-specific BNs. However, the CEG cannot be used to generalise discrete dynamic processes like the Dynamic BN, whose state space is infinite. We develop the DCEG to do this and demonstrate that it is a powerful and efficient representational framework for modelling discrete dynamic processes.

In Section 2 we formally define the infinite staged tree and the DCEG. We further introduce the Extended DCEG which attaches conditional holding times to each edge within the graph. We show that in certain simple scenarios there is a direct correspondence between the uncoloured version of these graphs and the state transition diagram of a semi-Markov process. In Section 3 we demonstrate that any general DBN lies in the class of DCEGs. We here also introduce the 2-time-slice DCEG, a special class of DCEGs, which, like the 2-time-slice DBN and its more general class of DBNs, imposes certain restrictions on the more general class of DCEGs. In Section 4 we show how to perform fast conjugate Bayesian estimation of these model classes,

extending the work in [2, 21] to this dynamic setting and demonstrate how a typical model can be scored. We conclude the paper with a short discussion.

2. Infinite Probability Trees and DCEGs

In this section we introduce the standard terminology used for CEGs and extend it to infinite trees and hence DCEGs. In the first subsection we derive the infinite staged tree, followed by a formal definition of the DCEG and a comparison between the DCEG and Markov processes. In the final subsection we extend the DCEG to not only describe the transitions between the vertices of the graph but also the time spent at each vertex.

2.1. Infinite Staged Trees

Definition 1. *Graph Let a* graph $\mathcal{G}$ *have vertex set* $V(\mathcal{G})$ *and a (directed) edge set* $E(\mathcal{G})$*, where for each* $e(v_i, v_j) \in E(\mathcal{G})$*, there exists a directed edge from* $v_i \to v_j$*,* $v_i, v_j \in V(\mathcal{G})$*. Call the vertex* v_i *a* parent *of* v_j *if* $e(v_i, v_j) \in E(\mathcal{G})$ *and let* $pa(v_j)$ *be the set of all parents of a vertex* v_j*. Also, call* v_k *a* child *of* v_i *if* $e(v_i, v_k) \in E(\mathcal{G})$ *and let* $ch(v_i)$ *be the set of all children of a vertex* v_i*. We say the graph is* infinite *when either the set* $V(\mathcal{G})$ *or the set* $E(\mathcal{G})$ *is infinite.*

Definition 2. *Tree*

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. A* leaf vertex *in* $V(\mathcal{T})$ *is a vertex with no children. Call a non-leaf vertex of a tree* $\mathcal{T}$ *a* situation *and denote the set of situations by* $S(\mathcal{T}) \subseteq V(\mathcal{T})$*.*

In this paper we only consider *event trees*, where all vertices are chance nodes and the edges of the tree label the possible events that happen. The path from the root vertex of the tree to a situation $s_i \in S(\mathcal{T})$ therefore represents a sequence of possible transitions or unfolding events that can occur, with the situation representing the state that is reached via those transitions. We further restrict ourselves to infinite trees where the number of situations $s_i \in S(\mathcal{T})$ is infinite but each situation $s_i \in S(\mathcal{T})$ has a finite number of edges, m_i, emanating from it.

Definition 3. *Floret*

A floret *is a subtree* $\mathcal{F}(s_i) = (V(\mathcal{F}(s_i)), E(\mathcal{F}(s_i)))$ *of* $\mathcal{T}$*,* $s_i \in S(\mathcal{T})$ *where:*

- *its vertex set* $V(\mathcal{F}(s_i))$ *consists of* $\{s_i\} \cup ch(s_i)$*, and*
- *its edge set* $E(\mathcal{F}(s_i))$ *consists of all the edges between* s_i *and its children in* $\mathcal{T}$*.*

Clearly an infinite event tree can be uniquely characterised by its florets, which retain the indexing of the vertices of $\mathcal{T}$. The edges of each floret can be labelled as $e_{s_i j} \in E(\mathcal{F}(s_i)), j = 1, \ldots, m_i$, where s_i has m_i children. As noted above, we can think of these edge labels as descriptions of the particular events or transitions that can occur after a unit reaches the root of the floret. In particular, we can also use the index $j = 1, \ldots, m_i$ to define a random variable taking values $\{x_1, \ldots, ...x_{m_i}\}$ associated with this floret.

Example 1 (continued). *Assume that the trekker decides every day whether to trek*

on that day or not. At the end of every day when he has not encountered danger or has avoided danger he can decide to possibly trek again the next day or to never trek again. An informal depiction of this event tree is given in Figure 3, where implicit further continuations of the tree are denoted by ....

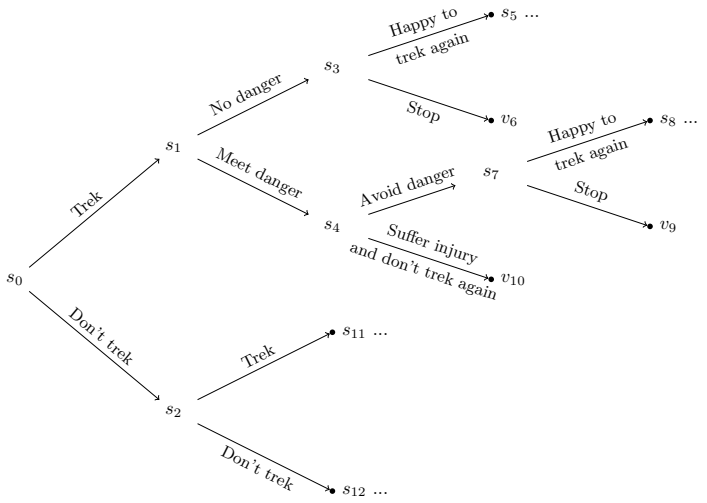

Figure 3: Trekking example: the beginning of the infinite tree, $\mathcal{T}$

In our example the edges $E(F(s_0))$ describe whether the trekker decides to trek $(e_{s_0 1})$ or not $(e_{s_0 2})$, while the floret of vertex s_1 describes, having decided to trek, whether the trekker meets danger $(e_{s_1 2})$ or not $(e_{s_1 1})$.

From the above example we can observe that each path within an infinite event tree is a sequence through time. To embellish this event tree into a probability tree, which specifies the development of a unit in the described population, we need to elicit the conditional probability vectors (CPVs) associated with each floret $F(s_i)$, given by

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

where $\pi_{s_i j} = P(e_{s_i j}|s_i)$ is the probability that the unit or process being modelled transitions from s_i along the j_{th} edge $e_{s_i j}$ and $\sum_{j=1}^{m_i} \pi_{s_i j} = 1$.

Collections of conditional independence (or Markovian) assumptions are intrinsic to most graphical models. It was shown in [1] that for an event tree these ideas can be captured by colouring the vertices and edges of the tree. This idea immediately extends to this class of infinite trees.

Definition 4. *Stage*

We say two situations s_i and s_k are in the same stage*, u, if and only if*

1. *there exists an isomorphism Φ_{ik} between the labels of $E(\mathcal{F}(s_i))$ and $E(\mathcal{F}(s_k))$, where $\Phi_{ik}(e_{s_i j}) = e_{s_k j}$, and*
2. $\pi_{s_i} = \pi_{s_k}$.

If s_i and s_k are in the same stage then we assign s_i, s_k and the pairs of edges $(e_{s_i j}, e_{s_k j})$ - as indexed above - the same colour. (See for example Figure 4.)

We can hence partition the situations of the tree $S(\mathcal{T})$ into stages, associated with a set of isomorphisms $\{\Phi_{ik} : s_i, s_k \in S(\mathcal{T})\}$.

Definition 5. *Staged Tree*

A staged tree *version of $\mathcal{T}$ is one where*

1. *all non-trivial situations and their edges are assigned a colour*
2. *situations in the same stage in $\mathcal{T}$ and their emanating edges are assigned the same colour, and*
3. *situations in different stages in $\mathcal{T}$ and their emanating edges are assigned different colours.*

Call U the stage partition of $\mathcal{T}$ and define the conditional probability vector (CPV) on stage u to be

$$\pi_u = (\pi_{u1}, \pi_{u2}, .., \pi_{um_u}), \tag{2}$$

where u has m_u emanating edges. When there is only a single situation in a stage, then we call this stage *trivial*. If U is the trivial partition, such that every situation is in a different stage, then the colouring contains no additional information about the process that is not contained in $\mathcal{T}$. However, [1] exhibits numerous examples of finite 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, where two situations are in the same stage if and only if they are associated with the same random variable and the values taken by the parents of this variable in the BN are the same (see Section 3).

A finer partition of the vertices in the infinite tree is given by the position partition. Let $\mathcal{T}(s_i)$ denote the full coloured subtree with root vertex s_i.

Definition 6. *Position*

Two situations s_i, s_k in the same stage, that is, $s_i, s_k \in u \in U$, are also in the same position *w if there is an isomorphic map Ψ_{ik} between the two coloured (potentially infinite) subtrees $\mathcal{T}(s_i) \to \mathcal{T}(s_k)$. We denote the set of positions by W.*

As above we can further define a CPV on each position:

$$\pi_w = (\pi_{w1}, \pi_{w2}, .., \pi_{wm_w}). \tag{3}$$

The definition hence requires that for two situations to be in the same position there must not only be a map between the edge sets $E(\mathcal{T}(s_i)) \to E(\mathcal{T}(s_k))$ of the two coloured subtrees but also the colours of any edges and vertices under this map must correspond. For example when s_i, s_k are a distance of one edge from a leaf node then $\mathcal{T}(s_i) = \mathcal{F}(s_i)$ and $\mathcal{T}(s_k) = \mathcal{F}(s_k)$ and so 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 for example all their children must have a parallel child in the same stage, and so on.

Surprisingly, the positions of an infinite tree $\mathcal{T}$ are sometimes associated with a coarser partition of its situations than a finite subtree of $\mathcal{T}$ with the same root. This is because in an infinite tree two situations lying on the same directed path from the root can be in the same position. This is impossible for two situations s_i, s_k in a finite tree: the tree rooted at a vertex further up a path must necessarily have fewer

vertices than the one closer to the root, so in particular no isomorphism between $\mathcal{T}(s_i)$ and $\mathcal{T}(s_k)$ can exist. We give examples below which exploit this phenomenon.

Example 1 (continued). *In the trekking example we may have the colouring of the tree as given in Figure 4. Hence we assume that the probability of deciding to trek again or not at the end of every day is the same when danger was not met as when danger was avoided.*

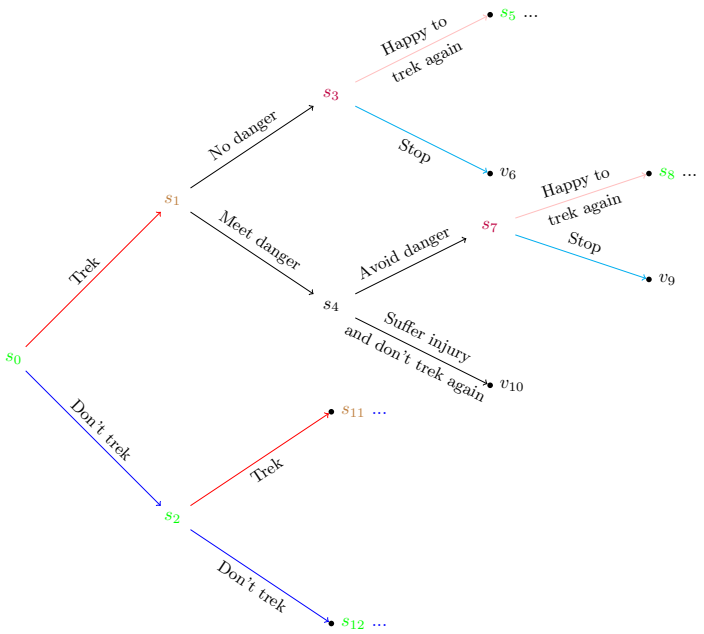

Figure 4: Trekking example: the beginning of the infinite staged tree, $\mathcal{T}$

Note that in this example the stage partition and the position partition of the situations coincides. Hence our stage and position partition is as follows:

$$w_0 = u_0 = \{s_0, s_2, s_5, s_8, s_{12} \ldots\}, w_1 = u_1 = \{s_1, s_{11}, \ldots\},$$
$$w_2 = u_2 = \{s_3, s_7, \ldots\}, w_3 = u_3 = \{s_4, \ldots\}. \tag{4}$$

From the definition of a position, w, given a unit lies in w, any information about how that unit arrived at w is irrelevant for predictions about its future development. As for the CEG, the positions therefore become the vertices of the new graph, the DCEG, which we use as a framework to support inference. Further, the colours code further symmetries described by the stages of the graph.

Note that, as for the BN and the DBN, we would normally plan to elicit the *structural equivalences* of the model - here the topology of the tree and stage structure associated with its colouring - *before* we elicit the associated conditional probability tables, which fully embellish the event tree into a probability tree. This would then allow the early interrogation and adaptation of the qualitative features of an elicited model before enhancing it with supporting probabilities. As for the BN and DBN these structural relationships can be evocatively and formally represented through the graph of the CEG and DCEG. In particular this graph can be used to explore and

critique the logical consequences of the elicited qualitative structure of the underlying process before the often time consuming task of quantifying the structure with specific probability tables.

2.2. Dynamic Chain Event Graphs

We can now define the DCEG, which depicts an infinite staged tree in a way analogous to the way the CEG represents structural equivalences in a finite tree.

Definition 7. *Dynamic Chain Event Graph*

A Dynamic Chain Event Graph *(DCEG)* $\mathcal{D} = (V(\mathcal{D}), E(\mathcal{D}))$ *of a staged tree* $\mathcal{T}$ *is a directed coloured graph with vertex set* $V(\mathcal{D}) = W$*, the set of positions of the staged tree* $\mathcal{T}$*, together with a single sink vertex,* w_∞*, comprising the leaf nodes of* T*, if these exist. The edge set* $E(\mathcal{D})$ *is given as follows: Let* $v \in w$ *be a single representative vertex of the position* w*. Then there is an edge from* w *to a position* $w' \in W$ *for each child* $v' \in ch(v), v' \in w'$ *in the tree* $\mathcal{T}$*. When two positions are also in the same stage then these and their edges are coloured in the same colour as the corresponding vertices and edges in the tree* $\mathcal{T}$*.*

We call the DCEG *simple* if the staged tree $\mathcal{T}$ is such that the set of positions equals the number of stages, $W = U$, and it is then uncoloured.

We note that when a tree is finite, a CEG is a DCEG. However the CEG is always acyclic, whilst a DCEG exhibits cycles when it has an infinite number of atoms but a finite graph. We illustrate below that in many applications, such as the one described here, the number of positions of a staged tree is finite even though the tree's vertex set is infinite. When this is the case the DCEG is a finite graph and therefore provides a succinct picture of the structural relationships in the process.

Example 1 (continued). *Figure 5 shows the corresponding DCEG of the staged tree given in Figure 4 with* $V(\mathcal{D})$ *given in equation 4. As the trekker decides every day whether to trek on that day or not, he could remain in position* w_0 *every day or move to* w_1*. At the end of every day he did not encounter danger or avoided danger he can decide to possibly trek again the next day, i.e. he returns to* w_0 *or he can decide to give up trekking for ever and hence goes to* w_∞ *where he remains. Observe that this DCEG in Figure 5 is simple because its stages and positions coincide and it is therefore uncoloured.*

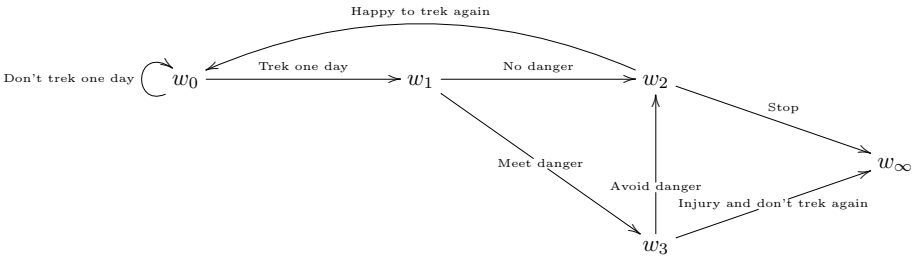

Figure 5: DCEG of $\mathcal{T}$

Note that when the graph of the DCEG has a finite number of positions its topology resembles the familiar state-transition diagram of a Markov process, where the

positions of the DCEG form the states of the Markov process. This correspondence is useful as many of the well-developed results on Markov processes can be simply extended to the DCEG. However, as mentioned at the end of section 2.1 the DCEG is usually constructed from a *description* of a process as a *staged tree* rather than from a prespecified Markov chain. There are also some differences in the graph of the DCEG and standard state-transition diagrams such as the one-to-one relationship between the atoms of the space of the DCEG and its paths and its colouring as will be illustrated in the examples below.

Example 2. *Example of a Markov Chain I*

Let $\{X_n : n \in \mathbb{N}\}$ *be a discrete-time Markov process on the state space* $\{a, b, c\}$ *with transition matrix* P *given by*

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

and with initial distribution $\alpha = (0.4, 0.4, 0.2)$*. Note that the transition probabilities from states* b *and* c *are the same. The state-transition diagram of the associated Markov process is given in Figure 6.*

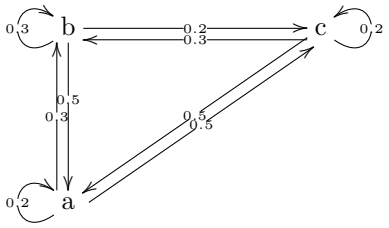

Figure 6: Example 2: State-transition diagram of a Markov process

However, the DCEG representation gives a different structure: The associated tree of the problem $\mathcal{T}$ *would have an infinite number of situations. These can be indexed as an initial situation* v_0*, the root of the tree, whose emanating edges represent the choice of initial state and by* $\{v_{i,n}, i = a, b, c, n = 0, 1, 2, \ldots\}$ *describing the states* $X_n = a, X_n = b, X_n = c$ *for* $n \in \mathbb{N}$*. However, by the Markov assumption and the transition probability matrix* P *the process only has* 3 *stages and the stage partition* U *is given by*

$$u_0 = \{v_0\}, u_a = \{v_{a,n}, n = 0, 1, 2, \ldots\}, u_{bc} = \{v_{b,n}, v_{c,n}, n = 0, 1, 2, \ldots\}.$$

Since $V(\mathcal{D})$ *is finite its DCEG can be drawn (Figure 7) even though the underlying tree is infinite. Note that in this example the DCEG is simple and hence the stages and positions coincide.*

Even here where the process is initially defined through a transition matrix, the graph of the DCEG automatically identifies states in the original description which have equivalent roles, here state b being identified with state c ,and illustrates the identical conditional probabilities associated with the two states, as $v_{b,n}$ and $v_{c,n}$, for $n = 1, 2, 3$ are in the same position. Further, the DCEG graph depicts explicitly the initial distribution of the process given by the edges emanating from w_0 and acknowledges the initial elicited distinctions of the states b and c because it can

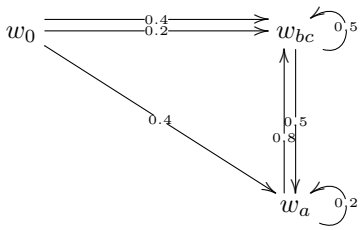

Figure 7: Example 2: DCEG representation of a Markov process

contain duplicate edges (here from w_0 to w_{bc}). These properties often have important interpretive value, as the DCEG can discover a different partition of the states of a variable or even help to construct new informative variables to represent a problem.

The DCEG may also provide a more expressive representation of a problem than its implicit state-transition diagram not only because of identifications like those illustrated above but also because it is often coloured. We illustrate this through a further example below.

Example 3. *Example of a Markov Chain II*

A coin is tossed independently, with probability $P(H) = \lambda$ of throwing heads and probability $P(T) = 1 - \lambda = \bar{\lambda}$ of throwing tails. The coin is tossed until three heads have appeared when the game terminates. Its DCEG has four positions describing whether 0, 1, 2 or 3 head have been tossed and is given in Figure 8.

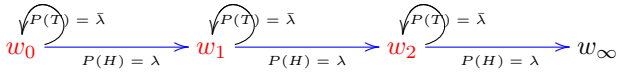

Figure 8: Example 3: DCEG representation of coin tossing example

Notice here that because each toss has the same probability λ of heads $\{w_0, w_1, w_2\}$ are all in the same stage and so its vertices w_0 w_1 and w_2 are coloured red and their corresponding edges are coloured blue and black.

This residual colouring, inherited from the staged tree allows us to further elaborate the structure of the transitions between these elicited states in a natural and consistent way. Further, it is the stage structure of the tree that supports the model selection algorithms of the CEG [2], hence allowing for more efficient learning procedures than obtainable by the position partition. We will demonstrate the way in which the stage partition is used for scoring different CEG structures in Section 4.

We will further demonstrate in Section 3 that many other important families of graphical models such as the DBN can be represented as a DCEG. So this framework contains an exciting potential for drawing together otherwise disparate classes of models under one umbrella.

2.3. DCEGs with Holding Time Distributions

In the previous subsection we have shown how an infinite tree can be elicited as a staged tree and consequently converted into a DCEG. The resulting representation produces an embellishment of the state-transition diagram of a discrete-time Markov process, as the DCEG provides additional information by colouring of the positions

which are also in the same stage and by allowing initially different states to be in the same position.

Given the graph of the DCEG we can trace the possible paths an individual may take and the associated events that may occur across time. So far we have implicitly assumed that we have regular steps such as a day, a week or a coin toss. In this case the time an individual stays in a particular position simply follows a geometric distribution, where the probability that an individual stays in position w for k time steps is equal to $P[e(w,w)|w]^k \times [1 - P\{e(w,w)|w\}]$. Hence in the trekking example (Example 1) we have that the holding time distribution on w_0 is geometric and all other holding time distributions are degenerate, as we assume that the consecutive events all occur within this time step (a single day). However, in many cases our process is unlikely to be governed by regular time steps and it is much more natural to think of the time steps to be event driven. A process like this is naturally 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. Motivated by this irregularity of events, we look at processes in which an individual stays a particular time at one vertex of the infinite tree and then moves along an edge to another vertex. We hence define in this section a generalisation of the DCEG, called the Extended DCEG, which attaches a conditional holding time distribution to each edge in the DCEG.

We call the time an individual stays in a situation s_i the holding time H_{s_i} associated with this situation. We can further also define the conditional holding times associated with each edge $e_{s_ij}, j = 1, \ldots, m_i$ in the tree, denoted by H_{s_ij}. This describes the time an individual stays at a situation s_i given that he moves along the edge e_{s_ij} next. Analogously to this we can further define holding times on the positions in the associated DCEG: We let H_w be the random variable describing the holding time on position $w \in W$ in the DCEG and $H_{wj}, j = 1, .., m_w$ the random variable describing the conditional time on w given the unit moves along the edge e_{wj} next.

In this paper we assume that all DCEGs are *time-homogeneous*. This means that the conditional holding time distributions for two situations are the same whenever they are in the same stage u. Hence, given the identity of the stage reached, the holding times are independent of the path taken. We denote the random variable of the conditional holding time associated with each stage by $H_{uj}, j = 1, \ldots, m_u$. Time-homogeneity then implies that when two positions are in the same stage u then their conditional holding time distributions are also the same. We note that an individual may spend a certain amount of time in position $w \in u$ before moving along the j_{th} edge to a position $w^{'}$ which is in the same stage. So an individual may make a transition into a different position but arrive at the same stage. A discussion of contexts when the holding time distribution may depend on the arrival time will be discussed to a later paper.

Definition 8. *Extended DCEG*

An Extended DCEG $\mathcal{D} = (V(\mathcal{D}), E(\mathcal{D}))$ *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_{uj}, to be passed through:*

$$F_{uj}(h) = P(H_{uj} \leq h), h \geq 0, \forall u \in U, j = 1, \ldots m_u. \tag{5}$$

Hence $F_{uj}(h)$ describes the time an individual stays in a position $w \in u$ before moving along the next edge e_{wj}.

Assume we have an individual passing through the DCEG and we record the positions he reaches. This unit is then a realisation of a stochastic process, $\{W_n, n \in \mathbb{N}\}$ with state space W and initial distribution $P(W_0 = w_0) = 1$. Further, let H_n be the holding time at the n_{th} position reached (after w_0) and E_n the n_{th} edge passed along.

Definition 9. *Semi-Markov DCEG*

An Extended DCEG is semi-Markov if

$$P(E_n, H_n \leq h | W_0, W_1, \ldots, W_n, E_0, E_1, \ldots, E_{n-1}, H_0, H_1, \ldots, H_{n-1}) = P(E_n, H_n \leq h | W_n). \quad (6)$$

Hence, the joint probability of the n_{th} holding time and the n_{th} edge passed along depends only on the current position of the individual.

As we are assuming a time-homogeneous DCEG we further have that

$$P(H_n \leq h | W_n = w, E_n = e_{wj}) = F_{uj}(h), w \in u, \quad (7)$$

and also

$$P(E_n = e_{wj} | W_n = w) = \pi_{uj}, w \in u, \quad (8)$$

Therefore,

$$P(E_n = e_{wj}, H_n \leq h | W_n = w) = P(E_n = e_{wj} | W_n = w) P(H_n \leq h | W_n = w, E_n = e_{wj}) = \pi_{uj} F_{uj}(h). \quad (9)$$

Finally, we can deduce that the joint probability density function of e_{wj} and h is given by

$$p(e_{wj}, h | w_i) = \pi_{uj} f_{uj}(h), \quad (10)$$

where $f_{uj}(.)$ is the probability density function of the holding times at stage u going along edge $e_{wj}, w \in u$.

A time-homogeneous and semi-Markov DCEG with stage partition U is hence fully specified by its set of conditional holding time distributions $\{F_{uj}(.) : u \in U\}$ and its collection of CPVs $\{\pi_u : u \in U\}$. Note that it is simple to extend the elicitation process on the probability tree mentioned at the end of section 2.1 to include holding times on each path of the subsequently constructed DCEG on which inference is performed.

Example 1 (continued). *Return again to the trekking example from Section 2.1 with a slightly different infinite tree given in Figure 9.*

We are no longer assuming that the trekker decides to trek or not at the start of every day, but that he spends some amount of time not trekking until he decides to trek again. This can be described by a holding time at position w_0. Further, the time until danger is met, danger is avoided or an injury is suffered, and the time to decide whether to continue or stop trekking may be of interest and we can define holding time distributions on these. The corresponding DCEG has the graph given in Figure

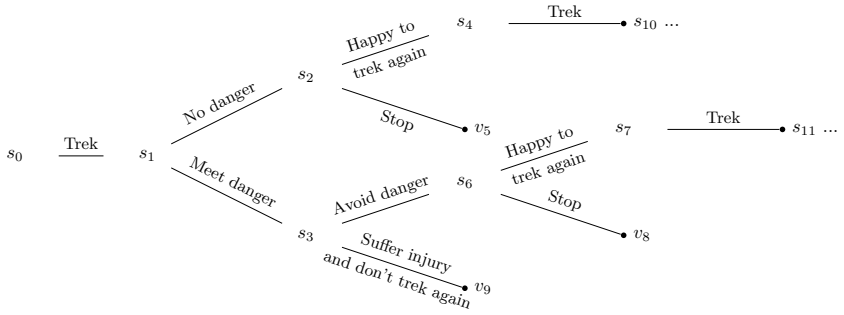

Figure 9: Variant of trekking example: infinite tree $\mathcal{T}^*$

10. The stages and positions are as follows:

$$w_0 = u_0 = \{s_0, s_4, s_7, \ldots\}, w_1 = u_1 = \{s_1, s_{10}, s_{11}, \ldots\},$$
$$w_2 = u_2 = \{s_2, s_6\}, w_3 = u_3 = \{s_3, \ldots\}, w_\infty = \{v_5, v_8, v_9, \ldots\}. \tag{11}$$

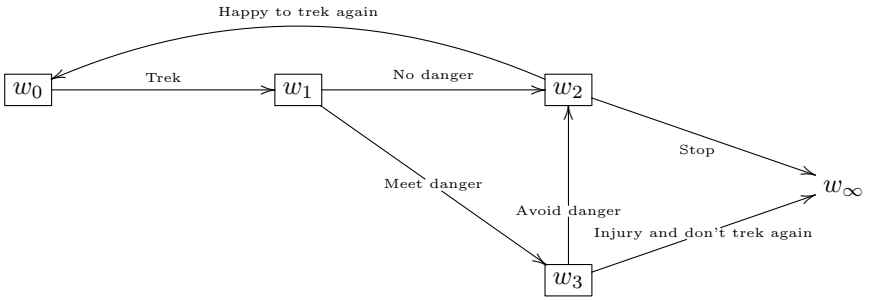

Figure 10: DCEG of $\mathcal{T}^*$

We frame the positions to illustrate that this is an Extended DCEG and that the conditional holding times are of interest. Note that this graph no longer contains loops from a position into itself as the probability of staying in the same position for a particular amount of time is instead described by the (conditional) holding times. We postpone the discussion of the holding time distribution for this example to Section 4.

Similar to the correspondence between the DCEG and Markov processes, the Extended DCEG is closely linked to semi-Markov processes [22, 23]. These are a generalisation of Markov processes that allow for the holding times to have any distribution instead of restricting them to have a geometric distribution (discrete-time Markov processes) or an exponential distribution (continuous-time Markov processes). We recall the definition of a semi-Markov process below:

Definition 10. *Semi-Markov Process [23]*

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

$$P(X_{n+1} = j, H_{n+1} \leq t | X_0, X_1, .., X_n, H_1, .., H_n) = P(X_{n+1} = j, H_{n+1} \leq t | X_n), \tag{12}$$

then $\{X_n, H_n\}$ *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 embedded Markov chain with transition probability matrix* $P = (p_{ij})$*, where* $p_{ij} = P(X_{n+1} = j | X_n = i)$*.*

Within Markov theory a semi-Markov process is usually specified by an initial distribution α and by its semi-Markov kernel Q whose ij_{th} entry is given by

$$Q_{ij}(t) = P(X_{n+1} = j, H_{n+1} \leq t | X_n = i). \tag{13}$$

We assume here that all Markov processes considered are time-homogeneous and hence the above equations do not depend on the index n. In order to illustrate the link between the Extended DCEG and semi-Markov processes we write the semi-Markov kernel as

$$Q_{ij}(t) = p_{ij} F_{ij}(t), \tag{14}$$

where

$$F_{ij}(t) = P(H_{n+1} \leq t | X_{n+1} = j, X_n = i) \tag{15}$$

is the conditional holding time distribution, i.e. the holding time at $X_n = i$ assuming that we move to $X_{n+1} = j$ next and p_{ij} is given in Definition 10.

We now show that the class of time-homogeneous semi-Markov processes is a subclass of the time-homogeneous Extended DCEG and hence that the Extended DCEG is a more general class of model:

Theorem 11. *Let an Extended DCEG be simple and let no two children lead from the same parent into the same child. Then the DCEG is a semi-Markov process with state space* W *(the set of positions), with conditional holding time distributions*

$$F_{w_i w_j}(t) = P(H_{w_i j} \leq t), \tag{16}$$

whenever $e_{w_i j} = e(w_i, w_j)$ *exists and* 0 *otherwise, and with the entries of the transition probability matrix of the embedded Markov Chain* $\{X_n, n \in \mathbb{N}\}$ *given by*

$$p_{w_i w_j} = P(e_{w_i j} | w_i), \tag{17}$$

if the edge $e_{w_i j} = e(w_i, w_j)$ *exists and* 0 *otherwise. If the position* w_0 *is a source node then the state-transition diagram of the semi-Markov process omits* w_0 *and the initial distribution is given by* π_{w_0}*. Otherwise the initial distribution assigns probability* 1 *to* w_0*.*

Proof. See Appendix ■

The correspondence above enables us to use the theory of semi-Markov processes to calculate for example the probability of staying in a position w_i for a certain time $t \geq 0$ and then moving to w_j, directly from the semi-Markov kernel. Similarly, we can calculate the probability that we are in position w_j at time t given that we were initially in position w_i from the transition matrix of the corresponding semi-Markov process, which can be derived as shown for example in [22].

3. Dynamic Bayesian Networks and the Two Time-Slice DCEG

Many formulations of stochastic processes can be represented by a DCEG with a finite graph. In this section we restrict ourselves to demonstrate this for just one

of these classes, the discrete DBN. It has been shown in [1] and [24] how a BN can be written as a staged tree and hence as a CEG. This can be simply extended to a dynamic setting and we show below how a DBN can be represented as an infinite staged tree and therefore as a DCEG. We formally write down the restrictions the BN and DBN impose on the staged tree structure and hence demonstrate the added flexibility of the DCEG by allowing for asymmetric dependence structures of the variables across time. It is also easy to check that many other processes such as dynamic context-specific BNs [16, 17] or dynamic Bayesian multinets [18, 19] are amenable to this representation. In subsection 3.2 we further introduce the two time-slice DCEG as a particular subclass of the DCEG. Note that in this section we are focusing again on the DCEG introduced in Section 2.2 with regular time-steps, where one-step transitions are known and holding times do not need to be explicitly considered.

3.1. The Relationship between DBNs and DCEGs

We first recall the definition of a general DBN:

Let $\{\mathbf{Z}_t : t \in I\}$ where $I = \{t_0, t_1, t_2, \ldots\}$ be a vector stochastic process. Assume that at each time point t, we have a vector of n_t variables $\mathbf{Z}_t = (Z_{1,t}, Z_{2,t}, \ldots, Z_{n_t,t})$, and that the components $Z_{p,t}, p = 1, ..., n_t$ all have a finite number of values. The variables $\mathbf{Z}_t$ then form a time-slice of the DBN for each time point t. Then in the most general case, by the definition of [25], the DBN on $\mathbf{Z}_t$ has an associated infinite acyclic directed graph $\mathcal{G}$ where the component $Z_{p,t}$ of $\mathbf{Z}_t$ has parents

$$pa(Z_{p,t}) = \{Z_{q,s} : t_0 \leq s \leq t, q \in \{1, 2, \ldots, n_s\}\} \tag{18}$$

So there is a directed edge into $Z_{p,t}$ from the components of vectors indexed by a time on or before it. In practice it is often assumed that $s = \{t-1, t\}$ such that the DBN is 1-Markov, assuming that a variable is only affected by variables of the previous time step and the current time step.

We demonstrate below how any general DBN can be written as an infinite staged tree and hence as a DCEG. We first show how to write the variables of the DBN as an infinite tree. We then define the conditional independence statements of the DBN by colouring the florets in the tree to form a stage partition of the situations.

Reindex the variables as $Z_k = Z_{p,t}, k = 1, 2, 3, \ldots$ so that, whenever $Z_i = Z_{q,s} \in pa(Z_{p,t})$, then the index $i < k$. This will ensure that parent variables come before children variables and time-slices come before each other. There is clearly always such an indexing because of the acyclicity and time element of $\mathcal{G}$. This gives a potential total ordering of the variables in $\{\mathbf{Z}_t : t \in I\}$ from which we choose one. We let $\mathbf{Z}^k = \{Z_i : i \leq k\}$.

Example 4. *Consider for example the following DBN on two binary variables $Z_{1,t}, Z_{2,t}$, given in Figure 11.*

We can now reindex the variables of the DBN as follows:

$$Z_1 = Z_{1,t_0}, Z_2 = Z_{2,t_0}, Z_3 = Z_{1,t_1}, Z_4 = Z_{2,t_1}, Z_5 = Z_{1,t_2}, Z_6 = Z_{2,t_2},$$

and thus

$$\mathbf{Z}^6 = \{Z_1, Z_2, Z_3, Z_4, Z_5, Z_6\}.$$

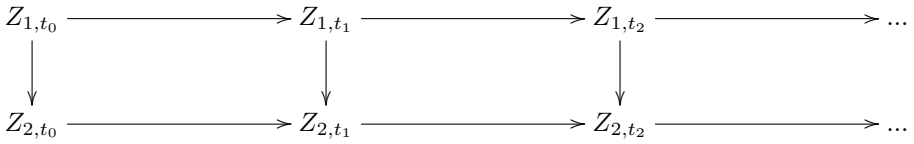

Figure 11: A simple DBN

We can then use the ordering described above to construct a corresponding infinite tree of the DBN: By the assumptions of the ordering the components up to index k can be represented by a finite event tree, which we denote by $\mathcal{T}_k = (V(\mathcal{T}_k), E(\mathcal{T}_k))$. Recall from Section 2.1 that each floret in the tree can be associated with a random variable Z_i and the edges $e_{ij}, j = 1, \ldots, m_i$ describe the m_i values in the sample space that this random variable can take. Hence the paths in the tree $\mathcal{T}_k$ correspond to the set of all combinations of values $\mathbf{Z}^k$ can take. Then a sequential construction of the stochastic process allows us to define a set of trees $\{\mathcal{T}_k\}_{k\geq 1}$, such that $\mathcal{T}_k$ is a subtree of $\mathcal{T}_{k+1}$, recursively as follows:

Let $L(\mathcal{T}_k) = V(\mathcal{T}_k) \setminus S(\mathcal{T}_k)$ be the set of leaf vertices of $\mathcal{T}_k$ and let $l_{ki} \in L(\mathcal{T}_k), i = 1, 2, ..., N_k$ be a single leaf vertex with $\mathcal{T}_k$ having N_k leaf vertices.

1. For $k = 1$, let $\mathcal{T}_1$ be the floret whose edges and receiving leaf vertices label the possible values of Z_1

   Example 4 (continued). *As we have defined $Z_1 = Z_{1,t_0}$, $\mathcal{T}_1$ hence corresponds to the tree given in Figure 12.*

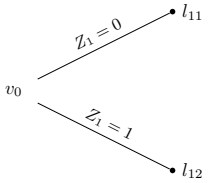

Figure 12: Illustration of $\mathcal{T}_1$

2. We define $E(\mathcal{T}_{k+1}) = E(\mathcal{T}_k) \cup E^+_{k+1}$, where

$$E^+_{k+1} = \{e_{l_{ki}j} : l_{ki} \in L(\mathcal{T}_k), j = 1, 2, ..., m_{k+1}\} \tag{19}$$

   is a set of $N_k \times m_{k+1}$ new edges, with m_{k+1} edges emanating from each vertex $l_{ki}, i = 1, 2, .., N_k$ describing the values the random variable Z_{k+1} can take. Now in $\mathcal{T}_{k+1}$ attach a new leaf vertex to each of the edges in E^+_{k+1}. Then

$$V^+_{k+1} = \{ch(l_{ki}) : l_{ki} \in L(\mathcal{T}_k)\} \tag{20}$$

   and $V(\mathcal{T}_{k+1}) = V(\mathcal{T}_k) \cup V^+_{k+1}$.

   Example 4 (continued). *So in our example to obtain $\mathcal{T}_2$ we attach two edges to each leaf vertex of $\mathcal{T}_1$ and attach a child to each new edge (see Figure 13).*

3. The infinite tree $\mathcal{T}(\mathcal{G})$ of this DBN is now simply defined as

$$\mathcal{T}(\mathcal{G}) = \lim_{k\to\infty} \mathcal{T}_k, \tag{21}$$

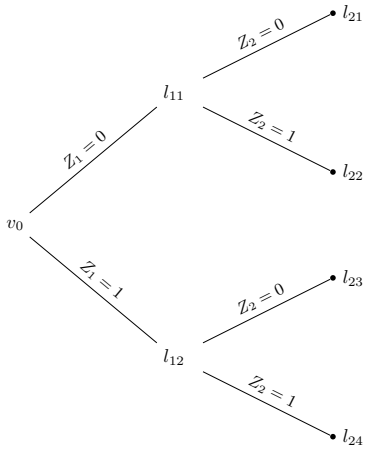

Figure 13: Illustration of $\mathcal{T}_2$

where the vertex set is

$$V(\mathcal{T}(\mathcal{G})) = V(\lim_{k\to\infty} \mathcal{T}_k) \tag{22}$$

and the edge set is given by

$$E(\mathcal{T}(\mathcal{G})) = E(\lim_{k\to\infty} \mathcal{T}_k) \tag{23}$$

The infinite length directed paths starting from the root of this tree correspond to the atoms of the sample space of the process described by $\mathcal{G}$.

We next represent the conditional independencies of the DBN by colouring the vertices and associated edges that are in the same stage as described in Section 2.1. The resulting staged tree then encodes the same conditional independencies as the DBN.

Notice that the vertex $l_{ki} \in V(\mathcal{T}_k) \in V(\mathcal{T}(\mathcal{G}))$ labels the conditioning history of the component Z_{k+1} based on the values of variables listed earlier in its indexing. By the definition of a DBN

$$Z_{k+1} \perp\!\!\!\perp \mathbf{Z}^k | pa(Z_{k+1}). \tag{24}$$

so by our definition l_{ki},$l_{kj} \in V(\mathcal{T})$ are in the same stage whenever

$$P(e_{l_{ki}j}|l_{ki}) = P(e_{l_{kj}j}|l_{kj}) \tag{25}$$

for all edges $e_{l_{ki}j}$ and $e_{l_{kj}j}$, $j = 1, \ldots, m_{k+1}$ or alternatively,

$$P(Z_{k+1} = z_{k+1}|l_{ki}) = P(Z_{k+1} = z_{k+1}|l_{kj}), \tag{26}$$

where z_{k+1} is a value the variable Z_{k+1} can take. If this is true then we assign the same colour to l_{ki} as to l_{kj}. For each possible value z_{k+1} we also assign the same colour to the edges in $\mathcal{T}$ corresponding to the events $\{Z_{k+1} = z_{k+1}|l_{ki}\}$ and $\{Z_{k+1} = z_{k+1}|l_{kj}\}$.

We have thus shown that any DBN can be written as an infinite staged tree.

Example 4 (continued). *In our example we obtain the following colouring on $\mathcal{T}_3$, representing the variables $Z_1 = Z_{1,t_0}, Z_2 = Z_{2,t_0}$ and $Z_3 = Z_{1,t_1}$, given in Figure 14.*

From the DBN in Figure 11 we have that $Z_{1,t_1} \perp\!\!\!\perp Z_{2,t_0} | Z_{1,t_0}$*, or equivalently* $Z_3 \perp\!\!\!\perp Z_2 | Z_1$*. Hence* $P(Z_3 = 0|l_{21}) = P(Z_3 = 0|l_{22})$ *and* $P(Z_3 = 1|l_{21}) = P(Z_3 = 1|l_{22})$ *and, similarly,* $P(Z_3 = 0|l_{23}) = P(Z_3 = 0|l_{24})$ *and* $P(Z_3 = 1|l_{23}) = P(Z_3 = 1|l_{24})$*, and therefore* l_{21} *and* l_{22} *as well as* l_{23} *and* l_{24} *have the same colouring attached to their nodes and corresponding edges.*

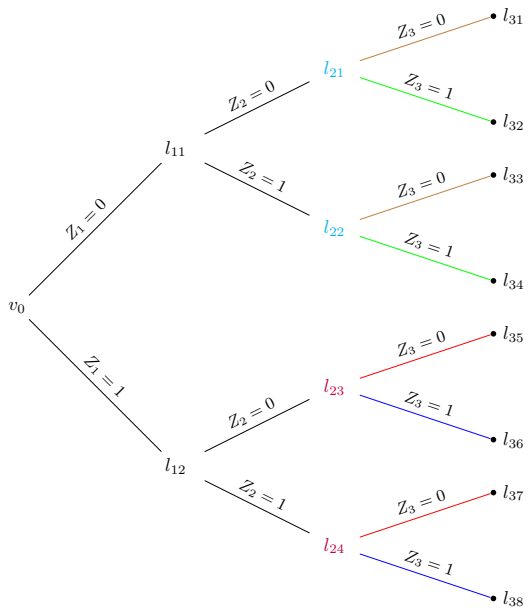

Figure 14: Illustration of $\mathcal{T}_3$

Note that the re-expression of the DBN as a staged tree emphasises how restrictive the general class of DBNs actually is in comparison to the DCEG. The usual classes of DBN only admit certain very specific families of stage partitions. In general in a standard BN, we have that two situations, s_i and s_j, describing a variable Z_k are in the same stage whenever the set of parent variables $pa(Z_k)$ take the same values on the corresponding paths leading to s_i and s_j in the tree. These restrictions are represented within the associated staged tree of the BN as follows:

1. Let two vertices describing a variable Z_k be in the same stage and let their paths differ in some of the values taken by the variables in $\mathbf{Z}^k$. The variables taking the same values on their respective paths is then $pa(Z_k)$. Then all vertices describing the variable Z_k are also in this stage whenever their associated paths differ by the same variables, $\mathbf{Z}^k \backslash pa(Z_k)$.
2. When two vertices describing a variable Z_k are in different stages, then all other vertices whose paths differ by the same variables are also in different stages.

When representing a general DBN as an infinite tree these restrictions are naturally extended across time-slices by the reindexing of the variables.

When the dependence structure is defined through conditional independencies then the DBN is topologically much simpler than the corresponding DCEG. But when, as is often the case, many combinations of values of states are logically impossible and the number of non-zero probability transitions between states is small then the DCEG depicts these zeros explicitly and can sometimes be topologically simpler than the

DBN. Consider, for example Figure 14. If the conditional probability tables (CPTs) of the BN state that $P(Z_2 = 1|Z_1 = 1) = 0$ then the edge describing this probability can be omitted from the tree and the tree is hence reduced to three quarters of its size. Hence unlike the BN and its dynamic analogue, as well as depicting independence relationships the DCEG also allow us to read zeros in the corresponding transition matrix, represented by missing edges in the tree. This is particularly helpful when representing processes which have many logical constraints.

3.2. The Two Time-Slice CEG

A widely used subclass of the general DBN is the two time-slice DBN (2T-DBN), which restricts the process described to be defined on two time-slices and the links between them. In this section we introduce a two time-slice DCEG (2T-DCEG) which imposes similar restrictions on the more general class of DCEGs. We first briefly review the 2T-DBN to demonstrate the analogy between these models and the 2T-DCEG.

The 2TDBN defined as in [26] makes the Markov assumption that a variable can only depend on contemporary variables and variables from the previous time-slice. Further, it demands that the dependence structure and the associated CPTs are the same across all time-slices for $t \geq t_1$. The 2T-DBN with graph $\mathcal{G}$ can therefore be completely defined according to [26] by the following:

1. The set of vertices $V(\mathcal{G})$,
2. its edge set, $E(\mathcal{G})$ consisting of edges within a time-slice t and between two time-slices t and $t + 1$,
3. the associated CPTs for the first time-slice t_0 and the CPTs for the $t + 1$ time-slice (with parents from time-slice t and $t + 1$).

Analogous to the 2T-DBN the two time-slice DCEG (2T-DCEG) describes a discrete-time process on $I = \{t_0, t_1, t_2, \ldots\}$ with its vertices being associated with adjacent time-slices. The 2T-DCEG then makes the additional assumption that the graphical structure and the CPVs associated with each stage, $\pi_u, u \in U$, are the same across the time-slices t, $t \geq t_1$. This further requires that the set of variables considered during each time-slice t is the same and also the variables have the same ordering in its associated infinite tree. The 2T-DCEG has, analogously to the DCEG, a set of positions associated with each variable Z_p, which describes the different states an individual can be in before the event $Z_p = z$ occurs (see Section 2.1). However, unlike the DCEG, the 2T-DCEG requires that at any time step $t \geq t_1$ we assume the same dependence structure between the variables and the same set of CPVs. Hence each variable Z_p will only have one associated set of positions describing the unfolding events, $Z_p = z$, for all time steps $t \geq t_1$. This contrasts with the DCEG which may have a different set of positions for every $Z_{p,t}$.

To summarise, the above restrictions are represented in the topology of the 2T-DCEG as follows:

- the variables $Z_{p,t}$ for $t \geq t_1$ are represented by a unique set of positions, and
- towards the end of every time slice the edges loop back to the start of the time slice

We will now illustrate these restrictions imposed onto the general DCEG structure

in the following real-world example.

Example 5. *We here consider a small subset of the Christchurch Health and Development Study, previously analysed in [27, 24], which studied around 1000 individuals and collected yearly information about events in each of these people's history over the first five years of the children's life. We here consider only the relationships of the following variables given below.*

- *Financial difficulty - a binary variable, describing whether the family is likely to have financial difficulties or not,*
- *Number of life events - a categorical variables distinguishing between* 0, $1-2$ *and* ≥ 3 *life events that may occur in one year,*
- *Hospital admission - a binary variable, describing whether the child is admitted to hospital or not.*

In this setting each time slice corresponds to a year of a child's life starting from when the child is one year old, $t_0 = 1$. *A plausible 2T-DCEG could be the one given in Figure 15. Note that this 2T-DCEG assumes that whether the individual is admitted to hospital or not does not affect the subsequent variables. This is evident from the double arrows from* w_3 *to* w_6, w_4 *to* w_7 *and* w_5 *to* w_8. *Also, observe that the variable describing the hospital admission is not included at time* $t = 0$, *as it does not provide additional information under this assumption.*

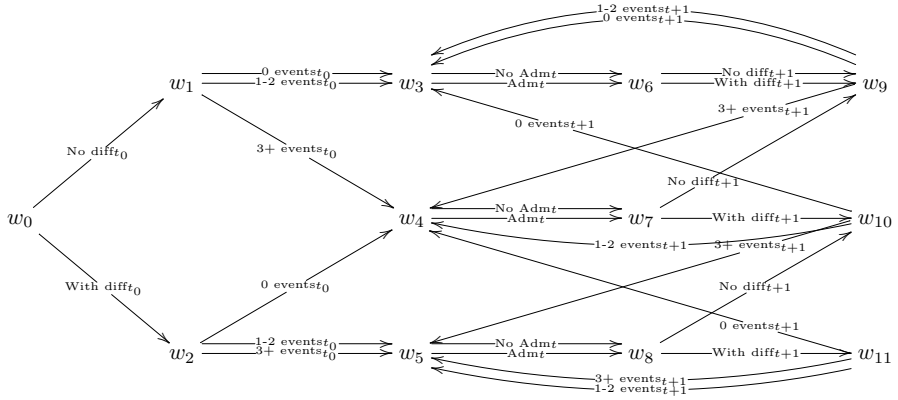

Figure 15: Two Time-Slice DCEG

We start at w_0 *in order to follow the path an individual might take through the DCEG across time. The first part of the graph describes the initial CPVs at time* t_0. *It is first resolved whether or not the family has financial difficulties (*$w_0 \to w_1$, $w_0 \to w_2$*) and whether the individual experiences* 0, $1-2$ *or* ≥ 3 *life events during this year (*$w_1 \to w_3$,$w_1 \to w_4$, $w_2 \to w_4$, $w_2 \to w_5$*). She then reaches one of the three positions* w_3, w_4 *and* w_5 *describing a 'health state' the individual is in before a hospital admission may occur. Independent of whether an admission has occurred or not she then moves to positions* w_6, w_7, w_8, *which describe the same three health states. Then, given the individual is in one of the three health states (*w_3, w_4, w_5*) at time* t, *for* $t \geq t_1$, *she traverses through the graph in the following year according to the financial difficulty in year* $t+1$ *and number of life events in year* $t+1$ *and ends up in one of the three previous health states again.*

By the definition of a DCEG the probability of an individual having a hospital admission at time t is given by $P(Adm = 1|w_3) = \pi_{w_3}$ or $P(Adm = 1|w_4) = \pi_{w_4}$

or $P(Adm = 1|w_5) = \pi_{w_5}$ depending on the position the individual is in at time t. These positions are reached depending on the number of life events and the financial difficulty in that year and the health state of the previous year, which is again determined by the financial difficulty and the number of life events of the year before. Hence the positions of the 2T-DCEG encode the entire history of an individual and we can trace back the full path the individual has taken through the graph. Note that this is a weaker assumption than that of the the 2T-DBN, whose CPTs only condition on the values of the variables in the current and previous year. So, although the representation of the 2T-DCEG is given only by an initial part and by the transitions from t to $t+1$ the model class can still take into account the full longitudinal history of the individual as the CPVs of the 2T-DCEG conditioned on the positions of the graph.

4. Bayesian Learning of the Parameters of a DCEG

Finally, we discuss the learning of the CPVs in a DCEG as well as the learning of the parameters of the holding time distributions in an Extended DCEG. Conjugate learning in non-dynamic CEGs, which can accommodate for both sampling schemes and causal experimental data, is now well documented [2, 28], where the methods closely resemble analogous learning in discrete BN's [29]. Even when these probabilities are believed to drift in time, learning schemes can sometimes still be devised in a closed form analysis (see [21]). We now show how we can extend the methods developed in [2] to the DCEG.

Recall that the DCEG has a set of stages $u \in U$ and that each position in u has m_u edges emanating from it in the graph of the DCEG. As defined in equation 2 we have associated with each stage u a CPV $\pi_u = (\pi_{u1}, \pi_{u2}, .., \pi_{m_u})$ and we denote the full set of CPVs by π. As introduced in Section 2.3, we can further attach a vector of conditional holding time distributions $(F_{u1}, F_{u2}, \ldots, F_{um_u})$ to each stage u with parameters $\lambda_u = (\lambda_1, \lambda_2, \ldots, \lambda_{m_u})$. We call the full set of parameters λ. Within a Bayesian framework we can then learn these parameters from data.

For each individual we can record the positions and edges he passes along and the holding times at each position. Recall from Section 2.3 that we let W_n be the random variable describing the n_{th} position reached, H_n the holding time at position W_n, and E_n the n_{th} edge passed along. Also, note that

$$p(w_1|w_0, e_0, h_0) = \ldots = p(w_N|w_0, e_0, h_0, \ldots, w_{n-1}, e_{n-1}, h_{n-1}) = 1, \qquad (27)$$

as given the previous position and the next edge the individual goes along uniquely determines the next position reached. We can then simplify the joint probability density of each individual using equation 27 and definition 10 to obtain

$$p(w_0, e_0, h_0, w_1, e_1, h_1, \ldots, w_N, e_N, h_N) = p(e_0, h_0|w_0)p(e_1, h_1|w_1)p(e_N, h_N|w_N) \qquad (28)$$

We know from Section 2.3 equation 10 that under time-homogeneity this is simply a product of the CPVs π_{uj} and the conditional holding time densities $f_{uj}(.)$ such that

$$p(w_0, e_0, h_0, w_1, e_1, h_1, \ldots, w_N, e_N, h_N) = \prod_{u \in U} \prod_{j=1}^{m_u} \pi_{uj}^{x_{uj}} \prod_{l=1}^{x_{uj}} f_{uj}(h_{ujl}), \qquad (29)$$

where the individual goes x_{uj} times along an edge $e_{wj}, w \in u$ each time staying a time h_{ujl} at the previous position.

We can now easily generalise this to a complete random sample of individuals going through the tree: We record the number of times the individuals pass along a position $w \in u$ and go along the j_{th} edge, $j = 1, .., m_u$, which we denote by N_{uj}. In addition, we let $\mathbf{h}_{uj}, j = 1, .., m_u, u \in U$ be the vector of conditional holding times for the individuals which arrive at stage u and move along the j_{th} edge next and we let h_{uji} be the holding time of the i_{th} pass along this edge. We denote the full set of holding times by $\mathbf{h} = \{\mathbf{h}_{uj}, u \in U, j = 1, .., m_u\}$ and the set of the number passes along each edge by $\mathbf{N} = \{N_{uj}, u \in U, j = 1, .., m_u\}$.

Then, immediately from the definition of a time-homogeneous and semi-Markov Extended DCEG $\mathcal{D}$, the likelihood $L(\pi, \lambda|\mathbf{N}, \mathbf{h}, \mathcal{D})$ of this random sample separates. Explicitly we have that

$$L(\pi, \lambda|\mathbf{N}, \mathbf{h}, \mathcal{D}) = L_1(\pi|\mathbf{N}, \mathcal{D})L_2(\lambda|\mathbf{h}, \mathbf{N}, \mathcal{D}). \tag{30}$$

Under random sampling, the first part of the likelihood can be written in the form

$$L_1(\pi|\mathbf{N}, \mathcal{D}) = \prod_{u \in U} L_u(\pi_u|\mathbf{N}_u, \mathcal{D}), \tag{31}$$

and

$$L_u(\pi_u|\mathbf{N}_u, \mathcal{D}) = \prod_{j=1}^{m_u} \pi_{uj}^{Nuj}, \tag{32}$$

where N_{uj} is as above. Note that $L_u(\pi_u|\mathbf{N}_u, \mathcal{D})$ is a multinomial likelihood with probability vector $\pi_u = (\pi_{u1}, \pi_{u2}, .., \pi_{um_u})$ on a sample of size $N = \sum_{j=1}^{m_u} N_{uj}$, the number of times stage u is reached.

The second component of the likelihood in equation (30) could be a composite of likelihoods associated with a variety of sampling distributions. Again given a complete random sample we have that:

$$L_2(\lambda|\mathbf{h}, \mathbf{N}, \mathcal{D}) = \prod_{u \in U} \prod_{j=1}^{m_u} L_{uj}(\lambda_{uj}|\mathbf{h}_{uj}, \mathbf{N}_{uj}, \mathcal{D}). \tag{33}$$

One convenient family of such sampling distributions for holding times, which we will see can support a conjugate analysis, is the Weibull distribution with a known shape parameter k, where $k = 1$ corresponds to the exponential family, when the likelihood takes the form

$$L_{uj}(\lambda_{uj}|\mathbf{h}_{uj}, \mathbf{N}_{uj}, \mathcal{D}) = \prod_{i=1}^{N_{uj}} \frac{k}{\lambda_{uj}^k} h_{uji}^{k-1} exp\left(-\left(\frac{h_{uji}}{\lambda_{uj}}\right)^k\right). \tag{34}$$

From equation 30 it is immediate that if λ and π are believed to be a priori independent so that

$$p(\pi, \lambda|\mathcal{D}) = p_1(\pi|\mathcal{D})p_2(\lambda|\mathcal{D}), \tag{35}$$

then the posterior density $p(\pi, \lambda|\mathbf{h}, \mathbf{N}, \mathcal{D})$ separates into

$$p(\pi, \lambda|\mathbf{h}, \mathbf{N}, \mathcal{D}) = p_1(\pi|\mathbf{N}, \mathcal{D})p_2(\lambda|\mathbf{h}, \mathbf{N}, \mathcal{D}) \tag{36}$$

and we can perform the updating of the CPVs, π, and the holding time parameters, λ, without reference to the other. Note that if the holding times are not recorded, then the second term in equation 30 is simply 1 and $p_2(\lambda|\mathbf{h}, \mathbf{N}, \mathcal{D}) = p_2(\lambda|\mathcal{D})$.

Learning the posterior $p_1(\pi|\mathbf{N}, \mathcal{D})$ is exactly analogous the standard CEG. Thus in [2, 29] it is shown for the CEG that under the assumptions that the $\pi_u, u \in U$ are a priori independent and that equivalent stages in different structures have the same prior, in a sense made explicit in these papers, demands that each stage prior must have a Dirichlet distribution independently. It can be checked that these arguments also apply to the DCEG. An obvious choice of prior to use for π_u is:

$$p_1(\pi|\mathcal{D}) = \prod_{u \in U} p_u(\pi_u|\mathcal{D}) = \prod_{u \in U} \frac{\Gamma(\sum_{j=1}^{m_u} \alpha_{uj})}{\prod_{j=1}^{m_u} \Gamma(\alpha_{uj})} \prod_{j=1}^{m_u} \pi_{uj}^{\alpha_{uj}-1}, \tag{37}$$

where each $\pi_u \sim Dir(\alpha_{u1}, \alpha_{u2}, ..., \alpha_{um_u})$. As with all Bayesian learning some care needs to be taken in the setting of the hyperparameter values. In the simplest case we assume that the paths taken on the associated infinite tree are a priori equally likely and we specify the hyperparameters associated with each floret accordingly. Then, given that the DCEG has an absorbing position w_∞ we can find, under the above assumptions, the $\alpha_u, u \in U$ of the DCEG structure $\mathcal{D}$ derived from the infinite tree by simply summing the hyperparameters of the situations merged. This direct analogue to [2] does however not work when no absorbing position exists, for then these sums diverge. We hence take a slightly different approach in this case and instead make use of the direct correspondence between the DCEG and Markov processes to find the limiting distribution of our process and derive the hyperparameters from this. This implicitly assumes that in this sense the prior beliefs are 'in equilibrium'.

Updating the prior distribution in equation 37 using the first component of the likelihood $L(\pi|\mathbf{N}, \mathcal{D})$ in equation 30 gives the posterior in closed form:

$$p_1(\pi|\mathbf{N}, \mathcal{D}) = \prod_{u \in U} p_u(\pi_u|\mathbf{N}_u, \mathcal{D}) = \prod_{u \in U} \frac{\Gamma(\sum_{j=1}^{m_u} \alpha_{uj} + N_{uj})}{\prod_{j=1}^{m_u} \Gamma(\alpha_{uj} + N_{uj})} \prod_{j=1}^{m_u} \pi_{uj}^{\alpha_{uj}+N_{uj}-1}. \tag{38}$$

Note that when the holding time distributions are identical across the model space the above conjugate analysis suffices to learn the parameters of the DCEG. We illustrate below how we can update the CPVs in a DCEG using the Christchurch example of Section 3.2.

Example 5 (continued). *Recall the 2T-DCEG of the previous section, which we repeat in Figure 16. Note that again the stages and positions of the graph coincide and hence learning the stage parameters is equivalent to learning the position parameters of the graph.*

To specify the stage priors, we determine the hyperparameters α_u of the Dirichlet distribution associated with each stage u as suggested above as follows: We first find the limiting distribution of the Markov process with state space

$$W = \{w_3, w_4, w_5, w_6, w_7, w_8, w_9, w_{10}, w_{11}\}$$

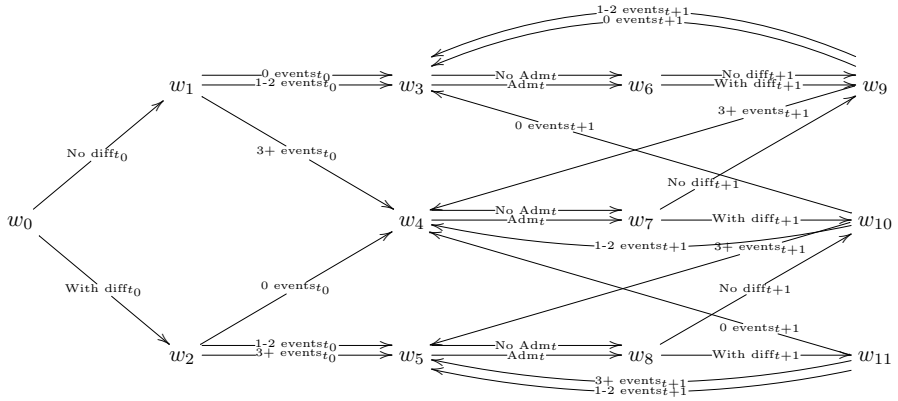

Figure 16: Two Time-Slice DCEG

and with transition probability matrix:

$$P = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{2}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & \frac{2}{3} & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Observe that the transition probability matrix assumed that all paths in the graph are equally likely. Solving the general balance equations we then deduce that $P(W = w_3) = P(W = w_6) = \frac{5}{9}$, $P(W = w_4) = P(W = w_7) = \frac{1}{3}$, $P(W = w_5) = P(W = w_8) = \frac{1}{9}$, $P(W = w_9) = \frac{13}{18}$, $P(W = w_{10}) = \frac{2}{9}$, $P(W = w_{11}) = \frac{1}{18}$, *which together with an equivalent sample size of 3 (equal to the largest number of categories a variable of the problem takes [30]), determines the strength of the prior on each stage. Further, assuming that the probabilities on the edges emanating from each position are uniform we can deduce the stage priors to be as given in Table 1.*

We can now update these priors separately and in closed form for each stage using the data available to us. We have complete data available from 1062 children born in Christchurch, New Zealand, for the first $2-5$ *years of their lives. We use the data from year* 2 *to update the initial positions* w_0, w_1 *and* w_2 *and then use the hospital admissions variable of year* 2, *as well as years* $3-5$, *to update the remaining CPVs. Doing so we obtain the following posterior distributions associated with each stage given in Table 1 with their corresponding means and* 95% *credible intervals. Thus, for example, the expected probability of an individual being admitted to hospital is* 0.07 *given he has reached position* w_3, *which means that he was previously in position* w_3 *and had fewer than* 3 *life events in the current year or that he was previously in state* w_4 *and then had no financial difficulties and less than* 3 *events in the current year or had financial difficulties but no life events in the current year. Similarly, we have that the probability of an admission when reaching* w_4 *is* 0.11 *and* 0.13 *when passing through* w_5.

We have therefore shown that given a multinomial likelihood and Dirichlet priors

Position	Prior	Posterior	Mean (95% credible interval)		
w_0	$Dir(\frac{5}{2}, \frac{1}{2})$	$Dir(875\frac{1}{2}, 189\frac{1}{2})$	$0.82(0.80, 0.84)$	$0.18(0.16, 0.20)$	
w_1	$Dir(\frac{5}{6}, \frac{5}{6}, \frac{5}{6})$	$Dir(135\frac{5}{6}, 436\frac{5}{6}, 302\frac{5}{6})$	$0.15(0.13, 0.18)$	$0.50(0.47, 0.53)$	$0.35(0.31, 0.38)$
w_2	$Dir(\frac{1}{6}, \frac{1}{6}, \frac{1}{6})$	$Dir(9\frac{1}{6}, 56\frac{1}{6}, 124\frac{1}{6})$	$0.05(0.02, 0.08)$	$0.30(0.23, 0.36)$	$0.65(0.58, 0.72)$
w_3	$Dir(\frac{5}{6}, \frac{5}{6})$	$Dir(1735\frac{5}{6}, 122\frac{5}{6})$	$0.93(0.92, 0.94)$	$0.07(0.06, 0.08)$	
w_4	$Dir(\frac{1}{2}, \frac{1}{2})$	$Dir(766\frac{1}{2}, 98\frac{1}{2})$	$0.89(0.86, 0.91)$	$0.11(0.09, 0.14)$	
w_5	$Dir(\frac{1}{6}, \frac{1}{6})$	$Dir(406\frac{1}{6}, 59\frac{1}{6})$	$0.87(0.84, 0.90)$	$0.13(0.10, 0.16)$	
w_6	$Dir(\frac{5}{6}, \frac{5}{6})$	$Dir(1679\frac{5}{6}, 178\frac{5}{6})$	$0.90(0.89, 0.92)$	$0.10(0.08, 0.11)$	
w_7	$Dir(\frac{1}{2}, \frac{1}{2})$	$Dir(700\frac{1}{2}, 164\frac{1}{2})$	$0.81(0.78, 0.84)$	$0.19(0.16, 0.22)$	
w_8	$Dir(\frac{1}{6}, \frac{1}{6})$	$Dir(227\frac{1}{6}, 238\frac{1}{6})$	$0.49(0.44, 0.53)$	$0.51(0.47, 0.56)$	
w_9	$Dir(\frac{13}{18}, \frac{13}{18}, \frac{13}{18})$	$Dir(516\frac{13}{18}, 989\frac{13}{18}, 352\frac{13}{18})$	$0.28(0.26, 0.30)$	$0.53(0.51, 0.55)$	$0.19(0.17, 0.21)$
w_{10}	$Dir(\frac{2}{9}, \frac{2}{9}, \frac{2}{9})$	$Dir(114\frac{2}{9}, 407\frac{2}{9}, 343\frac{2}{9})$	$0.13(0.11, 0.16)$	$0.47(0.44, 0.50)$	$0.40(0.36, 0.43)$
w_{11}	$Dir(\frac{1}{18}, \frac{1}{18}, \frac{1}{18})$	$Dir(29\frac{1}{18}, 181\frac{1}{18}, 255\frac{1}{18})$	$0.06(0.04, 0.09)$	$0.39(0.35, 0.43)$	$0.55(0.50, 0.59)$

Table 1: Posterior CPVs and 95% credible intervals

as specified above it is possible to define a prior to posterior analysis on the DCEG directly analogous to the BN parameter learning method described by [29] and the CEG learning defined by [2]. In particular, if we are comparing model classes of $\mathcal{D}$, we can also directly use the model selection algorithm developed in [2] on the DCEG, which uses the sum of the log Bayes Factor as a score function to find the maximum a posteriori model in this class.

To learn the parameters of the holding time distributions, we assume that the priors λ_{uj} are mutually independent. We then have that

$$p_2(\lambda|\mathcal{D}) = \prod_{u \in U} \prod_{j=1}^{m_u} p_{uj}(\lambda_{uj}|\mathcal{D}). \tag{39}$$

Putting Inverse-Gamma (IG) priors on each λ_{uj}^k, such that $\lambda_{uj}^k \sim IG(\alpha_{uj}, \beta_{uj})$ it is then easily checked that we can obtain the posteriors in closed form, such that each λ_{uj}^k has an independent $IG(\alpha_{uj} + N_{uj}, \beta_{uj} + \sum_{i=1}^{N_{uj}} h_{uji}^k)$ distribution.

We demonstrate using the trekking example (Example 1) how we can learn the parameters of the CPVs and of the holding time distributions separately and in closed form on a time-homogeneous and semi-Markov DCEG.

Example 1 (continued). *Recall the DCEG of the trekking example which is repeated in Figure 17 and whose stage and position partition is given in equation 11. We assume the DCEG is time-homogeneous and semi-Markov as in the above methodology. To first set up the Dirichlet priors on π_u and the Inverse-Gamma*

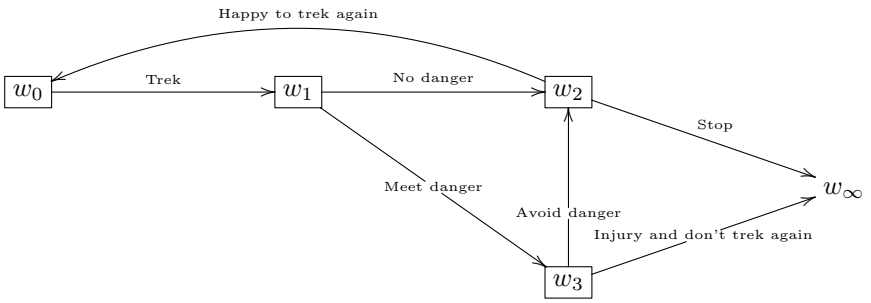

Figure 17: DCEG of trekking example

priors on λ_{uj} we again assume an uninformative prior on the paths of the associated tree and further specify an equivalent sample size of 14 *to ensure that the*

prior Inverse-Gamma distributions have a mean. To determine the hyperparameters α_u of the Dirichlet priors we can here use the standard approach of summing the hyperparameters of the situations in each stage, as, due to the sink node w_∞, the sum will not diverge as in the previous example. Recall from equation 11 that, for example, $u_1 = \{s_1, s_{10}, s_{11}, \ldots\}$. Then under the above assumptions and the tree structure in Figure 9 the situations in u_1 have the distributions: $v_1 \sim Dir(7,7), v_{11} \sim Dir(\frac{7}{4}, \frac{7}{4}), v_{12} \sim Dir(\frac{7}{8}, \frac{7}{8})$. Similarly, the next situations of u_1 will have the distributions $Dir(7 \times \frac{1}{4}^2, 7 \times \frac{1}{4}^2), Dir(7 \times \frac{1}{8}^2, 7 \times \frac{1}{8}^2), \ldots$. The infinite sum of the hyperparameters of these distributions are hence made up of two geometric and so we can obtain the hyperparameters of the prior on u_1 as the limit of these two series, such that we have $\pi_{u_1} \sim Dir(10\frac{1}{3}, 10\frac{1}{3})$. The hyperparameters of the remaining priors on u_2 and u_3 can be found in a similar way and are given together with the priors of the conditional holding times in Table 2.

Description	Holding time distribution	Prior
Hrs until trek	$H_{u_0 1} \sim Exp(\lambda_0)$	$\lambda_0 \sim IG(10\frac{1}{3}, 9\frac{1}{3})$
Meet danger	$N_{u_1} \sim Mult(\pi_{u_1})$	$\pi_{u_1} \sim Dir(10\frac{1}{3}, 10\frac{1}{3})$
Hrs until danger met	$H_{u_1 1} \sim Weibull(\lambda_1, k_1)$	$\lambda_1^{k_1} \sim IG(10\frac{1}{3}, 9\frac{1}{3})$
Hrs until finished trek/no danger met	$H_{u_1 2} \sim Weibull(\lambda_3, k_3)$	$\lambda_3^{k_3} \sim IG(10\frac{1}{3}, 9\frac{1}{3})$
Avoid danger	$N_{u_3} \sim Mult(\pi_{u_3})$	$\pi_{u_3} \sim Dir(4,4)$
Hrs until avoids danger	$H_{u_3 1} \sim Weibull(\lambda_2, k_2)$	$\lambda_2^{k_2} \sim IG(4,3)$
Hrs until injury	$H_{u_3 2} \sim Exp(\lambda_6)$	$\lambda_6 \sim IG(4,3)$
Happy to trek again	$N_{u_2} \sim Mult(\pi_{u_2})$	$\pi_{u_2} \sim Dir(6\frac{2}{3}, 6\frac{2}{3})$
Days until happy to trek	$H_{u_2 1} \sim Exp(\lambda_4)$	$\lambda_4 \sim IG(6\frac{2}{3}, 5\frac{2}{3})$
Hrs until decides to stop	$H_{u_2 2} \sim Exp(\lambda_5)$	$\lambda_5 \sim IG(6\frac{2}{3}, 5\frac{2}{3})$

Table 2: Prior distributions on CPVs and conditional holding times

In this example we propose exponential and Weibull holding time distributions associated with each edge in the graph. For example, it may be plausible to assume that the conditional holding time distribution of $H_{u_0 1}$, is an exponential distribution which describes the time until trekking and with scale parameter λ_0, which is the average number of hours until the individual starts trekking. We assume the more general Weibull distribution on $H_{u_1 1}$, $H_{u_1 2}$ and H_{u_3} with known shape parameters, where $k_1 = 2$, $k_2 = \frac{1}{2}$ and $k_3 = \frac{2}{3}$. We hence put Inverse-Gamma priors on λ_0, $\lambda_1^{k_1}$, $\lambda_2^{k_2}$, $\lambda_3^{k_3}$, λ_4, λ_5 and λ_6 and specify the priors such that we obtain a mean equal to 1 for all prior holding times. Further, we assume an equivalent sample size corresponding to the strength of the prior belief on the edge associated with each conditional holding time distribution (see Table 2).

We now simulate 1000 paths through the DCEG and record the edges and positions of each path as well as the holding time at each position passed through. The paths are simulated by assuming that the individual meets danger with probability $\frac{3}{4}$ and avoids danger, if danger is met, with probability $\frac{2}{3}$. The probability of trekking again is also $\frac{3}{4}$. Similarly, we simulate from an exponential distribution with parameter 5 to describe the time until trekking, which corresponds to the individual trekking on average after 5 hours. Further we simulate from a $Weibull(3,2)$ for the time until meeting danger, from a $Weibull(5, \frac{2}{3})$ for the time until the trek is finished and from a $Weibull(2, \frac{1}{2})$ to describe the time until danger is avoided. Finally, we choose exponential distributions with parameters 10, 3 and 2 for the days until the trekker is happy to trek again, the hours until deciding to stop for good and the hours until injury. We can then update the prior distributions given the random sample of individuals to obtain the following posterior distributions given in Table 3 together

with the mean and standard deviation of interest.

Description	Posterior	Mean	Standard deviation
Hrs until trek	$\lambda_0 \sim IG(1010\frac{1}{2}, 5018.58)$	4.97	0.16
Meet danger	$\pi_{u_1} \sim Dir(761\frac{1}{3}, 259\frac{1}{3})$	0.75	0.01
Hrs until danger met	$\lambda_1^{k_1} \sim IG(259\frac{1}{3}, 2056.59)$	2.50	0.08
Hrs until finished trek/no danger met	$\lambda_3^{k_3} \sim IG(761\frac{1}{3}, 2128.16)$	15.69	1.30
Avoid danger	$\pi_{u_3} \sim Dir(176, 81)$	0.68	0.03
Hrs until avoids danger	$\lambda_2^{k_2} \sim IG(176, 251.60)$	2.29	0.07
Hrs until injury	$\lambda_6 \sim IG(81, 150.39)$	1.88	0.21
Happy to trek again	$\pi_{u_2} \sim Dir(700\frac{2}{3}, 234\frac{2}{3})$	0.75	0.01
Days until happy to trek	$\lambda_4 \sim IG(701\frac{2}{3}, 7209.729)$	10.29	0.39
Hrs until decides to stop	$\lambda_5 \sim IG(234\frac{2}{3}, 629.64)$	2.69	0.18

Table 3: Posterior distributions on CPVs and conditional holding times with mean and standard deviation

The table shows that after learning the parameters the expected number of hours until the trekker decided to trek is 4.97 *hours. He is then expected to meet danger with probability* 0.75 *and the expected number of hours until this happens is* 2.5*. Having met danger,the expected probability of avoiding danger is* 0.68*. We can further deduce that danger is avoided on average after* 2.29 *hours and when an injury occurs this is expected to happen on average after* 1.87 *hours. If danger is not met the trekker finishes the trek average after* 15.69 *hours. After a safe return the trekker is happy to trek again with probability* 0.75 *after on average* 10.29 *days or decides to give up trekker after* 2.69 *hours.*

Because the estimation above is in closed form, marginal likelihood scoring methods are very quick and elegant. Thus, note that when the likelihood separates as in the above situations then the marginal likelihood of a DCEG structure given a complete random sample $L(\mathcal{D}|\mathbf{h}, \mathbf{N})$ also separates into into two parts one associated with the transitions and another with the holding times:

$$L(\mathcal{D}|\mathbf{h}, \mathbf{N}) = L_1(\mathcal{D}|\mathbf{N})L_2(\mathcal{D}|\mathbf{h}, \mathbf{N}). \tag{40}$$

Then, exactly analogously to the finite CEG the first component of the marginal likelihood of a DCEG takes the form:

$$L_1(\mathcal{D}|\mathbf{N}) = \prod_{u \in U} \frac{\Gamma(\sum_{j=1}^{m_u} \alpha_{uj})}{\Gamma(\sum_{j=1}^{m_u} \alpha_u + N_u)} \prod_{j=1}^{m_u} \frac{\Gamma(\alpha_{uj} + N_{uj})}{\Gamma(\alpha_{uj})}. \tag{41}$$

After a little algebra the second component of the marginal likelihood associated with, for example, exponential holding times distributions can be written as:

$$L_2(\mathcal{D}|\mathbf{h}, \mathbf{N}) = \prod_{u \in U} \prod_{j=1}^{m_u} \frac{\beta_{uj}^{\alpha_{uj}}}{\Gamma(\alpha_{uj})} \frac{\Gamma(\alpha_{uj} + N_{uj})}{\beta_{uj} + \sum_{j=1}^{N_{uj}} h_{uji}^{\alpha_{uj}+N_{uj}}}. \tag{42}$$

When the prior distributions on λ are the same for all DCEG structures the log marginal likelihood, *log* $L(\mathcal{D}|\mathbf{h}, \mathbf{N})$, can be written as a linear function of scores associated with different components of the models. In exactly the same way as in [24] we can then score the models and select the DCEG model with the highest maximum a posteriori score. Notice that the linearity of the score means that priors can be set so that two models with shared structure will score the same for that component of the sum, and so the difference in score will automatically ignore scores of shared components, just as for BNs and CEGs. So, because of conjugacy, the

framework we give here can also be used for a fast search to find a best fitting dynamic model within a given class. Of course, without further constraints, the size of the model class of DCEGs is vast. So, clever techniques need to be devised to traverse the model space and families of models developed to restrict the space. However, we are currently working with a number of examples in the dynamic domain and early results are promising and will be reported later.

5. Conclusion

We have demonstrated here that a dynamic version of the CEG is straightforward to develop and that this class enjoys most of the convenient properties of the CEG. It further usefully generalises the discrete DBN when the context demands it. Although we do not envisage the DCEG taking over from the DBN as a representational device and framework for structured stochastic propagation and learning we nevertheless believe that it provides a valuable complementary tool to alternative graphical models. It is particularly suited to domains where the levels of state vectors are numerous but the associated transitions are sparse, or when context-specific symmetries abound. The fact that their finite analogues express BNs as a special case and standard learning algorithms for these classes nest into each other means that the DCEG and DBN representations are particularly complementary: the first focuses on the micro structure of the transitions between states of the process whilst the other focuses on the macro elements of the relationships between relevant variables within the study domain. Further, the link between (semi-)Markov processes and DCEGs suggests that the technology of the processes can be simply extended to DCEGs, hence enabling the DCEG to be applicable to a wide range of domains.

Appendix

The proof of Theorem 11 is given below:

Proof. By equation 14 we know that a semi-Markov process can be specified by an initial distribution α, its conditional holding time distributions and the transition probability matrix $P = (p_{ij})$ of its embedded Markov Chain $\{X_n : n \in \mathbb{N}\}$. Let the state space of the semi-Markov process be
$W = \{w_0, w_1, w_2, .., w_\infty\}$, the set of positions in the DCEG. We then have that

$$p_{w_i w_j} = P(X_{n+1} = w_j | X_n = w_i) = P(e(w_i, w_j)|w_i), \tag{43}$$

whenever the edge $e(w_i, w_j)$ exists in the DCEG, where the probabilities $P(e(w_i, w_j)|w_i) = P(e_{w_i j}|w_i)$ are given by the CPVs of the DCEG. When the edge $e(w_i, w_j)$ does not exist then $p_{w_i w_j} = 0$. Relabelling the positions such that $w_0 = 1, w_1 = 2, ..w_\infty = n$, where n is the number of positions in the DCEG, then gives us the probability transition matrix P of the embedded Markov chain with state space $\{1, 2, .., n\}$ and ij_{th} entry p_{ij}. Substituting as above into equation 15 we further have that

$$\begin{aligned} F_{w_i w_j}(t) &= P(H_{n+1} \leq t | X_{n+1} = w_j, X_n = w_i) \\ &= P(H_{n+1} \leq t | X_n = w_i, e(w_i, w_j)) = P(H_{w_i j} \leq t), \end{aligned} \tag{44}$$

whenever $e(w_i, w_j)$ exists and 0 otherwise. Hence by the relabelling above, the conditional holding time distributions of the semi-Markov process are $\{F_{ij}(t) : i, j \in \{1, 2, .., n\}$. The initial distribution simply assigns probability 1 to w_0. However,

when w_0 is never reached again and simply serves as the starting point of the process, then the state-transition diagram of the semi-Markov process omits w_0 and the initial distribution becomes π_{w_0}. ■

Acknowledgements

The authors would like to thank John Horwood and the CHDS research group for providing the data set. One of the authors was funded by the EPSRC.

References

[1] J. Q. Smith, P. E. Anderson, Conditional independence and Chain Event Graphs, Artificial Intelligence 172 (2008) 42–68.
[2] G. Freeman, J. Q. Smith, Bayesian MAP model selection of Chain Event Graphs, Journal of Multivariate Analysis 102 (2011) 1152–1165.
[3] P. A. Thwaites, Causal identifiability via Chain Event Graphs, Artificial Intelligence 195 (2013) 291–315.
[4] L. M. Barclay, J. L. Hutton, J. Q. Smith, Chain Event Graphs for Informed Missingness, Bayesian Analysis (2013).
[5] C. Hitchcock, The intransitivity of causation revealed in equations and graphs, The Journal of Philosophy 98 (2001) 273–299.
[6] P. A. Thwaites, J. Q. Smith, E. Riccomagno, Causal analysis with Chain Event Graphs, Artificial Intelligence 174 (2010) 889–909.
[7] E. Riccomagno, J. Q. Smith, The geometry of causal probability trees that are algebraically constrained, Optimal Design and Related Areas in Optimization and Statistics (2009) 133–154.
[8] P. A. Thwaites, J. Q. Smith, Evaluating causal effects using Chain Event Graphs, in: Proceedings of PGM, 2006, pp. 293–300.
[9] P. A. Thwaites, J. Q. Smith, R. G. Cowell, Propagation using Chain Event Graphs, in: Proceedings of the Twenty-Fourth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-08), AUAI Press, Corvallis, Oregon, 2008, pp. 546–553.
[10] P. A. Thwaites, J. Q. Smith, Non-symmetric models, Chain Event Graphs and propagation, in: Proceedings of IPMU, 2006, pp. 2339–2347.
[11] S. French, D. R. Insua, Statistical decision theory: Kendall's Library of Statistics 9, Wiley, 2010.
[12] J. J. Oliver, Decision graphs-an extension of decision trees, in: Proceedings of the Fourth International Workshop on Artificial Intelligence and Statistics, 1993, pp. 343–350.
[13] M. Jaeger, Probabilistic Decision Graphs - combining verification and AI techniques for probabilistic inference, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 12 (2004) 19–42.
[14] M. Jaeger, J. D. Nielsen, T. Silander, Learning Probabilistic Decision Graphs, International Journal of Approximate Reasoning 42 (2006) 84–100.
[15] J. D. Nielsen, R. Rumí, A. Salmerón, Structural-EM for learning PDG models from incomplete data, International Journal of Approximate Reasoning 51 (2010) 515–530.
[16] C. Boutilier, N. Friedman, M. Goldszmidt, D. Koller, Context-specific independence in Bayesian Networks, in: Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence, 1996, Morgan Kaufmann Publishers Inc.,

1996, pp. 115–123.
[17] N. Friedman, M. Goldszmidt, Learning Bayesian Networks with local structure, in: M. I. Jordan (Ed.), Learning in Graphical Models, MIT Press, 1998, pp. 421–460.
[18] D. Geiger, D. Heckerman, Knowledge representation and inference in similarity networks and Bayesian multinets, Artificial Intelligence 82 (1996) 45–74.
[19] J. A. Bilmes, Dynamic Bayesian Multinets, Morgan Kaufmann Publishers Inc., 2000, pp. 38–45.
[20] P. A. Thwaites, J. Q. Smith, Separation theorems for Chain Event Graphs, CRiSM Research Report 11-09 (2011).
[21] G. Freeman, J. Q. Smith, Dynamic staged trees for discrete multivariate time series: forecasting, model selection and causal analysis, Bayesian Analysis 6 (2011) 279–305.
[22] V. S. Barbu, N. Limnios, Semi-Markov chains and hidden semi-Markov models toward applications: their use in reliability and DNA analysis, volume 191, Springer, 2008.
[23] J. Medhi, Stochastic processes, New Age International, 1994.
[24] L. M. Barclay, J. L. Hutton, J. Q. Smith, Refining a Bayesian Network using a Chain Event Graph, International Journal of Approximate Reasoning (2013). 10.1016/j.ijar.2013.05.006.
[25] K. P. Murphy, Machine learning: a probabilistic perspective, The MIT Press, 2012.
[26] K. B. Korb, A. E. Nicholson, Bayesian Artificial Intelligence, volume 1, cRc Press, 2004.
[27] D. M. Fergusson, L. J. Horwood, F. T. Shannon, Social and family factors in childhood hospital admission., Journal of Epidemiology and Community Health 40 (1986) 50.
[28] J. Q. Smith, Decision Analysis - Principles and Practice, Cambridge University Press, 2010.
[29] D. Heckerman, A tutorial on learning with Bayesian Networks, Innovations in Bayesian Networks (2008) 33–82.
[30] R. E. Neapolitan, Learning Bayesian Networks, Pearson Prentice Hall Upper Saddle River, 2004.