Using Chain Event Graphs to refine model selection

Peter Thwaites
University of Warwick, UK
Peter.Thwaites@warwick.ac.uk

Abstract

Chain Event Graphs (CEGs) are specifically designed to embody the conditional independence structure of problems whose state spaces are asymmetric and do not admit a natural product structure. The learning of CEGs is closely related to the learning of BNs, and if we use (for example) MAP model selection then where a model can be represented as both a BN and a CEG, the two methods assign this model the same score. If we suspect that a problem incorporates significant context-specific conditional independence structure we can use standard BN-based learning methods to select a good approximate model, and then use the CEG-based learning methods described here to further refine this model.

1 Introduction

The Chain Event Graph (CEG) (Smith and Anderson, 2008; Thwaites et al., 2008; Thwaites et al., 2010) is a graphical model which captures the conditional independence structure of problems which do not admit a natural product structure on their state spaces. Such problems often have no satisfactory representation as a BN or context-specific BN.

Specifically, a CEG is a function of an event tree. These trees (Shafer, 1996) are particularly suited to problems displaying asymmetry, but are not ideal for the representation of the conditional independence structure of a problem. The CEG has been developed to solve this fault.

A formal description and motivation for using CEGs, and an outline of some of their implicit conditional independence structure can be found in (Smith and Anderson, 2008). Three points from this paper are key to the ideas presented here. Firstly, problem asymmetries are represented explicitly in the topology of the CEG. Secondly, CEGs can be used to express a richer set of conditional independence statements not simultaneously expressible through a single BN. Lastly, the class of BNs is contained within that of CEGs. This is a property which we exploit here, since with appropriate prior settings, it follows that BN model selection procedures can be nested within those for CEGs.

Fast propagation algorithms for CEGs were developed in (Thwaites et al., 2008). These exploit the graph's embedded conditional independencies to factorize its mass function over local masses. In this paper we demonstrate how this factorization of the joint mass function over a given event space can also be used as a framework for searching over a space of promising candidate CEGs to discover models which provide good qualitative explanations of the underlying data generating process of a given data set. Because these search methods are similar to well known algorithms used for searching BNs we are able to use similar arguments for setting up hyperparameters over priors so that the priors over the model space decompose as collections of local beliefs.

In particular, as the sets of conditional independence statements expressible via a CEG are larger than the sets expressible via a BN, we can use CEG-based techniques to refine BN-based model selection. We first find one or more BNs which we believe adequately describe the problem, and then use the methods described in this

paper to ascertain whether there are context-specific adaptations of these models which are better reflections of the problem.

Section 2 briefly describes the process by which we create a CEG from an event tree. Section 3 introduces the techniques for learning CEGs. In section 4 we provide an example of how BN-based model selection can be refined by the use of CEG-based techniques. Further discussion appears in section 5.

2 Producing a CEG

Starting with an event tree $\mathcal{T}$ (vertex set $V(\mathcal{T})$, edge set $E(\mathcal{T})$), a probability tree can be specified by assigning probabilities to each member of $E(\mathcal{T})$.

Letting $\mathcal{T}(v)$ be the subtree of $\mathcal{T}$ rooted in the vertex v ($\in V(\mathcal{T})$), we say that the vertices v_1 and v_2 are in the same *position* if:

- the subtrees $\mathcal{T}(v_1)$ and $\mathcal{T}(v_2)$ have identical topologies,
- there exists a map between $\mathcal{T}(v_1)$ and $\mathcal{T}(v_2)$ such that corresponding edges in the two subtrees are labelled with the same outcomes (given different problem developments upto v_1 and v_2) and the same probabilities.

The set $K(\mathcal{T})$ of positions w partitions $V(\mathcal{T})$.

The CEG $\mathcal{C}$ is a coloured directed graph with vertex set $V(\mathcal{C}) = K(\mathcal{T}) \cup \{w_\infty\}$, and edge set $E(\mathcal{C})$. There exists an edge $e \in E(\mathcal{C})$ from w_1 to $w_2 \neq w_\infty$ for each vertex $v_2 \in w_2$ which is a child of a fixed representative $v_1 \in w_1$ for some $v_1 \in V(\mathcal{T})$, and an edge from w_1 to w_∞ for each leaf-node $v \in V(\mathcal{T})$ which is a child of a fixed representative $v_1 \in w_1$ for some $v_1 \in V(\mathcal{T})$.

The *floret* $F(w)$ of a position $w \in V(\mathcal{C})$ is w together with the set of outgoing edges from w. We say that the positions w_1 and w_2 are in the same *stage* u if:

- the florets $F(w_1)$ and $F(w_2)$ have identical topologies,
- there exists a map between $F(w_1)$ and $F(w_2)$ such that corresponding edges in the two florets are labelled with the same outcomes (given different problem developments upto v_1 and v_2) and the same probabilities.

For w_1, w_2 in the same stage the corresponding edges of $F(w_1)$ and $F(w_2)$ have the same *colour* (see positions w_1 and w_2 and their outgoing edges in Figure 2). For any $w \in u$ we can, without ambiguity let the *stage floret* $F(u)$ be u together with a set of edges labelled with the same events and probabilities as the outgoing edges of w.

The process of producing a CEG from a tree is illustrated in Example 1.

Example 1

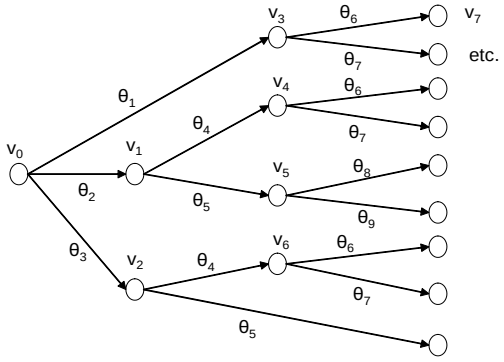


Figure 1: Tree for Example 1

Figure 1 shows an event tree $\mathcal{T}$ embellished with edge-probabilities. Edge event labels are not shown, but edges sharing a common probability label (eg. θ_4) correspond to the same event given a different history. The CEG $\mathcal{C}$ in Figure 2 is produced by combining the vertices $\{v_3, v_4, v_6\}$ into one position w_3, combining all leaf-nodes into a single sink-node w_∞, and relabelling vertices v_0, v_1, v_2, v_5 as w_0, w_1, w_2, w_4. The stages of the CEG are $u_0 = \{w_0\}$, $u_1 = \{w_1, w_2\}$, $u_2 = \{w_3\}$, $u_3 = \{w_4\}$. The edges leaving w_1 and w_2 are coloured as they lie in the same stage — their florets have identical topologies and corresponding edges are labelled with the same events and probabilities.

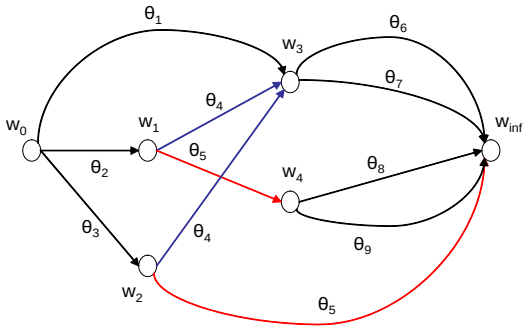


Figure 2: CEG for Example 1

Note that the CEG is specified through a particular event tree and statements about specific developments sharing the same distribution. Both of these properties can be expressed verbally in terms of a general explanation of the unfolding of events, and therefore have a meaning that transcends the particular instance.

3 Learning CEGs

In this paper we consider maximum a posteriori (MAP) model selection on the class of CEGs. Other methods exist for BNs and many of these extend to CEGs as straightforwardly as the extension described here.

As with BN-based modelling, if we have complete random sampling the likelihood for a CEG model separates into products of terms which are only a function of parameters associated with one component of the model. In the BN each term is associated with a variable and its parents; in the case of the CEG the model components are the stage florets. Furthermore, the term in the likelihood corresponding to a particular floret $F(u)$ is proportional to one obtained from multinomial sampling on the set of units arriving at u.

For each stage u we can label the edges in $F(u)$ by their probabilities under this model, so θ_{ui} labels the ith edge leaving any position which is a member of the stage u. We then let n_{ui} be the total number of sample units passing through an edge labelled θ_{ui}, and the likelihood for our CEG model is given by

$$L(\pi) = \prod_u \prod_i {\theta_{ui}}^{n_{ui}}$$

Assumptions of global and local independence together with the use of Dirichlet priors ensure conjugacy when learning BNs. To ensure the same with CEGs, we give the vectors of probabilities associated with the set of stage florets independent Dirichlet distributions. This gives prior and posterior distributions for the CEG model which are products of Dirichlet densities, and a marginal likelihood for $\mathcal{C}$ of

$$\prod_u \frac{\Gamma(\sum_i \alpha_{ui})}{\Gamma(\sum_i(\alpha_{ui}+n_{ui}))} \prod_i \frac{\Gamma(\alpha_{ui}+n_{ui})}{\Gamma(\alpha_{ui})} \qquad (1)$$

where α_{ui} are the exponents of our Dirichlet priors.

As $P(model \mid data) \propto P(data \mid model) \times P(model)$ we have to set prior probabilities for possible models as well as parameter priors. There are many choices for both these, but for accessibility in this paper we consider simple cases which have direct analogues in BN model selection. So, if there is no reason to do otherwise we let $P(model)$ be constant for all models in the candidate set of CEGs. Similarly we choose the case where hyperparameter priors are set to correspond to counts of dummy units through the CEG. We do this by putting a uniform prior over the root-to-sink paths of the CEG and assigning Dirichlet priors to each of the stage florets. It is straightforward to check (see for example (Freeman and Smith, 2009)) that for models expressible as both CEGs and BNs, the values given by expression (1) are then identical to those given by BN expression (2) using the prior settings suggested in (Cooper and Herskovits, 1992; Heckerman et al., 1995) etc.

$$\prod_{i \in V} \Big[\prod_j \frac{\Gamma(\sum_k \alpha_{ijk})}{\Gamma(\sum_k(\alpha_{ijk}+n_{ijk}))} \prod_k \frac{\Gamma(\alpha_{ijk}+n_{ijk})}{\Gamma(\alpha_{ijk})} \Big] \qquad (2)$$

Note that here i indexes the set of variables of the BN; k indexes the levels of the variable X_i;

and j indexes vectors of levels of the parental variables of X_i.

It is this result which allows us to use BN-based methods to narrow down the set of possible models before moving over to CEGs to use the techniques here presented to refine our search.

For the CEG in Figure 2, we put a uniform prior over the nine root-to-sink paths, and assign a $Di(2,4,3)$ prior to u_0, $Di(4,3)$ prior to $u_1 \equiv \{w_1, w_2\}$, $Di(3,3)$ prior to u_2, and $Di(1,1)$ prior to u_3. We then have $L(\pi)$ equal to

$$\begin{aligned}
&\frac{\Gamma(9)}{\Gamma(9+N)}\frac{\Gamma(2+n_{01})\Gamma(4+n_{02})\Gamma(3+n_{03})}{\Gamma(2)\Gamma(4)\Gamma(3)} \\
&\times\frac{\Gamma(7)}{\Gamma(7+n_{11}+n_{12})}\frac{\Gamma(4+n_{11})\Gamma(3+n_{12})}{\Gamma(4)\Gamma(3)} \\
&\times\frac{\Gamma(6)}{\Gamma(6+n_{21}+n_{22})}\frac{\Gamma(3+n_{21})\Gamma(3+n_{22})}{\Gamma(3)\Gamma(3)} \\
&\times\frac{\Gamma(2)}{\Gamma(2+n_{31}+n_{32})}\frac{\Gamma(1+n_{31})\Gamma(1+n_{32})}{\Gamma(1)\Gamma(1)}
\end{aligned}$$

where N is the sample size, and n_{11} (for example) is the total number of sample units leaving u_1 (ie. w_1 or w_2) via (in this case) a blue edge.

Note that, as in this example, CEGs can be used to depict models which admit known logical constraints. If we attempt to express the constraints of this example through a BN, we find that some variables have no outcomes given particular vectors of values of ancestral variables. We cannot simply set probabilities to zero in this instance as a Dirichlet distribution is then no longer appropriate and so the usual model selection procedure fails.

4 An Example

In this section we consider a simple example which demonstrates the versatility of our method. Our client is analyzing a medical data set relating to an inherited condition. A random sample of 100 (51 female, 49 male) people has been taken from a population who have had recent ancestors with the condition. For each individual in the sample a record has been kept of whether or not they displayed a particular symptom in their teens, and whether or not they then developed the condition in middle age. The data is given in Table 1, where $A = 0, 1$ corresponds to *female*, *male*; $B = 1$ corresponds to the individual displaying the symptom; and $C = 1$ corresponds to the individual developing the condition.

Table 1: Data for medical example

		A			
		0		1	
		B		B	
		0	1	0	1
C	0	33	6	10	12
	1	6	6	9	18

Eight possible BNs could be drawn for this problem, with directed edges present or absent between A & B, A & C, and B & C. These BNs represent eight possible models, which given the temporal ordering of the variables can be described by (a) full independence, (b) $A \to C$, $B \amalg (A, C)$, (c) $B \to C$, $A \amalg (B, C)$, (d) $A \to B$, $C \amalg (A, B)$, (e) $A \to C$, $B \to C$, $B \amalg A$, (f) $A \to B \to C$, $C \amalg A \mid B$, (g) $A \to B$, $A \to C$, $C \amalg B \mid A$, and (h) $A \to B \to C$, $A \to C$, full association. CEGs can also be drawn for these models, although as these are not asymmetric models, there is no advantage in doing so. For illustrative purposes the models (b), (d) and (f) are depicted as CEGs in Figure 3 (i), (ii) and (iii).

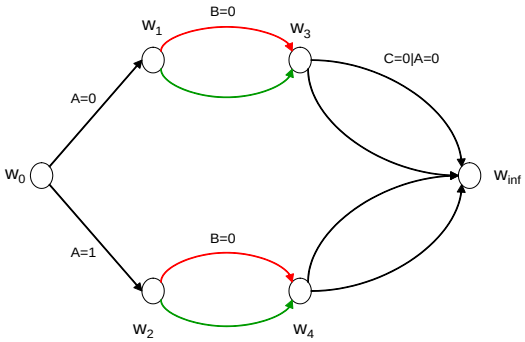


Figure 3 (i): $A \to C$, $B \amalg (A, C)$

The conditional independence properties of the models are easy to to read from the CEG. We can read, for example CEG (iii) as follows:

- as the edges leaving w_1 and w_2 are not coloured (ie. they carry different probabilities), these positions are not in the same stage, so $A \not\amalg B$,

- edges labelled $B = 0$ converge at w_3, so $C \amalg A \mid (B = 0)$. Similarly, edges labelled $B = 1$ converge at w_4, so $C \amalg A \mid (B = 1)$, and combining these we get $C \amalg A \mid B$.

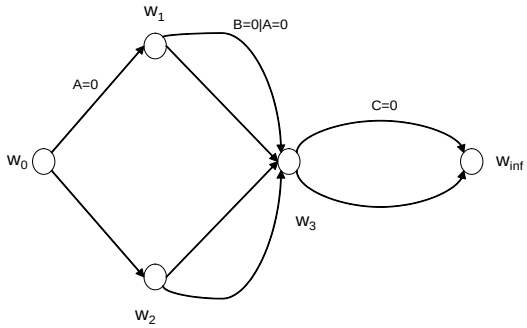
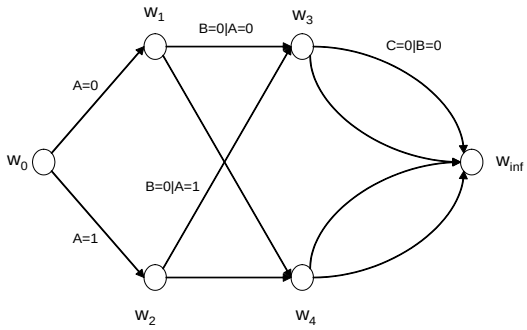
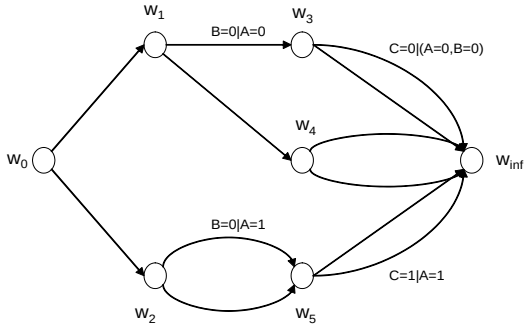


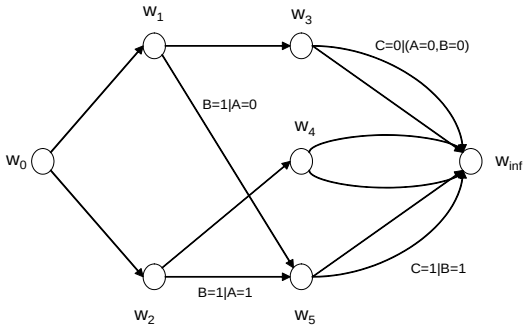
Figure 3 (ii): $A \to B$, $C \amalg (A, B)$



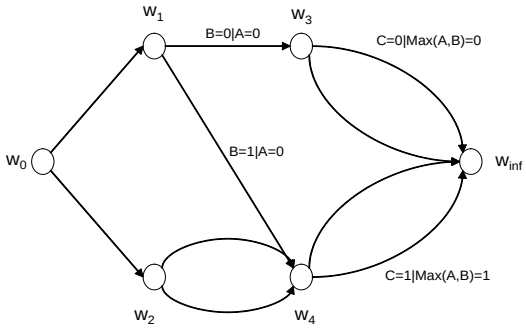
(iii): $A \to B \to C$, $C \amalg A \mid B$



(iv): $A \to B$, $C \amalg B \mid (A = 1)$



(v): $A \to B$, $C \amalg A \mid (B = 1)$



(vi): $A \to B$, $C \amalg (A, B) \mid Max(A, B)$

Our starting point is to search over the candidate set of eight BNs, and as our client has not expressed any preference for a particular model, we let $P(model)$ be constant for each model in the candidate set, which allows us to use $P(data \mid model)$ as a measure for $P(model \mid data)$. We then (as we are using MAP model selection) let the score of the model be the logarithm of its marginal likelihood (as expressed by (2)). Note that using CEGs and expression (1) would give us exactly the same scores as using BNs and (2). The scores for our eight models are given in Table 2. The model with the highest score is the MAP model for this candidate set.

Table 2

Model	Score	Model	Score
(a)	-208.44	(e)	-204.29
(b)	-204.80	(f)	-199.37
(c)	-205.02	(g)	-199.15
(d)	-202.79	(h)	-198.64

The lower scores for models (a), (b), (c) and (e) clearly indicate that B is directly dependent on A. Although model (h) has the highest score, the closeness of the scores for models (f) and (g), their proximity to the score for (h), and their distance from the score for (d) suggests that there is some context-specific conditional independence at work. Context-specific properties such as $C \amalg B \mid (A = 1)$ (there is one distribution for developing the condition given that gender is *male*) or $C \amalg A \mid (B = 1)$ (there is one distribution for developing the condition given that symptom was displayed) can be represented as context-specific BNs of the type described in, for example, (Boutilier et al., 1996; Poole and Zhang, 2003). They can also be represented elegantly as CEGs — these particular models are depicted in Figure 3 (iv) and (v) (which also reflect the established direct dependence of B on A). As we earlier read the CEG in Figure 3 (iii), we can read, for example CEG (v) as follows:

- w_1 and w_2 are not in the same stage, so $A \not\amalg B$,
- edges labelled $B = 1$ converge at a single position, so $C \amalg A \mid (B = 1)$, but edges labelled $B = 0$ do not, so we do not have $C \amalg A \mid (B = 0)$.

Note that the CEG portrays the context-specific conditional independence properties of the model in its topology — the context-specific BN does not. Also, although BN-based learning methods have been adapted for context-specific BNs (see for example (Feelders and van der Gaag, 2005)), our CEG-based methods work for all CEG models without the need for any adaptation.

Using CEGs to score models with context-specific properties of the sort described, we find that $C \amalg B \mid (A = 1)$ and $C \amalg A \mid (B = 1)$ are indeed improvements not just upon $C \amalg B \mid A$ and $C \amalg A \mid B$, but also upon *full association*, scoring -197.58 and -197.53 respectively. The closeness of these scores suggests that there may be a model with context-specific independence which cannot be expressed as simply as for these models, and which is better than both of them. In fact there are 30 possible CEG models for this problem, and this is without relaxing the edge ordering A, B, C. In fact the best model here is $C \amalg (A, B) \mid Max(A, B)$ (there is one distribution for developing the condition given that an individual is male OR displayed the symptom, and one distribution for developing the condition given that an individual is female AND did not display the symptom). This is shown as a CEG in Figure 3 (vi). It is not representable as a BN without transformation of the variables.

5 Discussion

In this paper we have concentrated on the principle of assigning a score to a member of a candidate class, rather than on algorithms for searching over this class. But as the example in the previous section demonstrates, not only is it very easy to establish the full candidate set of CEGs, it will also be straightforward to move between the members of this set when learning. The score for a CEG model decomposes into components associated with florets. When two CEGs contain the same floret, we assign this floret the same prior distribution in each model, and the separation of the likelihood means that this property is retained in the posterior distribution. As *similar* models will share a high proportion of florets, the scores for *similar* models will differ only in a small number of components. Efficient algorithms can therefore be created to search over the CEG model space (Freeman and Smith, 2009).

Various methods have been developed to restrict the search in BN model selection to subsets of the class of models (see for example (van Gerven and Lucas, 2004)). As what we are proposing is to use CEG model selection as a refining process, we can still utilise these methods before moving on to the class of CEGs. Also there are ways in which we can further restrict the search to explore subclasses of CEGs which are expected to provide good explanations of the data.

Because each model in the class of CEGs

is **qualitatively** expressed in any given context, the task of restricting the set of candidate CEGs is much easier than it might first appear. Thus for example, in the educational examples considered in (Freeman and Smith, 2009), the context demands that the underlying event tree is consistent with the order students study courses, and that certain vertices could never reasonably be combined into the same stage. These sorts of contextually defined constraints can readily be incorporated into customized search algorithms, and the efficiency of the search procedure improved. It is also not unusual for more quantitative information to be available, such as one type of stage combination being proportionately more probable than another. This can allow one to usefully further refine and improve the search, although then the framework the CEG provides is no longer totally qualitative.

We noted earlier that there is a wide choice of possible parameter priors available, and that we had chosen a particularly straightforward set with a direct analogue in BN model selection. Care does however need to be taken when choosing parameter priors if the model selection algorithm is to function efficiently. This issue has already been addressed by a number of authors for the case of BNs (see for example (Heckerman, 1998)) using concepts of distribution and independence equivalence, and parameter modularity to ensure plausibly consistent priors over this class. For a full Bayesian estimation with conjugate locally and globally independent priors, the class of BNs nests within the larger class of CEGs. If we require that all BNs within the subclass of CEGs we are studying continue to respect these independence rules, whilst also retaining our floret independence, then the choices of prior hyperparameters are limited analogously with the class of BNs. Using a result from (Geiger and Heckerman, 1997), it is shown in (Freeman and Smith, 2009) that for a significant class of CEGs, if we assign Markov equivalent models the same prior, then the joint distribution on the leaves of the underlying tree is necessarily a priori Dirichlet. Modularity conditions then result in floret distributions being Dirichlet and mutually independent.

In (Silander et al., 2007) it was demonstrated that MAP model selection on the class of BNs can be sensitive to how priors are set, even when these priors are conjugate product Dirichlets. Extending this idea to CEG model selection, it may be insufficient simply to state that we are setting a uniform Dirichlet prior on the root-to-sink paths; we may also need to exercise care in the choice of a scale parameter for this distribution. This requires an **explicit** evaluation of the overall strength of prior beliefs, which can then be specified via the *equivalent size* (count of dummy units) assigned in the prior to each root-to-leaf path of the underlying tree. As already noted, there are Bayesian model selection methods other than MAP which extend to CEGs. If the analyst does not feel sufficiently confident in making this evaluation, then for example using the Bayesian Information Criterion (BIC) could easily be modified for use with the set of CEG models.

Of course, just as with BNs, the conjugacy does not necessarily continue to hold when sampling is not complete. In this case approximate or numerical search algorithms need to be employed with consequent loss of accuracy or speed in scoring and comparing models. However in this case the methods for estimating BNs with missing values (see for example (Riggelsen, 2004)) can usually be extended so that they also apply to CEGs.

CEGs allow for the representation and analysis of problems whose state spaces are asymmetric and do not admit a natural product structure. In this paper we have shown that there are natural methods for learning CEGs which are closely related to the methods for learning BNs, and that we can use these CEG-based methods for further refining BN-based model selection.

Acknowledgments

This research has been funded by the UK Engineering and Physical Sciences Research Council as part of the project *Chain Event*

Graphs: Semantics and Inference (grant no. EP/F036752/1).

References

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

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

A. Feelders and L. van der Gaag. 2005. Learning Bayesian Network parameters with prior knowledge about context-specific qualitative influences. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence*, Arlington, Virginia.

G. Freeman and J. Q. Smith. 2009. Bayesian map model selection of Chain Event Graphs. Research Report 09–06, CRiSM.

D. Geiger and D. Heckerman. 1997. A characterization of the Dirichlet distribution through Global and Local independence. *Annals of Statistics*, 25:1344–1369.

D. Heckerman, D. Geiger, and D. Chickering. 1995. Learning Bayesian Networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243.

D. Heckerman. 1998. A tutorial on Learning with Bayesian Networks. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 301–354. MIT Press.

D. Poole and N. L. Zhang. 2003. Exploiting contextual independence in probabilistic inference. *Journal of Artificial Intelligence Research*, 18:263–313.

C. Riggelsen. 2004. Learning Bayesian Network parameters from incomplete data using importance sampling. In *Proceedings of the 2nd European Workshop on Probabilistic Graphical Models*, pages 169–176, Leiden.

G. Shafer. 1996. *The Art of Causal Conjecture*. MIT Press.

T. Silander, P. Kontkanen, and P. Myllymaki. 2007. On the sensitivity of the MAP Bayesian Network structure to the equivalent sample size parameter. In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence*, Vancouver.

J. Q. Smith and P. E. Anderson. 2008. Conditional independence and Chain Event Graphs. *Artificial Intelligence*, 172:42–68.

P. A. Thwaites, J. Q. Smith, and R. G. Cowell. 2008. Propagation using Chain Event Graphs. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, pages 546–553, Helsinki.

P. A. Thwaites, J. Q. Smith, and E. M. Riccomagno. 2010. Causal analysis with Chain Event Graphs. Accepted by *Artificial Intelligence*.

M. A. J. van Gerven and P. J. F. Lucas. 2004. Using background knowledge to construct Bayesian classifiers for data-poor domains. In *Proceedings of the 2nd European Workshop on Probabilistic Graphical Models*, Leiden.