Chain Event Graph MAP model selection

Peter A. Thwaites, **Guy Freeman** and **Jim Q. Smith**
Department of Statistics
University of Warwick
Coventry UK CV4 7AL

Abstract

When looking for general structure from a finite discrete data set it is quite common to search over the class of Bayesian Networks (BNs). The class of Chain Event Graph (CEG) models is however much more expressive and is particularly suited to depicting hypotheses about how situations might unfold. The CEG retains many of the desirable qualities of the BN. In particular it admits conjugate learning on its conditional probability parameters using product Dirichlet priors. The Bayes Factors associated with different CEG models can therefore be calculated in an explicit closed form, which means that search for the maximum a posteriori (MAP) model in this class can be enacted by evaluating the score function of successive models and optimizing. As with BNs, by choosing an appropriate prior over the model space, the conjugacy property ensures that this score function is linear in the different components of the CEG model. Local search algorithms can therefore be devised which unveil the rich class of candidate explanatory models, and allow us to select the most appropriate. In this paper we concentrate on this discovery process and upon the scoring of models within this class.

1 INTRODUCTION

The Chain Event Graph (CEG), introduced in Smith & Anderson (2008), Thwaites, Smith & Cowell (2008) and Smith, Riccomagno & Thwaites (2009), is a graphical model specifically designed to embody the conditional independence structure of problems whose state spaces are highly asymmetric and do not admit a natural product structure. There are many scenarios in medicine, biology and education where such asymmetries arise naturally (for examples see Smith & Anderson (2008)), and where the main features of the model class cannot be fully captured by a single BN or even a context specific BN. A key property of the CEG framework is that these graphical models are *qualitative* in their topologies – they encode sets of conditional independence statements about how things might happen, without prespecifying the probabilities associated with these events. Each CEG model can therefore be identified with a unique explanation of how situations might unfold.

For a detailed formal description and motivation for using a CEG model and an outline of some of its implicit conditional independence structure see Smith & Anderson (2008). In this paper it was shown that the CEG is a more expressive graphical model than the BN in that any asymmetries are represented explicitly in the topology of the CEG, and in that CEGs can be used to express a much richer set of conditional independence statements not simultaneously expressible through a single BN. It was also shown that the class of BNs is contained within that of CEGs. This is a property which we exploit later, since with appropriate prior settings, it follows that BN model selection procedures can be nested within those for CEGs.

The CEG is an event-based (rather than variable-based) graphical model, and is a function of an event tree. Any problem on a finite discrete data set can be modelled using an event tree, but they are particularly suited to problems with asymmetric state spaces. Unfortunately, it is almost impossible to read the conditional independence properties of a model from an event tree representation, as only trivial independencies are expressed within its topology. The CEG elegantly solves this problem, encoding a rich class of conditional independence statements through its edge and vertex structure.

So consider an event tree T with vertex set $V(T)$, di-

CRiSM Paper No. 09-07, www.warwick.ac.uk/go/crism

rected edge set $E(T)$, and $S(T) \subset V(T)$, the set of the tree's non-leaf vertices or *situations* (Shafer (1996)). A probability tree can then be specified by a transition matrix on $V(T)$, where absorbing states correspond to leaf-vertices. Transition probabilities are zero except for transitions to a situation's children (see Table 1).

Table 1: Part of the transition matrix for Example 1

	v_1	v_2	v_3	v_4	v_5	v_6	...	v_∞^1	v_∞^2	...
v_0	θ_1	θ_2	θ_3	0	0	0	...	0	0	...
v_1	0	0	0	θ_5	0	0	...	θ_4	0	...
v_2	0	0	0	0	θ_4	θ_5	...	0	0	...
⋮	⋮					⋮		⋮	⋮	

Let $T(v)$ be the subtree rooted in the situation v which contains all vertices after v in T. We say that v_1 and v_2 are in the same *position* if:

- the trees $T(v_1)$ and $T(v_2)$ are topologically identical,
- there is a map between $T(v_1)$ and $T(v_2)$ such that the edges in $T(v_2)$ are labelled, under this map, by the same probabilities as the corresponding edges in $T(v_1)$.

The set $W(T)$ of positions w partitions $S(T)$. The *transporter* CEG (Thwaites, Smith & Cowell 2008) is a directed graph with vertices $W(T) \cup \{w_\infty\}$, with an an edge e from w_1 to $w_2 \neq w_\infty$ for each situation $v_2 \in w_2$ which is a child of a fixed representative $v_1 \in w_1$ for some $v_1 \in S(T)$, and an edge from w_1 to w_∞ for each leaf-node $v \in V(T)$ which is a child of some fixed representative $v_1 \in w_1$ for some $v_1 \in S(T)$.

For the position w in our transporter CEG, we define the *floret* $F(w)$ to be w together with the set of outgoing edges from w. We say that w_1 and w_2 are in the same *stage* if:

- the florets $F(w_1)$ and $F(w_2)$ are topologically identical,
- there is a map between $F(w_1)$ and $F(w_2)$ such that the edges in $F(w_2)$ are labelled, under this map, by the same probabilities as the corresponding edges in $F(w_1)$.

The CEG $C(T)$ is then a mixed graph with vertex set $W(C)$ equal to the vertex set of the transporter CEG, directed edge set $E_d(C)$ equal to the edge set of the transporter CEG, and undirected edge set $E_u(C)$ consisting of edges which connect the component positions of each stage $u \in U(C)$, the set of stages. The CEG-construction process is illustrated in Example 1, and an example CEG in Figure 2.

Example 1

Consider the tree in Figure 1 which has 11 atoms (root-to-leaf paths). Symmetries in the tree allow us to store the distribution in 5 conditional tables which contain 11 (6 free) probabilities. The transporter CEG is produced by combining the vertices $\{v_4, v_5, v_7\}$ into one position w_4, the vertices $\{v_6, v_8\}$ into one position w_5, and all leaf-nodes into a single sink-node w_∞. The CEG C (Figure 2) has an undirected edge connecting the positions w_1 and w_2 as these lie in the same stage – their florets are topologically identical, and the edges of these florets carry the same probabilities.

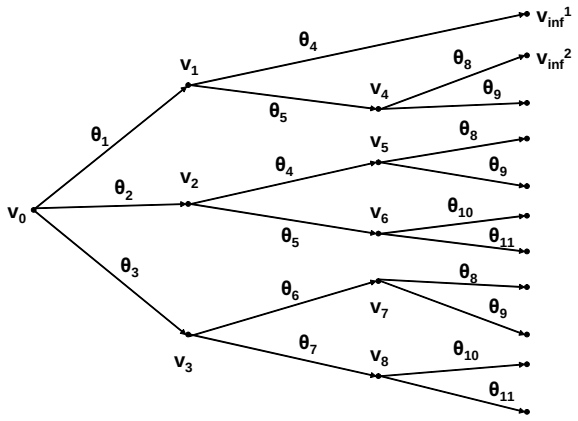

Figure 1: Tree for Example 1

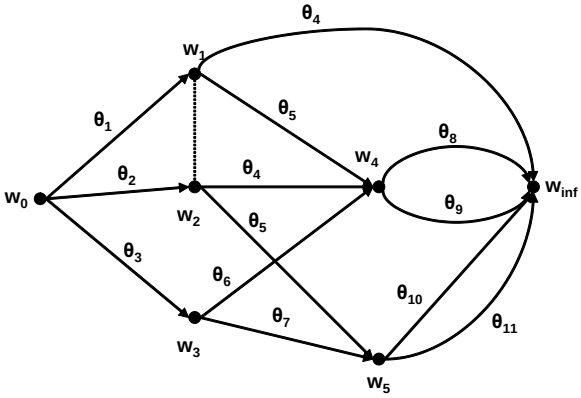

Figure 2: CEG for Example 1

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.

The analogue in the CEG of the clique in a BN is the floret. Fast propagation algorithms for a simple CEG were developed in Thwaites, Smith & Cowell (2008). These exploited the graph's embedded conditional independencies to factorize its mass function over local masses on florets. 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 that 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.

As the CEG can express a richer class of conditional independence structures than the BN, CEG model selection allows for the automatic identification of more subtle features of the data generating process than it would be possible to express (and therefore to evaluate) through the class of BNs. Simple examples of the types of structure that might exist and could be discovered are given below.

Section 2 introduces the techniques for learning CEGs and compares these with those for learning BNs. Section 3 consists of an example illustrating the advantages of searching over the extended candidate set available when learning CEGs, and section 4 contains further discussion of the theory.

2 LEARNING CEGs

The reason the CEG shares the conjugacy properties of the BN is that with complete random sampling the likelihood 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 component is the floret. Furthermore, the term in the likelihood corresponding to a particular floret is proportional to one obtained from multinomial sampling on the set of units arriving at the root of the floret.

From our CEG definition, if $w_1, w_2 \in u$ for some u, then the corresponding edges in the florets $F(w_1)$ and $F(w_2)$ carry the same probabilities. So, for each member u of the set of stages prescribed by the model under consideration for our CEG, we can label the edges leaving u by their probabilities under this model. We can then let x_{un} be the **total** number of sample units passing through an edge labelled π_{un}; and the likelihood $L(\boldsymbol{\pi})$ for our CEG model is given by

$$L(\boldsymbol{\pi}) = \prod_u \prod_n {\pi_{un}}^{x_{un}}$$

For BNs, the assumptions of local and global independence, and the use of Dirichlet priors ensures conjugacy. The analogue for CEGs is to give the vectors of probabilities associated with the stages independent Dirichlet distributions. Then the structure of the likelihood $L(\boldsymbol{\pi})$ results in prior and posterior distributions for the CEG model which are products of Dirichlet densities. The result of this conjugacy is that the marginal likelihood of each CEG is therefore the product of the marginal likelihoods of its component florets. Explicitly, the marginal likelihood of a CEG C is

$$\prod_u \frac{\Gamma(\sum_n \alpha_{un})}{\Gamma(\sum_n (\alpha_{un} + x_{un}))} \prod_n \frac{\Gamma(\alpha_{un} + x_{un})}{\Gamma(\alpha_{un})}$$

where, as above

- u indexes the stages of C
- n indexes the outgoing edges of each stage
- α_{un} are the exponents of our Dirichlet priors
- x_{un} are the data counts

As we are actually interested in $p(model \mid data)$, and this is proportional to $p(data \mid model) \times p(model)$, we need to set both parameter priors and prior probabilities for the possible models.

Care needs to be taken when choosing these parameters if the model selection algorithm is to function efficiently. We return to this issue in section 4, but note that many aspects have already been addressed by a number of authors for the special 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 (quite reasonably) 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. For example, if we search over the class of all CEGs whose underlying trees have a non prime number of leaves, then using a result from Geiger & Heckerman (1997), it can be shown that if we assign Markov

equivalent models the same prior, then the joint distribution on the leaves is necessarily a priori Dirichlet (see Freeman & Smith (2009)). Modularity conditions then result in floret distributions being Dirichlet and mutually independent.

Exactly analogously with BNs, parameter modularity in CEGs implies that whenever CEG models share some aspect of their topology, we assign this aspect the same prior distribution in each model. When such priors reflect our beliefs in a given context, this can reduce our problem dramatically to one of simply expressing prior beliefs about the possible floret distributions (ie. the local differences in model structure). As each CEG model is essentially a partition of the vertices in the underlying tree into sets of stages, this requirement ensures that when two partitions differ only in whether or not some subset of vertices belong to the same stage, the prior expressions for the models differ only in the term relating to this stage. The separation of the likelihood means that this local difference property is retained in the posterior distribution.

Now, our candidate set is much richer than the corresponding candidate BN set, and will probably contain models we have not previously considered in our analysis. Again, evoking modularity, if we have no information to suggest otherwise, we follow standard BN practice and let $p(model)$ be constant for all models in the class of CEGs. We now use the logarithm of the marginal likelihood of a CEG model as its score, and maximise this score over our set of candidate models to find the MAP model.

Our expression has the nice property that the difference in score between two models which are identical except for a particular subset of florets, is a function of the subscores only of the probability tables on the florets where they differ. Various fast deterministic and stochastic algorithms can therefore be derived to search over the model space, even when this is large – see Freeman & Smith (2009) for examples of these in the particular case where the underlying event tree is fixed. This property is of course shared by the class of BNs.

We set the priors of the hyperparameters so that they correspond to counts of dummy units through the graph. This can be done by setting a Dirichlet distribution on the root-to-sink paths, and for simplicity we choose a uniform distribution for this. It is then easy to check (see Freeman & Smith (2009)) that in the special case where the CEG is expressible as a BN, the CEG score above is equal to the standard score for a BN using the usual prior settings as recommended in, for example, Cooper & Herskovits (1992) and Heckerman, Geiger & Chickering (1995). As a comparison with our CEG-expression; given Dirichlet priors and a multivariate likelihood, the marginal likelihood on a BN is expressible as

$$\prod_{i\in V}\left[\prod_{m}\frac{\Gamma(\sum_n \alpha_{imn})}{\Gamma(\sum_n(\alpha_{imn}+x_{imn}))}\prod_{n}\frac{\Gamma(\alpha_{imn}+x_{imn})}{\Gamma(\alpha_{imn})}\right]$$

where

- i indexes the set of variables of the BN
- n indexes the levels of the variable X_i
- m indexes vectors of levels of the parental variables of X_i

The importance of this result is that were we first to search the space of BNs for the MAP model, then we could seamlessly refine this model using the CEG search score described above. Such embellishments will allow us to search over models containing context specific information or Noisy AND/OR gates. Furthermore any model we find will have an associated interpretation which can be stated in common language, and can be discussed and critiqued by our client/expert for its phenomenological plausibility.

For the CEG in Figure 2, we put a uniform prior over the 11 root-to-leaf paths, which in turn allows us to assign our stage priors as follows: we assign a $Di(3,4,4)$ prior to the stage identified by w_0, a $Di(3,4)$ prior to the stage $u_1 \equiv (w_1, w_2)$, a $Di(2,2)$ prior to each of the stages identified by w_3 and w_5, and a $Di(3,3)$ prior to the stage identified by w_4. We would then have a marginal likelihood of

$$\begin{aligned}
&\frac{\Gamma(11)}{\Gamma(11+N)}\frac{\Gamma(3+x_{01})\Gamma(4+x_{02})\Gamma(4+x_{03})}{\Gamma(3)\Gamma(4)\Gamma(4)}\\
\times\ &\frac{\Gamma(7)}{\Gamma(7+x_{01}+x_{02})}\frac{\Gamma(3+x_{14}+x_{24})\Gamma(4+x_{15}+x_{25})}{\Gamma(3)\Gamma(4)}\\
\times\ &\frac{\Gamma(4)}{\Gamma(4+x_{03})}\frac{\Gamma(2+x_{36})\Gamma(2+x_{37})}{\Gamma(2)\Gamma(2)}\\
\times\ &\frac{\Gamma(6)}{\Gamma(6+x_{15}+x_{24}+x_{36})}\frac{\Gamma(3+x_{48})\Gamma(3+x_{49})}{\Gamma(3)\Gamma(3)}\\
\times\ &\frac{\Gamma(4)}{\Gamma(4+x_{25}+x_{37})}\frac{\Gamma(2+x_{5\cdot 10})\Gamma(2+x_{5\cdot 11})}{\Gamma(2)\Gamma(2)}
\end{aligned}$$

where, with a slight abuse of notation, we let for example x_{24} be the data value associated with the edge leaving w_2 labelled θ_4; and where N is the sample size $= \sum_{n=1}^{3} x_{0n}$.

Note that, as in this example, CEGs can be used to depict models which admit known logical constraints.

If we attempt to express this particular constraint 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 procedures fails. Furthermore, this is one type of scenario which cannot be modelled adequately using the standard classes of context-specific BNs. By comparison, since such models exist within the class of CEG models, they can of course be revealed (and if appropriate, selected) by CEG-based conjugate search algorithms.

3 A SIMPLE SIMULATED MODEL

In this section we consider a simple example which demonstrates the versatility of our method. Our client is analysing 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 2, 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. Our client does not know whether displaying the symptom is independent of gender, but having looked at the data, believes that it is not.

Table 2: Data for example ($N = 100$)

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

Using his medical knowledge, our client has decided that the model lies in a candidate class of six, but is unwilling to express any preference for a particular model within this set.

In each of these six models B is not independent of A. The further conditional independence structure of the models is given by (i) $C \amalg (A, B)$, (ii) $C \amalg A \mid B$, (iii) $C \amalg B \mid A$, (iv) $C \amalg B \mid (A = 1)$ (there is one distribution for developing the condition given that gender is male), (v) $C \amalg A \mid (B = 1)$ (there is one distribution for developing the condition given that symptom was displayed), (vi) $C \amalg (A, B) \mid MAX(A, B)$ (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 – a Noisy OR gate).

The models are depicted in Figure 3. Only the first three of these models can be represented as BNs, with the fourth and fifth as context-specific BNs of the type described in, for example, Boutilier et al (1996) or Poole & Zhang (2003). The sixth would need us to create new variables in order for us to represent it as a BN – another example would be $C \amalg (A, B) \mid |A - B|$, which has a CEG similar to that of (ii), but with the edges leaving w_2 swapped so that $B = 1 \mid A = 1$ is the edge from w_2 to w_3, and $B = 0 \mid A = 1$ is the edge from w_2 to w_4.

We can read, for example CEG (ii) as follows:

- w_1 and w_2 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 $C \amalg A \mid B$.

In CEG (v) by contrast:

- 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 $C \amalg A \mid (B = 0)$.

The CEG portrays the context-specific conditional independence properties of the model in its topology – the context-specific BN does not.

Note that our client's candidate set is a restriction of the set of possible models – he has for instance dismissed models which encode statements such as $C \amalg B \mid (A = 0)$ or $C \amalg A \mid (B = 0)$ and all models where $A \amalg B$. In fact there are 15 possible models in the full candidate set if we require A to be a parent of B and B to be a temporal predecessor of C, and 30 if we relax the parental condition, but require that A is a temporal predecessor of B is a temporal predecessor of C. Note that there are only 4 possible BNs where A is a parent of B and B is a temporal predecessor of C, and 8 possible BNs where A is a temporal predecessor of B is a temporal predecessor of C. By using CEGs we can quickly have a clear idea of the full range of candidate models, and also our learning method works for all models in this range, including models such as $C \amalg (A, B) \mid MAX(A, B)$ or $C \amalg (A, B) \mid |A - B|$.

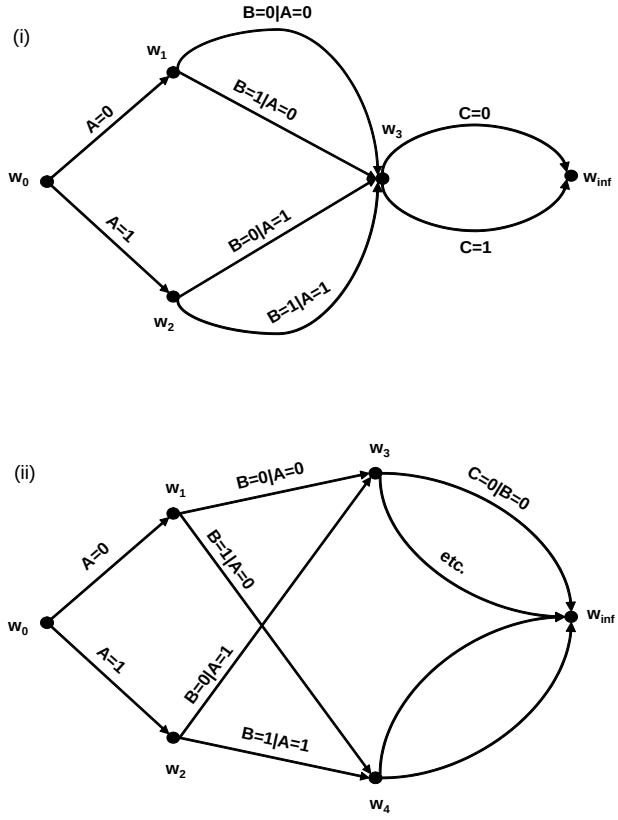

Figure 3: CEGs in our candidate set

As our client has no particular preference for any of the six models, it makes sense to let $p(model)$ be a constant value for all models in the candidate set, as suggested in section 2. This allows us to use $p(data \mid model)$ as a measure for $p(model \mid data)$, and we can then let the score of a model be its log marginal likelihood.

The three models expressible as BNs could of course be scored using the expression for BNs given above, and this would give us the same answer as our method using CEGs. But note that the BN-expressions for these models are more complex and less transparent than our CEG-expressions. We could perhaps use a learning method specifically adapted for context-specific BNs to score the fourth and fifth models (see for example Feelders & van der Gaag (2005)), but it is not evident how we would score the sixth model (consistently with the scoring of the other models) using a BN-based approach.

The score for model (i) decomposes into four components associated with the florets at w_0, w_1, w_2 and w_3. The components associated with the florets at w_0, w_1 and w_2 are retained in the remaining five models, so

the scores of the six models differ only in the components associated with the florets at $\{w_i\}_{i\geq 3}$. Scoring our 6 models we obtain -202.79, -199.37, -199.15, -197.58, -197.53 and -196.45. We can see that model (i) is the least appropriate, indicating that $C \not\perp\!\!\!\perp (A, B)$ and that there must be some sort of dependency of C on A and/or B. Models (iv), (v) and (vi) score better than models (ii) and (iii), indicating that this dependency is at best context-specific, and that the most appropriate model is not going to be expressible as a BN. In fact the best model in the candidate set is the Noisy OR gate, a model which could not be selected by a standard BN-based learning algorithm.

Looking at the CEGs in Figure 3, we can see that models (iv) and (vi) can be arrived at by making one alteration to model (iii), and that models (v) and (vi) can be arrived at by making one alteration to model (ii). It is easy to see how efficient algorithms could be created to search over the model space in this example.

Returning to the premise of our example, we share these results with our client, who then wants us to check whether a Noisy OR gate with $A \amalg B$ might score better than CEG model (vi). This model is depicted in Figure 4. The additional information in this CEG can be read as follows:

- there is an undirected edge connecting w_1 and w_2, so these two positions are in the same stage. Now positions in the same stage have their edges labelled identically, so the edges leaving w_1 and w_2 have labels that do not depend on the value of A. Consequently $A \amalg B$.

The score for this new model is -202.09, indicating that this model is not as good as model (vi). This is unsurprising given that the data in Table 1 suggests strongly that A is not independent of B.

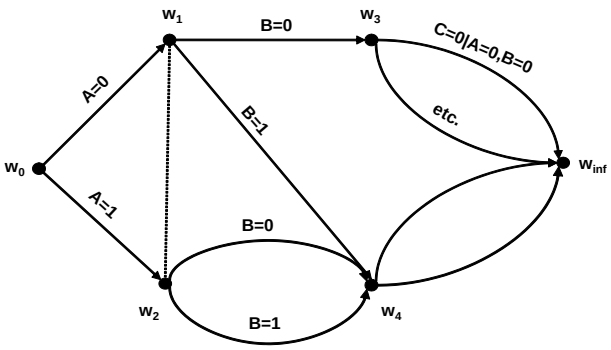

Figure 4: CEG for new $A \amalg B$ model

4 DISCUSSION

Clearly, searching over the class of CEGs is directly analogous to searching over the class of BNs, but the class of CEG models is much more expressive. This richness has an associated disadvantage – the class of all BNs is already difficult to search in large problems, and various methods have been developed to restrict the search to subsets of the class (see for example van Gerven & Lucas (2004), where the class of BNs that have edge-configurations consistent with a given spanning tree are searched). The number of possible CEGs available for even a small number of vertices is extremely large. Therefore, in even moderately sized problems it is usually efficacious to first restrict the model class to something smaller.

Because each model in this class is **qualitatively** expressed in any given context, this task 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. Thus, although more effort is needed to set up customized search spaces for CEGs than for BNs, we have found that the subsequent direct interpretability of any MAP model more than compensates for this effort.

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.

Silander et al (2007) have 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. If an analyst does not feel sufficiently confident in making this choice, we note that other Bayesian model selection methods (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. We will report on our findings on this topic in a later paper.

Lastly, it might be argued that context-specific BNs can be used to portray any set of conditional independence properties of a model, and that it would be a better use of resources developing improved learning methods for these graphs. In fact, as noted in section 2, there are significant sets of scenarios which cannot easily be modelled with context-specific BNs, which can none-the-less be modelled with CEGs. More importantly perhaps, an analyst modelling with BNs and their variants may not be aware just how many different models are available as possible explanations of the underlying data generating process of their data set. This is not a problem encountered by the analyst modelling with CEGs.

Acknowledgements

This research has been partly 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

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

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

[3] A. Feelders and L. van der Gaag. 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, 2005.

[4] G. Freeman and J. Q. Smith. Bayesian model selection of Chain Event Graphs. Research Report, CRiSM, 2009.

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

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

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

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

[9] C. Riggelsen. 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, 2004.

[10] G. Shafer. *The Art of Causal Conjecture*. MIT Press, 1996.

[11] T. Silander, P. Kontkanen, and P. Myllymaki. 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, 2007.

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

[13] J. Q. Smith, E. M. Riccomagno, and P. A. Thwaites. Causal analysis with Chain Event Graphs. Submitted to *Artificial Intelligence*, 2009.

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

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