Eliciting, Learning and Estimating Chain Event Graphs

Jim Smith (with Peter Thwaites, Guy Freeman, Bob Cowell and Eva Riccomagno) - with support from EPSRC

University of Warwick

April 2009

Advantages of Discrete Bayesian Networks

Bayesian networks are useful:

- for representing many models elegantly, expressively and formally.
- to graphically query their selected independence implications.
- as a framework for embellishing to a full probability model.
- as an interface between that probability model and its algebraic specification.
- for guiding fast propagation algorithms, conjugate learning.and model selection
- for extending models into causal structures.

Limitations of Bayesian Networks

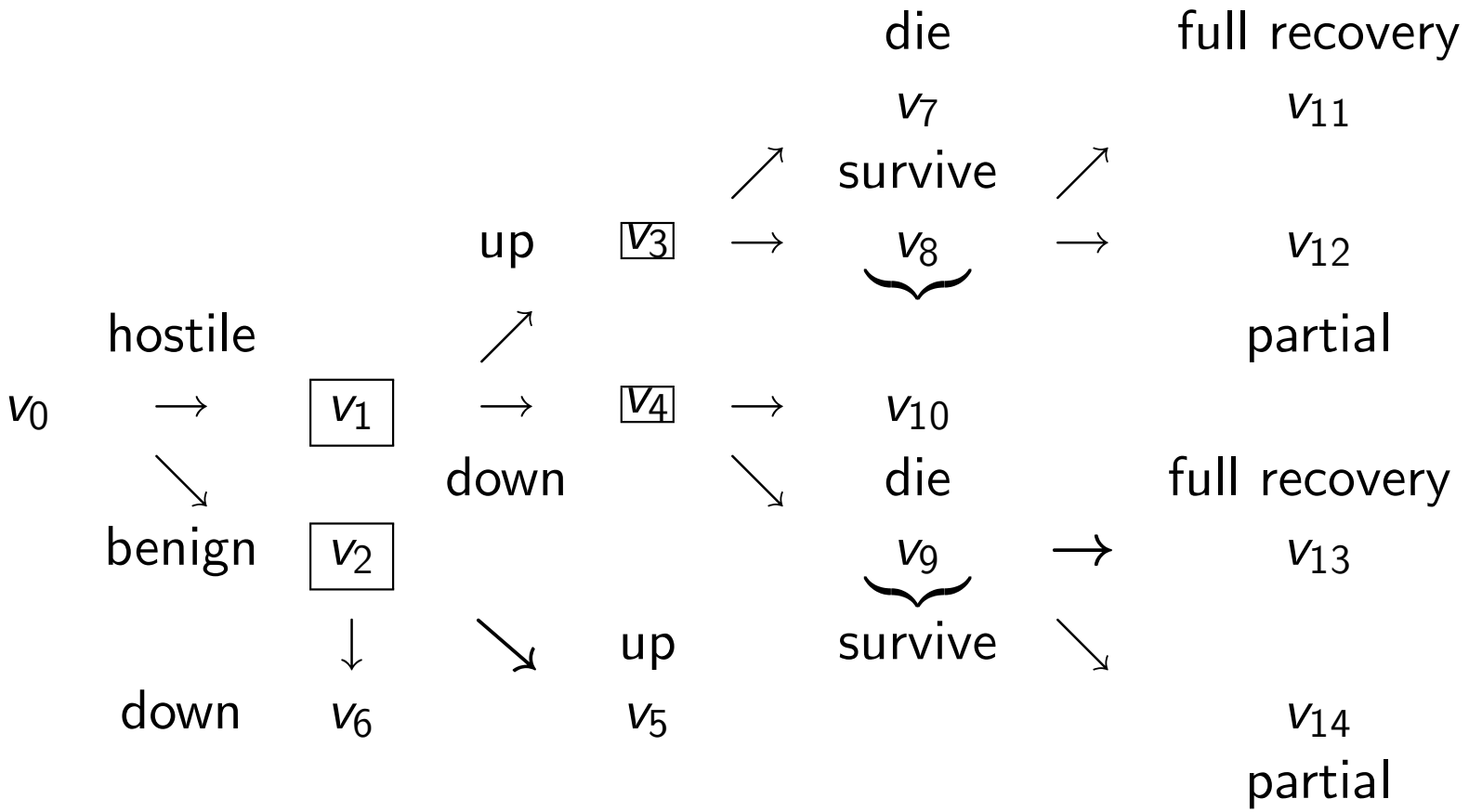
However!

- models specified using dependence relations between a preferred set of measurement variables.
- BN's not entirely natural for representing models explaining how things might happen.
- BN's do not represent the sample space in any way: often critical to estimation and selection issues.
- can only express certain types of probabilistic symmetry.
- the extensions to causal models are fragile and rather restrictive.

Advantages of an Event Tree

- The most natural expression of a model describing how things happen.
- Does not need a preferred set of measurement variables a priori.
- Explicitly represents the event space of a model, e.g. levels of variables.
- Asymmetries of the model space explicitly represented.
- Framework for probabilistic embellishment,estimation, selection and algebraic descriptions.
- Causal hypotheses much more richly expressed than in their BN analogues.

Example of an Event Tree



Limitations of Events Trees

However!

- Event trees of even moderately sized models are big!
- No inherent, topological expression of conditional independences.
- No non-trivial interrogation algorithms about dependence available.
- Do not provide a particularly efficient framework for propagation and learning.

Chain Event Graphs

- Typically topologically much simpler than event trees but still describe how things happen.
- Their paths represent fully the structure of the sample space.
- Expresses rich variety of dependence structures to be graphically queried.
- Embellish to a probability model and its associated algebraic rep.
- Like BNs provides a framework for fast propagation and conjugate learning.
- Almost as expressive of causal hypotheses as the event tree.

Constructing a CEG

Event tree $\longrightarrow$ Staged tree $\longrightarrow$ CEG [by positions and stages]

- Start with an event tree
- Convert it into a staged tree
- Then transform into a chain event graph by pooling positions and stages together

Example of a CEG

- Elicit *stages*: i.e. partition of situations with the same associated distribution

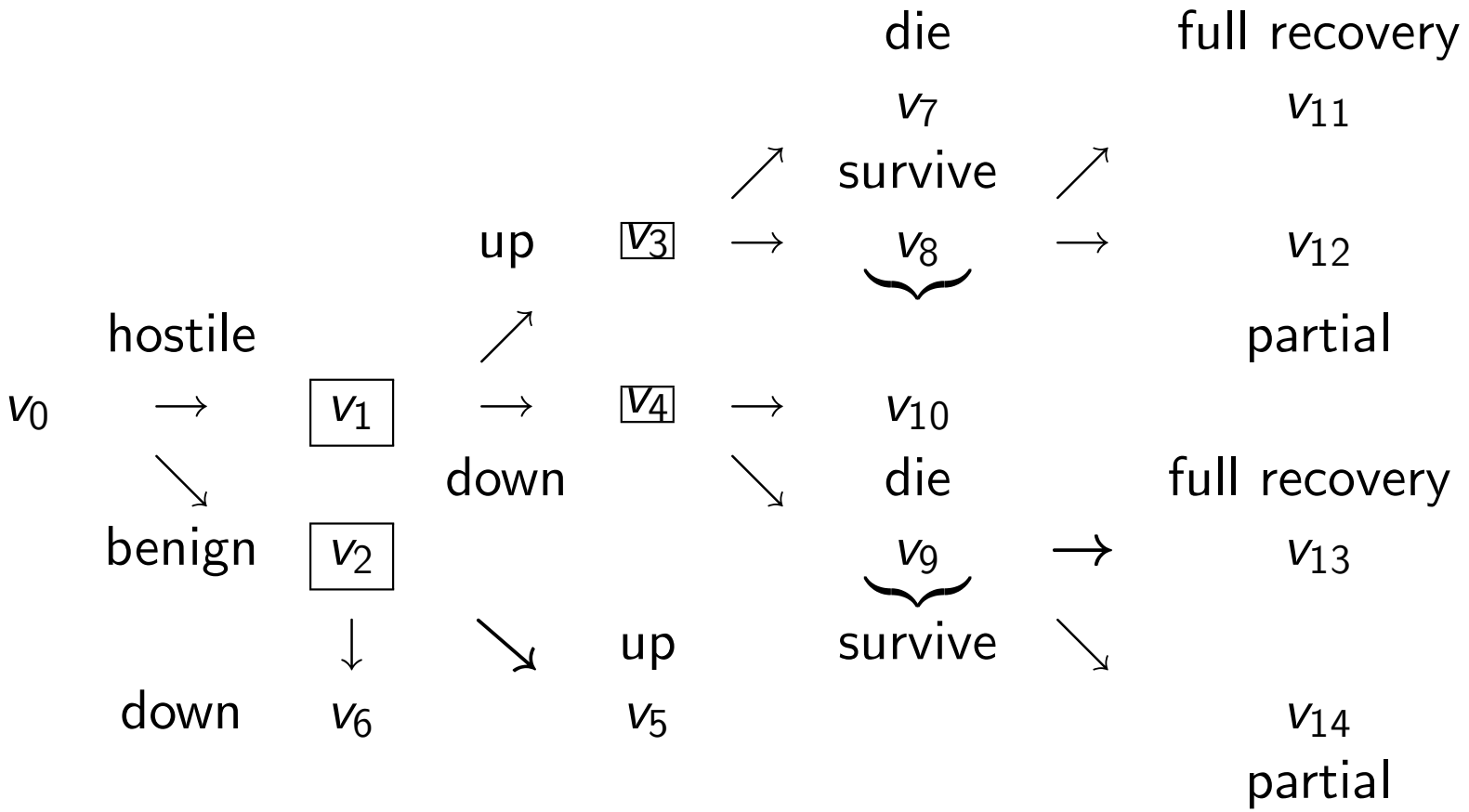
$$\begin{aligned} u_0 &= \{v_0\}, u_1 = \{v_1, v_2\}, u_2 = \{v_3, v_4\}, \\ u_3 &= \{v_8, v_9\}, u_\infty = \{leaves\} \end{aligned}$$

- Deduce *positions*: i.e. partition of situations with subsequent isomorphic trees

$$\begin{aligned} w_0 &= \{v_0\}, w_1 = \{v_1\}, w_2 = \{v_2\}, w_3 = \{v_3, v_4\}, \\ w_4 &= \{v_8, v_9\}, w_\infty = \{leaves\} \end{aligned}$$

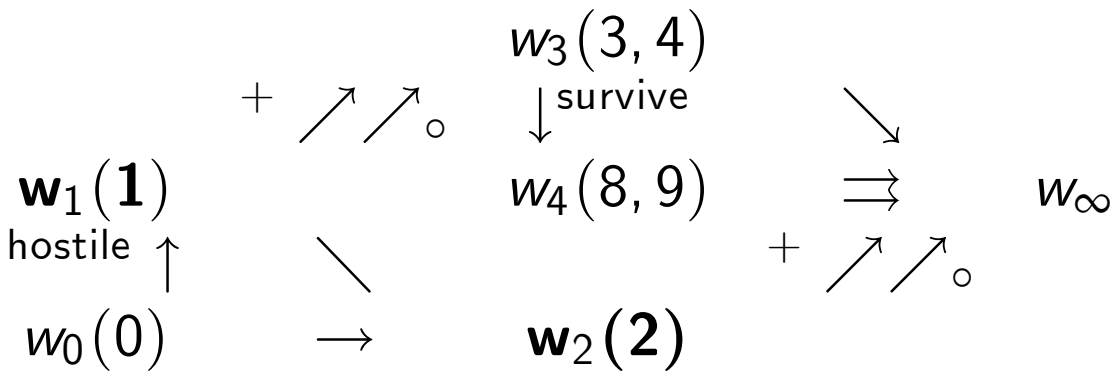
- Each position has an associated *floret*: that position and its emanating edges.
- Edges in florets of positions in the same stage are colored to convey isomorphism.

Example of an Event Tree



Example of a CEG

Draw CEG with vertices as positions
and undirected edges between stages.



Some Results about CEG's - Smith and Anderson(2008)

Theorem

If the random variables $X_1, X_2, \ldots, X_n$ with known sample spaces are fully expressed as a BN, G, or as a context specific BN G, and you know its CEG, C, then the random variables $X_1, X_2, \ldots, X_n$ and all their conditional independence structure together with their sample spaces can be retrieved from C.

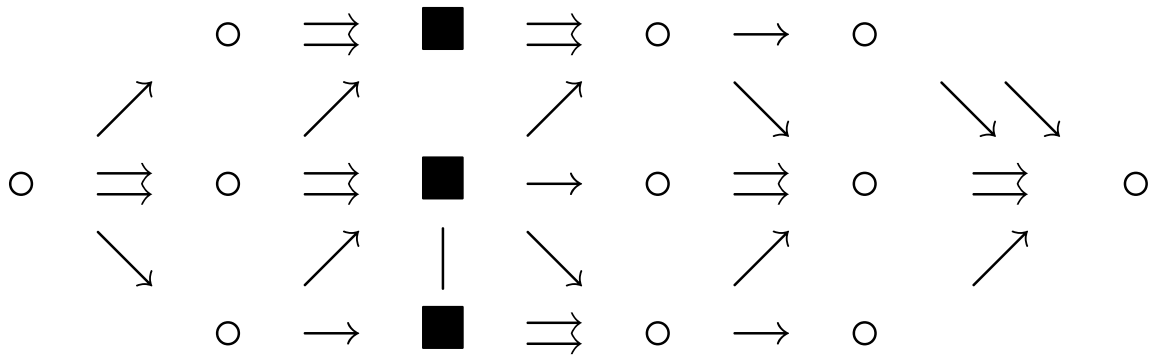
Theorem

$Downstream \amalg Upstream | \ w-Cut$

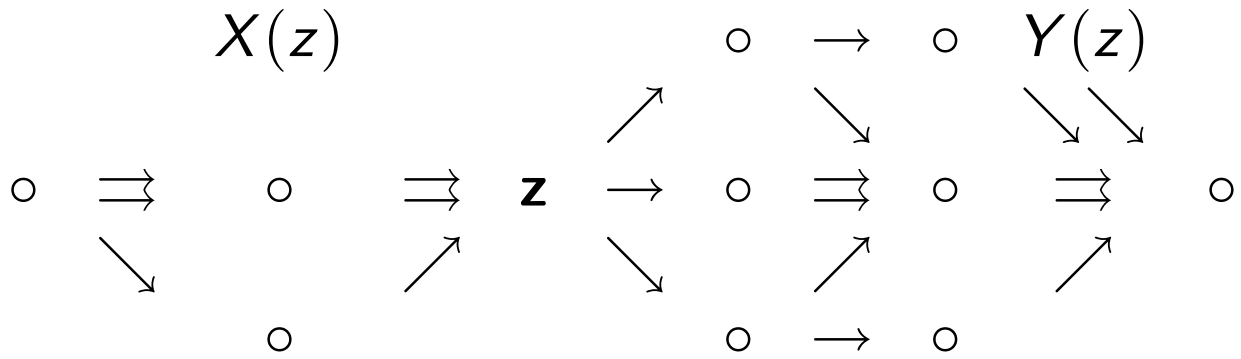
Theorem

$Children \amalg Upstream | u-Cut$

Example of a CEG with Cuts



Downstream $Y(z)$ independent of upstream $X(z)$ given cut $Z = z$.Cuts need not be orthogonal. So can construct dependence through functional relationships.

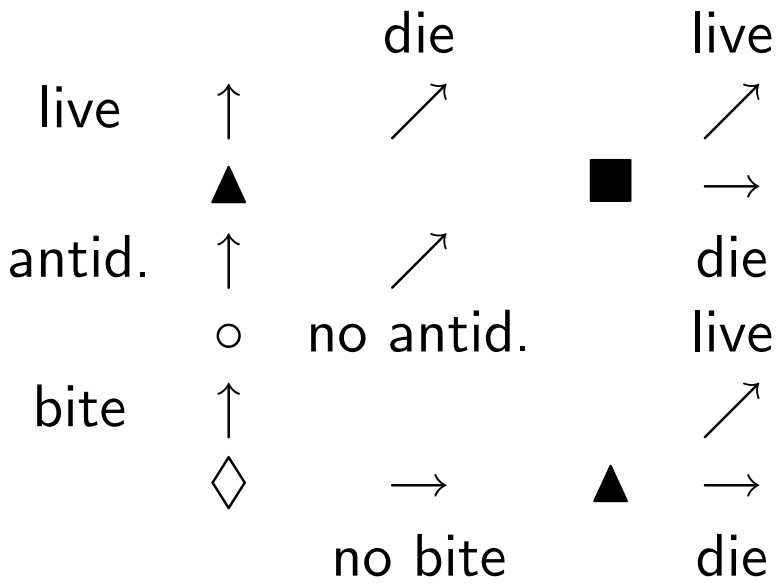


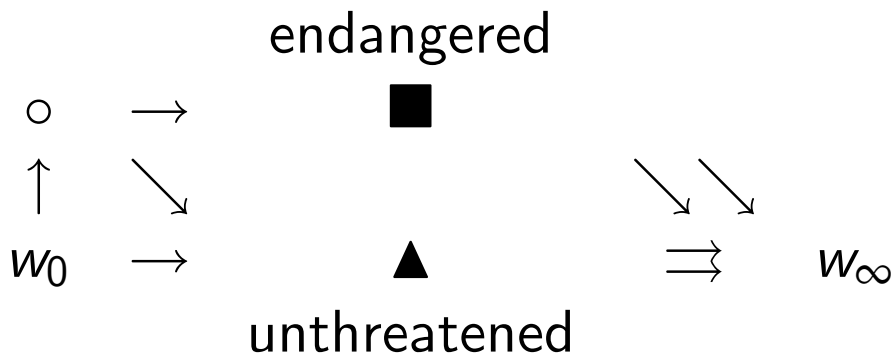
Towards Separation theorems: Thwaites and Smith(2009)

- The results above suggest that the there might be a necessary and sufficient separation theorem about rv's measurable with respect to the sigma field represented in the coloured graph.
- But what are the random variables to which such results apply and what form do the conditional independence statements take?
- The intrinsic variables about which such theorems apply are vectors of floret and incidence variables.
- Peter Thwaites and I have now proved various results of this type about regular CEG's.
- Results analogous to arc reversal results in BN's allow us to identify the equivalence classes of CEGs.
- Note that the topology of the CEG allow us to identify variables that are intrinsic.

Snake Bite Example

$X_1 \sim$ Bitten by snake, $X_2 \sim$ Carry and apply perfect antidote,$X_3 \sim$Die tomorrow..

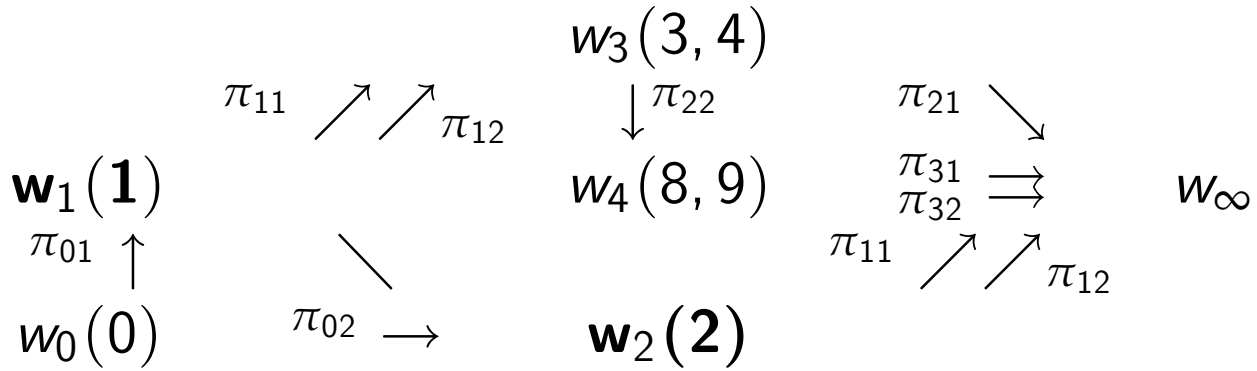




$X \sim$ not bitten/ bitten but apply antidote, $Y \sim (= X_3)$ live/die, $Z \sim$ unthreatened/endangered.
So from the CEG preferred variables exhibiting the conditional independence can be deduced from graph.

Probabilities on the gene CEG

- Embellish a CEG with probabilities just as in a tree.
- Note that the positions in the same stage have the same associated edge probabilities.
- Probabilities of atoms calculated by producting up edge probabilities on each root to leaf path.



Probabilities and Algebra of CEG's

- Each stage u has an associated simplex of probabilities $\{\pi_{i,u} : 1 \leq i \leq l_u\}$associated with its emanating edges in the CEG
- In our first example $l_u = 2$ and the root to sink probabilities are given by

$$p(v_5) = \pi_{20}\pi_{21} \qquad p(v_6) = \pi_{20}\pi_{11}$$
$$p(v_7) = \pi_{10}\pi_{11}\pi_{12} \qquad p(v_{10}) = \pi_{10}\pi_{21}\pi_{12}$$
$$p(v_{11}) = \pi_{10}\pi_{11}\pi_{22}\pi_{13} \qquad p(v_{12}) = \pi_{10}\pi_{11}\pi_{22}\pi_{23}$$
$$p(v_{13}) = \pi_{10}\pi_{21}\pi_{22}\pi_{23} \qquad p(v_{14}) = \pi_{10}\pi_{21}\pi_{22}\pi_{13}$$

- The probability of seeing of a value of a random variable on this space is the sum of these monomials. Note that the $8-$vector of atomic probabilities is constrained to lie in a 4 (rather than 7) dimensional space.
- Unlike the BN of the generating monomials need not be multilinear or homogeneous - in above they range from degree 2 to 4.

Conjugate Bayesian Inference on CEG's

- Because the likelihood separates, the class of regular CEG's admits simple conjugate learning.
- Explicitly the likelihood under complete random sampling is given by

$$\begin{aligned} l(\boldsymbol{\pi}) &= \prod_{u \in U} l_u(\boldsymbol{\pi}_u) \\ l_u(\boldsymbol{\pi}_u) &= \prod_{i \in u} \pi_{i,u}^{x(i,u)} \end{aligned}$$

  where $x(i,u)$ is the number of units entering stage u and proceeding along edge labelled (i,u). and $\sum_i \pi_{u,i} = 1$
- Independent Dirichlet priors $D(\boldsymbol{\alpha}(u))$ on the vectors $\boldsymbol{\pi}_u$ leads to independent Dirichlet $D(\boldsymbol{\alpha}^*(u))$ posteriors where

$$\alpha^*(i,u) = \alpha(i,u) + x(i,u)$$

Conjugate Bayesian Inference on CEG's

- Prior stage floret independence is a generalisation of local and global independence in BNs. Just as in Geiger and Heckerman(1997), floret independence, together with appropriate Markov equivalence characterises this product Dirichlet prior (see Freeman and Smith, 2009).
- Just like for BNs, non - ancestral sampling of a CEG data destroys conjugacy, but inference is no more difficult than for a BN.

Learning the topology of a CEG

.

- Choosing appropriate priors on model space and modular parameter priors over CEGs, for any CEG log marginal likelihood score is *linear* in stage components.
- Explicitly for $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$, let $s(\boldsymbol{\alpha}) = \log\Gamma(\sum_{i=1}^{k} \alpha_i)$ and $t(\boldsymbol{\alpha}) = \sum_{i=1}^{k} \log\Gamma(\alpha_i)$

$$\begin{aligned} \Psi(C) &= \log p(C) = \sum_{u \in C} \Psi_{u(c)} \\ \Psi_{u(c)} &= \sum s(\alpha(i,u)) - s(\alpha^*(i,u)) + t^*(\alpha(i,u)) - t(\alpha(i,u)) \end{aligned}$$

- Conjugacy and linearity implies e.g. MAP model selection using AHC or weighted MAX SAT is simple and fast over the albeit vast space class of CEG's (see Freeman and Smith,2009).
- Can use to embellish BN search to include context specific BNs.

Causal Bayesian Networks

- Recall that for causal BNs
  - Variables not downstream of X, a manipulated node, are unaffected by the manipulation.
  - X is set to the manipulated value $\hat{x}$ with probability 1.
  - Effect on downstream variables is identical to ordinary conditioning.
- But many manipulations don't follow these rules, e.g. "Whenever a unit is in set A of positions, take it to another position B".

Causal CEGs

- This can be implemented on a CEG by making paths through a position w pass along a designated edge to a designated position w', retaining all other floret distributions elsewhere.
- Similarly to Bayesian Networks:
  - Probabilities of edges not after w are unchanged.
  - An edge from w to w' forces w' after w.
  - Downstream probabilities after w' are unchanged.
- Generalizations of Pearl's Backdoor Theorem can be proven Riccomagno et al(2008).
  - Uses topology of the CEG to determine when the Bayes estimate of the effect of a manipulation is consistent, given partially observed data from the corresponding unmanipulated CEG.

Conclusions

- A CEG provides a useful graphical generalization of the BN, retaining all of the advantages of a BN other than compactness. These include model representation, support for interrogation, framework for fast propagation, conjugate analysis and fast model selection.
- It provides a much better framework than a BN for representing and analyzing the consequences of causal hypotheses.
- The increased expressiveness of CEG's are especially useful in applications in biology, education and forensic science.

THANK YOU FOR YOUR ATTENTION!!

Selected References of the authors

Riccomagno, E. and Smith, J.Q. (2009) "The Geometry of Causal Probability Trees that are Algebraically Constrained" in "Optimal Design and Related Areas in Optimization and Statistics" Eds L. Pronzato and A.Zhigljavsky, Springer 131-152

Smith, J.Q. and Anderson P.E. (2008) "Conditional independence and Chain Event Graphs" Artificial Intelligence, 172, 1, 42 - 68

Thwaites, P.,Smith, J.Q. and Cowell, R. (2008)" Propagation using Chain Event Graphs" Proceedings of the 24th Conference in UAI, Editors D. McAllester and P. Myllymaki, 546 -553.

Thwaites, P.E. and Smith, J.Q.(2006) "Evaluating Causal effects using Chain Event Graphs" Proc. 3rd E. W.on Probabilistic Graphical Models" Prague, 2006 293 -301

Riccomagno, E.M., Smith, J.Q. and Thwaites, P.(2008) "Causal Analysis with Chain Event Graphs" CRISM Res. Rep.

Freeman, G. and Smith, J.Q. (2009) "Bayesian Model Selection of Chain Event Graphs" CRISM Res.Rep. 09-06.

Thwaites, P.E. and Smith, J.Q.(2009) "A Separation Theorem for Chain

Formal definitions of stages and positions

- Two nodes v, v' are in the same stage u exactly when $X(v), X(v')$ have the same distribution under a bijection $\psi_u(v, v')$, where

$$\psi_u(v, v') : \mathbb{X}(v) = E(\mathcal{F}(v, \mathcal{T})) \longrightarrow \mathbb{X}(v') = E(\mathcal{F}(v', \mathcal{T}))$$

- In other words, the two nodes have identical probability distributions on their edges.
- Two nodes v, v' are in the same position w exactly when there exists a bijection $\phi_w(v, v')$ from $\Lambda(v, T)$, the set of paths in the tree from v to a leaf node, to $\Lambda(v', T)$,the set of paths from v' to a leaf node, such that all edges in all the paths are coloured, and that the sequence of colors in any path is the same as that in the path under the bijection.

Formal definition of a staged tree

- A staged tree is a tree with stage set $L(\mathcal{T})$ and edges coloured as follows:
  - When $v \in u \in L(\mathcal{T})$, but u contains only one node, all edges emanating from v are left uncoloured
  - When u contains more than one node, all edges emanating from v are coloured, such that two edges $e(v, v*), e(v', v'*)$ have the same colour if and only if $\psi_u(e(v, v*)) = e(v', v'*)$

Formal Definition of a Probability Graph

- The probability graph of a staged tree is a directed graph, possibly with some coloured edges. Each node represents a set of nodes from the probability tree in the same position in the staged tree
- Its edges are constructed as follows:
  - For each position w, choose a representative node $v(w)$. For each edge from $v(w)$ to $v'(w')$, construct a single edge $e(w, w')$, where $w' = w_\infty$ if v' is a leaf node in the tree; otherwise w' is the position of v'.
  - The colour of the edge is the colour of the edge between v and v'.
- So the number of edges in the probability tree is the same as in the staged tree.

A formal definition of the CEG

- The chain event graph is the mixed graph with
  - the same nodes as the probability graph;
  - the same directed edges as the probability graph; and
  - undirected edges drawn between different positions that are at the same stage
- The colors of the edges are also inherited from the probability graph

Manifest polynomials, identifiability and independence

- Conditional independences appear as usual in terms of factorization. Thus

$$\begin{aligned} & \pi_{21}^{-1}(p(v_5), p(v_{10}), p(v_{14}), p(v_{13})) \\ = \quad & \pi_{11}^{-1}(p(v_6), p(v_7), p(v_{11}), p(v_{12})) \\ = \quad & (\pi_{20}, \pi_{10}\pi_{12}, \pi_{10}\pi_{22}\pi_{13}, \pi_{10}\pi_{22}\pi_{23}) \end{aligned}$$

  so that under the appropriate identification of events (as can be read from the CEG)

$$X(\pm) \coprod \text{ rest}$$

- Now suppose we learn the distribution of a variable determining whether or not the organism survives unharmed. This probability is simply the value of a polynomial: $\pi_{20} + \pi_{10}\pi_{22}\pi_{13}$. The flats of this polynomial within the model space above, define the conditioned model space.