The Algebra of Causal Trees and Chain Event Graphs

Jim Smith (with Peter Thwaites and Eva Riccomagno),

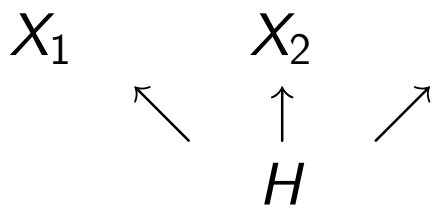
University of Warwick

August 2010

Introduction

- Algebraic geometry increasingly used to study the structure of discrete multivariate statistical models.
- Discrete graphical models e.g. Bayesian Networks especially useful.
- For many families of graphical model, each atomic probability in the joint pmf is a polynomial in certain "primitive" conditional probabilities.
- Inferring primitive probabilities from seeing the values of the marginal mass function of a measurable function then corresponds to the inverse map of a corresponding set of polynomials taking this value.
- Well known examples include the relationship between decomposable graphical models to toric ideals (see e.g. Garcia et al, 2005) and binary phylogenetic tree models and their semi algebraic characterisations, in terms of Mobius inversions and hyperdeterminants (see e.g. Zwiernik and Smith,2010).

A Typical example of a Binary Tree Model (Settimi and Smith,1998)



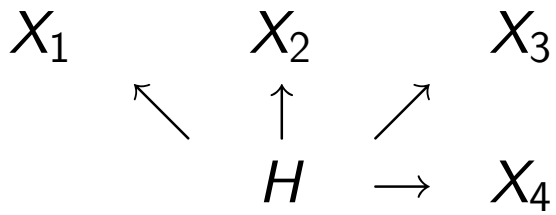
$$X_i = -1, 1, i = 1, 2, 3$$
$$H = -1, 1$$

- Form an algebraic point of view model on margin (X_1, X_2, X_3) saturated.(dim 7)
- However e.g. when $\mu_1 = \mu_2 = \mu_3$ then

$$\left(1 - \mu_H^2\right) \mu_{123}^2 - 4\mu_H^2 \mu_{12}\mu_{23}\mu_{13} = 0$$

- Probabilities real and in $[0, 1]$ $\Rightarrow$e.g $\mu_{12}\mu_{23}\mu_{13} > 0$ and $|\mu_{123}| \leq \frac{4}{3\sqrt{3}}$.
- Also observed means μ_1, μ_2, μ_3 induce further inequalily constraints.

More General Trees (Settimi and Smith,2000)



$$X_i = -1, 1, i = 1, 2, 3, 4$$
$$H = -1, 1$$

- This must now satisfy equations reducing dimension of (X_1, X_2, X_3, X_4).
- We lose 6 dimensions and demand moments satisfy 5 quadratic equations e.g.

$$\mu_{12}\mu_{34} - \mu_{14}\mu_{23} = 0$$

  and a quartic.
- Many other inequality constraints exist.
- Now marginals on binary phylogenetic trees fully characterized Allman and Rhodes(2010), Zwiernik and Smith(2010,2010a)

Contents of this talk

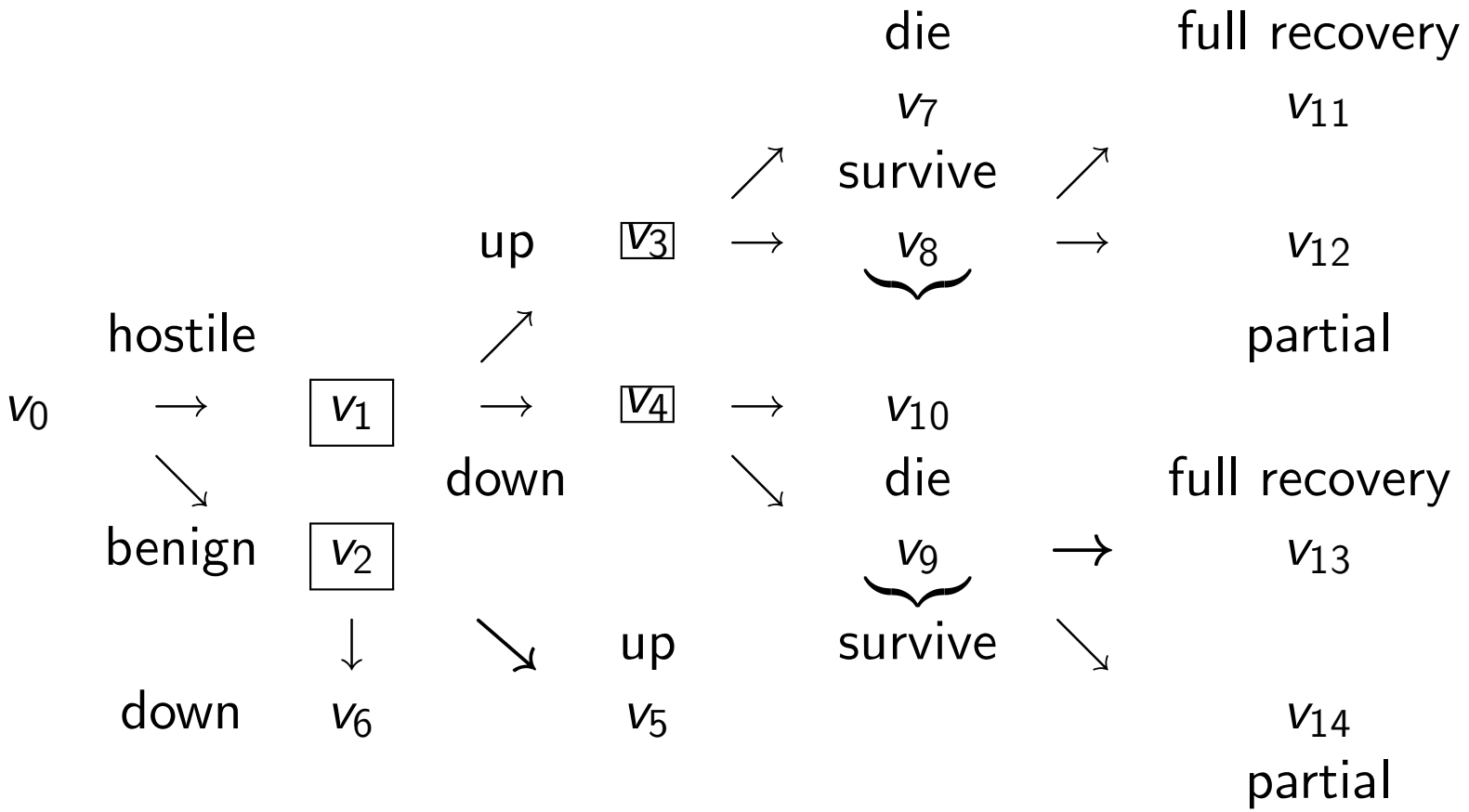
Today I will focus on a different class of graphical model based on event trees. They are a much richer class of models than the BN and so support much more varied polynomial forms in their algebraic formulations.

- Definition of an event tree and a chain event graph (CEG).
- Their corresponding polynomial representation.
- Inferential questions asked of the CEG.
- Causal questions associated with the CEG.
- Future challenges.

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 and algebraic descriptions.
- Causal hypotheses much more richly expressed than in their BN analogues.

Example of an Event Tree



Chain Event Graphs

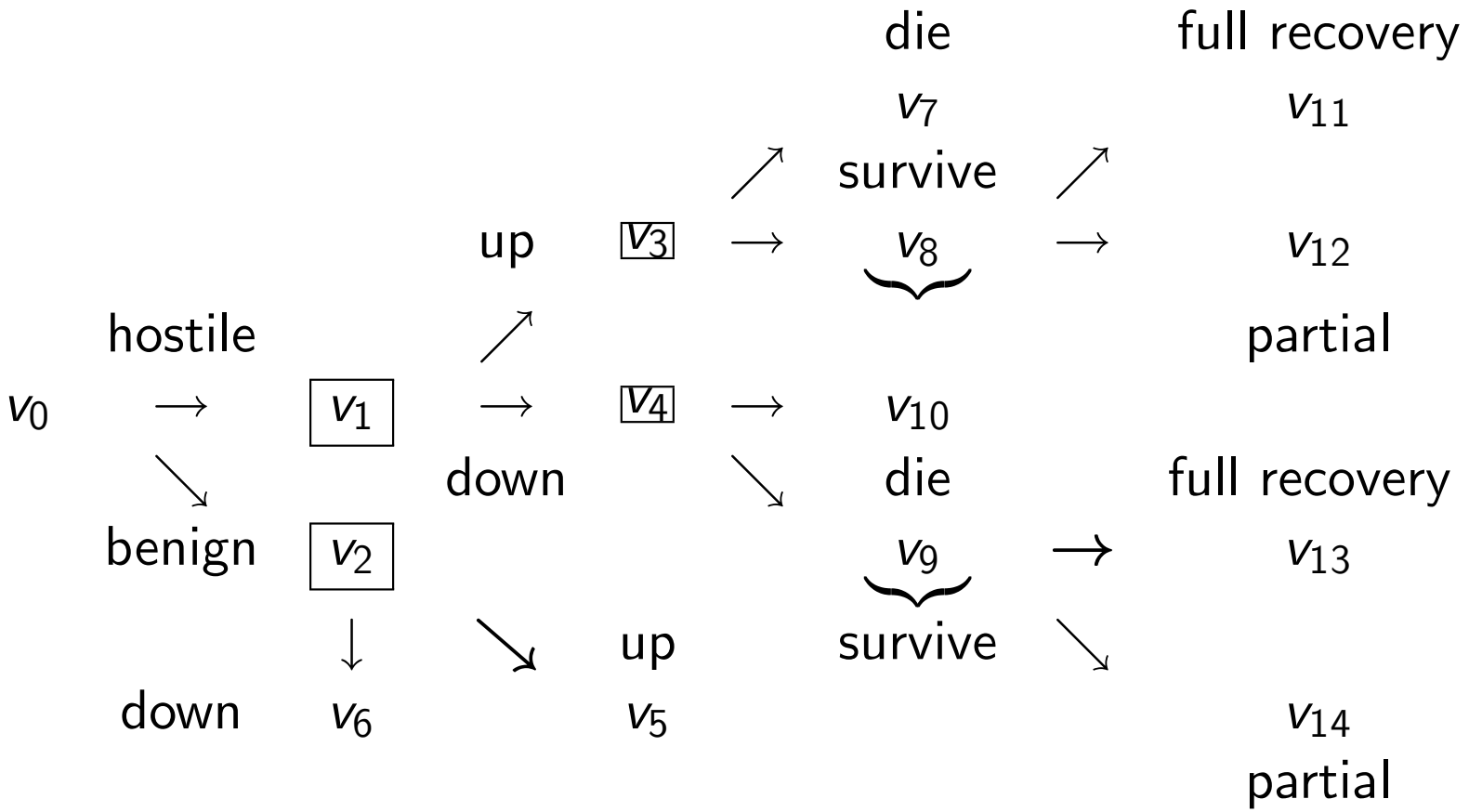
- Typically topologically much simpler than event trees but still describe how things happen.
- Their paths represent fully the structure of the sample space.
- Expresses rich variety of dependence structures to be graphically queried.
- Embellishes to a probability model and its associated algebraic 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 an Event Tree



Example of a CEG

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

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

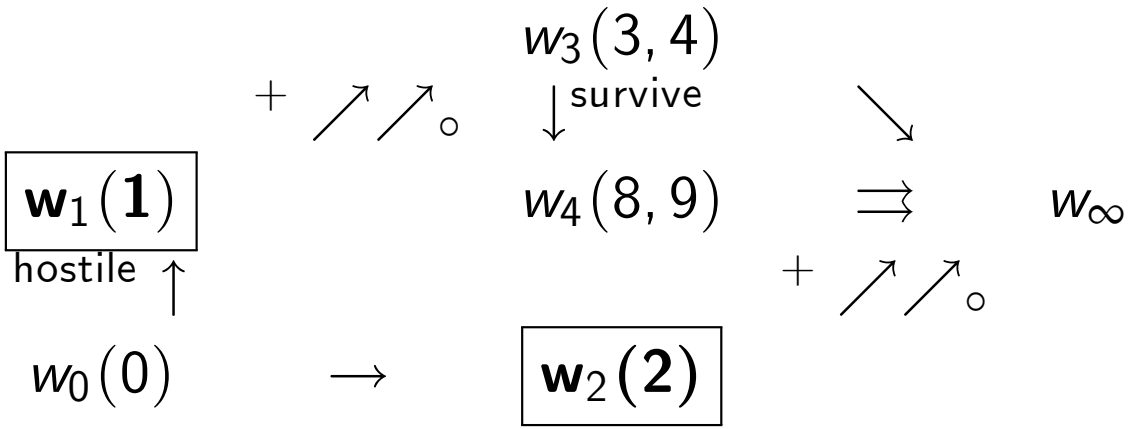
- Deduce *positions*: i.e. partition of situations with subsequent isomorphic trees

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

- 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 a CEG

Draw CEG with vertices as positions colour stages.



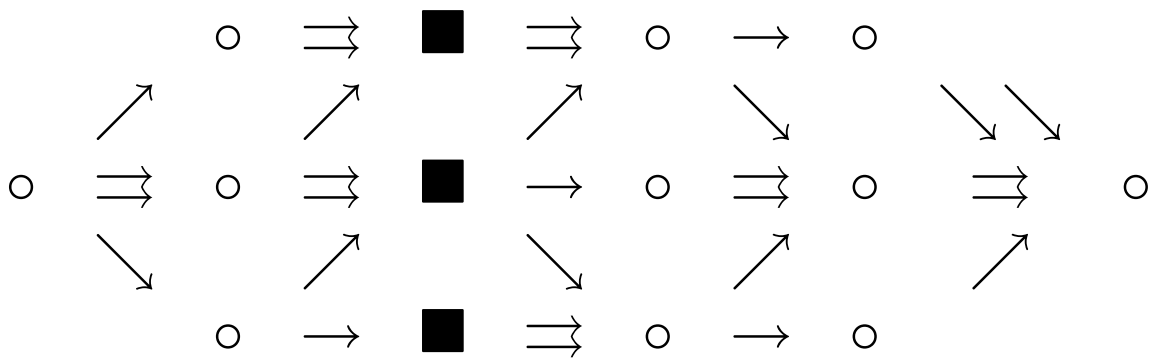
Properties of CEG's - Smith and Anderson(2008)

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.

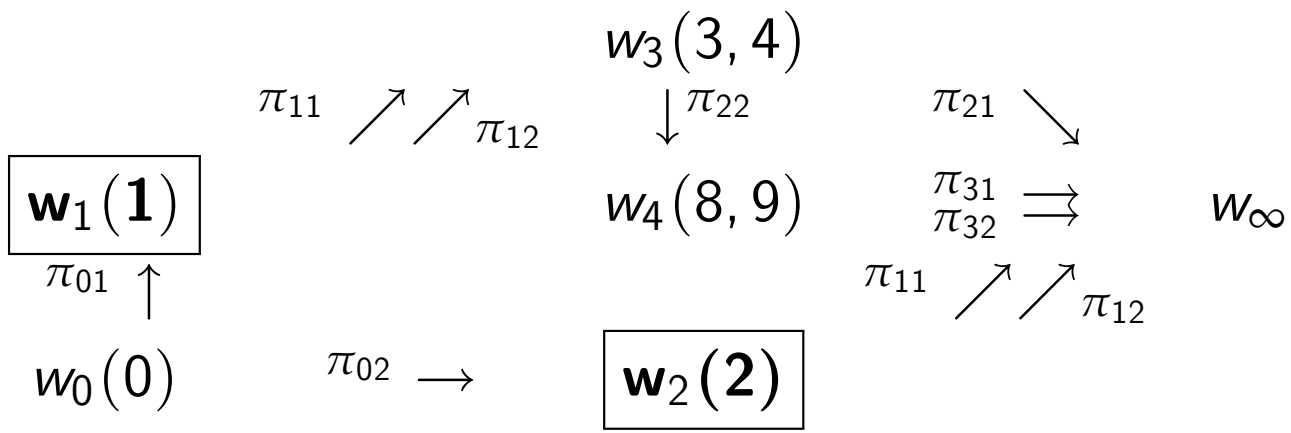
Theorem

Downstream $\amalg$ *Upstream*$|$ *w*$-$*Cut*



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 example $l_u = 2$ and the root to sink probabilities are given by

$$p(v_5) = \pi_{02}\pi_{21} \qquad p(v_6) = \pi_{02}\pi_{11}$$
$$p(v_7) = \pi_{01}\pi_{11}\pi_{21} \qquad p(v_{10}) = \pi_{01}\pi_{12}\pi_{21}$$
$$p(v_{11}) = \pi_{01}\pi_{11}\pi_{22}\pi_{31} \qquad p(v_{12}) = \pi_{01}\pi_{11}\pi_{22}\pi_{32}$$
$$p(v_{13}) = \pi_{01}\pi_{12}\pi_{22}\pi_{32} \qquad p(v_{14}) = \pi_{01}\pi_{12}\pi_{22}\pi_{31}$$

- The probability of learning the margin of a random variable on this space is to learn about some sums 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.

Manifest polynomials, identifiability and independence

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

$$\begin{aligned} & \pi_{21}^{-1}(p(v_5), p(v_{10}), p(v_{14}), p(v_{13})) \\ = \quad & \pi_{11}^{-1}(p(v_6), p(v_7), p(v_{11}), p(v_{12})) \\ = \quad & (\pi_{02}, \pi_{01}\pi_{21}, \pi_{01}\pi_{22}\pi_{31}, \pi_{01}\pi_{22}\pi_{32}) \end{aligned}$$

  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: $\pi_{02} + \pi_{01}\pi_{22}\pi_{31}$. The flats of this polynomial within the model space above, define the conditioned model space.

Causal Bayesian Networks

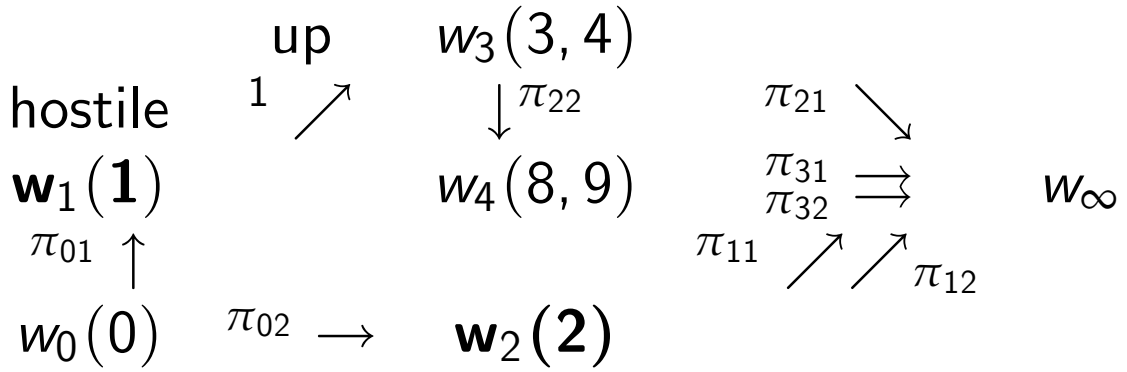
- Recall that for causal BNs
  - Variables not downstream of X, a manipulated node, are unaffected by the manipulation.
  - X is set to the manipulated value $\hat{x}$ with probability 1.
  - Effect on downstream variables is identical to ordinary conditioning.
- But many manipulations don't follow these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

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 Thwaites et al(2010).
  - 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.

An example of a causal CEG

Control: we plan to ensure gene is up regulated if in a hostile environment.



Now probability in the controlled system of being up regulated and dying is

$$p^*(v_7) = \pi_{01}\pi_{21}$$

In the uncontrolled system it is

$$p(v_7) = \pi_{01}\pi_{11}\pi_{21}$$

Often probabilities in a controlled system have a degree lower than in the unmanipulated system.

An example of a causal CEG

- Here we plan to ensure gene is up regulated if in a hostile environment the root to sink probabilities are given by

$$p^*(v_5) = \pi_{02}\pi_{21} \qquad p^*(v_6) = \pi_{02}\pi_{11}$$
$$p^*(v_7) = \pi_{01}\pi_{21} \qquad p^*(v_{10}) = 0$$
$$p^*(v_{11}) = \pi_{01}\pi_{22}\pi_{31} \qquad p^*(v_{12}) = \pi_{01}\pi_{11}\pi_{22}\pi_{32}$$
$$p^*(v_{13}) = 0 \qquad p^*(v_{14}) = 0$$

- Now suppose we want to estimate the probability of surviving unharmed if we force the gene to be up regulated if it is in a hostile environment. Then the probability that this occurs is

$$p^*(v_5) + p^*(v_6) + p^*(v_{11}) = \pi_{02} + \pi_{01}\pi_{22}\pi_{31}$$

- Note that this is the probability it would have survived if we did not control the environment. So we can identify this probability by observing the idle system.

Concluding Remarks

- The geometry of moderate dimensional polynomials is relatively well understood but only now being exploited to understand discrete statistical graphical models.
- In large problems, identifiability functions of parameters - like causal functions - as well as the inverse image of these observed functions when there is no identifiability, are giving valuable insight into reliability of estimation techniques and robustness of statistical inferences - especially in a Bayesian domain. We can **see** where the problems are!
- Two challenges in exploiting algebraic geometry for discrete statistical modeling are:
  1. Many results in algebraic geometry apply over polynomials over the complex field.
  2. Probabilities are positive and sum to one. The first issue demands a semi-algebraic rather than algebraic description.

THANK YOU FOR YOUR ATTENTION!!

Selected References of mine

Zwernik, P. and Smith, J.Q. (2010) "Tree-cumulants and the identifiability of Bayesian tree models" CRiSM Research Report (submitted)

Zwiernik, P. and Smith, J.Q. (2010) "The Geometry of Conditional Independence Tree Models with Hidden Variables" CRiSM Research Report (submitted)

Thwaites, P., Riccomagno, E.M.and Smith, J.Q. (2010) "Causal Analysis with Chain Event Graphs" Artificial Intelligence, 174, 889–909

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

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

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

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).

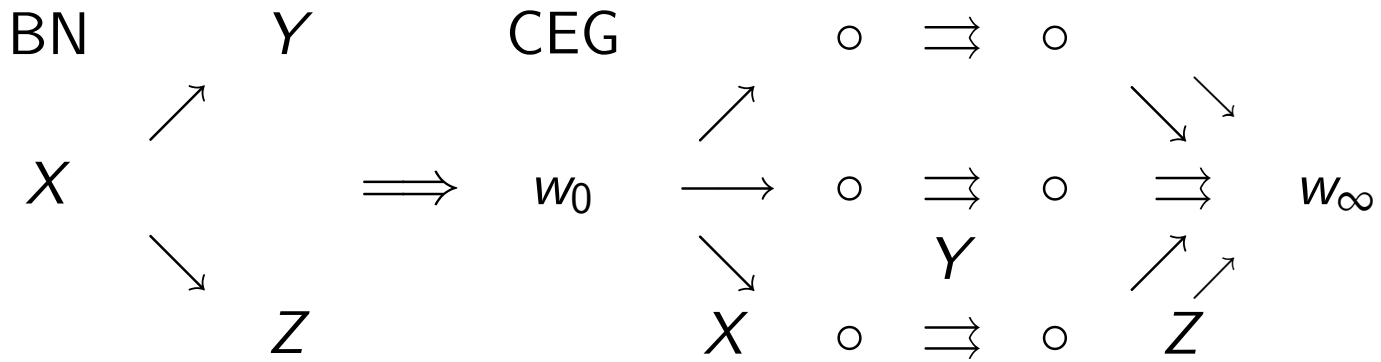
Challenges in searching for a CEG

.

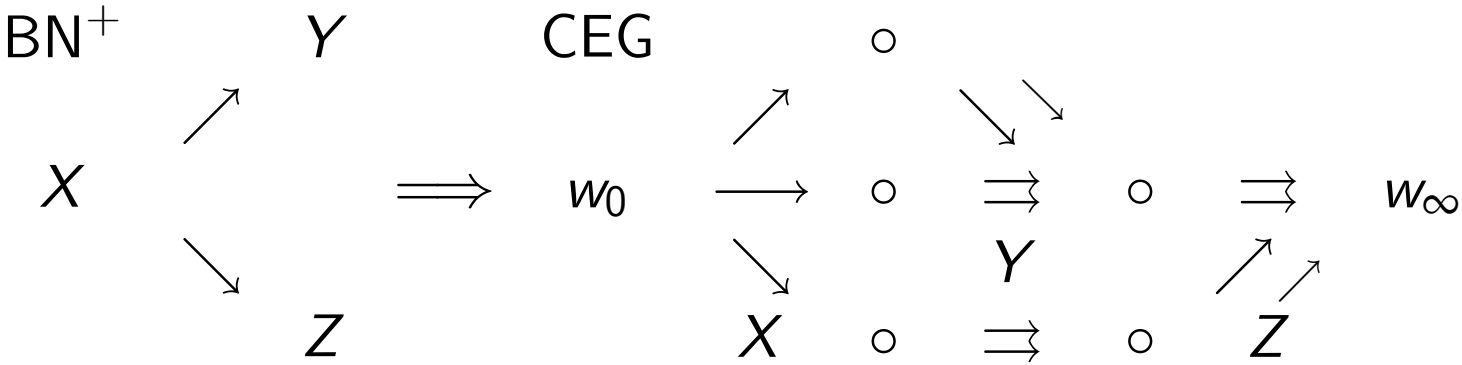
- CEG Search space of substantive problems is huge (orders of magnitude $>$ for BNs).
- However rationale behind the CEG helps design of intelligent search procedures.
- Often contextual information describing how things happen $\ll$ the size of space and makes methods feasible.
- E.g. in educational example (Freeman and Smith,2009) only need search over CEG's consistent with order of courses (event tree).
- CEG's can also be used to embellishing BN search. The score function above exactly corresponds to Bayes Factor score for BNs. So: search BN tree space $\longrightarrow$ search BN space associated with the best partial order $\longrightarrow$ search CEG embellishments.
- Without strong information, sparse tables seem to be combined to give MAP models with simple but unusual structure.

A CEG which extends a BN

.



but context specific BN^+ fits much better



(the distribution of Z is the same whether or not X takes a medium or large value)