# Estimating structure in graphical models

**Steffen L. Lauritzen**
**University of Oxford**

BS–IMS, Barcelona 2004

# Overview

- Introduction
- Conditional Independence Structures
- Estimating undirected graphs
- Estimating directed acyclic graphs
- Summary and perspectives

# Estimation of Structure

Advances in computing has set focus on *estimation of structure*:

- Model selection (e.g. subset selection in regression)
- System identification (engineering)
- Structural learning (AI or machine learning)

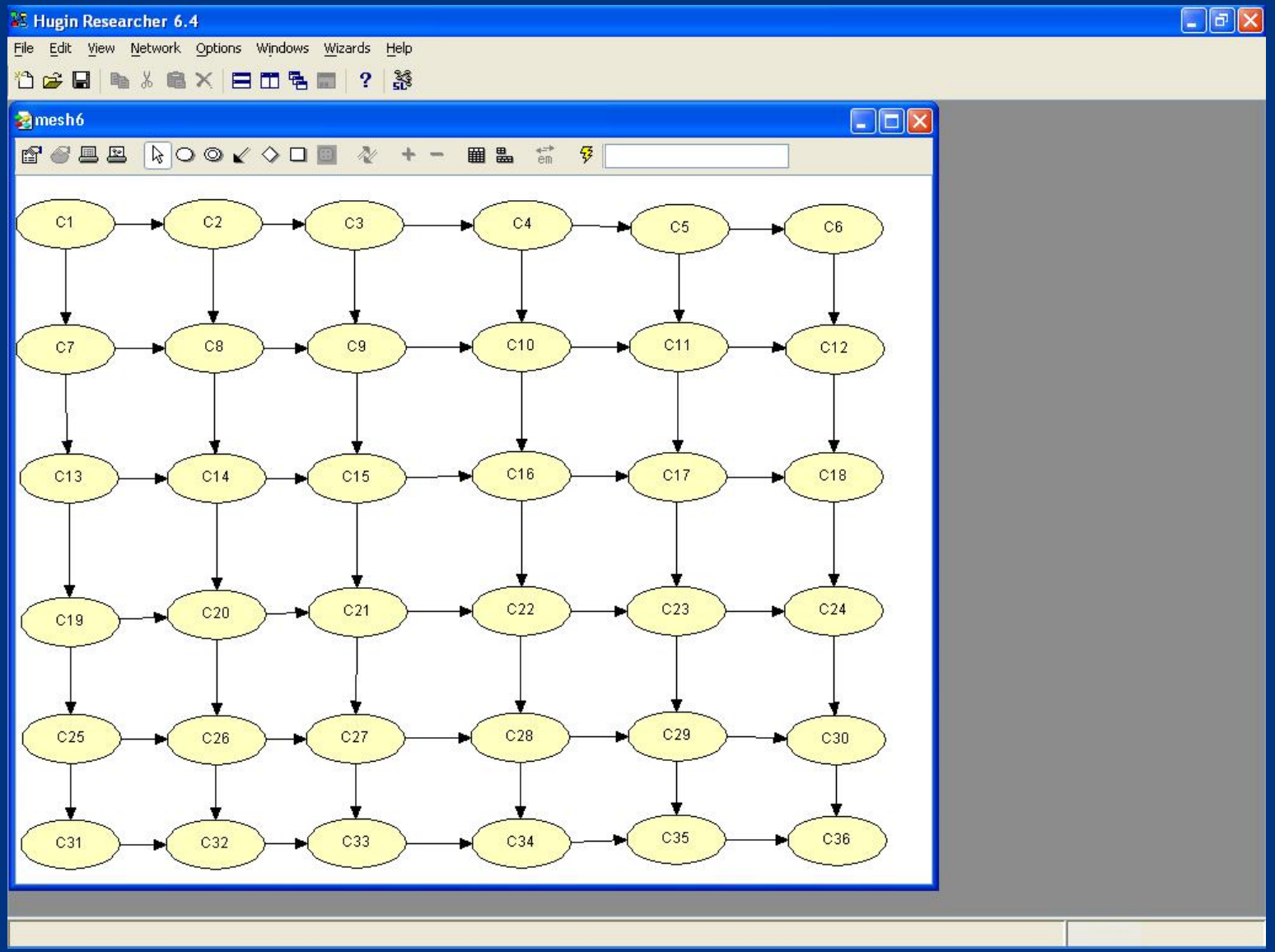
Graphical models describe conditional independence structures, so good case for formal analysis.

Methods must scale well with data size, as *many* structures and *huge* collections of data are to be considered.

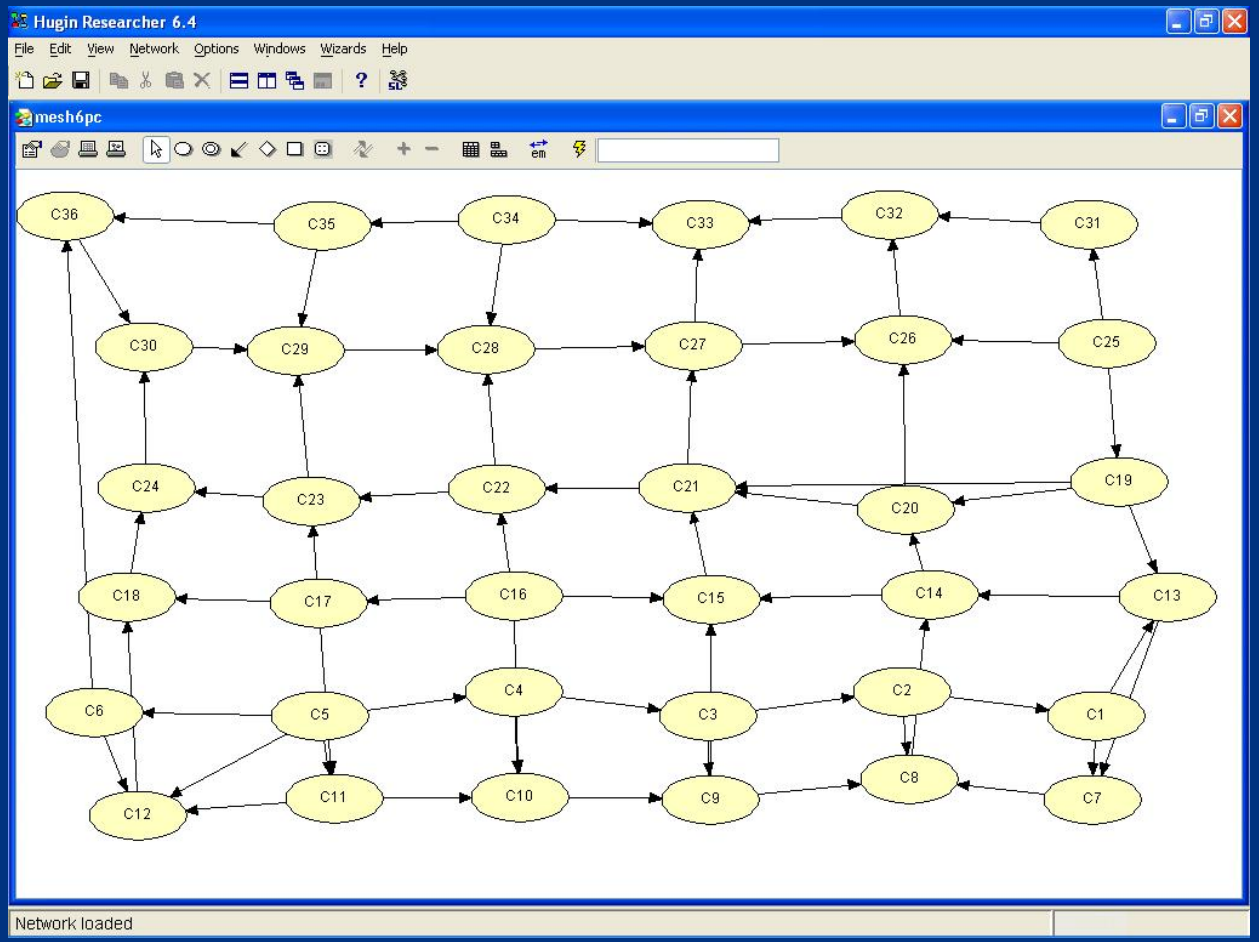
# Why estimation of structure?

- Parallel to e.g. density estimation
- Obtain quick overview of relations between variables in complex systems
- Data mining
- Gene regulatory networks
- Reconstructing family trees from DNA information
- Methods exist, but need better understanding of their statistical properties

# Markov mesh model

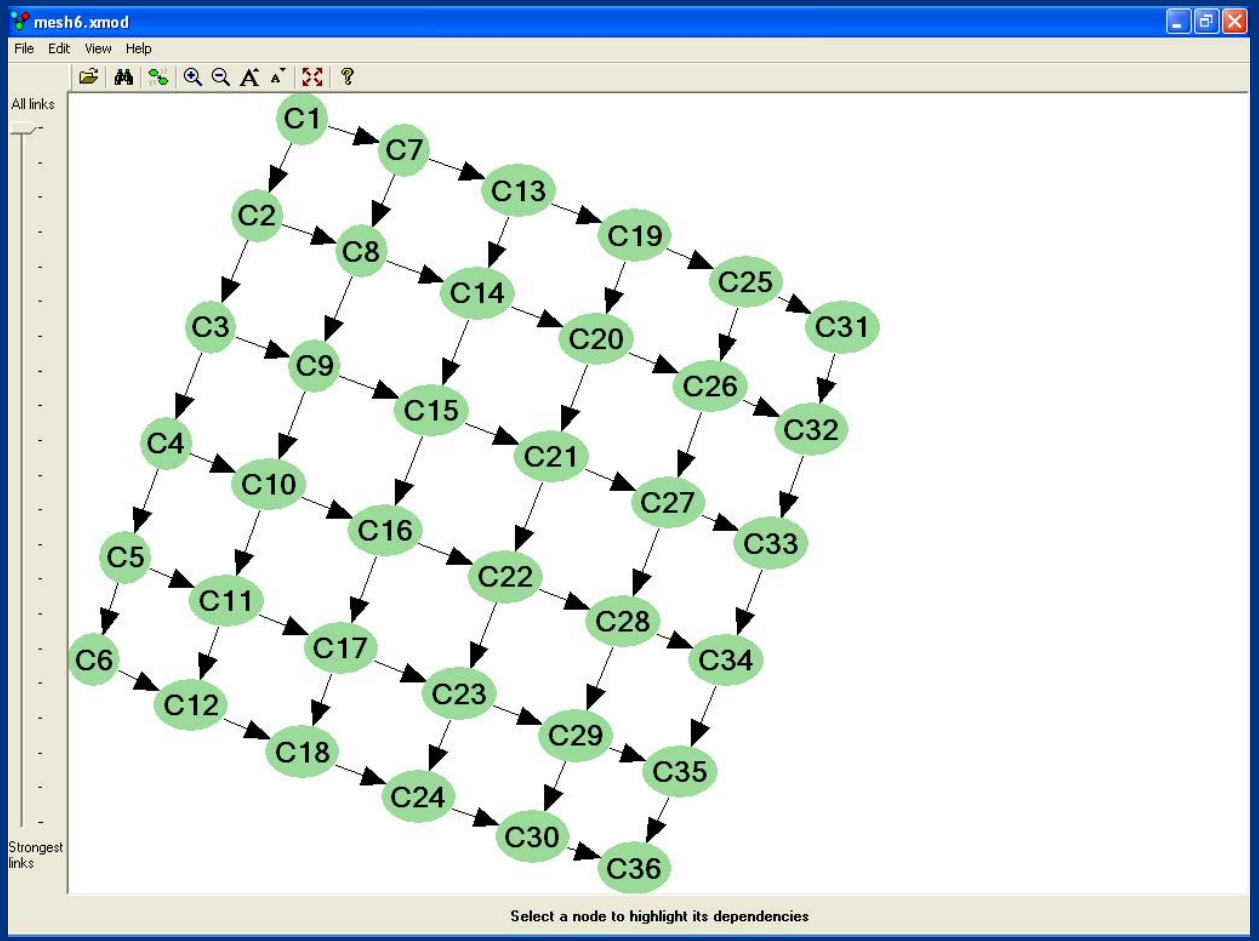


# PC algorithm



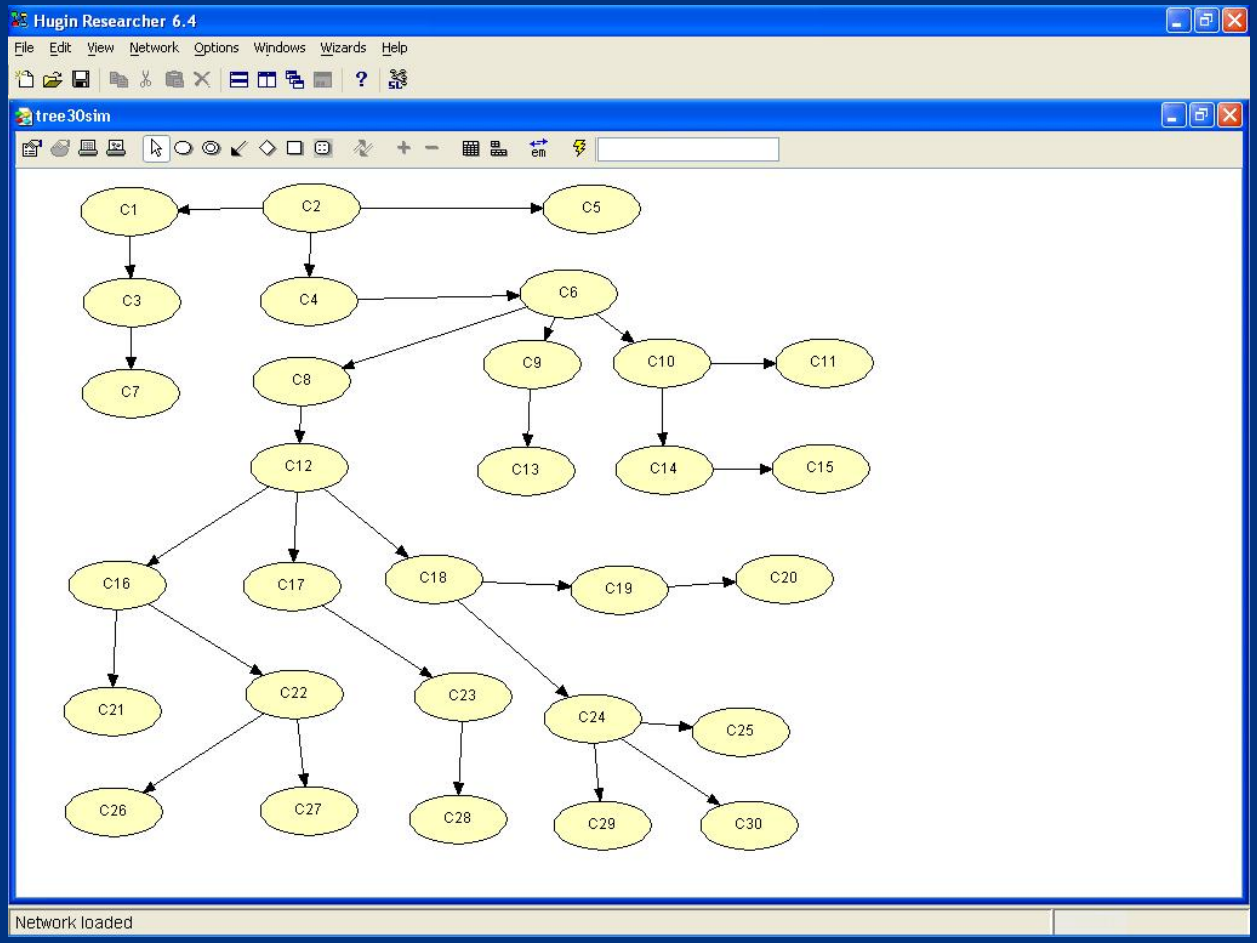

Crudest algorithm (HUGIN), 10000 simulated cases

# Bayesian GES



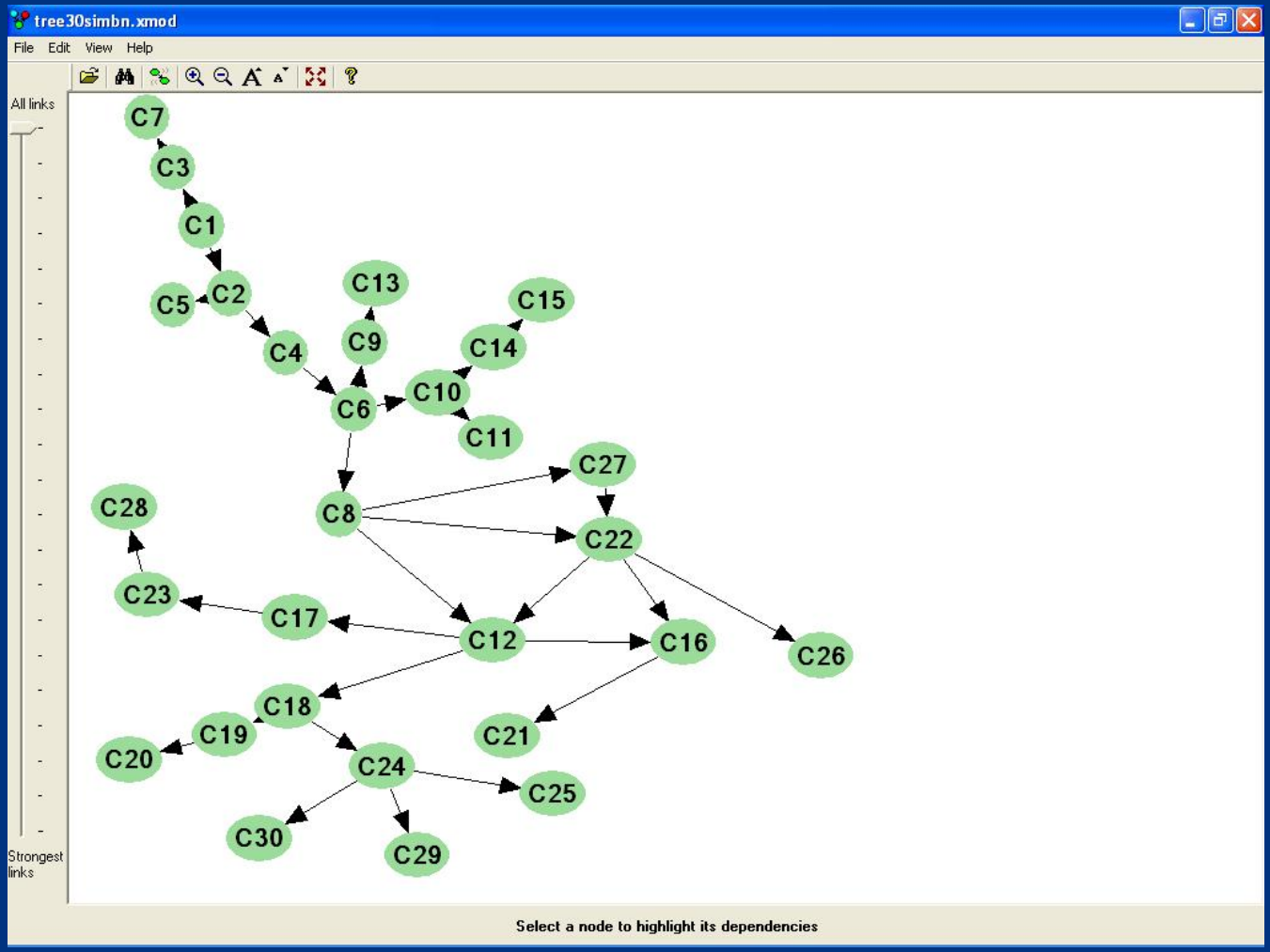

Crudest algorithm (WinMine), 10000 simulated cases

# Tree model

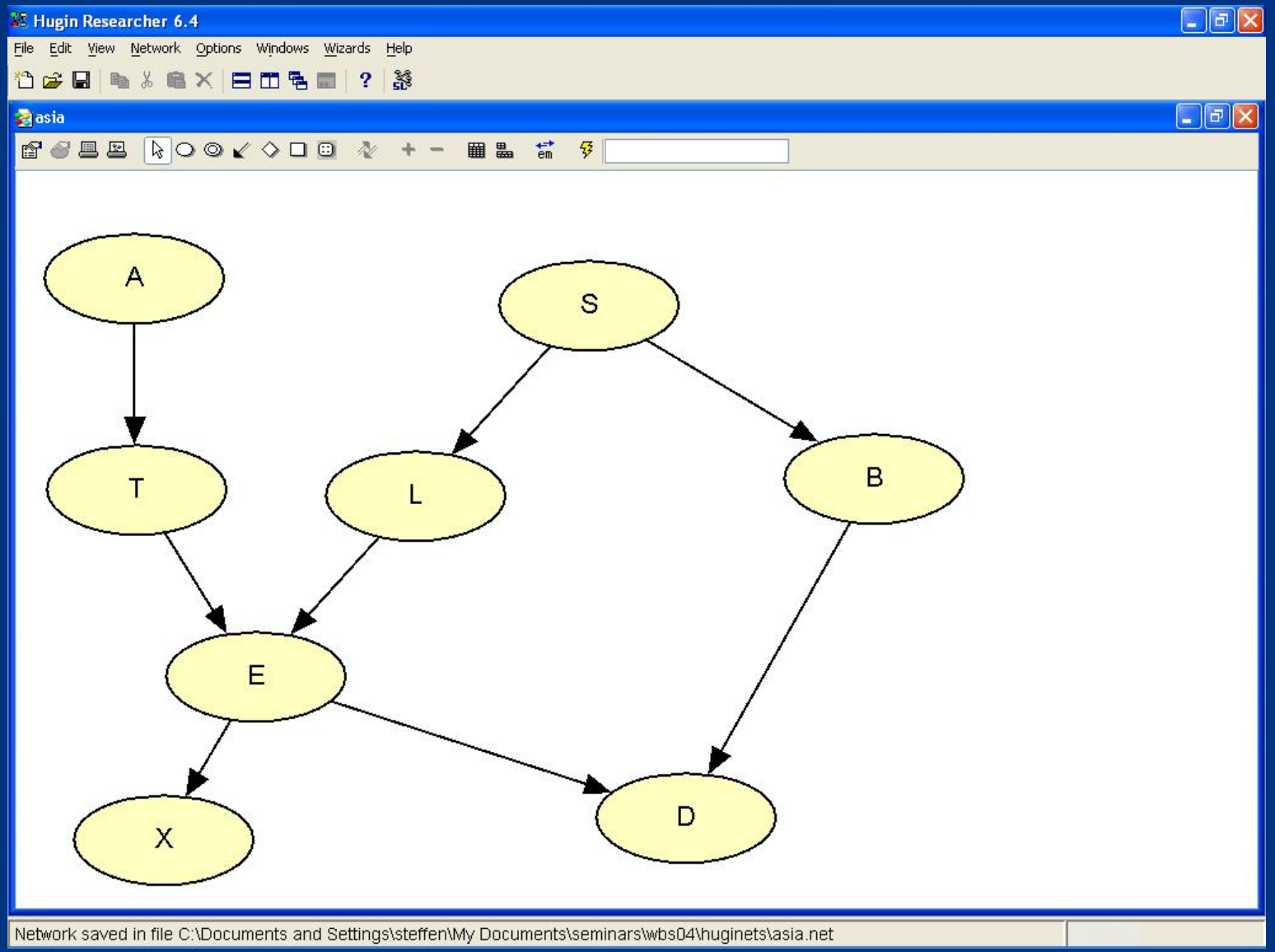



PC algorithm, 10000 cases, correct reconstruction

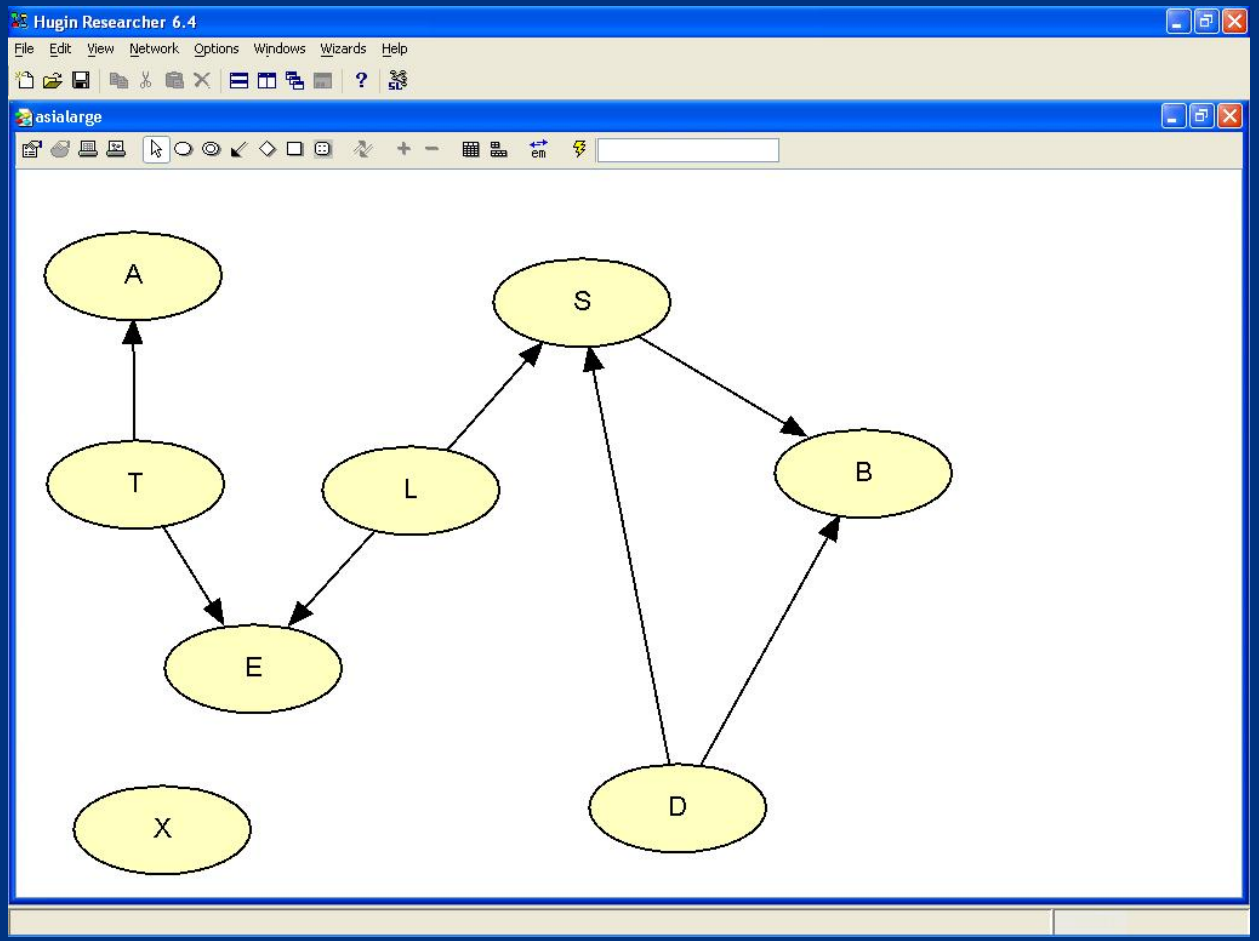
# Bayesian GES on tree



# Chest clinic

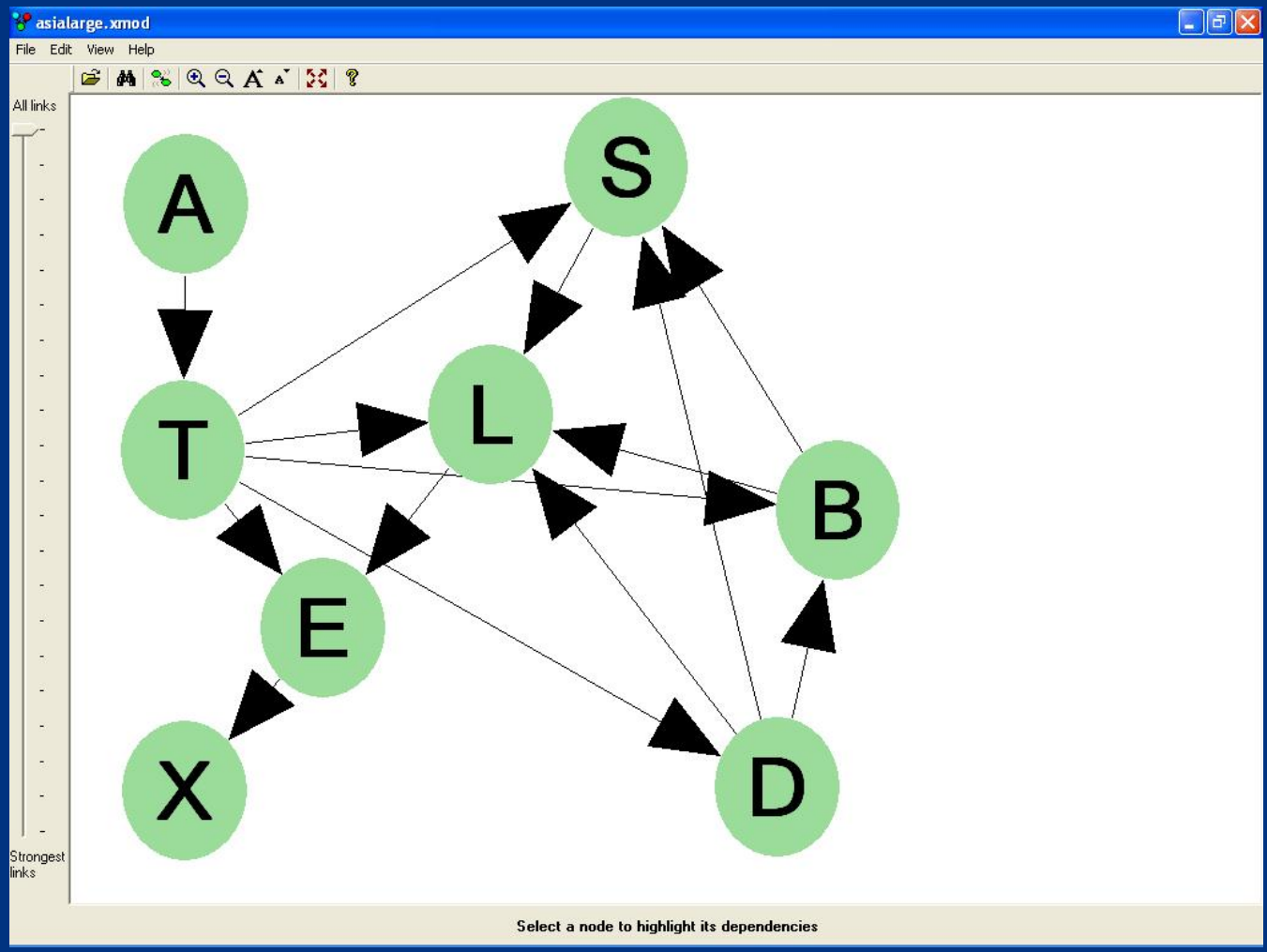


# PC algorithm




10000 simulated cases

# Bayesian GES



# Types of approach

- Methods for *judging adequacy of structure* such as
  - Tests of significance
  - Penalised likelihood scores

$$I_\kappa(\mathcal{M}) = \log \hat{L} - \kappa \dim(\mathcal{M})$$

    with $\kappa = 1$ for AIC Akaike (1974), or
    $\kappa = \frac{1}{2} \log N$ for BIC (Schwarz 1978).
  - Bayesian posterior probabilities
- *Search strategies* through space of possible structures, more or less based on *heuristics.*

# Conditional independence structures

- Undirected structures (Markov networks)
- Directed structures (Bayesian networks)
- Structures based upon chain graphs
- Other conditional independence structures
- Context dependent independence structures (split models)

*Lecture will focus on 'exact' results,* pertaining only to the first two of these.

# Conditional independence

$X$ and $Y$ are *conditionally independent* given $Z$ if $\mathcal{L}(X \mid Y, Z) = \mathcal{L}(X \mid Z)$ and we write $X \perp\!\!\!\perp Y \mid Z$. It holds

(C1) if $X \perp\!\!\!\perp Y \mid Z$ then $Y \perp\!\!\!\perp X \mid Z$;

(C2) if $X \perp\!\!\!\perp Y \mid Z$ and $U = g(Y)$, then $X \perp\!\!\!\perp U \mid Z$;

(C3) if $X \perp\!\!\!\perp Y \mid Z$ and $U = g(Y)$, then $X \perp\!\!\!\perp Y \mid (Z, U)$;

(C4) if $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp W \mid (Y, Z)$, then $X \perp\!\!\!\perp (Y, W) \mid Z$;

# Some notation

$\mathcal{G} = (V, E)$ is finite and simple undirected graph and $A \perp_{\mathcal{G}} B \,|\, S$ denotes that $S$ *separates* $A$ from $B$ in $\mathcal{G}$.

Density of $X = (X_v, v \in V)$ *factorizes w.r.t.* $\mathcal{G}$ if

$$f(x) = \prod_{a \in \mathcal{A}} \psi_a(x),$$

where $\mathcal{A}$ are *complete* subsets of $\mathcal{G}$ and $\psi_a(x)$ depends on $x_a$ only.

Then $P$ satisfies *the global Markov property:*

$$A \perp_{\mathcal{G}} B \,|\, S \implies A \perp\!\!\!\perp B \,|\, S$$

Consider sample $x^{(n)} = (x^1, \ldots, x^n)$ from $P$.

# Estimating trees

Assume $P$ factorizes w.r.t. an unknown *tree* $\tau$.

Chow and Liu (1968) showed *MLE $\hat{\tau}$ of $\mathcal{T}$ has maximal weight*, where the *weight* of $\tau$ is

$$w(\tau) = \sum_{e \in E(\tau)} H_n(e)$$

and $H_n(e)$ is the empirical *cross-entropy* or *mutual information* between endpoint variables of the edge $e = \{u, v\}$:

$$H_n(e) = \sum \frac{n(x_u, x_v)}{n} \log \frac{n(x_u, x_v)/n}{n(x_u)n(x_v)/n^2}.$$

# More on trees

*Fast algorithms* (Kruskal Jr. 1956) compute maximal weight spanning tree from weights $W = (w_{uv}, u, v \in V)$.

Chow and Wagner (1978) show *a.s. consistency in total variation of* $\hat{P}$*:* If $P$ factorises w.r.t. $\tau$, then

$$\sup_x |p(x) - \hat{p}(x)| \to 0 \text{ for } n \to \infty,$$

so *if* $\tau$ *is unique for* $P$, $\hat{\tau} = \tau$ *for all* $n > N$ *for some* $N$.

If $P$ does not factorize w.r.t. a tree, $\hat{P}$ *converges to closest tree-approximation* $\tilde{P}$ *to* $P$ (Kullback-Leibler distance).

# Gaussian Trees

If $X = (X_v, v \in V)$ is regular multivariate Gaussian, it factorizes w.r.t. an undirected graph if and only if its *concentration matrix* $K = \Sigma^{-1}$ satisfies

$$k_{uv} = 0 \iff u \not\sim v.$$

Results of Chow et al. are easily extended to Gaussian trees, with the weight of a tree determined as

$$w(\tau) = \prod_{e \in E(\tau)} (1 - r_e^2)^{-1/2},$$

with $r_e^2$ being the empirical correlation coeffient between $X_u$ and $X_v$.

# Special features of tree models

- *Direct likelihood methods* (ignoring penalty terms) lead to sensible results.
- Sampling distribution of tree MLE can be *simulated*
- *Penalty terms additive along branches*, so highest AIC or BIC scoring tree (forest) also available using a MWST algorithm.
- Pairwise marginal counts are *sufficient statistics* for the tree problem (empirical covariance matrix in the Gaussian case).

# Bayesian analysis

For $g$ in specified set of graphs, $\Theta_g$ is associated parameter space so that $P$ factorizes w.r.t. $g$ if and only if $P = P_\theta$ for some $\theta \in \Theta_g$.

$\pi_g$ is prior on $\Theta_g$. Prior $p(g)$ is uniform for simplicity.

Based on $x^{(n)}$, posterior distribution of $G$ is

$$p^*(g) = p(g \,|\, x^{(n)}) \propto p(x^{(n)} \,|\, g) = \int_{\Theta_g} p(x^{(n)} \,|\, g, \theta)\, \pi_g(d\theta).$$

Bayesian analysis looks for *MAP estimate* $g^*$ maximizing $p^*(g)$ or attempts to *sample from posterior*, using e.g. MCMC.

# Decomposable graphical models

For connected decomposable graphs and *strong hyper Markov priors* Dawid and Lauritzen (1993) show

$$p(x^{(n)} \mid g) = \frac{\prod_{C \in \mathcal{C}} p(x_C^{(n)})}{\prod_{S \in \mathcal{S}} p(x_S^{(n)})},$$

where each factor has explicit form. $\mathcal{C}$ are the *cliques* of $g$ and $\mathcal{S}$ the *separators* (mininal cutsets).

Trees are decomposable, so for trees we get

$$p(\tau \mid x^{(n)}) \propto \prod_{e \in E(\tau)} p(x_e^{(n)}).$$

# Bayesian analysis

*MAP estimates of trees can be computed* (also in Gaussian case).

Good direct algorithms exist for generating random spanning trees (Aldous 1990), so *full posterior analysis is possible for trees*.

*MCMC methods for exploring posteriors* of undirected graphs have been developed, e.g. (Giudici and Green 1999) and (Roverato 2002)

*Only heuristics available for MAP estimators* or maximizing penalized likelihoods such as AIC or BIC, for other than trees.

# Some challenges for undirected graphs

- Find feasible algorithm for (perfect) simulation from a distribution over decomposable graphs as

$$p(g) \propto \frac{\prod_{C \in \mathcal{C}} w(C)}{\prod_{S \in \mathcal{S}} w(S)},$$

  where $w(A), A \subseteq V$ are a prescribed set of positive weights.

- Find feasible algorithm for obtaining MAP in decomposable case. This may not be universally possible as problem most likely is NP-complete.

# Consistency

Haughton (1988) shows in some generality that *BIC structure estimation is consistent* in the sense that, asymptotically, BIC is maximized for the simplest structure compatible with the sampling distribution.

Also that BIC under same assumptions as above is an approximation to the Bayesian posterior, so the *same is true for the full posterior.*

Direct to show *consistency of BIC and Bayesian posterior for trees*, as in Chow and Wagner (1978).

Adapting martingale argument in Doob (1949) yields rather general posterior consistency.

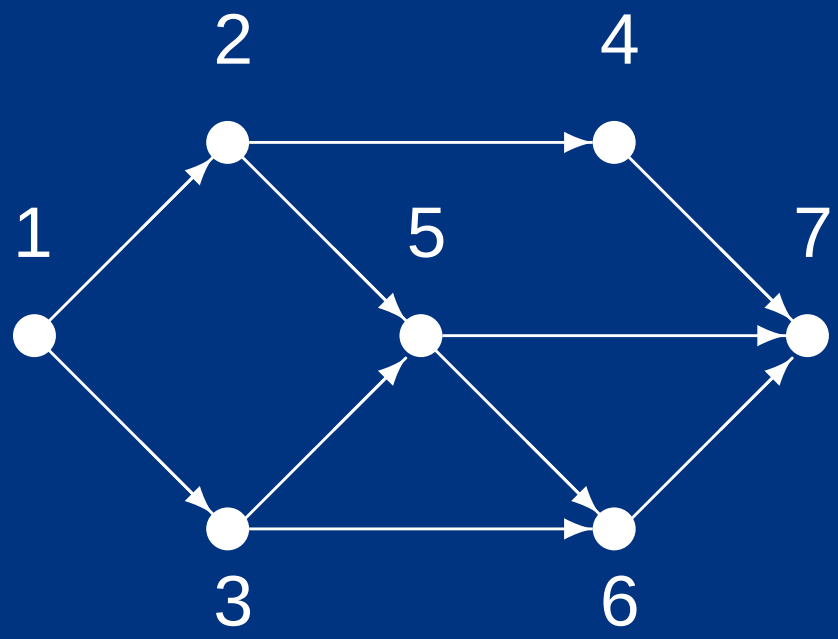
# Bayesian network

- Directed Acyclic Graph (DAG) $\mathcal{D} = (V, E)$.
- Nodes $V$ represent (random) variables $X_v, v \in V$
- Specify conditional distributions of children given parents: $p(x_v \mid x_{\mathrm{pa}(v)})$
- Joint distribution is then

$$p(x) = \prod_{v \in V} p(x_v \mid x_{\mathrm{pa}(v)}) \tag{1}$$

- Directed Markov properties identify conditional independence relations which follow from (1)

# Example of Bayesian network




Corresponds to the factorisation

$$\begin{aligned} p(x) &= p(x_1)p(x_2\,|\,x_1)p(x_3\,|\,x_1)p(x_4\,|\,x_2) \\ &\times\ p(x_5\,|\,x_2,x_3)p(x_6\,|\,x_3,x_5)p(x_7\,|\,x_4,x_5,x_6). \end{aligned}$$

# Global Directed Markov Property

The *factorization*

$$p(x) = \prod_{v \in V} p(x_v \,|\, x_{\mathrm{pa}(v)})$$

*is equivalent to*:

$$A \perp\!\!\!\perp B \mid S, \text{ whenever } A \perp_{\mathcal{D}} B \mid S$$

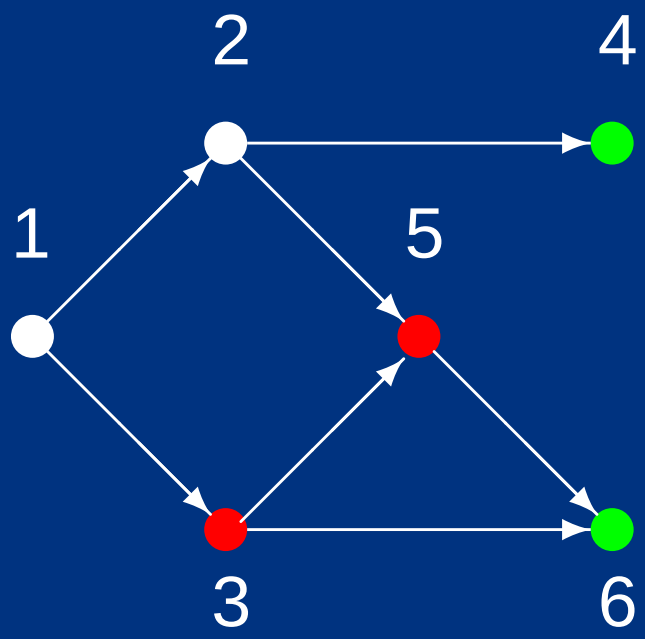
where $A \perp_{\mathcal{D}} B \mid S$ denotes that $A$ *and* $B$ *are* $d$*-separated by* $S$*.*

# Separation in DAGs

The $d$-separation $A \perp_{\mathcal{D}} B \mid S$ can e.g. be checked as follows:

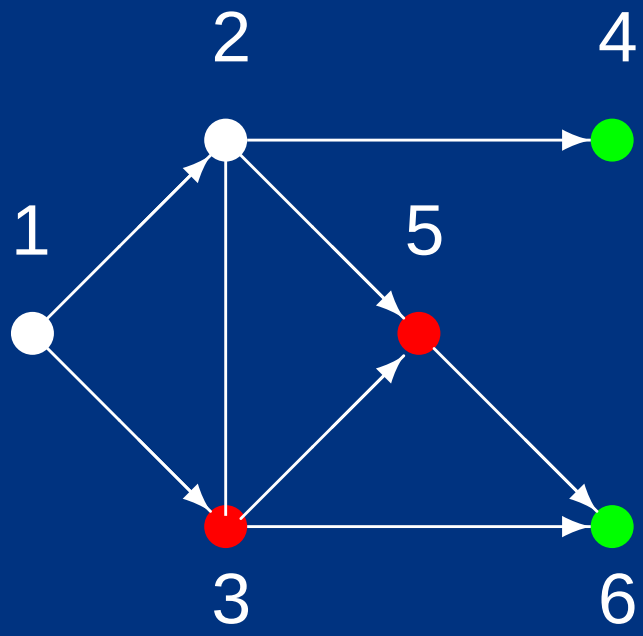
1. Reduce to subgraph induced by ancestral set of $A \cup B \cup S$
2. Add undirected edges between unmarried parents in this subgraph
3. Drop directions on all arrows
4. Then $A \perp_{\mathcal{D}} B \mid S$ *if and only if* $S$ *separates* $A$ *from* $B$ *in this undirected graph.*

# Forming ancestral set



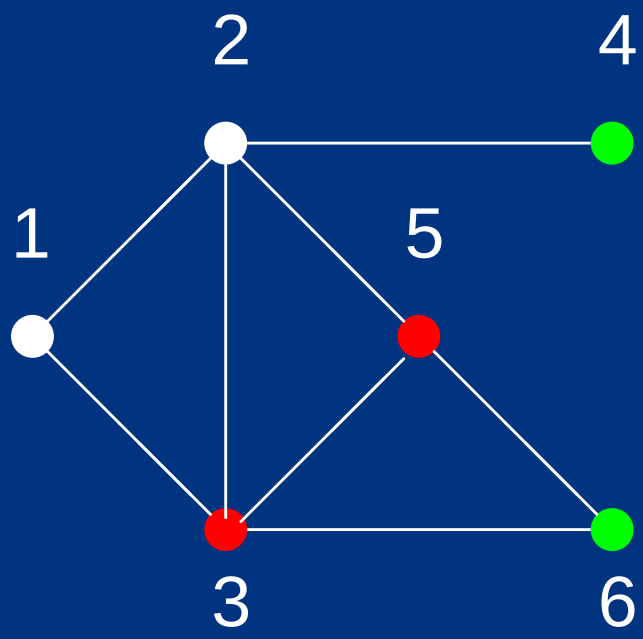

The subgraph induced by all ancestors of nodes involved in the query $4 \perp\!\!\!\perp 6 \,|\, 3, 5$?

# Adding links between unmarried parents




Adding an undirected edge between $2$ and $3$ with common child $5$ in subgraph induced by all ancestors of nodes involved in the query $4 \perp\!\!\!\perp 6 \mid 3, 5$?

# Dropping directions




Since $\{3, 5\}$ separates $4$ from $6$ in this graph, we conclude that $4 \perp\!\!\!\perp 6 \,|\, 3, 5$

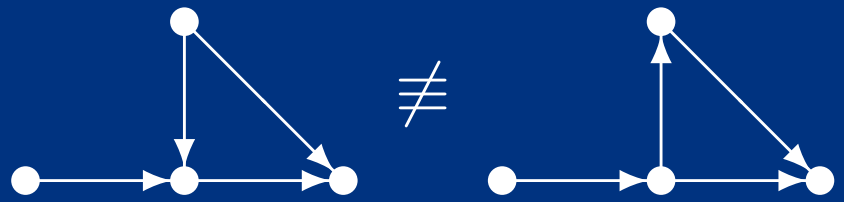
# Markov equivalence

*$\mathcal{D}$ and $\mathcal{D}'$ are equivalent if and only if:*

1. $\mathcal{D}$ and $\mathcal{D}'$ have same *skeleton* (ignoring directions)
2. $\mathcal{D}$ and $\mathcal{D}'$ have same unmarried parents

so

but



# Constraint-based search

**Step 1:** Identify skeleton using that, for $P$ faithful,

$$u \not\sim v \iff \exists S \subseteq V \setminus \{u, v\} : X_u \perp\!\!\!\perp X_v \mid X_S.$$

Begin with complete graph, check for $S = \emptyset$ and remove edges when independence holds. Then continue for increasing $|S|$.

*PC-algorithm* (Spirtes *et al*. 1993) exploits that only $S$ with $S \subseteq \mathrm{ne}(u)$ or $S \subseteq \mathrm{ne}(v)$ needs checking, $\mathrm{ne}$ refers to current skeleton.

**Step 2:** Identify directions to be consistent with independence relations found in Step 1.

# Faithfulness

A given distribution $P$ is in general compatible with a variety of structures, i.e. if $P$ corresponds to complete independence.

To identify a DAG structure something like the following must hold

$P$ is said to be *faithful* to $\mathcal{D}$ if

$$A \perp\!\!\!\perp B \mid X_S \iff A \perp_{\mathcal{D}} B \mid S.$$

*Most distributions are faithful.* More precisely, the non-faithful distributions form a Lebesgue null-set in parameter space associated with a DAG.

# Exact properties of PC-algorithm

*If $P$ is faithful to DAG $\mathcal{D}$, PC-algorithm finds $\mathcal{D}'$ equivalent to $\mathcal{D}$.*

It uses $N$ independence checks where $N$ is at most

$$N \leq 2\binom{|V|}{2}\sum_{i=0}^{d}\binom{|V|-1}{i} \leq \frac{|V|^{d+1}}{(d-1)!},$$

where $d$ is the maximal degree of any vertex in $\mathcal{D}$.

So worst case complexity is exponential, but *algorithm fast for sparse graphs.*

Sampling properties are less well understood although consistency results exist.

# Equivalence class searches

Searches directly in equivalence classes of DAGS.

Define *score function* $\sigma(\mathcal{D})$, with the property that

$$\mathcal{D} \equiv \mathcal{D}' \implies \sigma(\mathcal{D}) = \sigma(\mathcal{D}').$$

This holds e.g. if score function is *AIC or BIC or full Bayesian posterior with strong hyper Markov prior* (based upon Dirichlet or inverse Wishart distributions).

Equivalence class with maximal score is sought.

# Posterior distribution for DAG

For strong directed hyper Markov priors it holds that

$$p(x^{(n)} \mid d) = \prod_{v \in V} p(x_v^{(n)} \mid x_{\mathrm{pa}(v)}^{(n)})$$

so

$$p(d \mid x^{(n)}) \propto \prod_{v \in V} p(x_v^{(n)} \mid x_{\mathrm{pa}(v)}^{(n)}),$$

see e.g. Spiegelhalter and Lauritzen (1990), Cooper and Herskovits (1992), Heckerman *et al*. (1995)

*Challenge:* Find good algorithm for sampling from this full posterior.

# Greedy equivalence class search

1. Initialize with empty DAG
2. Repeatedly search among equivalence classes with a *single additional edge* and go to class with highest score - until no improvement.
3. Repeatedly search among equivalence classes with a *single edge less* and move to one with highest score - until no improvement.

*For BIC or fully Bayesian posterior score, this algorithm yields consistent estimate of equivalence class for* $P$*.*
(Chickering 2002)

# Some completely open questions

- What is the *speed of convergence* for consistent graph estimators?
- What are *natural neighbourhoods* to form confidence sets for graphs?
- Can some kind of *asymptotic distributions* be achieved?
- What are adaptive properties of *empirical Bayes* methods?

# Summary and further challenges

- Structure estimation is *practically possible but in need of a theory*
- More *exact results are needed*, to guide heuristics.
- For complexity reasons, consider *search algorithm itself as part of estimator*
- *Conceptual clarification* of properties of structure estimators needed
- Structures with *latent variables* constitute particularly challenging and important area.

# Some web addresses

*HUGIN:* www.hugin.com

*WinMine:* research.microsoft.com/~dmax/WinMine/Tooldoc.htm

*These overheads:*
www.stats.ox.ac.uk/~steffen/barcelona.pdf

Thank you!

# References


Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–23.

Aldous, D. (1990). A random walk construction of uniform spanning trees and uniform labelled trees. *SIAM Journal on Discrete Mathematics*, **3**, (4), 450–65.

Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of Machine Learning Research*, **3**, 507–54.

Chow, C. K. and Liu, C. N. (1968). Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, **14**, 462–7.

Chow, C. K. and Wagner, T. J. (1978). Consistency of an es-

timate of tree-dependent probability distributions. *IEEE Transactions on Information Theory*, **19**, 369–71.

Cooper, G. F. and Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, **9**, 309–47.

Dawid, A. P. and Lauritzen, S. L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *Annals of Statistics*, **21**, 1272–317.

Doob, J. L. (1949). Application of the theory of martingales. *Colloques Internationaux du Centre National de la Recherche Scientifique, Paris*, **13**, 22–7.

Giudici, P. and Green, P. J. (1999). Decomposable graphical Gaussian model determination. *Biometrika*, **86**, 785–801.

Haughton, D. (1988). On the choice of a model to fit the data

from an exponential family. *Annals of Statistics*, **16**, 342–55.

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

Kruskal Jr., J. B. (1956). On the shortest spanning subtree of a graph and the travelling salesman problem. *Proceedings of the American Mathematical Society*, **7**, 48–50.

Roverato, A. (2002). Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scandinavian Journal of Statistics*, **29**, 391–411.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–4.

Spiegelhalter, D. J. and Lauritzen, S. L. (1990). Sequential

updating of conditional probabilities on directed graphical structures. *Networks*, **20**, 579–605.

Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, prediction and search*. Springer-Verlag, New York. Reprinted by MIT Press.