Graphical models and Bayesian structural learning

Peter Green

UTS, Sydney
University of Bristol

ISBA 2016, Sardegna, 13 giugno 2016

Outline

Conditional independence

The key idea in understanding

- the structure of a multivariate distribution
- the structure of a sample of multivariate data

is conditional independence, a topic that has been extensively studied both in spatial statistics and in graphical modelling.

X and Y are conditionally independent given Z:

$$X \perp\!\!\!\perp Y \mid Z$$

means that if you already know the value of Z, learning that of Y tells you nothing more about X. Any dependence between X and Y is indirect, mediated through Z.

Conditional independence

The key idea in understanding

- the structure of a multivariate distribution
- the structure of a sample of multivariate data

is conditional independence, a topic that has been extensively studied both in spatial statistics and in graphical modelling.

X and Y are conditionally independent given Z:

$$X \perp\!\!\!\perp Y \mid Z$$

means that if you already know the value of Z, learning that of Y tells you nothing more about X. Any dependence between X and Y is indirect, mediated through Z.

Conditional independence, probabilistically

X and Y are conditionally independent given Z:

$$X \perp\!\!\!\perp Y \mid Z$$

means that if you already know the value of Z, learning that of Y tells you nothing more about X. Any dependence between X and Y is indirect, mediated through Z.

In terms of probability distributions, this means

$$p(x, y|z) = p(x|z)p(y|z)$$

It proves useful to represent conditional independences graphically.

Conditional independence, probabilistically

X and Y are conditionally independent given Z:

$$X \perp\!\!\!\perp Y \mid Z$$

means that if you already know the value of Z, learning that of Y tells you nothing more about X. Any dependence between X and Y is indirect, mediated through Z.

In terms of probability distributions, this means

$$p(x, y|z) = p(x|z)p(y|z)$$

It proves useful to represent conditional independences graphically.

Conditional independence, probabilistically

X and Y are conditionally independent given Z:

$$X \perp\!\!\!\perp Y \mid Z$$

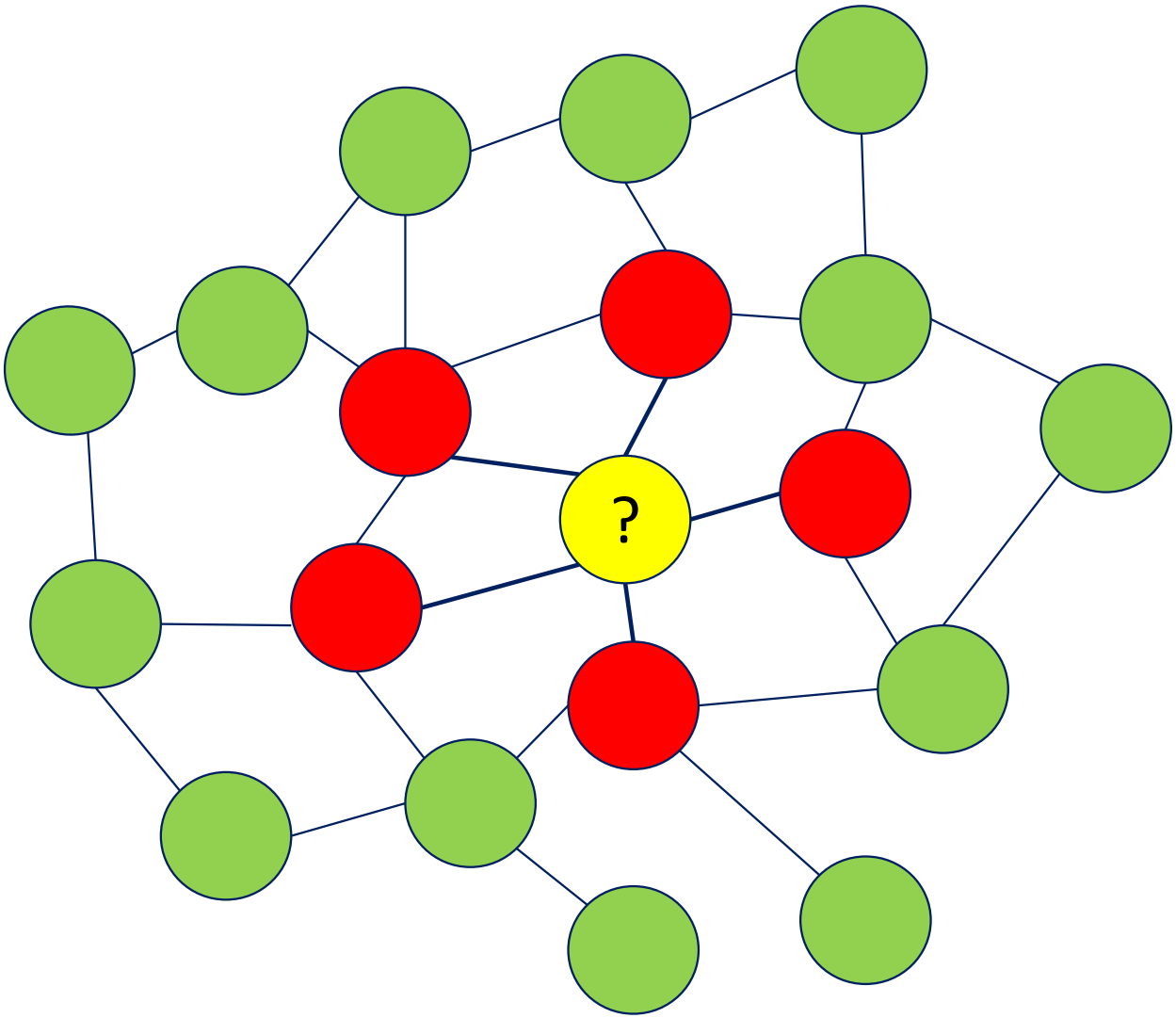
means that if you already know the value of Z, learning that of Y tells you nothing more about X. Any dependence between X and Y is indirect, mediated through Z.

In terms of probability distributions, this means

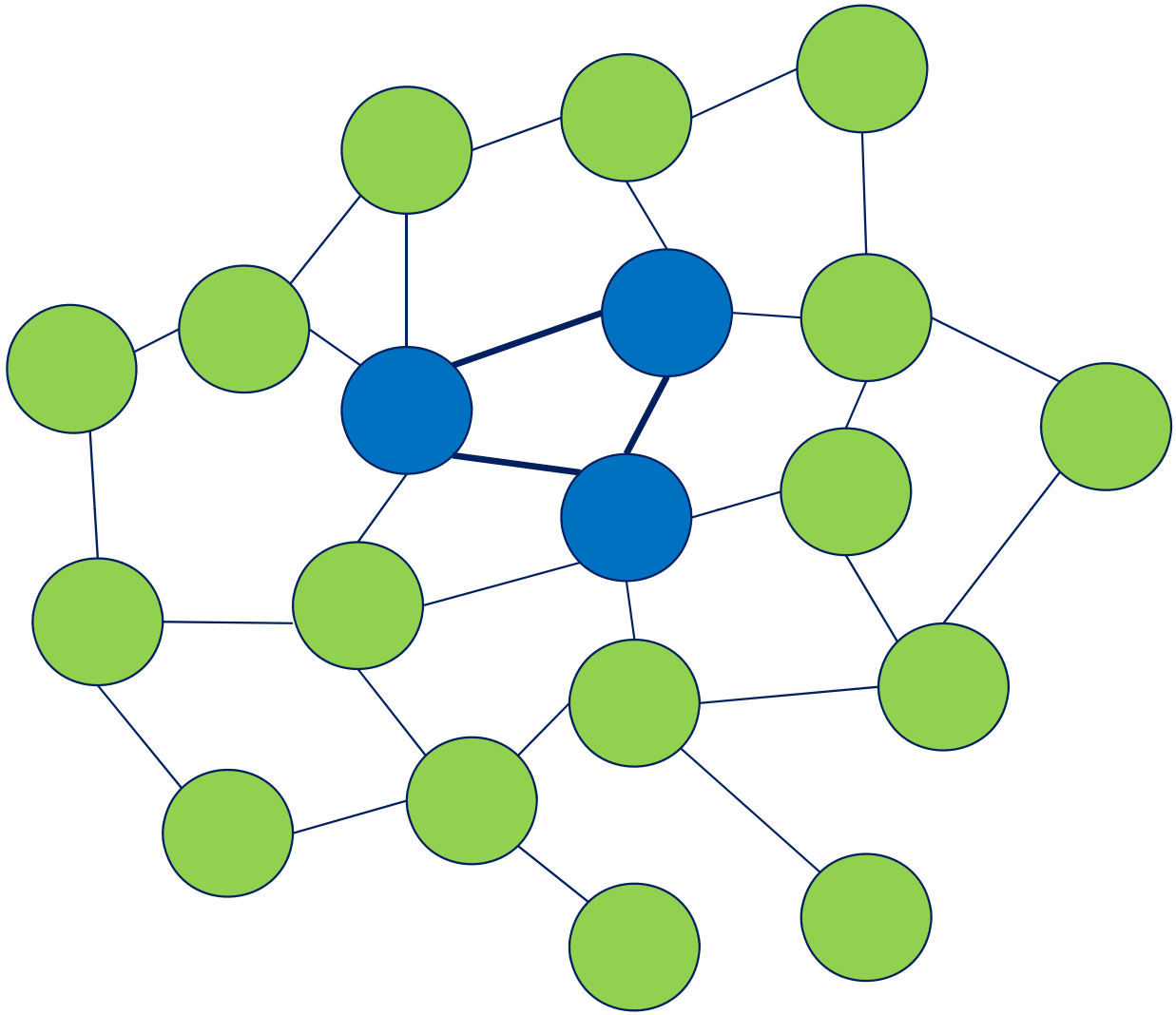
$$p(x, y|z) = p(x|z)p(y|z)$$

It proves useful to represent conditional independences graphically.

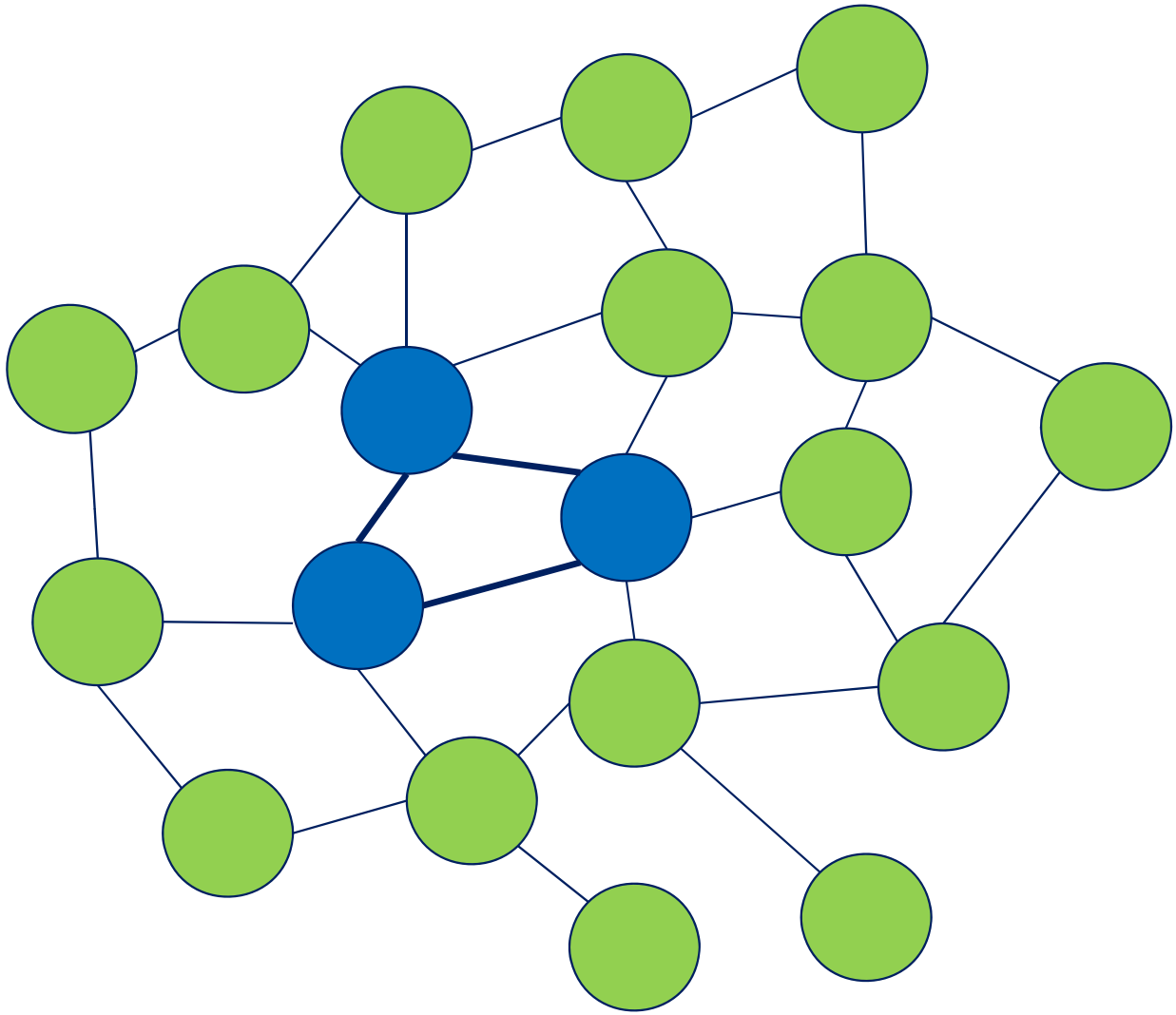
Markov random fields: the local Markov property



Markov random fields = Gibbs distributions



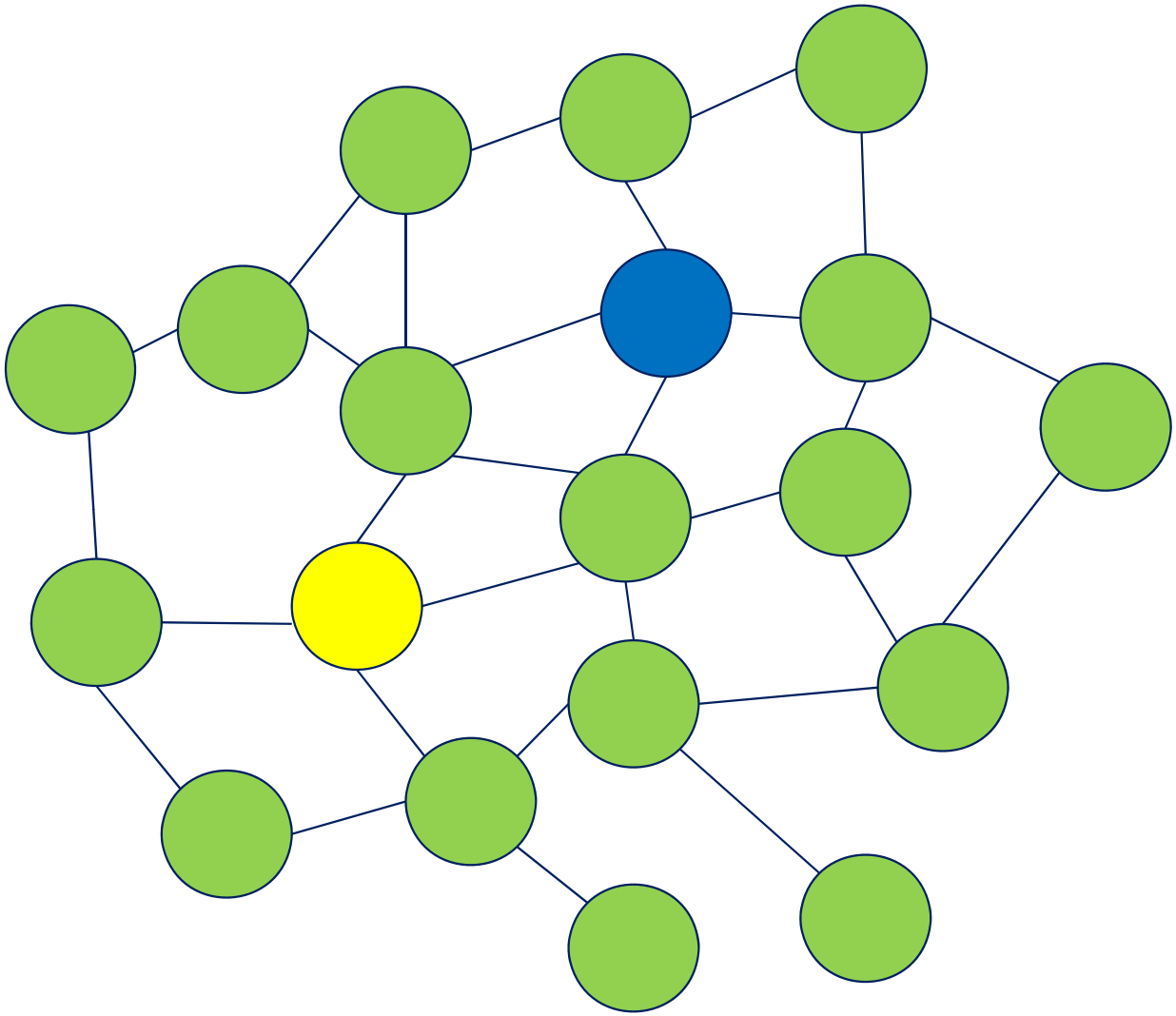
Markov random fields = Gibbs distributions



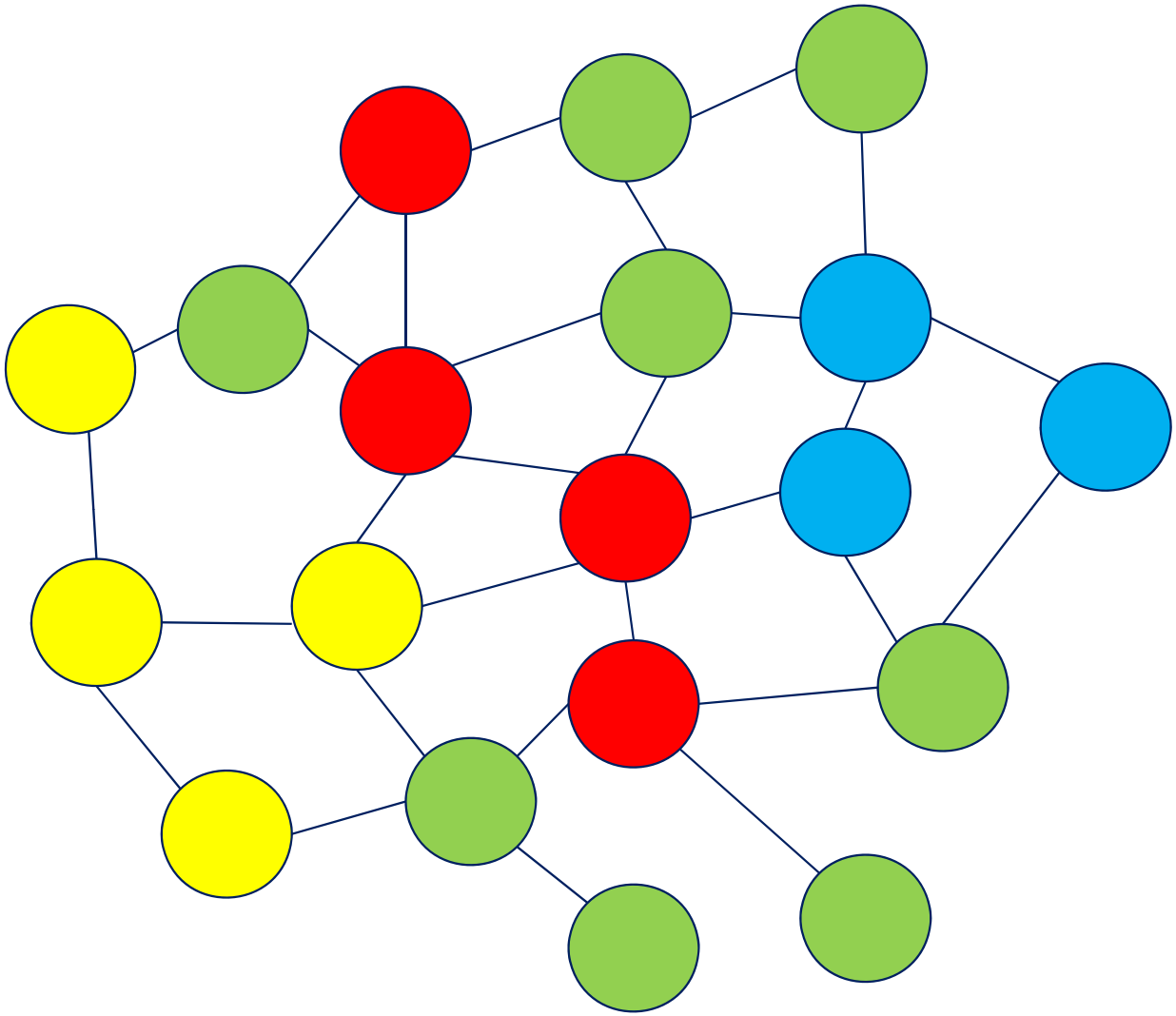
The Hammersley–Clifford theorem

The result that Markov random fields coincided with Gibbs distributions, under certain conditions, was known as the Hammersley–Clifford theorem (e.g. Besag, 1974).

Pairwise Markov property



Global Markov property



The Hammersley–Clifford theorem

The result that Markov random fields coincided with Gibbs distributions, under certain conditions, was known as the Hammersley–Clifford theorem (e.g. Besag, 1974).

Many years later, the theorem was superseded by a more complete understanding of Markov properties in undirected graphical models: we can distinguish Global, Local and Pairwise Markov properties, and relate all these to the Factorisation property of Gibbs distributions; in general

$$F \Longrightarrow G \Longrightarrow L \Longrightarrow P$$

Under an additional condition implied by positivity of the joint density, G, L and P are all equivalent, and if the density is also continuous, F is equivalent too.

The Hammersley–Clifford theorem

The result that Markov random fields coincided with Gibbs distributions, under certain conditions, was known as the Hammersley–Clifford theorem (e.g. Besag, 1974).

Many years later, the theorem was superseded by a more complete understanding of Markov properties in undirected graphical models: we can distinguish Global, Local and Pairwise Markov properties, and relate all these to the Factorisation property of Gibbs distributions; in general

$$F \implies G \implies L \implies P$$

Under an additional condition implied by positivity of the joint density, G, L and P are all equivalent, and if the density is also continuous, F is equivalent too.

The Hammersley–Clifford theorem

The result that Markov random fields coincided with Gibbs distributions, under certain conditions, was known as the Hammersley–Clifford theorem (e.g. Besag, 1974).

Many years later, the theorem was superseded by a more complete understanding of Markov properties in undirected graphical models: we can distinguish Global, Local and Pairwise Markov properties, and relate all these to the Factorisation property of Gibbs distributions; in general

$$F \implies G \implies L \implies P$$

Under an additional condition implied by positivity of the joint density, G, L and P are all equivalent, and if the density is also continuous, F is equivalent too.

Graphical models

The conditional independence graph $\mathcal{G}$ of a multivariate distribution (for a random vector X, say) tells us much about the structure of the distribution. $\mathcal{G} = (V, E)$ where the vertices V index the components of X, and there is an (undirected) edge between vertices i and j, written $i \sim j$

$$\text{unless} \quad X_i \perp\!\!\!\perp X_j \mid X_{V \setminus \{i,j\}}$$

Under the positivity condition, global and local Markov properties also hold.

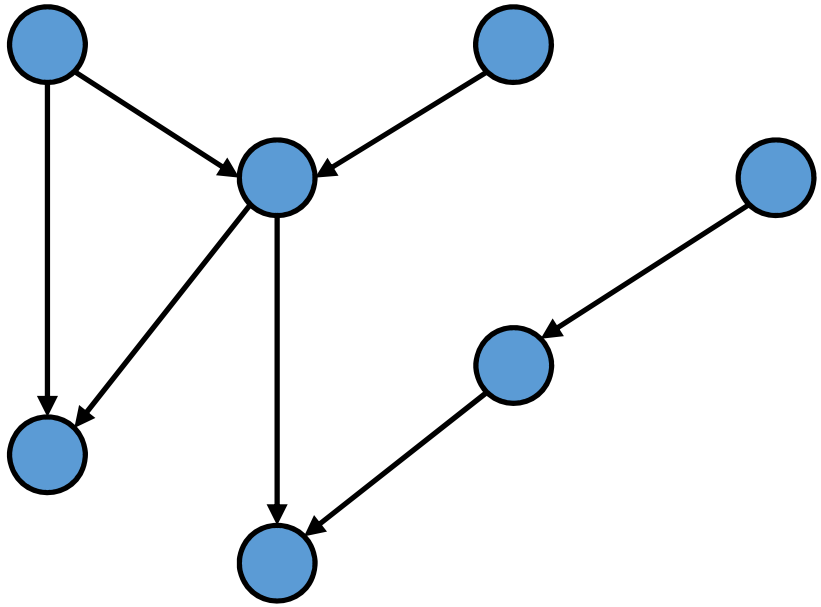
Graphical models

The conditional independence graph $\mathcal{G}$ of a multivariate distribution (for a random vector X, say) tells us much about the structure of the distribution. $\mathcal{G} = (V, E)$ where the vertices V index the components of X, and there is an (undirected) edge between vertices i and j, written $i \sim j$

$$\text{unless} \quad X_i \perp\!\!\!\perp X_j \mid X_{V \setminus \{i,j\}}$$

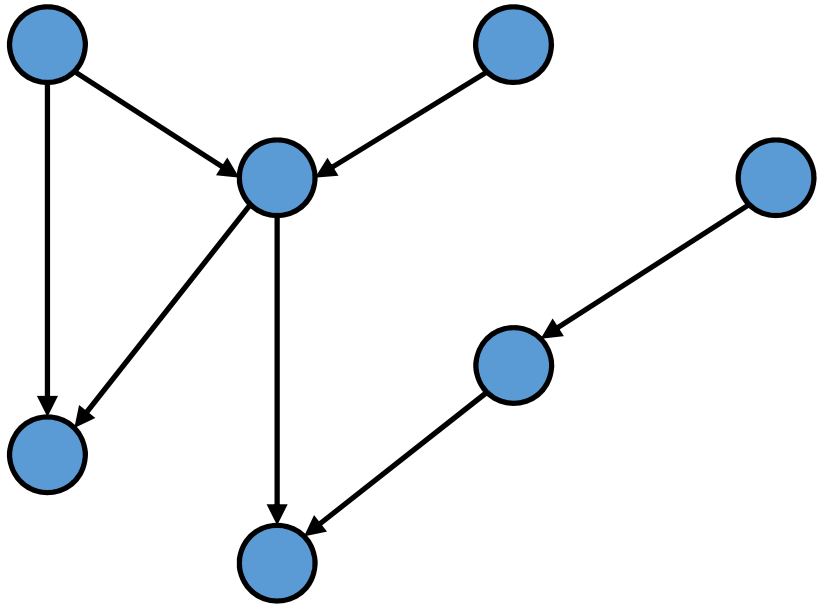
Under the positivity condition, global and local Markov properties also hold.

DAGs



A directed acyclic graph (DAG) is a directed graph in which there are no directed loops. Equivalently, it is a directed graph in which edges from a node can only go to a higher-numbered node.

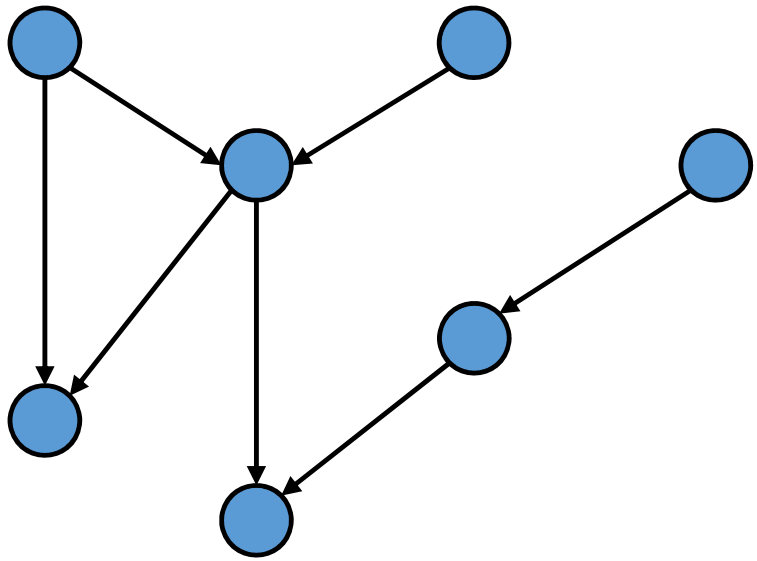
DAGs



A directed acyclic graph (DAG) is a directed graph in which there are no directed loops. Equivalently, it is a directed graph in which edges from a node can only go to a higher-numbered node.

Markov properties of DAGs

The directed local Markov property:

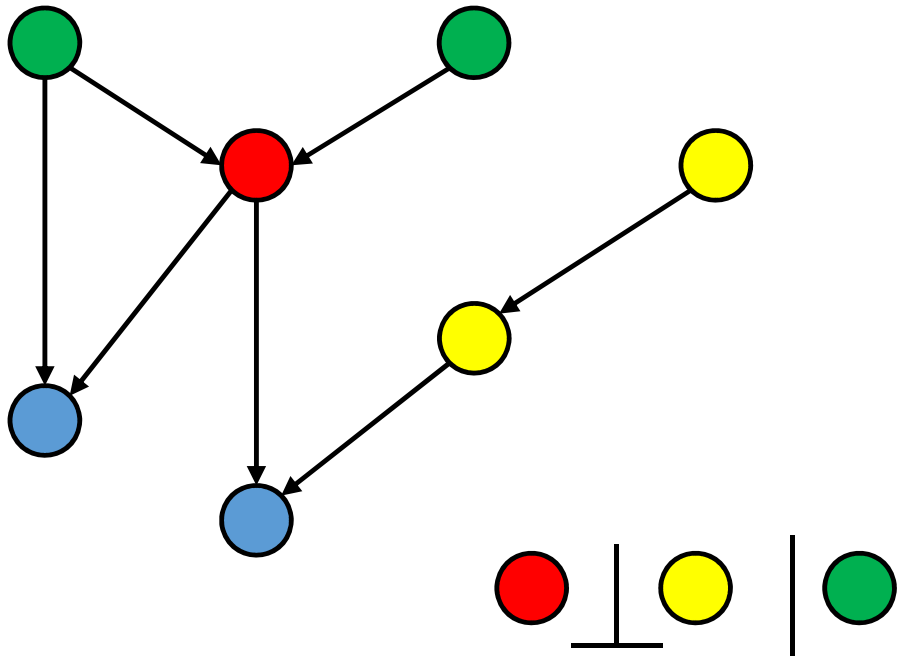


Variables are independent of their non-descendants, given their parents. There are also directed pairwise and global properties – the latter again involving separation. The directed factorisation property can be written

$$p(X) = \prod_{v \in V} p(X_v | X_{\mathrm{pa}(v)})$$

Markov properties of DAGs

The directed local Markov property:



Variables are independent of their non-descendants, given their parents. There are also directed pairwise and global properties – the latter again involving separation. The directed factorisation property can be written

$$p(X) = \prod_{v \in V} p(X_v | X_{\mathrm{pa}(v)})$$

Roles for graphs in statistics

- Visualisation
- Modelling
- Inference - understanding, discovery of structure
- Algorithms
- ...

Graphs driving algorithms

- MCMC algorithms: Gibbs and Metropolis–Hastings
- INLA
- SMC
- Probability propagation
- Other message-passing algorithms

Association and causality

As we all know, association and causality are different, an important distinction blurred by some who call all DAGs ‘causal’.

Graphical model ideas can play a key part in investigating causality, in particular elucidating exactly when it is legitimate to infer causal understanding from an observational study.

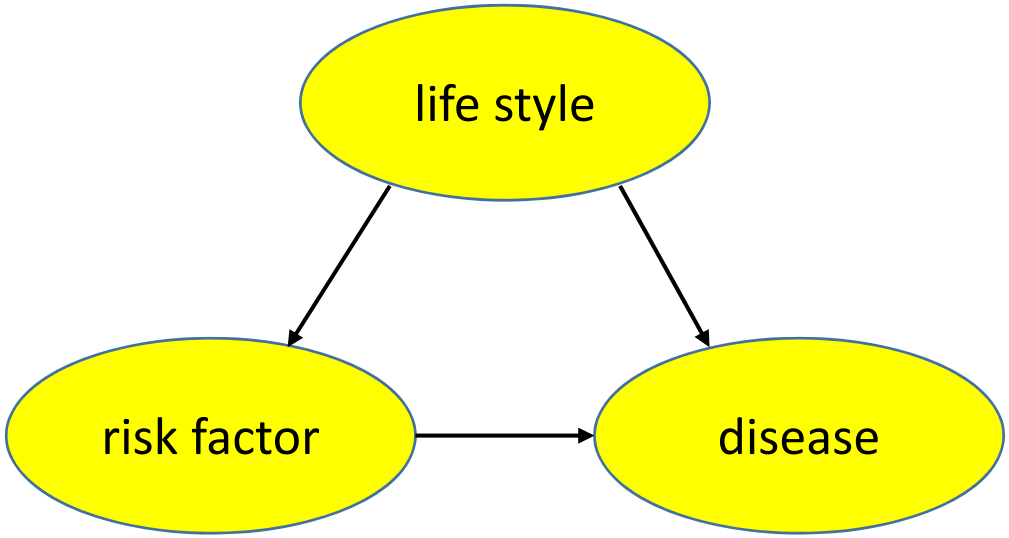
The key difficulty is the problem of confounding:

Association and causality

As we all know, association and causality are different, an important distinction blurred by some who call all DAGs ‘causal’.

Graphical model ideas can play a key part in investigating causality, in particular elucidating exactly when it is legitimate to infer causal understanding from an observational study.

The key difficulty is the problem of confounding:



Structural learning of undirected graphs

Given i.i.d. observations on X, we are often interested in inferring $\mathcal{G}$, the problem of structural learning (model selection, system identification). Why do we want to do this?

- $\mathcal{G}$ may be of direct interest (e.g. constructing pedigrees, gene networks)
- looking for parsimony (e.g. covariance selection for stability of estimation)
- just for visual understanding

This is a model selection problem; it entails search in a huge discrete model space: there are

$$2^{\binom{v}{2}}$$

graphs on v vertices.

Structural learning of undirected graphs

Given i.i.d. observations on X, we are often interested in inferring $\mathcal{G}$, the problem of structural learning (model selection, system identification). Why do we want to do this?

- $\mathcal{G}$ may be of direct interest (e.g. constructing pedigrees, gene networks)
- looking for parsimony (e.g. covariance selection for stability of estimation)
- just for visual understanding

This is a model selection problem; it entails search in a huge discrete model space: there are

$$2^{\binom{v}{2}}$$

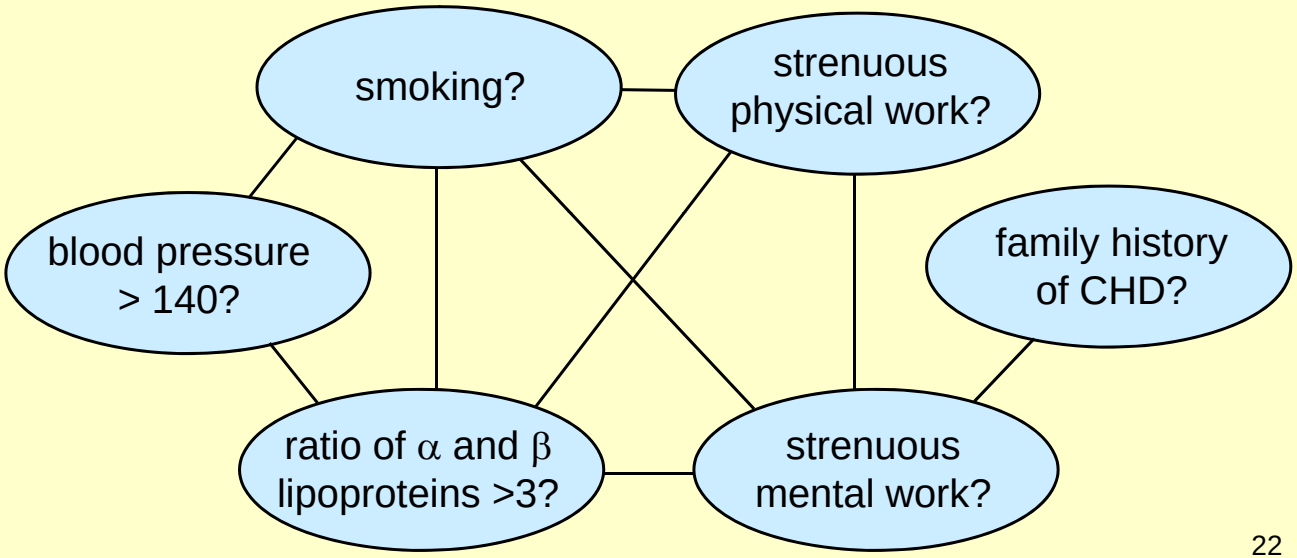
graphs on v vertices.

Contingency tables

Prognostic factors for coronary heart disease

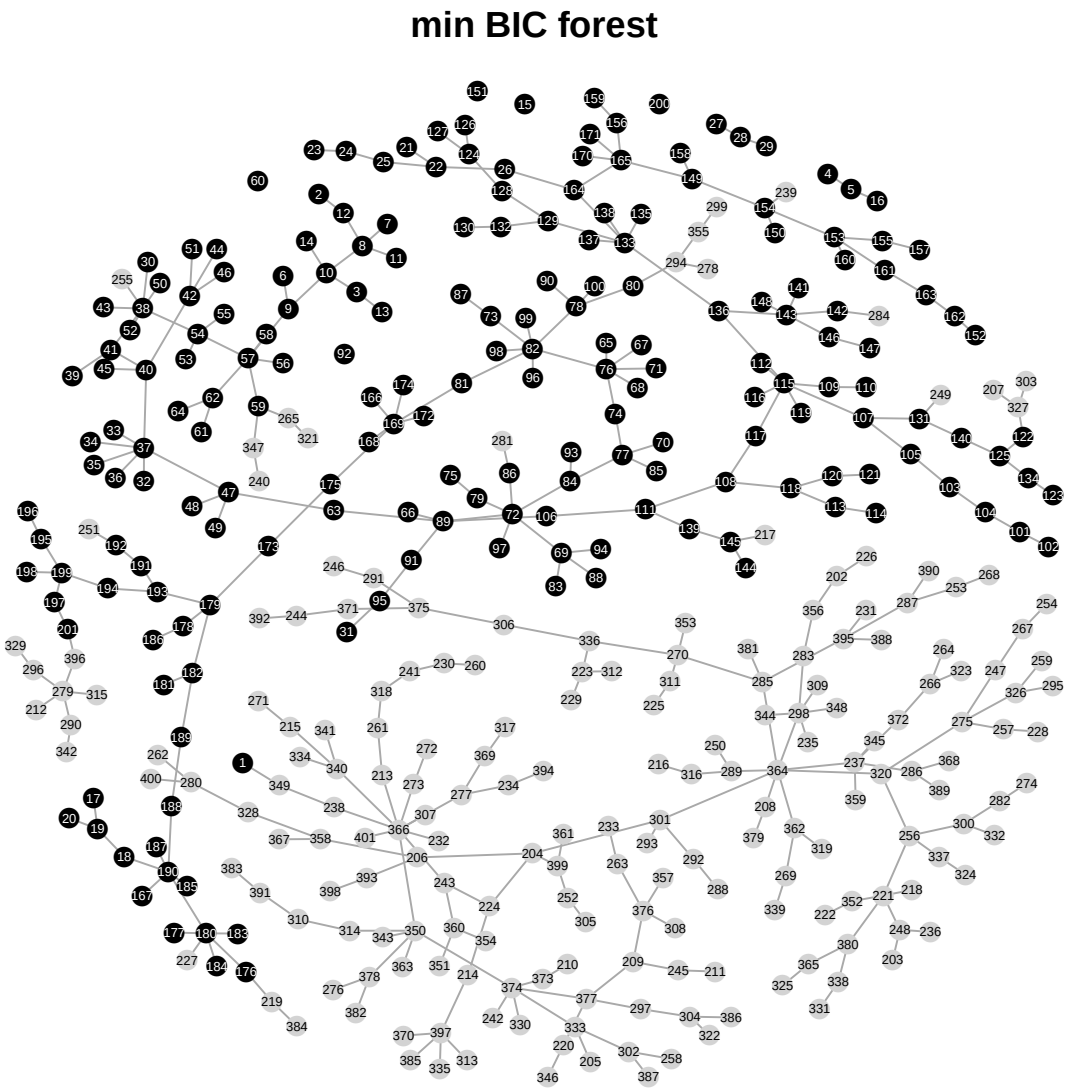
Analysis of a 2^6 contingency table

(Edwards & Havranek, *Biometrika,* 1985)



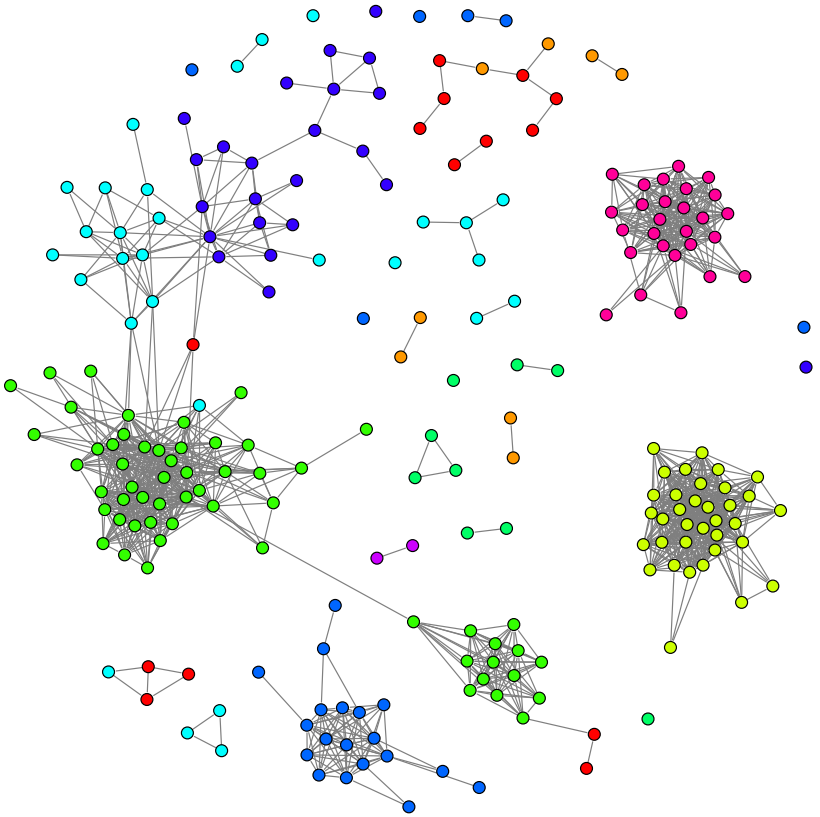
22

SNPs and gene expression

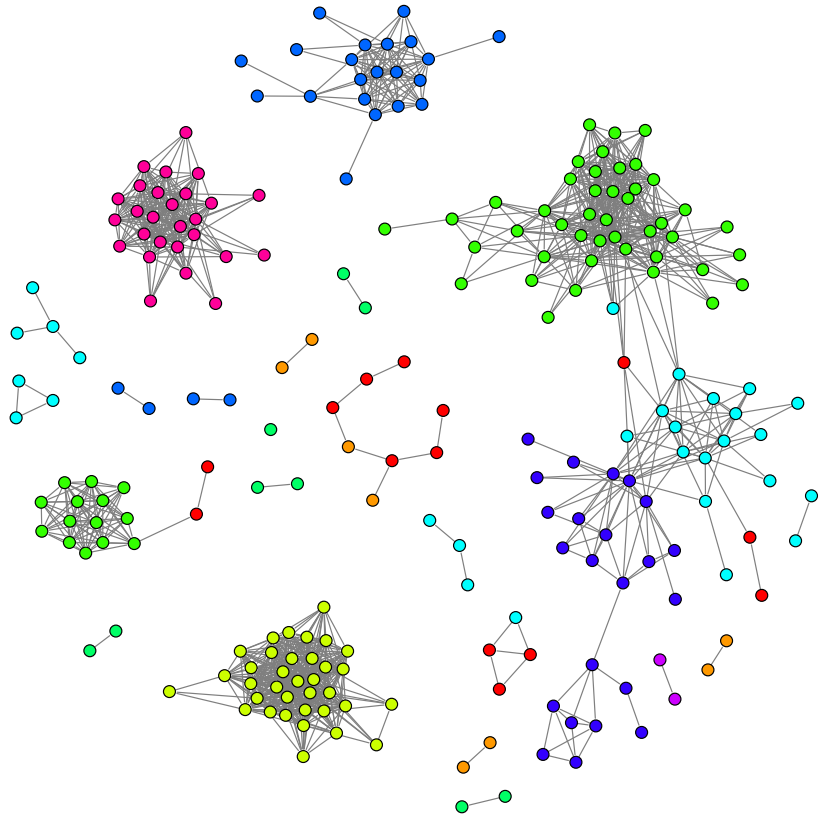


from Lauritzen (2012).

S&P 500 equity data



(a) glasso graph (1316 edges)

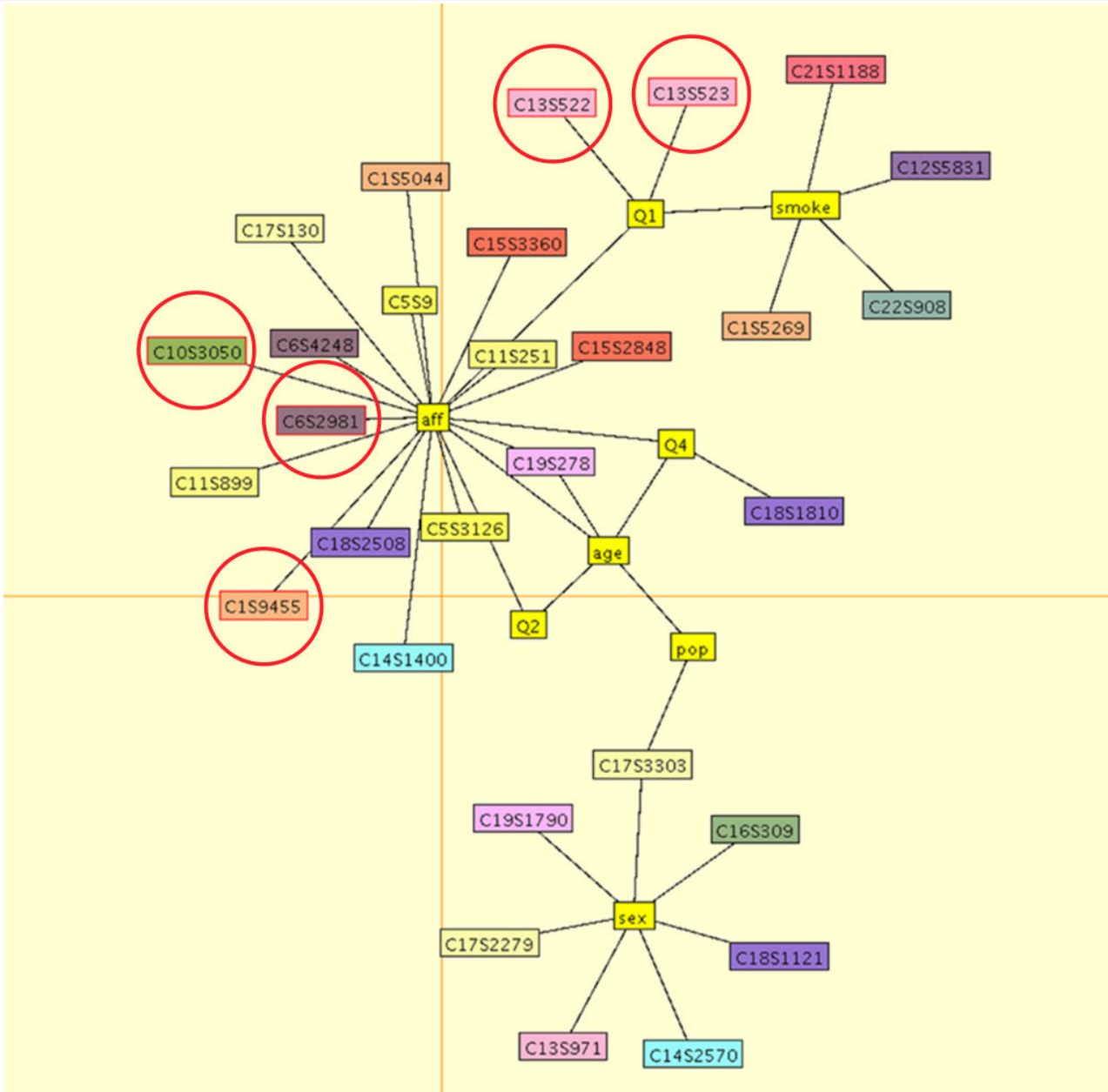


(b) nonparanormal graph (1316 edges)

from Lafferty, Liu, Wasserman (2012).

Genetic epidemiology

Graphical model fitted to contingency table relating disease status (aff), SNPs – with Linkage disequilibrium, covariates, and 4 quantitative traits. Abel & Thomas, GAW17.



What is structural learning really supposed to deliver?

Does the absence of an edge really signifies conditional independence, or simply insignificant dependence?

c.f. sparsity in regression modelling/variable selection

Structural learning

Today, we will emphasise Bayesian methods, and specifically those that in principle deliver exact or approximate posterior probabilities (for some or all graphs), not simply optimise a possibly-Bayesian objective function.

Except in very small problems, we typically restrict the space of graphs to be considered – e.g. to trees, forests, DAGs or decomposable graphs.

Structural learning

Today, we will emphasise Bayesian methods, and specifically those that in principle deliver exact or approximate posterior probabilities (for some or all graphs), not simply optimise a possibly-Bayesian objective function.

Except in very small problems, we typically restrict the space of graphs to be considered – e.g. to trees, forests, DAGs or decomposable graphs.

Decomposable graphical models

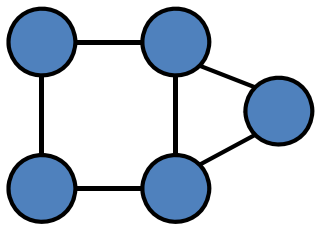
The case where $\mathcal{G}$ is decomposable has been much studied. Decomposability is a graph theory concept with statistical and computational implications.

Decomposable graphs are also known as triangulated or chordal: a graph is decomposable if and only if it has no chordless k-cycles for $k \geq 4$.

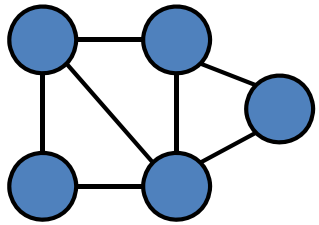
Decomposable graphical models

The case where $\mathcal{G}$ is decomposable has been much studied. Decomposability is a graph theory concept with statistical and computational implications.

Decomposable graphs are also known as triangulated or chordal: a graph is decomposable if and only if it has no chordless k-cycles for $k \geq 4$.



not decomposable



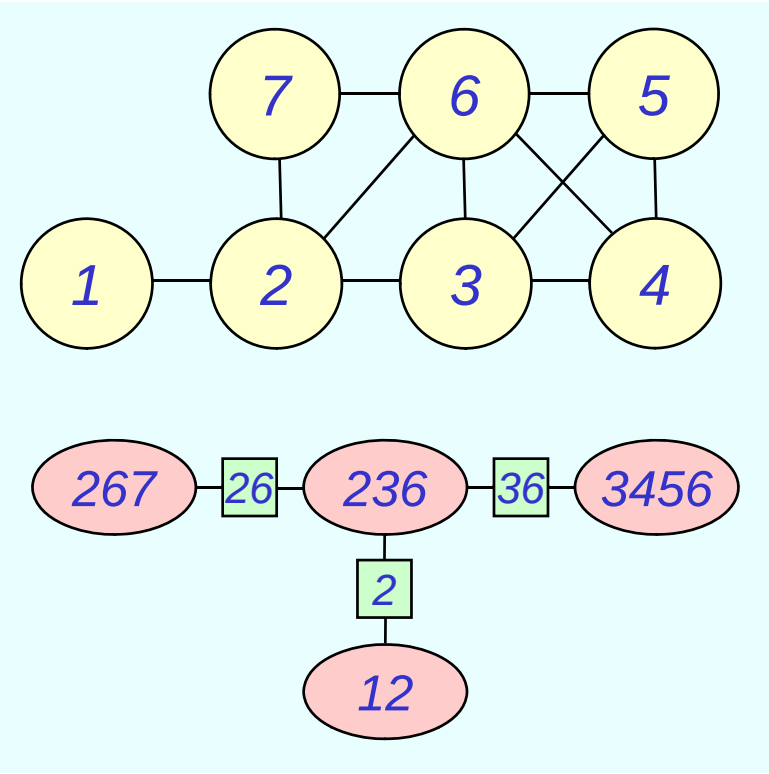
decomposable

Decomposability: junction trees

A graph is decomposable if and only if it has a junction tree representation.

A junction tree is a graph whose vertices are cliques (maximal complete subgraphs), with the property that the cliques containing any prescribed set of vertices forms a connected sub-tree.

We label the links of a junction tree with the separators, intersections of the adjacent cliques. There may be many junction trees for a given decomposable graph.

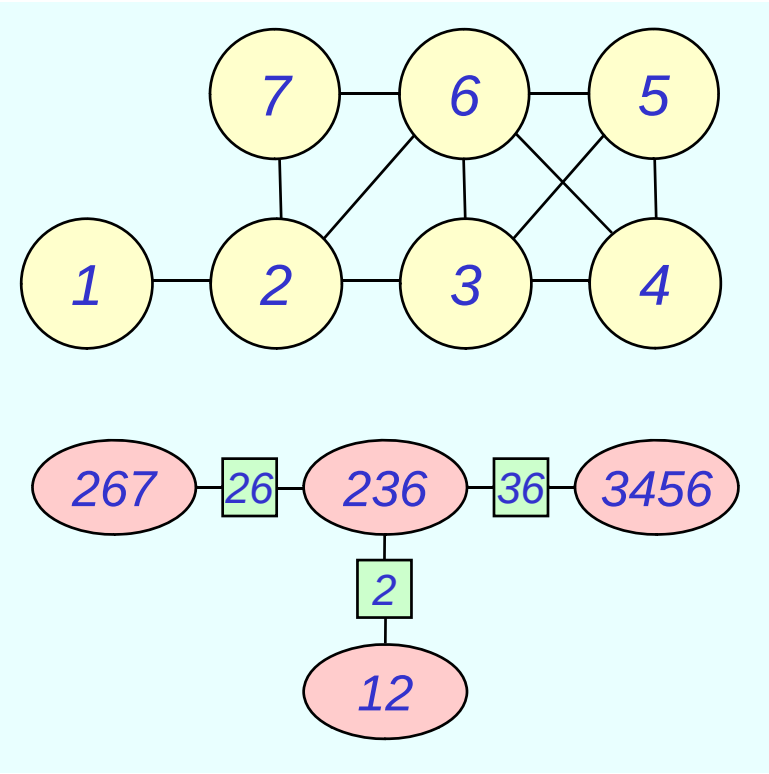


Decomposability: junction trees

A graph is decomposable if and only if it has a junction tree representation.

A junction tree is a graph whose vertices are cliques (maximal complete subgraphs), with the property that the cliques containing any prescribed set of vertices forms a connected sub-tree.

We label the links of a junction tree with the separators, intersections of the adjacent cliques. There may be many junction trees for a given decomposable graph.



Probabilistic significance of decomposability

If the distribution of a random vector X has a decomposable conditional independence graph, then it has a remarkable representation in terms of (often low-dimensional) marginals:

$$p(X) = \frac{\prod_{C \in \mathcal{C}} p(X_C)}{\prod_{S \in \mathcal{S}} p(X_S)}$$

This is the ultimate generalisation of the fact that for an ordinary Markov chain

$$p(X) = p(X_0) \prod_{i=1}^{N} p(X_i | X_{i-1}) = \frac{\prod_{i=1}^{N} p(X_{\{i-1,i\}})}{\prod_{i=2}^{N-1} p(X_{i-1})}$$

For a general decomposable graph, the same kind of factorisation follows the branches of the junction tree.

Probabilistic significance of decomposability

If the distribution of a random vector X has a decomposable conditional independence graph, then it has a remarkable representation in terms of (often low-dimensional) marginals:

$$p(X) = \frac{\prod_{C \in \mathcal{C}} p(X_C)}{\prod_{S \in \mathcal{S}} p(X_S)}$$

This is the ultimate generalisation of the fact that for an ordinary Markov chain

$$p(X) = p(X_0) \prod_{i=1}^{N} p(X_i | X_{i-1}) = \frac{\prod_{i=1}^{N} p(X_{\{i-1,i\}})}{\prod_{i=2}^{N-1} p(X_{i-1})}$$

For a general decomposable graph, the same kind of factorisation follows the branches of the junction tree.

Computational significance of decomposability

There are many consequences for computing with distributions on decomposable graphs, including junction tree algorithms (message passing/probability propagation) for Bayes nets (discrete graphical models).

Statistical significance of decomposability

Explicit Maximum likelihood estimates and exact tests for conditional independence for contingency tables and multivariate Gaussian distributions on decomposable graphs.

Clique–separator factorisation

$$p(X) = \frac{\prod_{C \in \mathcal{C}} p(X_C)}{\prod_{S \in \mathcal{S}} p(X_S)}$$

yields dramatic speed-ups in structural learning.

Dawid & Lauritzen's hyper-Markov laws - a framework for the construction of consistent prior distributions respecting the graphical structure.

Statistical significance of decomposability

Explicit Maximum likelihood estimates and exact tests for conditional independence for contingency tables and multivariate Gaussian distributions on decomposable graphs.

Clique–separator factorisation

$$p(X) = \frac{\prod_{C \in \mathcal{C}} p(X_C)}{\prod_{S \in \mathcal{S}} p(X_S)}$$

yields dramatic speed-ups in structural learning.

Dawid & Lauritzen's hyper-Markov laws - a framework for the construction of consistent prior distributions respecting the graphical structure.

Statistical significance of decomposability

Explicit Maximum likelihood estimates and exact tests for conditional independence for contingency tables and multivariate Gaussian distributions on decomposable graphs.

Clique–separator factorisation

$$p(X) = \frac{\prod_{C \in \mathcal{C}} p(X_C)}{\prod_{S \in \mathcal{S}} p(X_S)}$$

yields dramatic speed-ups in structural learning.

Dawid & Lauritzen's hyper-Markov laws - a framework for the construction of consistent prior distributions respecting the graphical structure.

How restrictive is decomposability?

How many graphs are decomposable?

There are $2^{\binom{v}{2}}$ graphs altogether on v vertices.

For $v \leq 3$ vertices, all are decomposable
for 4 vertices, $61/64$
for 6, $\approx 55\%$
for 8, $\approx 12\%$.

How restrictive is decomposability?

How many graphs are decomposable?

There are $2^{\binom{v}{2}}$ graphs altogether on v vertices.

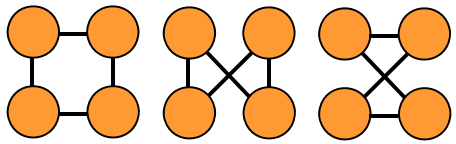
For $v \leq 3$ vertices, all are decomposable
for 4 vertices, $61/64$
for 6, $\approx 55\%$
for 8, $\approx 12\%$.

How restrictive is decomposability?

How many graphs are decomposable?

There are $2^{\binom{v}{2}}$ graphs altogether on v vertices.

For $v \leq 3$ vertices, all are decomposable
for 4 vertices, $61/64$
for 6, $\approx 55\%$
for 8, $\approx 12\%$.

The 3 non-decomposable 4-vertex graphs: 

Does that matter?

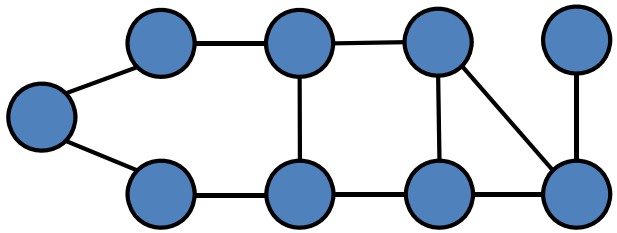
There is no reason why Nature should be kind enough to give us data from graphical models that are decomposable. . .

But given any (undirected) graphical model, we can add ('fill in') edges to make the graph decomposable.

Does that matter?

There is no reason why Nature should be kind enough to give us data from graphical models that are decomposable. . .

But given any (undirected) graphical model, we can add ('fill in') edges to make the graph decomposable.

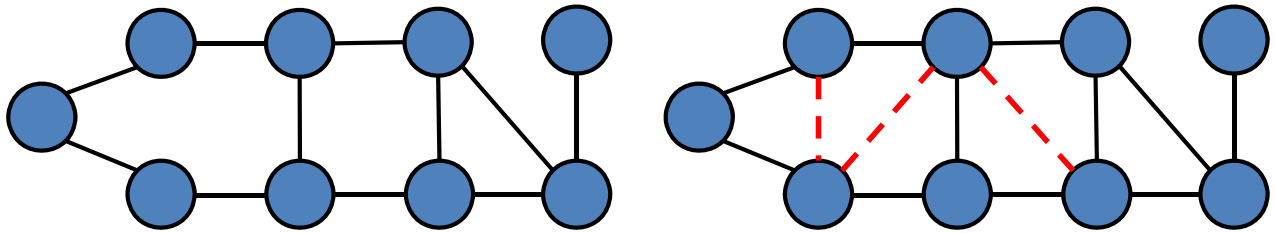


not decomposable

Does that matter?

There is no reason why Nature should be kind enough to give us data from graphical models that are decomposable. . .

But given any (undirected) graphical model, we can add ('fill in') edges to make the graph decomposable.



not decomposable

decomposable

So long as our model for the data, given the graph $\mathcal{G}$, allows arbitrarily small interactions, we will lose little by assuming decomposability – we will merely tend to infer (hopefully, slightly) more complicated graphs than necessary.

Bayesian model determination with non-decomposable graphs

What happens if the true graph is not decomposable, but you conduct Bayesian structural learning assuming it is?

Fitch, Jones and Massam (2014, *Bayesian Analysis*) show that (for the 0-mean Gaussian case, HIW prior on the concentration matrix), asymptotically,

- The posterior will converge to graphical structures that are minimal triangulations of the true graph.
- The marginal log likelihood ratio comparing different minimal triangulations is stochastically bounded and appears to remain data dependent regardless of the sample size.
- The covariance matrices corresponding to the different minimal triangulations are essentially equivalent, so model averaging is of minimal benefit.

Bayesian model determination with non-decomposable graphs

What happens if the true graph is not decomposable, but you conduct Bayesian structural learning assuming it is?

Fitch, Jones and Massam (2014, *Bayesian Analysis*) show that (for the 0-mean Gaussian case, HIW prior on the concentration matrix), asymptotically,

- The posterior will converge to graphical structures that are minimal triangulations of the true graph.
- The marginal log likelihood ratio comparing different minimal triangulations is stochastically bounded and appears to remain data dependent regardless of the sample size.
- The covariance matrices corresponding to the different minimal triangulations are essentially equivalent, so model averaging is of minimal benefit.

Bayesian model determination with non-decomposable graphs

What happens if the true graph is not decomposable, but you conduct Bayesian structural learning assuming it is?

Fitch, Jones and Massam (2014, *Bayesian Analysis*) show that (for the 0-mean Gaussian case, HIW prior on the concentration matrix), asymptotically,

- The posterior will converge to graphical structures that are minimal triangulations of the true graph.
- The marginal log likelihood ratio comparing different minimal triangulations is stochastically bounded and appears to remain data dependent regardless of the sample size.
- The covariance matrices corresponding to the different minimal triangulations are essentially equivalent, so model averaging is of minimal benefit.

Bayesian model determination with non-decomposable graphs

What happens if the true graph is not decomposable, but you conduct Bayesian structural learning assuming it is?

Fitch, Jones and Massam (2014, *Bayesian Analysis*) show that (for the 0-mean Gaussian case, HIW prior on the concentration matrix), asymptotically,

- The posterior will converge to graphical structures that are minimal triangulations of the true graph.
- The marginal log likelihood ratio comparing different minimal triangulations is stochastically bounded and appears to remain data dependent regardless of the sample size.
- The covariance matrices corresponding to the different minimal triangulations are essentially equivalent, so model averaging is of minimal benefit.

And assuming decomposability has tremendous advantages....

- Computational advantages in fitting the model
- Evaluating the fit
- Prediction
- Sampling data from fitted model

Bayesian graphical model determination

Given n i.i.d. samples $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ from a multivariate distribution on $\mathcal{R}^v$ parameterised by the graph $\mathcal{G}$ and parameters θ, the usual formulation takes the form

$$p(\mathcal{G}, \theta, \mathbf{X}) = \pi(\mathcal{G}) p(\theta|\mathcal{G}) p(\mathbf{X}|\mathcal{G}, \theta)$$

and we perform joint structural/quantitative learning by computing the posterior $p(\mathcal{G}, \theta|\mathbf{X}) \propto p(\mathcal{G}, \theta, \mathbf{X})$.

What prior on $\mathcal{G}$? And on $\theta|\mathcal{G}$?

Bayesian graphical model determination

Given n i.i.d. samples $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ from a multivariate distribution on $\mathcal{R}^v$ parameterised by the graph $\mathcal{G}$ and parameters θ, the usual formulation takes the form

$$p(\mathcal{G}, \theta, \mathbf{X}) = \pi(\mathcal{G}) p(\theta|\mathcal{G}) p(\mathbf{X}|\mathcal{G}, \theta)$$

and we perform joint structural/quantitative learning by computing the posterior $p(\mathcal{G}, \theta|\mathbf{X}) \propto p(\mathcal{G}, \theta, \mathbf{X})$.

What prior on $\mathcal{G}$? And on $\theta|\mathcal{G}$?

Priors on decomposable graphs

Many authors simply take a prior uniform over all valid graphs (makes sense even if we can't count them!), or a binomial model conditioned on decomposability.

Armstrong et al, 2009, advocate specifying prior over the size of the graph (which is then uniform conditional on size).

Is there a canonical (e.g. conjugate) approach?

Priors on decomposable graphs

Many authors simply take a prior uniform over all valid graphs (makes sense even if we can't count them!), or a binomial model conditioned on decomposability.

Armstrong et al, 2009, advocate specifying prior over the size of the graph (which is then uniform conditional on size).

Is there a canonical (e.g. conjugate) approach?

Conjugate priors on decomposable graphs

Recall that in any decomposable graphical model the likelihood has the form

$$p(X|\mathcal{G}) = \frac{\prod_{C\in\mathcal{C}} p(X_C|\mathcal{G})}{\prod_{S\in\mathcal{S}} p(X_S|\mathcal{G})}$$

So any prior on the graph $\mathcal{G}$ that factorises similarly as a product over cliques divided by a product over separators will be conjugate.

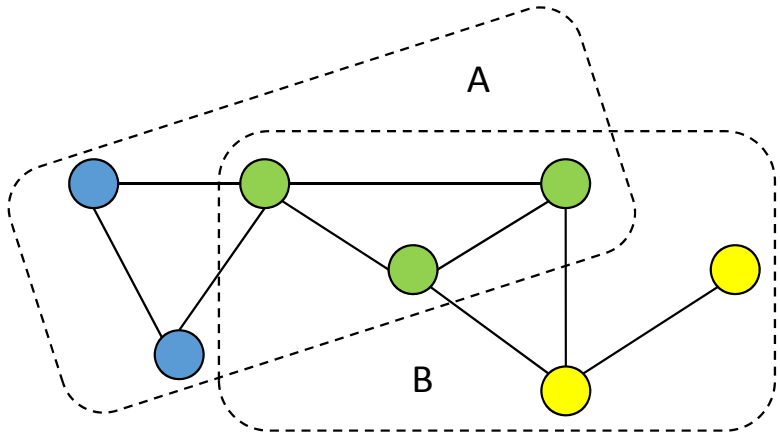
Byrne & Dawid's structural Markov property

A graph law $\pi(\mathcal{G})$ over the set $\mathfrak{U}$ of undirected decomposable graphs on V is *structurally Markov* (Byrne, 2011, Byrne & Dawid, 2015) if for any covering pair (A, B), we have :

$$\mathcal{G}_A \perp\!\!\!\perp \mathcal{G}_B \mid \{\mathcal{G} \in \mathfrak{U}(A, B)\} \quad [\pi],$$

where $\mathfrak{U}(A, B)$ is the set of decomposable graphs for which (A, B) is a decomposition.

- (A, B) is a covering pair if $A \cup B = V$
- (A, B) is a decomposition if $A \cap B$ is complete, and separates $A \setminus B$ and $B \setminus A$.



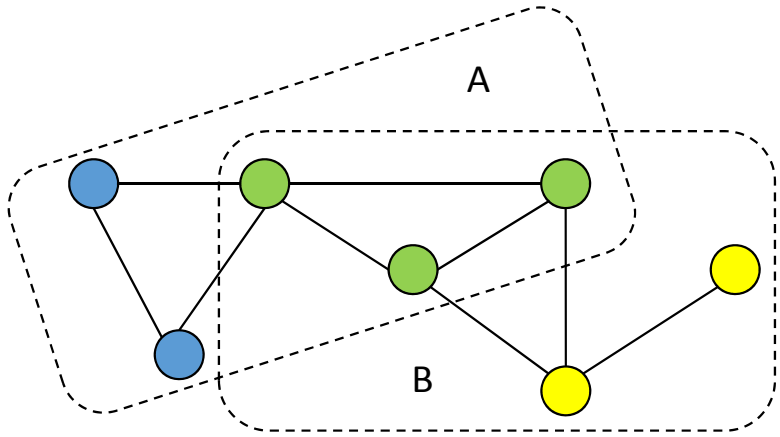
Byrne & Dawid's structural Markov property

A graph law $\pi(\mathcal{G})$ over the set $\mathfrak{U}$ of undirected decomposable graphs on V is *structurally Markov* (Byrne, 2011, Byrne & Dawid, 2015) if for any covering pair (A, B), we have :

$$\mathcal{G}_A \perp\!\!\!\perp \mathcal{G}_B \mid \{\mathcal{G} \in \mathfrak{U}(A, B)\} \quad [\pi],$$

where $\mathfrak{U}(A, B)$ is the set of decomposable graphs for which (A, B) is a decomposition.

- (A, B) is a covering pair if $A \cup B = V$
- (A, B) is a decomposition if $A \cap B$ is complete, and separates $A \setminus B$ and $B \setminus A$.



Byrne & Dawid's structural Markov property

A graph law $\pi(\mathcal{G})$ over the set $\mathfrak{U}$ of undirected decomposable graphs on V is *structurally Markov* (Byrne, 2011, Byrne & Dawid, 2015) if for any covering pair (A, B), we have :

$$\mathcal{G}_A \perp\!\!\!\perp \mathcal{G}_B \mid \{\mathcal{G} \in \mathfrak{U}(A, B)\} \quad [\pi],$$

where $\mathfrak{U}(A, B)$ is the set of decomposable graphs for which (A, B) is a decomposition.

Byrne & Dawid show that a graph law is structurally Markov if and only if has the form

$$\pi(\mathcal{G}) \propto \frac{\prod_{C \in \mathcal{C}} \phi_C}{\prod_{S \in \mathcal{S}} \phi_S}$$

where $\{\phi_A : A \subseteq V\}$ are arbitrary positive set-indexed parameters.

Byrne & Dawid's structural Markov property

A graph law $\pi(\mathcal{G})$ over the set $\mathfrak{U}$ of undirected decomposable graphs on V is *structurally Markov* (Byrne, 2011, Byrne & Dawid, 2015) if for any covering pair (A, B), we have :

$$\mathcal{G}_A \perp\!\!\!\perp \mathcal{G}_B \mid \{\mathcal{G} \in \mathfrak{U}(A, B)\} \quad [\pi],$$

where $\mathfrak{U}(A, B)$ is the set of decomposable graphs for which (A, B) is a decomposition.

Byrne & Dawid show that a graph law is structurally Markov if and only if has the form

$$\pi(\mathcal{G}) \propto \frac{\prod_{C \in \mathcal{C}} \phi_C}{\prod_{S \in \mathcal{S}} \phi_S}$$

where $\{\phi_A : A \subseteq V\}$ are arbitrary positive set-indexed parameters.

A new weak structural Markov property

A graph law $\pi(\mathcal{G})$ over the set $\mathfrak{U}$ of undirected decomposable graphs on V is *weakly structurally Markov (WSM)* if for any covering pair (A, B), we have :

$$\mathcal{G}_A \perp\!\!\!\perp \mathcal{G}_B \mid \{\mathcal{G} \in \mathfrak{U}^A(A, B)\} \quad [\pi],$$

where $\mathfrak{U}^A(A, B)$ is the set of decomposable graphs for which (A, B) is a decomposition, and $A \cap B$ is a clique in $\mathcal{G}_A$.

This places fewer conditional independence conditions on π, so potentially corresponds to a richer class of graph priors – but we will see that we can still say something concrete about the form of these laws.

A new weak structural Markov property

A graph law $\pi(\mathcal{G})$ over the set $\mathfrak{U}$ of undirected decomposable graphs on V is *weakly structurally Markov (WSM)* if for any covering pair (A, B), we have :

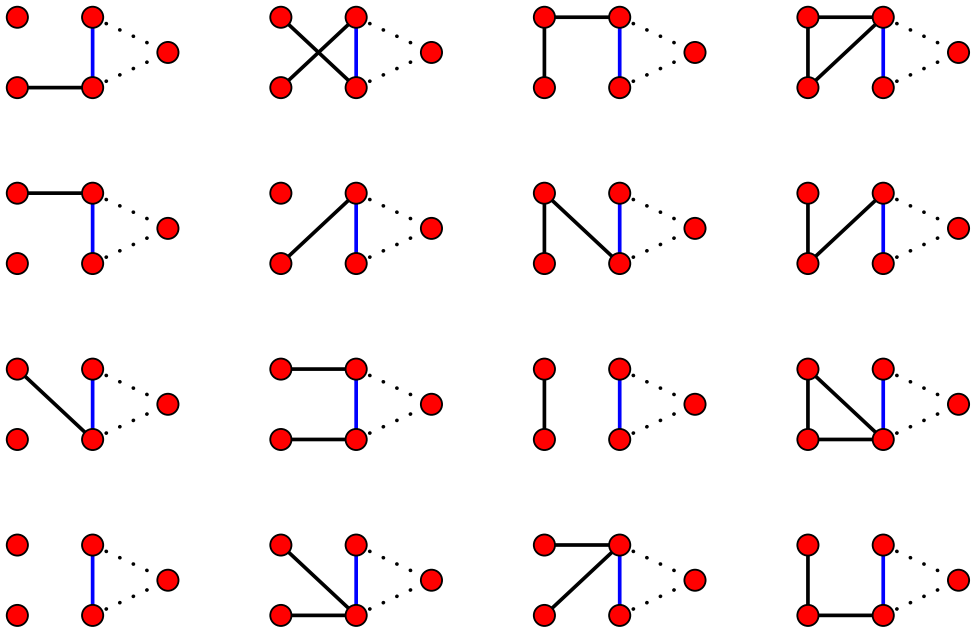
$$\mathcal{G}_A \perp\!\!\!\perp \mathcal{G}_B \mid \{\mathcal{G} \in \mathfrak{U}^A(A, B)\} \quad [\pi],$$

where $\mathfrak{U}^A(A, B)$ is the set of decomposable graphs for which (A, B) is a decomposition, and $A \cap B$ is a clique in $\mathcal{G}_A$.

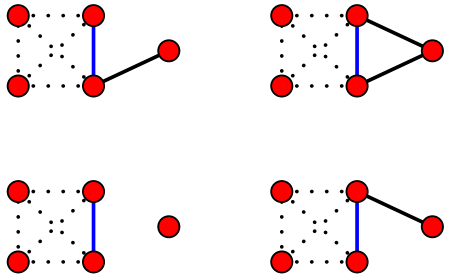
This places fewer conditional independence conditions on π, so potentially corresponds to a richer class of graph priors – but we will see that we can still say something concrete about the form of these laws.

A weak structural Markov property

16 possibilities for $\mathcal{G}_A$
(if $A \cap B$ remains a clique in $\mathcal{G}_A$)



4 possibilities for $\mathcal{G}_B$



$$\mathcal{G}_A \perp\!\!\!\perp \mathcal{G}_B \mid \{\mathcal{G} \in \mathfrak{U}^A(A,B)\} \quad [\pi],$$

Clique–separator factorisation graph laws

We can show that a graph law is weakly structurally Markov if and only if has the form

$$\pi(\mathcal{G}) \propto \frac{\prod_{C \in \mathcal{C}} \phi_C}{\prod_{S \in \mathcal{S}} \psi_S}$$

where $\{\phi_A : A \subseteq V\}, \{\psi_A : A \subseteq V\}$ are arbitrary positive set-indexed parameters.

This more general form allows valuable extra flexibility in prior specification; this class of priors has also been studied by Bornn and Caron (2011).

Clique–separator factorisation graph laws

We can show that a graph law is weakly structurally Markov if and only if has the form

$$\pi(\mathcal{G}) \propto \frac{\prod_{C \in \mathcal{C}} \phi_C}{\prod_{S \in \mathcal{S}} \psi_S}$$

where $\{\phi_A : A \subseteq V\}, \{\psi_A : A \subseteq V\}$ are arbitrary positive set-indexed parameters.

This more general form allows valuable extra flexibility in prior specification; this class of priors has also been studied by Bornn and Caron (2011).

Posterior using a prior with the weak structural Markov property

The posterior for $\mathcal{G}$ is

$$p(\mathcal{G}|X) \propto \frac{\prod_{C\in\mathcal{C}}[\phi_C p(X_C|\mathcal{G})]}{\prod_{S\in\mathcal{S}}[\psi_S p(X_S|\mathcal{G})]}$$

that is, a CSF law with parameters $\phi_A p(X_A|\mathcal{G})$ and $\psi_A p(X_A|\mathcal{G})$.

Bayesian decomposable graphical model determination

For trees, there are explicit finite algorithms for computing MAP estimates; also perfect simulation is possible for random spanning trees, so a full Bayesian analysis can be conducted.

It would be interesting to find a way to extend these ideas to decomposable graphs, but that has not so far been successful.

Bayesian decomposable graphical model determination

For decomposable graphs, joint structural/quantitative learning currently requires MCMC sampling of the posterior $p(\mathcal{G}, \theta|\mathbf{X}) \propto p(\mathcal{G}, \theta, \mathbf{X})$: this means running a Markov chain whose states have the form $(\mathcal{G}, \theta)$ – a graph and a vector of parameters.

This chain is constructed to have equilibrium distribution $p(\mathcal{G}, \theta|\mathbf{X})$ by ensuring that all moves have detailed balance with respect to this distribution, by using a Metropolis–Hastings sampler.

Pre-tests for maintaining decomposability

Conditions for maintaining decomposability in single-edge moves:

Frydenberg & Lauritzen Disconnecting x and y by removing an edge (x, y) from $\mathcal{G}$ will result in a decomposable graph if and only if x and y are contained in exactly one clique.

Giudici & Green Connecting x and y by adding an edge (x, y) to $\mathcal{G}$ will result in a decomposable graph if and only if x and y are contained in cliques that are adjacent in some junction tree of $\mathcal{G}$.

These extend to certain multiple-edge moves (completely connecting or disconnecting subsets of nodes).

These moves can be efficiently implemented using a junction-tree representation of the graph, since they make only local changes to the junction tree – but they can change its topology locally.

Pre-tests for maintaining decomposability

Conditions for maintaining decomposability in single-edge moves:

Frydenberg & Lauritzen Disconnecting x and y by removing an edge (x, y) from $\mathcal{G}$ will result in a decomposable graph if and only if x and y are contained in exactly one clique.

Giudici & Green Connecting x and y by adding an edge (x, y) to $\mathcal{G}$ will result in a decomposable graph if and only if x and y are contained in cliques that are adjacent in some junction tree of $\mathcal{G}$.

These extend to certain multiple-edge moves (completely connecting or disconnecting subsets of nodes).

These moves can be efficiently implemented using a junction-tree representation of the graph, since they make only local changes to the junction tree – but they can change its topology locally.

Pre-tests for maintaining decomposability

Conditions for maintaining decomposability in single-edge moves:

Frydenberg & Lauritzen Disconnecting x and y by removing an edge (x, y) from $\mathcal{G}$ will result in a decomposable graph if and only if x and y are contained in exactly one clique.

Giudici & Green Connecting x and y by adding an edge (x, y) to $\mathcal{G}$ will result in a decomposable graph if and only if x and y are contained in cliques that are adjacent in some junction tree of $\mathcal{G}$.

These extend to certain multiple-edge moves (completely connecting or disconnecting subsets of nodes).

These moves can be efficiently implemented using a junction-tree representation of the graph, since they make only local changes to the junction tree – but they can change its topology locally.

Pre-tests for maintaining decomposability

Conditions for maintaining decomposability in single-edge moves:

> Frydenberg & Lauritzen Disconnecting x and y by removing an edge (x, y) from $\mathcal{G}$ will result in a decomposable graph if and only if x and y are contained in exactly one clique.
>
> Giudici & Green Connecting x and y by adding an edge (x, y) to $\mathcal{G}$ will result in a decomposable graph if and only if x and y are contained in cliques that are adjacent in some junction tree of $\mathcal{G}$.

These extend to certain multiple-edge moves (completely connecting or disconnecting subsets of nodes).

These moves can be efficiently implemented using a junction-tree representation of the graph, since they make only local changes to the junction tree – but they can change its topology locally.

Using the junction tree as the state

We recently found a simple way to speed up sampling dramatically, by ruling out the need to change the topology of the junction tree – we do this by using directly the junction tree J as part of the model parameterisation, in place of the graph $\mathcal{G}$.

This means augmenting the model so that, conditional on $\mathcal{G}$, the junction tree J is a priori drawn uniformly from among all equivalent junction trees, thus replacing the prior $\pi(\mathcal{G})$ on decomposable graphs by

$$\widetilde{\pi}(J) = \frac{\pi(\mathcal{G}(J))}{\mu(\mathcal{G}(J))}$$

where $\mathcal{G}(J)$ is the decomposable graph determined by J and $\mu(\mathcal{G})$ is the number of equivalent junction trees representing $\mathcal{G}$.

Fortunately, we have an efficient local method for evaluating $\mu(\mathcal{G})$, and a local method for occasionally randomising over equivalent junction trees.

Using the junction tree as the state

We recently found a simple way to speed up sampling dramatically, by ruling out the need to change the topology of the junction tree – we do this by using directly the junction tree J as part of the model parameterisation, in place of the graph $\mathcal{G}$.

This means augmenting the model so that, conditional on $\mathcal{G}$, the junction tree J is a priori drawn uniformly from among all equivalent junction trees, thus replacing the prior $\pi(\mathcal{G})$ on decomposable graphs by

$$\widetilde{\pi}(J) = \frac{\pi(\mathcal{G}(J))}{\mu(\mathcal{G}(J))}$$

where $\mathcal{G}(J)$ is the decomposable graph determined by J and $\mu(\mathcal{G})$ is the number of equivalent junction trees representing $\mathcal{G}$.

Fortunately, we have an efficient local method for evaluating $\mu(\mathcal{G})$, and a local method for occasionally randomising over equivalent junction trees.

Using the junction tree as the state

We recently found a simple way to speed up sampling dramatically, by ruling out the need to change the topology of the junction tree – we do this by using directly the junction tree J as part of the model parameterisation, in place of the graph $\mathcal{G}$.

This means augmenting the model so that, conditional on $\mathcal{G}$, the junction tree J is a priori drawn uniformly from among all equivalent junction trees, thus replacing the prior $\pi(\mathcal{G})$ on decomposable graphs by

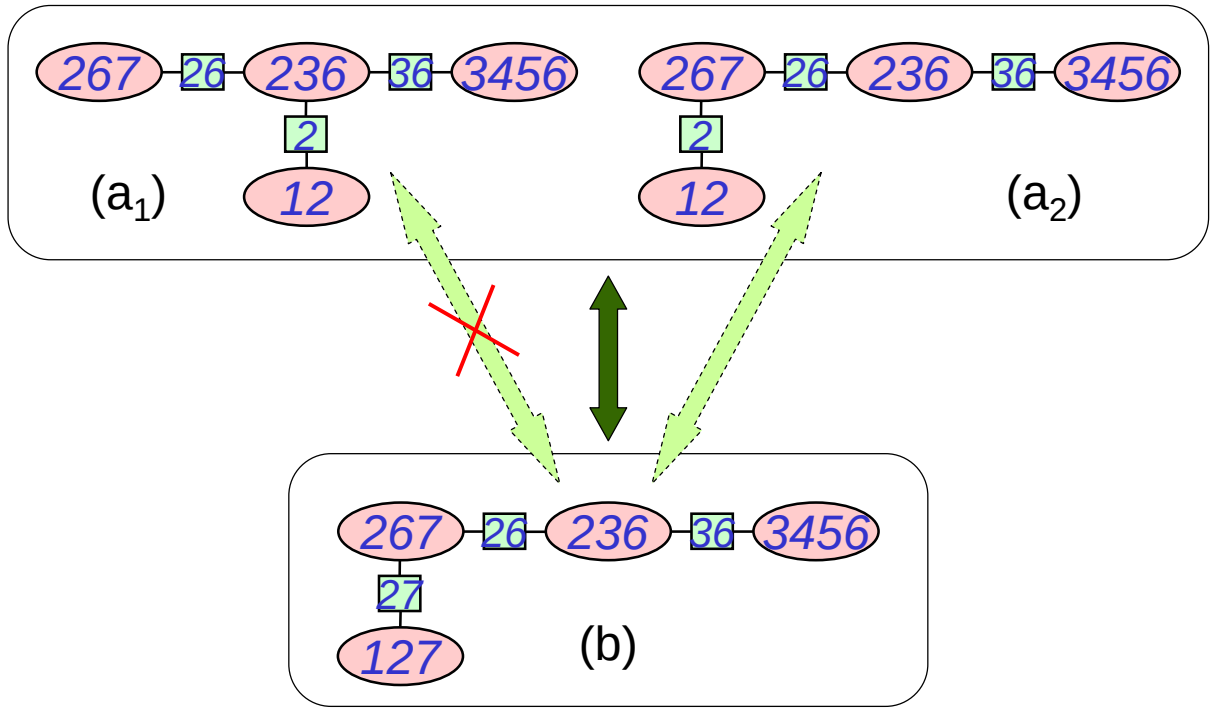
$$\widetilde{\pi}(J) = \frac{\pi(\mathcal{G}(J))}{\mu(\mathcal{G}(J))}$$

where $\mathcal{G}(J)$ is the decomposable graph determined by J and $\mu(\mathcal{G})$ is the number of equivalent junction trees representing $\mathcal{G}$.

Fortunately, we have an efficient local method for evaluating $\mu(\mathcal{G})$, and a local method for occasionally randomising over equivalent junction trees.

Using the junction tree as the state

Whether two decomposable graphs are adjacent in the junction tree representation depends on the choice of junction tree.



Sampling decomposable graphs for posterior simulation

- Pre-testing for maintenance of decomposability
- Simplification of likelihood ratios and prior ratios from properties of decomposability
- Using the junction tree as the state
- Multiple-edge moves, randomising over junction trees

allow effective sampling of moderate sized graphs (?50–100 vertices).

Other novel local moves have been proposed by Stingo and Marchetti (2015), working on a perfect ordering representation of the cliques of the graph, rather than the junction tree.

Sampling decomposable graphs for posterior simulation

- Pre-testing for maintenance of decomposability
- Simplification of likelihood ratios and prior ratios from properties of decomposability
- Using the junction tree as the state
- Multiple-edge moves, randomising over junction trees

allow effective sampling of moderate sized graphs (?50–100 vertices).

Other novel local moves have been proposed by Stingo and Marchetti (2015), working on a perfect ordering representation of the cliques of the graph, rather than the junction tree.

Updating parameters in decomposable graph models

Giudici and Green (1999) and other early work conducts joint structural and quantitative learning by sampling from the joint posterior of graph and parameters – usually a variable-dimension problem, necessitating reversible jump or equivalent trans-dimensional sampling.

To try to scale up to larger problems, almost all authors have used conjugate hyper-Markov priors for parameters, so that these can be integrated out. Metropolis ratios for graph updates then simplify into a locally-computable form.

Thus, hyper-inverse-Wishart priors are used for variance matrices in gaussian models (Giudici and Green, 1999, Dobra, 2004, Jones et al, 2005), and hyper-Dirichlet for multinomial cell probabilities (Tarantola, 2004).

For the gaussian case, there has been a lot of detailed work on choice of hyperparameters in hyper-inverse-Wishart priors (e.g. Rajaratnam, Massam and Carvalho, 2008, Carvalho and Scott, 2009, Ben-David et al, 2011).

Updating parameters in decomposable graph models

Giudici and Green (1999) and other early work conducts joint structural and quantitative learning by sampling from the joint posterior of graph and parameters – usually a variable-dimension problem, necessitating reversible jump or equivalent trans-dimensional sampling.

To try to scale up to larger problems, almost all authors have used conjugate hyper-Markov priors for parameters, so that these can be integrated out. Metropolis ratios for graph updates then simplify into a locally-computable form.

Thus, hyper-inverse-Wishart priors are used for variance matrices in gaussian models (Giudici and Green, 1999, Dobra, 2004, Jones et al, 2005), and hyper-Dirichlet for multinomial cell probabilities (Tarantola, 2004).

For the gaussian case, there has been a lot of detailed work on choice of hyperparameters in hyper-inverse-Wishart priors (e.g. Rajaratnam, Massam and Carvalho, 2008, Carvalho and Scott, 2009, Ben-David et al, 2011).

Updating parameters in decomposable graph models

Giudici and Green (1999) and other early work conducts joint structural and quantitative learning by sampling from the joint posterior of graph and parameters – usually a variable-dimension problem, necessitating reversible jump or equivalent trans-dimensional sampling.

To try to scale up to larger problems, almost all authors have used conjugate hyper-Markov priors for parameters, so that these can be integrated out. Metropolis ratios for graph updates then simplify into a locally-computable form.

Thus, hyper-inverse-Wishart priors are used for variance matrices in gaussian models (Giudici and Green, 1999, Dobra, 2004, Jones et al, 2005), and hyper-Dirichlet for multinomial cell probabilities (Tarantola, 2004).

For the gaussian case, there has been a lot of detailed work on choice of hyperparameters in hyper-inverse-Wishart priors (e.g. Rajaratnam, Massam and Carvalho, 2008, Carvalho and Scott, 2009, Ben-David et al, 2011).

Updating parameters in decomposable graph models

Giudici and Green (1999) and other early work conducts joint structural and quantitative learning by sampling from the joint posterior of graph and parameters – usually a variable-dimension problem, necessitating reversible jump or equivalent trans-dimensional sampling.

To try to scale up to larger problems, almost all authors have used conjugate hyper-Markov priors for parameters, so that these can be integrated out. Metropolis ratios for graph updates then simplify into a locally-computable form.

Thus, hyper-inverse-Wishart priors are used for variance matrices in gaussian models (Giudici and Green, 1999, Dobra, 2004, Jones et al, 2005), and hyper-Dirichlet for multinomial cell probabilities (Tarantola, 2004).

For the gaussian case, there has been a lot of detailed work on choice of hyperparameters in hyper-inverse-Wishart priors (e.g. Rajaratnam, Massam and Carvalho, 2008, Carvalho and Scott, 2009, Ben-David et al, 2011).

Learning decomposable graphs – beyond MCMC

There is accumulated evidence that MCMC sampling, using local moves, is

- good at recovering low-dimensional features of the posterior, such as edge-inclusion probabilities, but
- in larger graphs, poor at global properties, as measured by log-posteriors.

This motivates the powerful idea of using online estimation of edge-inclusion probabilities to drive a 'global' stochastic search heuristic - the key idea in FINCS (feature inclusion stochastic search), Scott and Carvalho (2008), which combines local moves with resampling from the past history and jumps to new regions of graph space.

Empirically, FINCS outperforms certain MCMC-based search methods in respect of hitting high-posterior-probability graphs and in prediction tasks, but we still don't know enough about how well different approaches cover regions with high total probability.

Learning decomposable graphs – beyond MCMC

There is accumulated evidence that MCMC sampling, using local moves, is

- good at recovering low-dimensional features of the posterior, such as edge-inclusion probabilities, but
- in larger graphs, poor at global properties, as measured by log-posteriors.

This motivates the powerful idea of using online estimation of edge-inclusion probabilities to drive a 'global' stochastic search heuristic - the key idea in FINCS (feature inclusion stochastic search), Scott and Carvalho (2008), which combines local moves with resampling from the past history and jumps to new regions of graph space.

Empirically, FINCS outperforms certain MCMC-based search methods in respect of hitting high-posterior-probability graphs and in prediction tasks, but we still don't know enough about how well different approaches cover regions with high total probability.

Learning decomposable graphs – beyond MCMC

There is accumulated evidence that MCMC sampling, using local moves, is

- good at recovering low-dimensional features of the posterior, such as edge-inclusion probabilities, but
- in larger graphs, poor at global properties, as measured by log-posteriors.

This motivates the powerful idea of using online estimation of edge-inclusion probabilities to drive a 'global' stochastic search heuristic - the key idea in FINCS (feature inclusion stochastic search), Scott and Carvalho (2008), which combines local moves with resampling from the past history and jumps to new regions of graph space.

Empirically, FINCS outperforms certain MCMC-based search methods in respect of hitting high-posterior-probability graphs and in prediction tasks, but we still don't know enough about how well different approaches cover regions with high total probability.

Non-decomposable graphs

... or rather, not necessarily decomposable graphs.

Even if $\mathcal{G}$ is not decomposable, we still have the prime component factorisation

$$p(X) = \frac{\prod_{P \in \mathcal{P}} p(X_P)}{\prod_{S \in \mathcal{S}} p(X_S)}$$

where the prime components P_i are the maximal subgraphs that cannot be decomposed: in a non-decomposable graph, at least one is not complete.

Non-decomposable graphs

... or rather, not necessarily decomposable graphs.

Even if $\mathcal{G}$ is not decomposable, we still have the prime component factorisation

$$p(X) = \frac{\prod_{P \in \mathcal{P}} p(X_P)}{\prod_{S \in \mathcal{S}} p(X_S)}$$

where the prime components P_i are the maximal subgraphs that cannot be decomposed: in a non-decomposable graph, at least one is not complete.

Bayesian model determination with non-decomposable graphs

The additional difficulties in sampling non-decomposable graphical models are (Jones *et al*, *Stat. Sci.*, 2005):

- The normalising constants in the non-complete prime component marginals do not have closed form, so we need Monte Carlo methods to estimate them.
- These Monte Carlo calculated values have high variance.
- When you make single-edge perturbations to the graph, there is no guarantee of significant cancellations in likelihood ratios.

These difficulties hugely increase computing time – in their experiments, 420 times for a 12-node, 15-edge example; 5500 times for 15-node, 26-edge example (this is for Gaussian models, using conjugate priors on variances).

Bayesian model determination with non-decomposable graphs

The additional difficulties in sampling non-decomposable graphical models are (Jones *et al*, *Stat. Sci.*, 2005):

- The normalising constants in the non-complete prime component marginals do not have closed form, so we need Monte Carlo methods to estimate them.
- These Monte Carlo calculated values have high variance.
- When you make single-edge perturbations to the graph, there is no guarantee of significant cancellations in likelihood ratios.

These difficulties hugely increase computing time – in their experiments, 420 times for a 12-node, 15-edge example; 5500 times for 15-node, 26-edge example (this is for Gaussian models, using conjugate priors on variances).

Bayesian model determination with non-decomposable graphs

Jones *et al* (*Stat. Sci.*, 2005) conclude that sampling from the posterior is not practical for problems with much more than 15 nodes – and resort to (fast) heuristics like stochastic shotgun search to identify a graph with high posterior probability instead.

Later samplers by Dobra, Lenkoski and Rodriguez (2011) and Wang and Li (2012) eliminate the need for MCMC for the normalising constants.

Bayesian model determination with non-decomposable graphs

Jones *et al* (*Stat. Sci.*, 2005) conclude that sampling from the posterior is not practical for problems with much more than 15 nodes – and resort to (fast) heuristics like stochastic shotgun search to identify a graph with high posterior probability instead.

Later samplers by Dobra, Lenkoski and Rodriguez (2011) and Wang and Li (2012) eliminate the need for MCMC for the normalising constants.

Stochastic shotgun search

As a heuristic for computing part of the posterior, Jones et al (2005) propose iterating on

1. start with a graph $\mathcal{G}$
2. select at random n_1 graphs that differ by one edge (neighbours), compute their unnormalized posterior probabilities and retain the top n_2
3. among the n_2 top neighbours, propose the ith graph $\mathcal{G}_i$ as a new starting graph for an MCMC update with probability proportional to p_i^α where p_i is the unnormalized posterior probability of graph $\mathcal{G}_i$ and α is an annealing parameter.
4. return to step 2 and iterate. Maintain a list of the overall best n_3 graphs visited.

If the top n_3 graphs really capture most of the posterior probability, we might hope to get a good idea of the true posterior distribution.

This SSS method is more fully explored in Hans et al (2007); see also Ben-David et al (2011).

Stochastic shotgun search

As a heuristic for computing part of the posterior, Jones et al (2005) propose iterating on

1. start with a graph $\mathcal{G}$
2. select at random n_1 graphs that differ by one edge (neighbours), compute their unnormalized posterior probabilities and retain the top n_2
3. among the n_2 top neighbours, propose the ith graph $\mathcal{G}_i$ as a new starting graph for an MCMC update with probability proportional to p_i^α where p_i is the unnormalized posterior probability of graph $\mathcal{G}_i$ and α is an annealing parameter.
4. return to step 2 and iterate. Maintain a list of the overall best n_3 graphs visited.

If the top n_3 graphs really capture most of the posterior probability, we might hope to get a good idea of the true posterior distribution.

This SSS method is more fully explored in Hans et al (2007); see also Ben-David et al (2011).

Stochastic shotgun search

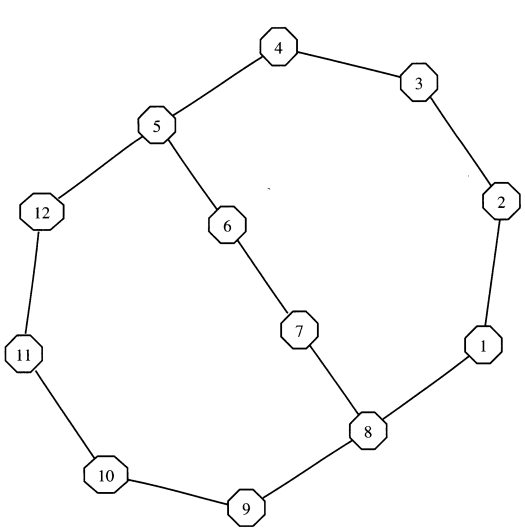
As a heuristic for computing part of the posterior, Jones et al (2005) propose iterating on

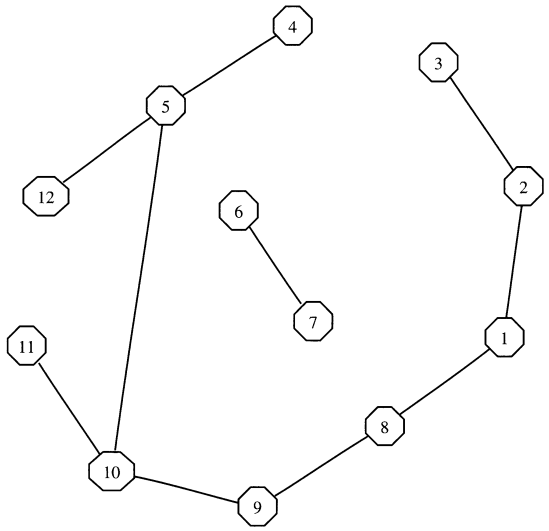
1. start with a graph $\mathcal{G}$
2. select at random n_1 graphs that differ by one edge (neighbours), compute their unnormalized posterior probabilities and retain the top n_2
3. among the n_2 top neighbours, propose the ith graph $\mathcal{G}_i$ as a new starting graph for an MCMC update with probability proportional to p_i^α where p_i is the unnormalized posterior probability of graph $\mathcal{G}_i$ and α is an annealing parameter.
4. return to step 2 and iterate. Maintain a list of the overall best n_3 graphs visited.

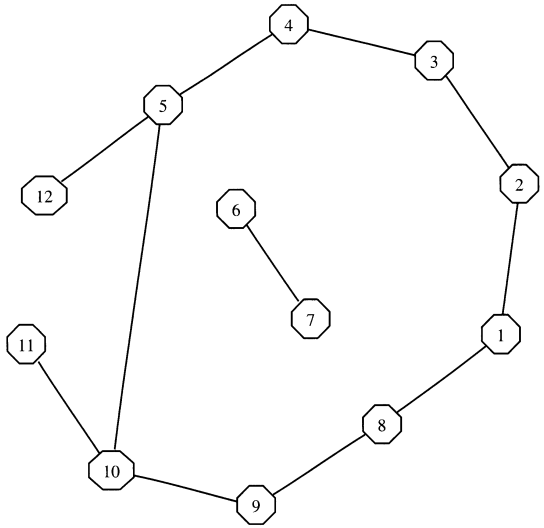
If the top n_3 graphs really capture most of the posterior probability, we might hope to get a good idea of the true posterior distribution.

This SSS method is more fully explored in Hans et al (2007); see also Ben-David et al (2011).

Examples of Stochastic shotgun search vs MCMC



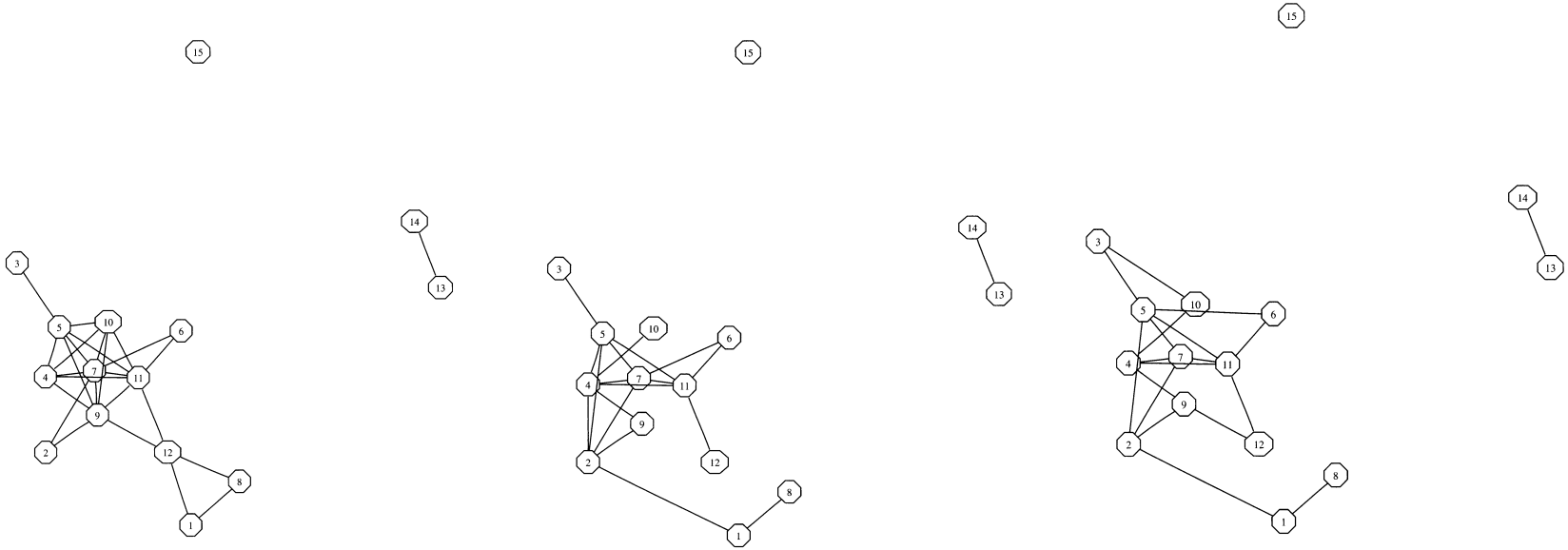




True, max posterior probability graphs: decomposable, unrestricted

Method[a]	Run time (sec)	Max log posterior	Graphs to first top graph visit	Time to first top graph visit
MH-d	36	−2591.18	912	1
SSS-d	183	−2591.18	792	2
MH-u	15,220	−2590.94	415	2
SSS-u	2773	−2590.94	13,266	5

Examples of Stochastic shotgun search vs MCMC



True, max posterior probability graphs: decomposable, unrestricted

Method[a]	Run time (sec)	Max log posterior	Graphs to first top graph visit	Time to first top graph visit
MH-d	93	15633.76	349,484	36
SSS-d	234	15633.76	33,495	9
MH-u	513,077	15633.83	666,425	309, 222
SSS-u	5930	15636.38	82,845	112

A sparse regression approach

. . . to gaussian model determination with non-decomposable graphs.

In a multivariate gaussian distribution, all conditional distributions are of course gaussian linear regressions. Dobra et al (2004) propose estimating these conditional regressions from data using (Bayesian) sparse regression techniques.

To ensure consistency of the estimated conditionals, one approach is take a pre-fixed order of the variables, and regress each one only on its predecessors. Sparse regressions then estimate a DAG, and we can deliver a corresponding undirected graph.

Non-Bayesian sparse regression methods such as the lasso have also been used in this way (Meinshausen and Buhlmann, 2006, Yuan and Lin, 2007)

A sparse regression approach

. . . to gaussian model determination with non-decomposable graphs.

In a multivariate gaussian distribution, all conditional distributions are of course gaussian linear regressions. Dobra et al (2004) propose estimating these conditional regressions from data using (Bayesian) sparse regression techniques.

To ensure consistency of the estimated conditionals, one approach is take a pre-fixed order of the variables, and regress each one only on its predecessors. Sparse regressions then estimate a DAG, and we can deliver a corresponding undirected graph.

Non-Bayesian sparse regression methods such as the lasso have also been used in this way (Meinshausen and Buhlmann, 2006, Yuan and Lin, 2007)

A sparse regression approach

. . . to gaussian model determination with non-decomposable graphs.

In a multivariate gaussian distribution, all conditional distributions are of course gaussian linear regressions. Dobra et al (2004) propose estimating these conditional regressions from data using (Bayesian) sparse regression techniques.

To ensure consistency of the estimated conditionals, one approach is take a pre-fixed order of the variables, and regress each one only on its predecessors. Sparse regressions then estimate a DAG, and we can deliver a corresponding undirected graph.

Non-Bayesian sparse regression methods such as the lasso have also been used in this way (Meinshausen and Buhlmann, 2006, Yuan and Lin, 2007)

A sparse regression approach

. . . to gaussian model determination with non-decomposable graphs.

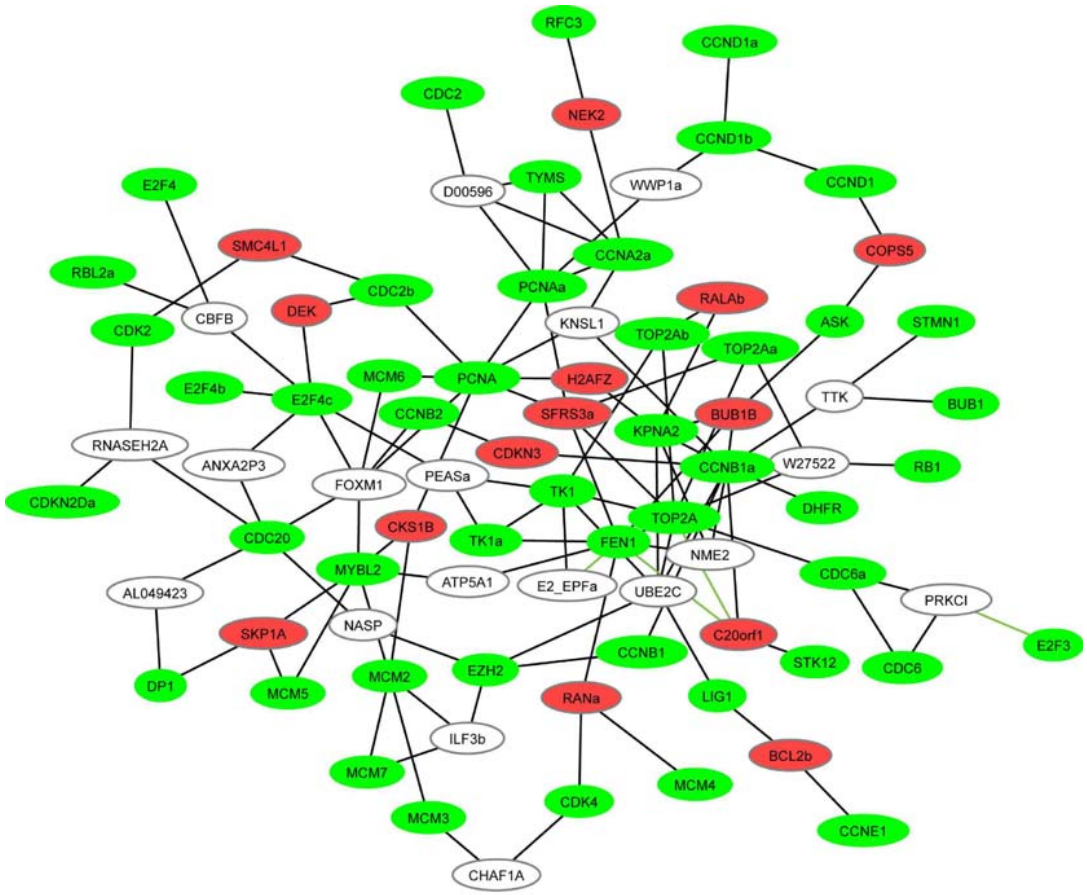
In a multivariate gaussian distribution, all conditional distributions are of course gaussian linear regressions. Dobra et al (2004) propose estimating these conditional regressions from data using (Bayesian) sparse regression techniques.

To ensure consistency of the estimated conditionals, one approach is take a pre-fixed order of the variables, and regress each one only on its predecessors. Sparse regressions then estimate a DAG, and we can deliver a corresponding undirected graph.

Non-Bayesian sparse regression methods such as the lasso have also been used in this way (Meinshausen and Buhlmann, 2006, Yuan and Lin, 2007)

A sparse regression approach

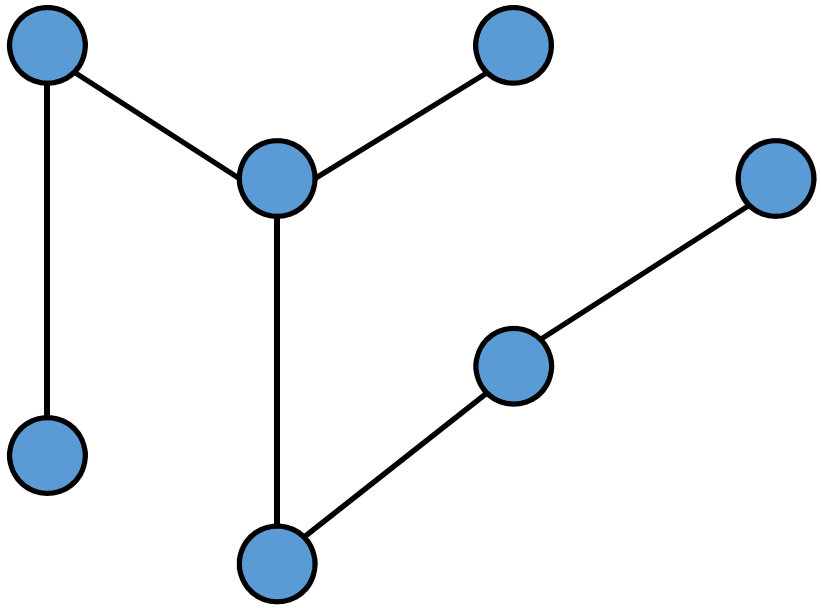
Subgraph containing two genes of interest in an example on expression data on 12588 genes in a breast cancer study, by Dobra et al (2004)



Other important contributions on non-decomposable graphs

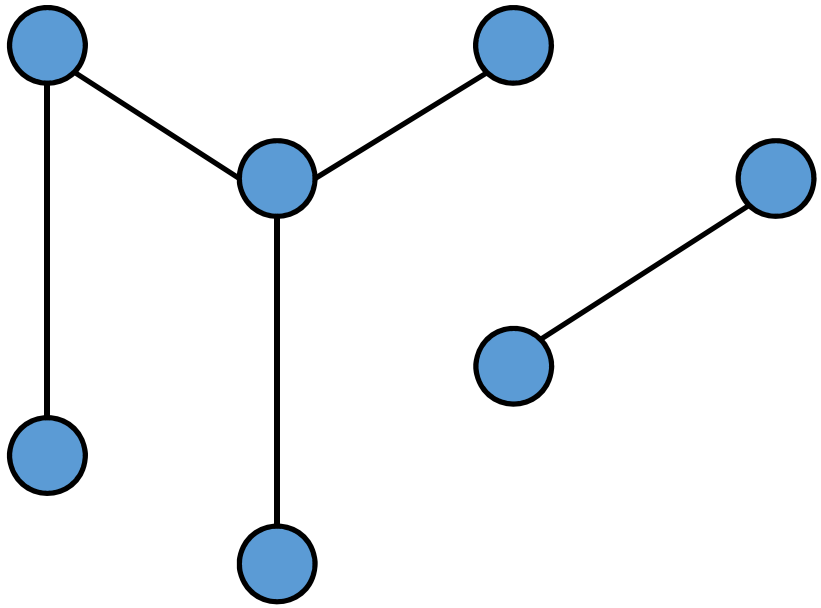
- Carvalho, Massam and West (2007), Wang and Carvalho (2010), Mitsakakis, Massam and Escobar (2011) – sampling Σ.
- Lenkoski and Dobra (2011) – MOSS, a random uphill search method, constructs list of most probable graphs, subject to max clique size.
- Stingo and Marchetti (2015) adapt their perfect-ordering-representation to an approximate method for non-decomposable graphs, using a special mixture prior that models 'nearly decomposable' graphs.

Trees



A tree is a connected undirected graph with no loops.

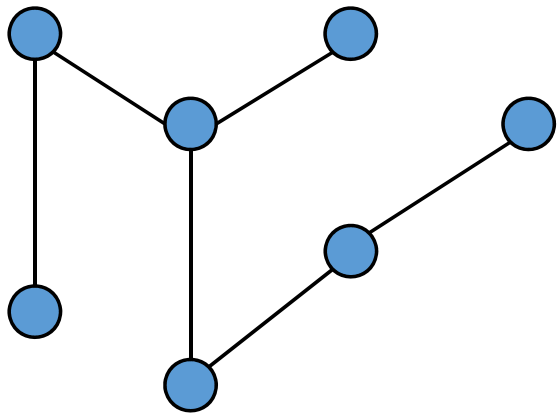
Forests



A forest is an undirected graph with no loops.

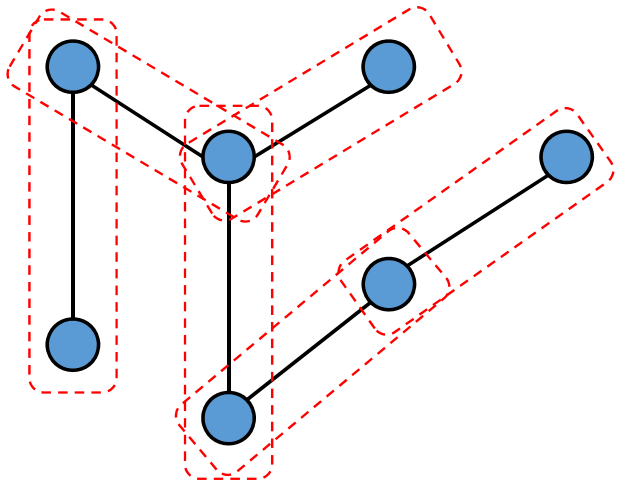
Trees and forests

Trees and forests are decomposable graphs, of course: for a tree, the junction tree is essentially isomorphic to the tree itself.



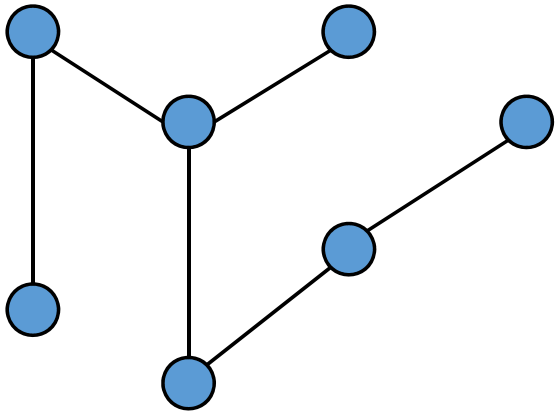
Trees and forests

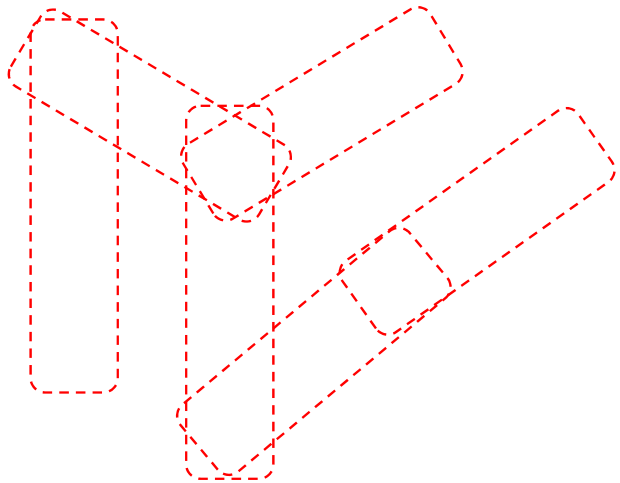
Trees and forests are decomposable graphs, of course: for a tree, the junction tree is essentially isomorphic to the tree itself.



Trees and forests

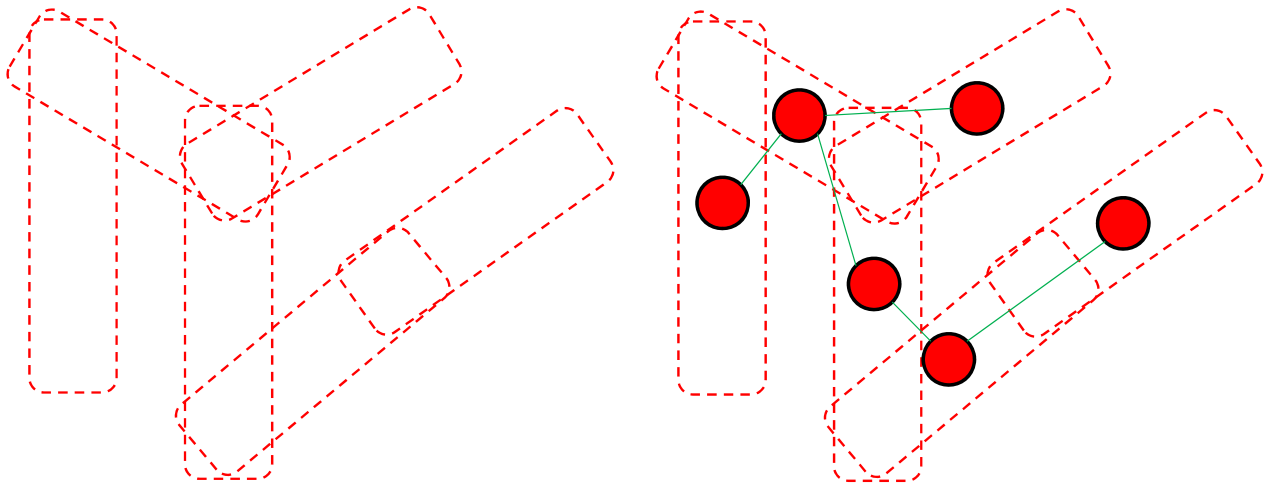
Trees and forests are decomposable graphs, of course: for a tree, the junction tree is essentially isomorphic to the tree itself.





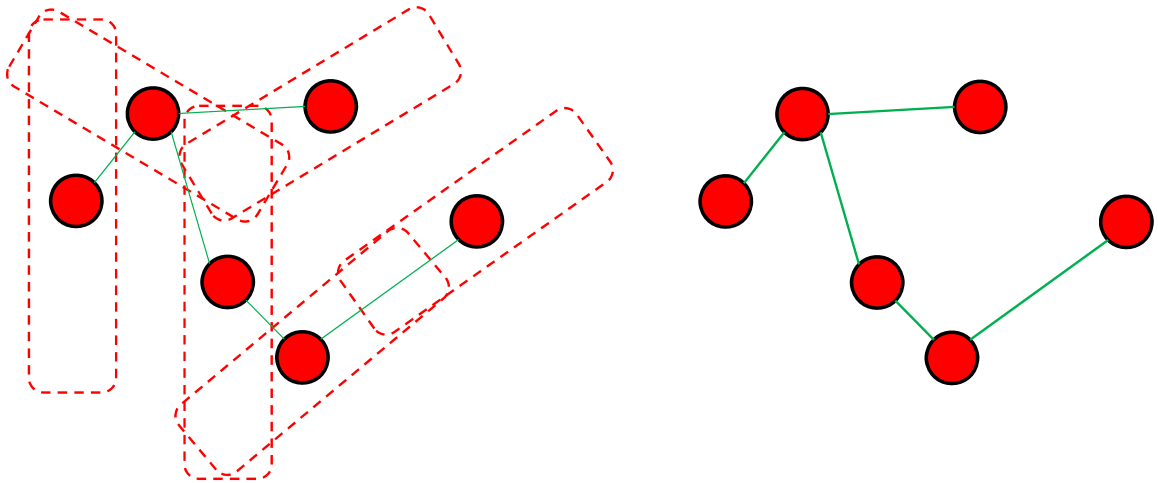
Trees and forests

Trees and forests are decomposable graphs, of course: for a tree, the junction tree is essentially isomorphic to the tree itself.



Trees and forests

Trees and forests are decomposable graphs, of course: for a tree, the junction tree is essentially isomorphic to the tree itself.



Trees and forests

If $\mathcal{G}$ is a tree $\mathcal{T}$, then

$$P(X|\mathcal{T}) = \frac{\prod_{C\in\mathcal{C}} p(X_C|\mathcal{T})}{\prod_{S\in\mathcal{S}} p(X_S|\mathcal{T})} = \frac{\prod_{e\in\mathcal{E}(\mathcal{T})} p(X_e|\mathcal{T})}{\prod_{v\in V_i} p(X_v|\mathcal{T})}$$

where V_i are the non-leaf (interior) vertices, and $\mathcal{E}(\mathcal{T})$ are the edges of $\mathcal{T}$, so

$$p(\mathcal{T}|X) \propto p(\mathcal{T}) \prod_{(u,v)\in\mathcal{E}(\mathcal{T})} \frac{p(X_{u,v})}{p(X_u)p(X_v)}$$

. . . a product of Bayes factors for dependence along edges.

Perfect simulation for posterior on trees and forests

For trees,

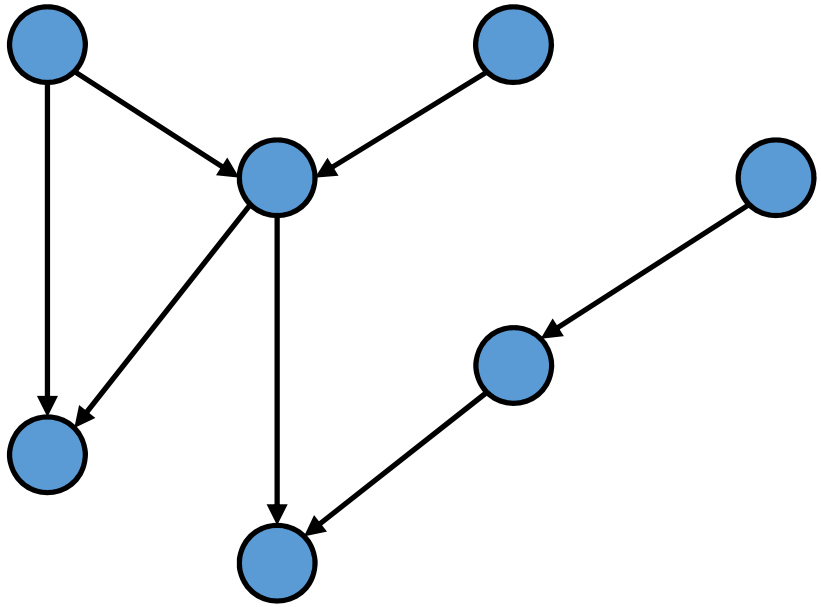
$$p(\mathcal{T}|X) \propto p(\mathcal{T}) \prod_{(u,v)\in\mathcal{E}(\mathcal{T})} \frac{p(X_{u,v})}{p(X_u)p(X_v)}$$

is amenable to perfect simulation (cf Propp and Wilson, 1998), using algorithms for random spanning trees (Guénoche, 1983; Broder, 1989; Aldous, 1990).

This extends to tree priors $p(\mathcal{T})$ that are 'decomposable' - i.e. factorise as products of weights on edges (Meilă & Jaakkola, 2006).

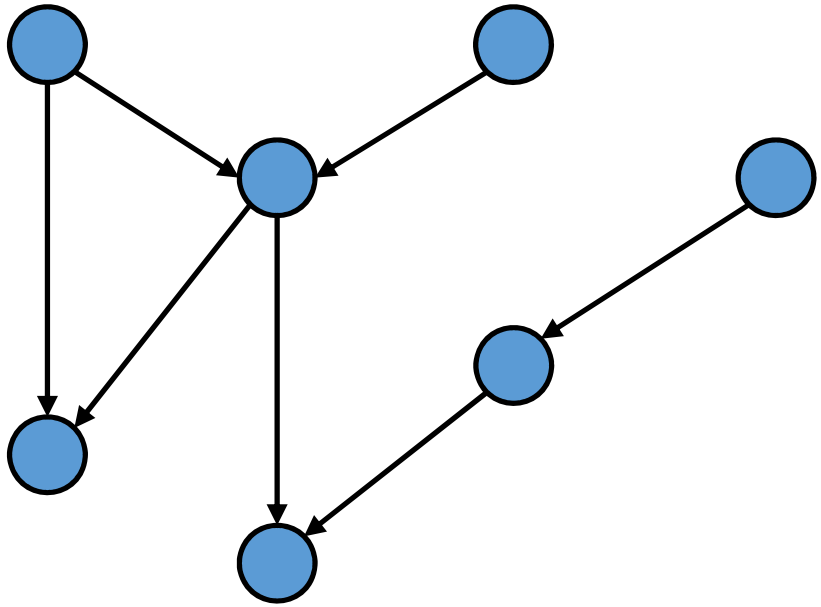
Perfect sampling from random forest distributions of similar form is possible but much harder (Dai, 1998).

DAGs



A directed acyclic graph (DAG) is a directed graph in which there are no directed loops. Equivalently, it is a directed graph in which edges from a node can only go to a higher-numbered node.

DAGs



A directed acyclic graph (DAG) is a directed graph in which there are no directed loops. Equivalently, it is a directed graph in which edges from a node can only go to a higher-numbered node.

DAGs

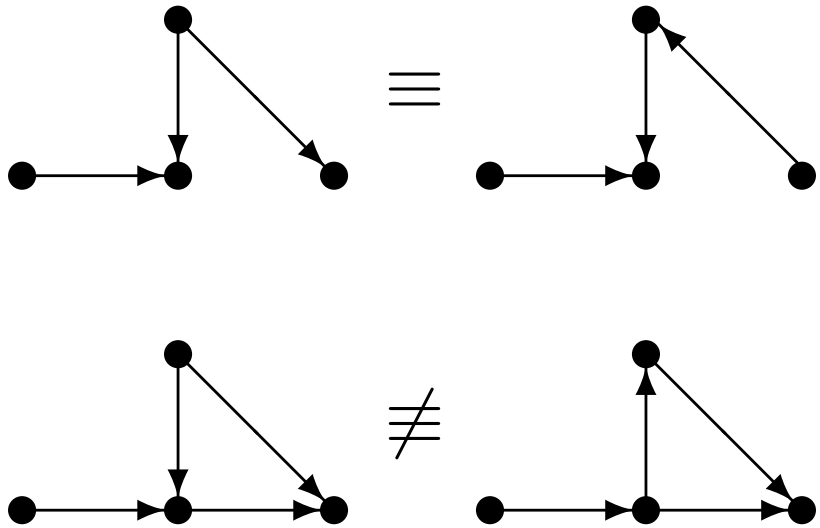
.. particularly important because of (real or more likely imagined) causal interpretation.

When the variables are all categorical, these models are often called Bayes(ian) net(work)s – a big focus of attention in the machine learning community, because of use in expert systems and availability of very fast algorithms.

(So in literature search, look for both DAGs and BNs).

Markov equivalence

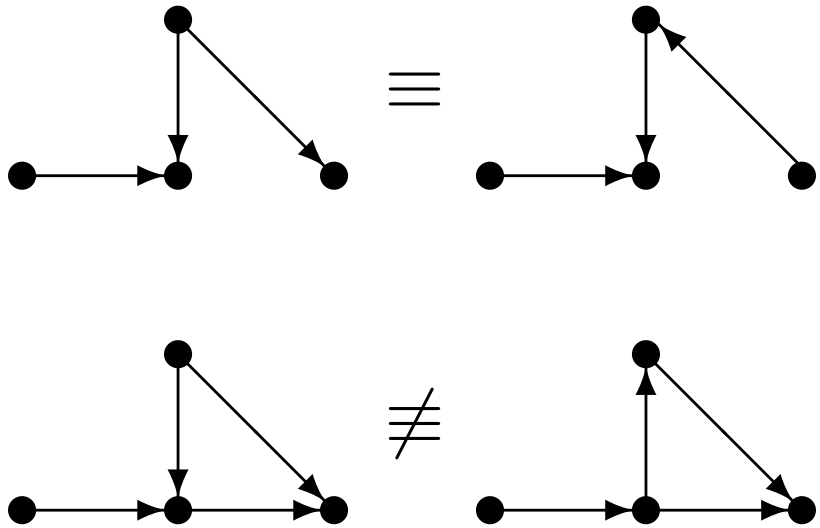
Two DAGs are Markov equivalent if they imply exactly the same set of conditional independences.



A (graphical) test: they are Markov equivalent if they have the same skeleton (graph ignoring directions) and the same sets of unmarried parents (Verma and Pearl, 1990).

Markov equivalence

Two DAGs are Markov equivalent if they imply exactly the same set of conditional independences.

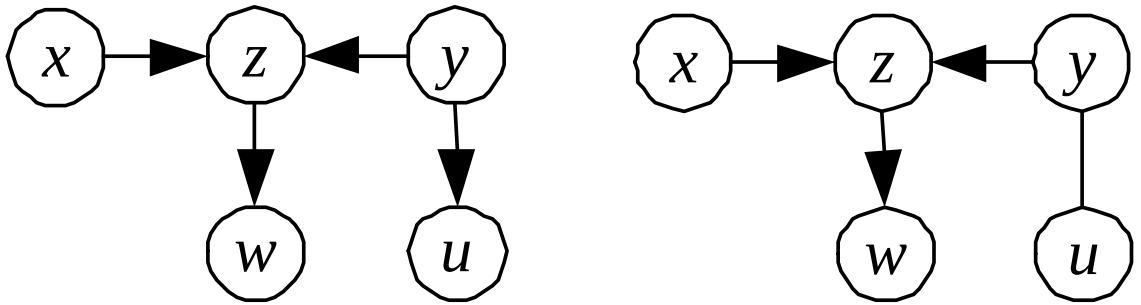


A (graphical) test: they are Markov equivalent if they have the same skeleton (graph ignoring directions) and the same sets of unmarried parents (Verma and Pearl, 1990).

Markov equivalence

Markov equivalence classes of DAGs can be represented graphically, in a form known as maximally oriented graphs (Meek, 1995), essential graphs (Andersson et al., 1997 or completed PDAGs (Chickering, 2002)).

A completed PDAG is a chain graph (graph with edges and arrows, but no directed cycles) that has the same skeleton as the original DAG, and edges that are directed if and only if they are directed in every equivalent DAG.

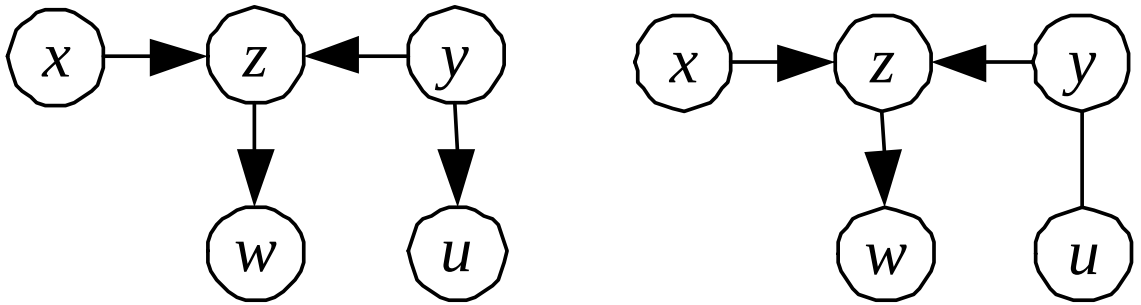


A DAG and its completed PDAG.

Markov equivalence

Markov equivalence classes of DAGs can be represented graphically, in a form known as maximally oriented graphs (Meek, 1995), essential graphs (Andersson et al., 1997 or completed PDAGs (Chickering, 2002)).

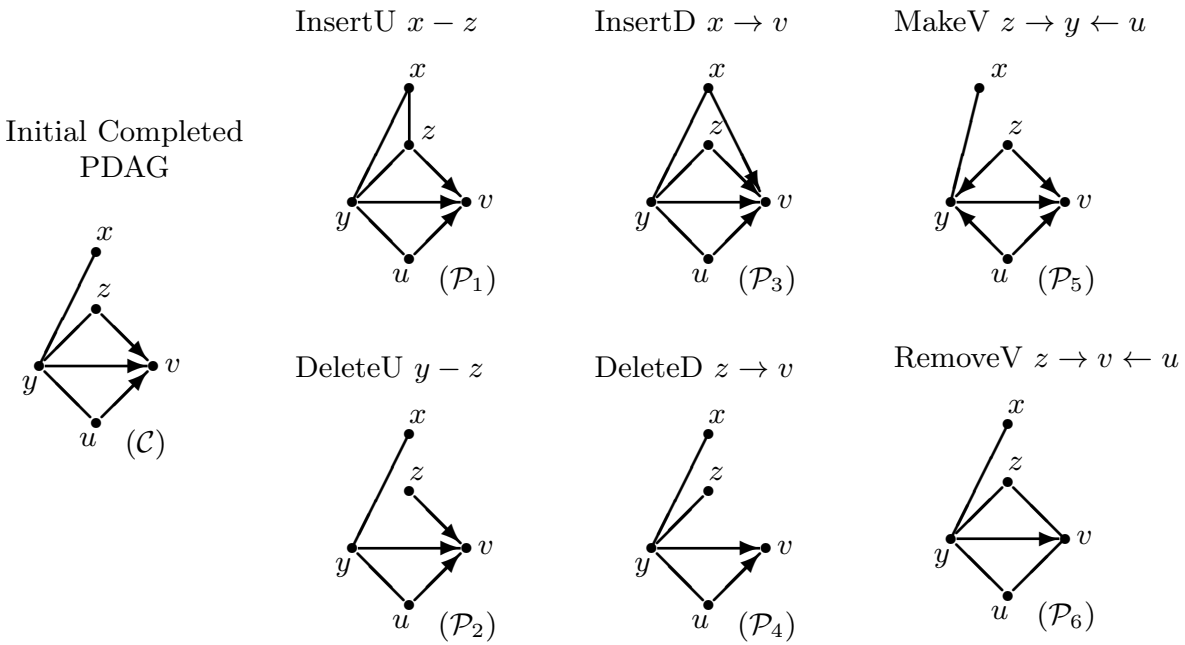
A completed PDAG is a chain graph (graph with edges and arrows, but no directed cycles) that has the same skeleton as the original DAG, and edges that are directed if and only if they are directed in every equivalent DAG.



A DAG and its completed PDAG.

MCMC on completed PDAGs

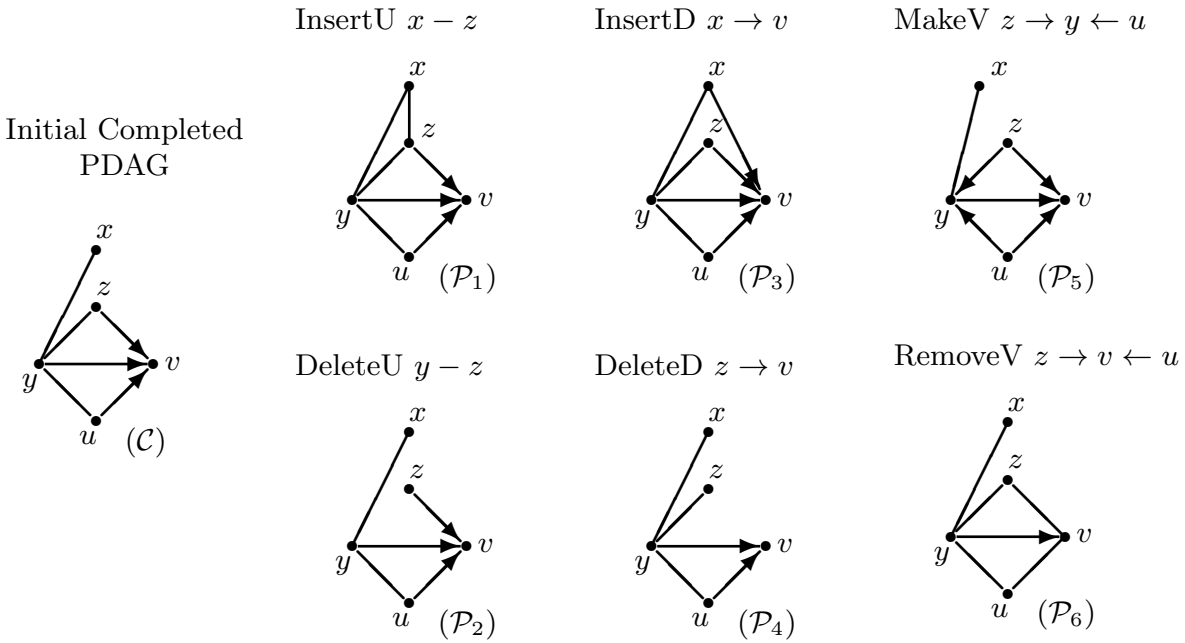
He, Jia & Yu (2013) construct a reversible Markov chain on Markov equivalence classes of DAGs, that is scalable to large sparse graphs. It uses operators (MCMC moves) chosen from these:



restricting to those that are reversible at the current state. Builds on Chickering (2002) and earlier MCMC from Madigan et al (1996), Perlman (2000) and Peña (2013). Potentially useful for Bayesian computation.

MCMC on completed PDAGs

He, Jia & Yu (2013) construct a reversible Markov chain on Markov equivalence classes of DAGs, that is scalable to large sparse graphs. It uses operators (MCMC moves) chosen from these:



restricting to those that are reversible at the current state. Builds on Chickering (2002) and earlier MCMC from Madigan et al (1996), Perlman (2000) and Peña (2013). Potentially useful for Bayesian computation.

Wrap-up

- Bayesian structural learning (in the sense of delivering posterior distributions) from i.i.d. data is a simply stated task
- . . . that proves amazingly hard to deliver computationally except on a modest scale
- . . . even with special choices of priors, etc.
- . . . in spite of the best efforts of a lot of creative people!
- We don't really know what heuristics and approximations are delivering, quantitatively.
- Perhaps time to focus effort (algorithmic and theoretical) more on edge-inclusion probabilities and other low-dimensional margins?
- . . . and more effort on relaxation of the hard constraints of conditional independence?

Wrap-up

- Bayesian structural learning (in the sense of delivering posterior distributions) from i.i.d. data is a simply stated task
- . . . that proves amazingly hard to deliver computationally except on a modest scale
- . . . even with special choices of priors, etc.
- . . . in spite of the best efforts of a lot of creative people!
- We don't really know what heuristics and approximations are delivering, quantitatively.
- Perhaps time to focus effort (algorithmic and theoretical) more on edge-inclusion probabilities and other low-dimensional margins?
- . . . and more effort on relaxation of the hard constraints of conditional independence?

Wrap-up

- Bayesian structural learning (in the sense of delivering posterior distributions) from i.i.d. data is a simply stated task
- . . . that proves amazingly hard to deliver computationally except on a modest scale
- . . . even with special choices of priors, etc.
- . . . in spite of the best efforts of a lot of creative people!
- We don't really know what heuristics and approximations are delivering, quantitatively.
- Perhaps time to focus effort (algorithmic and theoretical) more on edge-inclusion probabilities and other low-dimensional margins?
- . . . and more effort on relaxation of the hard constraints of conditional independence?

Wrap-up

- Bayesian structural learning (in the sense of delivering posterior distributions) from i.i.d. data is a simply stated task
- . . . that proves amazingly hard to deliver computationally except on a modest scale
- . . . even with special choices of priors, etc.
- . . . in spite of the best efforts of a lot of creative people!
- We don't really know what heuristics and approximations are delivering, quantitatively.
- Perhaps time to focus effort (algorithmic and theoretical) more on edge-inclusion probabilities and other low-dimensional margins?
- . . . and more effort on relaxation of the hard constraints of conditional independence?

Wrap-up

- Bayesian structural learning (in the sense of delivering posterior distributions) from i.i.d. data is a simply stated task
- . . . that proves amazingly hard to deliver computationally except on a modest scale
- . . . even with special choices of priors, etc.
- . . . in spite of the best efforts of a lot of creative people!
- We don't really know what heuristics and approximations are delivering, quantitatively.
- Perhaps time to focus effort (algorithmic and theoretical) more on edge-inclusion probabilities and other low-dimensional margins?
- . . . and more effort on relaxation of the hard constraints of conditional independence?

Wrap-up

- Bayesian structural learning (in the sense of delivering posterior distributions) from i.i.d. data is a simply stated task
- . . . that proves amazingly hard to deliver computationally except on a modest scale
- . . . even with special choices of priors, etc.
- . . . in spite of the best efforts of a lot of creative people!
- We don't really know what heuristics and approximations are delivering, quantitatively.
- Perhaps time to focus effort (algorithmic and theoretical) more on edge-inclusion probabilities and other low-dimensional margins?
- . . . and more effort on relaxation of the hard constraints of conditional independence?

Wrap-up

- Bayesian structural learning (in the sense of delivering posterior distributions) from i.i.d. data is a simply stated task
- . . . that proves amazingly hard to deliver computationally except on a modest scale
- . . . even with special choices of priors, etc.
- . . . in spite of the best efforts of a lot of creative people!
- We don't really know what heuristics and approximations are delivering, quantitatively.
- Perhaps time to focus effort (algorithmic and theoretical) more on edge-inclusion probabilities and other low-dimensional margins?
- . . . and more effort on relaxation of the hard constraints of conditional independence?

Contacting me:

- Email: P.J.Green@bristol.ac.uk
- Webpage: www.stats.bris.ac.uk/∼peter/

Bibliography

Abel, H.J. & Thomas, A.(2011) Accuracy and computational efficiency of a graphical modeling approach to linkage disequilibrium estimation. *Statist. Applic. Genet. Molec. Biol.* 10:Article 5.

Andersson, S. A., Madigan, D. & Perlman, M. D. (1997). A characterization of Markov equivalence classes for acyclic digraphs. *The Annals of Statistics*, 25:505–541.

Armstrong, H., Carter, C. K., Wong, K. F. K. & Kohn, R. (2009). Bayesian covariance matrix estimation using a mixture of decomposable graphical models. *Statistics and Computing*, 19:303-316.

Atay-Kayis, A. & Massam, H. (2005). A Monte Carlo method for computing the marginal likelihood in nondecomposable graphical Gaussian models. *Biometrika* 92:317–35.

Beal, M. & Gharamani, Z. (2003). The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. *Bayesian statistics 7*: 453–464.

Ben-David, E., Li, T., Massam, H. and Rajaratnam, B. (2011). High dimensional Bayesian inference for Gaussian directed acyclic graph models. *ArXiv*: 1109.4371.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B*, 36:192–236.

Byrne, S. (2011). Hyper and Structural Markov Laws for Graphical Models. PhD thesis, Statistical Laboratory, University of Cambridge.

Byrne, S. & Dawid, A. P. (2015). Structural Markov graph laws for Bayesian model uncertainty. *The Annals of Statistics*, 43:1647–1681.

Carvalho, C. M., Massam, H. and West, M. (2007). Simulation of hyper-inverse Wishart distributions in graphical models. *Biometrika*, 94:647–659.

Bibliography

Carvalho, C. M. and Scott, J. G. (2009). Objective Bayesian model selection in Gaussian graphical models. *Biometrika*, 96:497–512.

Chickering, M. (2002). Learning equivalence classes of Bayesian-network structures. *Journal of Machine Learning Research*, 2:445–498.

Chickering, M., Heckerman, D. & Meek, C. (1997) A Bayesian approach to learning Bayesian networks with local structure. *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*.

Consonni, G. & Leucari, V. (2001). Model determination for directed acyclic graphs. *Journal of the Royal Statistical Society. Series D*, 50:243–256.

Cooper, G. F. & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347.

Corander, J., Ekdahl, M. and Koski, T. (2008). Parallel interacting MCMC for learning of topologies of graphical models. *Data Min Knowl Disc*, 17:431–456.

Dai, H. (2008). Perfect sampling methods for random forests. *Advances in Applied Probability*, 40:897–917.

Dawid, A. P. & Lauritzen, S. L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. The *Annals of Statistics*, 21:1272–1317.

Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G., and West, M. (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, 90:196–212.

Dobra, A., Lenkoski, A. and Rodriguez, A. (2011). Bayesian inference for general gaussian graphical models with application to multivariate lattice data. *Journal of the American Statistical Association*, 106:1418–1433.

Bibliography

Dobra, A. and Lenkoski, A. (2011). Copula gaussian graphical models and their application to modeling functional disability data. *The Annals of Applied Statistics*, 5:969–993.

Edwards, D., & Havránek, T. (1985). A fast procedure for model search in multidimensional contingency tables. *Biometrika*, 72:339–351.

Fitch, A. Marie, Jones, M. Beatrix and Massam, Hélène (2014). The performance of covariance selection methods that consider decomposable models only. *Bayesian Analysis*, 9:659–684.

Fronk, E. M. & Giudici, P. (2004). Markov Chain Monte Carlo model selection for DAG models. *Statistical Methods and Applications*, 13:259–273.

Friedman, N. & Koller, D. (2003). Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning*, 50:95–125.

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

Goudie, R. J. B. and Mukherjee, S. (2016). A Gibbs sampler for learning DAGs. *Journal of Machine Learning Research*, 17:1–39.

Green, P. J., Hjort, N. L. & Richardson, S. (2003) (eds). *Highly structured stochastic systems*. Oxford University Press.

Green, P. J. & Thomas, A. (2013). Sampling decomposable graphs using a Markov chain on junction trees. *Biometrika*, 100:91–110.

Hans, C., Dobra, A. and West, M. (2007) Shotgun stochastic search in regression with many variables. *J. American Statistical Association*, 102:507–516.

He, Y., Jia, J. & Yu, B. (2013). Reversible MCMC on Markov equivalence classes of sparse directed acyclic graphs. *The Annals of Statistics*, 41:1742–1779.

Bibliography

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

Højsgaard, S., Edwards, D., & Lauritzen, S. (2012). *Graphical models with R*. Springer.

Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C. & West, M. (2005). Experiments in stochastic computation for high dimensional graphical models. *Statist. Sci.*, 20:388–400.

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.

Lafferty, J., Liu, H. & Wasserman, L. (2012). Sparse nonparametric graphical models. *Statistical Science*, 27: 519–537.

Lauritzen, S. L. (1996). it Graphical Models. Oxford: Clarendon Press.

Lauritzen, S. L. (2012). *Wald lectures*, Istanbul, 2012: http://www.stats.ox.ac.uk/∼steffen.

Lauritzen, S. L. & Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B*, 50:157-224.

Lenkoski, A. and Dobra, A. (2011). Computational aspects related to inference in gaussian graphical models with the g-Wishart prior. *Journal of Computational and Graphical Statistics*, 20:140–157.

Meilă, Marina, & Jordan, Michael (2000). Learning with Mixtures of Trees. *Journal of Machine Learning Research*, 1:1-48.

Meilă, Marina, & Jaakkola, Tommi (2006). Tractable Bayesian learning of tree belief networks. *Statistics and Computing*, 16.1:77–92.

Bibliography

Mitsakakis, N., Massam, H. and Escobar, M. D. (2011). A Metropolis-Hastings based method for sampling from the G-Wishart distribution in Gaussian graphical models. *Electronic Journal of Statistics*, 5:18–30.

Rajaratnam, B., Massam, H. and Carvalho, C. M. (2008). Flexible covariance estimation in graphical gaussian models. *Annals of Statistics*, 36:2818–2849.

Rios, F. L. (2015). Bayesian structure learning in graphical models. PhD thesis, University of Stockholm.

Scott, J. G. and Carvalho, C. M. (2008). Feature-inclusion stochastic search for gaussian graphical models. *Journal of Computational and Graphical Statistics*, 17:790–808.

Spiegelhalter, D. J. & Lauritzen, S. L. (1990). Sequential updating of conditional probabilities on directed graphical structures. *Networks*, 20:579–605.

Stingo, F., and Marchetti, G. M. (2015). Efficient local updates for undirected graphical models. *Statistics and Computing* 25:159–171.

Tarantola, Claudia (2004). MCMC model determination for discrete graphical models. *Statistical Modelling*, 4:39–61.

Thomas, A. & Green, P. J. (2009). Enumerating the junction trees of a decomposable graph. *J. Comp. Graph. Statist.*, 18:930–940.

Thomas, A. & Green, P. J. (2009). Enumerating the decomposable neighbours of a decomposable graph under a simple perturbation scheme. *Computational Statistics and Data Analysis*, 53:1232–1238.

Wang, H. and Carvalho, C. M. (2010). Simulation of hyper-inverse Wishart distributions for non-decomposable graphs. *Electronic Journal of Statistics*, 4:1470–1475.

Wang, H. (2012). Bayesian graphical lasso models and efficient posterior computation. *Bayesian Analysis*, 7:867–886.

Wang, H. and Li, S. (2012). Efficient Gaussian graphical model determination under G-Wishart prior distributions. *Electronic Journal of Statistics*, 6:168–198.