Bayesian inference about decomposable graphs

Peter Green[1] Alun Thomas[2]

[1]UTS, Sydney
University of Bristol

[2]University of Utah

University of Bristol, 28 February 2014

Outline

1. Conditional independence
2. Decomposable graphs
3. Bayesian model determination in decomposable graphs
   - Priors on decomposable graphs
   - Sampling junction trees
4. Examples/demonstrations
5. Non-decomposable graphs

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.

It proves useful to represent conditional independences graphically.

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.

It proves useful to represent conditional independences graphically.

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.

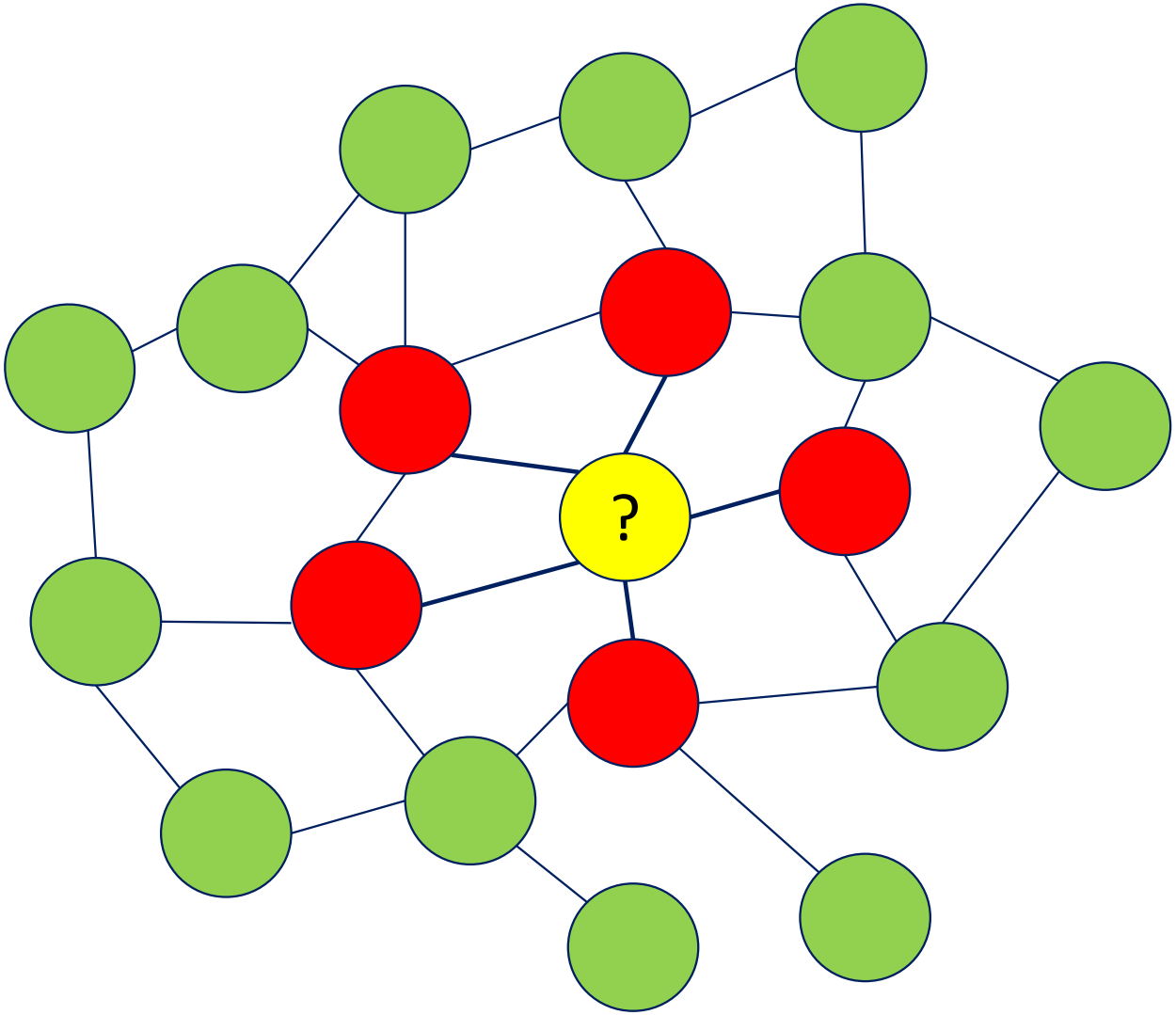
X and Y are conditionally independent given Z:

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

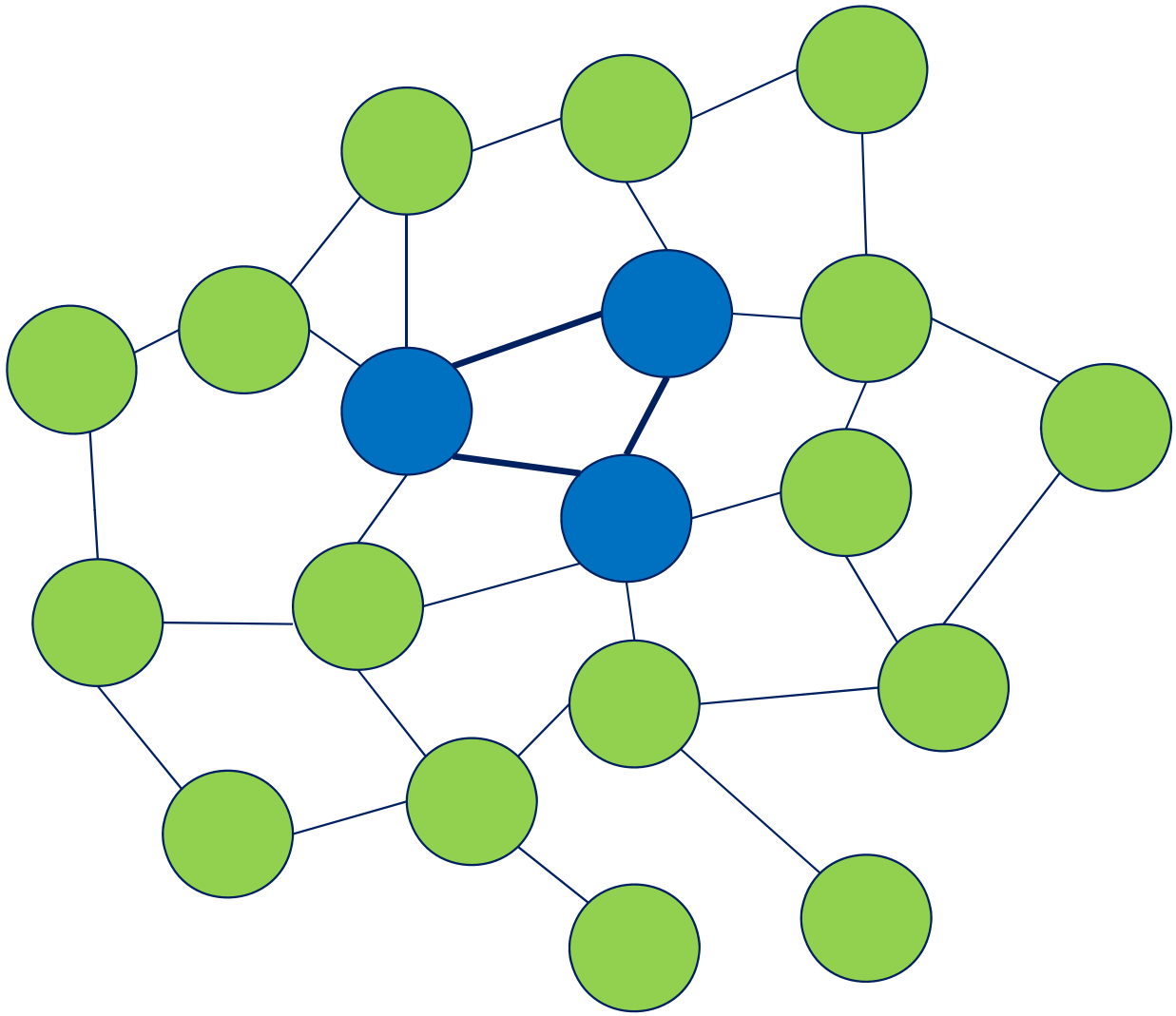
means that if you already know the value of Z, learning that of Y tells you nothing more about X. Any dependence between X and Y is indirect, mediated through Z.

It proves useful to represent conditional independences graphically.

Markov random fields: the local Markov property



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.

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

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

and under an additional condition implied by positivity they are all equivalent.

The Hammersley–Clifford theorem

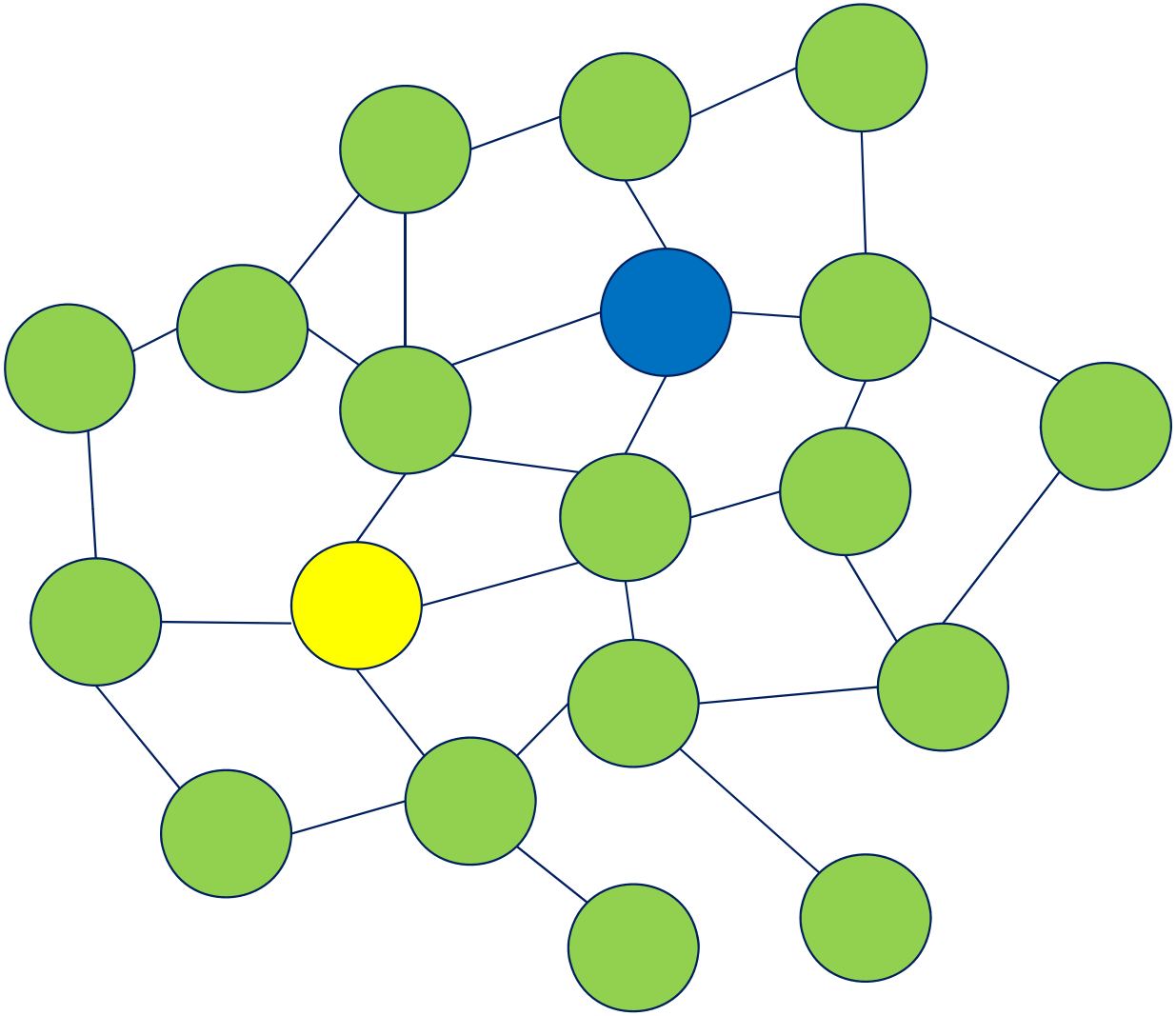
The result that Markov random fields coincided with Gibbs distributions, under certain conditions, was known as the Hammersley–Clifford theorem.

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

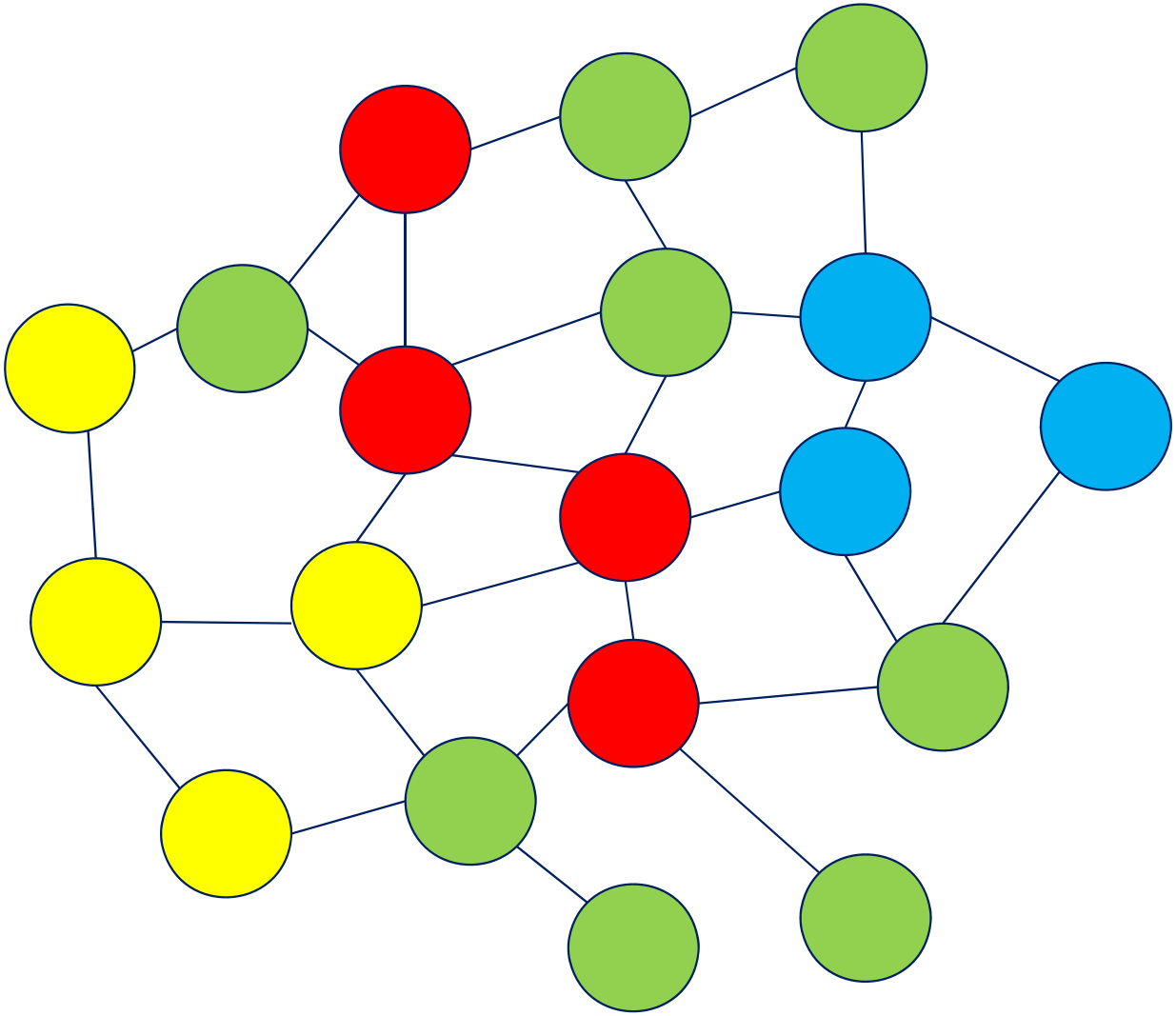
$$F \implies G \implies L \implies P$$

and under an additional condition implied by positivity they are all equivalent.

Pairwise Markov property



Global Markov property



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 conditions (positivity is sufficient), 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 conditions (positivity is sufficient), global and local Markov properties also hold.

Structural learning

Given i.i.d. observations on X, we are often interested in inferring $\mathcal{G}$, the problem of structural learning.

$\mathcal{G}$ may be of direct interest; also determining $\mathcal{G}$ as part of inference about covariance is a way of imposing parsimony.

This entails search in a huge discrete model space: there are

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

graphs on v vertices.

Structural learning

Given i.i.d. observations on X, we are often interested in inferring $\mathcal{G}$, the problem of structural learning.

$\mathcal{G}$ may be of direct interest; also determining $\mathcal{G}$ as part of inference about covariance is a way of imposing parsimony.

This entails search in a huge discrete model space: there are

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

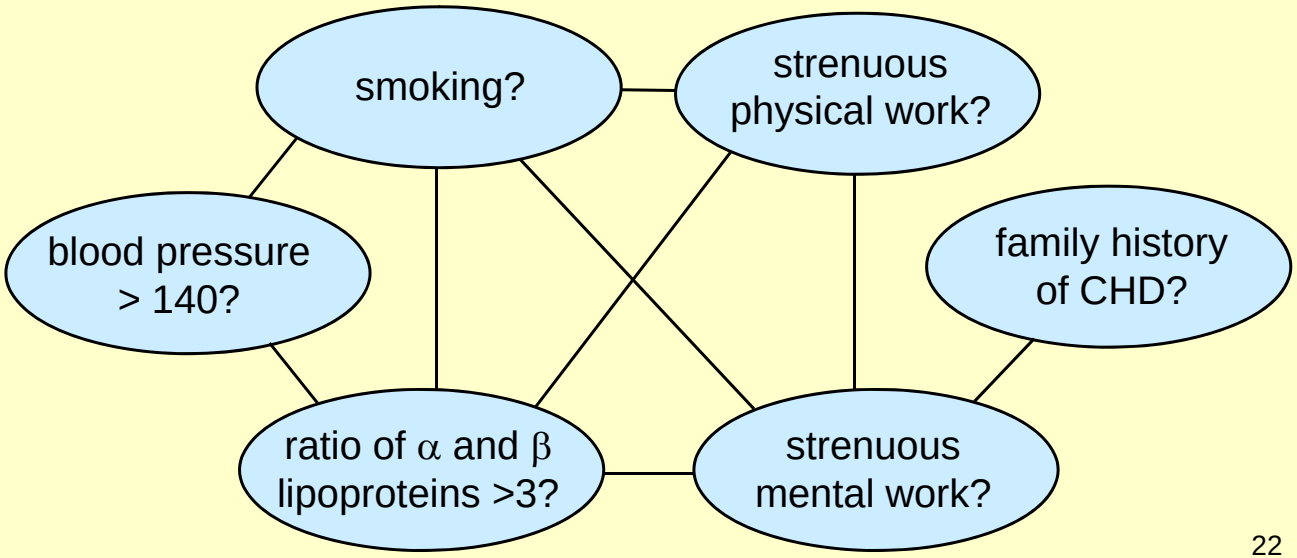
graphs on v vertices.

Contingency tables

Prognostic factors for coronary heart disease

Analysis of a 2^6 contingency table

(Edwards & Havranek, *Biometrika,* 1985)



22

Structural learning

The main approaches

- Score-based methods (e.g. optimisation of a penalised likelihood, such as `glasso` or BIC)
- Contraint-based methods (querying conditional independences, e.g. PC algorithm)
- Bayesian methods (deliver posterior probabilities over graphs (and parameters))

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

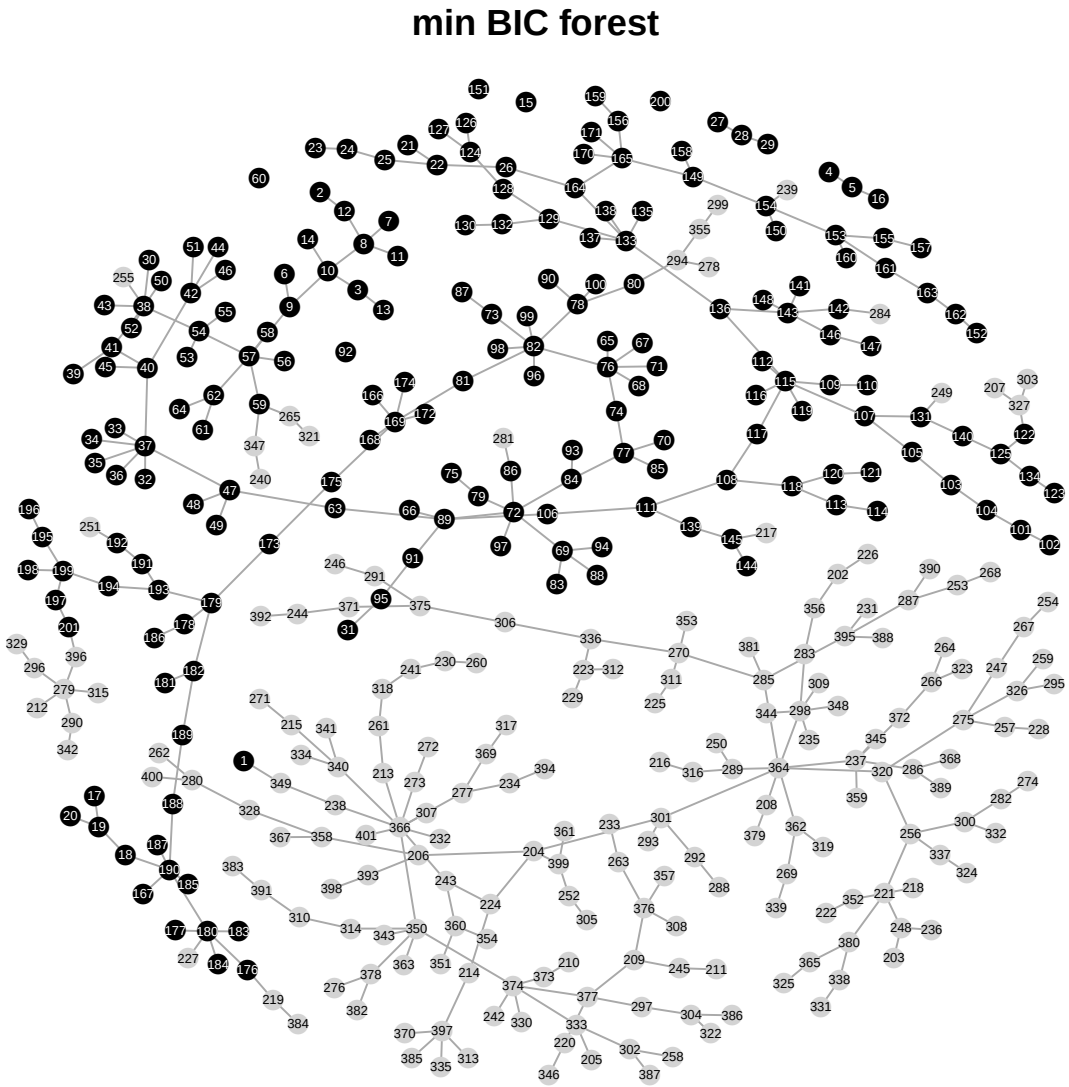
Structural learning

The main approaches

- Score-based methods (e.g. optimisation of a penalised likelihood, such as `glasso` or BIC)
- Contraint-based methods (querying conditional independences, e.g. PC algorithm)
- Bayesian methods (deliver posterior probabilities over graphs (and parameters))

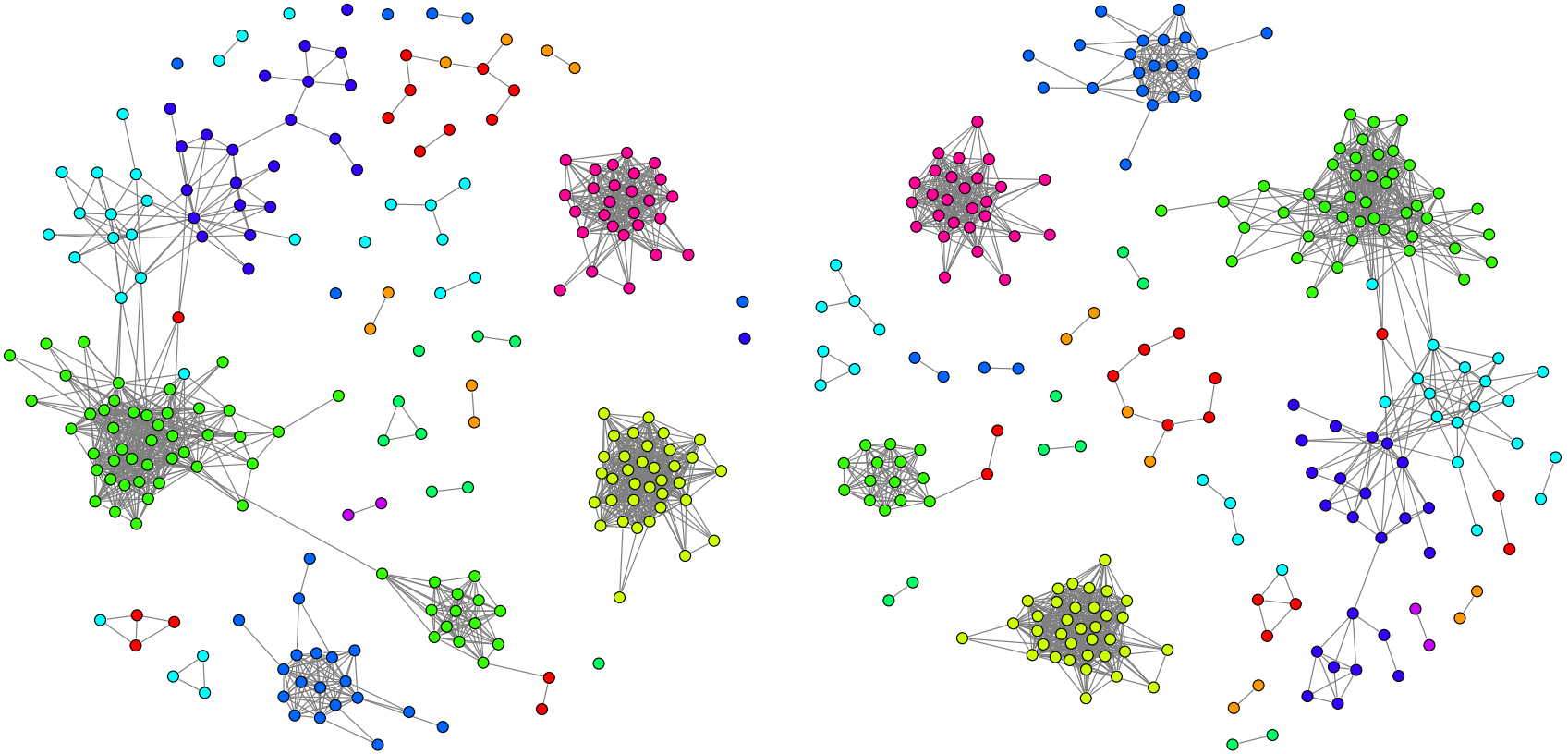
Except in very small problems, we typically restrict the space of graphs to be considered – e.g. to trees, forests, DAGs or decomposable graphs.

SNPs and gene expression



from Lauritzen (2012).

S&P 500 equity data



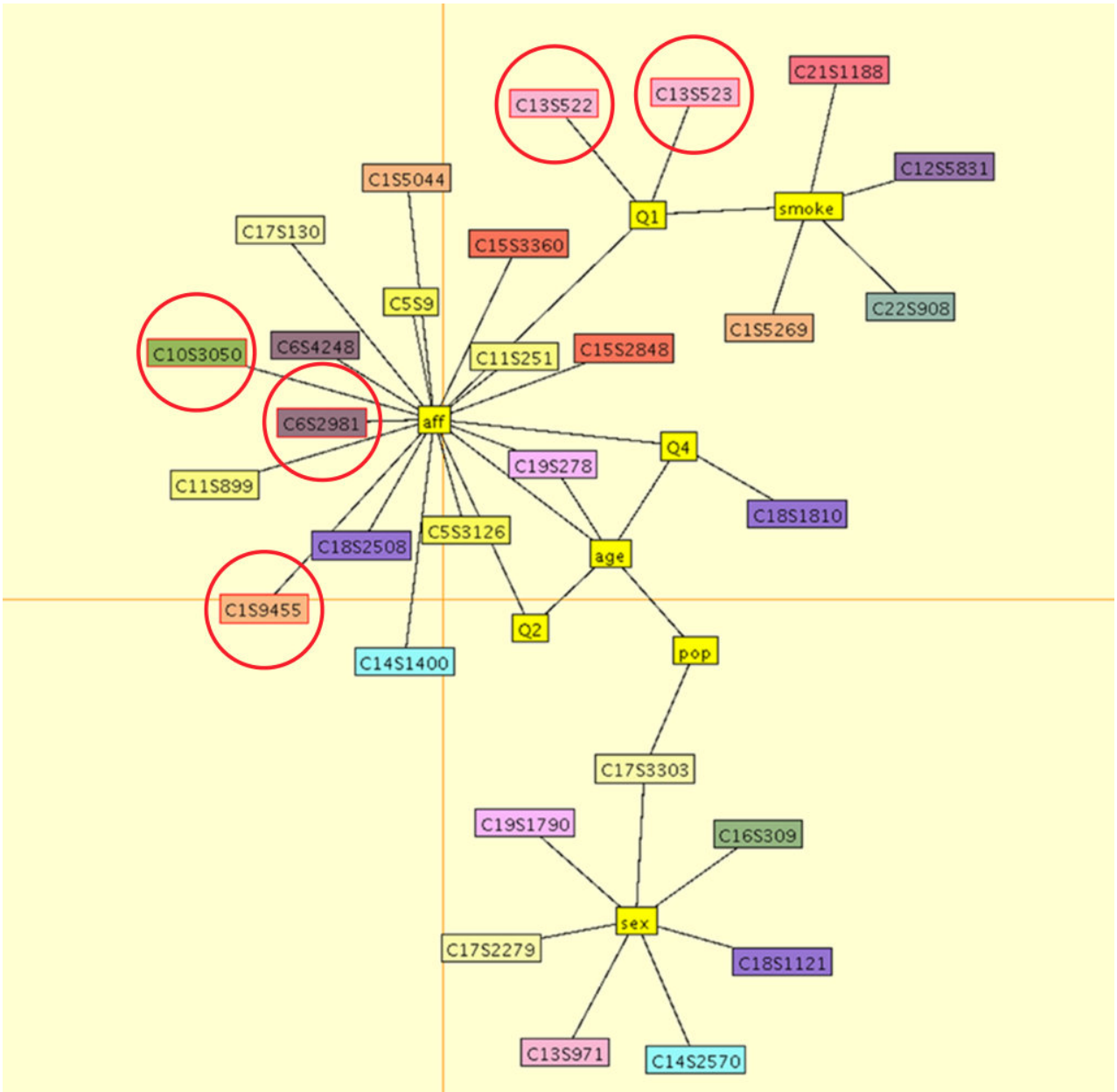
(a) glasso graph (1316 edges)

(b) nonparanormal graph (1316 edges)

from Lafferty, Liu, Wasserman (2012).

Genetic epidemiology

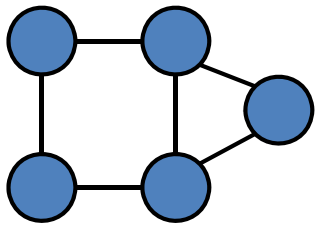
Graphical model fitted to contingency table relating disease status (aff), SNPs – with Linkage disequilibrium, covariates, and 4 quantitative traits. Abel & Thomas, GAW17.



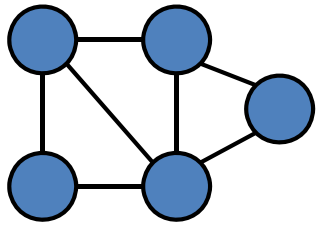
Decomposable graphical models

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

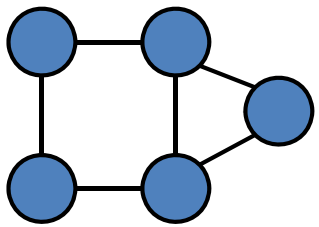


decomposable

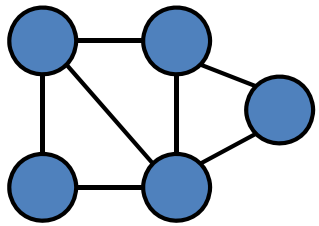
Decomposable graphical models

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

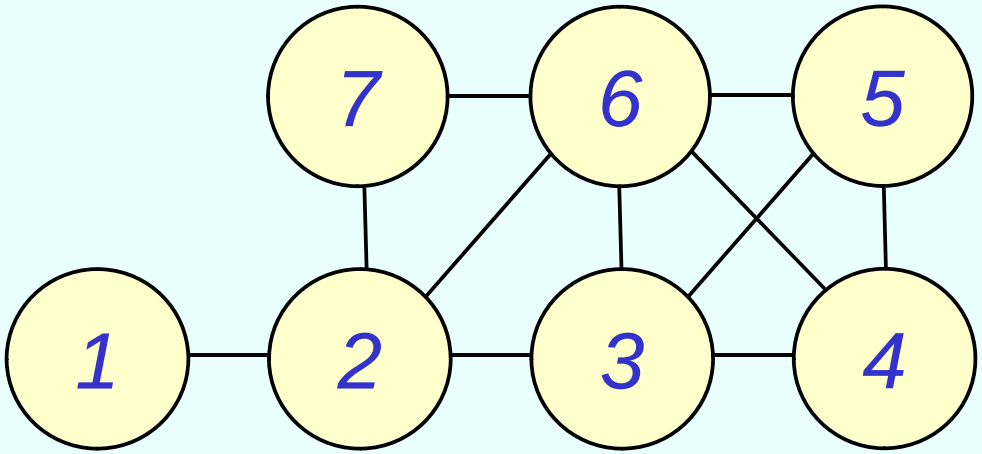
A graph is decomposable if and only if it has a junction tree representation.

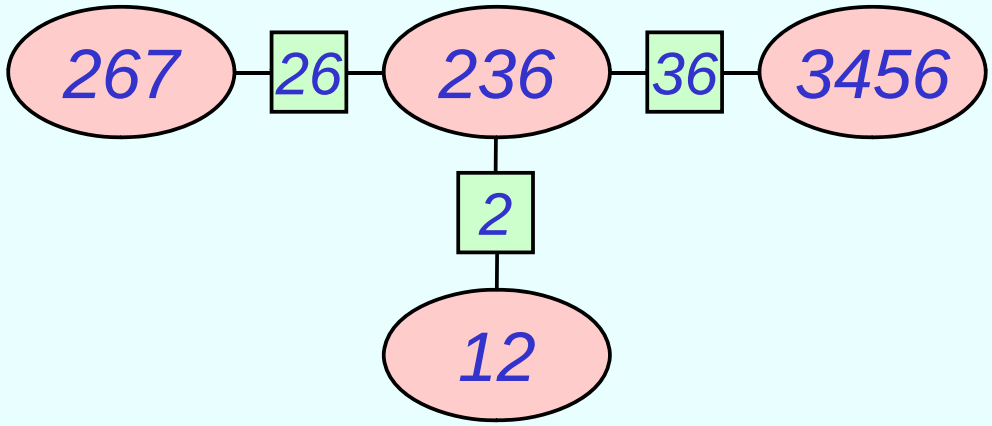
A junction tree is a graph whose vertices are cliques (maximal complete subgraphs), with the property that the cliques containing any prescribed set of vertices forms a connected sub-tree.

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.

A small decomposable graph

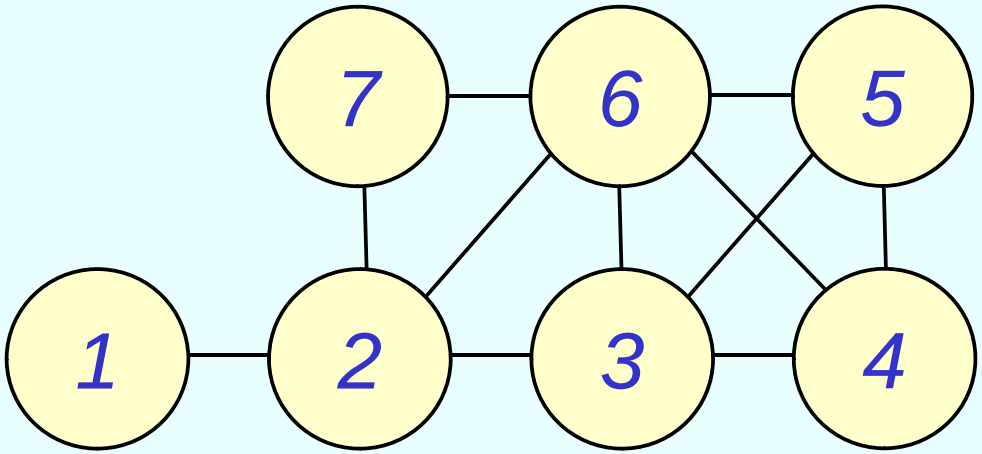
Non-uniqueness of junction tree

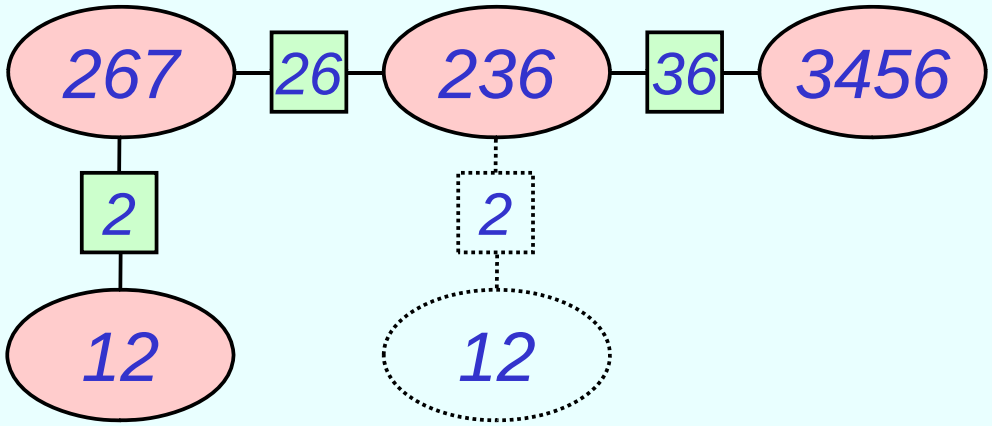




A small decomposable graph

Non-uniqueness of 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.

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.

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

Clique–separator factorisation yields dramatic speed-ups in structural learning.

Statistical significance of decomposability

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

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

Clique–separator factorisation yields dramatic speed-ups in structural learning.

Statistical significance of decomposability

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

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

Clique–separator factorisation yields dramatic speed-ups in structural learning.

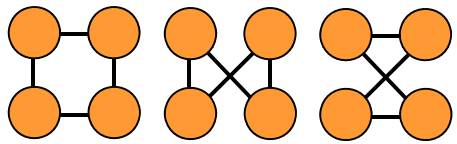
How restrictive is decomposability?

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?

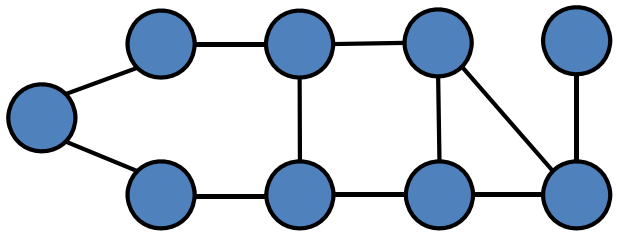
There is no reason why Nature should be kind enough to give us data from graphical models that are decomposable. . .

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.

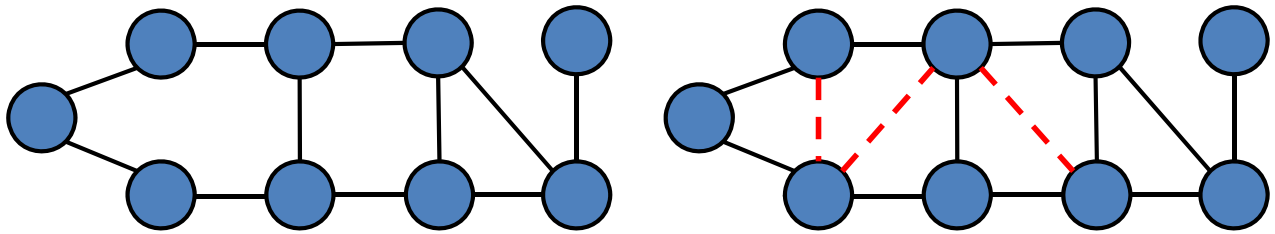


not decomposable

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.

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, a typical 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 $p(\mathcal{G}, \theta|\mathbf{X}) \propto p(\mathcal{G}, \theta, \mathbf{X})$.

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'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) if for any covering pair (A, B), we have :

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

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'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) if for any covering pair (A, B), we have :

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

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'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) if for any covering pair (A, B), we have :

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

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

Byrne shows 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'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) if for any covering pair (A, B), we have :

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

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

Byrne shows 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 :

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

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 :

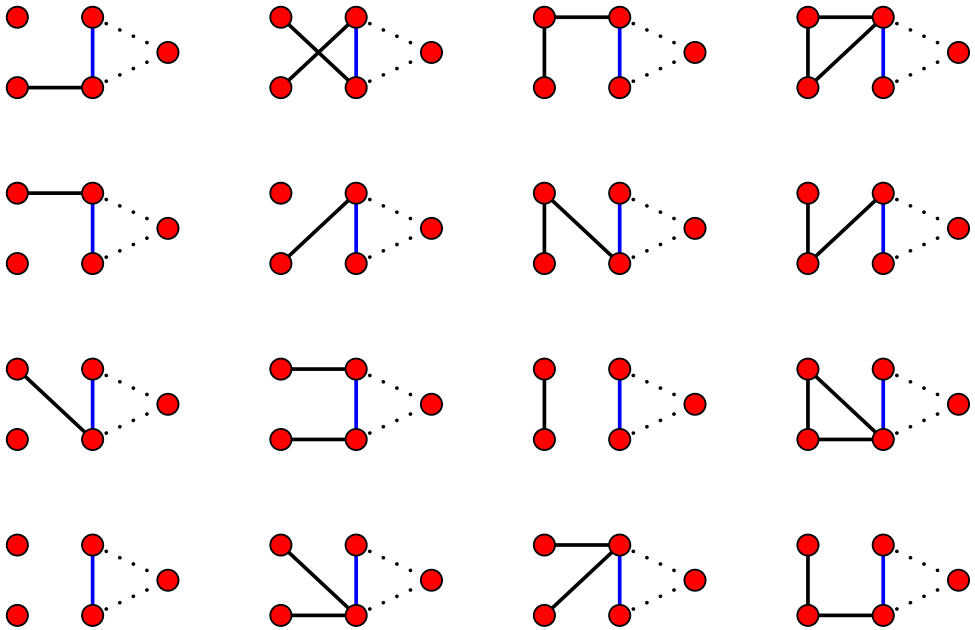
$$\mathcal{G}_A \perp \mathcal{G}_B \mid \{\mathcal{G} \in \mathfrak{U}^A(A, B)\} \quad [\pi],$$

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 $\mathcal{G}_A$.

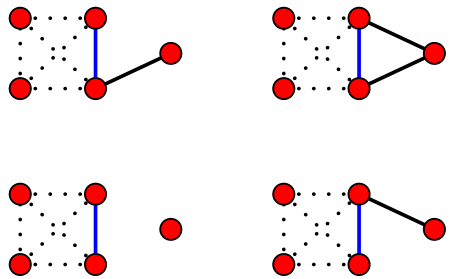
This places fewer conditional independence conditions on π, so potentially corresponds to a richer class of graph priors – but we will see that we can still say something concrete about the form of these laws.

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 $\mathcal{G}_B$



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

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

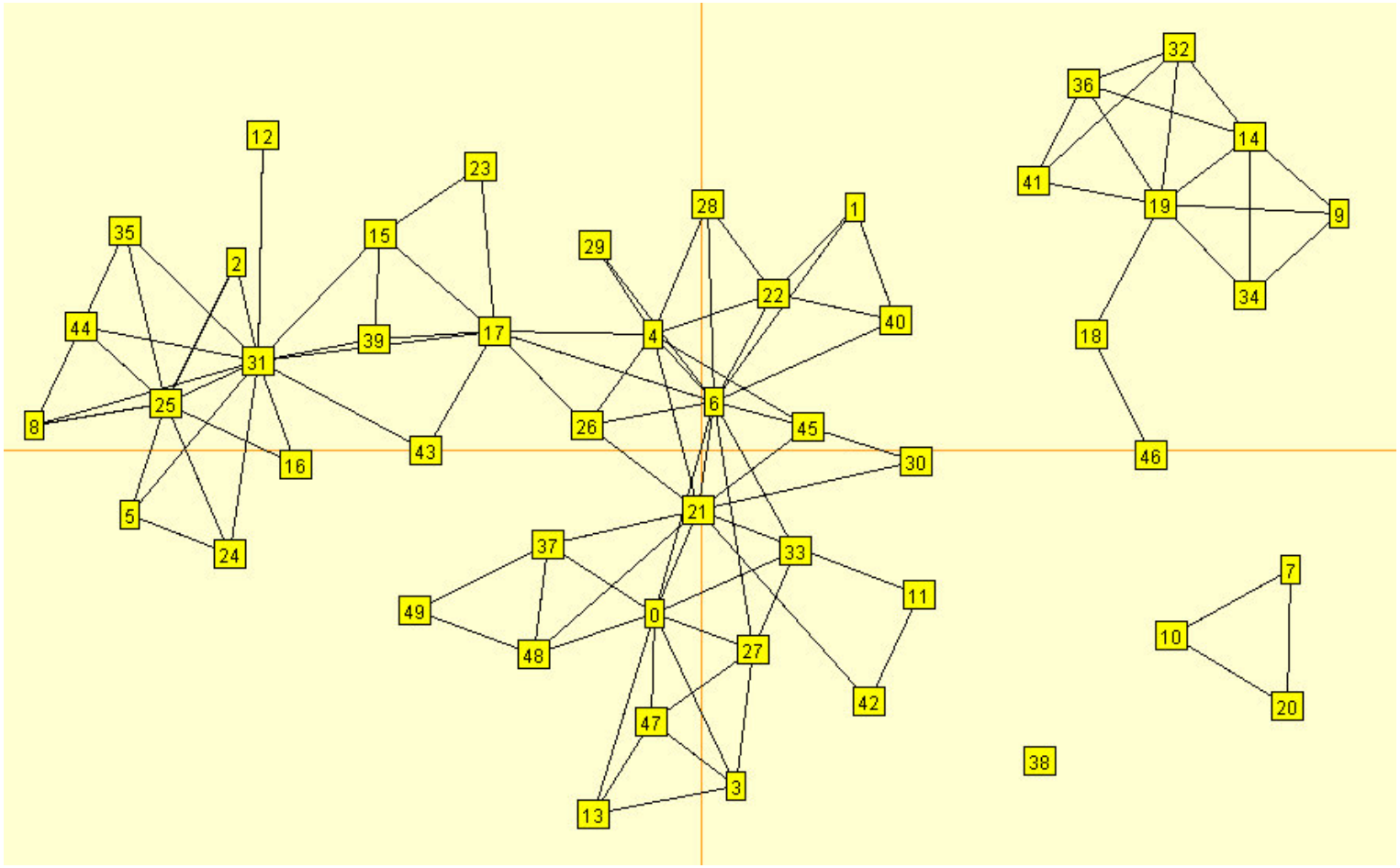
We can show that a graph law is weakly structurally Markov if and only if has the form

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

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

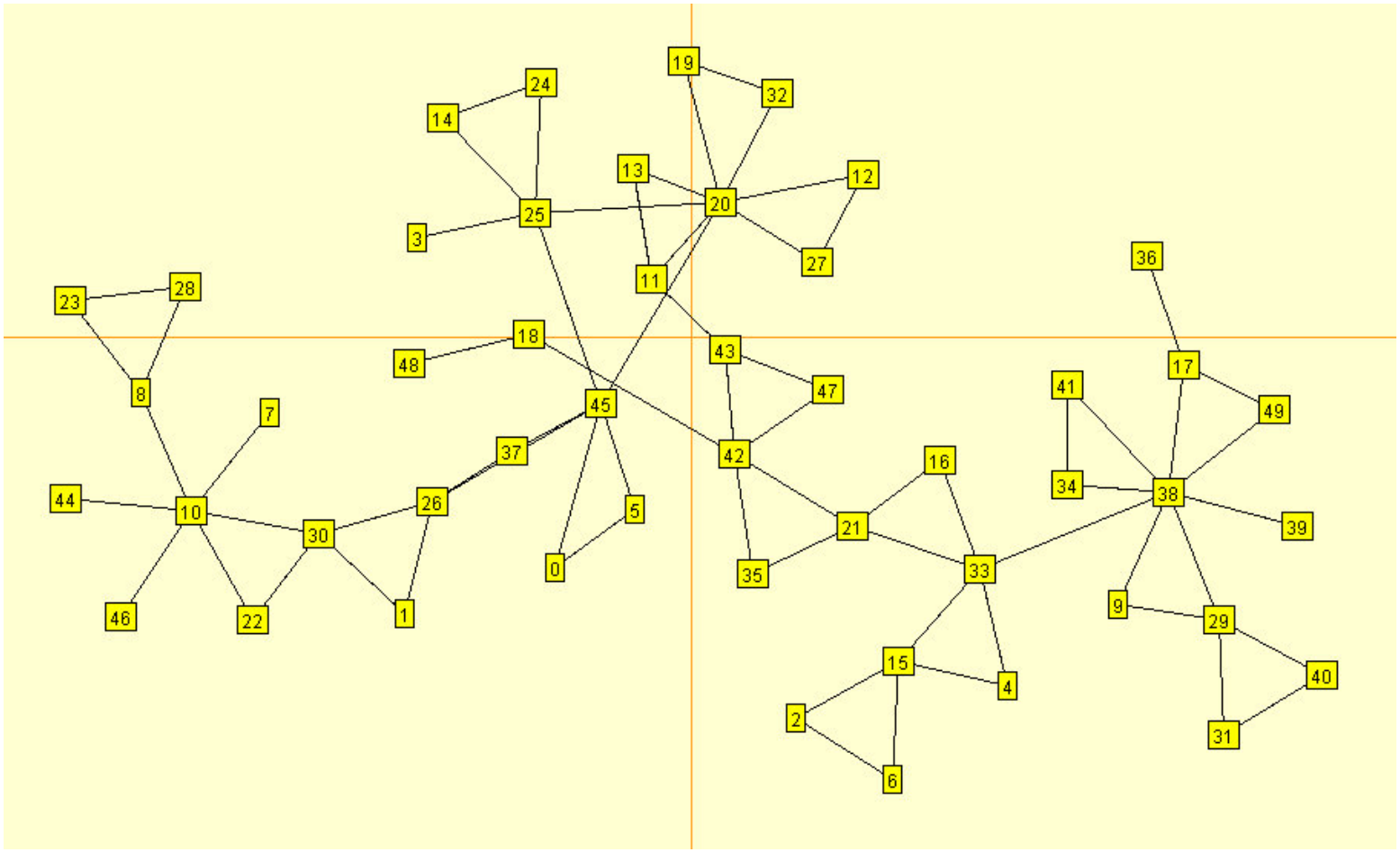
This more general form allows valuable extra flexibility in prior specification; this class of priors has also been studied by Bornn and Caron (2011).

Example sample from a CSF graph law



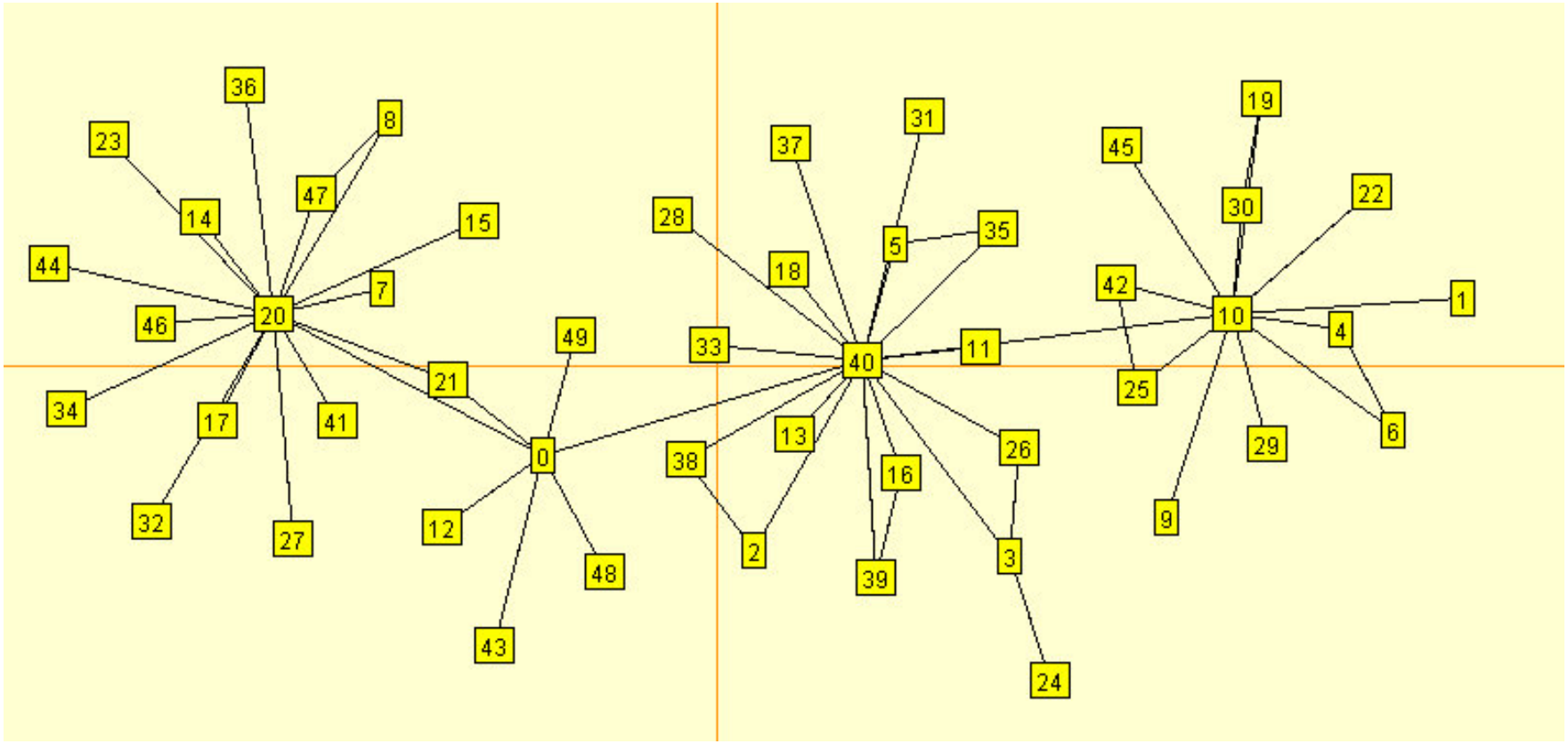
$\phi_C = \exp(4(|C| - 1))$ for $|C| \leq 4$, else 0; $\psi_S = \exp(4|S|)$

Example sample from a CSF graph law



$\phi_C = \exp(4(|C| - 1))$ for $|C| \leq 4$, else 0;
$\psi_S = \exp(4)$ for $|S| = 1$, else ∞

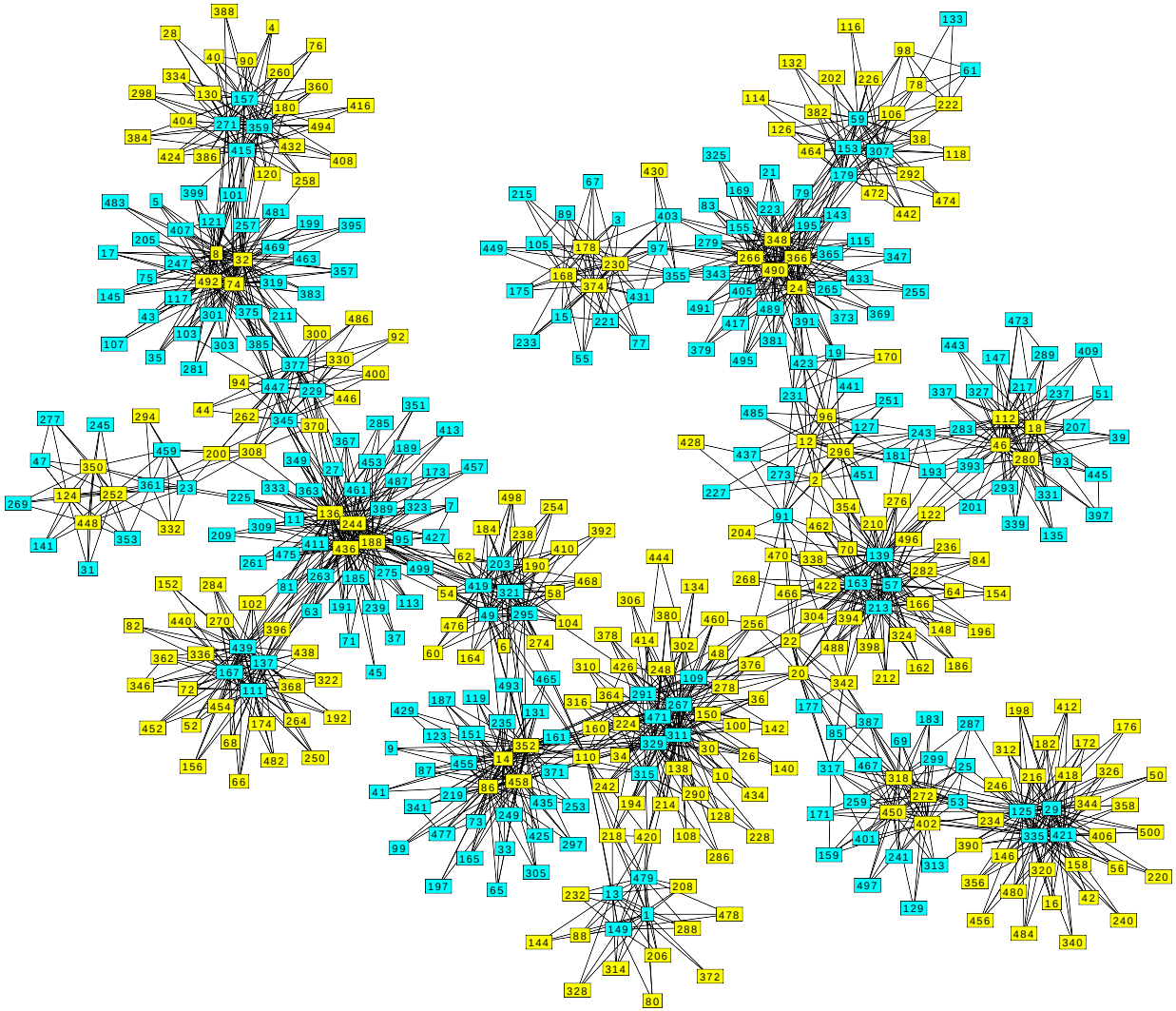
Example sample from a CSF graph law



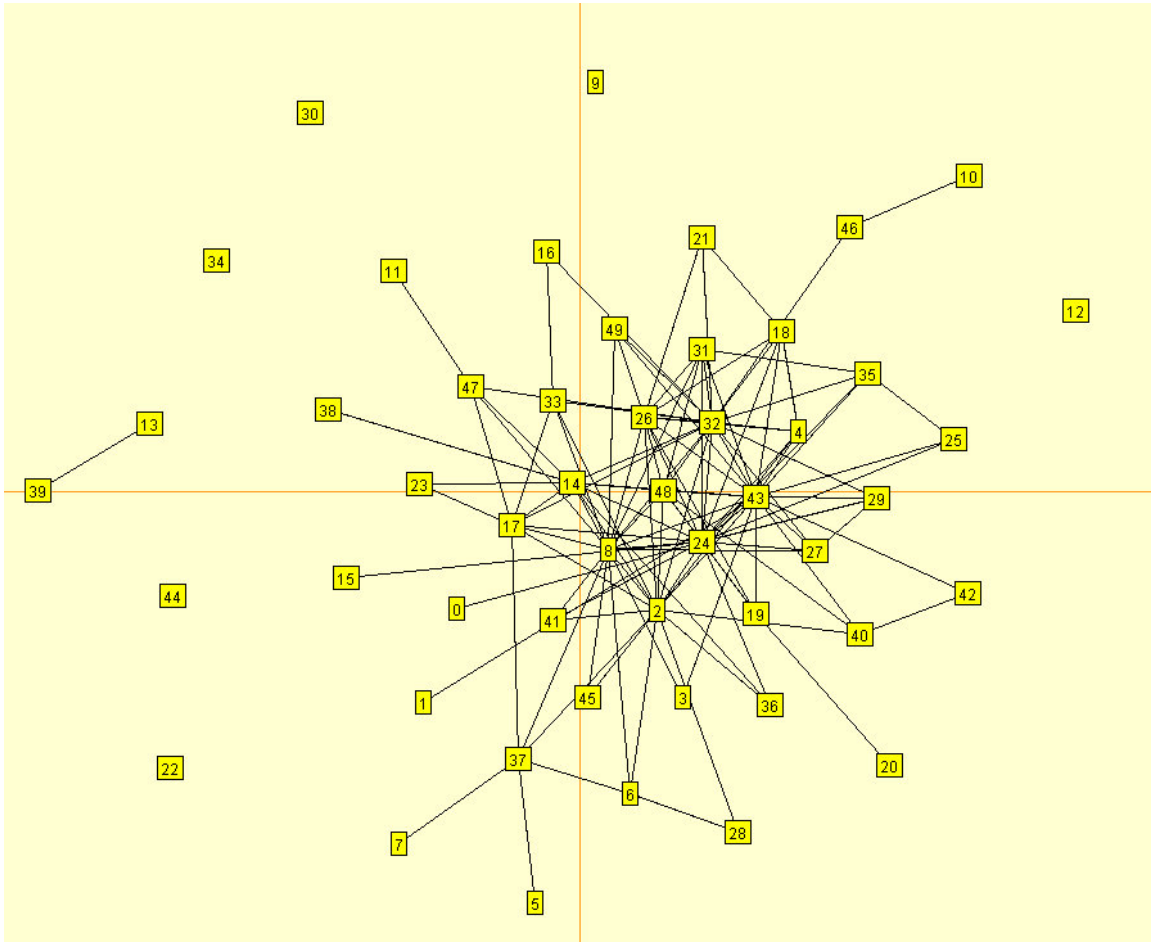
$\phi_C = \exp(4(|C| - 1) + 3\#\{v \in C : \mod(v, 10) = 0\})$ for $|C| \leq 4$, else 0;

$\psi_S = \exp(4)$ for $|S| = 1$, else ∞

Example sample from a CSF graph law



Example sample from an edge-penalty graph law



$\phi_C = \psi_C = \exp(-\alpha|C|(|C|-1)/2)$ with $\alpha = .75$ – i.e.
$\pi(\mathcal{G}) \propto \exp(-\alpha \#\text{edges})$

WSM=CSF – sketch proof

Consider a particular junction tree of $\mathcal{G}$, with junction tree links connecting C_j to $C_{h(j)}$ via separator S_j. For each j, let R_j be any subset of $C_{h(j)}$ that is a proper superset of S_j.

The conditional independence assertions of WSM imply both

- For any choice of such $\{R_j\}$, we have

$$\pi(\mathcal{G}) = \prod_j \pi(\mathcal{G}^{(C_j)}) \times \prod_{j \geq 2} \frac{\pi(\mathcal{G}^{(R_j, C_j)})}{\pi(\mathcal{G}^{(R_j)})\pi(\mathcal{G}^{(C_j)})}$$

  where $\mathcal{G}^{(\ldots)}$ is the graph with cliques ....

- $\pi(\mathcal{G}^{(R,C)})/\pi(\mathcal{G}^{(R)})\pi(\mathcal{G}^{(C)})$ depends only on S, for all sets of vertices R, C for which $R \cup C = V$ and $R \cap C = S$, and where both R and C are strict supersets of S.

WSM=CSF – sketch proof

Consider a particular junction tree of $\mathcal{G}$, with junction tree links connecting C_j to $C_{h(j)}$ via separator S_j. For each j, let R_j be any subset of $C_{h(j)}$ that is a proper superset of S_j.

The conditional independence assertions of WSM imply both

- For any choice of such $\{R_j\}$, we have

$$\pi(\mathcal{G}) = \prod_j \pi(\mathcal{G}^{(C_j)}) \times \prod_{j \geq 2} \frac{\pi(\mathcal{G}^{(R_j, C_j)})}{\pi(\mathcal{G}^{(R_j)})\pi(\mathcal{G}^{(C_j)})}$$

  where $\mathcal{G}^{(\ldots)}$ is the graph with cliques ....
- $\pi(\mathcal{G}^{(R,C)})/\pi(\mathcal{G}^{(R)})\pi(\mathcal{G}^{(C)})$ depends only on S, for all sets of vertices R, C for which $R \cup C = V$ and $R \cap C = S$, and where both R and C are strict supersets of S.

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 therefore 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.

See Giudici & G (1999) (Gaussian case) and Giudici, G & Tarantola (2000) (contingency table case). These assume parameter priors $p(\theta|\mathcal{G})$ that are consistent across $\mathcal{G}$, using the Dawid & Lauritzen hyper-Markov laws.

Bayesian decomposable graphical model determination

For decomposable graphs, joint structural/quantitative learning therefore 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.

See Giudici & G (1999) (Gaussian case) and Giudici, G & Tarantola (2000) (contingency table case). These assume parameter priors $p(\theta|\mathcal{G})$ that are consistent across $\mathcal{G}$, using the Dawid & Lauritzen hyper-Markov laws.

Bayesian decomposable graphical model determination

Typically, a move involves proposing a single-edge perturbation to the graph $\mathcal{G}$, together with appropriate changes to θ. In MCMC sampling using single-edge moves, a junction tree representation of the current $\mathcal{G}$ permits both

- cheap pre-testing that the proposed new graph $\mathcal{G}'$ is decomposable
- fast local updating of the graph from $\mathcal{G}$ to $\mathcal{G}'$ when the move passes the Metropolis–Hastings acceptance test

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 $\mathcal{G}$.

Pre-tests for maintaining decomposability

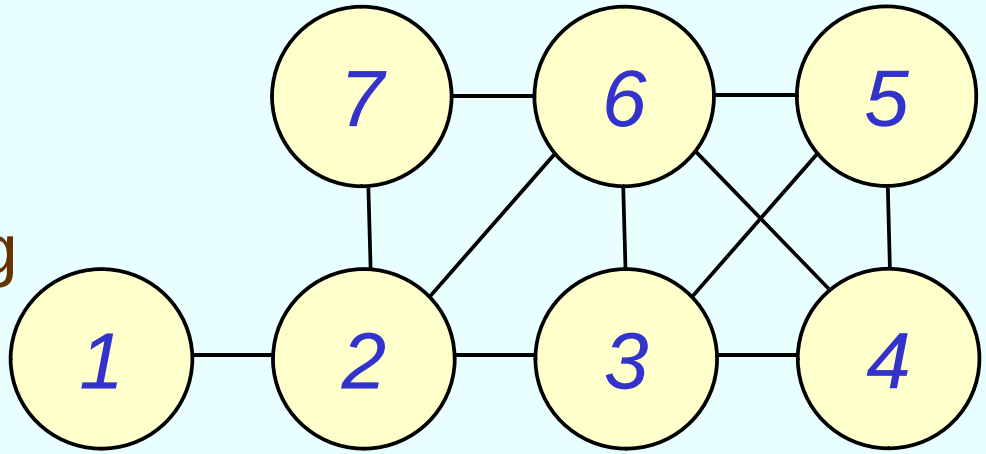
Conditions for maintaining decomposability in single-edge moves:

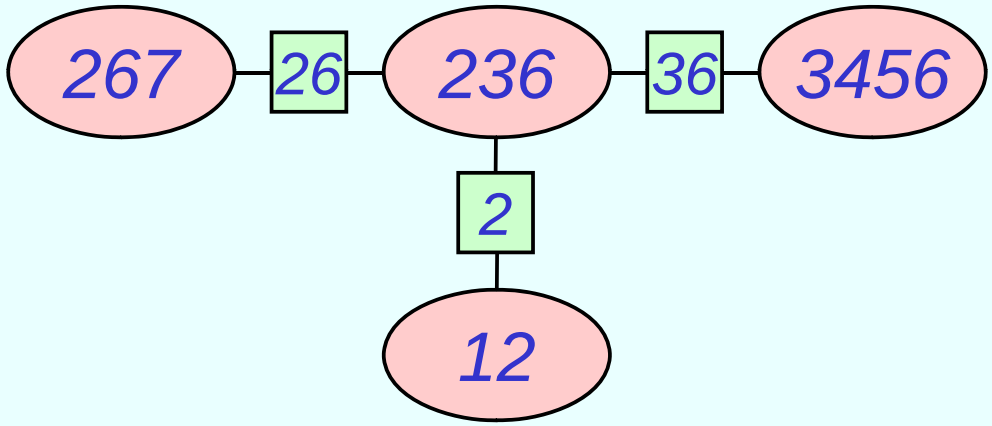
Frydenberg & Lauritzen Disconnecting x and y by removing an edge (x, y) from $\mathcal{G}$ will result in a decomposable graph if and only if x and y are contained in exactly one clique.

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 $\mathcal{G}$.

Single edge move

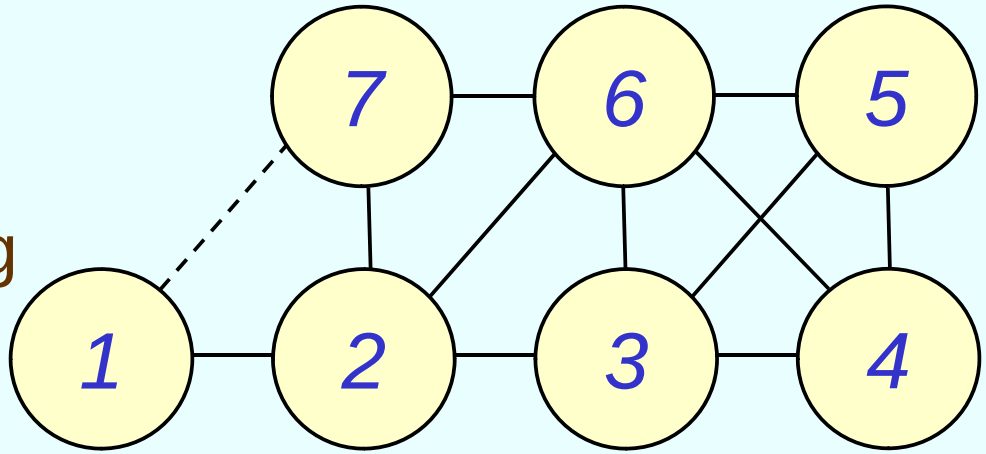
Once the test is complete, actually committing to adding or deleting the edge is little work



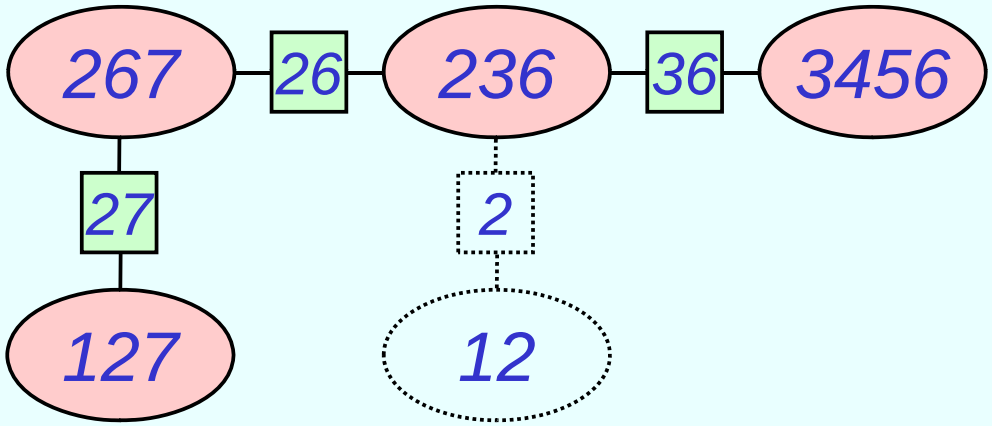


Single edge move

Once the test is complete, actually committing to adding or deleting the edge is little work



It makes only a (relatively) local change to the junction tree



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 $\mu(\mathcal{G})$.

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 $\mu(\mathcal{G})$.

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 $\mu(\mathcal{G})$.

Using the junction tree as the state

Trade-off between

- faster, more restrictive choice of proposed vertex pairs (x, y) specifying edges to be added/deleted, and avoidance of the manipulation from one junction tree to another, and
- we do not allow some edge moves that would yield a decomposable graph, because the junction tree needs to be manipulated, so the space of possible (junction tree) states of the chain is less connected.

Certain multiple-edge moves that change the topology of the junction tree in a simple way can also be included.

Using the junction tree as the state

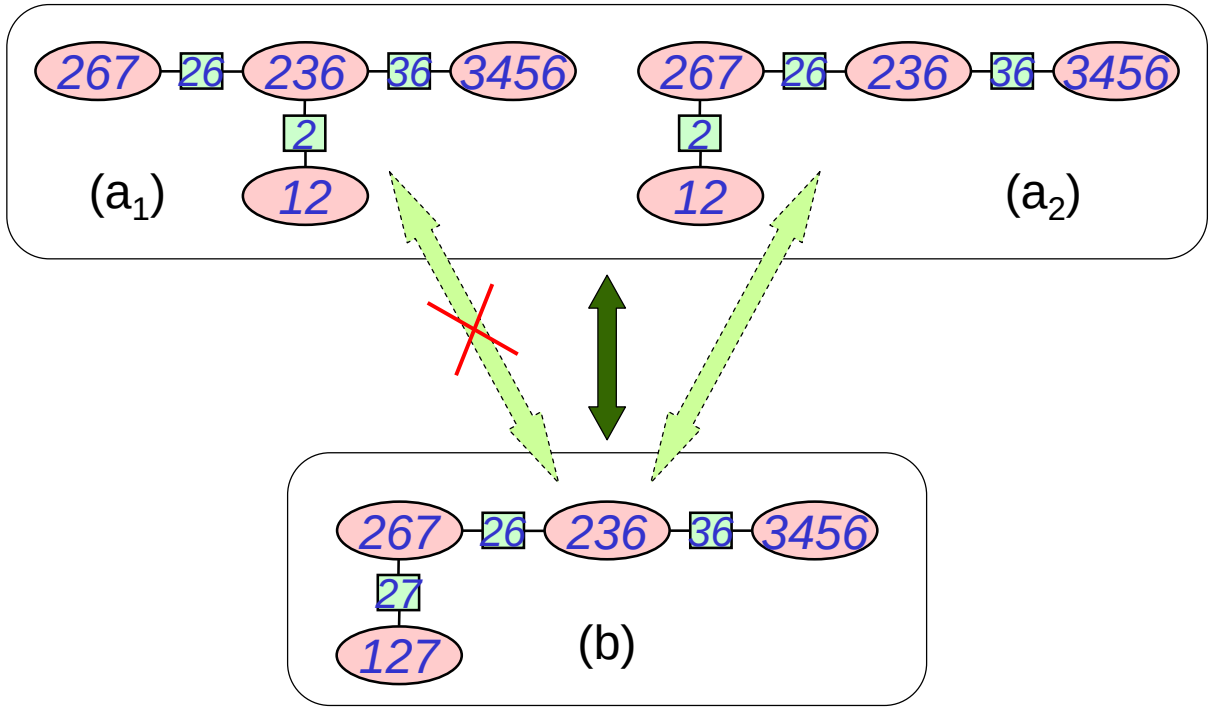
Trade-off between

- faster, more restrictive choice of proposed vertex pairs (x, y) specifying edges to be added/deleted, and avoidance of the manipulation from one junction tree to another, and
- we do not allow some edge moves that would yield a decomposable graph, because the junction tree needs to be manipulated, so the space of possible (junction tree) states of the chain is less connected.

Certain multiple-edge moves that change the topology of the junction tree in a simple way can also be included.

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.



A graphical Gaussian intra-class model

Given a decomposable graph $\mathcal{G}$ on v vertices labelled $1, 2, \ldots, v$, and real scalar parameters $\sigma^2 > 0$ and ρ, we define a non-negative definite matrix $V = V_{\mathcal{G}}(\sigma^2, \rho)$ by

$$V_{ij} = \begin{cases} \sigma^2 & \text{if } i = j \\ \rho\sigma^2 & \text{if } (i,j) \text{ is an edge in } \mathcal{G}, \end{cases}$$

and $(V^{-1})_{ij} = 0$ if (i,j) is not an edge in $\mathcal{G}$.

By Grone et al (1984), since $\mathcal{G}$ is decomposable and V restricted to each clique is positive definite, V exists and is unique, in fact the unique completion of the specified entries that is positive definite; it is the variance matrix of a v–variate Gaussian distribution for which $\mathcal{G}$ is the conditional independence graph. We call this the graphical Gaussian intra-class model (GGIM).

A graphical Gaussian intra-class model

Given a decomposable graph $\mathcal{G}$ on v vertices labelled $1, 2, \ldots, v$, and real scalar parameters $\sigma^2 > 0$ and ρ, we define a non-negative definite matrix $V = V_{\mathcal{G}}(\sigma^2, \rho)$ by

$$V_{ij} = \begin{cases} \sigma^2 & \text{if } i = j \\ \rho\sigma^2 & \text{if } (i,j) \text{ is an edge in } \mathcal{G}, \end{cases}$$

and $(V^{-1})_{ij} = 0$ if (i,j) is not an edge in $\mathcal{G}$.

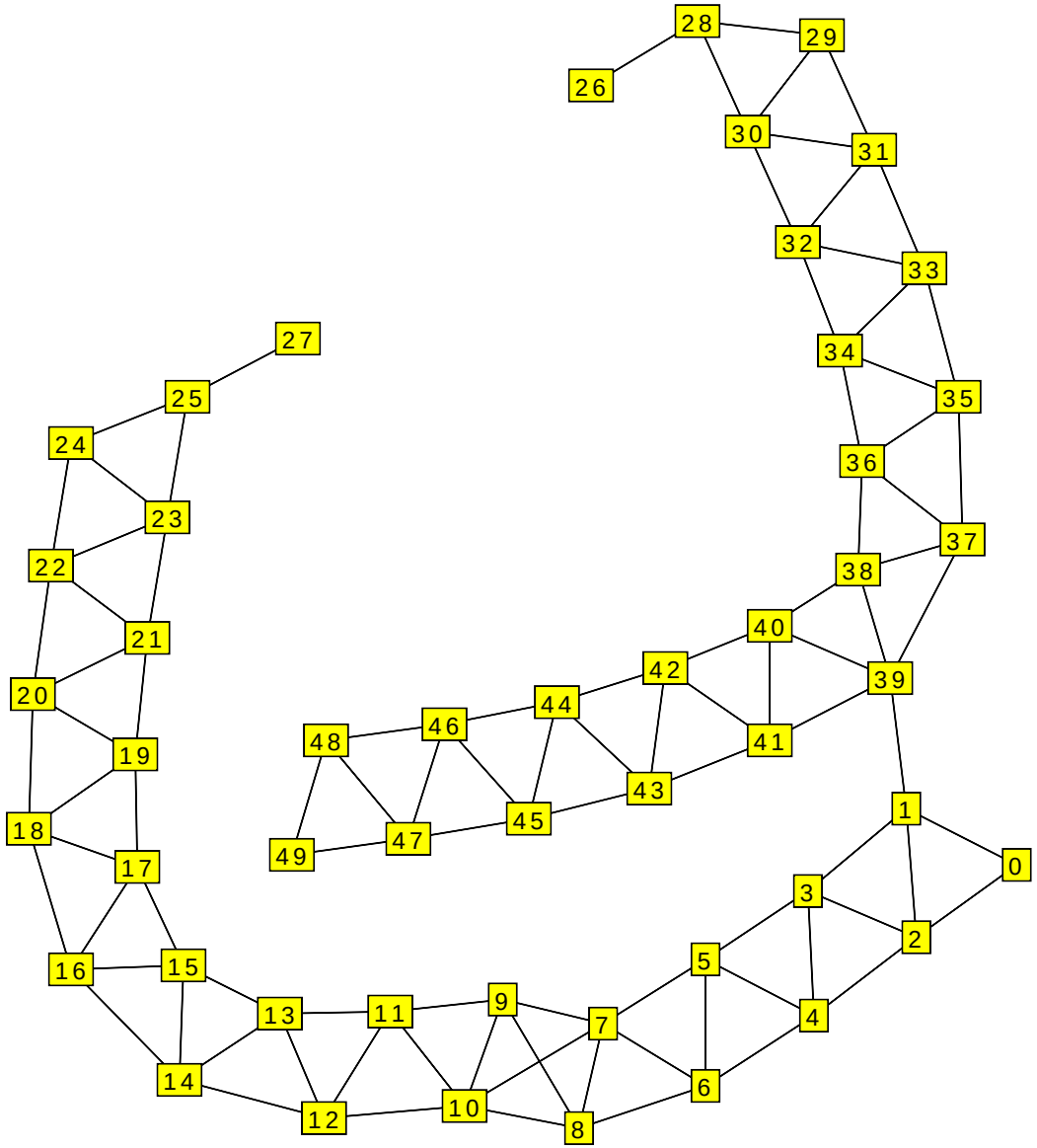
By Grone et al (1984), since $\mathcal{G}$ is decomposable and V restricted to each clique is positive definite, V exists and is unique, in fact the unique completion of the specified entries that is positive definite; it is the variance matrix of a v–variate Gaussian distribution for which $\mathcal{G}$ is the conditional independence graph. We call this the graphical Gaussian intra-class model (GGIM).

A 50-vertex graphical Gaussian intra-class model

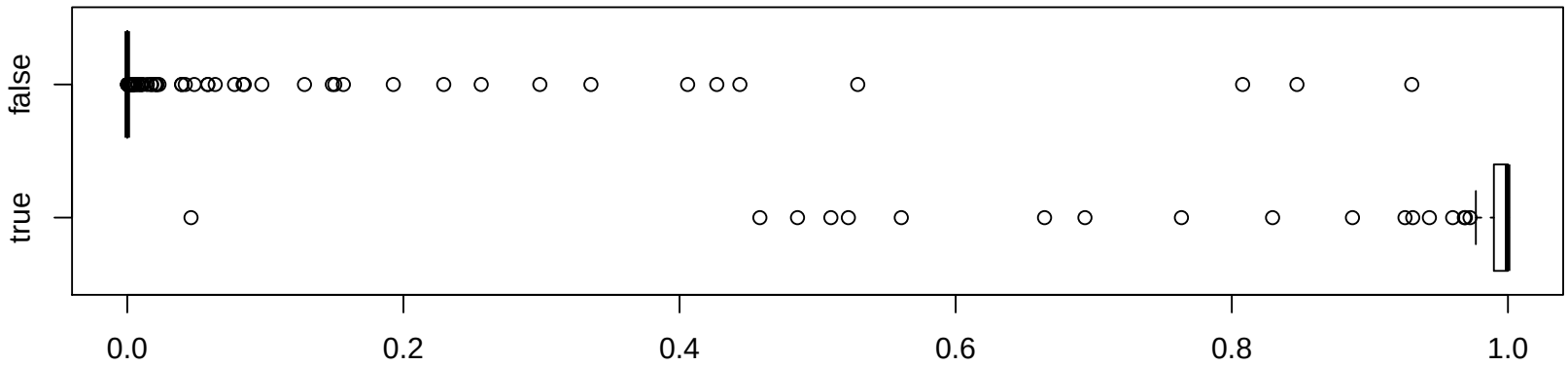
We simulated 1000 GGIM observations on 50 variables with $\sigma^2 = 30$ and $\rho = 0.2$. We used a second order Markov Chain graphical structure, that is, $(V^{-1})_{ij} = 0$ for all i and j such that $|i - j| > 2$.

A 50-vertex graphical Gaussian intra-class model

A graph typical of the type sampled early in their runs by all three samplers for the GGIM model. The edge between variables 1 and 39 is spurious, and has to be removed before the correct edges near variables 25 and 26 can be added.



A 50-vertex graphical Gaussian intra-class model



(posterior probabilities of edge presence for the 95 ‘true’ edges and the 1130 ‘false’ ones)

A graphical model for Linkage Disequilibrium

Abel & Thomas (*SAGMB*, 2011), Thomas & Camp (*Amer J Hum Gen*, 2004)

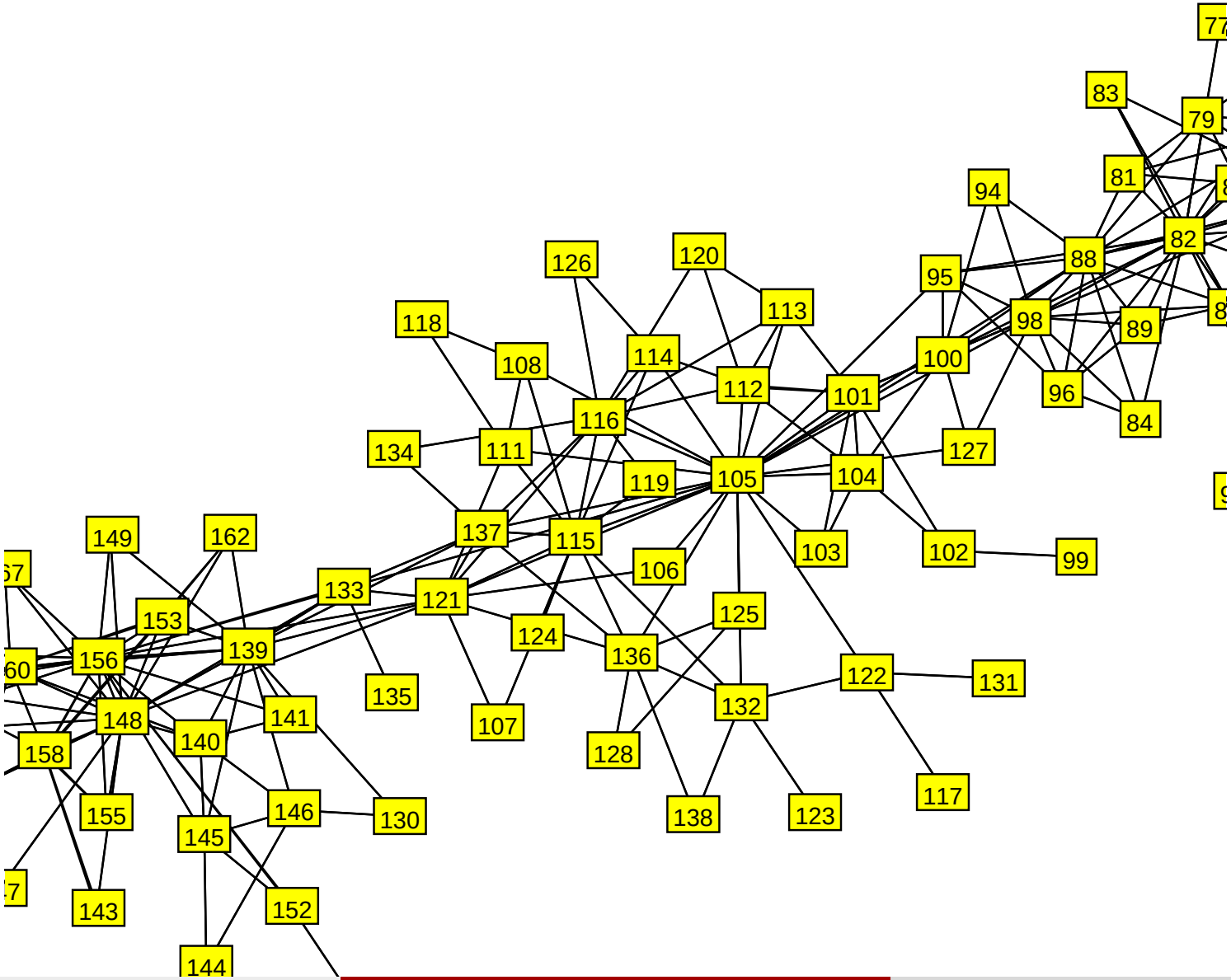
- genotype SNP data, unphased
- multinomial model, unknown graph (coding LD)
- assumes graph decomposable, sets max edge length (e.g. 15–40)
- multinomial cell probabilities maximised out
- sampler alternates between updating graph, and imputing phase and other missing data
- Abel & Thomas demonstrate up to 100,000 loci on 60 individuals, and on 500 loci up to 12,500 individuals.
- Example: first 500 loci on chromosome 1 for the 60 unrelated parents in the original HapMap Yoruba data set

A graphical model for Linkage Disequilibrium

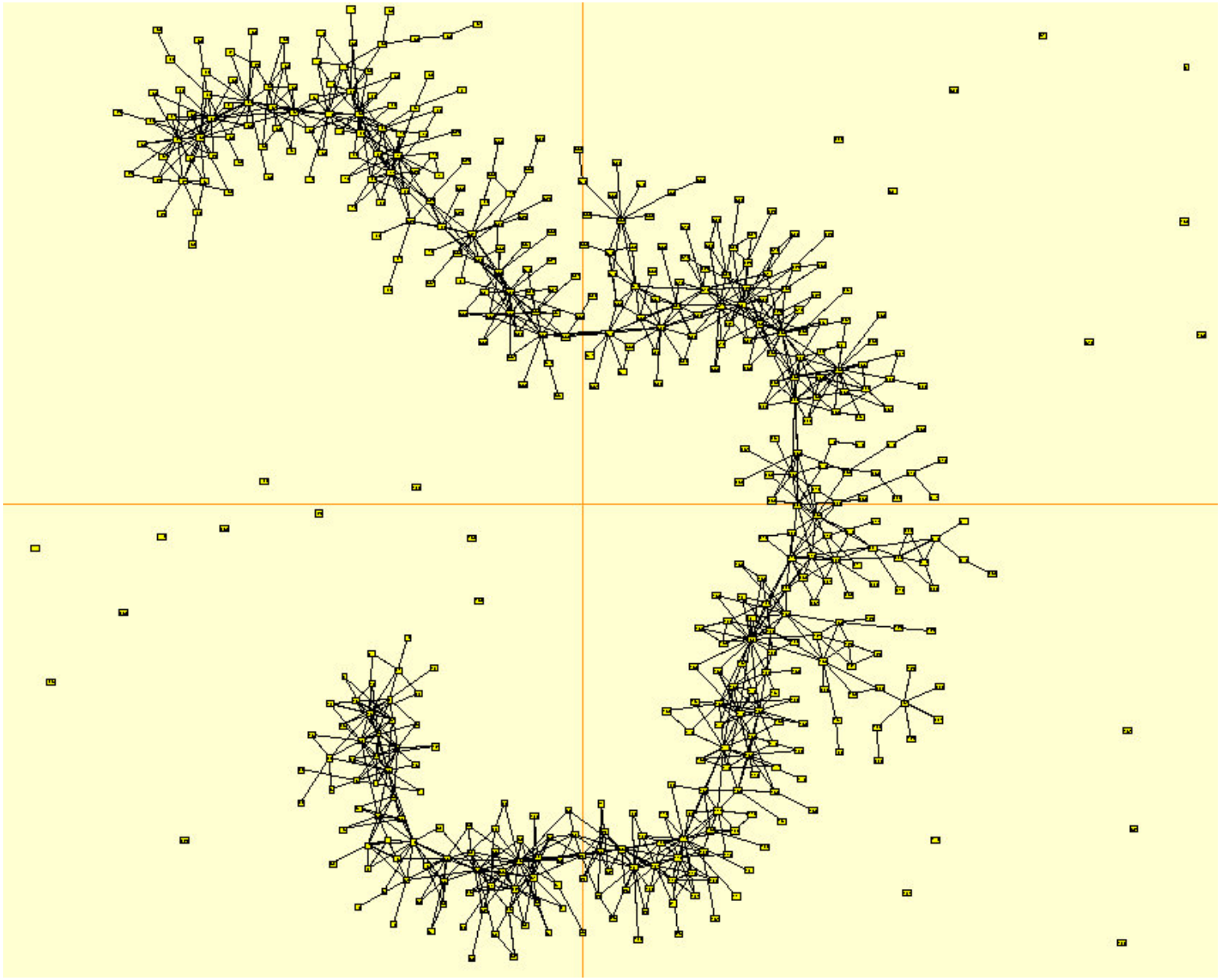
Abel & Thomas (*SAGMB*, 2011), Thomas & Camp (*Amer J Hum Gen*, 2004)

- genotype SNP data, unphased
- multinomial model, unknown graph (coding LD)
- assumes graph decomposable, sets max edge length (e.g. 15–40)
- multinomial cell probabilities maximised out
- sampler alternates between updating graph, and imputing phase and other missing data
- Abel & Thomas demonstrate up to 100,000 loci on 60 individuals, and on 500 loci up to 12,500 individuals.
- Example: first 500 loci on chromosome 1 for the 60 unrelated parents in the original HapMap Yoruba data set

A graphical model for Linkage Disequilibrium



A graphical model for Linkage Disequilibrium



A graphical model for Linkage Disequilibrium

- implemented with some clever use of moving windows
- attains cross-validation accuracy in imputing missing genotypes comparable to best alternatives
- more effort and flexibility in modelling allelic associations may be important in difficult data sets
- computationally expensive
- efficient use of memory
  - very large panels
  - fitted model phases haplotypes and imputes missing data on a huge scale very quickly

Non-decomposable graphs

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):

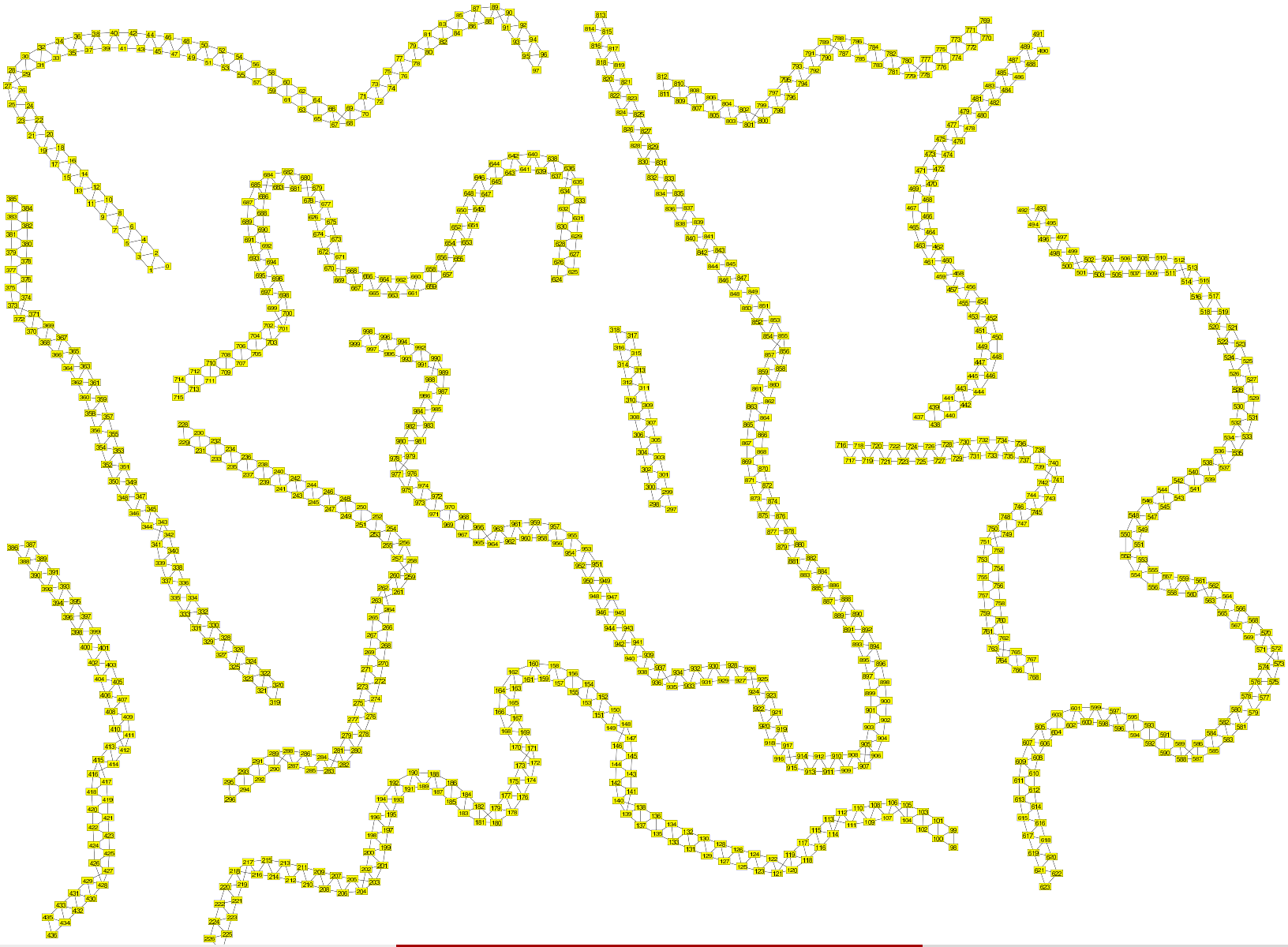
- The normalising constants in the non-complete prime component marginals do not have closed form, so we need Monte Carlo methods to estimate them.
- These Monte Carlo calculated values have high variance.
- When you make single-edge perturbations to the graph, there is no guarantee of significant cancellations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

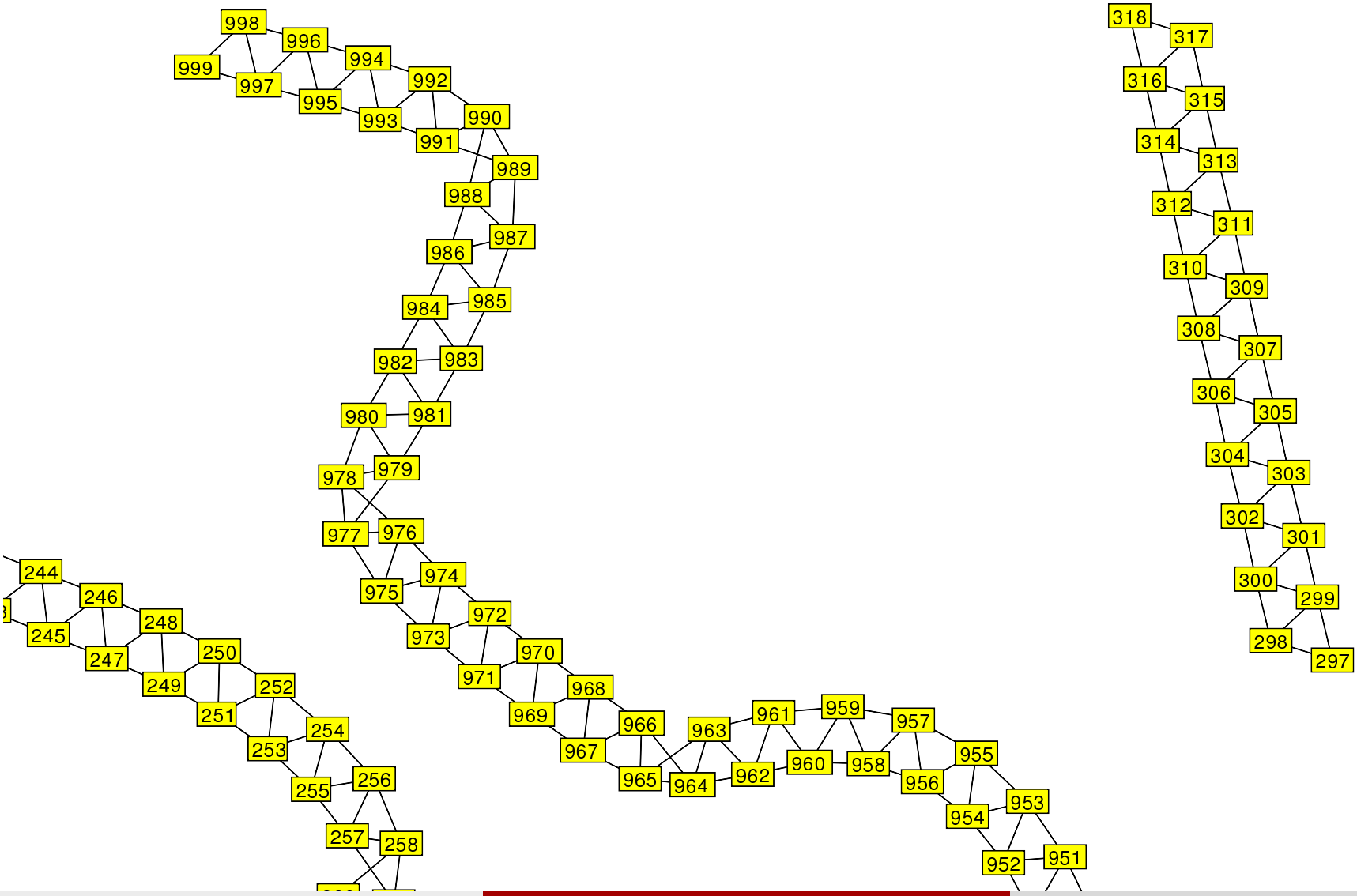
Bayesian model determination with non-decomposable graphs

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.

A 1000-vertex graphical Gaussian intra-class model



A 1000-vertex graphical Gaussian intra-class model



Some issues for future work

- Parallelisation
- Latent variables
- ‘Nearly decomposable’ graphs?
- Decision theory approach to delivering ‘optimal’ graph – with loss function on presence of individual edges – ignore decomposability constraint?
- Perfect simulation

- "Sampling decomposable graphs using a Markov chain on junction trees", by Green and Thomas, *Biometrika*, 2013; arXiv:1104.4079
- "Enumerating the junction trees of a decomposable graph", *JCGS*, 2009, by Thomas and Green
- "Enumerating the decomposable neighbours of a decomposable graph under a simple perturbation scheme", *CSDA*, 2009, by Thomas and Green
- Webpage: www.stats.bris.ac.uk/∼peter/
- Email: P.J.Green@bristol.ac.uk
- Steffen Lauritzen's Wald lectures, Istanbul, 2012: http://www.stats.ox.ac.uk/∼steffen