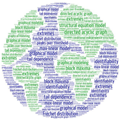
Bayesian inference with and about decomposable models

Peter Green

UTS, Sydney
University of Bristol

Conditional Independence Structures and Extremes,
München, der 12.Oktober 2016
Joint work with Alun Thomas, Utah



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:

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

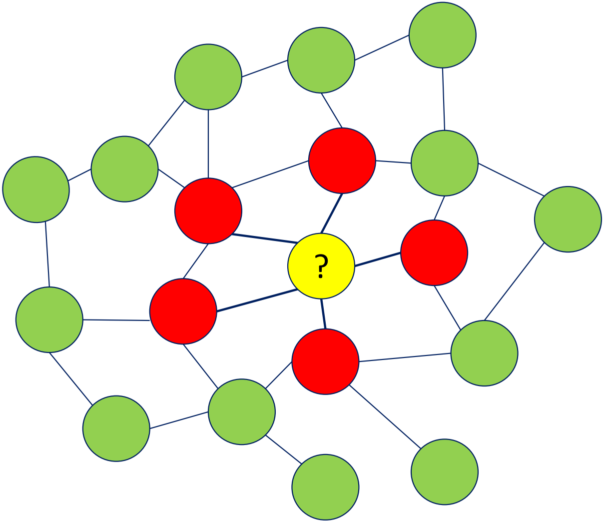
means that if you already know the value of Z, learning that of Y tells you nothing more about X. Any dependence between X and Y is indirect, mediated through Z.

In terms of probability distributions, this means

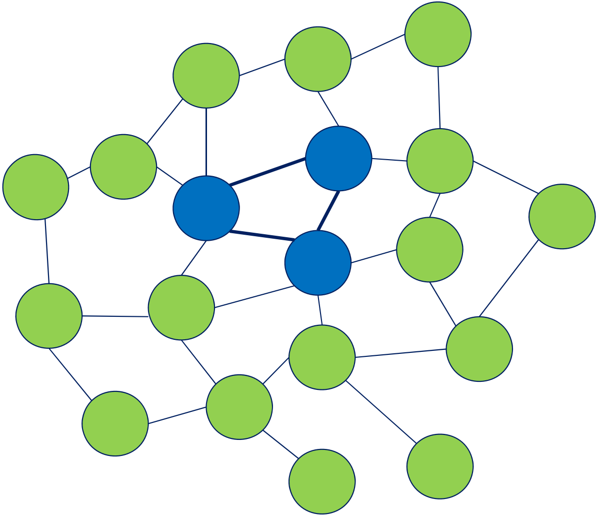
$$p(x, y|z) = p(x|z)p(y|z)$$

It proves useful to represent conditional independences graphically.

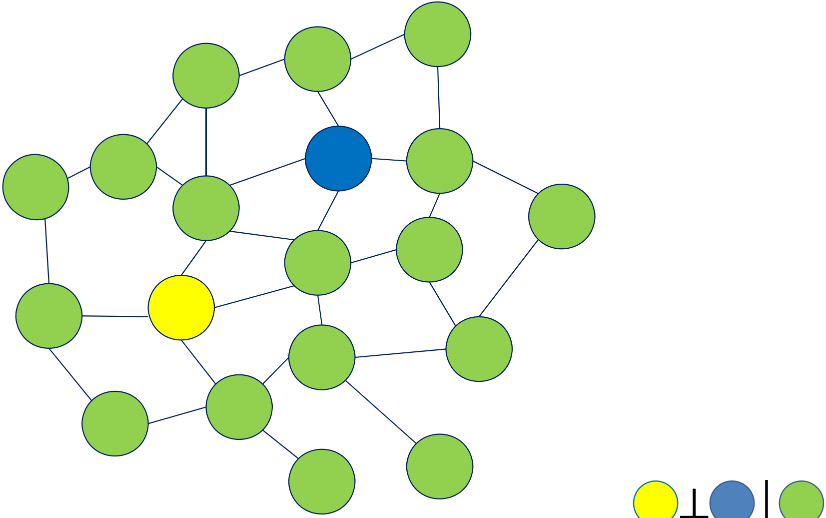
Markov random fields: the local Markov property



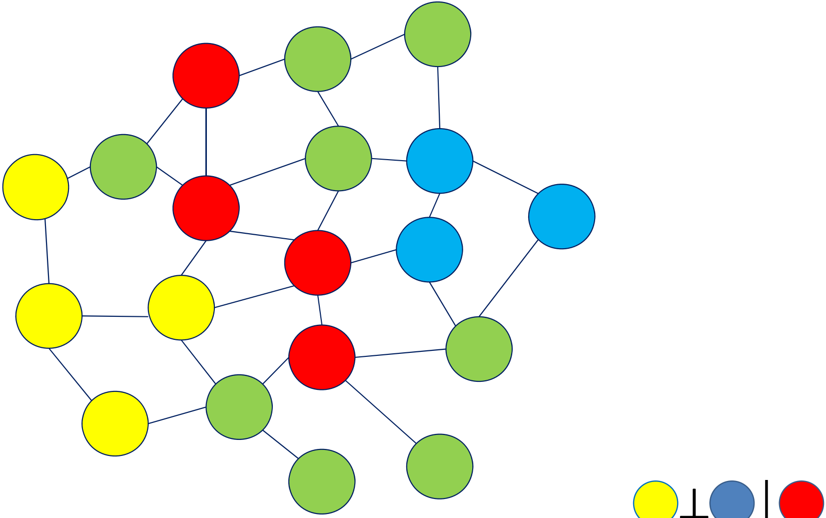
Markov random fields = Gibbs distributions



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

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

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.

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.

Structural learning

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

In this talk, we will talk about some aspects of 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 – and we restrict to decomposable graphs.

Structural learning

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

In this talk, we will talk about some aspects of 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 – and we restrict to decomposable graphs.

Decomposable graphical models

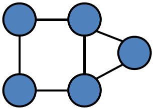
The case where $\mathcal{G}$ is decomposable has been much studied. Decomposability is a graph theory concept with statistical and computational implications.

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

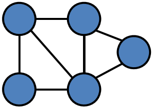
Decomposable graphical models

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



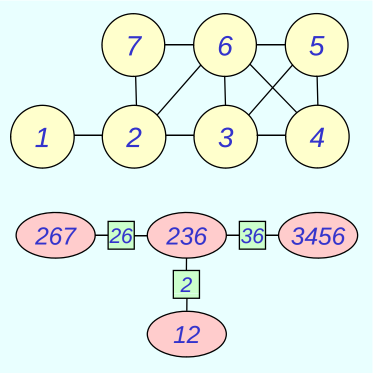
decomposable

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.

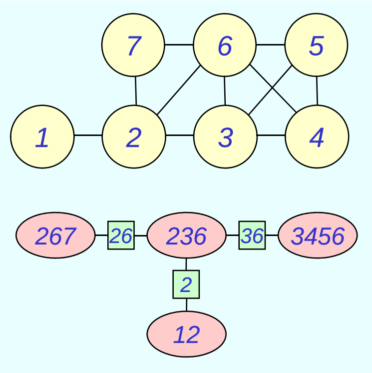


Decomposability: junction trees

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

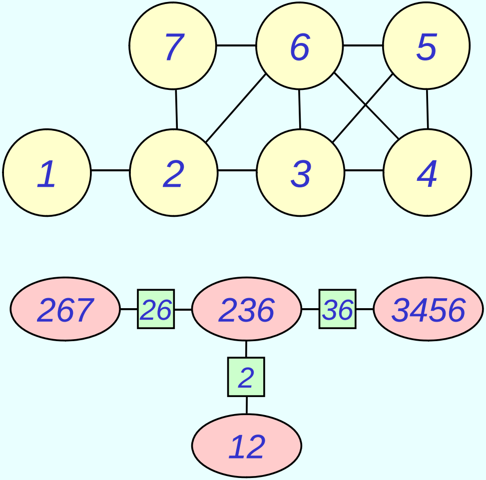
A junction tree is a graph whose vertices are cliques (maximal complete subgraphs), with the property that the cliques containing any prescribed set of vertices forms a connected sub-tree.

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.



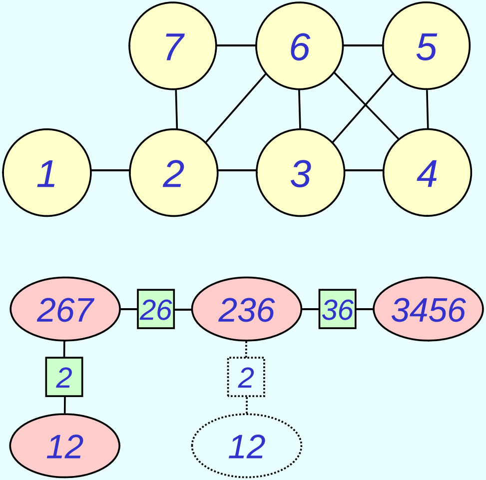
A small decomposable graph

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 pairwise 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 pairwise 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.

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.

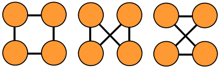
For $v \leq 3$ vertices, all are decomposable
for 4 vertices, $61/64$
for 6, $\approx 55\%$
for 8, $\approx 12\%$.

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?

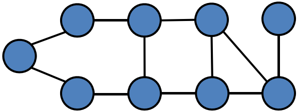
There is no reason why Nature should be kind enough to give us data from graphical models that are decomposable…

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.

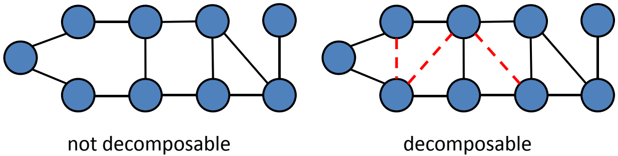


not decomposable

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.



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.

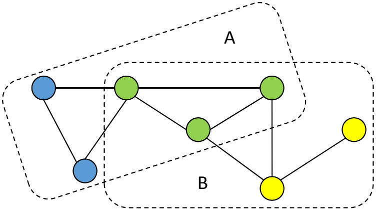
Byrne & Dawid's structural Markov property

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 $B \setminus A$.



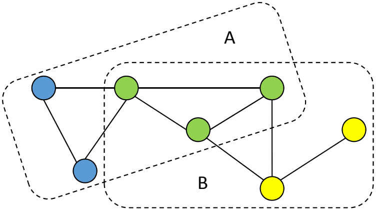
Byrne & Dawid's structural Markov property

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 :

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

where $\mathfrak{U}^\star(A, B)$ 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 :

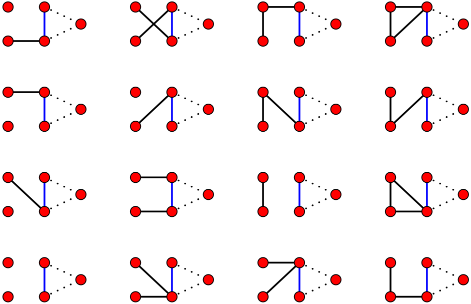
$$\mathcal{G}_A \perp\!\!\!\perp \mathcal{G}_B \mid \{\mathcal{G} \in \mathfrak{U}^\star(A, B)\} \quad [\pi],$$

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

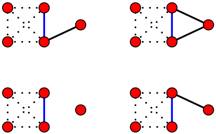
This places fewer conditional independence conditions on π, so potentially corresponds to a richer class of graph priors – but we will see that we can still say something concrete about the form of these laws.

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 $\mathcal{G}_B$



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

Clique–separator factorisation graph laws

Theorem 1

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

Theorem 1

A graph law is weakly structurally Markov if and only if has the form

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

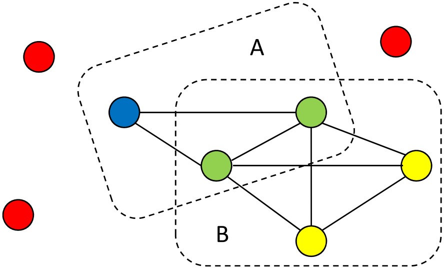
where $\{\phi_A : A \subseteq V\}, \{\psi_A : A \subseteq V\}$ *are arbitrary positive set-indexed parameters.*

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

WSM $=$ CSF proof: notation

A decomposable graph is determined by its cliques (maximal complete subgraphs). We write $\mathcal{G}^{(C_1, C_2, \ldots)}$ for the decomposable graph with cliques $C_1, C_2, \ldots$ (without ambiguity we can omit singleton cliques from the list).

In particular, $\mathcal{G}^{(A)}$ is the graph on V that is complete in A and empty otherwise, and $\mathcal{G}^{(A,B)}$ is the graph on V that is complete on both A and B and empty otherwise.

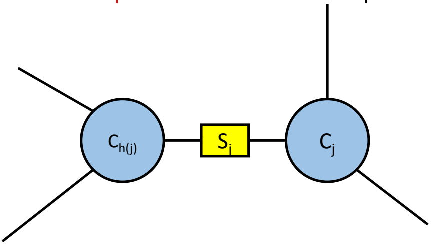


WSM = CSF proof: perfect ordering

An ordering of the cliques of an undirected graph, $(C_1, C_2, \ldots, C_c)$ is perfect if for each $j = 2, 3, \ldots, c$, there exists $h = h(j)$ such that

$$S_j = C_j \cap \bigcup_{i=1}^{j-1} C_i \subseteq C_h$$

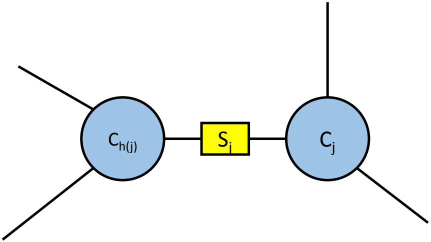
An undirected graph is decomposable if it admits a perfect ordering.



WSM = CSF proof: pluperfect ordering

A perfect ordering is called pluperfect if for each $j = 2, 3, \ldots, c$, S_j is not a proper subset of any other separator that was available, that is, could have been chosen at this point to connect $\bigcup_{i<j} C_i$ to some C_k, $k > j$.

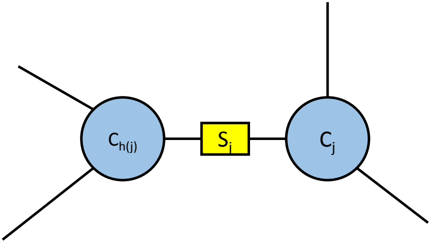
If there is a perfect ordering, there is a pluperfect one.



WSM = CSF proof: pluperfect ordering

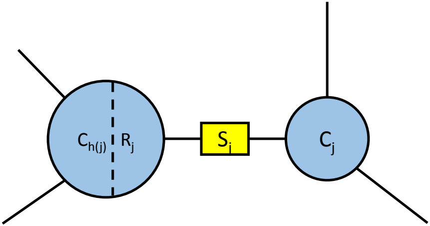
A perfect ordering is called pluperfect if for each $j = 2, 3, \ldots, c$, S_j is not a proper subset of any other separator that was available, that is, could have been chosen at this point to connect $\bigcup_{i<j} C_i$ to some C_k, $k > j$.

If there is a perfect ordering, there is a pluperfect one.



WSM = CSF proof: pluperfect ordering

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



WSM $=$ CSF – outline

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

- $\pi(\mathcal{G}^{(R_1,R_2)})/\pi(\mathcal{G}^{(R_1)})\pi(\mathcal{G}^{(R_2)})$ depends only on S, for all sets of vertices R_1, R_2 that are strict supersets of S, and for which $R_1 \cup R_2 \subseteq V$ and $R_1 \cap R_2 = S$.

WSM = CSF – outline

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

- $\pi(\mathcal{G}^{(R_1,R_2)})/\pi(\mathcal{G}^{(R_1)})\pi(\mathcal{G}^{(R_2)})$ depends only on S, for all sets of vertices R_1, R_2 that are strict supersets of S, and for which $R_1 \cup R_2 \subseteq V$ and $R_1 \cap R_2 = S$.

WSM $=$ CSF proof

Lemma 1

Let π be the density of a WSM graph law on V, and let $\mathcal{G}$ be a decomposable graph on V.

Consider a particular pluperfect ordering and junction tree of the cliques of $\mathcal{G}$, and suppose that the links of the junction tree connect C_j to $C_{h(j)}$ via separator S_j for each $j = 2, 3, \ldots, J$, where J is the number of cliques, and $h(j) \leq j - 1$.

For each such j, let R_j be any subset of $C_{h(j)}$ that is a proper superset of S_j. Then for any choice of such $\{R_j\}$, we have

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

WSM = CSF proof

Proof of Lemma 1.

Let $B = \cup_{i=1}^{j-1} C_i$ and $A = (V \setminus B) \cup R_j$. Then (A, B) is a decomposition and $A \cap B = R_j$. This intersection $A \cap B$ may or may not be a clique in $\mathcal{G}_B$ but is always a clique in $\mathcal{G}_A$. For this choice of (A, B), $\mathcal{G} \in \mathfrak{U}^\star(A, B)$, so under WSM, we have $\mathcal{G}_A \perp \mathcal{G}_B$, implying the 'cross-over identity'

$$\pi(\mathcal{G}^{(C_1,\dots,C_j)})\pi(\mathcal{G}^{(R_j)}) = \pi(\mathcal{G}^{(C_1,\dots,C_{j-1})})\pi(\mathcal{G}^{(R_j,C_j)}).$$

We can therefore write

$$\pi(\mathcal{G}) = \pi(\mathcal{G}^{(C_1)}) \prod_{j\geq 2} \frac{\pi(\mathcal{G}^{(C_1,\dots,C_j)})}{\pi(\mathcal{G}^{(C_1,\dots,C_{j-1})})} = \pi(\mathcal{G}^{(C_1)}) \prod_{j\geq 2} \frac{\pi(\mathcal{G}^{(R_j,C_j)})}{\pi(\mathcal{G}^{(R_j)})}$$

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

□

WSM = CSF proof

Proof of Lemma 1.

Let $B = \cup_{i=1}^{j-1} C_i$ and $A = (V \setminus B) \cup R_j$. Then (A, B) is a decomposition and $A \cap B = R_j$. This intersection $A \cap B$ may or may not be a clique in $\mathcal{G}_B$ but is always a clique in $\mathcal{G}_A$. For this choice of (A, B), $\mathcal{G} \in \mathfrak{U}^\star(A, B)$, so under WSM, we have $\mathcal{G}_A \perp \mathcal{G}_B$, implying the 'cross-over identity'

$$\pi(\mathcal{G}^{(C_1,\ldots,C_j)})\pi(\mathcal{G}^{(R_j)}) = \pi(\mathcal{G}^{(C_1,\ldots,C_{j-1})})\pi(\mathcal{G}^{(R_j,C_j)}).$$

We can therefore write

$$\pi(\mathcal{G}) = \pi(\mathcal{G}^{(C_1)}) \prod_{j\geq 2} \frac{\pi(\mathcal{G}^{(C_1,\ldots,C_j)})}{\pi(\mathcal{G}^{(C_1,\ldots,C_{j-1})})} = \pi(\mathcal{G}^{(C_1)}) \prod_{j\geq 2} \frac{\pi(\mathcal{G}^{(R_j,C_j)})}{\pi(\mathcal{G}^{(R_j)})}$$

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

□

WSM = CSF proof

Proof of Lemma 1.

Let $B = \cup_{i=1}^{j-1} C_i$ and $A = (V \setminus B) \cup R_j$. Then (A, B) is a decomposition and $A \cap B = R_j$. This intersection $A \cap B$ may or may not be a clique in $\mathcal{G}_B$ but is always a clique in $\mathcal{G}_A$. For this choice of (A, B), $\mathcal{G} \in \mathfrak{U}^\star(A, B)$, so under WSM, we have $\mathcal{G}_A \perp \mathcal{G}_B$, implying the 'cross-over identity'

$$\pi(\mathcal{G}^{(C_1,\dots,C_j)})\pi(\mathcal{G}^{(R_j)}) = \pi(\mathcal{G}^{(C_1,\dots,C_{j-1})})\pi(\mathcal{G}^{(R_j,C_j)}).$$

We can therefore write

$$\pi(\mathcal{G}) = \pi(\mathcal{G}^{(C_1)}) \prod_{j\geq 2} \frac{\pi(\mathcal{G}^{(C_1,\dots,C_j)})}{\pi(\mathcal{G}^{(C_1,\dots,C_{j-1})})} = \pi(\mathcal{G}^{(C_1)}) \prod_{j\geq 2} \frac{\pi(\mathcal{G}^{(R_j,C_j)})}{\pi(\mathcal{G}^{(R_j)})}$$

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

$\square$

WSM = CSF proof

Lemma 2

Let π be the density of a WSM graph law on V, and let $\mathcal{G}$ be a decomposable graph on V, with $|V| = n$, and let S be any subset of the vertices V with $|S| \leq n - 2$.

Then $\pi(\mathcal{G}^{(R_1,R_2)})/\pi(\mathcal{G}^{(R_1)})\pi(\mathcal{G}^{(R_2)})$ depends only on S, for all sets of vertices R_1, R_2 that are strict supersets of S, and for which $R_1 \cup R_2 \subseteq V$ and $R_1 \cap R_2 = S$.

WSM = CSF proof

Proof of Lemma 2.

Apply Lemma 1 to the graph $\mathcal{G}^{(R_1,R_2)}$, which has cliques R_1, R_2 and separator S. We have

$$\pi(\mathcal{G}^{(R_1,R_2)}) = \pi(\mathcal{G}^{(R_1)})\pi(\mathcal{G}^{(R_2)}) \times \frac{\pi(\mathcal{G}^{(R,R_2)})}{\pi(\mathcal{G}^{(R)})\pi(\mathcal{G}^{(R_2)})}, \text{ that is,}$$

$$\frac{\pi(\mathcal{G}^{(R_1,R_2)})}{\pi(\mathcal{G}^{(R_1)})\pi(\mathcal{G}^{(R_2)})} = \frac{\pi(\mathcal{G}^{(R,R_2)})}{\pi(\mathcal{G}^{(R)})\pi(\mathcal{G}^{(R_2)})},$$

for any R with $S \subset R \subseteq R_1$. This means that any vertices may be added to or removed from R_1, or by symmetry to or from R_2, without changing the value of $\pi(\mathcal{G}^{(R_1,R_2)})/\{\pi(\mathcal{G}^{(R_1)})\pi(\mathcal{G}^{(R_2)})\}$, providing it remains true that $R_1 \cup R_2 \subseteq V$, $R_1 \cap R_2 = S$, $R_1 \supset S$ and $R_2 \supset S$. □

WSM = CSF proof

Proof of Lemma 2.

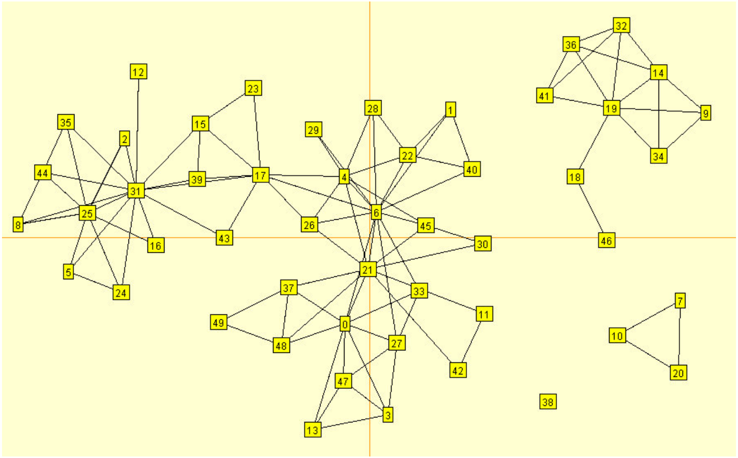
Apply Lemma 1 to the graph $\mathcal{G}^{(R_1,R_2)}$, which has cliques R_1, R_2 and separator S. We have

$$\pi(\mathcal{G}^{(R_1,R_2)}) = \pi(\mathcal{G}^{(R_1)})\pi(\mathcal{G}^{(R_2)}) \times \frac{\pi(\mathcal{G}^{(R,R_2)})}{\pi(\mathcal{G}^{(R)})\pi(\mathcal{G}^{(R_2)})}, \text{ that is,}$$

$$\frac{\pi(\mathcal{G}^{(R_1,R_2)})}{\pi(\mathcal{G}^{(R_1)})\pi(\mathcal{G}^{(R_2)})} = \frac{\pi(\mathcal{G}^{(R,R_2)})}{\pi(\mathcal{G}^{(R)})\pi(\mathcal{G}^{(R_2)})},$$

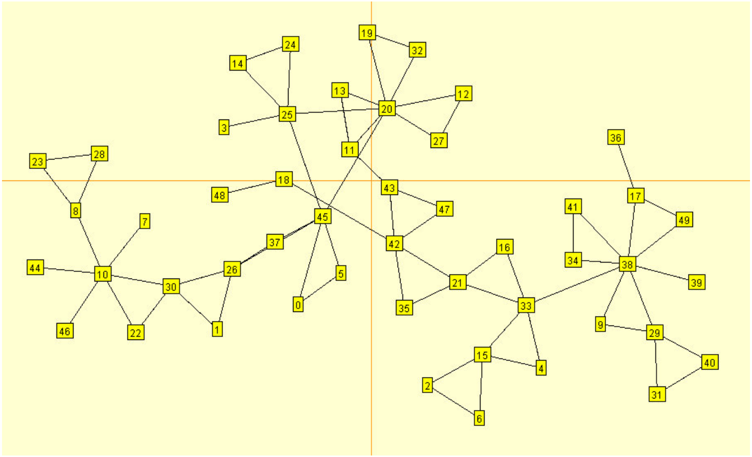
for any R with $S \subset R \subseteq R_1$. This means that any vertices may be added to or removed from R_1, or by symmetry to or from R_2, without changing the value of $\pi(\mathcal{G}^{(R_1,R_2)})/\{\pi(\mathcal{G}^{(R_1)})\pi(\mathcal{G}^{(R_2)})\}$, providing it remains true that $R_1 \cup R_2 \subseteq V$, $R_1 \cap R_2 = S$, $R_1 \supset S$ and $R_2 \supset S$. □

Example sample from a CSF graph law



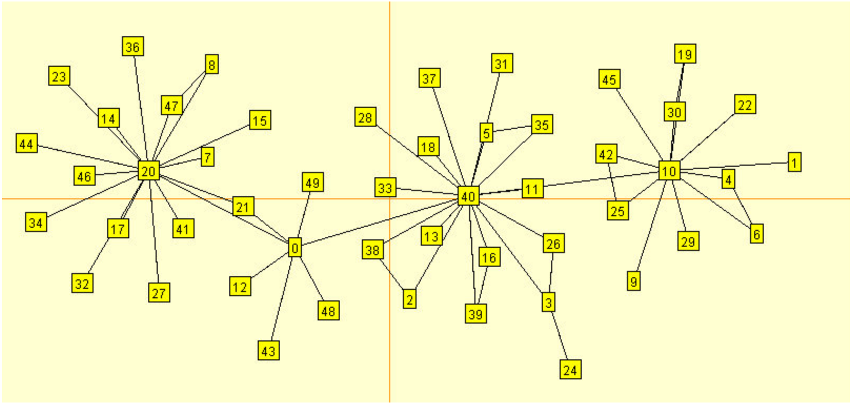
$\phi_C = \exp(4(|C| - 1))$ for $|C| \leq 4$, else 0; $\psi_S = \exp(4|S|)$

Example sample from a CSF graph law



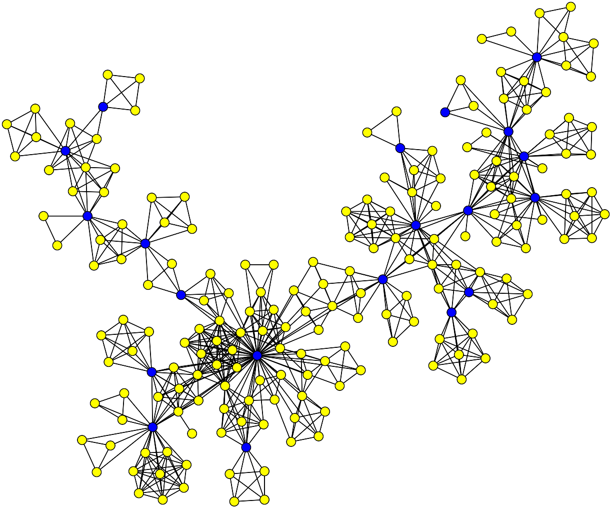
$\phi_C = \exp(4(|C|-1))$ for $|C| \leq 4$, else 0;
$\psi_S = \exp(4)$ for $|S| = 1$, else ∞

Example sample from a CSF graph law

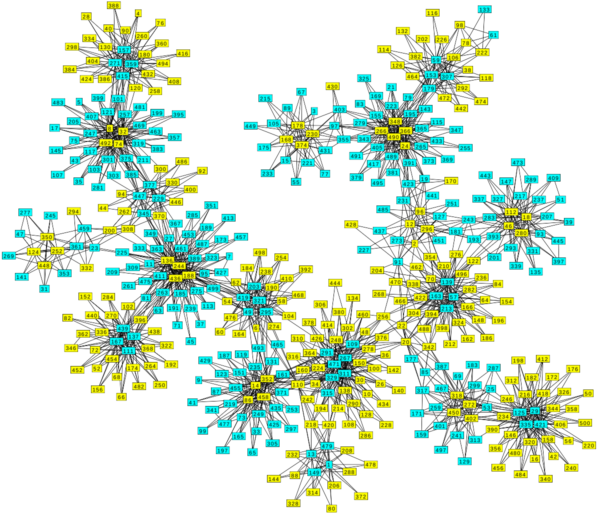


$$\phi_C = \exp(4(|C| - 1) + 3\#\{v \in C : \mod(v, 10) = 0\}) \text{ for } |C| \le 4, \text{ else } 0;$$
$$\psi_S = \exp(4) \text{ for} |S| = 1, \text{ else } \infty$$

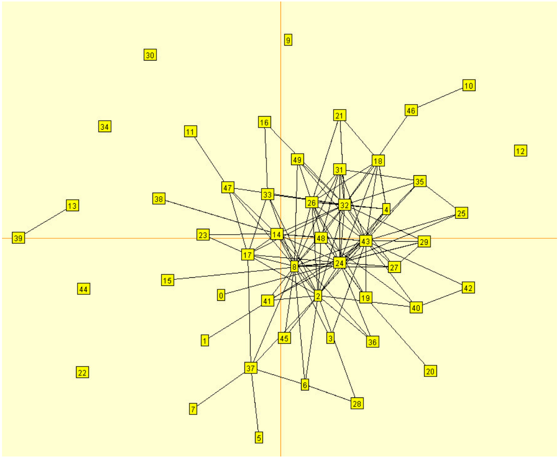
Example sample from a CSF graph law



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. $\pi(\mathcal{G}) \propto \exp(-\alpha\#\text{edges})$