# Junction Tree Algorithm[1]

Dmitry Adamskiy, David Barber

University College London

[1] These slides accompany the book *Bayesian Reasoning and Machine Learning*. The book and demos can be downloaded from www.cs.ucl.ac.uk/staff/D.Barber/brml. Feedback and corrections are also available on the site. Feel free to adapt these slides for your own purposes, but please include a link the above website.

# A general purpose inference algorithm (?)

## Applicability

- The JTA deals with ‘marginal’ inference in multiply-connected structures.
- The JTA can handle both Belief and Markov Networks.

---

## Efficiency

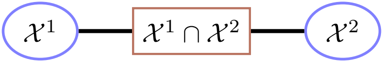
- The complexity of the JTA can be very high if the graph is multiply connected.
- Provides an upper bound on the computational complexity.
- May be that there are some problems for which much more efficient algorithms exist than the JTA.

# Clique Graph

A clique graph consists of a set of potentials, $\phi_1(\mathcal{X}^1), \ldots, \phi_n(\mathcal{X}^n)$ each defined on a set of variables $\mathcal{X}^i$. For neighbouring cliques on the graph, defined on sets of variables $\mathcal{X}^i$ and $\mathcal{X}^j$, the intersection $\mathcal{X}^s = \mathcal{X}^i \cap \mathcal{X}^j$ is called the separator and has a corresponding potential $\phi_s(\mathcal{X}^s)$. A clique graph represents the function

$$\frac{\prod_c \phi_c(\mathcal{X}^c)}{\prod_s \phi_s(\mathcal{X}^s)}$$

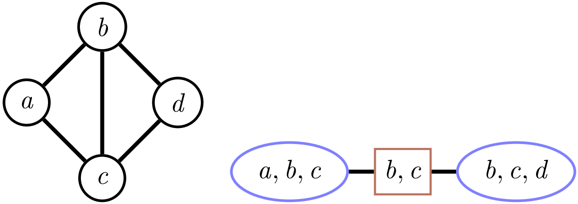
## Example




$$\frac{\phi(\mathcal{X}^1)\phi(\mathcal{X}^2)}{\phi(\mathcal{X}^1 \cap \mathcal{X}^2)}$$

# Markov Net $\rightarrow$ Clique Graph

$$p(a, b, c, d) = \frac{\phi(a, b, c)\phi(b, c, d)}{Z}$$




(a) (b)

Figure: (a) Markov network $\phi(a, b, c)\phi(b, c, d)$. (b) Clique graph representation of (a).

---

## Clique potential assignments

- The separator potential may be set to the normalisation constant $Z$.
- Cliques have potentials $\phi(a, b, c)$ and $\phi(b, c, d)$.

# Transformation

$$p(a,b,c,d) = \frac{\phi(a,b,c)\phi(b,c,d)}{Z}$$

By summing we have

$$Zp(a,b,c) = \phi(a,b,c)\sum_d \phi(b,c,d), \qquad Zp(b,c,d) = \phi(b,c,d)\sum_a \phi(a,b,c)$$

Multiplying the two expressions, we have

$$\begin{aligned} Z^2 p(a,b,c)p(b,c,d) &= \left(\phi(a,b,c)\sum_d \phi(b,c,d)\right)\left(\phi(b,c,d)\sum_a \phi(a,b,c)\right) \\ &= Z^2 p(a,b,c,d)\sum_{a,d} p(a,b,c,d) \end{aligned}$$

In other words

$$p(a,b,c,d) = \frac{p(a,b,c)p(b,c,d)}{p(c,b)}$$

---

## Clique potential assignments

- The separator potential may be set to $p(b,c)$.
- Cliques have potentials $p(a,b,c)$ and $p(b,c,d)$.

The cliques and separators contain the marginal distributions.

# Markov $\rightarrow$ Clique Graph

## The transformation

$$\phi(a,b,c) \rightarrow p(a,b,c)$$
$$\phi(b,c,d) \rightarrow p(b,c,d)$$
$$Z \rightarrow p(c,b)$$

The usefulness of this representation is that if we are interested in the marginal $p(a,b,c)$, this can be read off from the transformed clique potential.

---

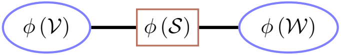
## JTA

- The JTA is a systematic way of transforming the clique graph potentials so that at the end of the transformation the new potentials contain the marginals of the distribution.
- The JTA will work by a sequence of local transformations
- Each local transformation will leave the Clique representation invariant.

# Absorption

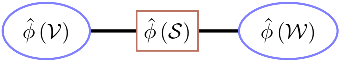
Consider neighbouring cliques $\mathcal{V}$ and $\mathcal{W}$, sharing the variables $\mathcal{S}$ in common. In this case, the distribution on the variables $\mathcal{X} = \mathcal{V} \cup \mathcal{W}$ is

$$p(\mathcal{X}) = \frac{\phi(\mathcal{V})\phi(\mathcal{W})}{\phi(\mathcal{S})}$$




and our aim is to find a new representation

$$p(\mathcal{X}) = \frac{\hat{\phi}(\mathcal{V})\hat{\phi}(\mathcal{W})}{\hat{\phi}(\mathcal{S})}$$




in which the potentials are given by

$$\hat{\phi}(\mathcal{V}) = p(\mathcal{V}), \quad \hat{\phi}(\mathcal{W}) = p(\mathcal{W}), \quad \hat{\phi}(\mathcal{S}) = p(\mathcal{S})$$

We can explicitly work out the new potentials as function of the old potentials:

$$p(\mathcal{W}) = \sum_{\mathcal{V}\setminus\mathcal{S}} p(\mathcal{X}) = \sum_{\mathcal{V}\setminus\mathcal{S}} \frac{\phi(\mathcal{V})\phi(\mathcal{W})}{\phi(\mathcal{S})} = \phi(\mathcal{W})\frac{\sum_{\mathcal{V}\setminus\mathcal{S}} \phi(\mathcal{V})}{\phi(\mathcal{S})}$$

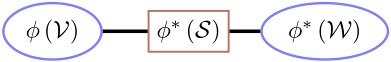
and

$$p(\mathcal{V}) = \sum_{\mathcal{W}\setminus\mathcal{S}} p(\mathcal{X}) = \sum_{\mathcal{W}\setminus\mathcal{S}} \frac{\phi(\mathcal{V})\phi(\mathcal{W})}{\phi(\mathcal{S})} = \phi(\mathcal{V})\frac{\sum_{\mathcal{W}\setminus\mathcal{S}} \phi(\mathcal{W})}{\phi(\mathcal{S})}$$

# Absorption

We say that the clique $\mathcal{W}$ 'absorbs' information from clique $\mathcal{V}$. First we define a new separator

$$\phi^{*}(\mathcal{S}) = \sum_{\mathcal{V} \setminus \mathcal{S}} \phi(\mathcal{V})$$




and refine the $\mathcal{W}$ potential using

$$\phi^{*}(\mathcal{W}) = \phi(\mathcal{W}) \frac{\phi^{*}(\mathcal{S})}{\phi(\mathcal{S})}$$

---

## Invariance

The advantage of this interpretation is that the new representation is still a valid clique graph representation of the distribution since

$$\frac{\phi(\mathcal{V})\phi^{*}(\mathcal{W})}{\phi^{*}(\mathcal{S})} = \frac{\phi(\mathcal{V})\phi(\mathcal{W})\frac{\phi^{*}(\mathcal{S})}{\phi(\mathcal{S})}}{\phi^{*}(\mathcal{S})} = \frac{\phi(\mathcal{V})\phi(\mathcal{W})}{\phi(\mathcal{S})} = p(\mathcal{X})$$

# Absorption

After $\mathcal{W}$ absorbs information from $\mathcal{V}$ then $\phi^*(\mathcal{W})$ contains the marginal $p(\mathcal{W})$. Similarly, after $\mathcal{V}$ absorbs information from $\mathcal{W}$ then $\phi^*(\mathcal{V})$ contains the marginal $p(\mathcal{V})$. After the separator $\mathcal{S}$ has participated in absorption along both directions, then the separator potential will contain $p(\mathcal{S})$.

---

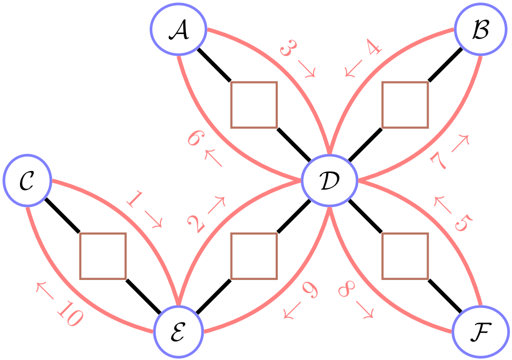
## Proof

$$\phi^{**}(\mathcal{S}) = \sum_{\mathcal{W}\backslash\mathcal{S}} \phi^*(\mathcal{W}) = \sum_{\mathcal{W}\backslash\mathcal{S}} \frac{\phi(\mathcal{W})\phi^*(\mathcal{S})}{\phi(\mathcal{S})} = \sum_{\{\mathcal{W}\cup\mathcal{V}\}\backslash\mathcal{S}} \frac{\phi(\mathcal{W})\phi(\mathcal{V})}{\phi(\mathcal{S})} = p(\mathcal{S})$$

Continuing, we have the new potential $\phi^*(\mathcal{V})$ given by

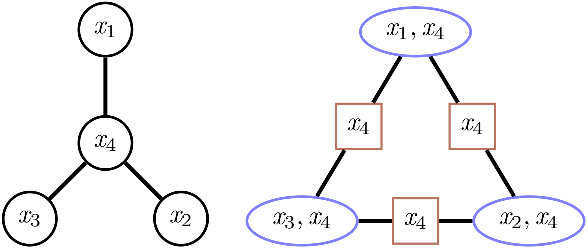
$$\begin{aligned}\phi^*(\mathcal{V}) &= \frac{\phi(\mathcal{V})\phi^{**}(\mathcal{S})}{\phi^*(\mathcal{S})} = \frac{\phi(\mathcal{V})\sum_{\mathcal{W}\backslash\mathcal{S}}\phi(\mathcal{W})\phi^*(\mathcal{S})/\phi(\mathcal{S})}{\phi^*(\mathcal{S})} \\ &= \frac{\sum_{\mathcal{W}\backslash\mathcal{S}}\phi(\mathcal{V})\phi(\mathcal{W})}{\phi(\mathcal{S})} = p(\mathcal{V})\end{aligned}$$

# Absorption Schedule on a Clique Tree




- For a valid schedule, messages can only be passed to a neighbour when all other messages have been received.
- More than one valid schedule may exist.

# Forming a Clique Tree




(a) (b)

$$p(x_1, x_2, x_3, x_4) = \phi(x_1, x_4)\phi(x_2, x_4)\phi(x_3, x_4)$$

- The clique graph of this singly-connected Markov network is multiply-connected, where the separator potentials are all set to unity.
- For absorption to work, we need a singly-connected clique graph.

# Forming a Clique Tree

$$p(x_1, x_2, x_3, x_4) = \phi(x_1, x_4)\phi(x_2, x_4)\phi(x_3, x_4)$$

Reexpress the Markov network in terms of marginals. First we have the relations

$$p(x_1, x_4) = \sum_{x_2, x_3} p(x_1, x_2, x_3, x_4) = \phi(x_1, x_4) \sum_{x_2} \phi(x_2, x_4) \sum_{x_3} \phi(x_3, x_4)$$

$$p(x_2, x_4) = \sum_{x_1, x_3} p(x_1, x_2, x_3, x_4) = \phi(x_2, x_4) \sum_{x_1} \phi(x_1, x_4) \sum_{x_3} \phi(x_3, x_4)$$

$$p(x_3, x_4) = \sum_{x_1, x_2} p(x_1, x_2, x_3, x_4) = \phi(x_3, x_4) \sum_{x_1} \phi(x_1, x_4) \sum_{x_2} \phi(x_2, x_4)$$

Taking the product of the three marginals, we have

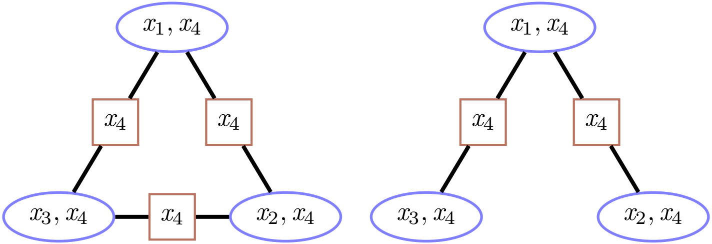
$$\begin{aligned} & p(x_1, x_4)p(x_2, x_4)p(x_3, x_4) \\ & \quad = \phi(x_1, x_4)\phi(x_2, x_4)\phi(x_3, x_4) \underbrace{\left( \sum_{x_1} \phi(x_1, x_4) \sum_{x_2} \phi(x_2, x_4) \sum_{x_3} \phi(x_3, x_4) \right)^2}_{p(x_4)^2} \end{aligned}$$

# Forming a Clique Tree

This means that the Markov network can be expressed in terms of marginals as

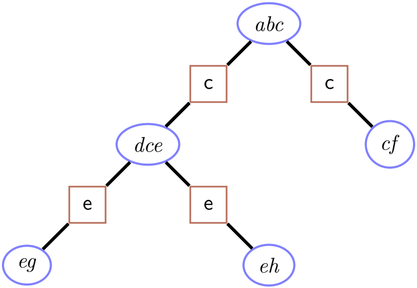
$$p(x_1, x_2, x_3, x_4) = \frac{p(x_1, x_4)p(x_2, x_4)p(x_3, x_4)}{p(x_4)p(x_4)}$$

Hence a valid clique graph is also given by




- If a variable (here $x_4$) occurs on every separator in a clique graph loop, one can remove that variable from an arbitrarily chosen separator in the loop.
- Provided that the original Markov network is singly-connected, one can always form a clique tree in this manner.

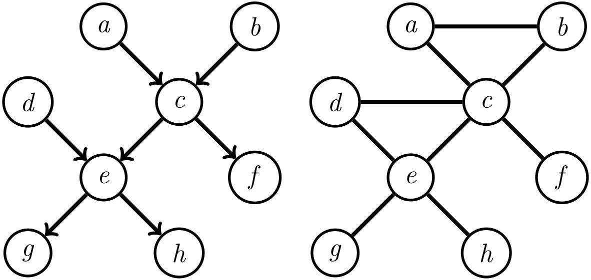
# Junction Tree




## Running Intersection Property

- A Clique Tree is a Junction Tree if, for each pair of nodes, $\mathcal{V}$ and $\mathcal{W}$, all nodes on the path between $\mathcal{V}$ and $\mathcal{W}$ contain the intersection $\mathcal{V} \cap \mathcal{W}$.
- Any singly-connected Markov Network can be transformed into a Junction Tree.
- Thanks to the running intersection property, local consistency of marginals propagates to global marginal consistency.

# Belief Net → Markov Net

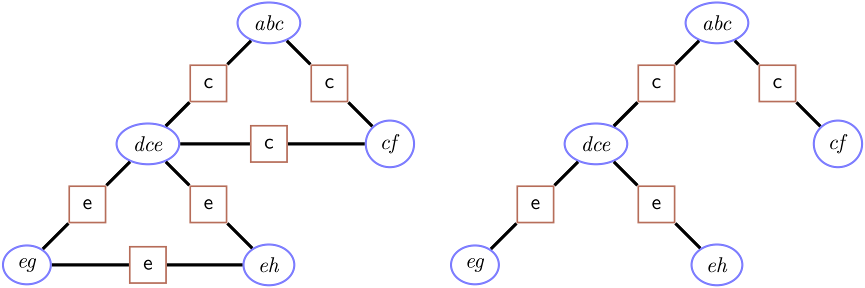



---

## Moralisation

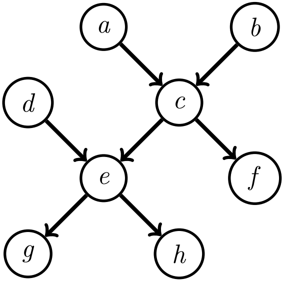
Form a link between all unmarried parents.

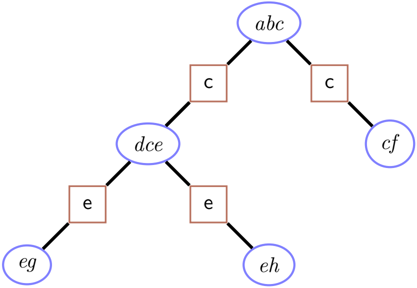
# Markov Net → Junction Tree




- Form the clique graph
- Identify a maximal weight spanning tree of the clique graph. (The weight of the edge is the number of variables in the separator)

# Absorption



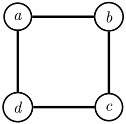




- Assign potentials to JT cliques.

$$\begin{aligned}\phi(abc) &= p(a)p(b)p(c|a,b), \quad \phi(dce) = p(d)p(e|d,c) \\ \phi(cf) &= p(f|c), \quad \phi(eg) = p(g|e), \quad \phi(eh) = p(h|e)\end{aligned}$$

  All separator potentials are initialised to unity. Note that in some instances it can be that a junction tree clique is assigned to unity.
- Carry out absorption using a valid schedule.
- Marginals can then be read off the transformed potentials.

# Multiply-Connected Markov Nets




$$p(a, b, c, d) = \phi(a, b)\phi(b, c)\phi(c, d)\phi(d, a)$$

Let's first try to make a clique graph. We have a choice about which variable first to marginalise over. Let's choose $d$:

$$p(a, b, c) = \phi(a, b)\phi(b, c)\sum_d \phi(c, d)\phi(d, a)$$

We can express the joint in terms of the marginals using

$$p(a, b, c, d) = \frac{p(a, b, c)}{\sum_d \phi(c, d)\phi(d, a)}\phi(c, d)\phi(d, a)$$

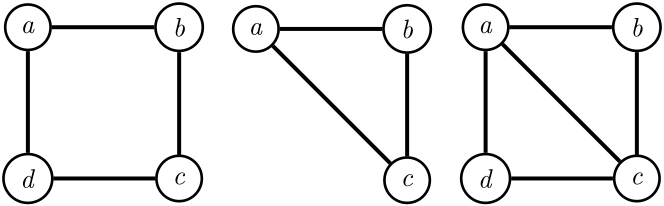
Furthermore,

$$p(a, c, d) = \phi(c, d)\phi(d, a)\sum_b \phi(a, b)\phi(b, c)$$

Plugging this into the above equation, we have

$$p(a, b, c, d) = \frac{p(a, b, c)p(a, c, d)}{\sum_d \phi(c, d)\phi(d, a)\sum_b \phi(a, b)\phi(b, c)} = \frac{p(a, b, c)p(a, c, d)}{p(a, c)}.$$

# Induced representation

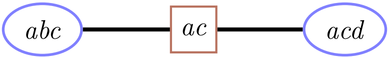



left An undirected graph with a loop.

middle Eliminating node $d$ adds a link between $a$ and $c$ in the marginal subgraph.

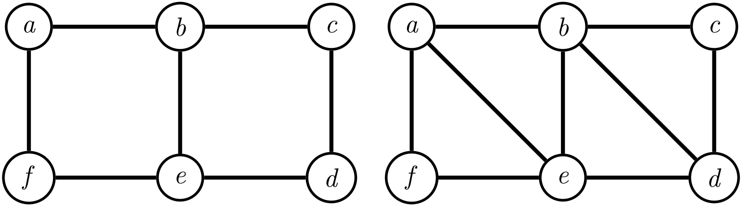
right Induced representation of the joint.

below Junction tree.



# Triangulation

In a triangulated graph, every loop of length 4 or more must have a chord (a shortcut). Such graphs are called decomposable.



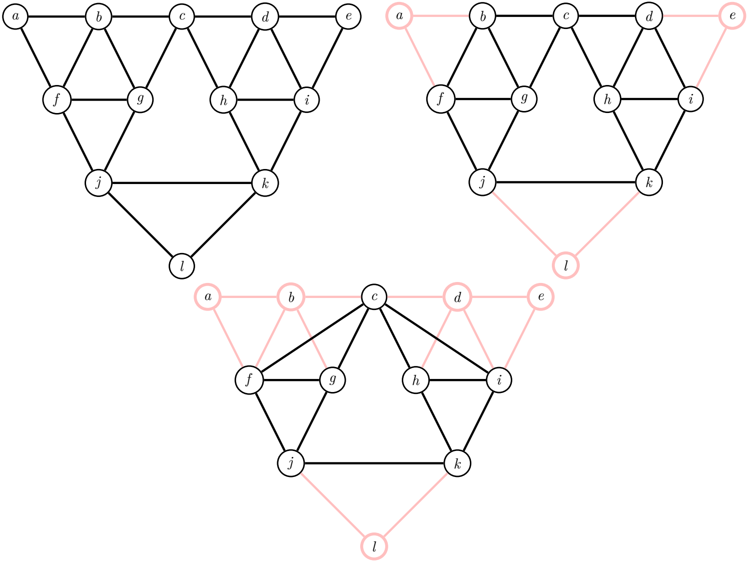

Left: a non-decomposable graph. Right: triangulated version.

# Triangulation via Variable Elimination

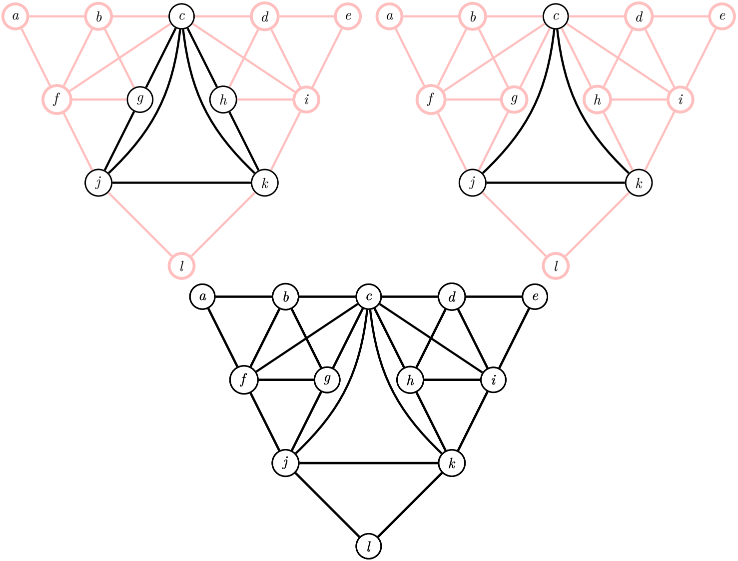
- Repeat:
- Select any non-deleted node $x$ in the graph
- Add links to all the neighbours of $x$.
- Node $x$ is then deleted.
- Until all nodes have been deleted

This procedure guarantees a triangulated graph. There are many other triangulation algorithms. No known efficient way to find the 'best' triangulation (the one with the smallest cliques).

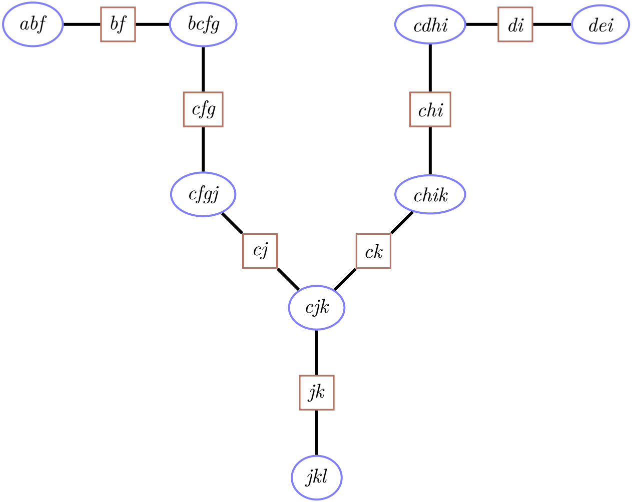
# Triangulation via Variable Elimination



# Triangulation via Variable Elimination



# The Junction Tree




This satisfies the running intersection property.

# The JTA

**Moralisation** Marry the parents. This is required only for directed distributions. Note that all the parents of a variable are married together – a common error is to marry only the 'neighbouring' parents.

**Triangulation** Ensure that every loop of length 4 or more has a chord.

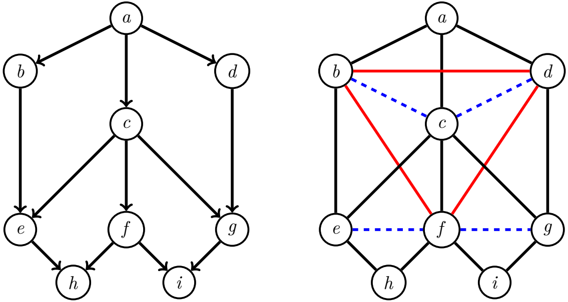
**Junction Tree** Form a junction tree from cliques of the triangulated graph, removing any unnecessary links in a loop on the clique graph. Algorithmically, this can be achieved by finding a tree with maximal spanning weight with weight $w_{ij}$ given by the number of variables in the separator between cliques $i$ and $j$.

**Potential Assignment** Assign potentials to junction tree cliques and set the separator potentials to unity.

**Message Propagation** Carry out absorption until updates have been passed along both directions of every link on the JT.

The clique marginals can then be read off from the JT.

# Example




Left: Original loopy Belief Network.
Right: The moralisation links (dashed) are between nodes $e$ and $f$ and between nodes $f$ and $g$. The other additional links come from triangulation. The clique size of the resulting clique tree (not shown) is four.

See also demoJTree.m for an example with the famous 'chest clinic' network.

# Remarks

- For discrete variables, the computational complexity of the JTA is exponential in the largest clique size.
- There may exist more efficient algorithms in particular cases. One particular special case is that of marginal inference for a binary variable MRF on a two-dimensional lattice containing only pure quadratic interactions. In this case the complexity of computing a marginal inference is $O(n^3)$ where $n$ is the number of variables in the distribution. This is in contrast to the pessimistic exponential complexity suggested by the JTA.
- One might think that the only class of distributions for which essentially a linear time algorithm is available are singly-connected distributions. However, there are decomposable graphs for which the cliques have limited size meaning that inference is tractable. For example an extended version of the 'ladder'graph has a simple induced decomposable representation. These structures are hyper trees.
- By replacing summation with maximisation, we can perform max-absorption to compute the most likely joint state – this is the union of the most likely clique states.