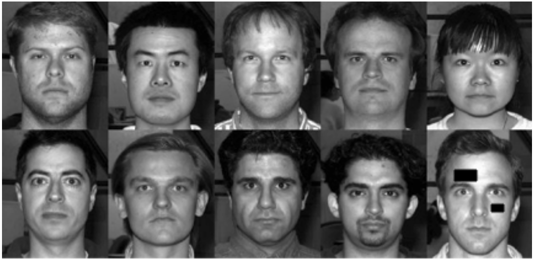
Local Computations with Probabilities on Graphical Structures

Lauritzen and Spiegelhalter, *JRSS-B*, 1988

October 10, 2014

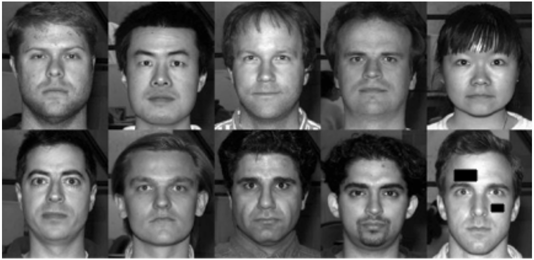
Typical Problems: Pictures

Given some images of faces with missing part. Want to find most likely values of obscured pixels.



Typical Problems: Pictures

Given some images of faces with missing part. Want to find most likely values of obscured pixels.



Typical small image has 100×100 pixels; modelling full joint distribution of 10,000 random variables is not realistic.

Typical Problems: Expert Systems

Consider a medical diagnosis tool, with list of symptoms, diseases, patient history:

- patient has a cough;
- patient is a non-smoker;
- chest x-ray has dark patches;
- $\cdots$

Given partial information, want to know probability patient has tuberculosis.

Typical Problems: Expert Systems

Consider a medical diagnosis tool, with list of symptoms, diseases, patient history:

- patient has a cough;
- patient is a non-smoker;
- chest x-ray has dark patches;
- $\cdots$

Given partial information, want to know probability patient has tuberculosis.

With 50 variables (even if only binary) calculation of marginal from $\sim 10^{15}$ joint probabilities becomes challenging.

Typical Problems: Expert Systems

Consider a medical diagnosis tool, with list of symptoms, diseases, patient history:

- patient has a cough;
- patient is a non-smoker;
- chest x-ray has dark patches;
- $\cdots$

Given partial information, want to know probability patient has tuberculosis.

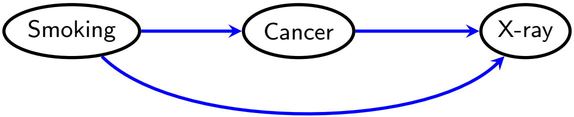
With 50 variables (even if only binary) calculation of marginal from $\sim 10^{15}$ joint probabilities becomes challenging.

Idea is to assume each variable is dependent only upon some others, and perform computations locally.

Outline

1. Directed Acyclic Graphs
2. Undirected Graphical Models
3. Junction Tree Algorithms
4. Project

A Simple DAG



A Simple DAG

We might hypothesise that smoking increases the chance of X-ray shadow *only* through possible development of lung cancer.

A Simple DAG

We might hypothesise that smoking increases the chance of X-ray shadow *only* through possible development of lung cancer.

This implies that the chances of X-ray shadow are independent of smoking status, *given* whether or not the patient has cancer.

A Simple DAG

We might hypothesise that smoking increases the chance of X-ray shadow *only* through possible development of lung cancer.

This implies that the chances of X-ray shadow are independent of smoking status, *given* whether or not the patient has cancer.

In other words,

$$p(\text{X-ray} \mid \text{Smoking}, \text{Cancer}) = p(\text{X-ray} \mid \text{Cancer}).$$

Factorisation

We can always write a joint probability distribution in factors

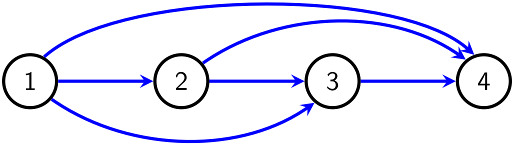
$$p(x_1, \ldots, x_k) = p(x_1)\, p(x_2 \mid x_1)\, p(x_3 \mid x_1, x_2) \cdots p(x_k \mid x_1, \ldots, x_{k-1}).$$

Factorisation

We can always write a joint probability distribution in factors

$$p(x_1, \ldots, x_k) = p(x_1)\, p(x_2 \mid x_1)\, p(x_3 \mid x_1, x_2) \cdots p(x_k \mid x_1, \ldots, x_{k-1}).$$

Draw a graph on $1, \ldots, k$ with edges $i \to j$ for each $i < j$.

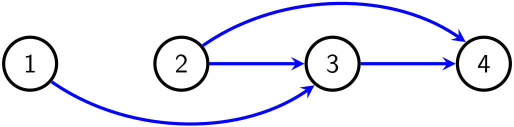


Factorisation

We can always write a joint probability distribution in factors

$$p(x_1, \dots, x_k) = p(x_1)\, p(x_2 \mid x_1)\, p(x_3 \mid x_1, x_2) \cdots p(x_k \mid x_1, \dots, x_{k-1}).$$

Draw a graph on $1, \dots, k$ with edges $i \to j$ for each $i < j$.



Remove $i \to j$ if x_i does not appear in conditioning set for x_j.

E.g.:

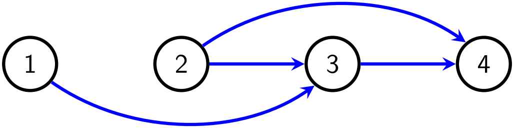
$$p(x_2 \mid x_1) = p(x_2)$$
$$p(x_4 \mid x_1,\, x_2,\, x_3) = p(x_4 \mid x_2,\, x_3)$$

Factorisation

We can always write a joint probability distribution in factors

$$p(x_1, \ldots, x_k) = p(x_1)\, p(x_2 \mid x_1)\, p(x_3 \mid x_1, x_2) \cdots p(x_k \mid x_1, \ldots, x_{k-1}).$$

Draw a graph on $1, \ldots, k$ with edges $i \to j$ for each $i < j$.



Remove $i \to j$ if x_i does not appear in conditioning set for x_j.

E.g.:

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

This is equivalent to $X_2 \perp\!\!\!\perp X_1$ and $X_4 \perp\!\!\!\perp X_1 \mid X_2, X_3$.

Some Terminology

If $v \rightarrow w$ we say v is a **parent** of w. The parents of w are $\text{pa}_{\mathcal{G}}(w)$.

Some Terminology

If $v \to w$ we say v is a **parent** of w. The parents of w are $\mathrm{pa}_{\mathcal{G}}(w)$.

We tend to abuse notation and equate the vertex (v) with the random variable (X_v).

Some Terminology

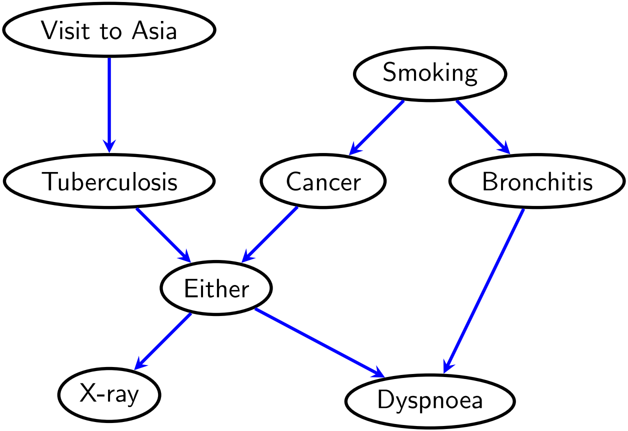
If $v \to w$ we say v is a **parent** of w. The parents of w are $\mathrm{pa}_{\mathcal{G}}(w)$.

We tend to abuse notation and equate the vertex (v) with the random variable (X_v).

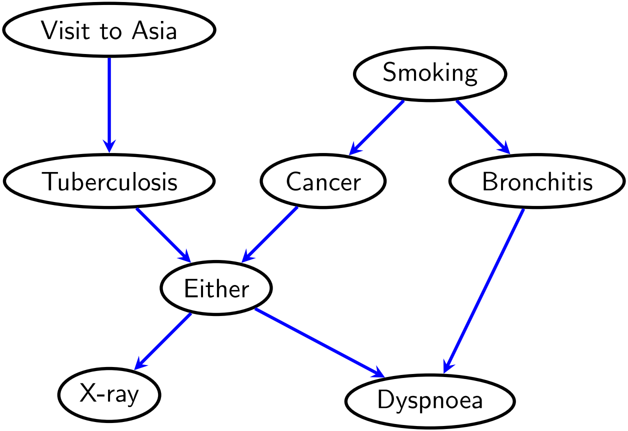
A joint probability density p **factorizes according to** $\mathcal{G}$ if

$$p(x_1, \dots, x_k) = \prod_{i=1}^{k} p(x_i \mid x_{\mathrm{pa}(i)}).$$

Example


Visit to Asia
Smoking
Tuberculosis
Cancer
Bronchitis
Either
X-ray
Dyspnoea

Example



Only need to record conditional probability tables for each node.

Calculations

Suppose we want to know the conditional distribution of Bronchitis given a negative x-ray and a visit to Asia. Then

Calculations

Suppose we want to know the conditional distribution of Bronchitis given a negative x-ray and a visit to Asia. Then

Well:

$$
\begin{aligned}
p(x,a,b) &= \sum_{s,t,c,d,e} p(a)p(s)p(t|a)p(c|s)p(b|s)p(e|t,c)p(d|e,b)p(x|e) \\
&= p(a)\sum_t p(t|a)\sum_e p(x|e)\sum_c p(e|t,c)\cdots \\
&\qquad \cdots\sum_s p(s)p(c|s)p(b|s)\sum_d p(d|e,b)
\end{aligned}
$$

Calculations

Suppose we want to know the conditional distribution of Bronchitis given a negative x-ray and a visit to Asia. Then

Well:

$$
\begin{aligned}
p(x,a,b) &= \sum_{s,t,c,d,e} p(a)p(s)p(t|a)p(c|s)p(b|s)p(e|t,c)p(d|e,b)p(x|e) \\
&= p(a)\sum_t p(t|a)\sum_e p(x|e)\sum_c p(e|t,c)\cdots \\
&\qquad \cdots\sum_s p(s)p(c|s)p(b|s)\sum_d p(d|e,b)
\end{aligned}
$$

this isn't unique:

$$
\begin{aligned}
&= p(a)\sum_t p(t|a)\sum_e p(x|e)\sum_s p(s)p(b|s)\cdots \\
&\qquad \sum_c p(e|t,c)p(c|s)\sum_d p(d|e,b).
\end{aligned}
$$

Calculations

Suppose we want to know the conditional distribution of Bronchitis given a negative x-ray and a visit to Asia. Then

Well:

$$
\begin{aligned}
p(x, a, b) &= \sum_{s,t,c,d,e} p(a)p(s)p(t|a)p(c|s)p(b|s)p(e|t, c)p(d|e, b)p(x|e) \\
&= p(a) \sum_t p(t|a) \sum_e p(x|e) \sum_c p(e|t, c) \cdots \\
&\qquad \cdots \sum_s p(s)p(c|s)p(b|s) \sum_d p(d|e, b)
\end{aligned}
$$

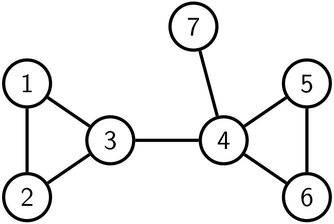
this isn't unique:

$$
\begin{aligned}
&= p(a) \sum_t p(t|a) \sum_e p(x|e) \sum_s p(s)p(b|s) \cdots \\
&\qquad \sum_c p(e|t, c)p(c|s) \sum_d p(d|e, b).
\end{aligned}
$$

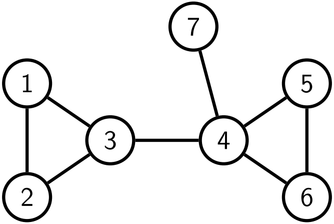
So in general, how do we know which sums to 'push' into the middle?

Outline

Undirected Graphs

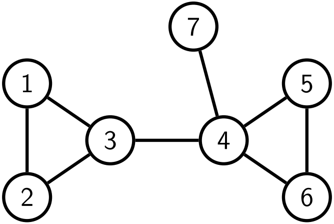


Undirected Graphs



Many ways in which a distribution can 'obey' an undirected graph. Let $\mathcal{C}(\mathcal{G})$ be the **cliques** (maximal fully connected subgraphs).

Undirected Graphs



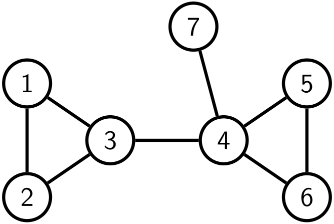
Many ways in which a distribution can 'obey' an undirected graph. Let $\mathcal{C}(\mathcal{G})$ be the **cliques** (maximal fully connected subgraphs).

$P \in \mathcal{P}_f(\mathcal{G})$ **factorizes** according to $\mathcal{G}$ if

$$p(x_1, \ldots, x_k) = \prod_{C \in \mathcal{C}(\mathcal{G})} \psi_C(x_C)$$

for some potential functions $\psi_C \geq 0$.

Undirected Graphs



Many ways in which a distribution can 'obey' an undirected graph. Let $\mathcal{C}(\mathcal{G})$ be the **cliques** (maximal fully connected subgraphs).

$P \in \mathcal{P}_f(\mathcal{G})$ **factorizes** according to $\mathcal{G}$ if

$$p(x_1, \ldots, x_k) = \prod_{C \in \mathcal{C}(\mathcal{G})} \psi_C(x_C)$$

for some potential functions $\psi_C \geq 0$.

So $p = \psi_{123} \cdot \psi_{34} \cdot \psi_{456} \cdot \psi_{47}$ above.

Decomposable Graphs

Say a graph is **decomposable** if either it is complete, or there exist three disjoint non-empty subsets $A \cup B \cup S = V$, where:

(i) $\mathcal{G}_S$ is complete;

(ii) S separates A from B in $\mathcal{G}$;

(iii) $\mathcal{G}_{A \cup S}$ and $\mathcal{G}_{B \cup S}$ are decomposable.

Decomposable Graphs

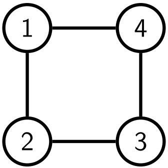
Say a graph is **decomposable** if either it is complete, or there exist three disjoint non-empty subsets $A \cup B \cup S = V$, where:

(i) $\mathcal{G}_S$ is complete;

(ii) S separates A from B in $\mathcal{G}$;

(iii) $\mathcal{G}_{A \cup S}$ and $\mathcal{G}_{B \cup S}$ are decomposable.

Doesn't always work:



Decomposable Graphs

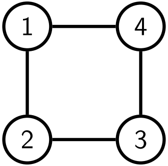
Say a graph is **decomposable** if either it is complete, or there exist three disjoint non-empty subsets $A \cup B \cup S = V$, where:

(i) $\mathcal{G}_S$ is complete;

(ii) S separates A from B in $\mathcal{G}$;

(iii) $\mathcal{G}_{A \cup S}$ and $\mathcal{G}_{B \cup S}$ are decomposable.

Doesn't always work:



Theorem

An undirected graph $\mathcal{G}$ is decomposable iff it contains no chordless cycles of length ≥ 4.

Decomposable Graphs

If a graph is decomposable as (A, S, B) then

$$p(x_V) = f(x_A, x_S) \cdot g(x_B, x_S)$$

Decomposable Graphs

If a graph is decomposable as (A, S, B) then

$$\begin{aligned} p(x_V) &= f(x_A, x_S) \cdot g(x_B, x_S) \\ &= p(x_S) \cdot p(x_A \mid x_S) \cdot p(x_B \mid x_S) \\ &= \frac{p(x_A, x_S) \cdot p(x_B, x_S)}{p(x_S)}. \end{aligned}$$

Decomposable Graphs

If a graph is decomposable as (A, S, B) then

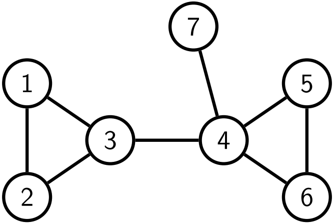
$$\begin{aligned} p(x_V) &= f(x_A, x_S) \cdot g(x_B, x_S) \\ &= p(x_S) \cdot p(x_A \mid x_S) \cdot p(x_B \mid x_S) \\ &= \frac{p(x_A, x_S) \cdot p(x_B, x_S)}{p(x_S)}. \end{aligned}$$

So by induction

$$= \frac{\prod_{C \in \mathcal{C}} p_C(x_C)}{\prod_{S \in \mathcal{S}} p_S(x_S)}$$

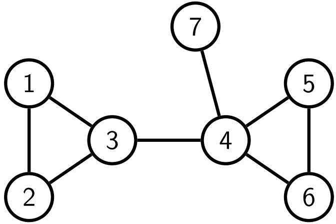
where $\mathcal{S}$ is a collection of separators (not necessarily disjoint).

Forming a Junction Tree

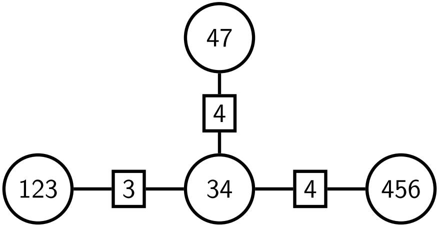


If a graph is decomposable, can form a **junction tree** of cliques and separators:

Forming a Junction Tree



If a graph is decomposable, can form a **junction tree** of cliques and separators:



Idea

Let $\mathcal{T}$ be a junction tree. Initially have

$$p(x_V) = \frac{\prod_C \psi_C(x_C)}{\prod_S \psi_S(x_S)}$$

for cliques C and separators S (separators may not be unique).

Idea

Let $\mathcal{T}$ be a junction tree. Initially have

$$p(x_V) = \frac{\prod_C \psi_C(x_C)}{\prod_S \psi_S(x_S)}$$

for cliques C and separators S (separators may not be unique).

Lemma

If the ψ_C and ψ_S functions have consistent margins, then $\psi_C(x_C) = p_C(x_C)$.

Idea

Let $\mathcal{T}$ be a junction tree. Initially have

$$p(x_V) = \frac{\prod_C \psi_C(x_C)}{\prod_S \psi_S(x_S)}$$

for cliques C and separators S (separators may not be unique).

Lemma

If the ψ_C and ψ_S functions have consistent margins, then $\psi_C(x_C) = p_C(x_C)$.

If margins are not consistent, then we can make them so!

Message Passing

To 'pass a message' from clique C to clique D (with separator S), set:

- $\psi'_S(x_S) = \sum_{x_{C \setminus S}} \psi_C(x_C)$
- $\psi'_D(x_D) = \psi_D(x_D) \frac{\psi'_S(x_S)}{\psi_S(x_S)}$

Message Passing

To 'pass a message' from clique C to clique D (with separator S), set:

- $\psi'_S(x_S) = \sum_{x_{C \setminus S}} \psi_C(x_C)$
- $\psi'_D(x_D) = \psi_D(x_D) \frac{\psi'_S(x_S)}{\psi_S(x_S)}$

Lemma

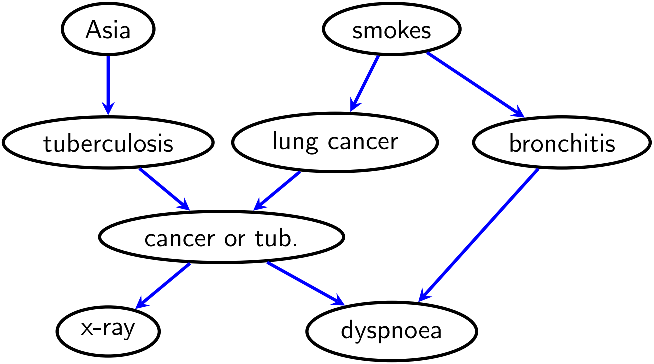
After passing a message from C to D, the probability distribution remains unchanged, and

$$\sum_{x_{C \setminus S}} \psi_C(x_C) = \psi'_S(x_S).$$

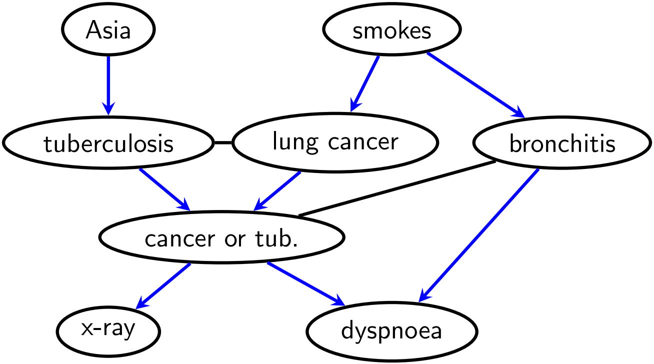
If, in addition, $\sum_{x_{D \setminus S}} \psi_D(x_D) = \psi_S(x_S)$, then

$$\sum_{x_{D \setminus S}} \psi'_D(x_D) = \psi'_S(x_S).$$

From DAGs to Decomposable Graphs

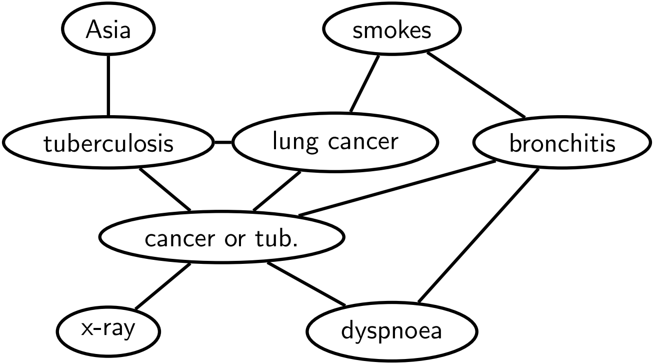


From DAGs to Decomposable Graphs



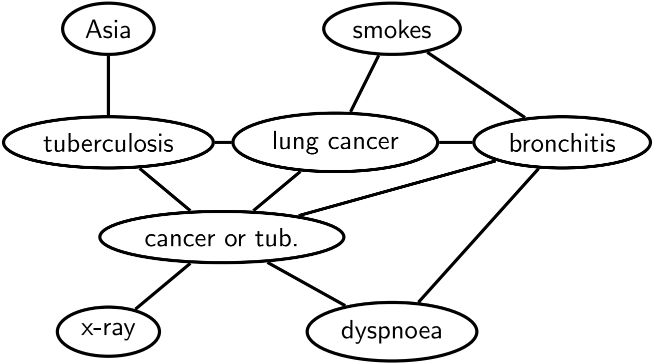
1. 'Marry' parents.

From DAGs to Decomposable Graphs



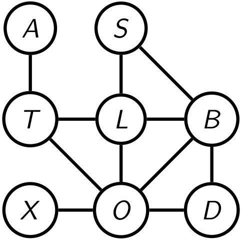
1. ‘Marry’ parents.
2. Drop arrows.

From DAGs to Decomposable Graphs



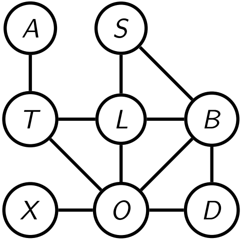
1. 'Marry' parents.
2. Drop arrows.
3. Triangulate (not unique or easy!)

Forming a Junction Tree

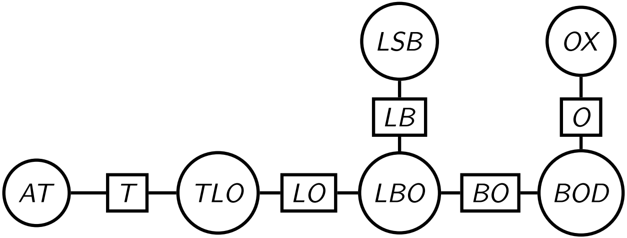


From the decomposable graph, form a junction tree:

Forming a Junction Tree



From the decomposable graph, form a junction tree:



Initialization

Suppose we have conditional probability tables $p(x_v \mid x_{\mathrm{pa}(v)})$. Then

$$p(x_V) = \prod_v p(x_v \mid x_{\mathrm{pa}(v)})$$

and each set $\{v\} \cup \mathrm{pa}(v)$ is contained in a clique.

Initialization

Suppose we have conditional probability tables $p(x_v \mid x_{\mathrm{pa}(v)})$. Then

$$p(x_V) = \prod_v p(x_v \mid x_{\mathrm{pa}(v)})$$

and each set $\{v\} \cup \mathrm{pa}(v)$ is contained in a clique.

So can set $\prod_C \psi_C(x_C) = \prod_v p(x_v \mid x_{\mathrm{pa}(v)})$ and $\psi_S(x_S) = 1$ to get valid ψ representation.

Initialization

Suppose we have conditional probability tables $p(x_v \mid x_{\mathrm{pa}(v)})$. Then

$$p(x_V) = \prod_v p(x_v \mid x_{\mathrm{pa}(v)})$$

and each set $\{v\} \cup \mathrm{pa}(v)$ is contained in a clique.

So can set $\prod_C \psi_C(x_C) = \prod_v p(x_v \mid x_{\mathrm{pa}(v)})$ and $\psi_S(x_S) = 1$ to get valid ψ representation.

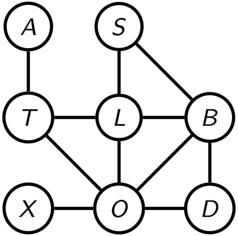
Not locally consistent, but:

Theorem

After collecting and distributing messages in a junction tree, the potentials are locally consistent.

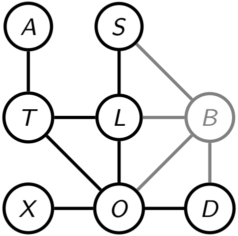
Outline

Conditioning



$$p(a,t,s,l,b,o,d,x) = \psi(a,t) \cdot \psi(t,l,o) \cdot \psi(s,l,b) \cdot \psi(b,o,d) \cdot \psi(x,o)$$

Conditioning

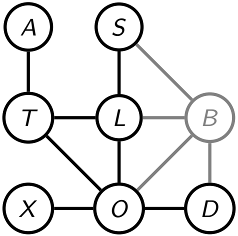


$$p(a,t,s,l,b,o,d,x) = \psi(a,t)\cdot\psi(t,l,o)\cdot\psi(s,l,b)\cdot\psi(b,o,d)\cdot\psi(x,o)$$

Graph structure is preserved under conditioning:

$$p(a,t,s,l,o,d,x \mid b) \propto \psi(a,t)\cdot\psi(t,l,o)\cdot\psi^*(s,l)\cdot\psi^*(o,d)\cdot\psi(x,o).$$

Conditioning



$$p(a, t, s, l, b, o, d, x) = \psi(a, t) \cdot \psi(t, l, o) \cdot \psi(s, l, b) \cdot \psi(b, o, d) \cdot \psi(x, o)$$

Graph structure is preserved under conditioning:

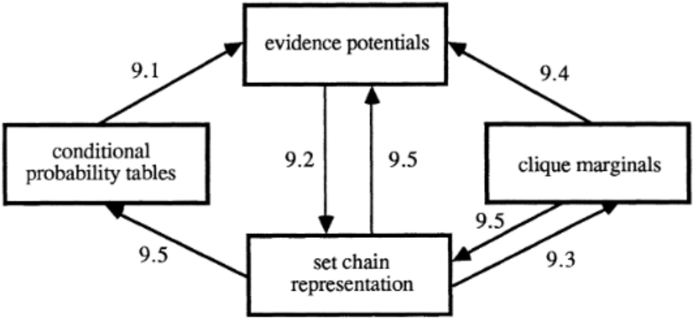
$$p(a, t, s, l, o, d, x \mid b) \propto \psi(a, t) \cdot \psi(t, l, o) \cdot \psi^*(s, l) \cdot \psi^*(o, d) \cdot \psi(x, o).$$

Gives something proportional to conditional distribution: use structure to calculate normalizing constant.

Outline

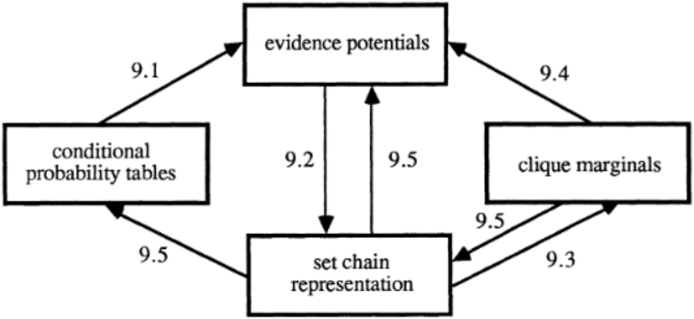
Project Idea

The idea would be to produce a collection of functions for switching between different representations of a DAG model/Bayesian Network.



Project Idea

The idea would be to produce a collection of functions for switching between different representations of a DAG model/Bayesian Network.



Can also write functions for the introduction of evidence, and efficient calculation of arbitrary conditional and marginal distributions.

[Won't attempt to find good decomposable representations.]