Advanced Statistical Topics 2001-02

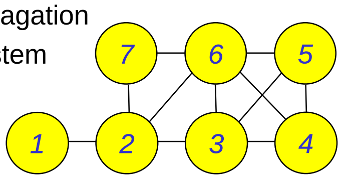
Module 4:

Probabilistic expert systems

A. Introduction

Module outline

- Information, uncertainty and probability
- Motivating examples
- Graphical models
- Probability propagation
- The HUGIN system



Motivating examples

- Simple applications of Bayes' theorem
- Markov chains and random walks
- Bayesian hierarchical models
- Forensic genetics
- Expert systems in medical and engineering diagnosis

The 'Asia' (chest-clinic) example

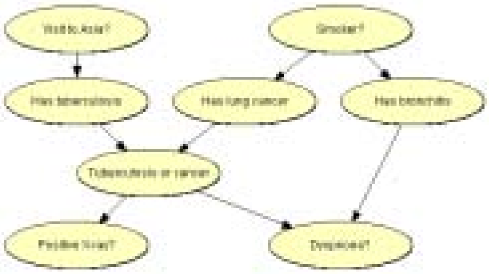
Shortness-of-breath (dyspnoea) may be due to tuberculosis, lung cancer, bronchitis, more than one of these diseases or none of them.

A recent visit to Asia increases the risk of tuberculosis, while smoking is known to be a risk factor for both lung cancer and bronchitis.

The results of a single chest X-ray do not discriminate between lung cancer and tuberculosis, as neither does the presence or absence of dyspnoea.

+2

Visual representation of the Asia example - a graphical model



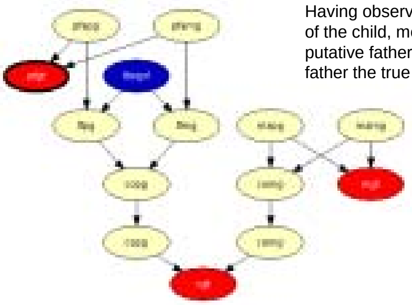
The ‘Asia’ (chest-clinic) example

Now … a patient presents with shortness-of-breath (dyspnoea) …. How can the physician use available tests (X-ray) and enquiries about the patient’s history (smoking, visits to Asia) to help to diagnose which, if any, of tuberculosis, lung cancer, or bronchitis is the patient probably suffering from?

An example from forensic genetics

DNA profiling based on STR’s (single tandem repeats) are finding many uses in forensics, for identifying suspects, deciding paternity, etc. Can we use Mendelian genetics and Bayes’ theorem to make probabilistic inference in such cases?

Graphical model for a paternity enquiry - allowing mutation



Having observed the genotype of the child, mother and putative father, is the putative father the true father?

Surgical rankings

- 12 hospitals carry out different numbers of a certain type of operation:

  47, 148, 119, 810, 211, 196, 148, 215, 207, 97, 256, 360 respectively.

- They are differently successful, and there are:

  0, 18, 8, 46, 8, 13, 9, 31, 14, 8, 29, 24 fatalities, respectively.

Surgical rankings, continued

- What inference can we draw about the relative qualities of the hospitals based on these data?
- Does knowing the mortality at one hospital tell us anything at all about the other hospitals - that is, can we ‘pool’ information?

B. Key ideas

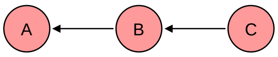
Key ideas in exact probability calculation in complex systems

- Graphical model (usually a directed acyclic graph)
- Conditional independence graph
- Decomposability
- Probability propagation: 'message-passing'

Let's motivate this with some simple examples….

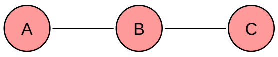
+1

Directed acyclic graph (DAG)



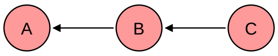
… indicating that model is specified by *p(C)*, *p(B|C)* and *p(A|B): p(A,B,C) = p(A|B)p(B|C)p(C)*

The corresponding Conditional independence graph (CIG) is

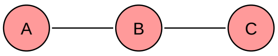


… encoding various conditional independence assumptions, e.g. *p(A,C|B) = p(A|B)p(C|B)*

DAG



CIG



$$p(A,B,C) = p(A,B)p(C \mid A,B) = p(A,B)p(C \mid B)$$

$$= \frac{p(A,B)p(B,C)}{p(B)}$$

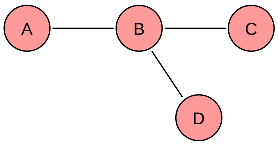
true for any *A, B, C*

since $C \perp A \mid B$

definition of *p(C|B)*

+4

CIG



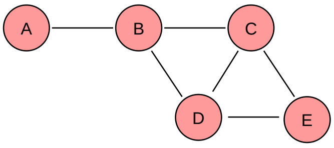
$$p(A,B,C,D) = p(A,B)p(C \mid A,B)p(D \mid A,B,C)$$

$$= p(A,B)p(C \mid B)p(D \mid B)$$

$$= \frac{p(A,B)p(B,C)p(B,D)}{p(B)p(B)}$$

+2

CIG



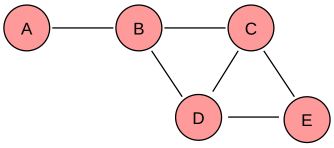
$$p(A,B,C,D,E) = p(A,B)p(C,D \mid A,B)p(E \mid A,B,C,D)$$

$$= p(A,B)p(C,D \mid B)p(E \mid C,D)$$

$$= \frac{p(A,B)p(B,C,D)p(C,D,E)}{p(B)p(C,D)}$$

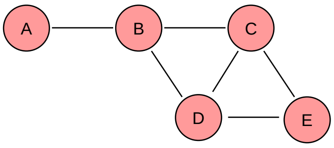
+2

CIG



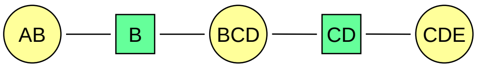
$$p(A,B,C,D,E) = \frac{p(A,B)p(B,C,D)p(C,D,E)}{p(B)p(C,D)}$$

CIG



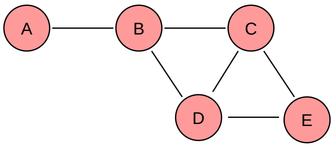
$$p(A,B,C,D,E) = \frac{p(A,B)p(B,C,D)p(C,D,E)}{p(B)p(C,D)} = \frac{\prod_{\text{cliques}C} p(X_C)}{\prod_{\text{separators}S} p(X_S)}$$

JT

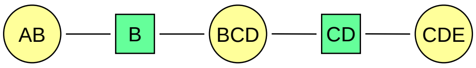


+1

CIG



JT



$$p(A,B,C=c,D,E) = \frac{p(A,B)p(B,C=c,D)p(C=c,D,E)}{p(B)p(C=c,D)}$$

+1

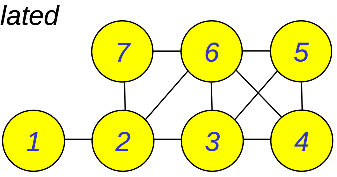
Decomposability

An important concept in processing information through undirected graphs is *decomposability*

(= graph triangulated

= no chordless

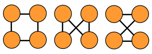
≥ 4-cycles)



Is decomposability a serious constraint?

out of $2^{\binom{n}{2}}$

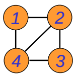
- How many graphs are decomposable?

Number of vertices	Proportion of graphs that are decomposable
≤ 3	all
4	61/64 – all but: 
6	~80%
16	~45%

- Models using decomposable graphs are ‘dense’

Is decomposability any use?

- Maximum likelihood estimates can be computed exactly in decomposable models

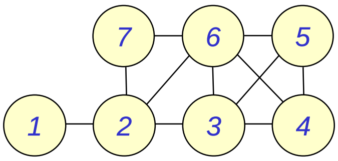


$$\Longrightarrow \hat{E}(N_{ijkl}) = \frac{n_{ij+l}\,n_{+jkl}}{n_{+j+l}}$$

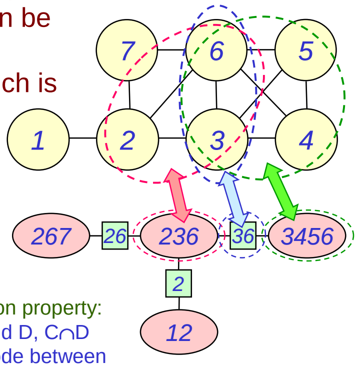
- Decomposability is a key to the ‘message passing’ algorithms for probabilistic expert systems (and peeling genetic pedigrees)

Cliques

A *clique* is a *maximal complete subgraph*: here the cliques are {1,2},{2,6,7}, {2,3,6}, and {3,4,5,6}



A graph is decomposable if and only if it can be represented by a junction tree (which is not unique)



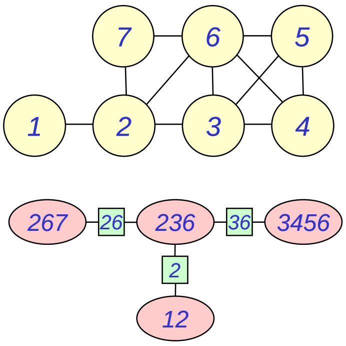
a clique

another clique

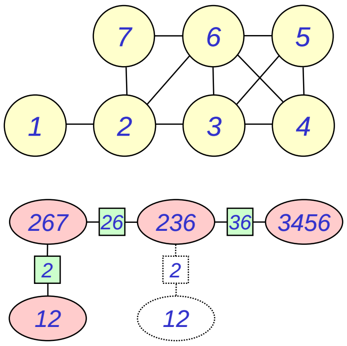
a separator

The running intersection property:
For any 2 cliques C and D, C∩D is a subset of every node between them in the junction tree

Non-uniqueness of junction tree



Non-uniqueness of junction tree



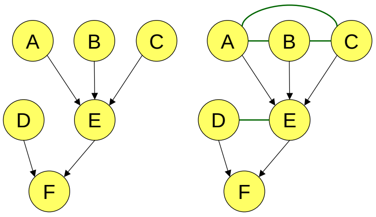
C. The works

Exact probability calculation in complex systems

0. Start with a directed acyclic graph
1. Find corresponding Conditional Independence Graph
2. Ensure decomposability
3. Probability propagation: ‘message-passing’

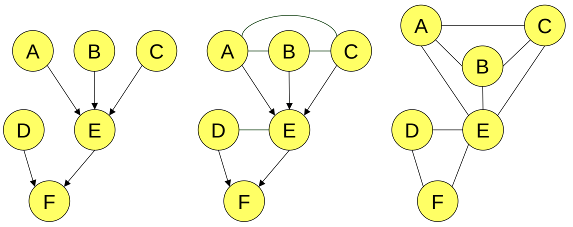
1. Finding the (undirected) conditional independence graph for a given DAG

- Step 1: moralise (parents must marry)

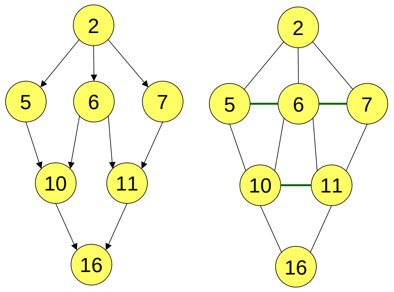


1. Finding the (undirected) conditional independence graph for a given DAG

- Step 2: drop directions

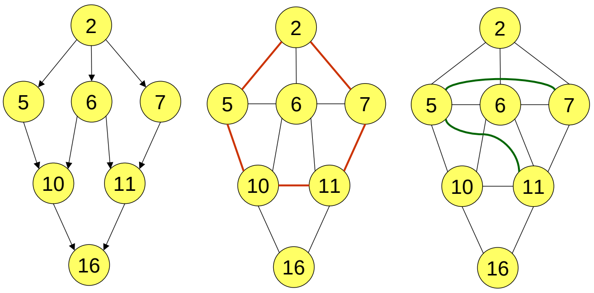


2. Ensuring decomposability

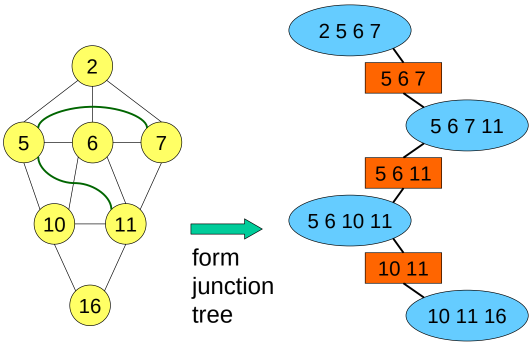


2. Ensuring decomposability

…. triangulate



3. Probability propagation



form
junction
tree

If the distribution *p(X)* has a decomposable CIG, then it can be written in the following potential representation form:

$$p(X) = \frac{\prod_{cliquesC} \psi(X_C)}{\prod_{separatorsS} \psi(X_S)}$$

the individual terms are called potentials;
the representation is not unique

The potential representation

$$p(X) = \frac{\prod_{cliquesC} \psi(X_C)}{\prod_{separatorsS} \psi(X_S)}$$

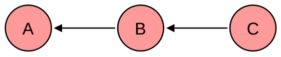
can easily be initialised by
• assigning each DAG factor $p(X_v \mid X_{pa(v)})$ to (one of) the clique(s) containing *v* & *pa(v)*
• setting all separator terms to 1

We can then manipulate the individual potentials, maintaining the identity

$$p(X) = \frac{\prod_{cliquesC} \psi(X_C)}{\prod_{separatorsS} \psi(X_S)}$$

- first until the potentials give the clique and separator marginals,
- and subsequently so they give the marginals, conditional on given data.
- The manipulations are done by ‘message-passing’ along the branches of the junction tree

DAG



A|B	A=0	A=1
B=0	3/4	1/4
B=1	2/3	1/3

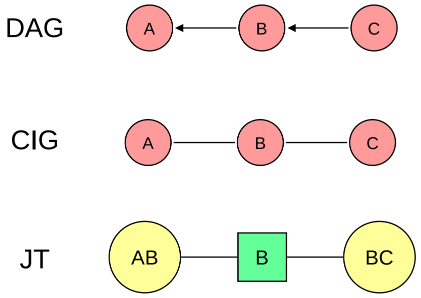
B|C	B=0	B=1
C=0	3/7	4/7
C=1	1/3	2/3

C=0	.7
C=1	.3

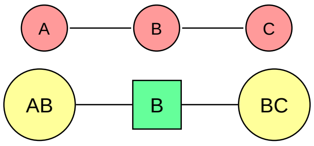
p(A,B,C) = p(A|B)p(B|C)p(C)

Wish to find p(B|A=0) , p(C|A=0)

Problem setup



Transformation of graph



	A=0	A=1
B=0	3/4	1/4
B=1	2/3	1/3

B=0	1
B=1	1

	C=0	C=1
B=0	.3	.1
B=1	.4	.2

A|B	A=0	A=1
B=0	3/4	1/4
B=1	2/3	1/3

B|C	B=0	B=1
C=0	3/7	4/7
C=1	1/3	2/3

C=0	.7
C=1	.3

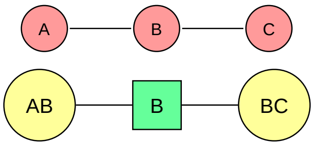
Initialisation of potential representation

We now have a valid potential representation

$$p(X) = \frac{\prod_{cliquesC} \psi(X_C)}{\prod_{separatorsS} \psi(X_S)}$$

$$p(A,B,C) = \frac{\psi(A,B)\psi(B,C)}{\psi(B)}$$

but individual potentials are not yet marginal distributions



	A=0	A=1
B=0	3/4	1/4
B=1	2/3	1/3

B=0	1
B=1	1

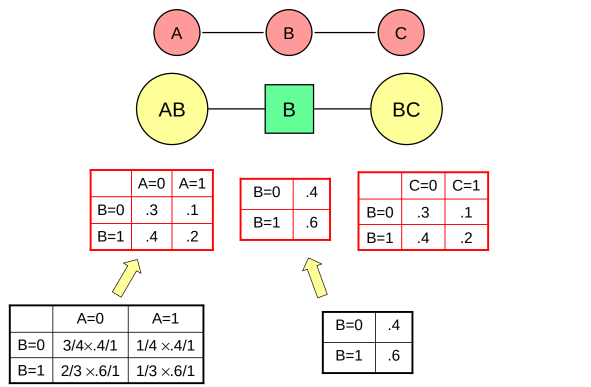
	C=0	C=1
B=0	.3	.1
B=1	.4	.2

	A=0	A=1
B=0	3/4×.4/1	1/4 ×.4/1
B=1	2/3 ×.6/1	1/3 ×.6/1

B=0	.4
B=1	.6

Passing message from BC *to* AB *(1)*

multiply
marginalise



	A=0	A=1
B=0	.3	.1
B=1	.4	.2

B=0	.4
B=1	.6

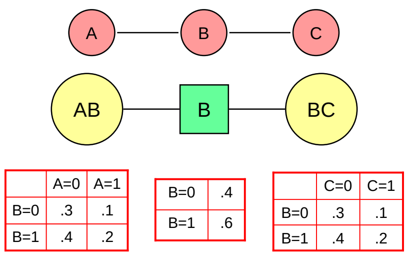
	C=0	C=1
B=0	.3	.1
B=1	.4	.2

	A=0	A=1
B=0	3/4×.4/1	1/4 ×.4/1
B=1	2/3 ×.6/1	1/3 ×.6/1

B=0	.4
B=1	.6

Passing message from BC *to* AB *(2)*

assign



	A=0	A=1
B=0	.3	.1
B=1	.4	.2

B=0	.4
B=1	.6

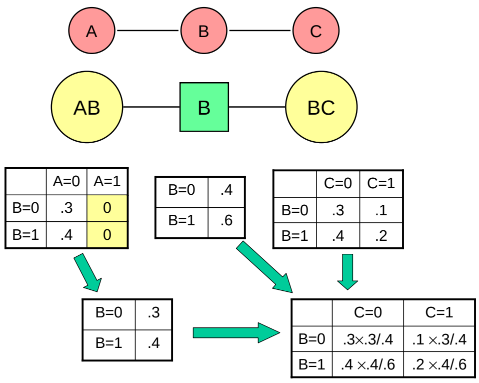
	C=0	C=1
B=0	.3	.1
B=1	.4	.2

After equilibration - marginal tables

We now have a valid potential representation where individual potentials *are* marginals:

$$p(X)=\frac{\prod_{\text{cliques}C} p(X_C)}{\prod_{\text{separators}S} p(X_S)}$$

$$p(A,B,C)=\frac{p(A,B)p(B,C)}{p(B)}$$



	A=0	A=1
B=0	.3	0
B=1	.4	0

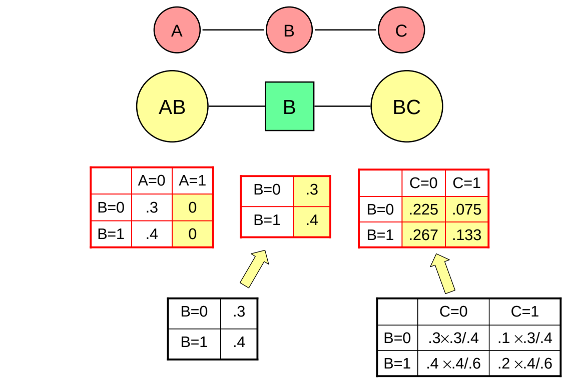
B=0	.4
B=1	.6

	C=0	C=1
B=0	.3	.1
B=1	.4	.2

B=0	.3
B=1	.4

	C=0	C=1
B=0	.3×.3/.4	.1 ×.3/.4
B=1	.4 ×.4/.6	.2 ×.4/.6

Propagating evidence (1)



	A=0	A=1
B=0	.3	0
B=1	.4	0

B=0	.3
B=1	.4

	C=0	C=1
B=0	.225	.075
B=1	.267	.133

B=0	.3
B=1	.4

	C=0	C=1
B=0	.3×.3/.4	.1 ×.3/.4
B=1	.4 ×.4/.6	.2 ×.4/.6

Propagating evidence (2)

We now have a valid potential representation

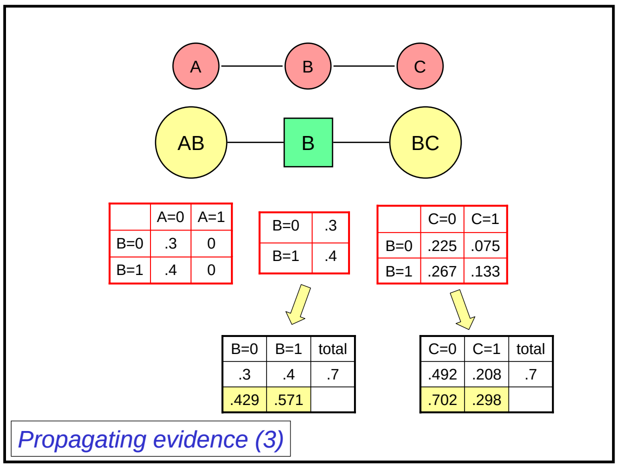
$$p(X)=\frac{\prod_{\text{cliques}C} \psi(X_C)}{\prod_{\text{separators}S} \psi(X_S)}$$

$$p(A,B,C)=\frac{\psi(A,B)\psi(B,C)}{\psi(B)}$$

where

$$\psi(X_E)=p(X_E\cap\{A=0\})$$

for any clique or separator *E*



	A=0	A=1
B=0	.3	0
B=1	.4	0

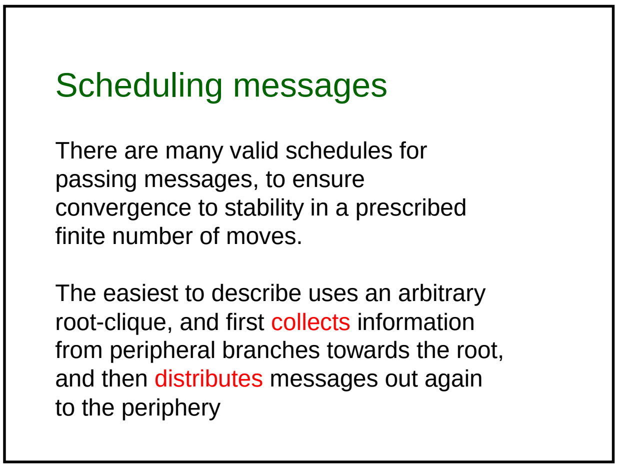
B=0	.3
B=1	.4

	C=0	C=1
B=0	.225	.075
B=1	.267	.133

B=0	B=1	total
.3	.4	.7
.429	.571	

C=0	C=1	total
.492	.208	.7
.702	.298	

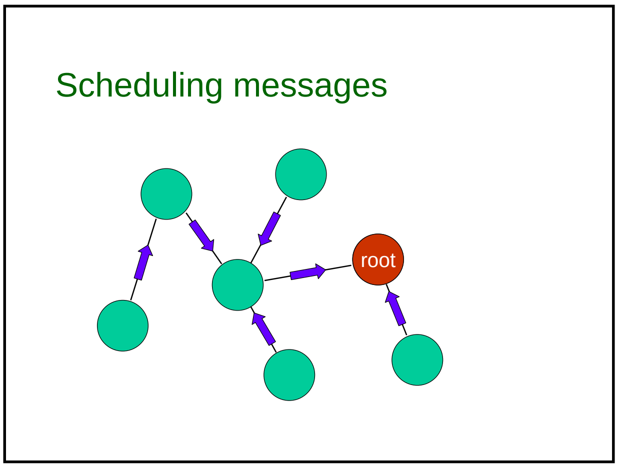
Propagating evidence (3)



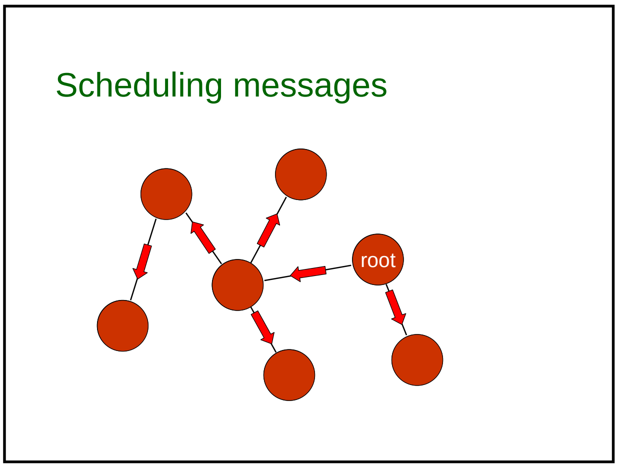
Scheduling messages

There are many valid schedules for passing messages, to ensure convergence to stability in a prescribed finite number of moves.

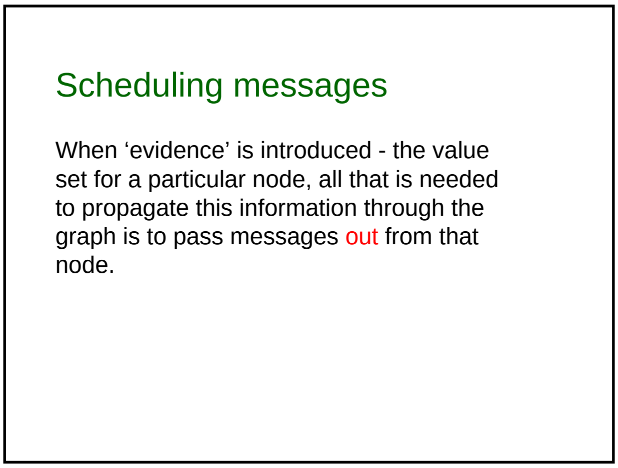
The easiest to describe uses an arbitrary root-clique, and first collects information from peripheral branches towards the root, and then distributes messages out again to the periphery



Scheduling messages



Scheduling messages



Scheduling messages

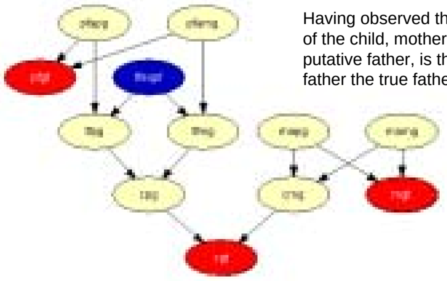
When ‘evidence’ is introduced - the value set for a particular node, all that is needed to propagate this information through the graph is to pass messages out from that node.

D. Applications

An example from forensic genetics

DNA profiling based on STR’s (single tandem repeats) are finding many uses in forensics, for identifying suspects, deciding paternity, etc. Can we use Mendelian genetics and Bayes’ theorem to make probabilistic inference in such cases?

Graphical model for a paternity enquiry - neglecting mutation

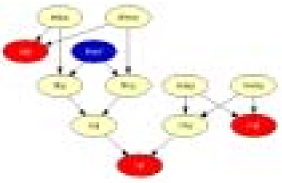


Having observed the genotype of the child, mother and putative father, is the putative father the true father?

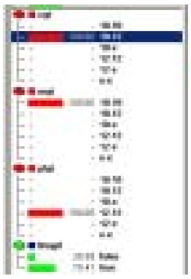
Graphical model for a paternity enquiry - neglecting mutation

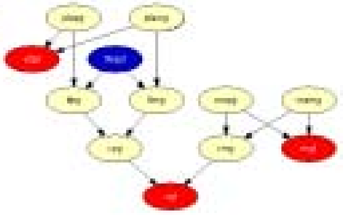
Having observed the genotype of the child, mother and putative father, is the putative father the true father?

Suppose we are looking at a gene with only 3 alleles - 10, 12 and ‘x’, with population frequencies 28.4%, 25.9%, 45.6% - the child is 10-12, the mother 10-10, the putative father 12-12



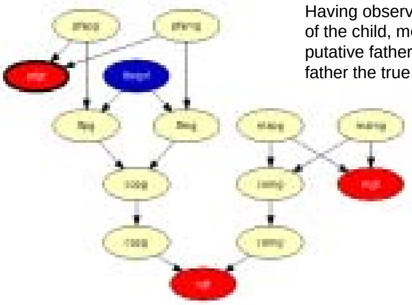
Graphical model for a paternity enquiry - neglecting mutation





⇒ we’re 79.4% sure the putative father is the true father

Graphical model for a paternity enquiry - allowing mutation



Having observed the genotype of the child, mother and putative father, is the putative father the true father?

DNA forensics example

(thanks to Julia Mortera)

- A blood stain is found at a crime scene
- A body is found somewhere else!
- There is a suspect
- DNA profiles on all three - crime scene sample is a ‘mixed trace’: is it a mix of the victim and the suspect?

DNA forensics in Hugin

- Disaggregate problem in terms of paternal and maternal genes of both victim and suspect.
- Assume Hardy-Weinberg equilibrium
- We have profiles on 8 STR markers - treated as independent (linkage equilibrium)

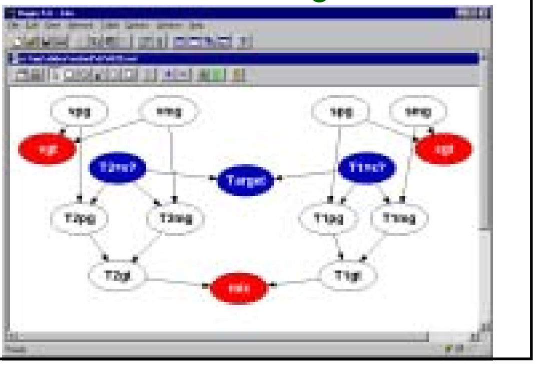
DNA forensics

The data:

Marker	Victim	Suspect	Crime scene
D3S1358	18 18	16 16	16 18
VWA	17 17	17 18	17 18
TH01	6 7	6 7	6 7
TPOX	8 8	8 11	8 11
D5S818	12 13	12 12	12 13
D13S317	8 8	8 11	8 11
FGA	22 26	24 25	22 24 25 26
D7S820	8 10	8 11	8 10 11

2 of 8 markers show more than 2 alleles at crime scene ⇒mixture of 2 or more people

DNA forensics in Hugin

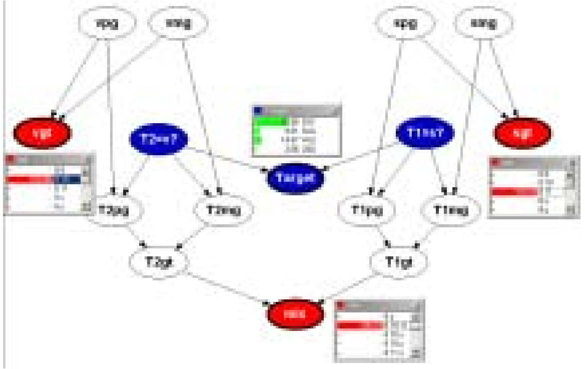


DNA forensics

Population gene frequencies for D7S820 (used as ‘prior’ on ‘founder’ nodes):

Allele	probability
8	.185
10	.135
11	.234
x	.233
y	.214

hugin



DNA forensics

Results (suspect+victim vs. unknown+victim):

Marker	Victim	Suspect	Crime scene	Likelihood ratio (sv/uv)
D3S1358	18 18	16 16	16 18	11.35
VWA	17 17	17 18	17 18	15.43
TH01	6 7	6 7	6 7	5.48
TPOX	8 8	8 11	8 11	3.00
D5S818	12 13	12 12	12 13	14.79
D13S317	8 8	8 11	8 11	24.45
FGA	22 26	24 25	22 24 25 26	76.92
D7S820	8 10	8 11	8 10 11	4.90
overall				3.93×10^{8}

Surgical rankings

- 12 hospitals carry out different numbers of a certain type of operation:

  47, 148, 119, 810, 211, 196, 148, 215, 207, 97, 256, 360 respectively.
- They are differently successful, and there are:

  0, 18, 8, 46, 8, 13, 9, 31, 14, 8, 29, 24 fatalities, respectively.

Surgical rankings, continued

- What inference can we draw about the relative qualities of the hospitals based on these data?
- A natural model is to say the number of deaths y_i in hospital i has a Binomial distribution $y_i \sim Bin(n_i,p_i)$ where the n_i are the numbers of operations, and it is the p_i that we want to make inference about.

Surgical rankings, continued

- How to model the p_i?
- We do not want to assume they are all the same.
- But they are not necessarily `completely different'.
- In a Bayesian approach, we can say that the p_i are random variables, drawn from a common distribution.

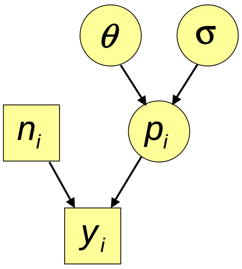
Surgical rankings, continued

- Specifically, we could take

$$\log\frac{p_i}{1-p_i} \sim N(\theta,\sigma^2)$$

- If θ and σ^2 are fixed numbers, then inference about p_i only depends on y_i (and n_i, θ and σ^2).

Graph for surgical rankings

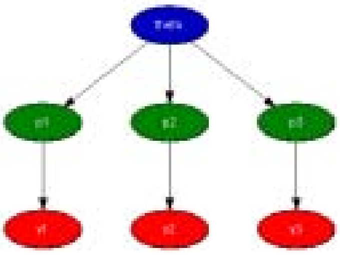


Surgical rankings, continued

- But don't you think that knowing that p_1=0.08, say, would tell you *something* about p_2?
- Putting prior distributions on θ and σ^2 allows `borrowing strength' between data from different hospitals

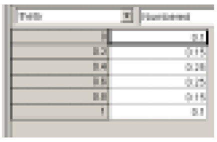
Surgical rankings - simplified

3 hospitals, *p* discrete, only one hyperparameter



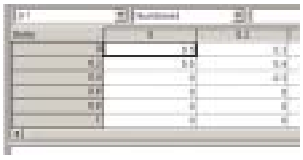
Surgical rankings - simplified

prior for θ



Theta	Numbered
0	0.1
0.2	0.15
0.4	0.25
0.6	0.25
0.8	0.15
1	0.1

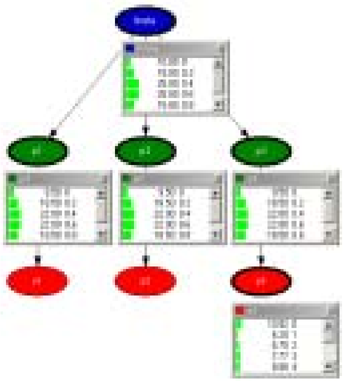
prior for p_i given θ



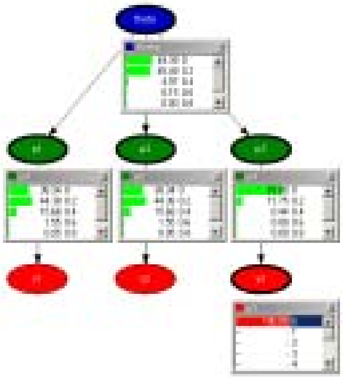
p1 Numbered

theta	0	0.2
0	0.5	0.1
0.2	0.5	0.4
0.4	0	0.1
0.6	0	0
0.8	0	0
1	0	0

Surgical rankings



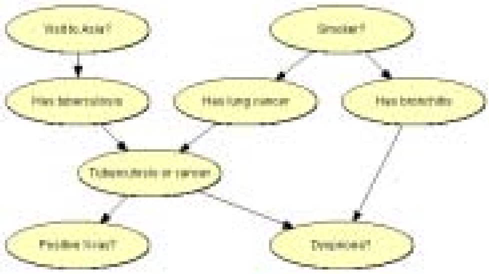
Surgical rankings



The ‘Asia’ (chest-clinic) example

Shortness-of-breath (dyspnoea) may be due to tuberculosis, lung cancer, bronchitis, more than one of these diseases or none of them. A recent visit to Asia increases the risk of tuberculosis, while smoking is known to be a risk factor for both lung cancer and bronchitis. The results of a single chest X-ray do not discriminate between lung cancer and tuberculosis, as neither does the presence or absence of dyspnoea.

Visual representation of the Asia example - a graphical model



The ‘Asia’ (chest-clinic) example

Now … a patient presents with shortness-of-breath (dyspnoea) …. How can the physician use available tests (X-ray) and enquiries about the patient’s history (smoking, visits to Asia) to help to diagnose which, if any, of tuberculosis, lung cancer, or bronchitis is the patient probably suffering from?

E. Proofs

E. Proofs

Factorisation of joint distribution,
forming potential representation,
when graph is decomposable

Decomposability

The following are equivalent

- *G* is decomposable
- *G* is triangulated (or chordal)
- The cliques of *G* may be ‘perfectly numbered’ to satisfy the running intersection property

$$C_i \cap \bigcup_{j<i} C_j \subseteq C_{i^*} \forall i = 2,3,\ldots,k$$

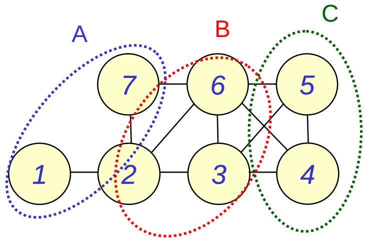
where $i^* \in \{1,2,\ldots,i-1\}$

Decomposability

G is decomposable means that either

- *G* is complete, or
- *G* admits a proper decomposition (*A*,*B*,*C*), that is:
  - *B* separates *A* and *C*
  - *B* is complete, *A* and *C* are non-empty
  - the subgraphs $G_{A\cup B}$ and $G_{B\cup C}$ are decomposable

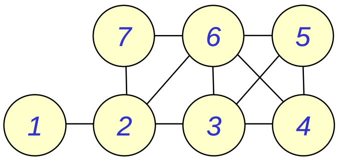
A decomposable graph



Decomposability

G is triangulated or chordal means that

- *G* has no loops of 4 or more vertices without a chord



Decomposability

The running intersection property

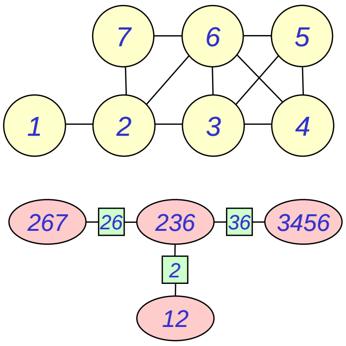
$$C_i \cap \bigcup_{j<i} C_j \subseteq C_{i^*} \forall i = 2,3,\ldots,k$$
$$i^* \in \{1,2,\ldots,i-1\}$$

is what allows the construction of the junction tree and the possibility of probability propagation

The junction tree

For *i*=2,3,…,*k*, join C_i to C_{i^*}, labelling the edge by S_i

A decomposable graph and (one of) its junction tree(s)



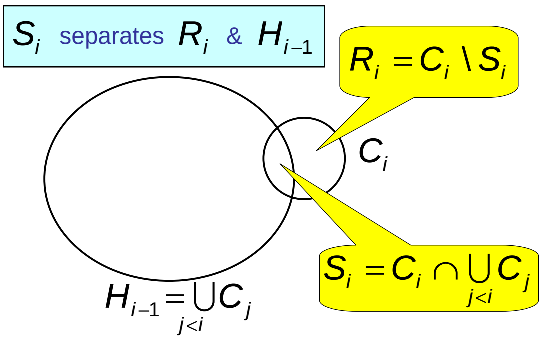
Decomposability

In $C_i \cap \bigcup_{j<i} C_j \subseteq C_{i^*} \forall i = 2,3,\ldots,k$

let
$$S_i = C_i \cap \bigcup_{j<i} C_j$$
$$R_i = C_i \setminus S_i$$
$$H_{i-1} = \bigcup_{j<i} C_j$$

then
$$S_i = C_i \cap H_{i-1} \subseteq C_{i^*} \forall i = 2,3,\ldots,k$$

Decomposability



S_i separates R_i & H_{i-1}

$R_i = C_i \setminus S_i$

C_i

$S_i = C_i \cap \bigcup_{j<i} C_j$

$H_{i-1} = \bigcup_{j<i} C_j$

Factorisation of joint distribution

Recall $H_{i-1} = \bigcup_{j<i} C_j$, then

$$p(V) = p(H_1)p(C_2 \setminus H_1 \mid H_1)\times$$

$$p(C_3 \setminus H_2 \mid H_2)...p(C_k \setminus H_{k-1} \mid H_{k-1})$$

but the typical factor is

$$p(C_i \setminus H_{i-1} \mid H_{i-1}) = p(R_i \mid H_{i-1})$$

$$= p(R_i \mid S_i) = \frac{p(R_i,S_i)}{p(S_i)} = \frac{p(C_i)}{p(S_i)}$$

Factorisation of joint distribution

So

$$p(V) = \frac{\prod_{i=1}^{k} p(C_i)}{\prod_{i=2}^{k} p(S_i)}$$

as required

E. Proofs

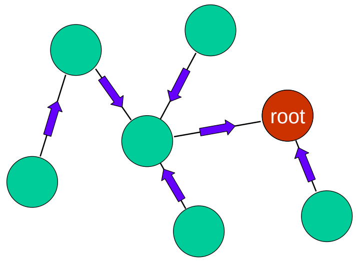
The collect/distribute schedule ensures equilibrium in message-passing

Scheduling messages

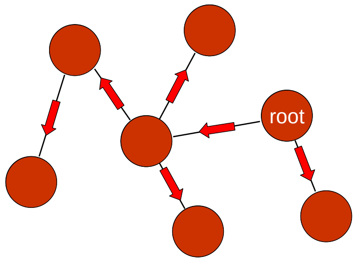
There are many valid schedules for passing messages, to ensure convergence to stability in a prescribed finite number of moves.

The easiest to describe uses an arbitrary root-clique, and first collects information from peripheral branches towards the root, and then distributes messages out again to the periphery

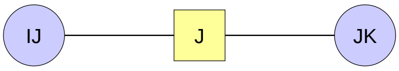
Scheduling messages



Scheduling messages



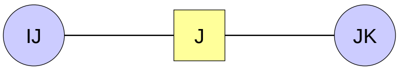
Consider a single edge of the junction tree



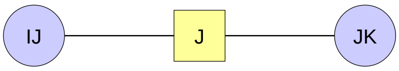
(I, J and K may be vectors)

- Edge is in equilibrium if J table is equal to J marginal in both IJ and JK tables
- Tree is in equilibrium if every edge is

Consider a single edge of the junction tree



Messages are [1] passed into IJ, then [2] from IJ to JK, then [3] from JK to root and back to JK, then [4] from JK to IJ, then [5] from IJ to ‘leaves’ of tree.

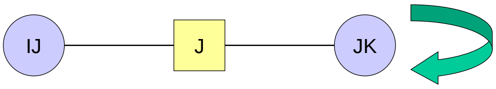


$$a_{ij} \qquad b_j \qquad c_{jk}$$

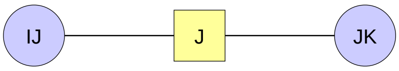
State after message passed from IJ to JK

$$a_{ij} \qquad \sum_i a_{ij} \qquad \frac{c_{jk}\sum_i a_{ij}}{b_j}$$

Messages passed from JK to root and back to JK



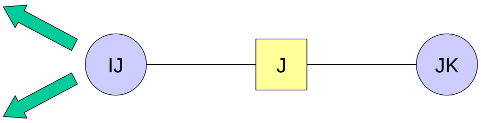
As a result, JK table gets multiplied by a term indexed by *(j,k)* - but not *i*



$$a_{ij} \qquad \sum_i a_{ij} \qquad \frac{c_{jk}\sum_i a_{ij}}{b_j}\times d_{jk}$$

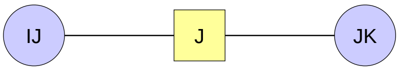
$$\frac{\left(\sum_k c_{jk}d_{jk}\right)\left(a_{ij}\right)}{b_j} \qquad \frac{\left(\sum_k c_{jk}d_{jk}\right)\left(\sum_i a_{ij}\right)}{b_j} \qquad \frac{\left(c_{jk}d_{jk}\right)\left(\sum_i a_{ij}\right)}{b_j}$$

Messages passed from IJ back to leaves



IJ, J and JK tables are not changed again

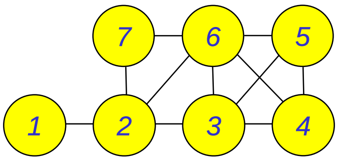
Final tables



$$\frac{\left(\sum_k c_{jk} d_{jk}\right)(a_{ij})}{b_j} \qquad \frac{\left(\sum_k c_{jk} d_{jk}\right)\left(\sum_i a_{ij}\right)}{b_j} \qquad \frac{(c_{jk} d_{jk})\left(\sum_i a_{ij}\right)}{b_j}$$

- satisfy equilibrium conditions

Software



- The HUGIN system: freeware version (Hugin Lite 5.7): http://www.stats.bris.ac.uk/~peter/Hugin57.zip
- Grappa (suite of R functions) http://www.stats.bris.ac.uk/~peter/Grappa

Module outline

- Information, uncertainty and probability
- Motivating examples
- Graphical models
- Probability propagation
- The HUGIN system