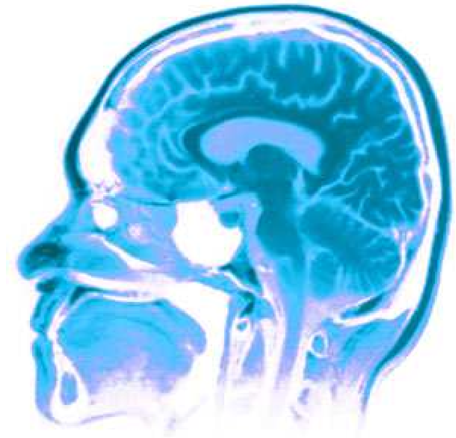
CPSC540



Directed Graphical Models

Nando de Freitas

2011

KPM Book Sections: *6*

$$P(B|A) = \frac{P(A|B)\,P(B)}{P(A)}$$

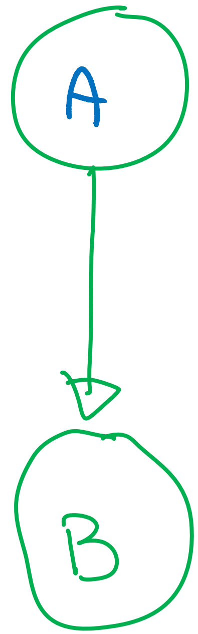
∴ since $\sum_B P(B|A) = 1$ (sum over the possible values of B)

we have

$$P(A) = \sum_B P(A|B)\,P(B) = \sum_B P(A,B)$$

That is,

$$P(B|A) = \frac{P(A|B)\,P(B)}{\sum_B P(A|B)\,P(B)}$$



0	1
0.4	0.6

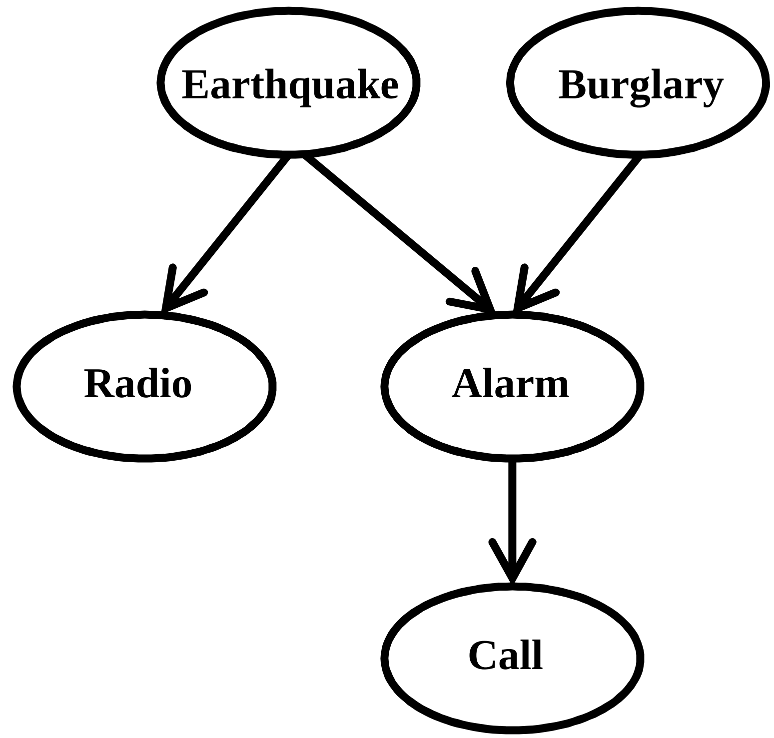
$= P(A)$

A \ B	0	1
0	0.1	0.9
1	0.5	0.5

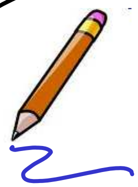
$= P(B \mid A)$

Directed graphical models

- Directed acyclic graph (DAG)
  - **Nodes** – random variables
  - **Edges** – direct influence ("causation")
- X_i independent of $X_{\text{ancestors}} \mid X_{\text{parents}}$
- Simplifies chain rule by using conditional independencies



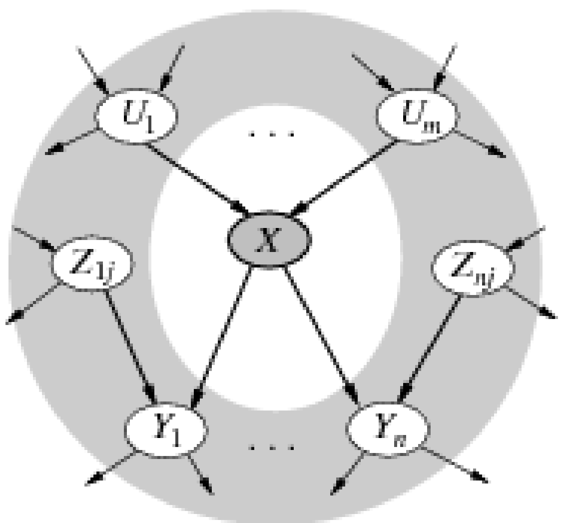
$$
\begin{aligned}
&P(B, E, A, R, C) \\
&= P(B)P(E|\cancel{B})P(A|B, E)P(R|\cancel{A}, \cancel{B}, E)P(C|\cancel{R}, A, \cancel{B}, \cancel{E}) \\
&= P(B)P(E)P(A|B, E)P(R|E)P(C|A)
\end{aligned}
$$



Markov blankets for DAGs

- The Markov blanket of a node is the set that renders it independent of the rest of the graph.
- This is the parents, children and co-parents.

$$p(X_i|X_{-i}) = \frac{p(X_i, X_{-i})}{\sum_x p(X_i, X_{-i})}$$



$$= \frac{p(X_i, U_{1:n}, Y_{1:m}, Z_{1:m}, R)}{\sum_x p(x, U_{1:n}, Y_{1:m}, Z_{1:m}, R)}$$

$$= \frac{p(X_i|U_{1:n})[\prod_j p(Y_j|X_i, Z_j)]P(U_{1:n}, Z_{1:m}, R)}{\sum_x p(X_i = x|U_{1:n})[\prod_j p(Y_j|X_i = x, Z_j)]P(U_{1:n}, Z_{1:m}, R)}$$

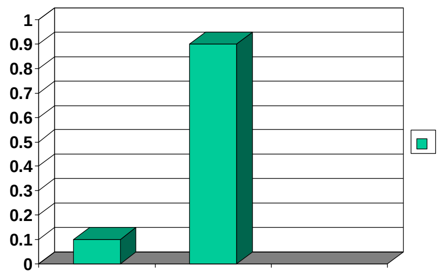
$$= \frac{p(X_i|U_{1:n})[\prod_j p(Y_j|X_i, Z_j)]}{\sum_x p(X_i = x|U_{1:n})[\prod_j p(Y_j|X_i = x, Z_j)]}$$

$$p(X_i|X_{-i}) \propto p(X_i|Pa(X_i)) \prod_{Y_j \in ch(X_i)} p(Y_j|Pa(Y_j))$$

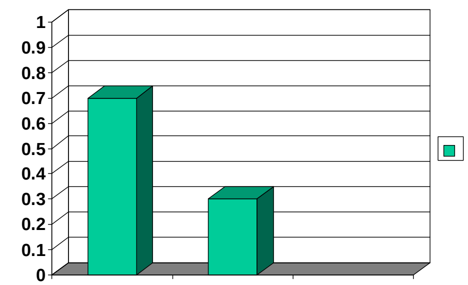
Useful for Gibbs sampling

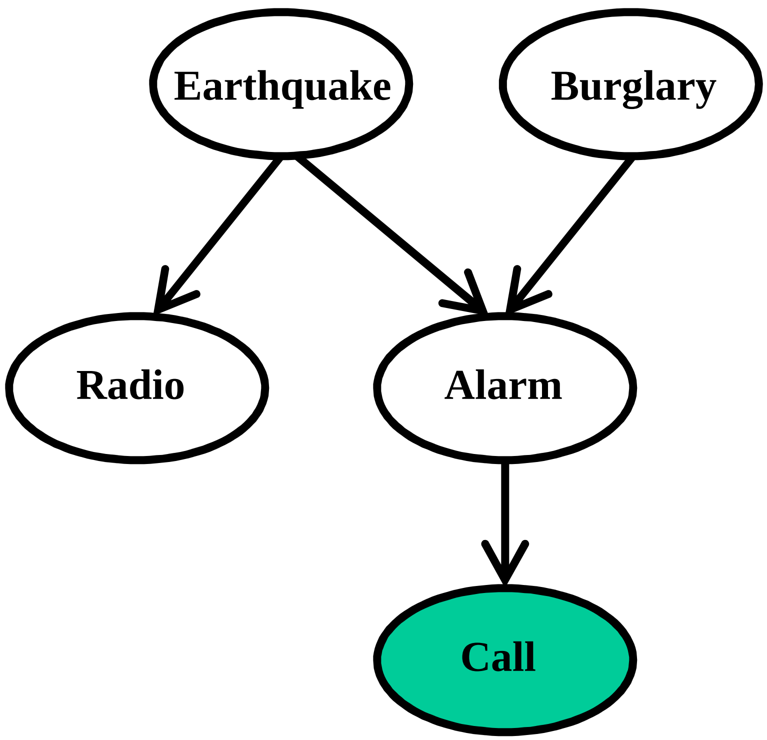
Inference

P(E=t|C=t)=0.1



P(B=t|C=t) = 0.7

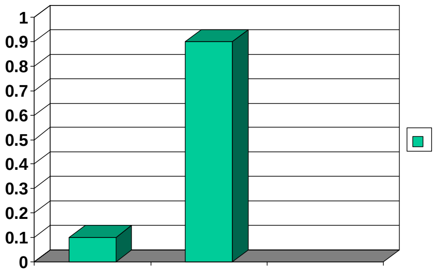




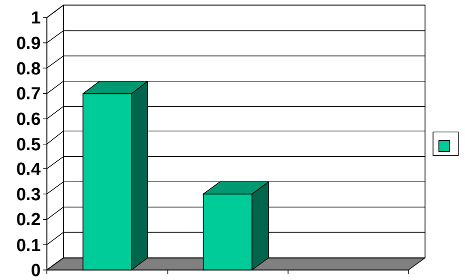
C=t

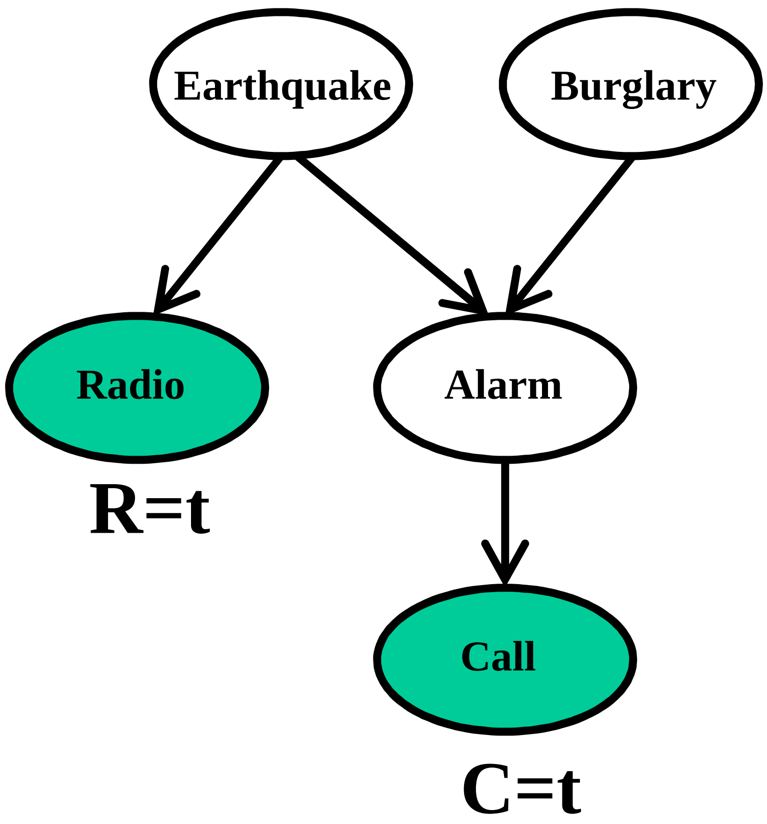
Inference

P(E=t|C=t)=0.1

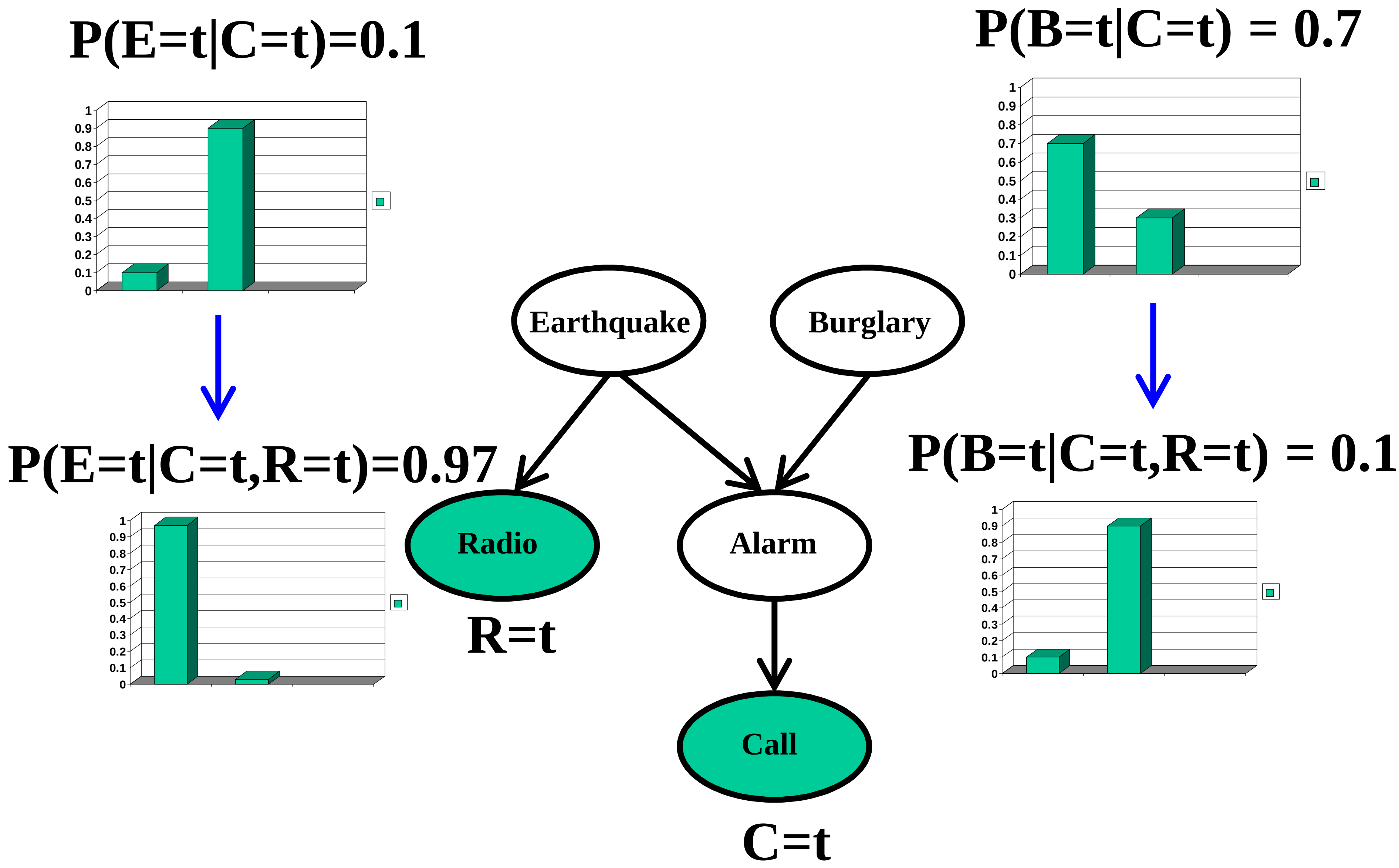


P(B=t|C=t) = 0.7

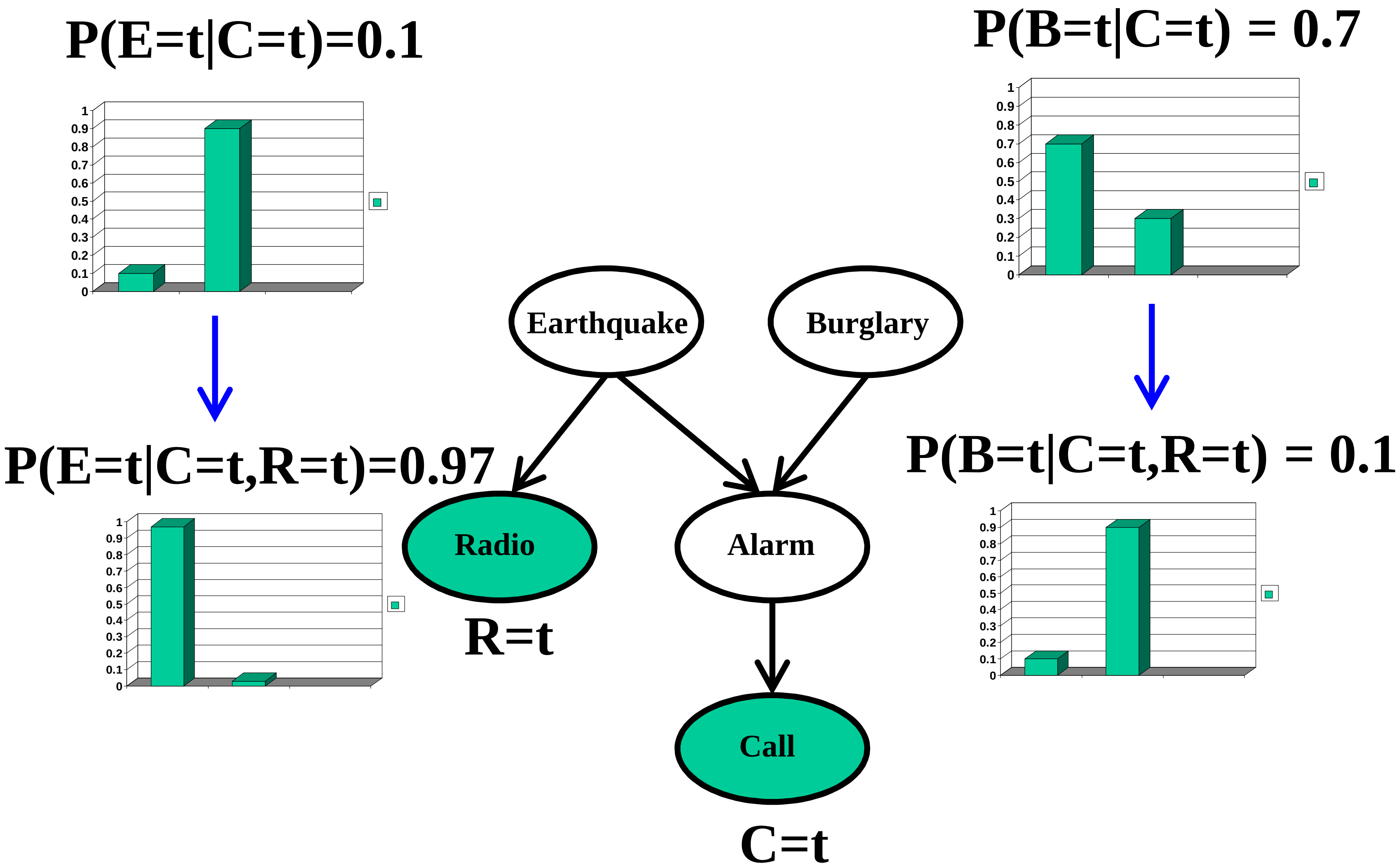




Inference


P(E=t|C=t)=0.1
P(B=t|C=t) = 0.7
P(E=t|C=t,R=t)=0.97
P(B=t|C=t,R=t) = 0.1
Earthquake
Burglary
Radio
Alarm
Call
R=t
C=t

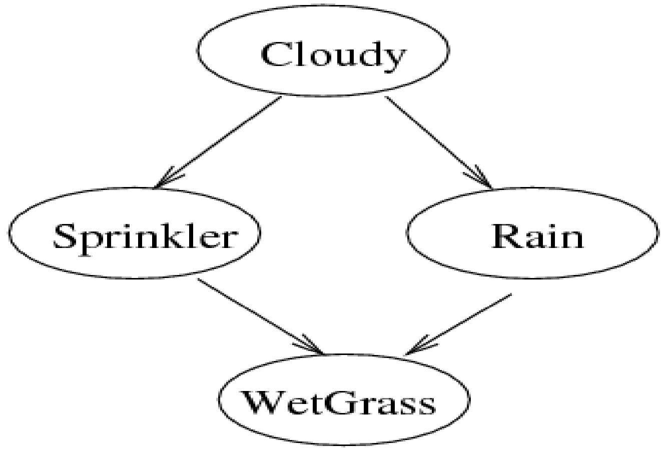
Inference



Explaining away effect

Example model

P(C=F)	P(C=T)
0.5	0.5



C	P(S=F)	P(S=T)
F	0.5	0.5
T	0.9	0.1

C	P(R=F)	P(R=T)
F	0.8	0.2
T	0.2	0.8

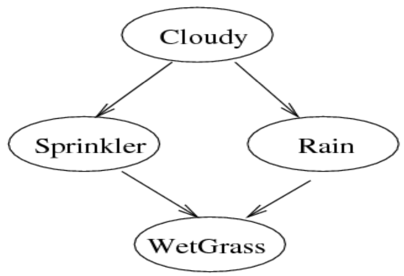
S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

$$p(C,S,R,W) = p(C)p(S|C)p(R|C)p(W|S,R)$$

Joint distribution

$$p(C,S,R,W) = p(C)p(S|C)p(R|C)p(W|S,R)$$

P(C=F)	P(C=T)
0.5	0.5



C	P(S=F)	P(S=T)
F	0.5	0.5
T	0.9	0.1

C	P(R=F)	P(R=T)
F	0.8	0.2
T	0.2	0.8

S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

c	s	r	w	prob
0	0	0	0	0.200
0	0	0	1	0.000
0	0	1	0	0.005
0	0	1	1	0.045
0	1	0	0	0.020
0	1	0	1	0.180
0	1	1	0	0.001
0	1	1	1	0.050
1	0	0	0	0.090
1	0	0	1	0.000
1	0	1	0	0.036
1	0	1	1	0.324
1	1	0	0	0.001
1	1	0	1	0.009
1	1	1	0	0.000
1	1	1	1	0.040

Inference

$$p(S=1) = \sum_{c=0}^{1}\sum_{r=0}^{1}\sum_{w=0}^{1} p(C=c, S=1, R=r, W=w) = 0.3$$

- Prior that sprinkler is on

$$p(S=1|W=1) = \frac{p(S=1, W=1)}{p(W=1)} = 0.43$$

- Posterior that sprinkler is on given that grass is wet

$$p(S=1|W=1, R=1) = \frac{p(S=1, W=1, R=1)}{p(W=1, R=1)} = 0.19$$

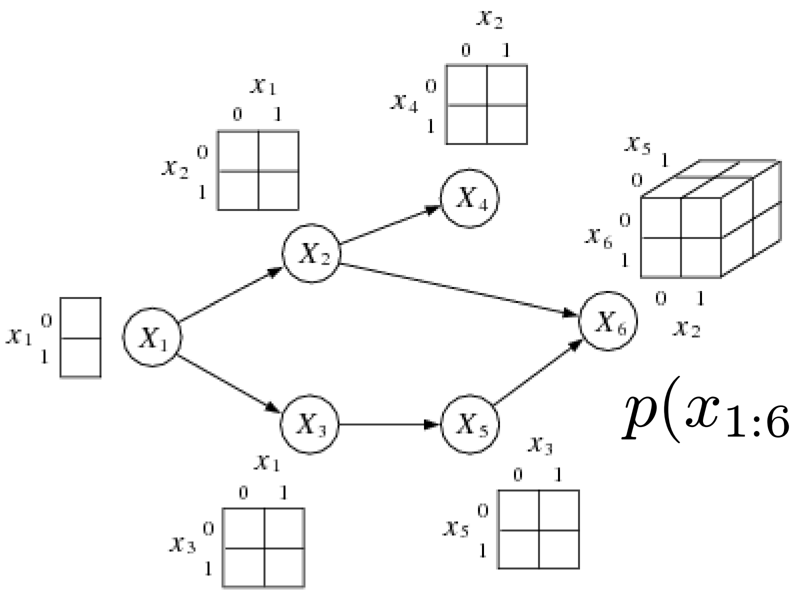
- Posterior that sprinkler is on given that grass is wet and it is raining **Explaining away!**

Directed graphical models

- A prob distribution factorizes according to a DAG if it can be written as

$$p(\mathbf{x}) = \prod_{j=1}^{d} p(x_j | \mathbf{x}_{\pi_j})$$

where π_j are the parents of j , and the nodes are ordered topologically (parents before children).



Each row of the conditional probability table (CPT) defines the distribution over the child's values given its parents values. The model is locally normalized.

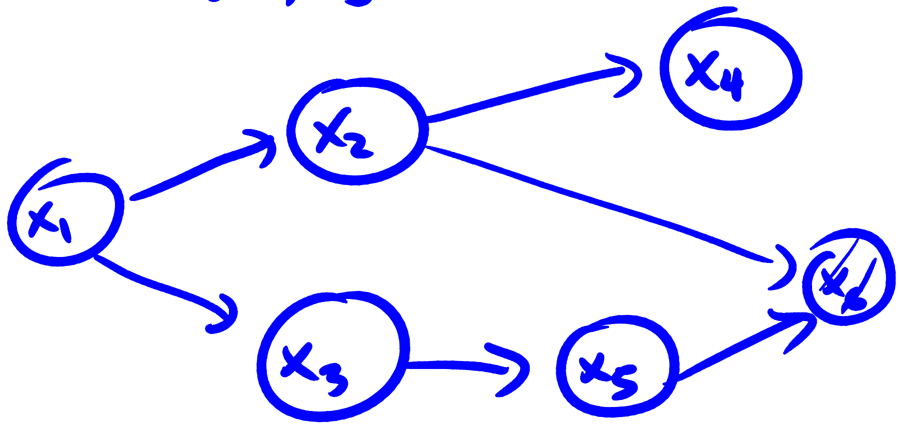
$$p(x_{1:6}) = p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)$$
$$p(x_5|x_3)p(x_6|x_2, x_5)$$

$ab + ac$

$a(b+c)$

Efficient inference in DAGs

$x_i \in \{0,1\}$



$P(x_1 | x_6 = 1) = ?$

$$P(x_1 | x_6 = 1) = \frac{P(x_1, x_6 = 1)}{P(x_6 = 1)}$$

$$P(x_1 | x_6 = 1) = \frac{\sum_{x_2} \sum_{x_3} \sum_{x_4} \sum_{x_5} P(x_1 x_2 x_3 x_4 x_5 x_6 = 1)}{\sum_{x_{1:5}} P(x_{1:5}, x_6 = 1)}$$

Efficient inference in DAGs

Efficient inference in DAGs

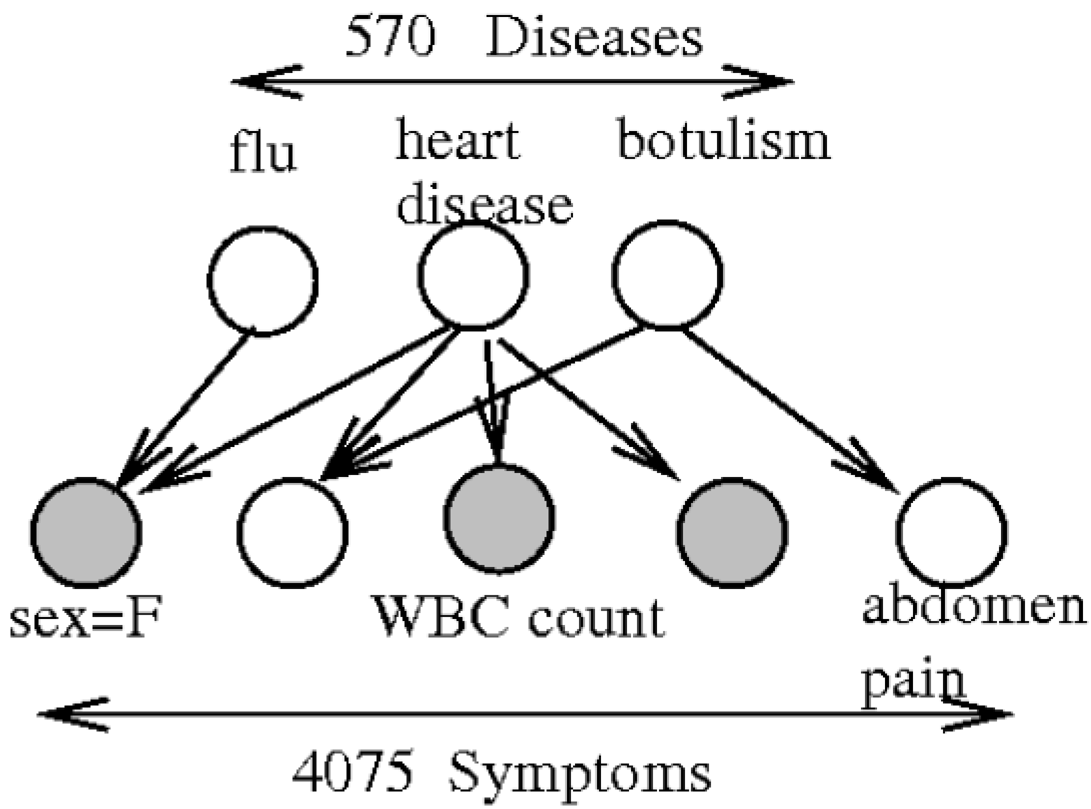
Junction tree algorithm

The idea of replacing sums of products ($ac+ab$) by products of sums ($a(b+c)$) is at the heart of most inference algorithms. For exact inference, in Gaussian and discrete networks of reasonable size, we use the **junction tree algorithm**. This algorithm involves two steps:

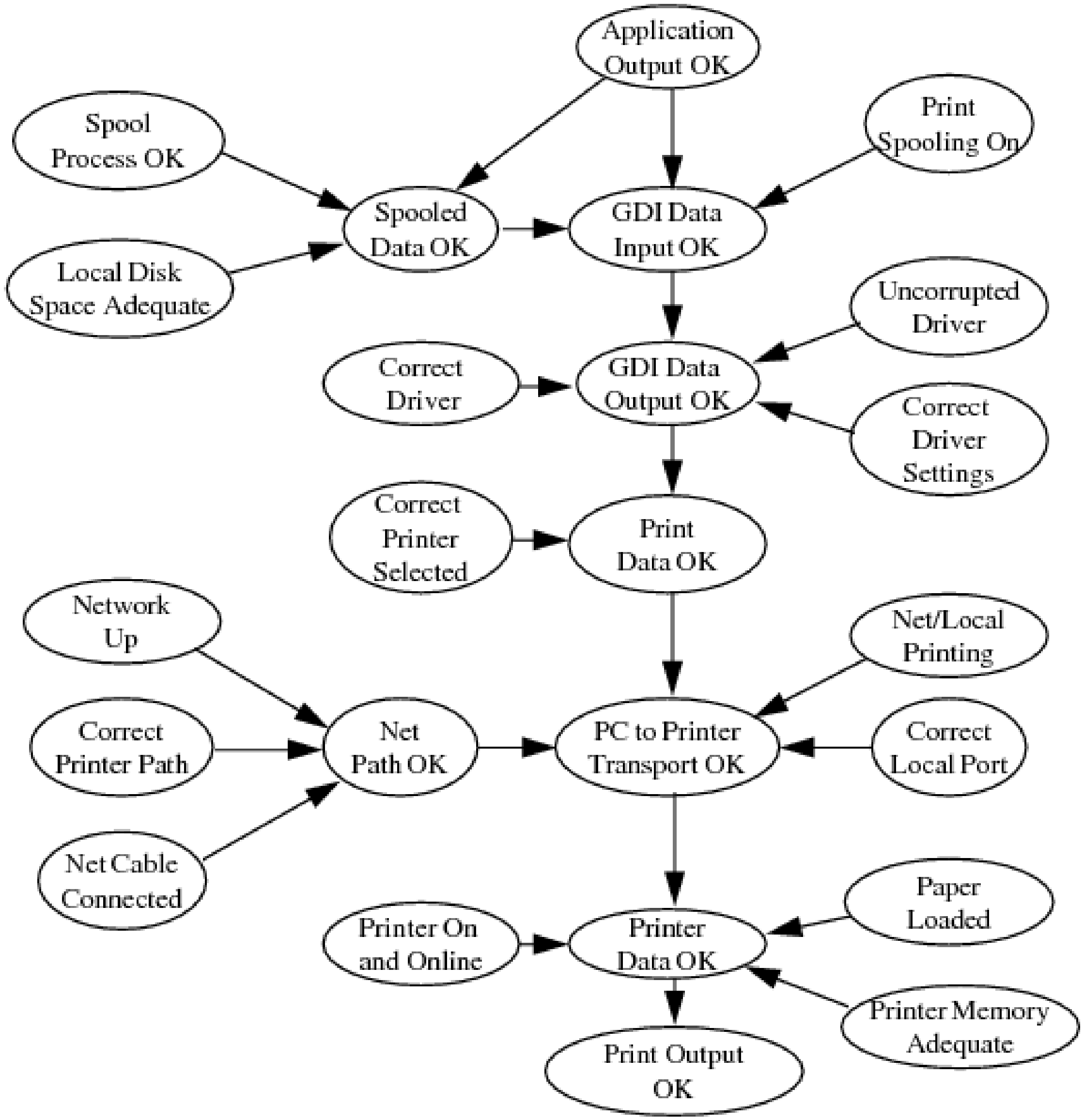
1. Converting the directed graph to an undirected graph called the junction tree.

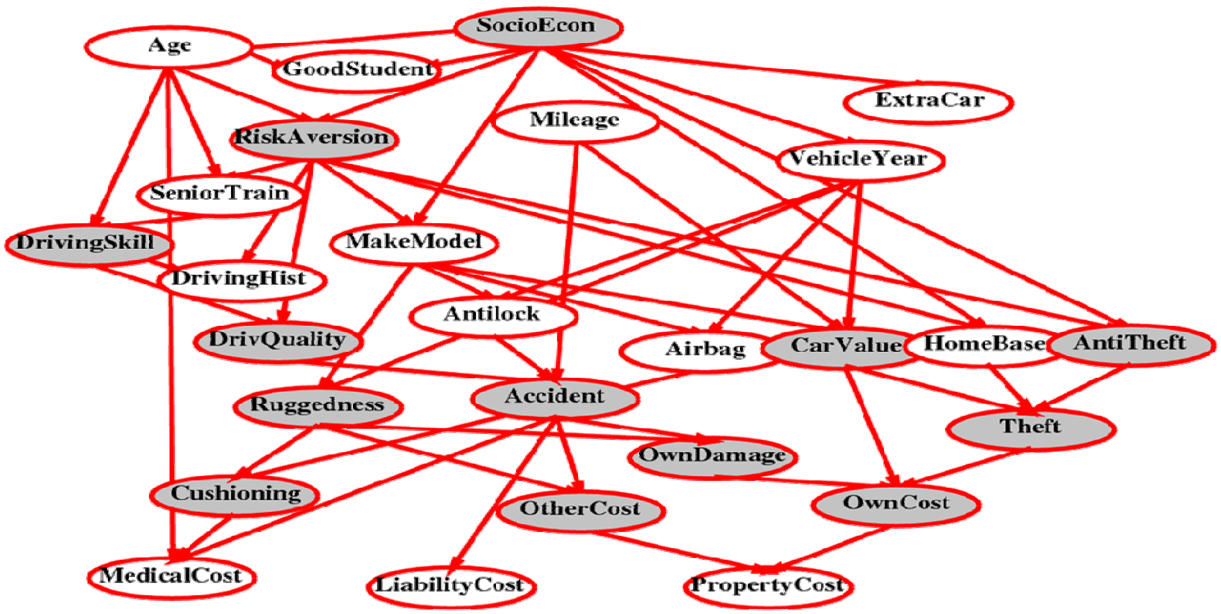
2. Running belief propagation. That is, replace sums of products by products of sums.

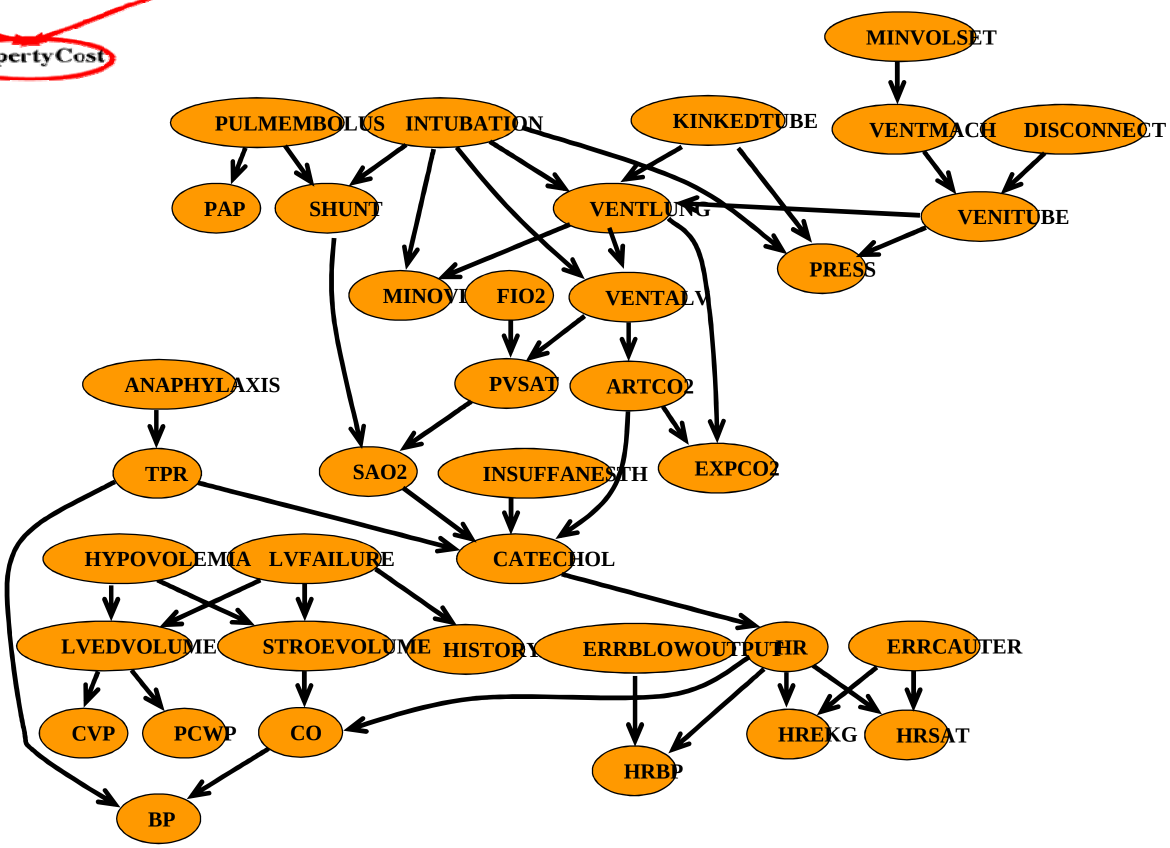
Diagnoses


570 Diseases
flu
heart disease
botulism
sex=F
WBC count
abdomen pain
4075 Symptoms

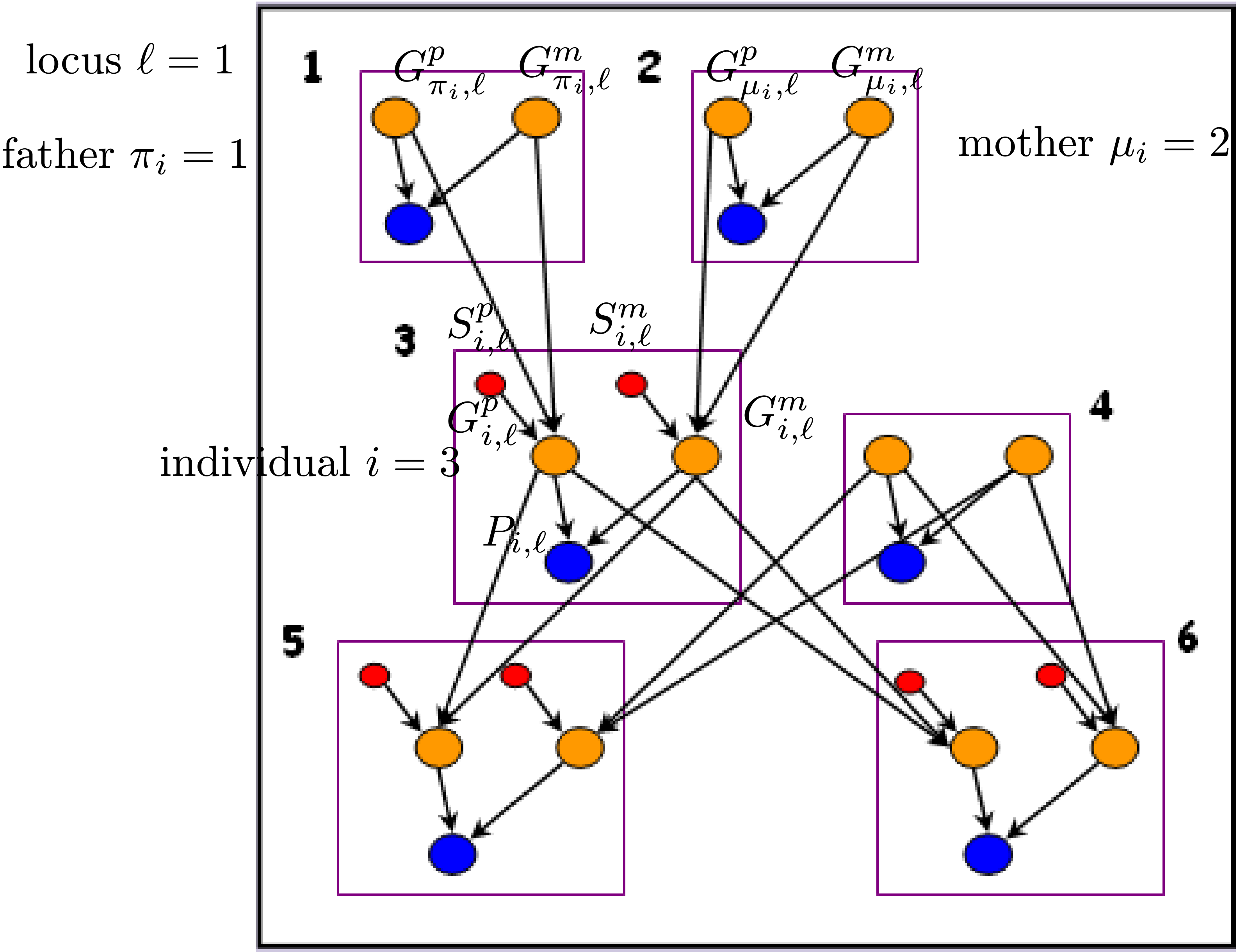
Microsoft Windows Printer Troubleshooter



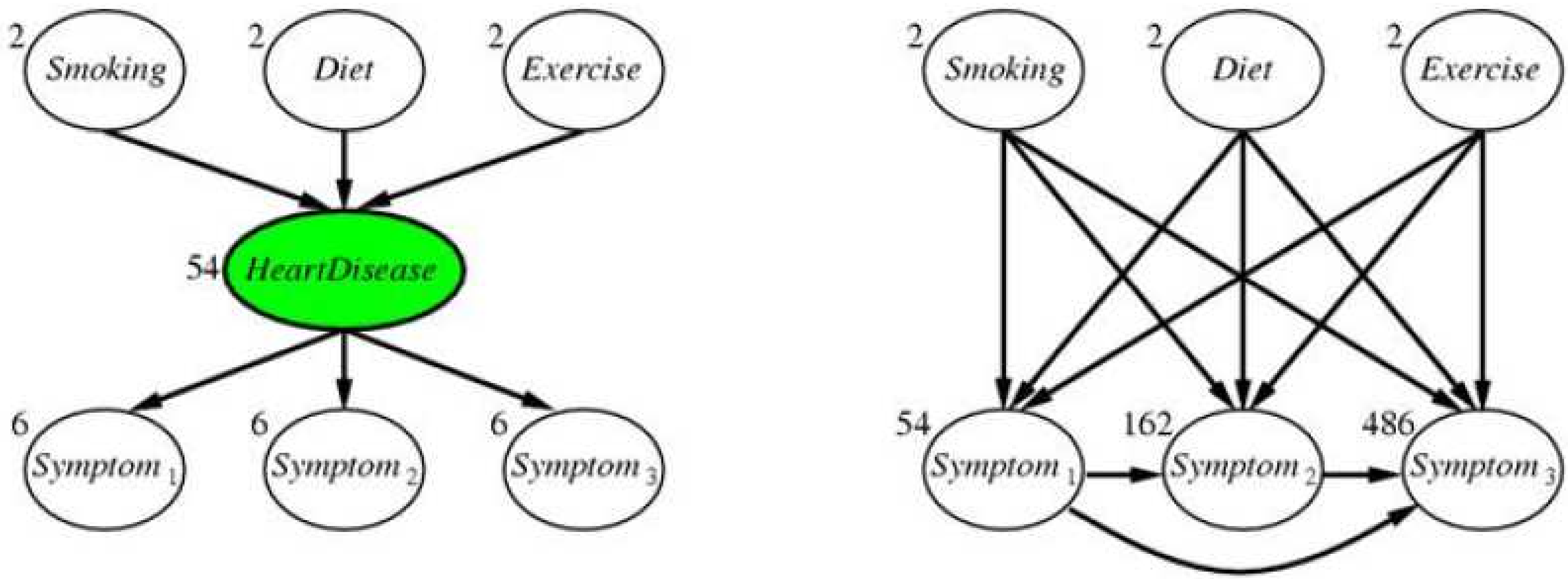

Age
SocioEcon
GoodStudent
ExtraCar
Mileage
RiskAversion
VehicleYear
SeniorTrain
DrivingSkill
MakeModel
DrivingHist
Antilock
DrivQuality
Airbag
CarValue
HomeBase
AntiTheft
Ruggedness
Accident
Theft
OwnDamage
Cushioning
OtherCost
OwnCost
MedicalCost
LiabilityCost
PropertyCost


MINVOLSET
PULMEMBOLUS
INTUBATION
KINKEDTUBE
VENTMACH
DISCONNECT
PAP
SHUNT
VENTLUNG
VENITUBE
PRESS
MINOVL
FIO2
VENTALV
ANAPHYLAXIS
PVSAT
ARTCO2
TPR
SAO2
INSUFFANESTH
EXPCO2
HYPOVOLEMIA
LVFAILURE
CATECHOL
LVEDVOLUME
STROEVOLUME
HISTORY
ERRBLOWOUTPUT
HR
ERRCAUTER
CVP
PCWP
CO
HREKG
HRSAT
HRBP
BP

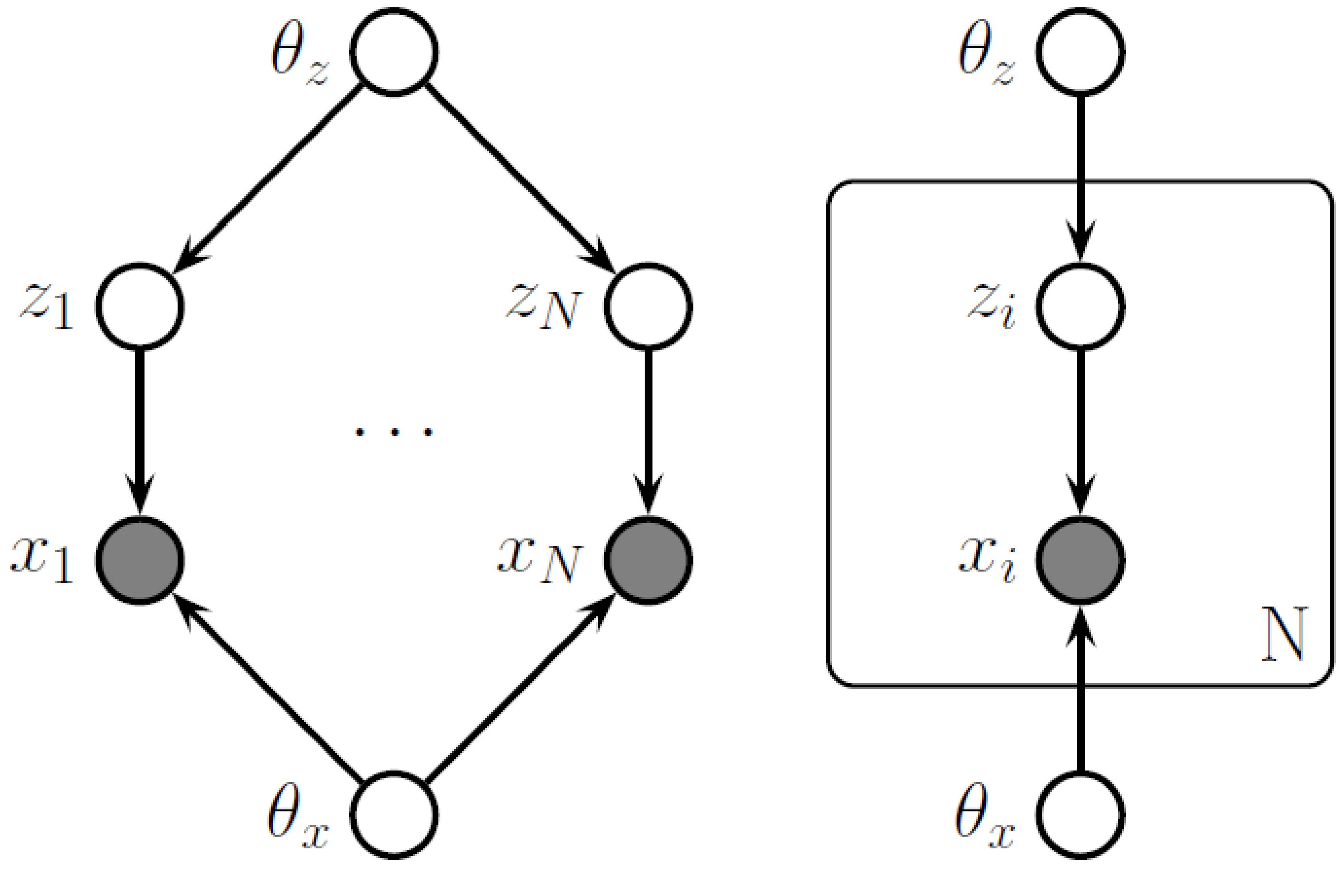
Pedigree tree



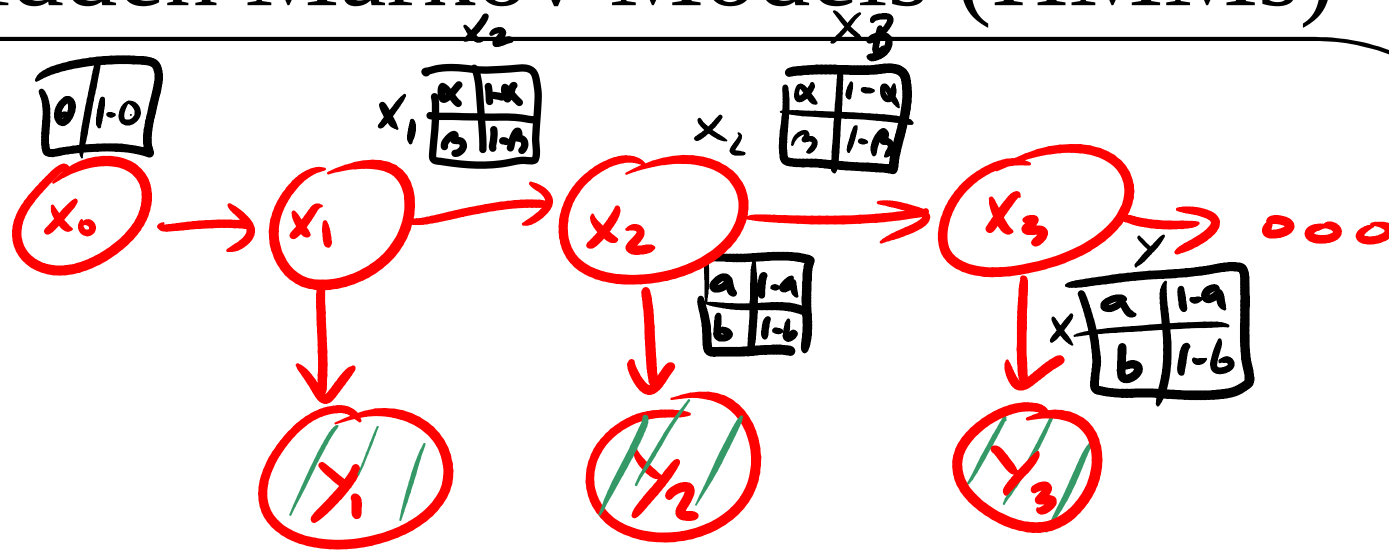
Latent variables



Plates



Hidden Markov Models (HMMs)



What is $P(x_2 \mid y_1, y_2)$?

Given: $P(x_t \mid x_{t-1}) =$

	0	1
0	α	$1-\alpha$
1	β	$1-\beta$

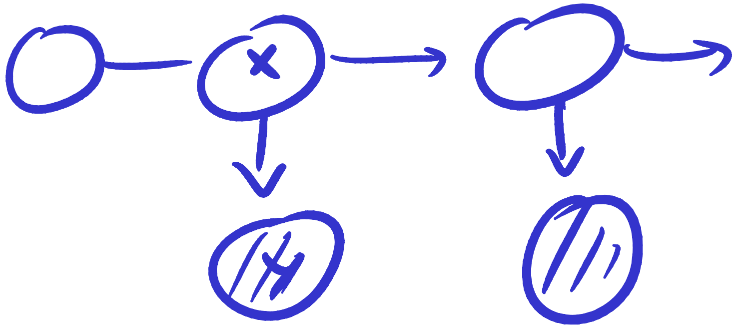
dynamic model

$P(y_t \mid x_t) =$

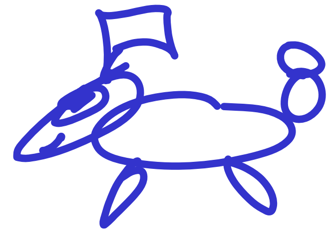
	a	1-a
	b	1-b

Observation model

Hidden Markov Models (HMMs)



AIBO



$x \in \{\text{sad}, \text{happy}\}$

$y \in \{\text{watch TV}, \text{sleeping}, \text{crying}, \text{socializing}\}$

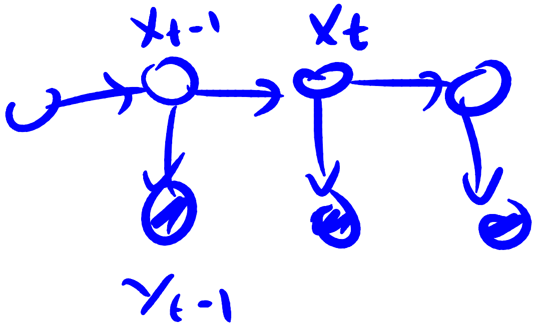
$P(x_t \mid x_{t-1}) =$

	s	h
s	0.8	0.2
h	0.09	0.91

	w	s	c	soc
S	0.3	0	0.1	0.6
H	0.3	0.7	0	0

$P(y_t \mid x_t)$

Hidden Markov Models (HMMs)



$$x_t \perp\!\!\!\perp y_{t-1} \mid x_{t-1}$$

$P(x_t \mid y_{1:t})$ is what we want

① $$P(x_t \mid y_{1:t-1}) = \sum_{x_{t-1}} P(x_t, x_{t-1} \mid y_{1:t-1})$$

$$= \sum_{x_{t-1}} P(x_t \mid x_{t-1}) P(x_{t-1} \mid y_{1:t-1})$$

Hidden Markov Models (HMMs)

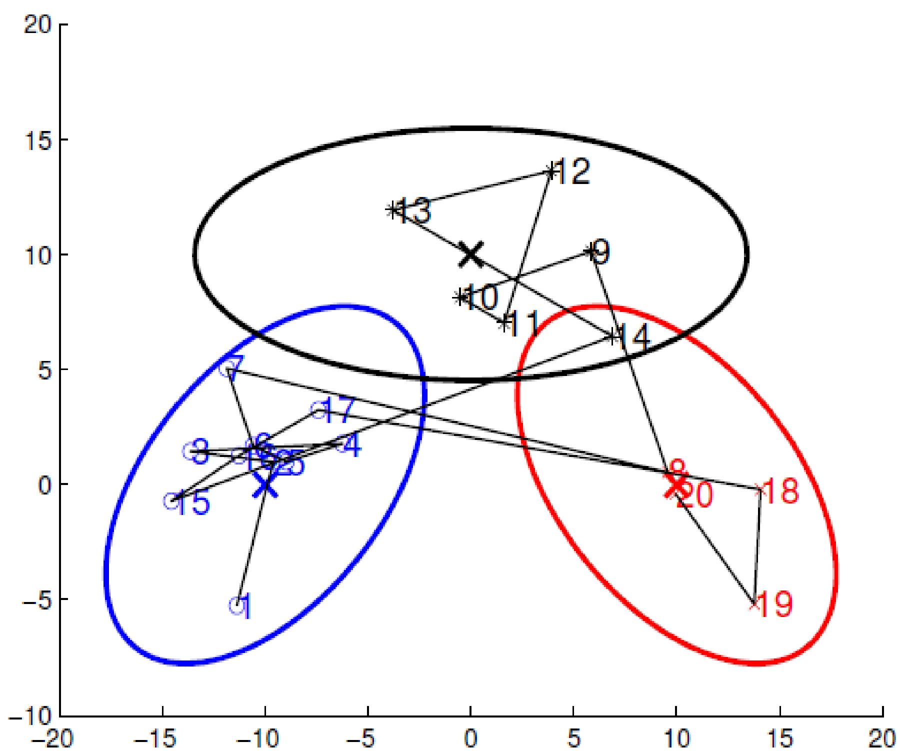
$$P(x_t \mid y_{1:t-1}) = \sum_{x_{t-1}} P(x_t \mid x_{t-1}) P(x_{t-1} \mid y_{1:t-1})$$

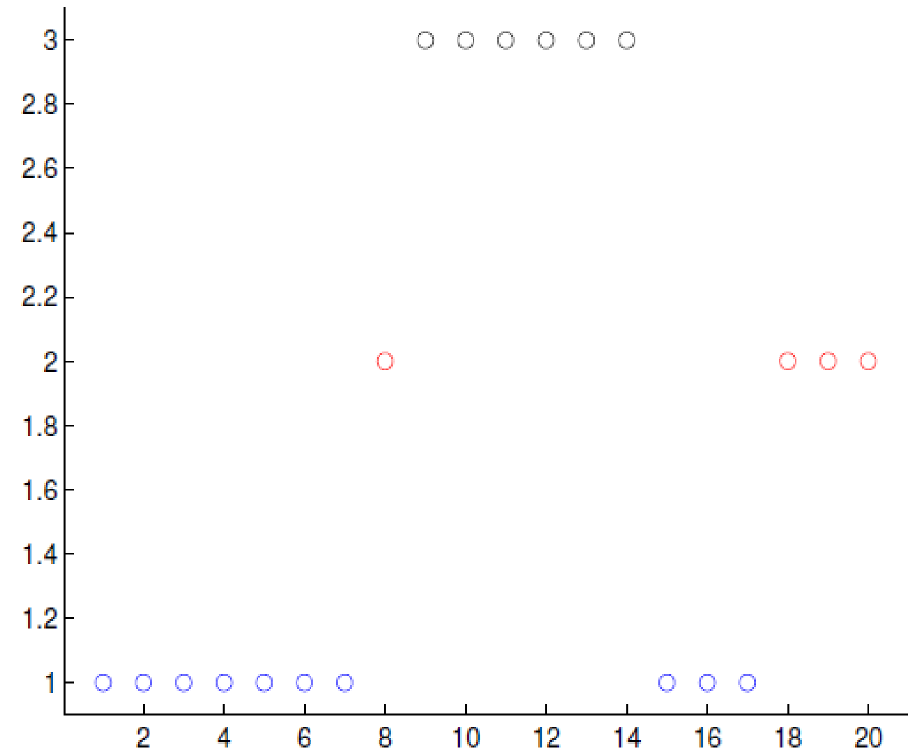
$$P(x_t \mid y_{1:t}) = P(x_t \mid y_t\, y_{1:t-1})$$

$$= \frac{P(y_t \mid x_t, y_{1:t-1}) P(x_t \mid y_{1:t-1})}{P(y_t \mid y_{1:t-1})}$$

$$= \frac{P(y_t \mid x_t) P(x_t \mid y_{1:t-1})}{P(y_t \mid y_{1:t-1})}$$

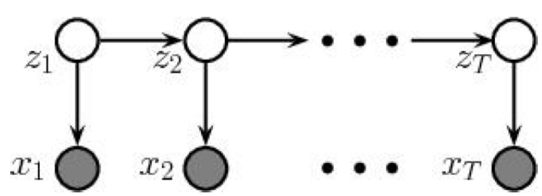
HMMs



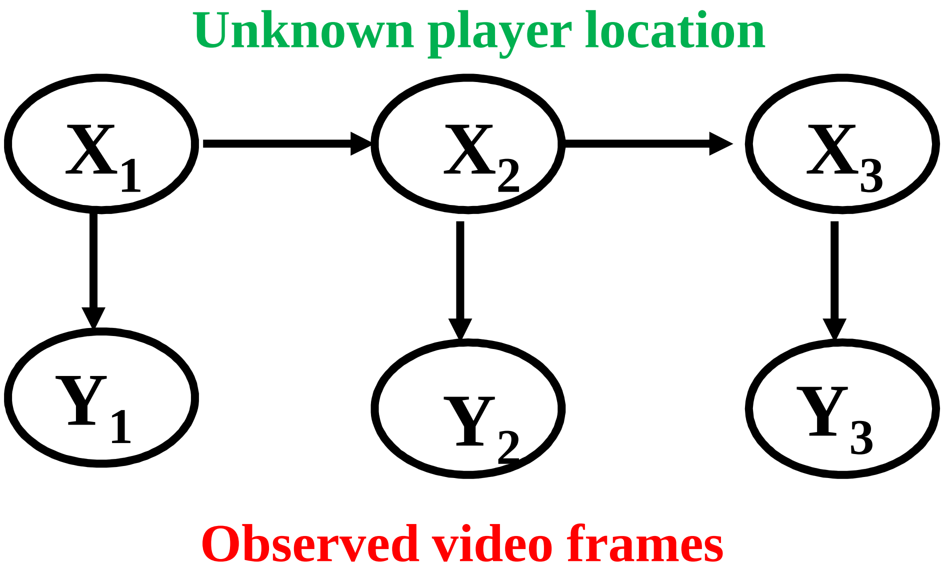


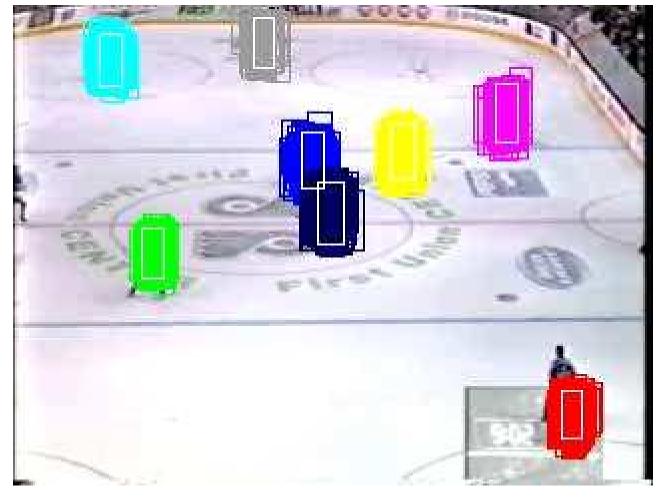
$$p(\mathbf{x}_t | z_t = k, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}_t | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

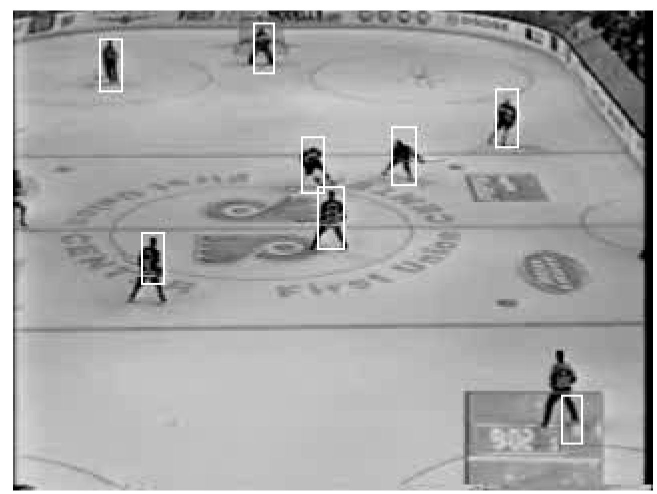
$$p(z_t = k | z_{t-1} = j, \boldsymbol{\theta}) = A(j, k)$$



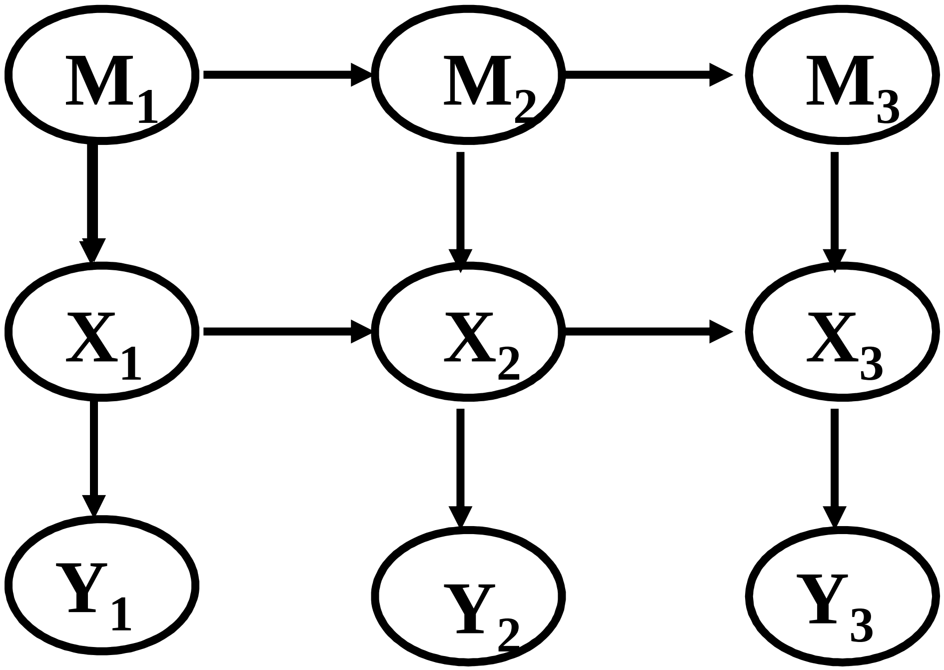
Dynamic Bayesian networks







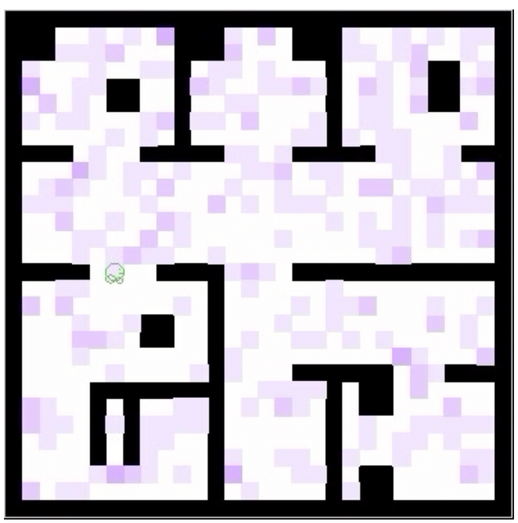
Dynamic Bayesian networks

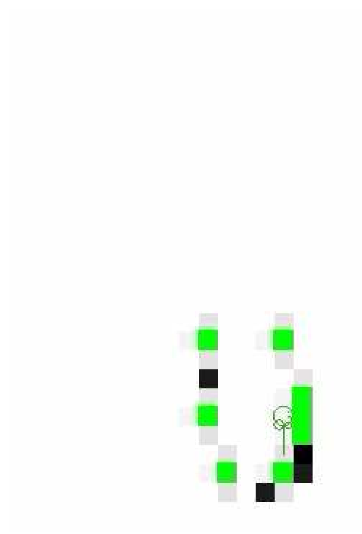


Unknown map

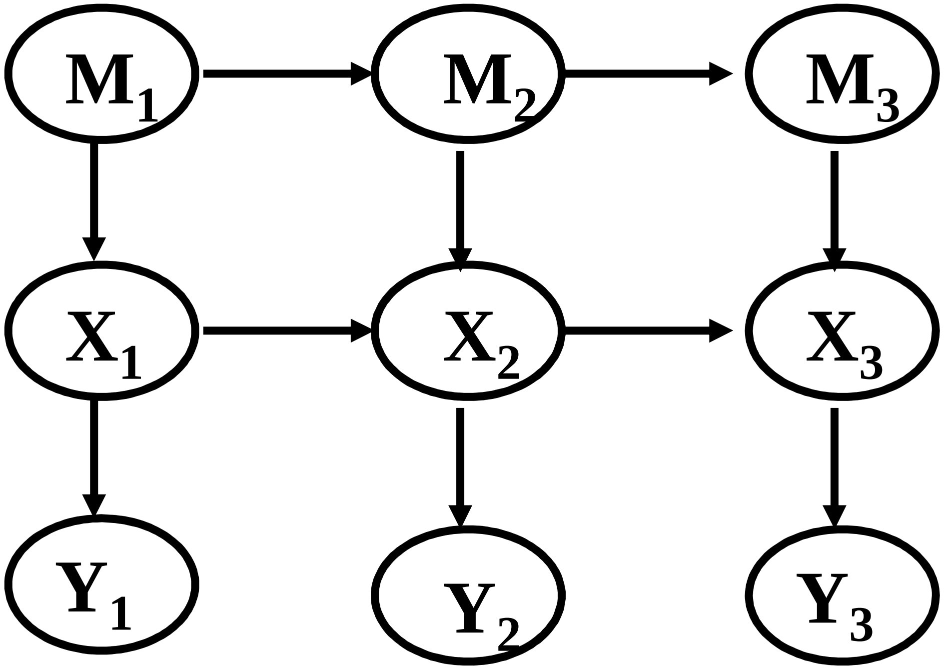
Unknown robot location

Sensor readings





Dynamic Bayesian networks



Unknown internal discrete state

Unknown continuous signals

Sensor readings

Sequential problems

$$p\left(x_0\right)$$

$$p\left(x_t \middle| x_{t-1}\right) \quad \text{for } t \geq 1$$

$$p\left(y_t \middle| x_t\right) \quad \text{for } t \geq 1$$

- **Filtering**: Compute $p(x_t|y_{1:t})$.
- **Prediction**: Compute $p(x_{t+\tau}|y_{1:t})$.
- **Smoothing**: Compute $p(x_{t-\tau}|y_{1:t})$.

Sequential problems

$$\textit{Prediction:}\ p\left(x_t \middle| y_{1:t-1}\right) = \int p\left(x_t \middle| x_{t-1}\right) p\left(x_{t-1} \middle| y_{1:t-1}\right) dx_{t-1}$$

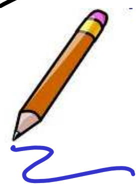
$$\textit{Updating:}\ p\left(x_t \middle| y_{1:t}\right) = \frac{p\left(y_t \middle| x_t\right) p\left(x_t \middle| y_{1:t-1}\right)}{\int p\left(y_t \middle| x_t\right) p\left(x_t \middle| y_{1:t-1}\right) dx_t}$$

Kalman Filtering

We consider the following dynamic state space model:

$$
\begin{aligned}
x_t &= Ax_{t-1} + Bw_t + Fu_t \\
y_t &= Cx_t + Dv_t + Gu_t,
\end{aligned}
$$

where $y_t \in \mathbb{R}^{n_y}$ denotes the observations, $x_t \in \mathbb{R}^{n_x}$ denotes the unknown Gaussian states, $u_t \in \mathcal{U}$ is a known control signal, the parameters (A, B, C, D, F, G) are known matrices and the initial mean and covariance of x_t are μ_0, Σ_0. The noise processes are $i.i.d$ Gaussian: $w_t \sim \mathcal{N}(0, I)$ and $v_t \sim \mathcal{N}(0, I)$. Our model implies the continuous densities



Kalman Filtering

$$\mu_{t|t-1} \triangleq \mathbb{E}\left(x_t \middle| y_{1:t-1}\right)$$

$$\mu_{t|t} \triangleq \mathbb{E}\left(x_t \middle| y_{1:t}\right)$$

$$y_{t|t-1} \triangleq \mathbb{E}\left(y_t \middle| y_{1:t-1}\right)$$

$$\Sigma_{t|t-1} \triangleq cov\left(x_t \middle| y_{1:t-1}\right)$$

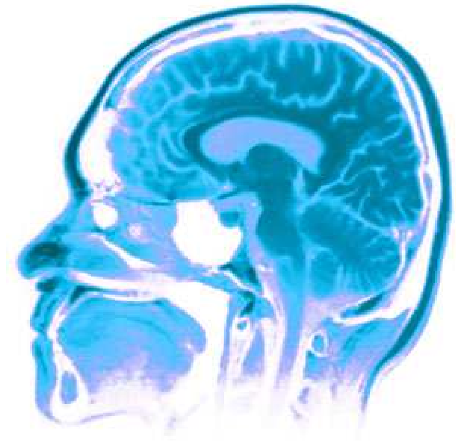
$$\Sigma_{t|t} \triangleq cov\left(x_t \middle| y_{1:t}\right)$$

$$S_t \triangleq cov\left(y_t \middle| y_{1:t-1}\right)$$

$$p\left(y_t \middle| y_{1:t-1}\right) = \mathcal{N}\left(y_t; y_{t|t-1}, S_t\right)$$

$$\begin{aligned}
\Sigma_{t|t-1} &= A\Sigma_{t-1|t-1}A' + BB' \\
S_t &= C\Sigma_{t|t-1}C' + DD' \\
y_{t|t-1} &= C\mu_{t|t-1} + Gu_t \\
\mu_{t|t} &= \mu_{t|t-1} + \Sigma_{t|t-1}C^{\mathrm{T}}S_t^{-1}(y_t - y_{t|t-1}) \\
\Sigma_{t|t} &= \Sigma_{t|t-1} - \Sigma_{t|t-1}C'S_t^{-1}C\Sigma_{t|t-1}
\end{aligned}$$

Next class



More on Directed Graphical Models

Nando de Freitas
2011
KPM Book Sections: *6.*