Adapted from AIMA slides

Bayesian networks

Peter Antal
antal@mit.bme.hu

Outline

- Reminder: inference in the joint distribution
- Reminder: properties of independencies, independence models
- Independencies in representation & inference
  - Example: Naive Bayesian networks
  - Example: Hidden Markov models
- Bayesian networks
- A construction/learning method

Inference by enumeration

Every question about a domain can be answered by the joint distribution.

Typically, we are interested in the posterior joint distribution of the query variables **Y** given specific values **e** for the evidence variables **E**

Let the hidden variables be **H** = **X** - **Y** - **E**

Then the required summation of joint entries is done by summing out the hidden variables:

$\mathbf{P}(\mathbf{Y} \mid \mathbf{E} = \mathbf{e}) = \alpha \mathbf{P}(\mathbf{Y}, \mathbf{E} = \mathbf{e}) = \alpha \Sigma_h \mathbf{P}(\mathbf{Y}, \mathbf{E} = \mathbf{e}, \mathbf{H} = \mathbf{h})$

- The terms in the summation are joint entries because **Y**, **E** and **H** together exhaust the set of random variables
- Obvious problems:
  1. Worst-case time complexity $O(d^n)$ where d is the largest arity
  2. Space complexity $O(d^n)$ to store the joint distribution
  3. How to find the numbers for $O(d^n)$ entries?

Properties of independence

a Symmetry: The observational probabilistic conditional independence is symmetric.

$$I_p(\mathbf{X};\mathbf{Y}|\mathbf{Z}) \text{ iff } I_p(\mathbf{Y};\mathbf{X}|\mathbf{Z})$$

b Decomposition: Any part of an irrelevant information is irrelevant.

$$I_p(\mathbf{X};\mathbf{Y}\cup\mathbf{W}|\mathbf{Z}) \Rightarrow I_p(\mathbf{X};\mathbf{Y}|\mathbf{Z}) \text{ and } I_p(\mathbf{X};\mathbf{W}|\mathbf{Z})$$

c Weak union: Irrelevant information remains irrelevant after learning (other) irrelevant information.

$$I_p(\mathbf{X};\mathbf{Y}\cup\mathbf{W}|\mathbf{Z}) \Rightarrow I_p(\mathbf{X};\mathbf{Y}|\mathbf{Z}\cup\mathbf{W})$$

d Contraction: Irrelevant information remains irrelevant after forgetting (other) irrelevant information.

$$I_p(\mathbf{X};\mathbf{Y}|\mathbf{Z}) \text{ and } I_p(\mathbf{X};\mathbf{W}|\mathbf{Z}\cup\mathbf{Y}) \Rightarrow I_p(\mathbf{X};\mathbf{Y}\cup\mathbf{W}|\mathbf{Z})$$

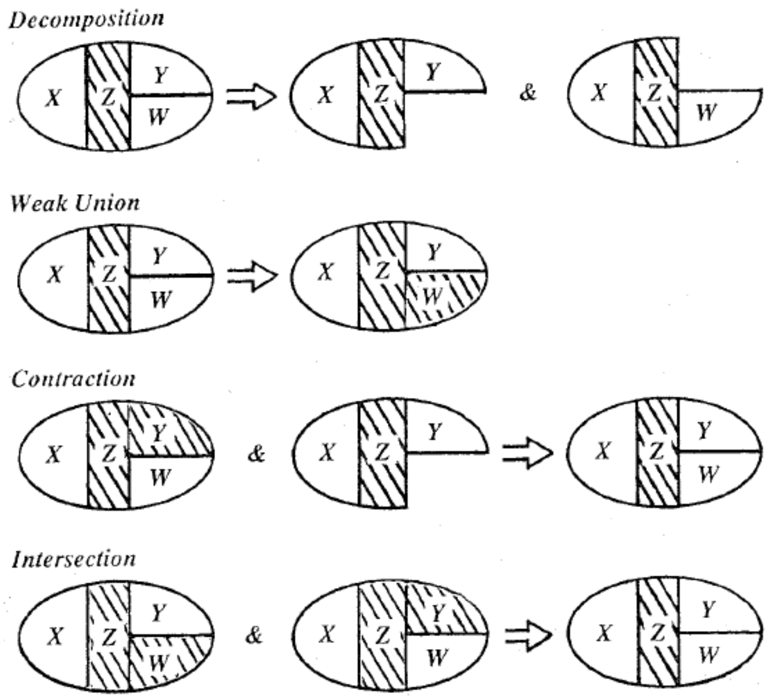
e Intersection: Symmetric irrelevance implies joint irrelevance if there are no dependencies.

$$I_p(\mathbf{X};\mathbf{Y}|\mathbf{Z}\cup\mathbf{W}) \text{ and } I_p(\mathbf{X};\mathbf{W}|\mathbf{Z}\cup\mathbf{Y}) \Rightarrow I_p(\mathbf{X};\mathbf{Y}\cup\mathbf{W}|\mathbf{Z})$$

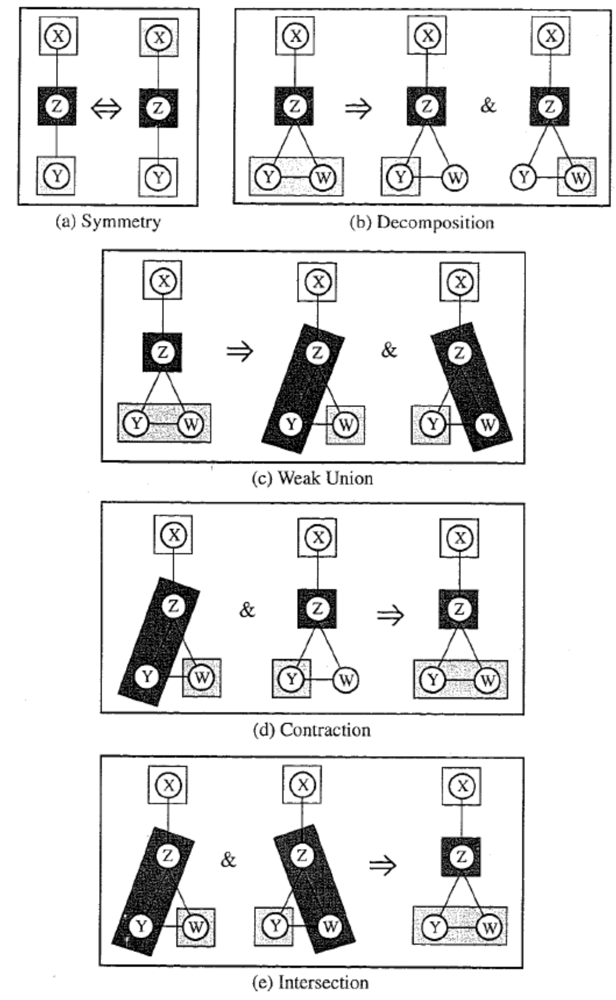
Semi-graphoids, graphoids

Semi-graphoids (SG): Symmetry, Decomposition, Weak Union, Contraction (holds in all probability distribution). SG is sound, but incomplete inference.

Graphoids: Semi-graphoids+Intersection (holds only in strictly positive distribution)



J.Pearl: Probabilistic Reasoning in intelligent systems, 1998



The independence model of a distribution

The independence map (model) M of a distribution P is the set of the valid independence triplets:

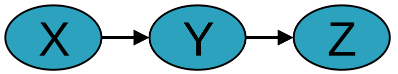
$M_P=\{I_{P,1}(X_1;Y_1|Z_1),\ldots, I_{P,K}(X_K;Y_K|Z_K)\}$

If P(X,Y,Z) is a Markov chain, then

$M_P=\{D(X;Y), D(Y;Z), I(X;Z|Y)\}$

Normally/almost always: $D(X;Z)$

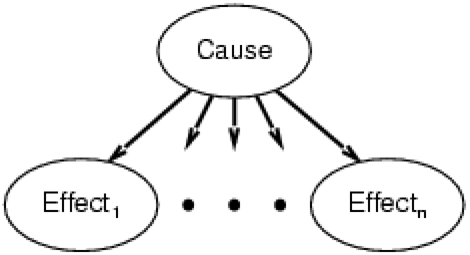
Exceptionally: $I(X;Z)$



Naive Bayesian network

Assumptions:

1, Two types of nodes: a cause and effects.



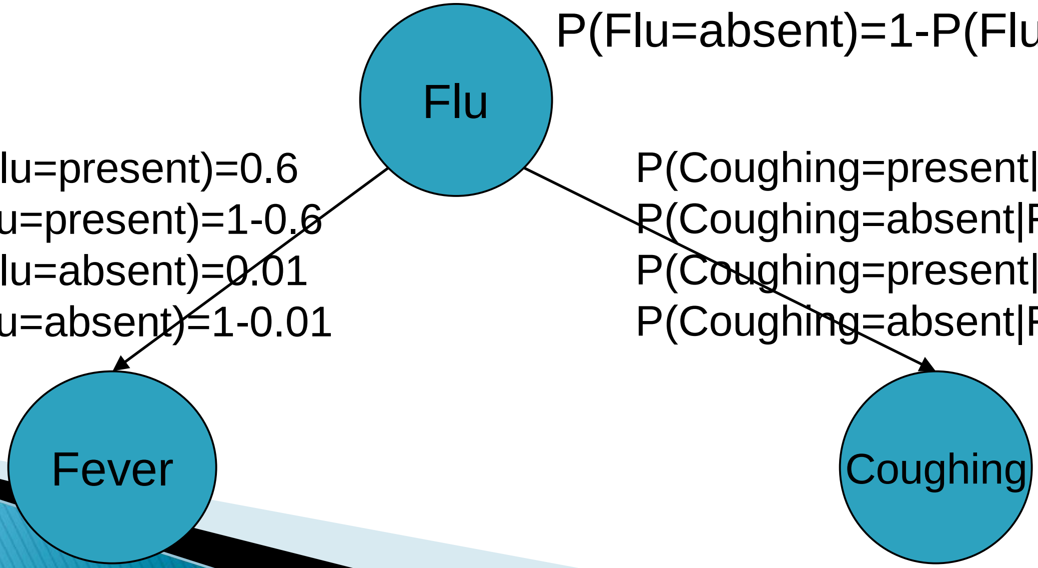
2, Effects are conditionally independent of each other given their cause.

Variables (nodes)

Flu: present/absent
FeverAbove38C: present/absent
Coughing: present/absent

Model

P(Flu=present)=0.001
P(Flu=absent)=1-P(Flu=present)



P(Fever=present|Flu=present)=0.6
P(Fever=absent|Flu=present)=1-0.6
P(Fever=present|Flu=absent)=0.01
P(Fever=absent|Flu=absent)=1-0.01

P(Coughing=present|Flu=present)=0.3
P(Coughing=absent|Flu=present)=1-0.7
P(Coughing=present|Flu=absent)=0.02
P(Coughing=absent|Flu=absent)=1-0.02

Naive Bayesian network (NBN)

Decomposition of the joint:

$P(Y,X_1,..,X_n) \quad = P(Y)\prod_i P(X_i,|Y, X_1,..,X_{i-1})$ //by the chain rule

$\quad = P(Y)\prod_i P(X_i,|Y)$ // by the N-BN assumption

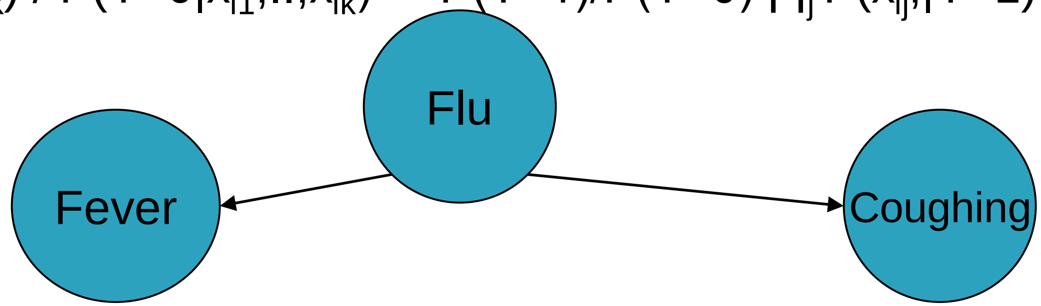
2n+1 parameteres!

Diagnostic inference:

$P(Y|x_{i1},..,x_{ik}) \quad = P(Y)\prod_j P(x_{ij},|Y) \,/\, P(x_{i1},..,x_{ik})$

If Y is binary, then the odds

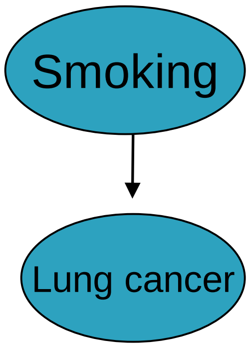
$P(Y=1|x_{i1},..,x_{ik}) \,/\, P(Y=0|x_{i1},..,x_{ik}) \; = P(Y=1)/P(Y=0) \prod_j P(x_{ij},|Y=1) \,/\, P(x_{ij},|Y=0)$



$$p(Flu = present \,|\, Fever = absent, Coughing = present)$$

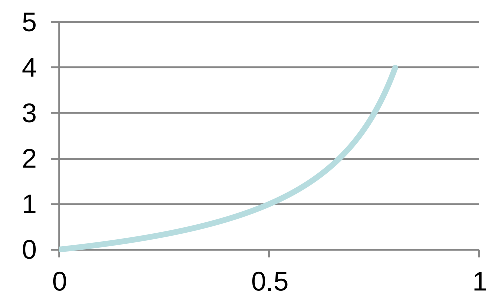
$$\propto p(Flu = present)\, p(Fever = absent \,|\, Flu = present)\, p(Coughing = present \,|\, Flu = present)$$

Conditional probabilities, odds, odds ratios

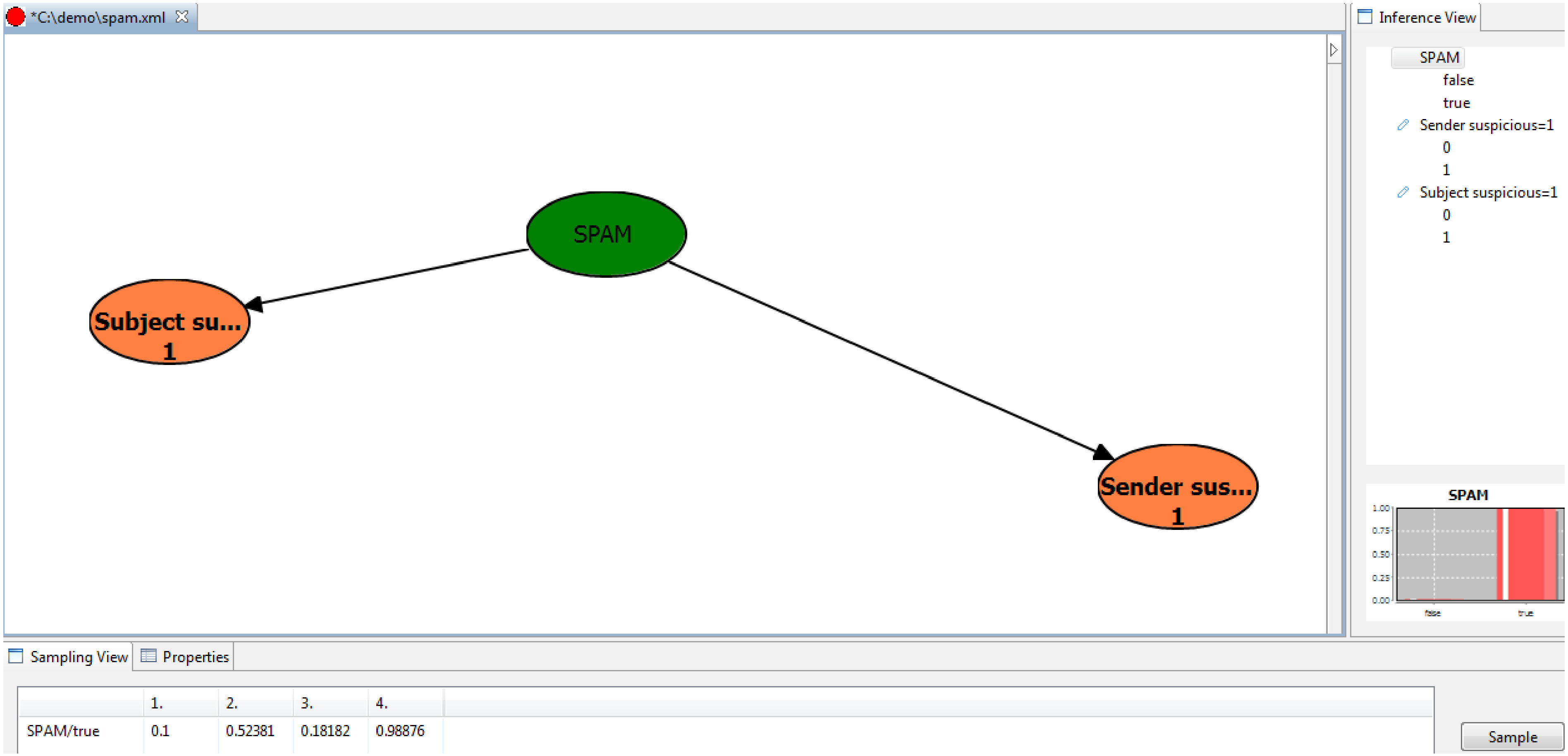


	¬S	S	
¬LC	P(¬S, ¬LC)	P(S, ¬LC)	P(¬LC)
LC	P(¬S, LC)	P(S, LC)	P(LC)
	P(¬S)	P(S)	

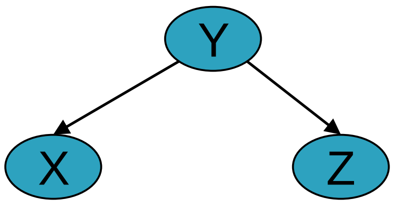
Probability:
P(LC)
Conditional probabilities (e.g., probability of LC given S)**:**
P(LC| ¬S)= ??? P(LC| S)= ??? P(LC)
Odds:
$[0,1] \rightarrow [0,\infty]$: Odds(p)=p/(1-p)
O(LC| ¬S)= ??? O(LC| S)
Odds Ratio (OR) Independent of prevalence!
OR(LC,S)=O(LC| S)/O(LC| ¬S)



Example: Construct a spam filter



The independence map of a N-BN



If P(Y,X,Z) is a naive Bayesian network, then
M_P={D(X;Y), D(Y;Z), I(X;Z|Y)}
Normally/almost always: D(X;Z)
Exceptionally: I(X;Z)

Learning of N-BNs?

- Bayesian learning?
- Identification of
  - parameters,
  - structure?

Hidden Markov Models (HMMs)

The world changes; we need to track and predict it

Diabetes management vs vehicle diagnosis

Basic idea: copy state and evidence variables for each time step

$\mathbf{X}_t$ = set of unobservable state variables at time t
e.g., $BloodSugar_t$, $StomachContents_t$, etc.

$\mathbf{E}_t$ = set of observable evidence variables at time t
e.g., $MeasuredBloodSugar_t$, $PulseRate_t$, $FoodEaten_t$

This assumes **discrete time**; step size depends on problem

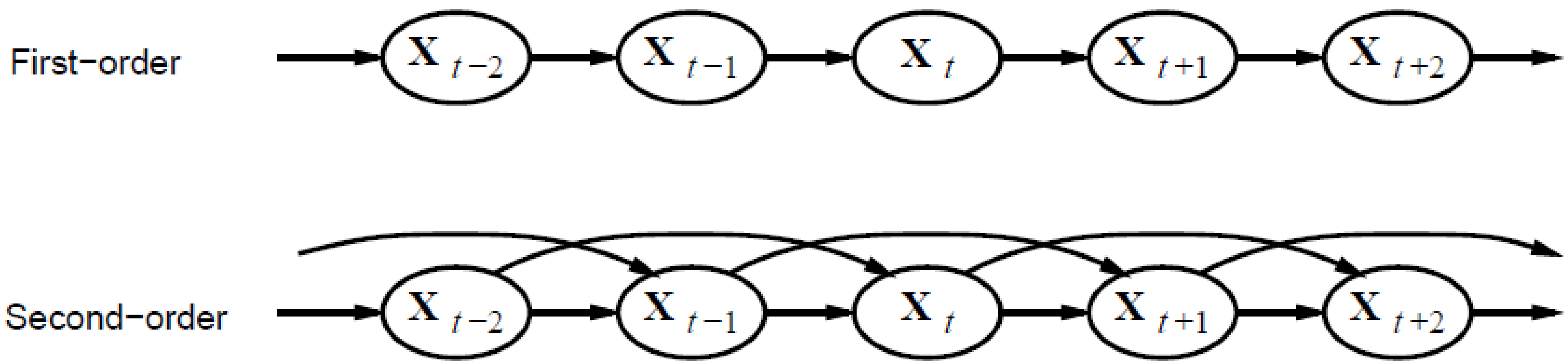
Notation: $\mathbf{X}_{a:b} = \mathbf{X}_a, \mathbf{X}_{a+1}, \ldots, \mathbf{X}_{b-1}, \mathbf{X}_b$

Markov chains

Markov assumption: $\mathbf{X}_t$ depends on **bounded** subset of $\mathbf{X}_{0:t-1}$

First-order Markov process: $\mathbf{P}(\mathbf{X}_t|\mathbf{X}_{0:t-1}) = \mathbf{P}(\mathbf{X}_t|\mathbf{X}_{t-1})$

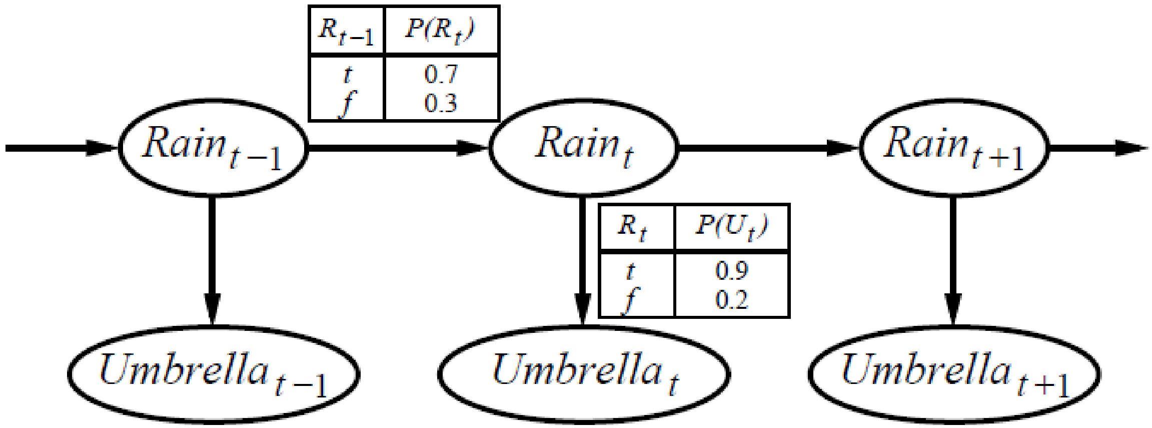
Second-order Markov process: $\mathbf{P}(\mathbf{X}_t|\mathbf{X}_{0:t-1}) = \mathbf{P}(\mathbf{X}_t|\mathbf{X}_{t-2}, \mathbf{X}_{t-1})$



Sensor Markov assumption: $\mathbf{P}(\mathbf{E}_t|\mathbf{X}_{0:t}, \mathbf{E}_{0:t-1}) = \mathbf{P}(\mathbf{E}_t|\mathbf{X}_t)$

Stationary process: transition model $\mathbf{P}(\mathbf{X}_t|\mathbf{X}_{t-1})$ and sensor model $\mathbf{P}(\mathbf{E}_t|\mathbf{X}_t)$ fixed for all t

Example



First-order Markov assumption not exactly true in real world!

Possible fixes:

1. **Increase order** of Markov process
2. **Augment state**, e.g., add $Temp_t$, $Pressure_t$

Example: robot motion.

Augment position and velocity with $Battery_t$

Inference in HMMs

Filtering: $\mathbf{P}(\mathbf{X}_t|\mathbf{e}_{1:t})$

belief state—input to the decision process of a rational agent

Prediction: $\mathbf{P}(\mathbf{X}_{t+k}|\mathbf{e}_{1:t})$ for $k > 0$

evaluation of possible action sequences;
like filtering without the evidence

Smoothing: $\mathbf{P}(\mathbf{X}_k|\mathbf{e}_{1:t})$ for $0 \leq k < t$

better estimate of past states, essential for learning

Most likely explanation: $\arg\max_{\mathbf{x}_{1:t}} P(\mathbf{x}_{1:t}|\mathbf{e}_{1:t})$

speech recognition, decoding with a noisy channel

Filtering

Aim: devise a **recursive** state estimation algorithm:

$$\mathbf{P}(\mathbf{X}_{t+1}|\mathbf{e}_{1:t+1}) = f(\mathbf{e}_{t+1}, \mathbf{P}(\mathbf{X}_t|\mathbf{e}_{1:t}))$$

$$\begin{aligned}\mathbf{P}(\mathbf{X}_{t+1}|\mathbf{e}_{1:t+1}) &= \mathbf{P}(\mathbf{X}_{t+1}|\mathbf{e}_{1:t}, \mathbf{e}_{t+1}) \\ &= \alpha\mathbf{P}(\mathbf{e}_{t+1}|\mathbf{X}_{t+1}, \mathbf{e}_{1:t})\mathbf{P}(\mathbf{X}_{t+1}|\mathbf{e}_{1:t}) \\ &= \alpha\mathbf{P}(\mathbf{e}_{t+1}|\mathbf{X}_{t+1})\mathbf{P}(\mathbf{X}_{t+1}|\mathbf{e}_{1:t})\end{aligned}$$

I.e., prediction + estimation. Prediction by summing out $\mathbf{X}_t$:

$$\begin{aligned}\mathbf{P}(\mathbf{X}_{t+1}|\mathbf{e}_{1:t+1}) &= \alpha\mathbf{P}(\mathbf{e}_{t+1}|\mathbf{X}_{t+1})\Sigma_{\mathbf{x}_t}\mathbf{P}(\mathbf{X}_{t+1}|\mathbf{x}_t, \mathbf{e}_{1:t})P(\mathbf{x}_t|\mathbf{e}_{1:t}) \\ &= \alpha\mathbf{P}(\mathbf{e}_{t+1}|\mathbf{X}_{t+1})\Sigma_{\mathbf{x}_t}\mathbf{P}(\mathbf{X}_{t+1}|\mathbf{x}_t)P(\mathbf{x}_t|\mathbf{e}_{1:t})\end{aligned}$$

$\mathbf{f}_{1:t+1} = \text{FORWARD}(\mathbf{f}_{1:t}, \mathbf{e}_{t+1})$ where $\mathbf{f}_{1:t} = \mathbf{P}(\mathbf{X}_t|\mathbf{e}_{1:t})$

Time and space **constant** (independent of t)

Learning HMMs

- Parameter learning
- Structure learning
  - HMMs are equivalent with stochastic finite state automatons

Bayesian networks

- A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions

- Syntax:
  - a set of nodes, one per variable
  -
  - a directed, acyclic graph (link ≈ "directly influences")
  - a conditional distribution for each node given its parents:

$$\mathbf{P}(X_i \mid Parents(X_i))$$

- In the simplest case, conditional distribution represented as a conditional probability table (CPT) giving the distribution over X_i for each combination of parent values

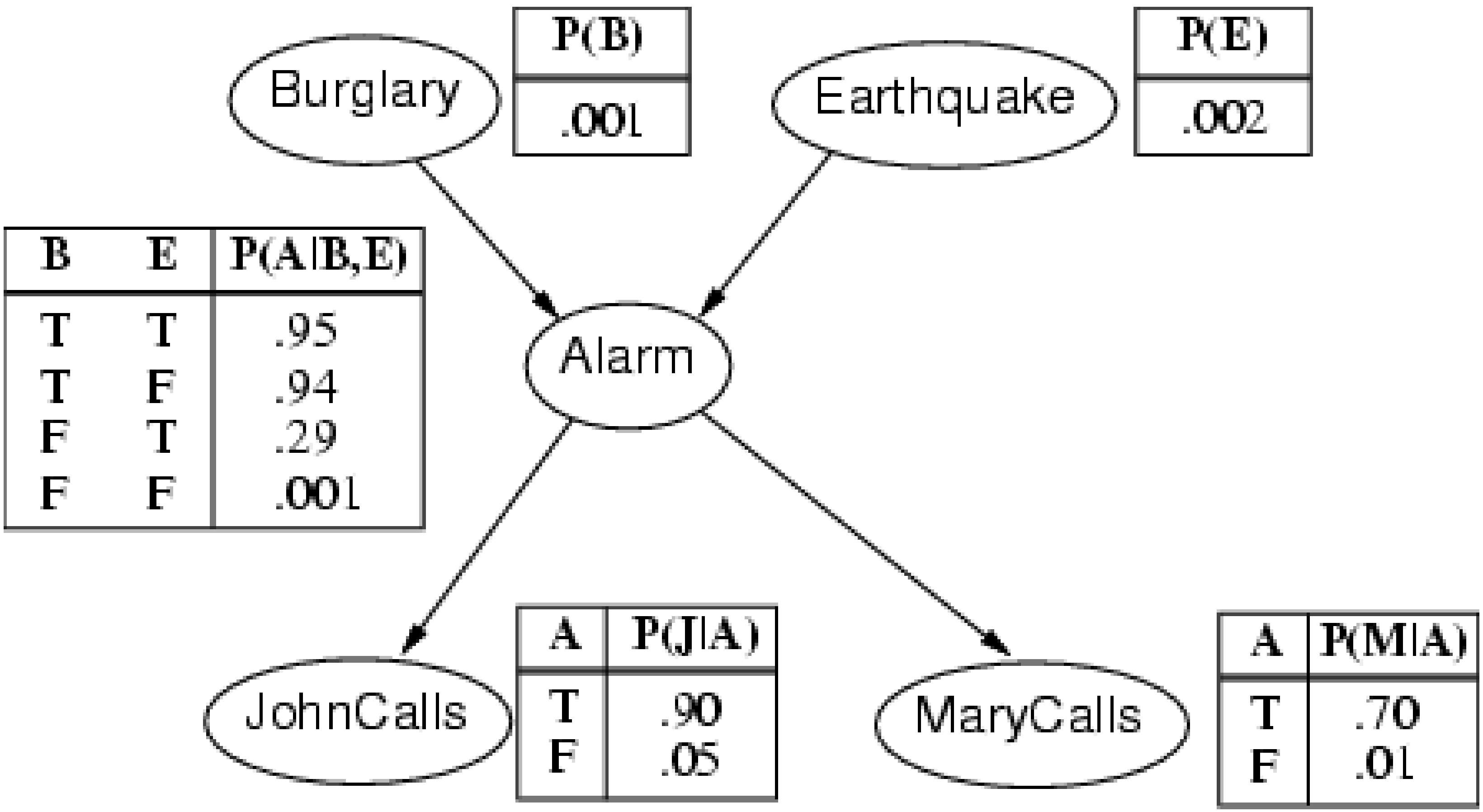
Example

- I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes it's set off by minor earthquakes. Is there a burglar?

- Variables: *Burglary*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*

- Network topology reflects "causal" knowledge:
  - A burglar can set the alarm off
  - An earthquake can set the alarm off
  - The alarm can cause Mary to call
  - The alarm can cause John to call

Example contd.



P(B)
.001

P(E)
.002

B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J|A)
T	.90
F	.05

A	P(M|A)
T	.70
F	.01

Learning Bayesian networks

- 1. Choose an ordering of variables $X_1, \ldots, X_n$
- 2. For $i = 1$ to n
  - add X_i to the network
  - select parents from $X_1, \ldots, X_{i-1}$ such that

$$\mathbf{P}(X_i \mid Parents(X_i)) = \mathbf{P}(X_i \mid X_1, \ldots X_{i-1})$$

This choice of parents guarantees:

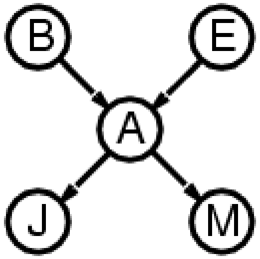
$$\mathbf{P}(X_1, \ldots, X_n) = \pi_{i=1}^{n} \mathbf{P}(X_i \mid X_1, \ldots, X_{i-1}) \quad //\text{(chain rule)}$$

$$= \pi_{i=1}^{n} \mathbf{P}(X_i \mid Parents(X_i)) \quad //\text{(by construction)}$$

Ordered Markov condition (OMC) is satisfied: P obeys the OMC w.r.t. G.

Compactness

- A CPT for Boolean X_i with k Boolean parents has 2^k rows for the combinations of parent values

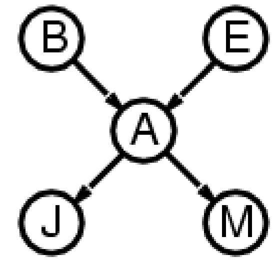


- Each row requires one number p for $X_i = true$ (the number for $X_i = false$ is just $1-p$)
- If each variable has no more than k parents, the complete network requires $O(n \cdot 2^k)$ numbers
- I.e., grows linearly with n, vs. $O(2^n)$ for the full joint distribution
- For burglary net, $1 + 1 + 4 + 2 + 2 = 10$ numbers (vs. $2^5-1 = 31$)

Semantics

The full joint distribution is defined as the product of the local conditional distributions:

$$\mathbf{P}(X_1, \ldots, X_n) = \pi_{i=1}^{n} \mathbf{P}(X_i \mid Parents(X_i))$$



e.g., $\mathbf{P}(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$

$$= \mathbf{P}(j \mid a)\, \mathbf{P}(m \mid a)\, \mathbf{P}(a \mid \neg b, \neg e)\, \mathbf{P}(\neg b)\, \mathbf{P}(\neg e)$$

Inference in BNs

Singly connected networks (or polytrees):

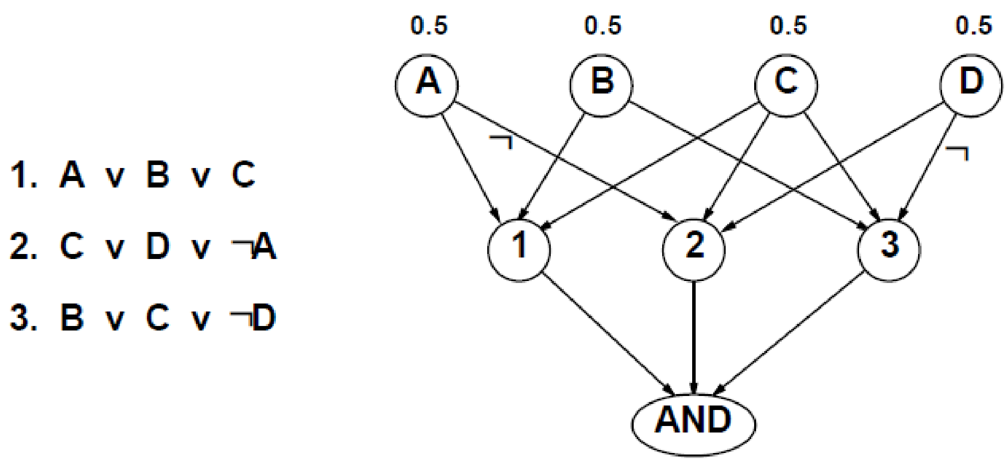
– any two nodes are connected by at most one (undirected) path

– time and space cost of exact inference $O(d^k n)$

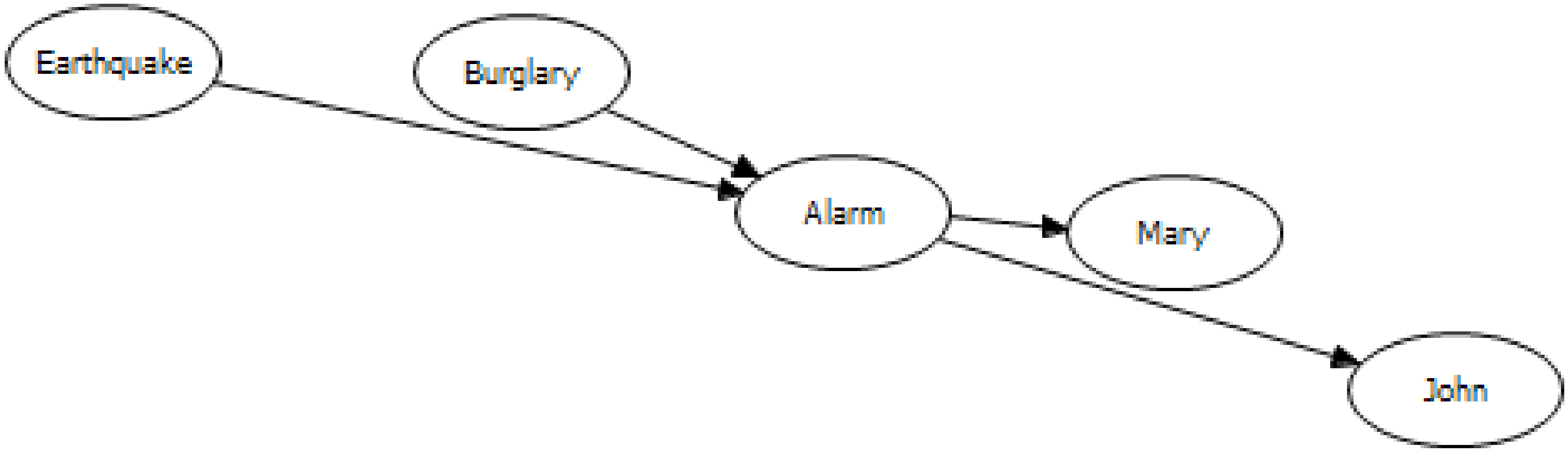
Multiply connected networks:

– can reduce 3SAT to exact inference: 0<p(AND)? $\Rightarrow$ NP-hard

– equivalent to **counting** 3SAT models $\Rightarrow$ #P-complete



Learning with an „acausal" ordering



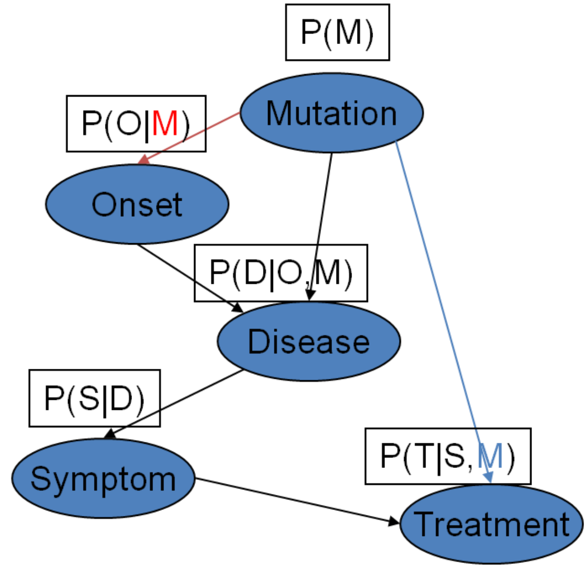
1. Choose an ordering of variables $X_1, \ldots, X_n$
2. For $i = 1$ to n
   add X_i to the network
   select parents from $X_1, \ldots, X_{i-1}$ such that

$$\mathbf{P}(X_i \mid Parents(X_i)) = \mathbf{P}(X_i \mid X_1, \ldots X_{i-1})$$

Bayesian networks: interpretations

3. Concise representation of joint distributions

$$P(M,O,D,S,T) =$$
$$P(M)P(O|M)P(D|O,M)P(S|D)P(T|S,M)$$



1. Causal model

$$M_P=\{I_{P,1}(X_1;Y_1|Z_1),\ldots\}$$

2. Graphical representation of (in)dependencies

Summary

- Conditional independencies, independence model
- Naive Bayesian networks
- Hidden Markov Models
- (General) Bayesian networks
- A method for BN structure learning