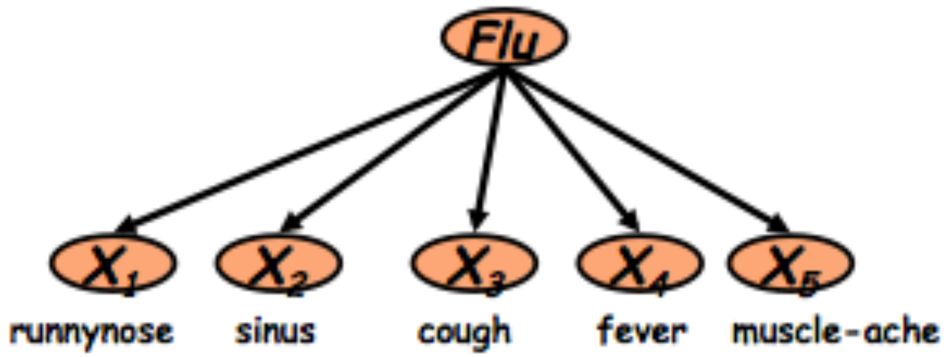
Bayesian Networks

The Naive Bayes Classifier



- **Conditional Independence Assumption:** features are **independent** of each other **given the class**:

$$P(X_1,\ldots,X_5 \mid C) = P(X_1 \mid C) \bullet P(X_2 \mid C) \bullet \cdots \bullet P(X_5 \mid C)$$

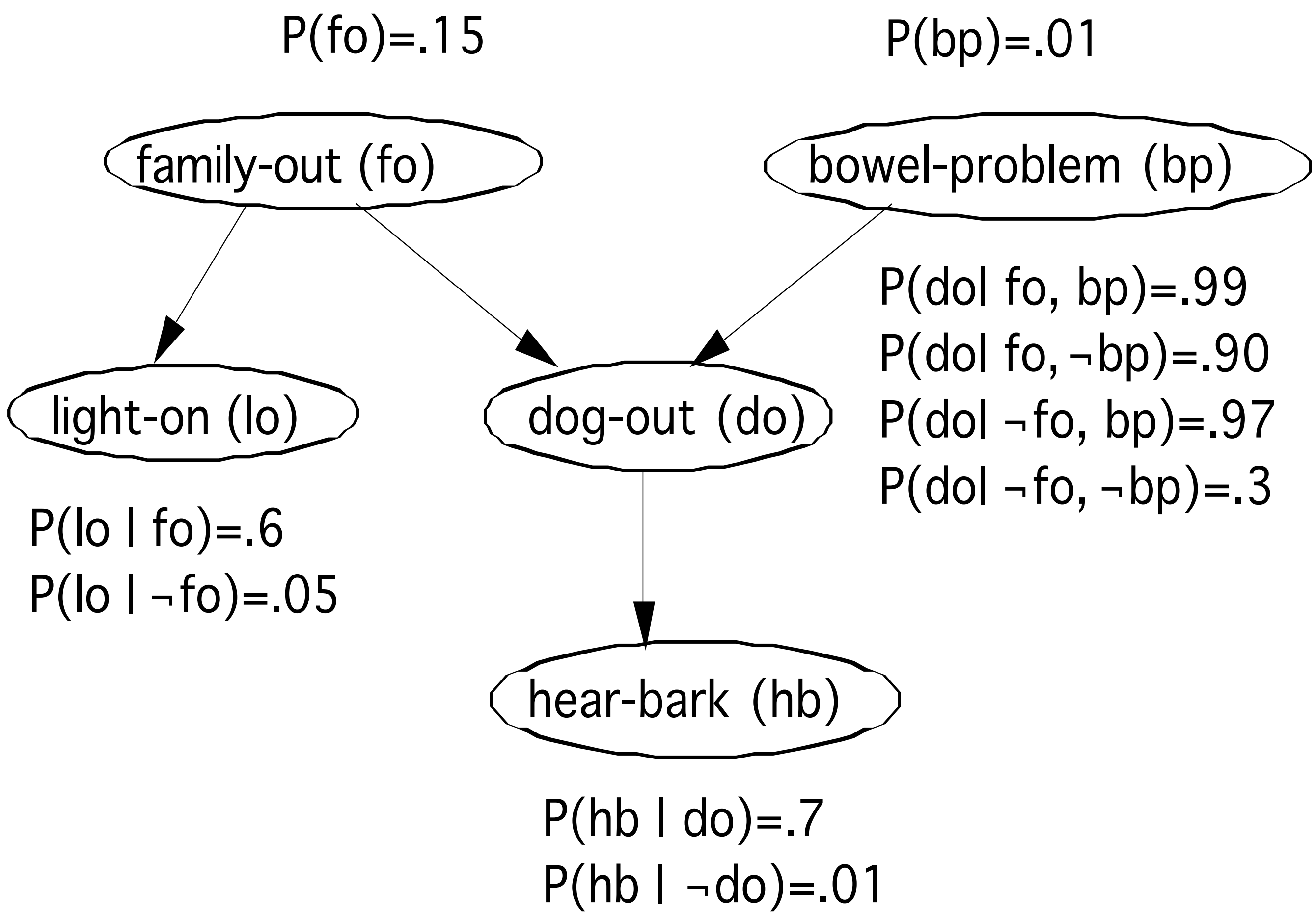
Bayesian Network

- Independence assumptions
  - Seems to be necessary for probabilistic inference to be practical.
- Naïve Bayes Method
  - Makes independence assumptions that are often not true
  - Also called Idiot Bayes Method for this reason.
- Bayesian Network
  - Explicitly models the independence relationships in the data.
  - Use these independence relationships to make probabilistic inferences.
  - Also known as: Belief Net, Bayes Net, Causal Net, …

Bayesian Networks: Definition

- Bayesian networks are directed acyclic graphs (DAGs).
- Nodes in Bayesian networks represent random variables, which are normally assumed to take on discrete values.
- The links of the network represent direct probabilistic influence.
- The structure of the network represents the probabilistic dependence/independence relationships between the random variables represented by the nodes.

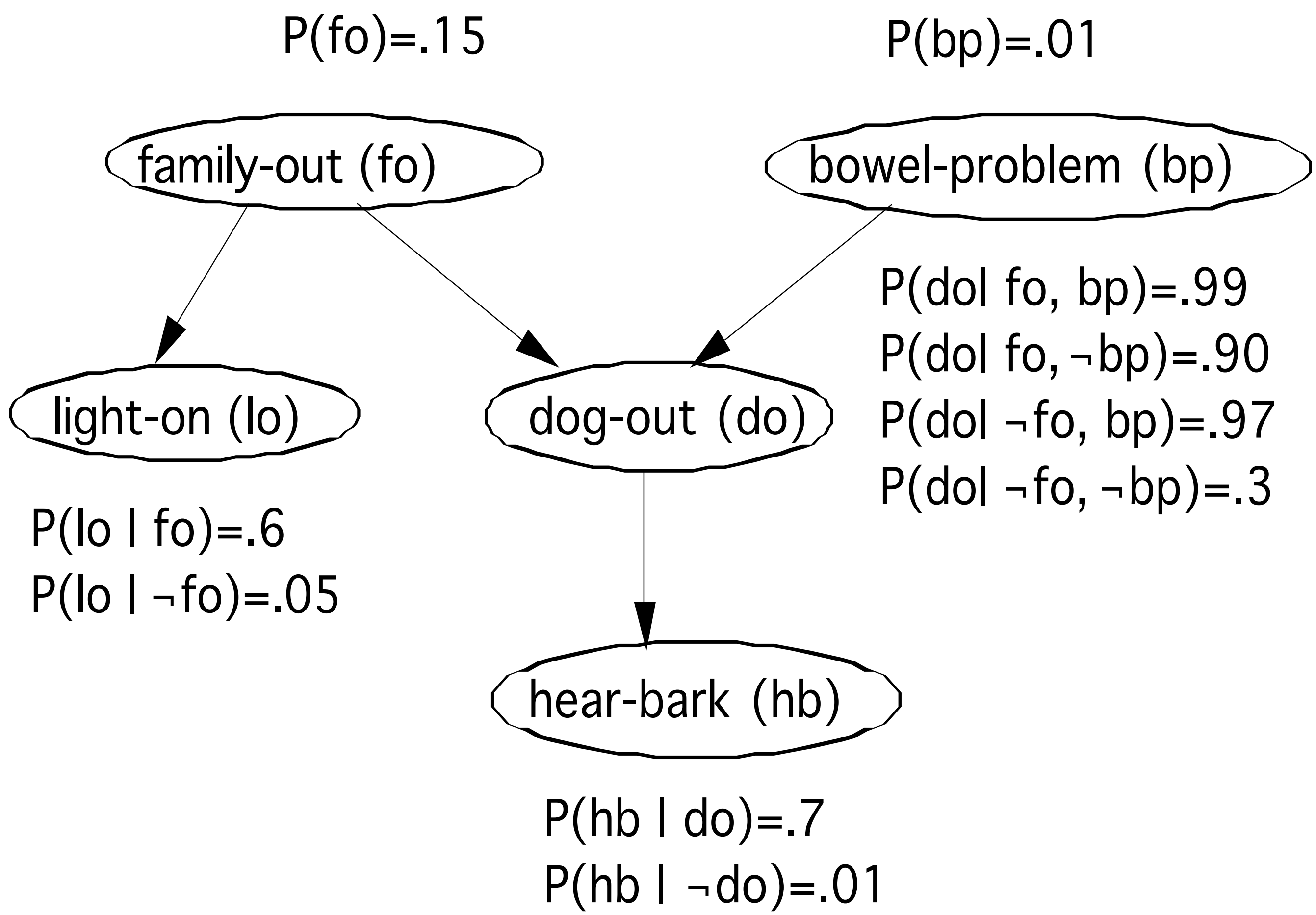
Example


P(fo)=.15
P(bp)=.01
family-out (fo)
bowel-problem (bp)
P(do| fo, bp)=.99
P(do| fo, ¬bp)=.90
P(do| ¬fo, bp)=.97
P(do| ¬fo, ¬bp)=.3
light-on (lo)
dog-out (do)
P(lo | fo)=.6
P(lo | ¬fo)=.05
hear-bark (hb)
P(hb | do)=.7
P(hb | ¬do)=.01

Bayesian Network: Probabilities

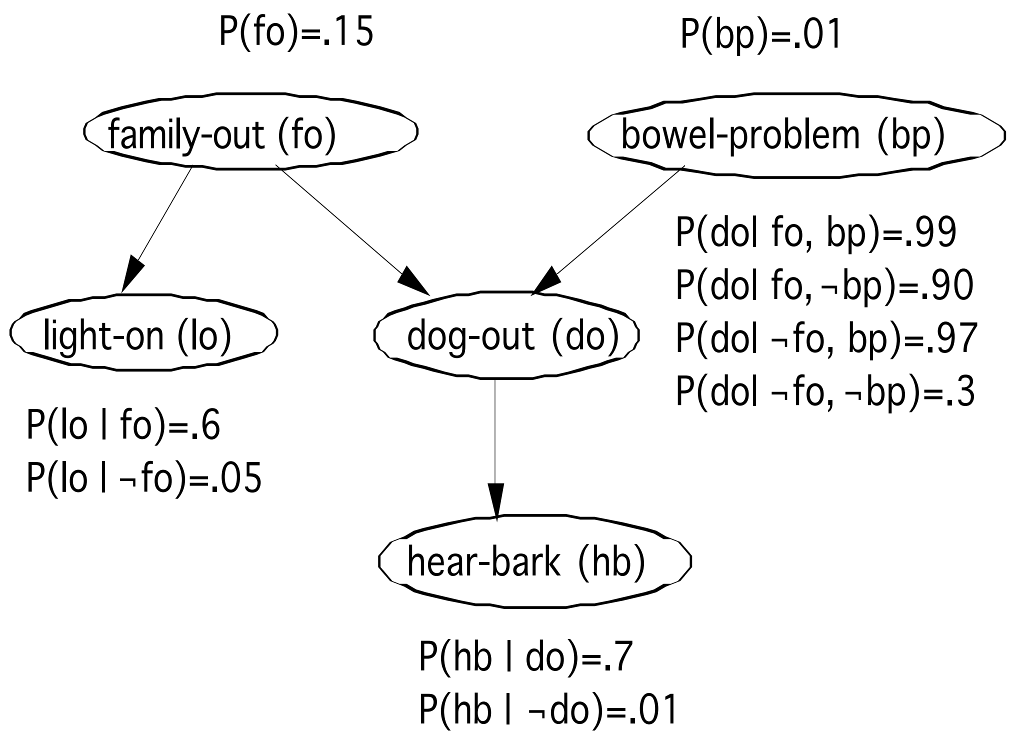
- The nodes and links are quantified with probability distributions.
- The root nodes (those with no ancestors) are assigned prior probability distributions.
- The other nodes are assigned with the conditional probability distribution of the node given its parents.

Example


P(fo)=.15
P(bp)=.01
family-out (fo)
bowel-problem (bp)
light-on (lo)
dog-out (do)
hear-bark (hb)
P(do| fo, bp)=.99
P(do| fo, ¬bp)=.90
P(do| ¬fo, bp)=.97
P(do| ¬fo, ¬bp)=.3
P(lo | fo)=.6
P(lo | ¬fo)=.05
P(hb | do)=.7
P(hb | ¬do)=.01

Probabilistic Inference

Network represents the joint probability over all the variables

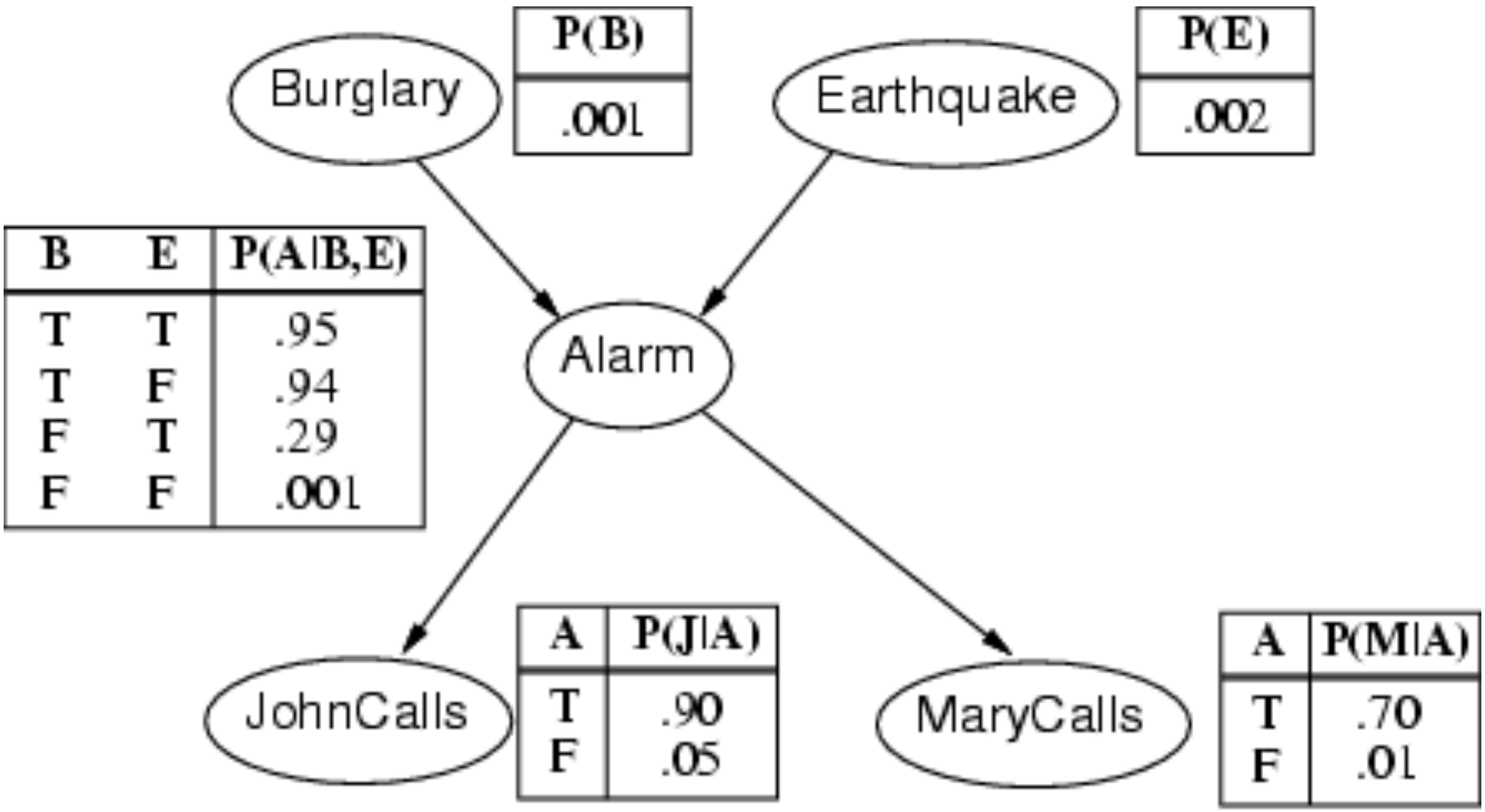


$$
\begin{aligned}
P(hb,do,lo,fo,bp) &= P(hb|do,lo,fo,bp)*P(do,lo,fo,bp) \\
&= P(hb|do,lo,fo,bp)*P(do|lo,fo,bp)*P(lo,fo,bp) \\
&= P(hb|do,lo,fo,bp)*P(do|lo,fo,bp)*P(lo|fo,bp)P(fo,bp) \\
&= P(hb|do,lo,fo,bp)*P(do|lo,fo,bp)*P(lo|fo,bp)P(fo|bp)P(bp)
\end{aligned}
$$

Compactness

- How many numbers are required to build a Bayes Net?
  - For a Boolean variable X_i with k Boolean parents, how many rows in the CPT?
  - If each variable has no more than k parents and there are n nodes in the network, how many numbers required?
- How many numbers required to specify the full joint distribution?

Another Example



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

Constructing Bayesian networks

- 1. Choose an ordering of variables $X_1, \dots, X_n$
- 2. For i = 1 to n
  - add X_i to the network
  - select parents from $X_1, \dots, X_{i-1}$ such that
    $$\mathbf{P}(X_i \mid Parents(X_i)) = \mathbf{P}(X_i \mid X_1, \dots X_{i-1})$$

This choice of parents guarantees:

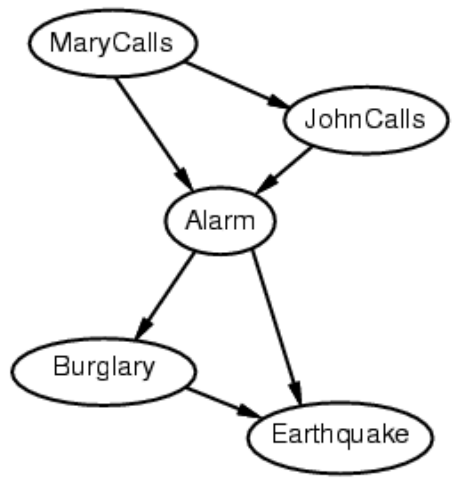
$$\mathbf{P}(X_1, \dots, X_n) = \pi_{i=1}^{n} \mathbf{P}(X_i \mid X_1, \dots, X_{i-1}) \text{ (chain rule)}$$

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

Example

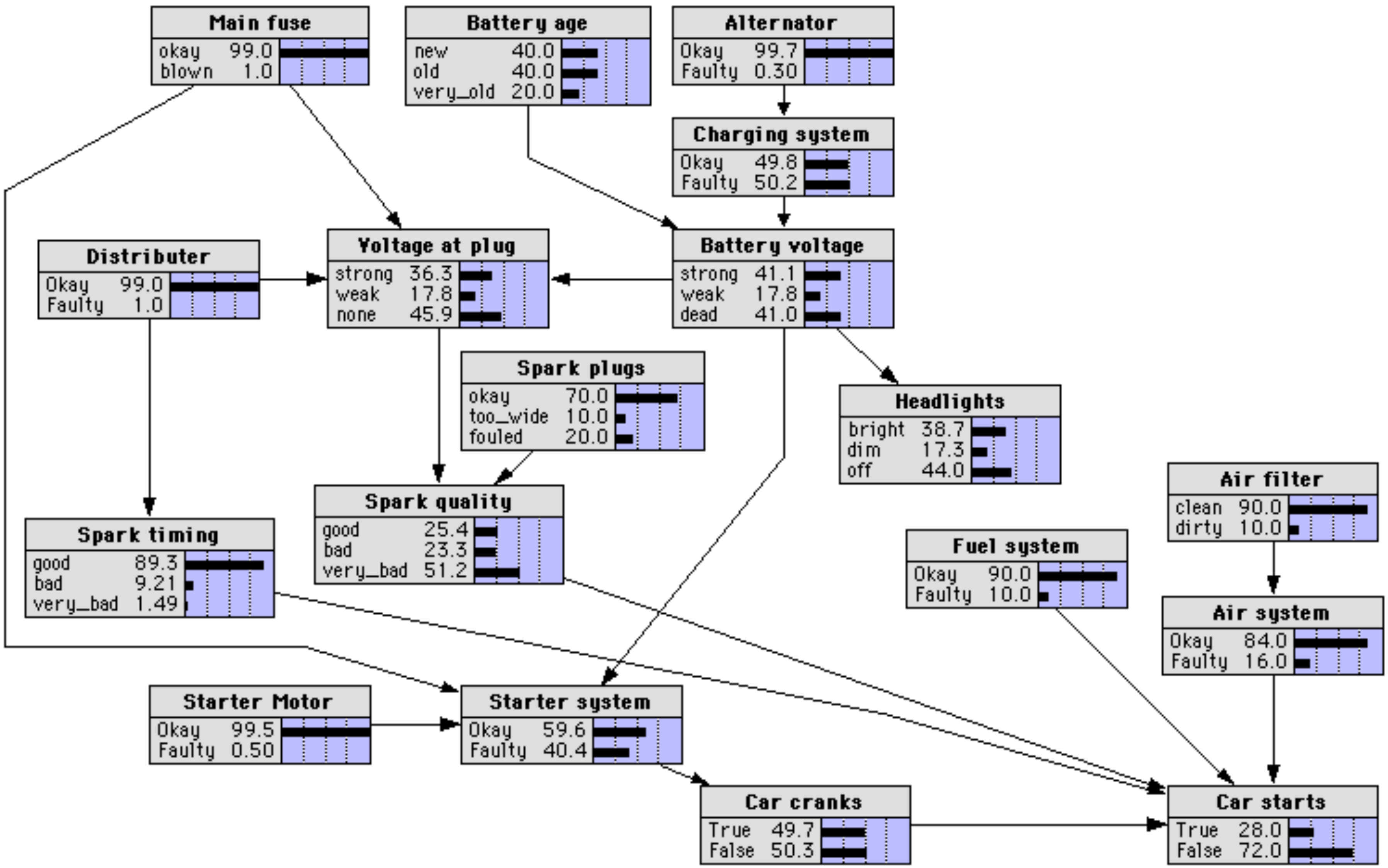
- Suppose we choose the ordering M, J, A, B, E

Example contd.

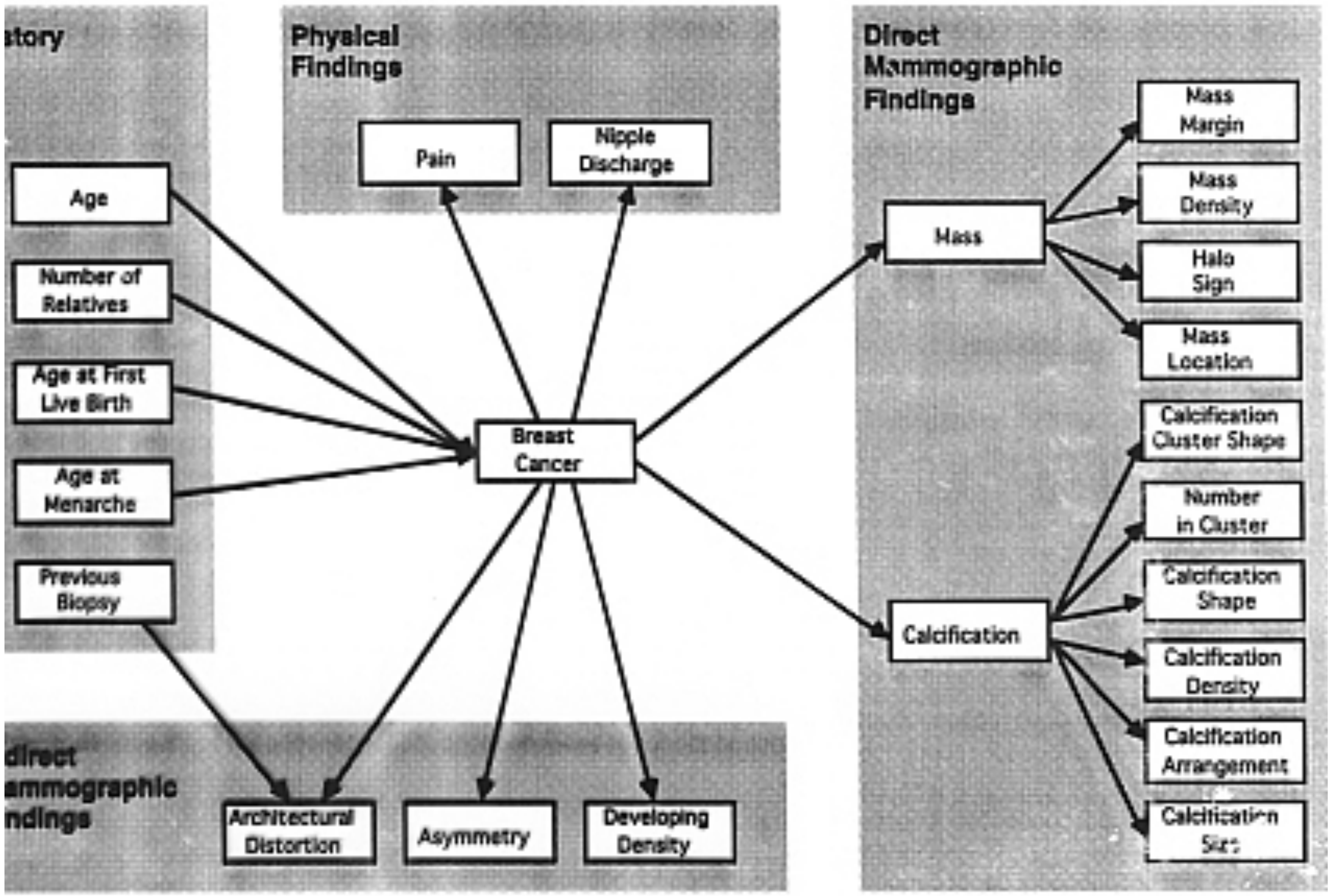


- Deciding conditional independence is hard in noncausal directions
- (Causal models and conditional independence seem hardwired for humans!)
- Network is less compact: 1 + 2 + 4 + 2 + 4 = 13 numbers needed

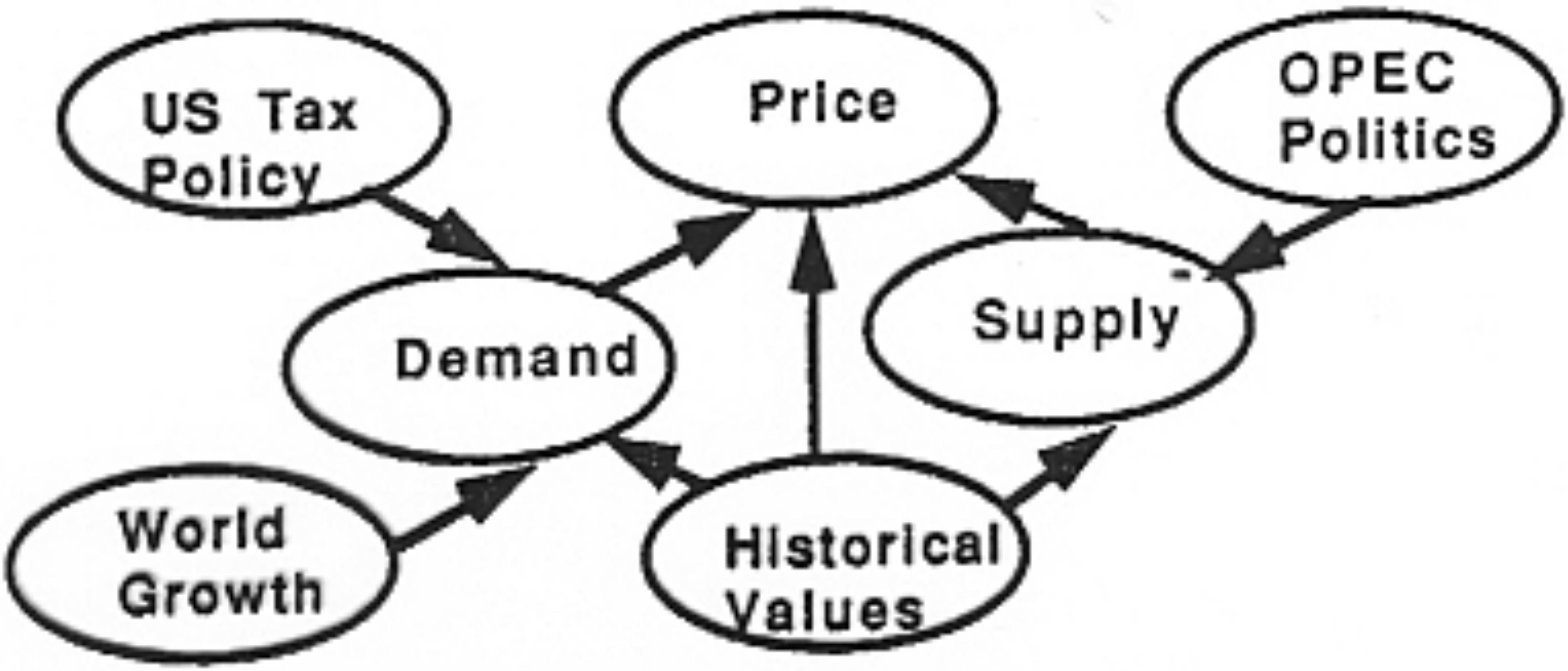
Example: Car Diagnosis



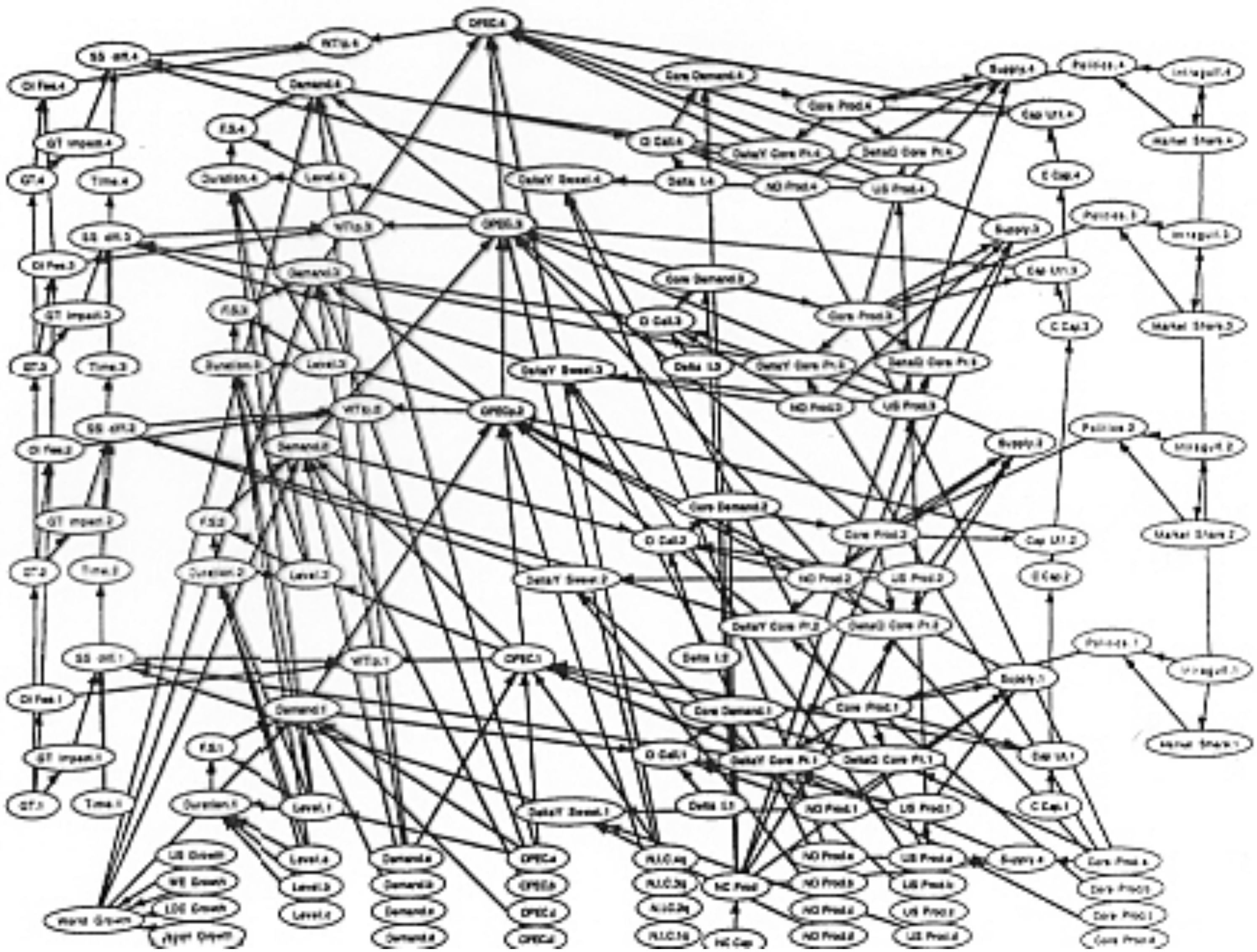
MammoNet


story
Physical Findings
Direct Mammographic Findings
Age
Number of Relatives
Age at First Live Birth
Age at Menarche
Previous Biopsy
Pain
Nipple Discharge
Breast Cancer
Mass
Mass Margin
Mass Density
Halo Sign
Mass Location
Calcification
Calcification Cluster Shape
Number in Cluster
Calcification Shape
Calcification Density
Calcification Arrangement
direct ammographic ndings
Architectural Distortion
Asymmetry
Developing Density

ARCO1: Forecasting Oil Prices

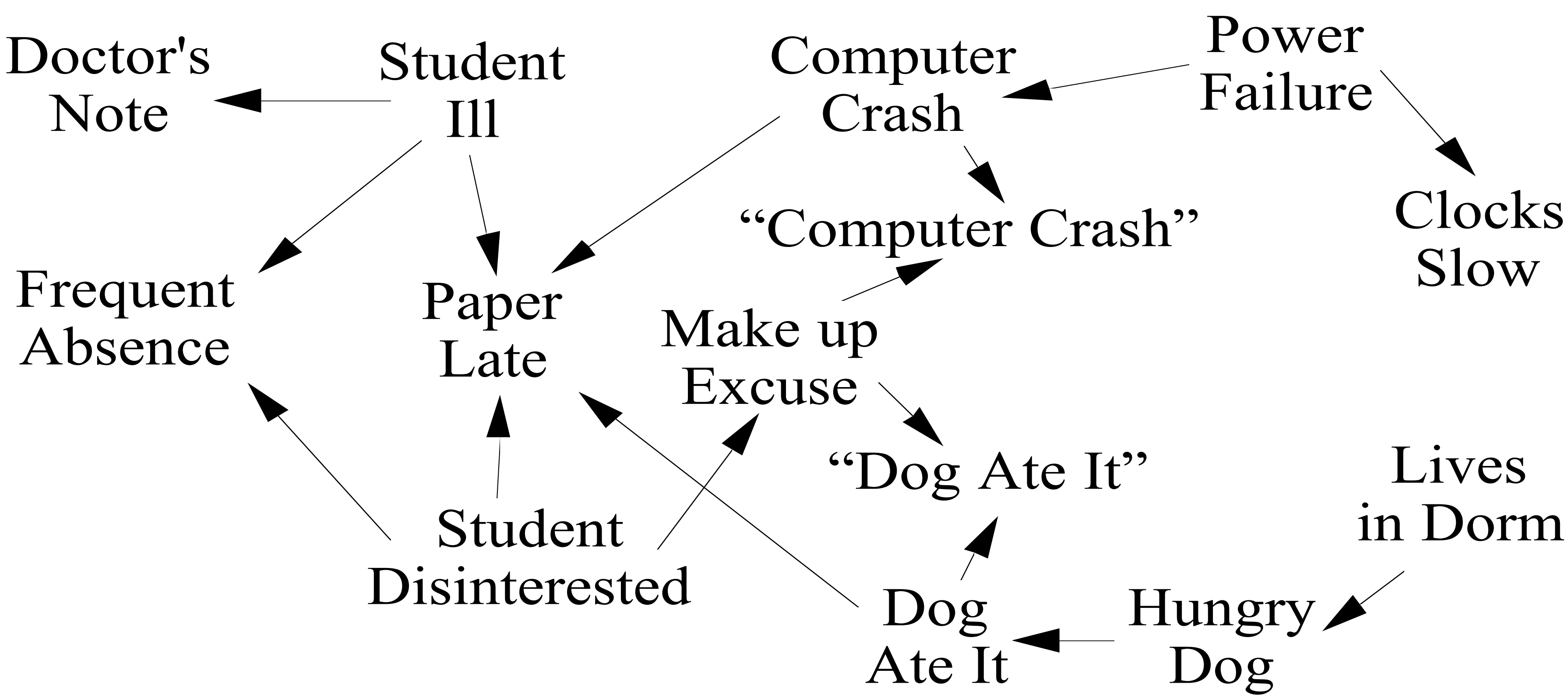


ARCO1: Forecasting Oil Prices



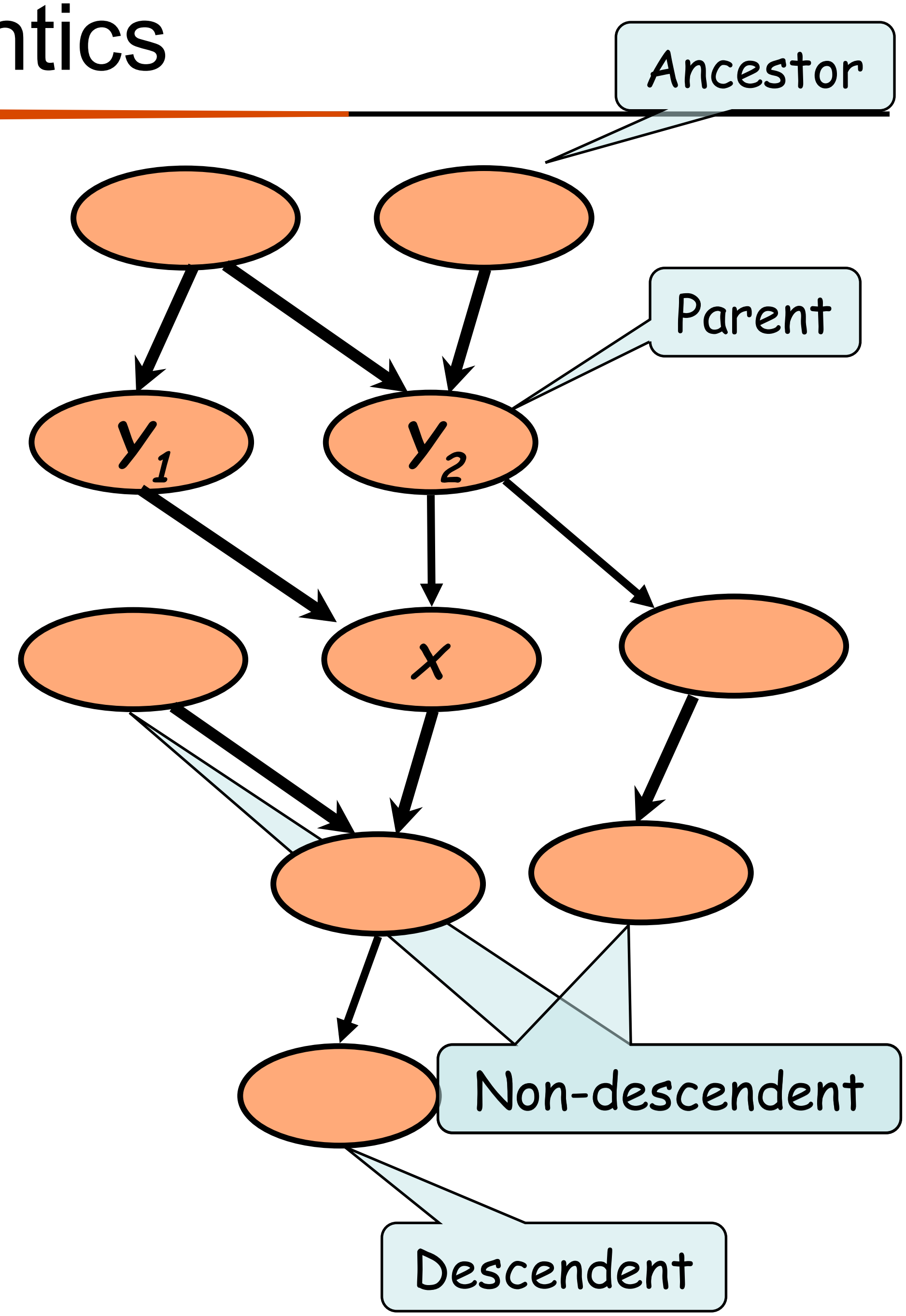
Construct a Bayes Net…

- …to reason about why a student's homework was late
- Choose your variables. Consider things like:
  - Why was the homework late?
  - What kind of a student is this person?
  - What other behaviors might you observe?
- Then build the network

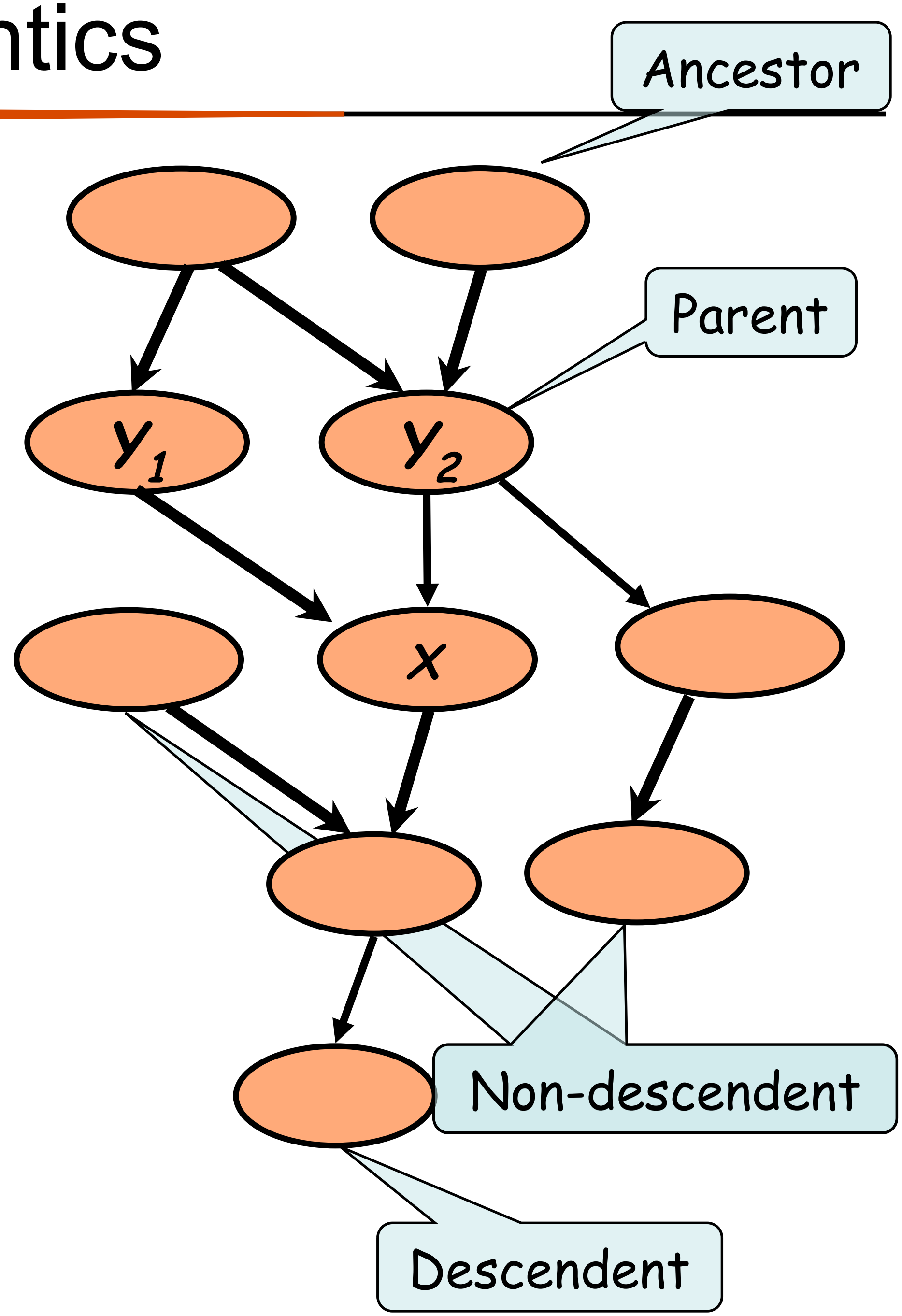


Doctor's Note
Student Ill
Computer Crash
Power Failure
"Computer Crash"
Clocks Slow
Frequent Absence
Paper Late
Make up Excuse
"Dog Ate It"
Lives in Dorm
Student Disinterested
Dog Ate It
Hungry Dog

Topological Semantics

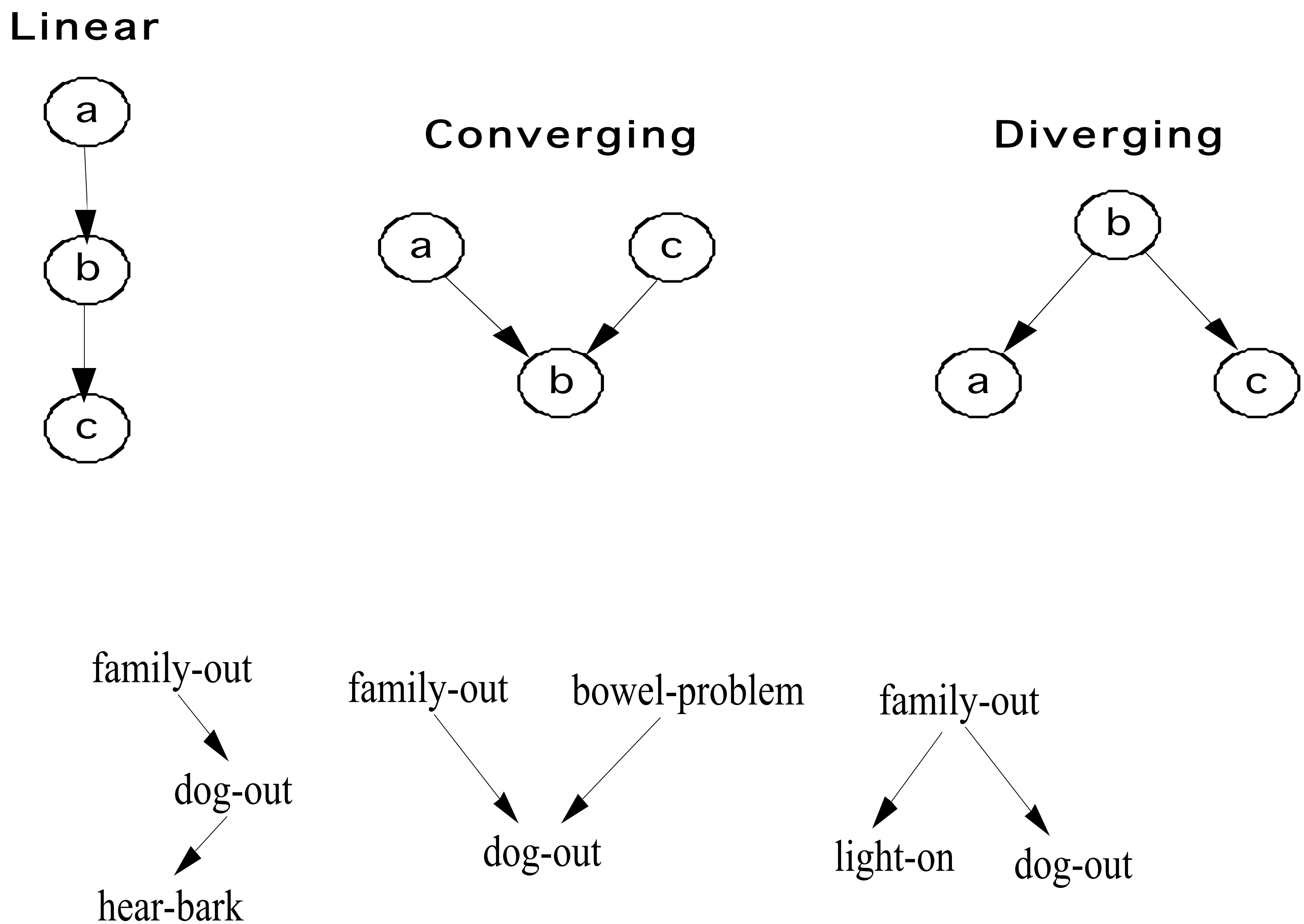


Topological Semantics



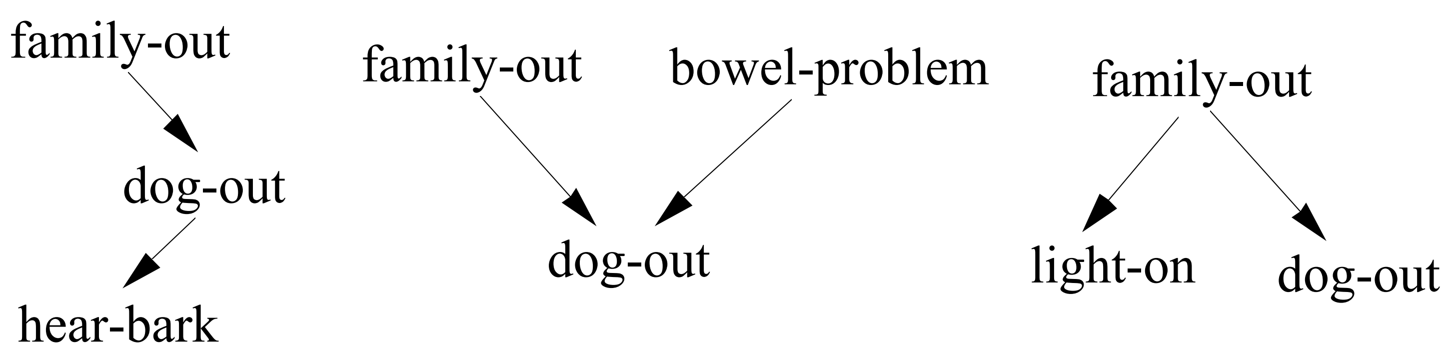
Local Markov Assumption:
A variable X is independent of its non-descendents given its parents

Three Types of Connections

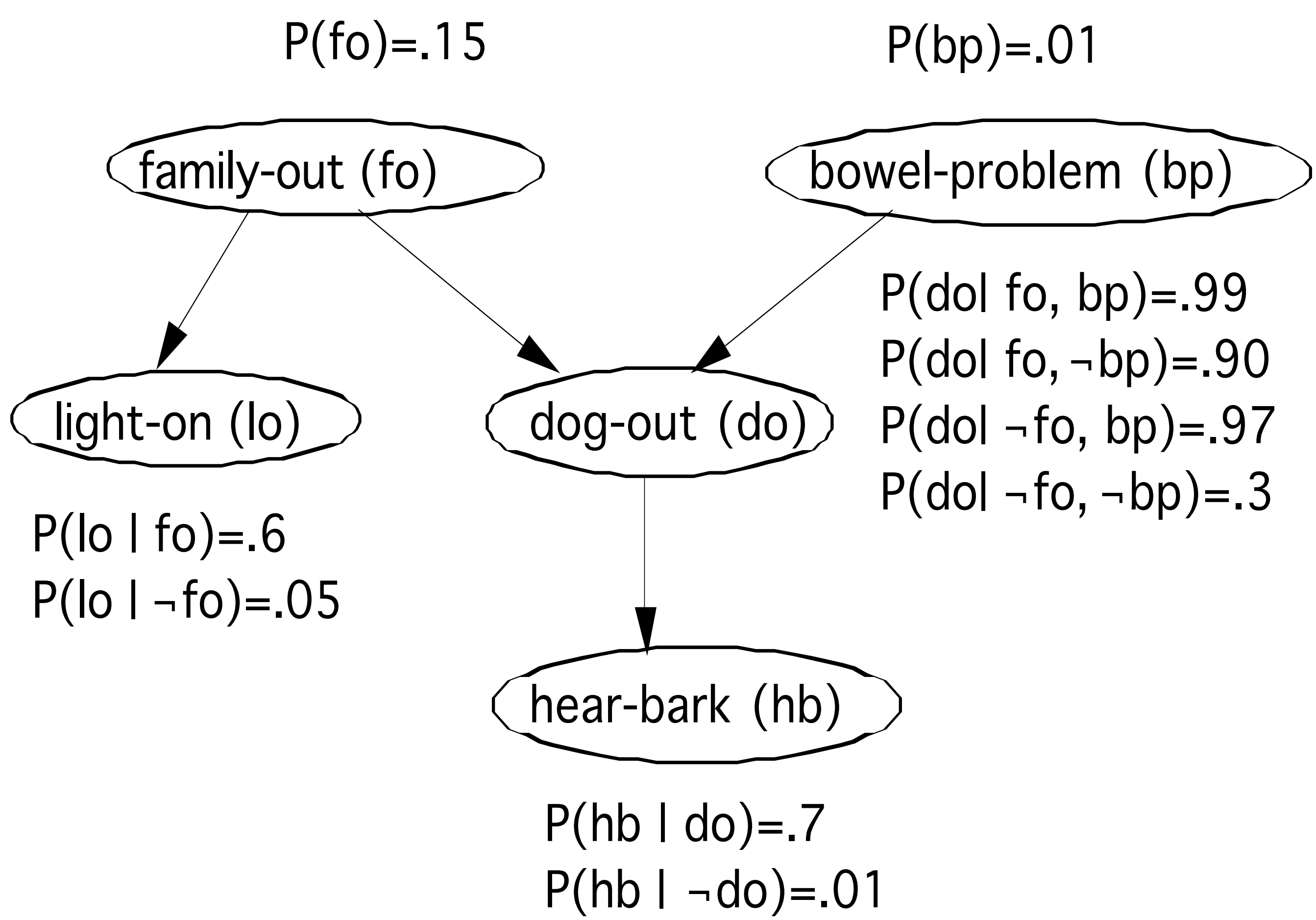


Connection Pattern and Independence

- **Linear connection**: The two end variables are dependent on each other. The middle variable renders them independent.
- **Converging connection**: The two end variables are independent of each other. The middle variable renders them dependent.
- **Divergent connection**: The two end variables are dependent on each other. The middle variable renders them independent.



Bayes nets represent joint probabilities



$P(fo, \neg lo, do, hb, \neg bp) = ?$

Inference in Bayesian Networks

- The inputs to a Bayesian Network evaluation algorithm is a set of evidences: e.g.,

  E = { hear-bark=true, lights-on=true }

- The outputs of Bayesian Network evaluation algorithm are
  - Simple queries

    $$P(X_i \mid E)$$

    where X_i is a variable in the network.
  - conjunctive queries:

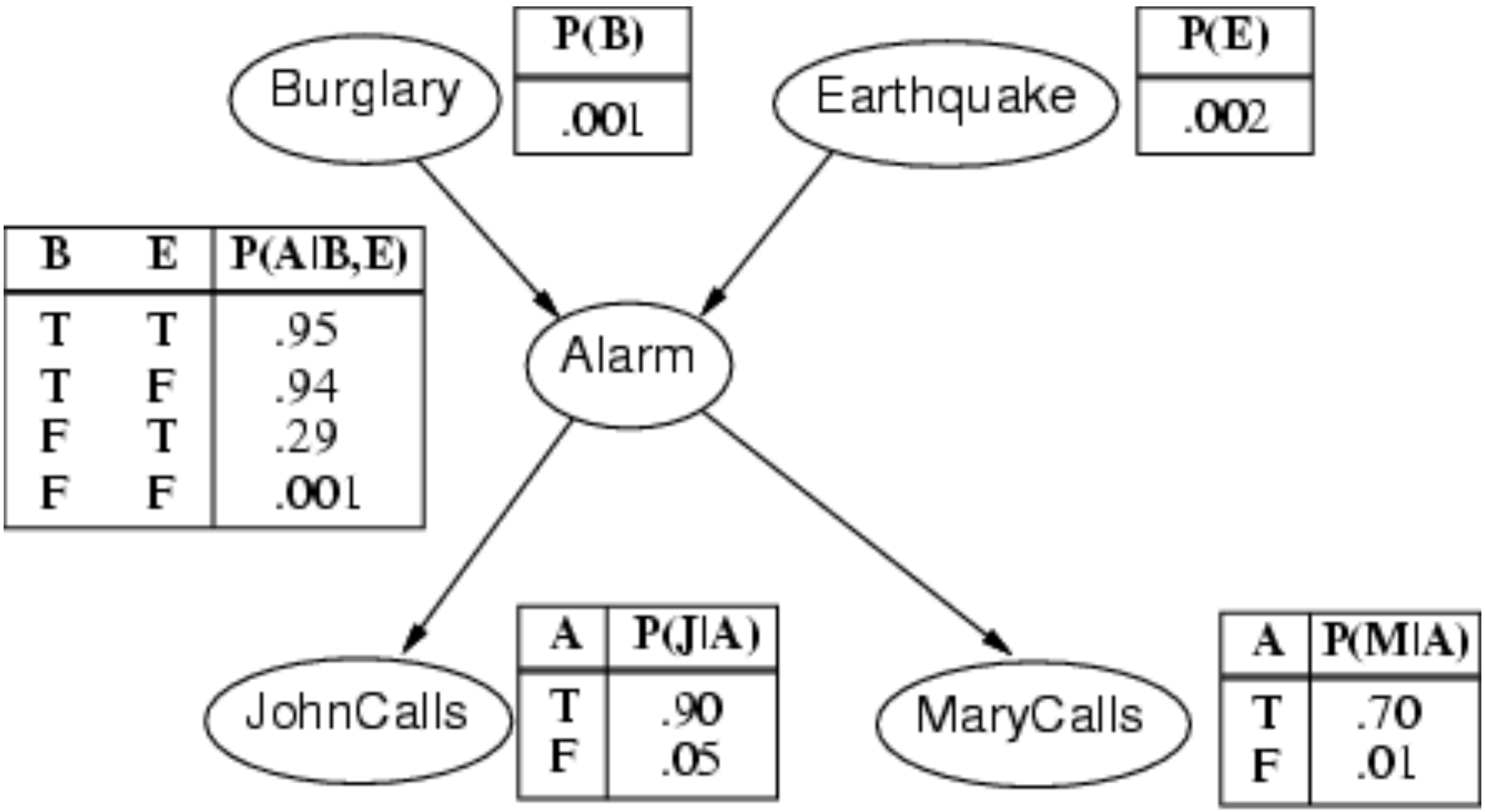
    $P(X_i,X_j \mid E)$

Inference Overview

- Exact Inference: Today
  - Enumeration
  - Variable Elimination
  - Belief Propagation

- Approximate Inference: Tuesday

Reminder: Burglary network



P(B)
.001

P(E)
.002

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

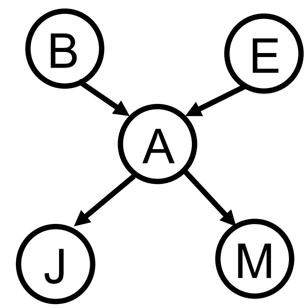
A	P(J|A)
T	.90
F	.05

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

Inference by Enumeration

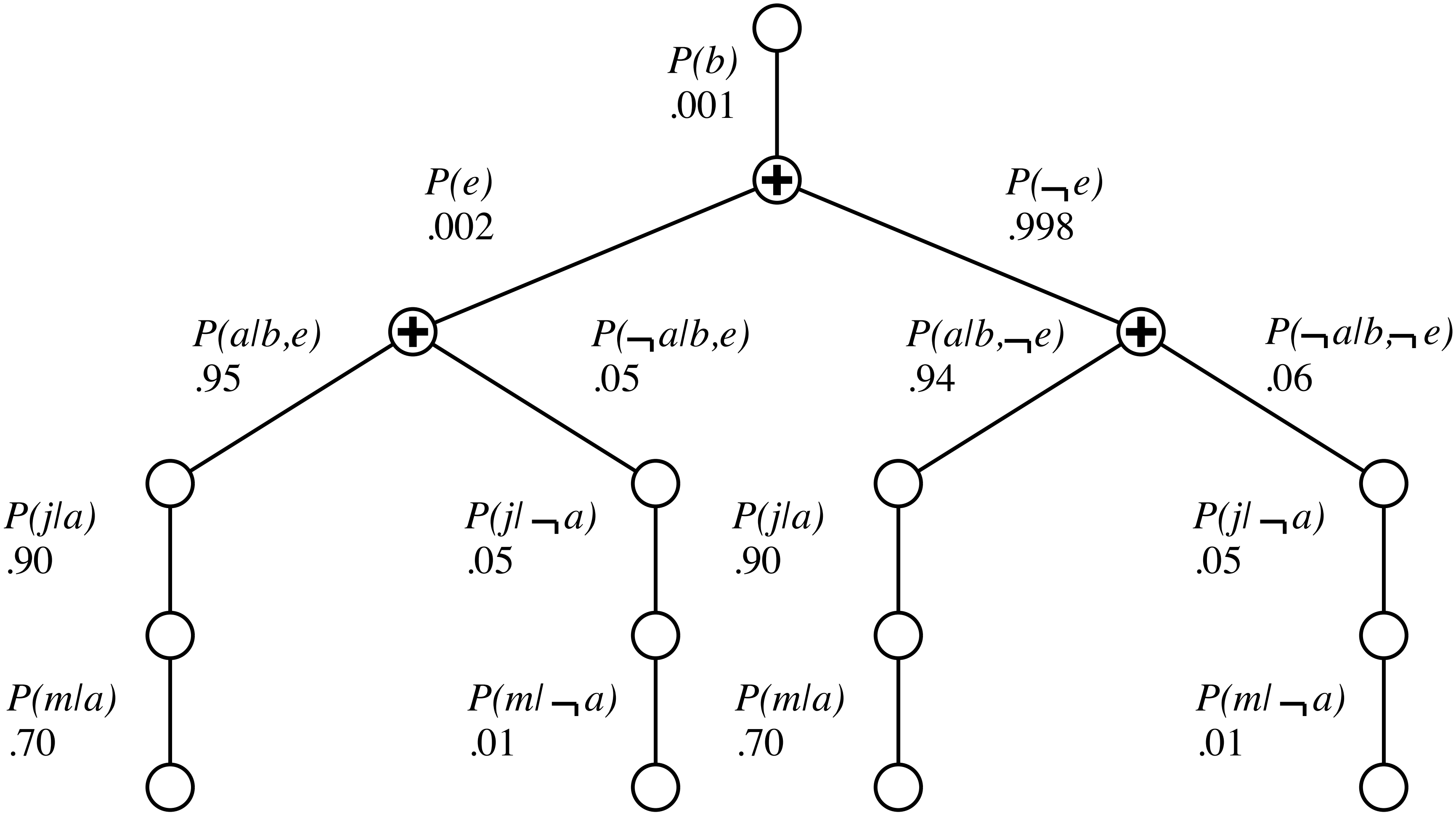
- Bayes Nets represent a joint probability

Simple query on the burglary network:



$$P(B|j,m)$$
$$= P(B,j,m) \,/\, P(j,m)$$
$$= \alpha P(B,j,m)$$
$$= \alpha \sum_e \sum_a P(B,e,a,j,m)$$

Evaluation Tree



Inference By Variable Elimination

- Carry out sums from right to left, storing intermediate results to avoid recomputation

$$
\begin{aligned}
\mathbf{P}(B|j,m) \\
&= \alpha \underbrace{\mathbf{P}(B)}_{B} \textstyle\sum_e \underbrace{P(e)}_{E} \textstyle\sum_a \underbrace{\mathbf{P}(a|B,e)}_{A} \underbrace{P(j|a)}_{J} \underbrace{P(m|a)}_{M} \\
&= \alpha \mathbf{P}(B) \textstyle\sum_e P(e) \textstyle\sum_a \mathbf{P}(a|B,e) P(j|a) f_M(a) \\
&= \alpha \mathbf{P}(B) \textstyle\sum_e P(e) \textstyle\sum_a \mathbf{P}(a|B,e) f_J(a) f_M(a) \\
&= \alpha \mathbf{P}(B) \textstyle\sum_e P(e) \textstyle\sum_a f_A(a,b,e) f_J(a) f_M(a) \\
&= \alpha \mathbf{P}(B) \textstyle\sum_e P(e) f_{\bar{A}JM}(b,e) \text{ (sum out } A\text{)} \\
&= \alpha \mathbf{P}(B) f_{\bar{E}\bar{A}JM}(b) \text{ (sum out } E\text{)} \\
&= \alpha f_B(b) \times f_{\bar{E}\bar{A}JM}(b)
\end{aligned}
$$

Variable Elimination: Details

- To sum out a variable from a product of factors:
  - Move any constant factors outside the sum
  - Add up submatrices in pointwise product of remaining factors

$$\Sigma_x f_1 \times \cdots \times f_k = f_1 \times \cdots \times f_i \, \Sigma_x \, f_{i+1} \times \cdots \times f_k = f_1 \times \cdots \times f_i \times f_{\bar{X}}$$

$f_1 \ldots f_i$ do not depend on x

Pointwise product of factors f_1 and f_2:

$$f_1(x_1, \ldots, x_j, y_1, \ldots, y_k) \times f_2(y_1, \ldots, y_k, z_1, \ldots, z_l)$$
$$= f(x_1, \ldots, x_j, y_1, \ldots, y_k, z_1, \ldots, z_l)$$

E.g., $f_1(a, b) \times f_2(b, c) = f(a, b, c)$

Pointwise Product

Pointwise product of factors f_1 and f_2:

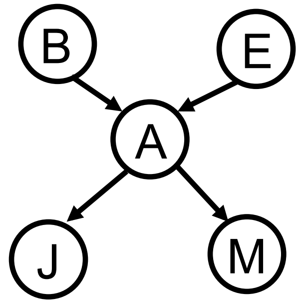
$$f_1(x_1, \ldots, x_j, y_1, \ldots, y_k) \times f_2(y_1, \ldots, y_k, z_1, \ldots, z_l)$$
$$= f(x_1, \ldots, x_j, y_1, \ldots, y_k, z_1, \ldots, z_l)$$

E.g., $f_1(a, b) \times f_2(b, c) = f(a, b, c)$

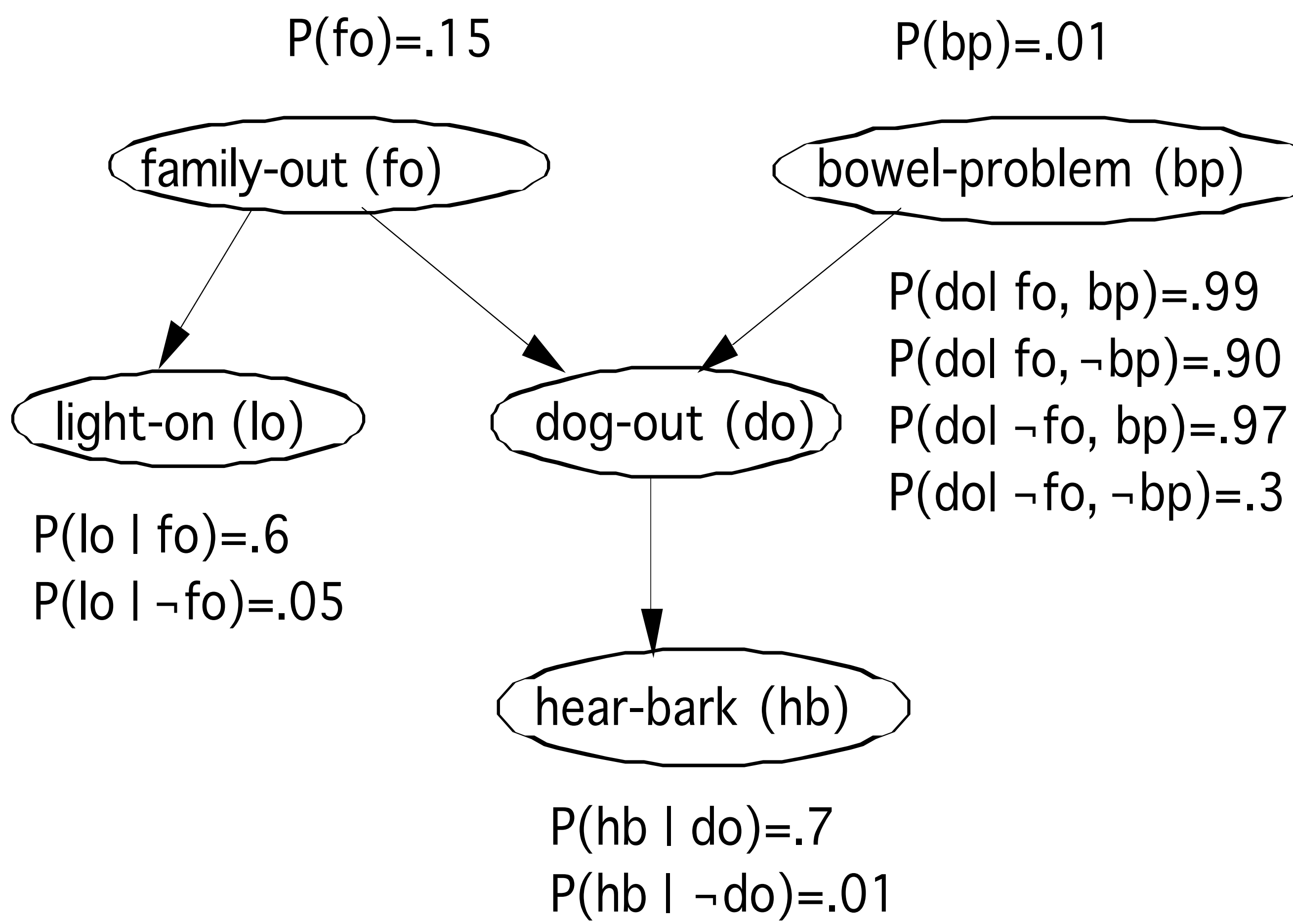
A	B	f_1(A,B)	B	C	f_2(B,C)	A	B	C	f_3(A,B,C)
T	T	.3	T	T	.2	T	T	T	
T	F	.7	T	F	.8	T	T	F	
F	T	.9	F	T	.6	T	F	T	
F	F	.1	F	F	.4	T	F	F	
						F	T	T	
						F	T	F	
						F	F	T	
						F	F	F	

Irrelevant variables

- What if you want to know P(J|b)?



$$P(J \mid b) = \alpha P(b) \sum_e P(e) \sum_a P(a \mid b, e) P(J \mid a) \sum_m P(m \mid a)$$

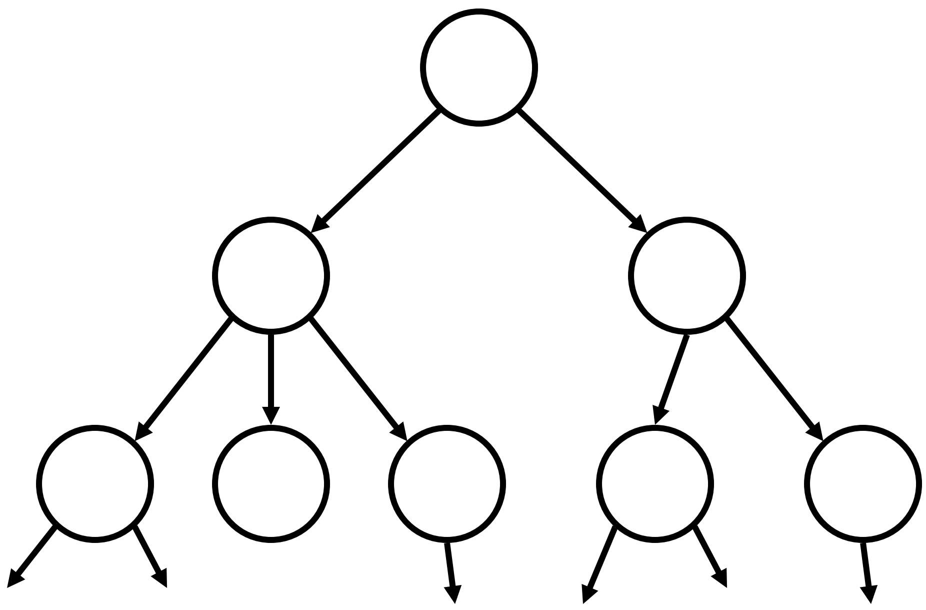

P(fo)=.15
P(bp)=.01
family-out (fo)
bowel-problem (bp)
P(do| fo, bp)=.99
P(do| fo, ¬bp)=.90
P(do| ¬fo, bp)=.97
P(do| ¬fo, ¬bp)=.3
light-on (lo)
dog-out (do)
P(lo | fo)=.6
P(lo | ¬fo)=.05
hear-bark (hb)
P(hb | do)=.7
P(hb | ¬do)=.01

So is VE any better than Enumeration?

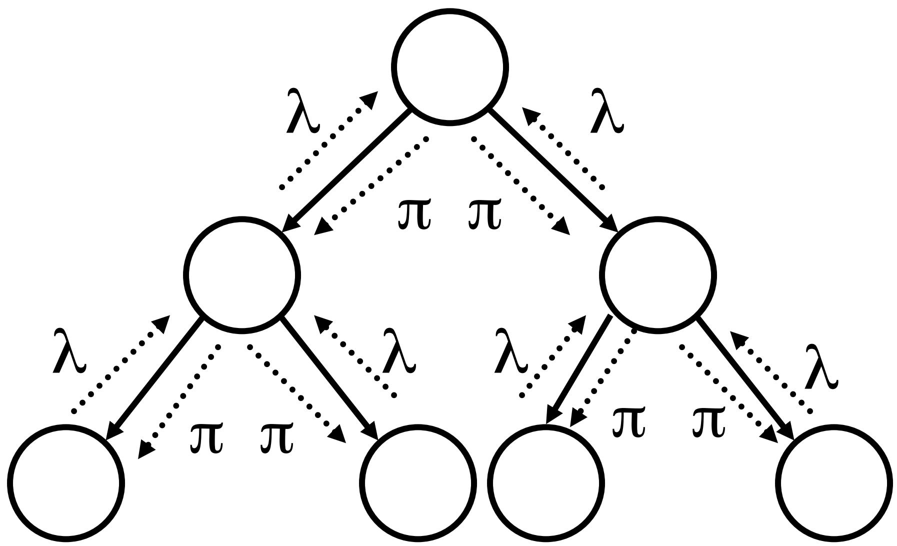
- Yes and No…
  - For singly-connected networks (poly-trees), YES
  - In general, NO…

Bayesian Network Inference

- **But**…inference is still tractable in some cases.
- Special case: trees (each node has one parent)
- VE is LINEAR in this case



Another Algorithm: Belief Propagation



www.cs.cmu.edu/~awm/381/lec/bayesinfer/bayesinf.ppt

Conjunctive queries

- What if we want, e.g., P(A, B | C) instead of just marginal distributions P(A | C) and P(B | C)?
- Just use chain rule (successive applications of the product rule):
  - $P(A, B \mid C) = P(A \mid B,C)\, P(B \mid C)$
  - Each of the latter probabilities can be computed using the technique just discussed.

Summary

- Bayesian networks encode independence assumptions
- Inference in Bayes Nets is inherently difficult, unless the network has special structure
- Variable Elimination and Belief Propagation save time by:
  - Exploiting independence
  - Storing intermediate results