

# For Friday

- Finish chapter 14
- Homework:
  - Chapter 13, exercises 6 and 8
- Wumpus program due

# Program 2

- Any questions?

# Probability

- Questions?

# Problems with Probabilistic Reasoning

- If no assumptions of independence are made, then an exponential number of parameters is needed for sound probabilistic reasoning.
- There is almost never enough data or patience to reliably estimate so many very specific parameters.
- If a blanket assumption of conditional independence is made, efficient probabilistic reasoning is possible, but such a strong assumption is rarely warranted.

# Practical Naïve Bayes

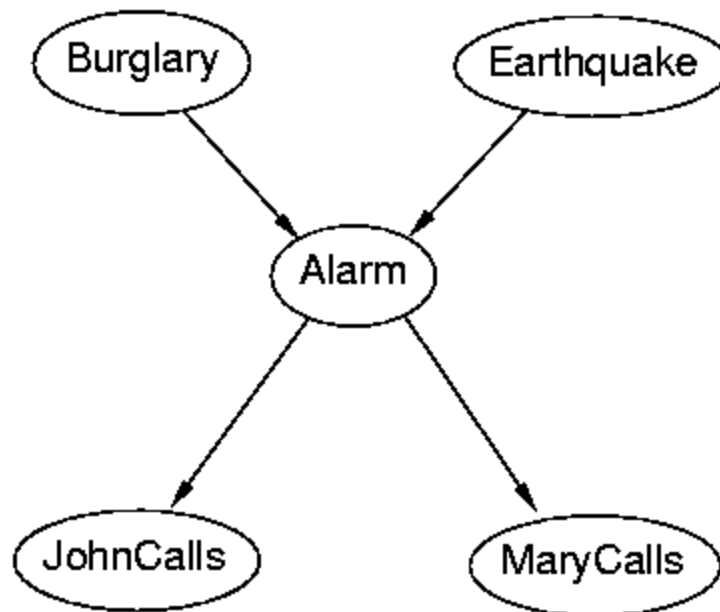
- We're going to assume independence, so what numbers do we need?
- Where do the numbers come from?

# Bayesian Networks

- Bayesian networks (belief network, probabilistic network, causal network) use a directed acyclic graph (DAG) to specify the direct (causal) dependencies between variables and thereby allow for limited assumptions of independence.
- The number of parameters need for a Bayesian network are generally much less compared to making no independence assumptions.

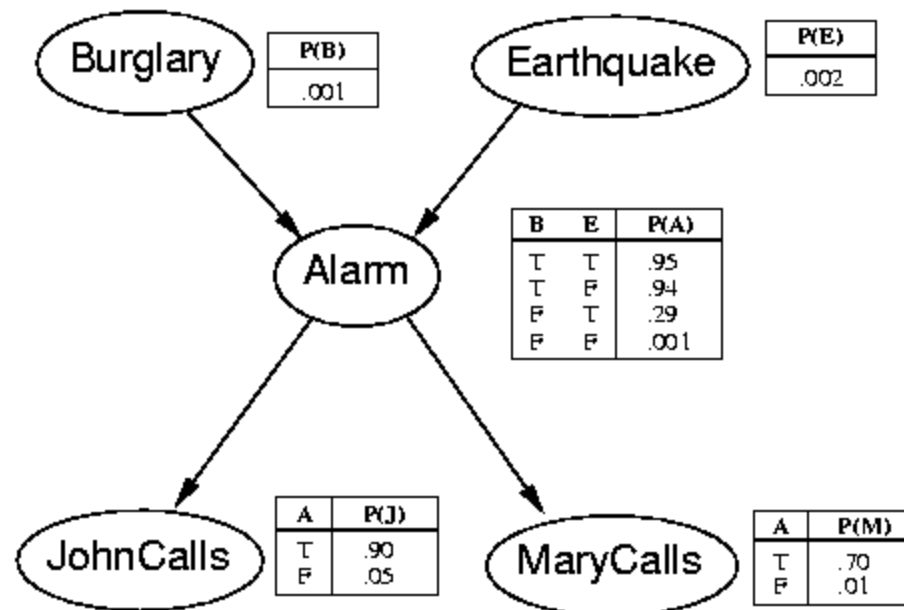
## Bayesian Networks

- Each random variable is represented with a node.
- Directed edges indicate direct causal influences.



## Conditional Probability Tables

- Each node in the network has a **conditional probability table (CPT)** that gives the probability of each of its values given every possible combination of values for its parents (called a **conditioning case**).
- Roots (sources) of the graph that have no parents are given prior probabilities.



# More on CPTs

- Probability of false is not given since rows must sum to 1.
- Requires 10 parameters rather than  $2^5 = 32$  (actually only 31 since all 32 values must sum to 1)
- Therefore, the number of probabilities needed for a node is exponential in the number of parents (the **fan-in**).

## Joint Probability Distribution

- A Bayesian network implicitly defines a joint distribution

$P(x_1, \dots, x_n)$  shorthand for  $P((X_1 = x_1) \wedge \dots \wedge (X_n = x_n))$

$$P(x_1, \dots, x_n) = \prod_{i=1}^n P(x_i | \text{Parents}(X_i))$$

- Example

$$\begin{aligned} & P(J \wedge M \wedge A \wedge \neg B \wedge \neg E) \\ &= P(J|A)P(M|A)P(A|\neg B \wedge \neg E)P(\neg B)P(\neg E) \\ &= 0.9 \times 0.70 \times 0.001 \times 0.999 \times 0.998 = 0.00062 \end{aligned}$$

- Therefore, an inefficient approach to doing probabilistic inference with a Bayes net is to first calculate the complete joint distribution and then use this to calculate any needed unconditioned or conditional probabilities.

# Noisy-Or Nodes

- To avoid specifying the complete CPT, special nodes that make assumptions about the style of interaction can be used.
- A noisy-or node assumes that the parents are **independent causes** that are noisy, i.e. there is some probability that they will not cause the effect.
- The noise parameter for each cause indicates the probability it will **not** cause the effect.
- Probability that the effect is not present is the product of the noise parameters of all the parent nodes that are true (since independence is assumed).  
$$P(\text{Fever}|\text{Cold}) = 0.4, P(\text{Fever}|\text{Flu}) = 0.8, P(\text{Fever}|\text{Malaria}) = 0.9$$
$$P(\text{Fever} | \text{Cold} \wedge \text{Flu} \wedge \neg\text{Malaria}) = 1 - 0.6 * 0.2 = 0.88$$
- Number of parameters needed is **linear** in fan-in rather than **exponential**.

# Independencies

- If removing a subset of nodes  $S$  from the network renders nodes  $X_i$  and  $X_j$  disconnected, then  $X_i$  and  $X_j$  are independent given  $S$ , i.e.

$$P(X_i | X_j, S) = P(X_i | S)$$

- However, this is too strict a criteria for conditional independence since two nodes will still be considered independent if there simply exists some variable that depends on both. (i.e. Burglary and Earthquake should be considered independent since the both cause Alarm)

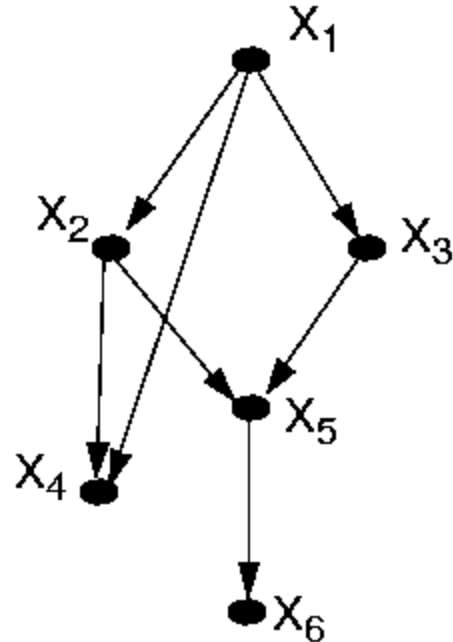
- Unless we know something about a **common effect** of two “independent causes” or a descendent of a common effect, then they can be considered independent.
- For example, **if** we know nothing else, Earthquake and Burglary are **independent**.
- However, if we have information about a common effect (or descendent thereof) then the two “independent” causes become probabilistically linked since evidence for one cause can “explain away” the other.
- If we know the alarm went off, then it makes earthquake and burglary dependent since evidence for earthquake decreases belief in burglary and vice versa.

# Types of Connections

- Given a triplet of variables  $x, y, z$  where  $x$  is connected to  $z$  via  $y$ , there are 3 possible connection types:
  - tail-to-tail:  $x \leftarrow y \rightarrow z$
  - head-to-tail:  $x \leftarrow y \leftarrow z$ , or  $x \rightarrow y \rightarrow z$
  - head-to-head:  $x \rightarrow y \leftarrow z$
- For tail-to-tail and head-to-tail connections,  $x$  and  $z$  are independent given  $y$ .
- For head-to-head connections,  $x$  and  $z$  are “marginally independent” but may become dependent given the value of  $y$  or one of its descendants (through “explaining away”).

# Separation

- A subset of variables  $S$  is said to **separate**  $X$  from  $Y$  if all (undirected) paths between  $X$  and  $Y$  are separated by  $S$ .
- A path  $P$  is separated by a subset of variables  $S$  if at least one pair of successive links along  $P$  is **blocked** by  $S$ .
- Two links meeting head-to-tail or tail-to-tail at a node  $Z$  are blocked by  $S$  if  $Z$  is in  $S$ .
- Two links meeting head-to-head at a node  $Z$  are blocked by  $S$  if neither  $Z$  nor any of its descendants are in  $S$ .



$X_2$  and  $X_3$  are separated by  $\{X_1\}$  and  $\{X_1, X_4\}$

$X_2$  and  $X_3$  are not separated by  $\{X_1, X_6\}$  since  $X_6$  as a descendent of  $X_5$ , unblocks the head-to-head connection at  $X_5$ .

Does  $\{X_1\}$  separate  $X_4$  from  $X_3$ ?

Does  $\{X_1, X_6\}$  separate  $X_4$  from  $X_3$ ?

Does  $\{X_5\}$  separate  $X_1$  from  $X_6$ ?

Does  $\{X_2\}$  separate  $X_4$  from  $X_6$ ?

# Probabilistic Inference

- Given known values for some evidence variables, we want to determine the **posterior probability** of some query variables.
- **Example:** Given that John calls, what is the probability that there is a Burglary?
- John calls **90%** of the time there is a burglary and the alarm detects **94%** of burglaries, so people generally think it should be fairly high (80-90%). But this ignores the **prior probability** of John calling. John also calls **5%** of the time when there is no alarm. So over the course of **1,000** days we expect one burglary and John will probably call. But John will also call with a false report **50** times during **1,000** days on average. So the call is about **50 times more likely** to be a false report
- **$P(\text{Burglary} \mid \text{JohnCalls}) \sim 0.02$ .**
- Actual probability is **0.016** since the alarm is not perfect (an earthquake could have set it off or it could have just went off on its own). Of course even if there was no alarm and John called incorrectly, there could have been an undetected burglary anyway, but this is very unlikely.

# Types of Inference

- **Diagnostic (evidential, abductive):** From effect to cause.

$$P(\text{Burglary} \mid \text{JohnCalls}) = 0.016$$

$$P(\text{Burglary} \mid \text{JohnCalls} \wedge \text{MaryCalls}) = 0.29$$

$$P(\text{Alarm} \mid \text{JohnCalls} \wedge \text{MaryCalls}) = 0.76$$

$$P(\text{Earthquake} \mid \text{JohnCalls} \wedge \text{MaryCalls}) = 0.18$$

- **Causal (predictive):** From cause to effect

$$P(\text{JohnCalls} \mid \text{Burglary}) = 0.86$$

$$P(\text{MaryCalls} \mid \text{Burglary}) = 0.67$$

# More Types of Inference

- **Intercausal (explaining away):** Between causes of a common effect.

$$P(\text{Burglary} \mid \text{Alarm}) = 0.376$$

$$P(\text{Burglary} \mid \text{Alarm} \wedge \text{Earthquake}) = 0.003$$

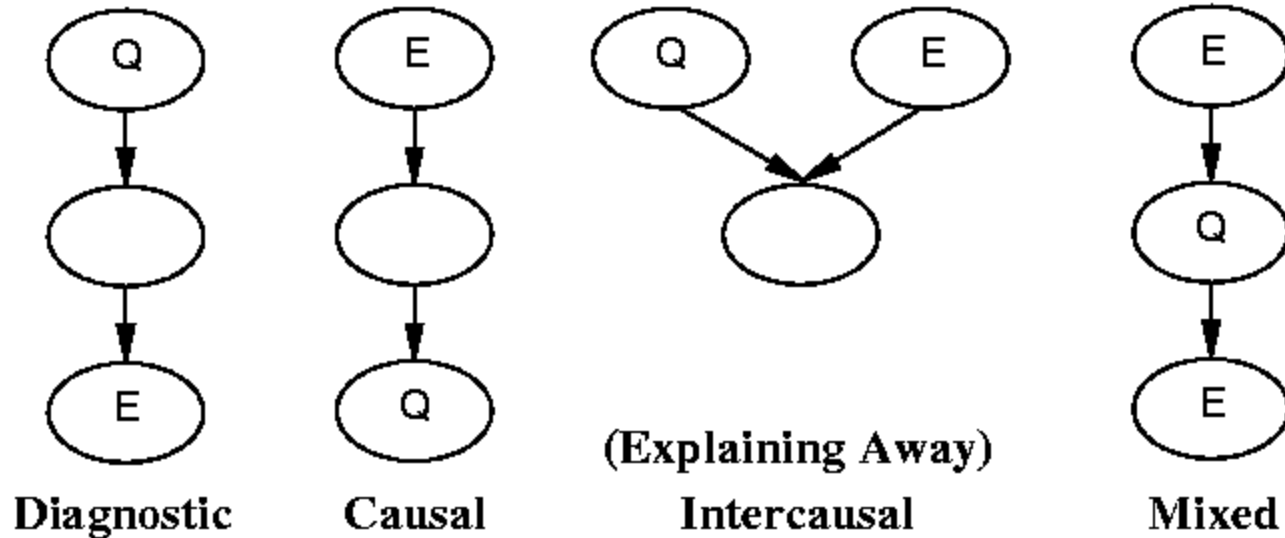
- **Mixed:** Two or more of the above combined (diagnostic and causal)

$$P(\text{Alarm} \mid \text{JohnCalls} \wedge \neg \text{Earthquake}) = 0.03$$

(diagnostic and intercausal)

$$P(\text{Burglary} \mid \text{JohnCalls} \wedge \neg \text{Earthquake}) = 0.017$$

## Types of Inference (figure)



# Inference Algorithms

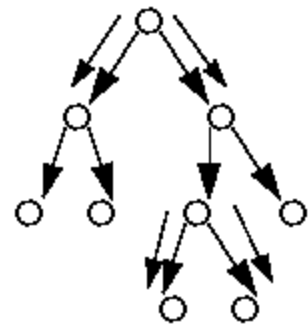
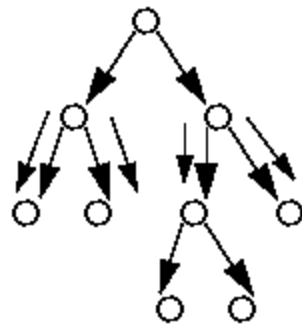
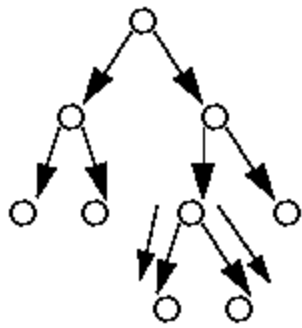
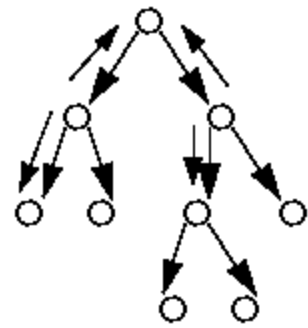
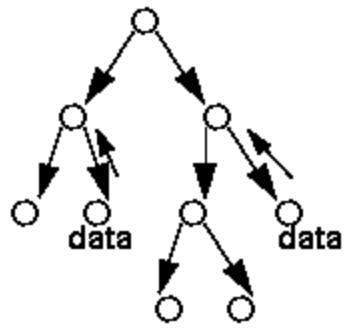
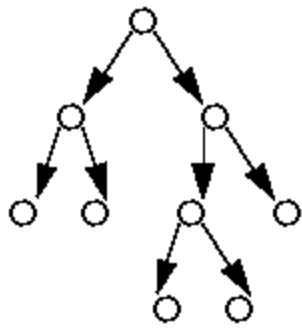
- Most inference algorithms for Bayes nets are not goal-directed and calculate posterior probabilities for all other variables.
- In general, the problem of Bayes net inference is NP-hard (exponential in the size of the graph).

# Polytree Inference

- For singly-connected networks or polytrees, in which there are no undirected loops (there is at most one undirected path between any two nodes), polynomial (linear) time algorithms are known.
- Details of inference algorithms are somewhat mathematically complex, but algorithms for polytrees are structurally quite simple and employ simple propagation of values through the graph.

# Belief Propagation

- Belief propagation and updating involves transmitting two types of messages between neighboring nodes:
  - $\lambda$  messages are sent from children to parents and involve the strength of evidential support for a node.
  - $\pi$  messages are sent from parents to children and involve the strength of causal support.



# Propagation Details

- Each node  $B$  acts as a simple processor which maintains a vector  $\lambda(B)$  for the total evidential support for each value of the corresponding variable and an analogous vector  $\pi(B)$  for the total causal support.
- The belief vector  $BEL(B)$  for a node, which maintains the probability for each value, is calculated as the normalized product:

$$BEL(B) = \alpha \lambda(B) \pi(B)$$

# Propogation Details (cont.)

- Computation at each node involve  $\lambda$  and  $\pi$  message vectors sent between nodes and consists of simple matrix calculations using the CPT to update belief (the  $\lambda$  and  $\pi$  node vectors) for each node based on new evidence.
- Assumes CPT for each node is a matrix (M) with a column for each value of the variable and a row for each conditioning case (all rows must sum to 1).

		value of Alarm	
		T	F
Values of Burglary & Earthquake	T T	0.95	0.05
	T F	0.94	0.06
	F T	0.29	0.71
	F F	0.001	0.999

- Propagation algorithm is simplest for trees in which each node has only one parent (i.e. one cause).
- To initialize,  $\lambda(B)$  for all leaf nodes is set to all 1's and  $\pi(B)$  of all root nodes is set to the priors given in the CPT. Belief based on the root priors is then propagated down the tree to all leaves to establish priors for all nodes.
- Evidence is then added incrementally and the effects propagated to other nodes.