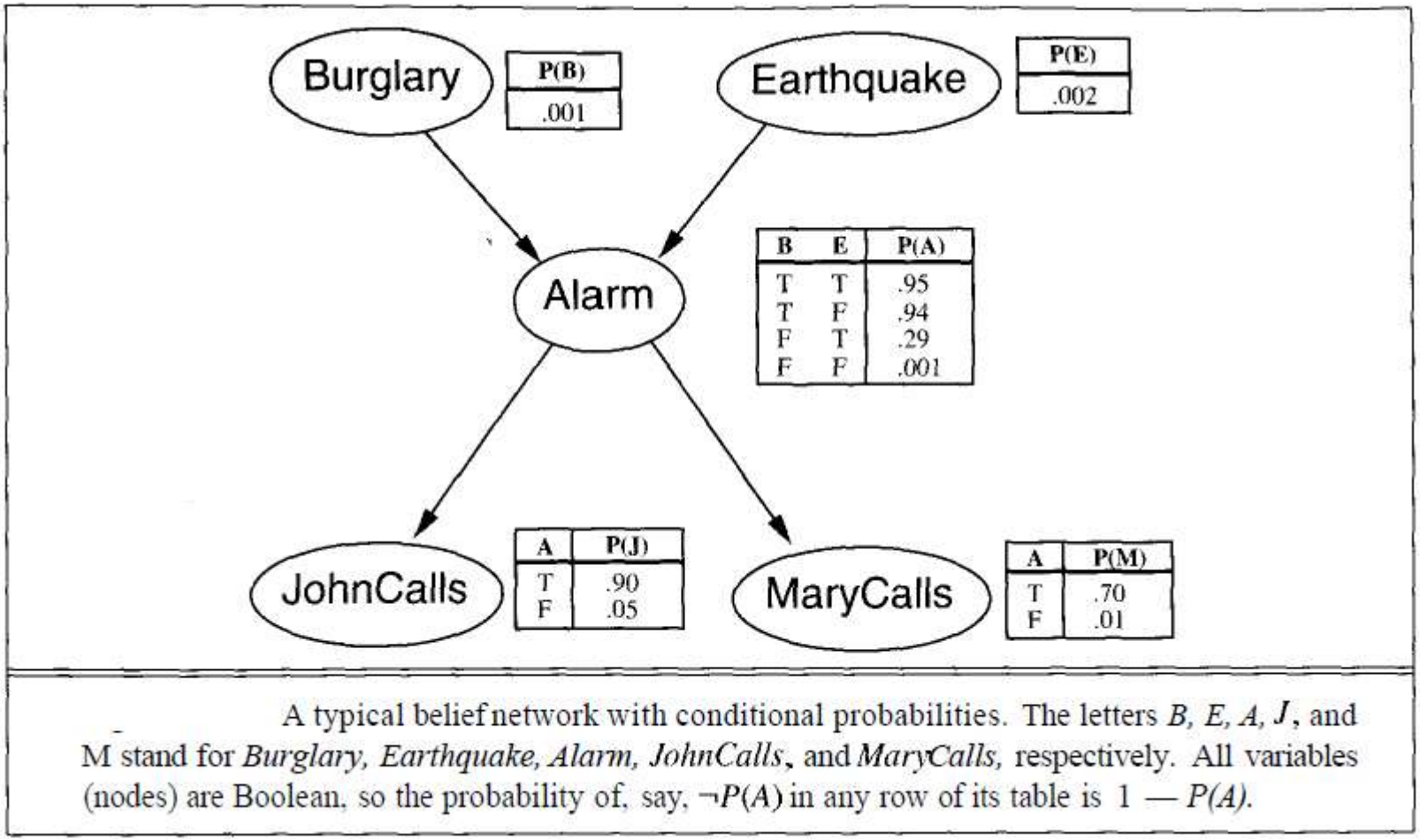
LU32-Bayesian networks – The semantics of Bayesian Networks

LU Objectives	
To explain the method for constructing Bayesian network	
To represent conditional distribution efficiently	
LU Outcomes	CO : 5
Construct Bayesian network	
Represent conditional distributions	

A typical Bayesian or Belief Networks



A typical belief network with conditional probabilities. The letters *B*, *E*, *A*, *J*, and M stand for *Burglary*, *Earthquake*, *Alarm*, *JohnCalls*, and *MaryCalls*, respectively. All variables (nodes) are Boolean, so the probability of, say, $\neg P(A)$ in any row of its table is $1 - P(A)$.

Semantics of Bayesian Networks

- There are two ways in which one can understand the semantics of Bayesian networks.

  1. The first is to see the network as a representation of the joint probability distribution.

  2. The second is to view it as an encoding of a collection of conditional independence statements.

- Two views are equivalent, but the first turns out to be helpful in understanding how to *construct networks,* whereas the second is helpful in designing inference procedures.

Representing the full joint distribution

- A Bayesian network is a directed acyclic graph with some numeric parameters attached to each node.
- One way to define what the network means is to define the way in which it represents a specific joint distribution over all the variables.
- To do this, we first need to retract (temporarily) what we said earlier about the parameters.
- Parameters correspond to conditional probabilities $P(X_i \mid \text{Parents}(X_i))$ but until we assign semantics to the network as a whole, they are just numbers $\theta(X_i \mid \text{Parents}(X_i))$

Representing the full joint distribution

- Joint distribution is the probability of a conjunction of particular assignments to each variable

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} \theta(x_i \,|\, parents(X_i))$$

- So joint distribution is represented by the product of the appropriate elements of the conditional probability tables (CPTs) in the Bayesian network
- So the equation can be rewritten as

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \,|\, parents(X_i))$$

Representing the full joint distribution

- To calculate the probability that the alarm has sounded, but neither a burglary nor an earthquake has occurred, and both John and Mary call, multiply entries from the joint distribution as

$$P(j,m, a,\neg b,\neg e) = P(j \mid a)P(m \mid a)P(a \mid \neg b \wedge \neg e)P(\neg b)P(\neg e)$$
$$= 0.90 \times 0.70 \times 0.001 \times 0.999 \times 0.998 = 0.000628$$

- So joint distribution is represented by the product of the appropriate elements of the conditional probability tables (CPTs) in the Bayesian network
-

Method for constructing the Bayesian Network

Conjunctive probability

$$P(x_1, \ldots, x_n) = P(x_n|x_{n-1}, \ldots, x_1)P(x_{n-1}, \ldots, x_1)$$

$$P(x_1, \ldots, x_n) = P(x_n|x_{n-1}, \ldots, x_1)P(x_{n-1}|x_{n-2}, \ldots, x_1) \cdots P(x_2|x_1)P(x_1)$$

$$= \prod_{i=1}^{n} P(x_i|x_{i-1}, \ldots, x_1)$$

This identity is called the **chain rule.**

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

Method for constructing the Bayesian Network ..contd

The general procedure for incremental network construction is as follows:

1. Choose the set of relevant variables X, that describe the domain.
2. Choose an ordering for the variables.
3. While there are variables left:
   a. Pick a variable X_i and add a node to the network for it.
   b. Set ***Parents(X_i)*** to some minimal set of nodes already in the net such that the conditional independence property is satisfied
   c. Define the conditional probability table for X_i

Method for constructing the Bayesian Network ..contd

- The belief network is a correct representation of the domain only if each node is conditionally independent of its predecessors in the node ordering, given its parents.

- Hence, in order to construct a belief network with the correct structure for the domain, the parents are chosen for each node such that this property holds.

- Intuitively, the parents of node X_i, should contain all those nodes in X_1, ..., X_{i-1} that directly influence X_i.

  For example, suppose for the network in Figure except for the choice of parents for *MaryCalls*.

  ***MaryCalls* is certainly influenced by whether or not there is a *Burglary* or an *Earthquake*, but it is not *directly* influenced.**

- The knowledge of the domain tells that these events only influence Mary's calling behavior through their effect on the alarm.

- Also, given the state of the alarm, whether or not John calls has no influence on Mary's calling. *Hence*

  P(MaryCalls,JohnCalls, Alarm, Earthquake, Burglary) = P(MaryCalls|Alarm)

Method for constructing the Bayesian Network ..contd

- Because each node is only connected to earlier nodes, this construction method guarantees that the network is **acyclic**.
- Another important property of belief networks is that they contain **no redundant probability** values, except perhaps for one entry in each row of each conditional probability table.
- This means that **it is impossible for the knowledge engineer or domain expert to create a belief network that violates the axioms of probability.**

Compactness and node ordering

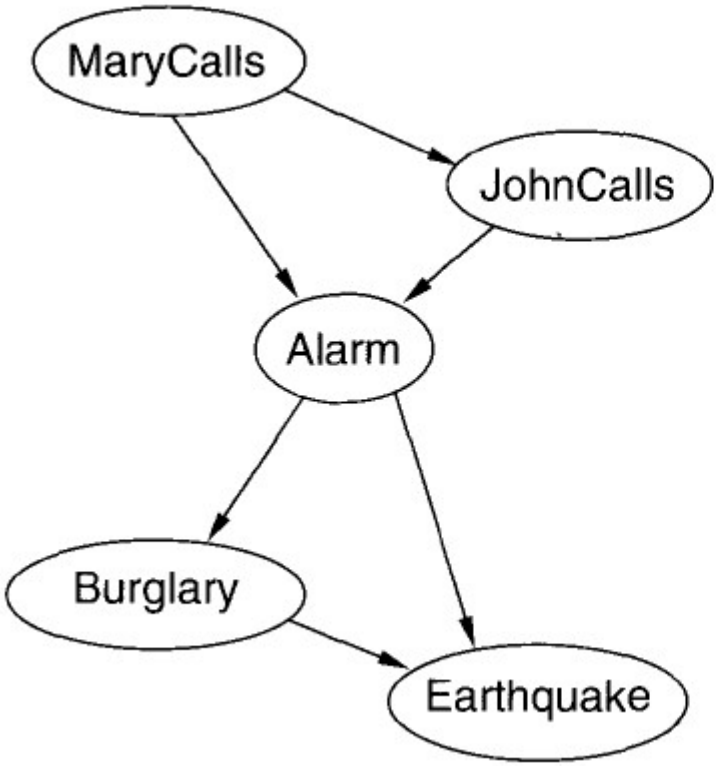
- The compact form the belief network enables the handling the large number of pieces of evidences without the **exponential growth of complexity** (locally structured or sparse)
- In a locally structured system, each subcomponent interacts directly with only a bounded number of other components, regardless of the total number of components.
- Local structure is usually associated with linear rather than exponential growth in complexity.
- If there are **n** variables (assuming **Boolean variables**) and if each variable is directly influenced by the **k** variables, the complexity becomes $n.2^k$.

  For example: Consider a belief network with 20 nodes and each node has maximum of 5 parents, the belief or Bayesian network has $20 * 2^5 = 640$ numbers (in fully joint network has over million numbers)

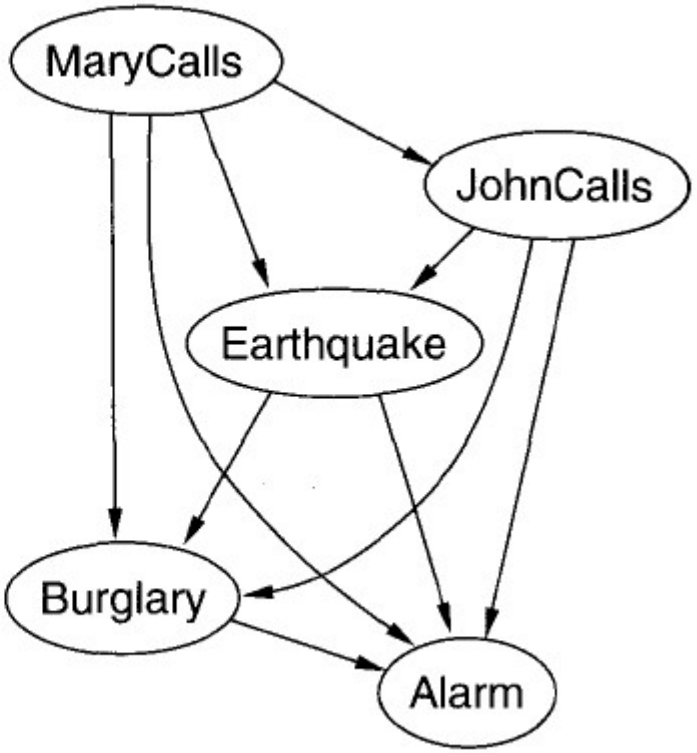
  Exercise: Find out how many number if n=30 (960 !!)

- If the link from Earthquake to JohnCalls and MaryCalls (enlarging the table) depends on the comparing the importance of getting more accurate probabilities with the cost of specifying the extra information (increased complexity)
- *The correct order to add nodes is to add the "**root causes**" first, then the variables they **influence**,* and so on until the leaves are reached

Placing of nodes in wrong order - Example



Case 1



Case 2

Case 1- Order: MaryCalls, JohnCalls, Alarm, Burglary, Earthquake
Case 2 – Order: MaryCalls, JohnCalls, Earthquake , Burglary, Alarm (requires 31 distinct probabilities to be specified

Conditional Independence relations to Bayesian Networks

- Topological semantics that specifies the conditional independence relationships encoded by the graph structure and from these the numerical semantics can be derived.
- The topological semantics is given by either of the following two specifications
  1. A node is **conditionally independent** of its **non-descendants**, given its parents. For example, JohnCalls is independent of Burglary and Earthquake, given the value of Alarm
  2. A node is **conditionally independent** of **all other nodes** in the network, given its parents, children, and children's parents (**Markov Blanket**). For example, Burglary is independent of JohnCalls and MaryCalls, given Alarm and Earthquake

A general criterion that "**A set of nodes X is independent of another set of nodes Y given a third set Z** " is called **d-seperation**

EFFICIENT REPRESENTATION OF CONDITIONAL DISTRIBUTIONS

- Uncertain relationships can often be characterized by noisy logical relationships. The standard example is the noisy-OR relation, which is a generalization of the logical OR.
- The noisy-OR model allows for uncertainty about the ability of each parent to cause the child to be true.
- The causal relationship between parent and child may be inhibited.
- The model makes two assumptions.
  1. First, it assumes that all the possible causes are listed. (If some are missing, we can always add a leak node that covers "miscellaneous causes.").
  2. Second, it assumes that inhibition of each parent is independent of inhibition of any other parents

EFFICIENT REPRESENTATION OF CONDITIONAL DISTRIBUTIONS

- For example, let fever be the child for the parents cold, flu and malaria.
- Fever is false if and only if all its true parents are inhibited.
- The probability of child is the product of the inhibition probabilities q for each parent.
- Let us suppose the individual inhibition probabilities are as follows:

  q_{cold} = P(¬fever | cold,¬flu,¬malaria) = 0.6 ,
  q_{flu} = P(¬fever | ¬cold, flu,¬malaria) = 0.2 ,
  $q_{malaria}$ = P(¬fever | ¬cold,¬flu, malaria) = 0.1 .
- From this information and the noisy-OR assumptions, the entire CPT can be built using the formula:

$$P(x_i \mid parents(X_i)) = 1 - \prod_{\{j:X_j = true\}} q_j$$

- The product is taken over the parents that are set to true for that row of the CPT

EFFICIENT REPRESENTATION OF CONDITIONAL DISTRIBUTIONS

$Cold$	Flu	$Malaria$	$P(Fever)$	$P(\neg Fever)$
F	F	F	0.0	1.0
F	F	T	0.9	**0.1**
F	T	F	0.8	**0.2**
F	T	T	0.98	$0.02 = 0.2 \times 0.1$
T	F	F	0.4	**0.6**
T	F	T	0.94	$0.06 = 0.6 \times 0.1$
T	T	F	0.88	$0.12 = 0.6 \times 0.2$
T	T	T	0.988	$0.012 = 0.6 \times 0.2 \times 0.1$

Bayesian nets with continuous variables

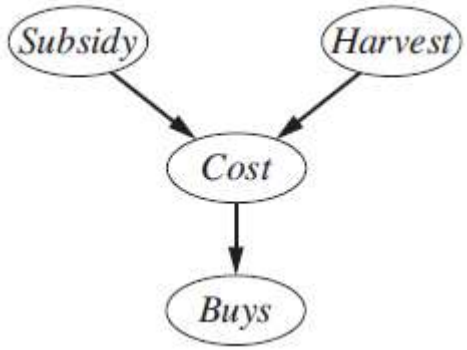
- Many real-world problems involve continuous quantities, such as height, mass, temperature, and money; in fact, much of statistics deals with random variables whose domains are continuous.
- Continuous variables have an infinite number of possible values, so it is impossible to specify conditional probabilities explicitly for each value.
- One method to avoid is discretization which is dividing up the possible values into a fixed set of intervals.
- Discretization often results in a considerable loss of accuracy and very large CPTs.
- One common method is to define standard families of probability density functions that are specified by a finite number of parameters.
- For example, a Gaussian (or normal) distribution $N(\mu, \sigma^2)(x)$ has the mean μ and the variance σ^2 as parameters.

Bayesian nets with continuous variables

- One another method is called a nonparametric representation.
- This is to define the conditional distribution implicitly with a collection of instances, each containing specific values of the parent and child variables.
- A network with both discrete and continuous variables is called a hybrid Bayesian network.
- In a hybrid network, two new kinds of distributions are used:
  1. The conditional distribution for a continuous variable given discrete or continuous parents;
  2. The conditional distribution for a discrete variable given continuous parents

Hybrid Bayesian Network

- Consider the example in which a customer buys some fruit depending on its cost,which depends in turn on the size of the harvest and whether the government's subsidy scheme is operating.



- The variable Cost is continuous and has continuous (Harvest) and discrete(Subsidy) parents;
- The variable Buys is discrete and has a continuous parent

Hybrid Bayesian Network

- For the Cost variable, we need to specify P(Cost | Harvest, Subsidy).
- Here the discrete parent is handled by enumeration—that is, by specifying both P(Cost | Harvest, subsidy) and P(Cost | Harvest, ¬subsidy).
- To handle Harvest, we specify how the distribution over the cost c depends on the continuous value h of Harvest. I.e. we specify the parameters of the cost distribution as a function of h.
- We can use Linear Gaussian distribution, in which the child has a Gaussian distribution whose mean μ varies linearly with the value of the parent and whose standard deviation σ is fixed.
- We need two distributions, one for subsidy and one for ¬subsidy, with different parameters

$$P(c \mid h, subsidy) = N(a_t h + b_t, \sigma_t^2)(c) = \frac{1}{\sigma_t \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{c-(a_t h+b_t)}{\sigma_t}\right)^2}$$

$$P(c \mid h, \neg subsidy) = N(a_f h + b_f, \sigma_f^2)(c) = \frac{1}{\sigma_f \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{c-(a_f h+b_f)}{\sigma_f}\right)^2}$$

Hybrid Bayesian Network

- For the distributions of discrete variables with continuous parents.
- Consider, for example, the Buys node.
- It seems reasonable to assume that the customer will buy if the cost is low and will not buy if it is high and that the probability of buying varies smoothly in some intermediate region.
- In other words, the conditional distribution is like a “soft” threshold function.
- One way to make soft thresholds is to use the integral of the standard normal distribution:

$$\Phi(x) = \int_{-\infty}^{x} N(0,1)(x)dx$$

- The probability of Buys given Cost is

$$P(buys \mid Cost = c) = \Phi((-c + \mu)/\sigma)$$

- This is known as Probit Distribution

Hybrid Bayesian Network

- Another method is the logit distribution , It uses the logistic function $1/(1 + e^{-x})$ to produce a soft threshold

$$P(buys \mid Cost = c) = \frac{1}{1 + exp(-2\frac{-c+\mu}{\sigma})}$$