Artificial Intelligence

Bayesian Networks

Christopher Simpkins

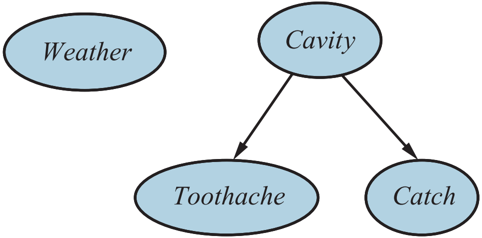
Kennesaw State University

Representation of Uncertain Knowledge

A Bayesian network is a directed graph in which each node is annotated with quantitative probability information. The full specification is as follows:

1. Each node corresponds to a random variable, which may be discrete or continuous.
2. Directed links or arrows connect pairs of nodes. If there is an arrow from node X to node Y, then X is said to be a parent of Y. The graph has no directed cycles and hence is a directed acyclic graph, or DAG.
3. Each node X_i has associated probability information $\theta(X_i|Parents(X_i))$ that quantifies the effect of the parents on the node using a finite number of parameters.

Bayesian Network Topology



- The topology of the network – the set of nodes and links – specifies the conditional independence relationships that hold in the domain.
  - $Toothache$ and $Catch$ are conditionally independent given $Cavity$.
- Intuitiveley, and arrow $X \to Y$ means X has a direct influence on Y – so parents should be *causes* of effects.

It is usually easy for a domain expert to decide what direct influences exist in the domain – much easier, in fact, than actually specifying the probabilities themselves.

Example: Earthquake vs. Burglary

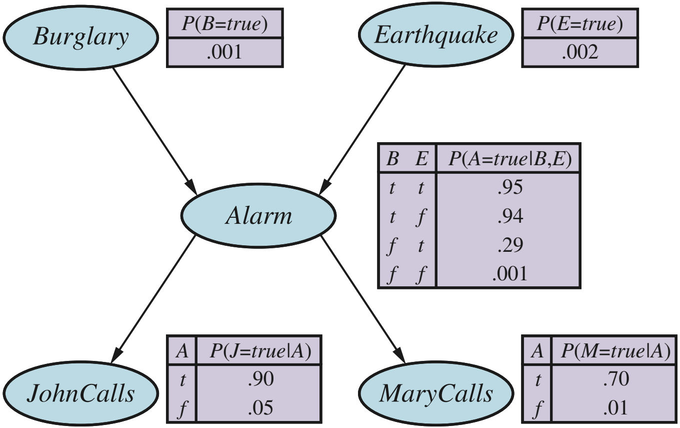
You have a burglar alarm fairly reliable at detecting a burglary, but is occasionally set off by minor earthquakes.

- Two neighbors, John and Mary, who promise to call when they hear the alarm.
- John nearly always calls when he hears the alarm, but sometimes confuses the telephone ringing with the alarm and calls then, too.
- Mary, on the other hand, likes rather loud music and often misses the alarm altogether.

Given the evidence of who has or has not called, we would like to estimate the probability of a burglary.

Bayes Net for Earthquake vs. Burglary Reasoning

The *syntax* of a Bayes net consists of a directed acyclic graph (DAG) with some local probability information attached to each node.



P(B=true)
.001

P(E=true)
.002

B	E	P(A=true|B,E)
t	t	.95
t	f	.94
f	t	.29
f	f	.001

A	P(J=true|A)
t	.90
f	.05

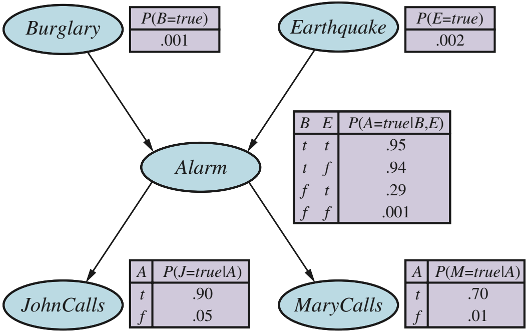
A	P(M=true|A)
t	.70
f	.01

The full joint distribution for all the variables is defined by the topology and the local probability information recorded in conditional probability tables (CPTs).

Conditional Probability Tables

Each row in a CPT contains the conditional probability of each node value for a conditioning case. A conditioning case is a possible combination of values for the parent nodes – a miniature possible world.

- Each row must sum to 1, because the entries represent an exhaustive set of cases for the variable.
- For Boolean variables, once you know the probability of $true$ is p, the probability of $false$ must be $1 - p$, so we often omit the second number.
- In general, a table for a Boolean variable with k Boolean parents contains 2^k independently specifiable probabilities.
- A node with no parents has only one row, representing the prior probabilities of each possible value of the variable.



P(B=true)
.001

P(E=true)
.002

B	E	P(A=true|B,E)
t	t	.95
t	f	.94
f	t	.29
f	f	.001

A	P(J=true|A)
t	.90
f	.05

A	P(M=true|A)
t	.70
f	.01

Many Possibilities, Few Variables of Interest

Network does not explicitly represent Mary currently listening to loud music or telephone ringing and confusing John.

- These factors are summarized in the uncertainty associated with the links from $Alarm$ to $JohnCalls$ and $MaryCalls$.
- This shows both laziness and ignorance in operation: a lot of work to find out the likelihood of those factors, and we have no reasonable way to obtain the relevant information anyway.

The probabilities actually summarize a potentially infinite set of circumstances:

- The alarm might fail to sound (humidity, power failure, dead battery, cut wires, a dead mouse stuck in the bell, etc.) or
- John or Mary might fail to call and report it (out to lunch, on vacation, temporarily deaf, passing helicopter, etc.).

In this way, a small agent can cope with a very large world, at least approximately.

Semantics of Bayesian Networks

The *semantics* defines how the syntax – a DAG with local probabilities – corresponds to a joint distribution over the variables of the network.

A Bayes net contains:

- n variables, $X_1, \ldots, X_n$, and
- (implicit) joint distributions $Pr(X_1 = x_1 \wedge \cdots \wedge X_n = x_n)$, or $Pr(x_1, \ldots, x_n)$.

Each entry in the joint distribution is defined by:

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

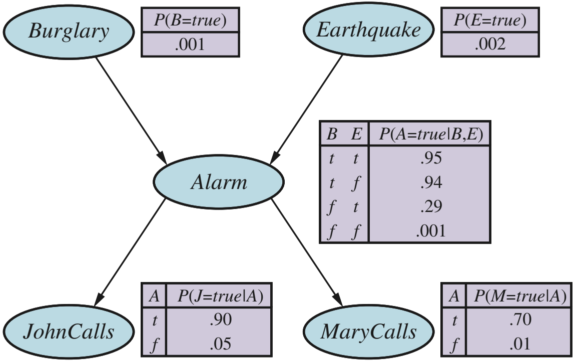
where $parents(X_i)$ denotes the values of $Parents(X_i)$ that appear in $x_1, dots, x_n$.

So each entry in the joint distribution is the product of appropriate elements of the local CPTs in the Bayes net.

Example: An Alarm, Two Calls, but No Events of Interest

What is the probability that John and Mary call, but no earthquake or burglary occur?

- Take each entry i in each CPT θ to mean $Pr(x_1|parents(X_i))$
  - The entries in the CPTs must be accurate conditional probabilities for the variables given their parents for the Bayes net to be useful in performing probabilistic inference.
- Use those values in the calculation of the joint probability using Using the abbreviations j, m, a, b and e:



Burglary

P(B=true)
.001

Earthquake

P(E=true)
.002

Alarm

B	E	P(A=true|B,E)
t	t	.95
t	f	.94
f	t	.29
f	f	.001

JohnCalls

A	P(J=true|A)
t	.90
f	.05

MaryCalls

A	P(M=true|A)
t	.70
f	.01

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

$$\begin{aligned} Pr(j, m, a, \neg b, \neg e) &= Pr(j|a)Pr(m|a)Pr(a|\neg b, \neg e)Pr(\neg b)Pr(\neg e) \\ &= 0.90 \times 0.70 \times 0.001 \times 0.99 \times 0.98 \\ &= 0.000628 \end{aligned}$$

Constructing Bayesian Networks

A Bayesian network is a correct representation of the domain only if each node is conditionally independent of its other predecessors in the node ordering, given its parents. Mathematically, if $Parents(X_i) \subseteq \{X_{i-1}, \ldots, X_1\}$, then:

$$Pr(X_i \mid X_{i-1}, \ldots, X_i) = Pr(X_i | Parents(X_i)) \tag{13.3}$$

We can construct a valid Bayes net with this methodology:

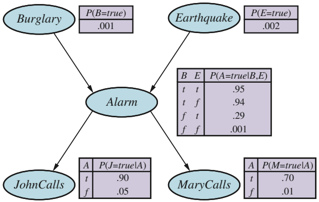
1. Nodes: First determine the set of variables that are required to model the domain. Now order them, $\{X_1, \ldots, X_n\}$. Any order will work, but the resulting network will be more compact if the variables are ordered such that causes precede effects.
   - For the Burglary-Earthquake domain, B, E, A, J, M or E, B, A, M, J work.
2. Links: For $i = 1$ to n do:
   - Choose a minimal set of parents for X_i from $X_1, \ldots, X_{i-1}$, such that Equation (13.3) is satisfied.
   - For each parent insert a link from the parent to X_i.
   - CPTs: Write down the conditional probability table, $P(X_i|Parents(X_i))$.

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

Knowledge Engineering Considerations in Bayes Nets

Suppose we have completed the network except for the choice of parents for $MaryCalls$. We know:

- $MaryCalls$ is influenced by Burglary or Earthquake, but not directly. Our domain knowledge tells us that these events influence Mary's calling behavior only through their effect on the alarm.
- Also, given the state of the alarm, whether John calls has no influence on Mary's calling.



P(B=true)
.001

P(E=true)
.002

B	E	P(A=true|B,E)
t	t	.95
t	f	.94
f	t	.29
f	f	.001

A	P(J=true|A)
t	.90
f	.05

A	P(M=true|A)
t	.70
f	.01

Formally, we believe that the following conditional independence statement holds:

$$Pr(MaryCalls \mid JohnCalls, Alarm, Earthquake, Burglary) = Pr(MaryCalls \mid Alarm).$$

Thus, Alarm will be the only parent node for MaryCalls.

Because

- each node is connected only to earlier nodes, this construction method guarantees that the network is acyclic, and
- there are no redundant probability values,

there is no chance for inconsistency: it is impossible for the knowledge engineer or domain expert to create a Bayesian network that violates the axioms of probability.

Compactness of Bayesian Network Models

Bayes nets are complete and nonredundant, but their *compactness* is their “secret sauce.”

- Bayes nets are a kind of **locally structured**, or **sparse** system.
  - In locally structured systems, each component interacts only with a bounded number, k, of other components, where $k \ll n$, the total number of components.
  - Local structure typically results in linear rather than exponential complexity.

Example: assume boolean variables.

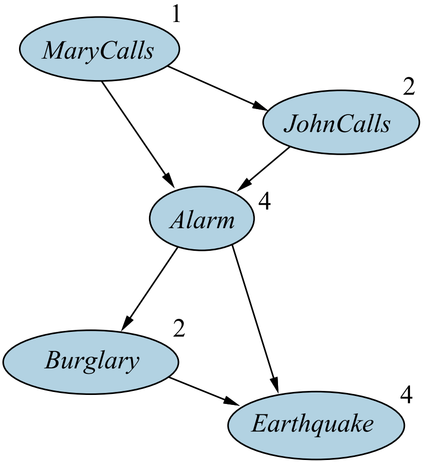
- With n variables, the full joint distribution is 2^n numbers.
  - With 30 variables that's $2^{30} = 1,073,741,824$ numbers.
- With n variables, each of which is influenced by k parents, each CPT is 2^k numbers and a full Bayesian network has $n \cdot 2^k$ numbers.
  - With 30 variables that's $30 \cdot 2^5 =$ 960 numbers.

Effects of Node Ordering

We saw earlier that B, E, A, J, M is a good node ordering because causes come before effects. With this node ordering we get 10 conditional probabilities. What if we choose a different ordering, like M, J, A, B, E?

- Add $MaryCalls$: No parents.
- Add $JohnCalls$: If Mary calls, alarm probably sounded, which increases prob that John calls. So $JohnCalls$ needs $MaryCalls$ as a parent.
- Add $Alarm$: The more calls, the more likely alarm sounded, so $Alarm$ needs $MaryCalls$ and $JohnCalls$ as parents.
- Add $Burglary$: Given knowledge of alarm, calls irrelevant: $Pr(Burglary \mid Alarm, JohnCalls, MaryCalls) = Pr(Burglary \mid Alarm)$. So just $Alarm$ as parent.
- Add $Earthquake$: With alarm, earthquake more likely. But burglary explains alarm, and probability of an earthquake only slightly higher than normal. So need both $Alarm$ and $Burglary$ as parents.

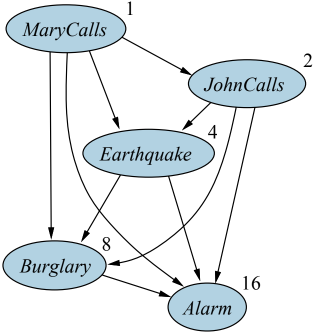
We end up with 13 conditional probabilities.



Knowledge Engineering in Probabilistic Systems

There is still knowledge engineering involved, which includes design choices.

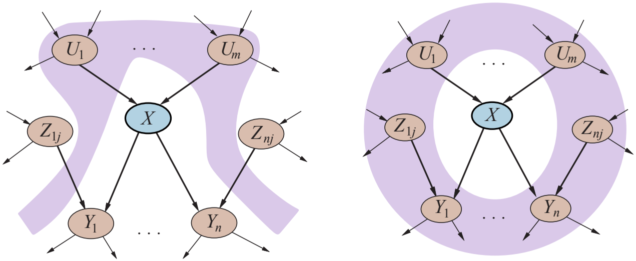
- You may choose to trade a small amount of accuracy for the improved performance of leaving out an influence variable.
  - If there's a large earthquake, John and Mary probably won't call even if they hear the alarm. Could choose to include links from $EarthQuake$ to $JohnCalls$ and $MaryCalls$. This would increase accuracy slightly, but probably bnot worth the extra complexity.
- And as we just saw, node ordering matters, potentially quite a bit.
  - With node ordering M, J, E, B, A the resulting Bayes net requires 31 probabilities – same as the full joint distribution.



Any Bayes net represents the same joint distibution, but some are far more efficient.

> *Key takeaway: stick to causal models. They are easier to specify, easier to get right, and they lead to more efficient Bayes nets.*

Conditional Independence Relations



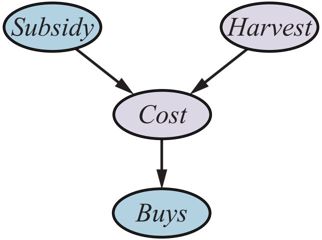
(a) (b)

- (a) Each variable is conditionally independent of its non-descendants, given its parents.
- (b) A variable is conditionally independent of all other nodes in the network, given its parents, children, and children's parents – that is, given its **Markov blanket**.

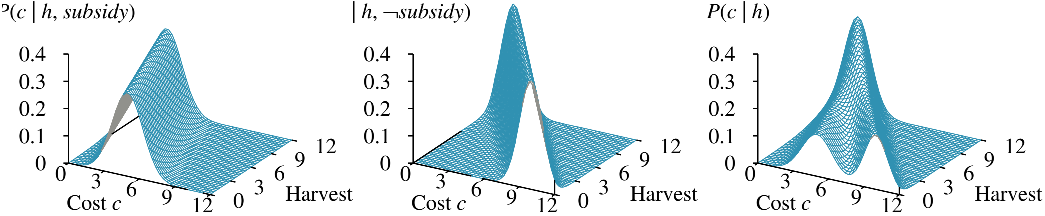
For example, the variable $Burglary$ is independent of $JohnCalls$ and $MaryCalls$, given $Alarm$ and $Earthquake$.

Hybrid Bayesian Networks

Bayesian Networks with Discrete and Continuous Variables

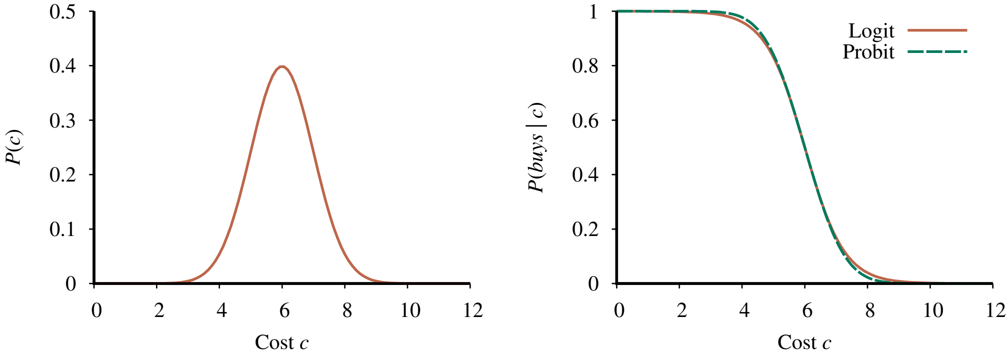


Linear-Gaussian Conditional Distributions



(a) (b) (c)

Soft Thresholding for Continuous Parents



(a)

(b)

Closing Thoughts

- A Bayesian network is a directed acyclic graph whose nodes correspond to random variables; each node has a conditional distribution for the node, given its parents.
- Bayesian networks provide a concise way to represent conditional independence relationships in the domain.
- A Bayesian network specifies a joint probability distribution over its variables. The probability of any given assignment to all the variables is defined as the product of the corresponding entries in the local conditional distributions. A Bayesian network is often exponentially smaller than an explicitly enumerated joint distribution.
- Many conditional distributions can be represented compactly by canonical families of distributions. Hybrid Bayesian networks, which include both discrete and continuous variables, use a variety of canonical distributions.