Bayesian Networks

Chapter 14

Section 1, 2, 4

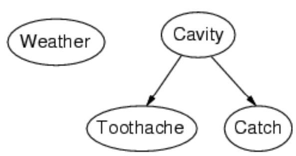
Bayesian networks

- 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")
  - if there is a link from x to y, x is said to be a parent of y
  - 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

- Topology of network encodes conditional independence assertions:

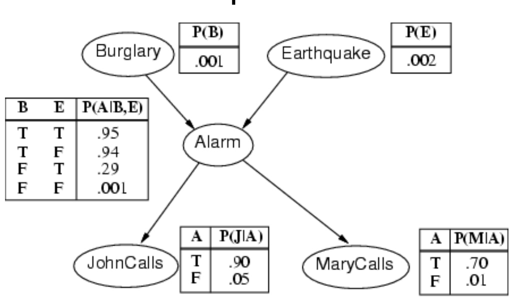


- *Weather* is independent of the other variables
- *Toothache* and *Catch* are conditionally independent given *Cavity*

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

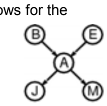
B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

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

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

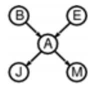
Compactness

- 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:

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

Thus each entry in the joint distribution is represented by the product of the appropriate elements of the conditional probability tables in the Bayesian network.

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

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

$= 0.90 * 0.70 * 0.001 * 0.999 * 0.998 = 0.00062$

Back to the dentist example ...

- We now represent the world of the dentist D using three propositions – Cavity, Toothache, and PCatch
- D's belief state consists of 2^3 = 8 states each with some probability:
  {cavity^toothache^pcatch,
  ¬cavity^toothache^pcatch,
  cavity^ ¬toothache^pcatch,...}

The belief state is defined by the full joint probability of the propositions

	toothache		¬toothache	
	pcatch	¬pcatch	pcatch	¬ pcatch
cavity	0.108	0.012	0.072	0.008
¬ cavity	0.016	0.064	0.144	0.576

Probabilistic Inference

	toothache		¬toothache	
	pcatch	¬ pcatch	pcatch	¬ pcatch
cavity	0.108	0.012	0.072	0.008
¬ cavity	0.016	0.064	0.144	0.576

P(cavity v toothache) = 0.108 + 0.012 + ...
= 0.28

Probabilistic Inference

	toothache		¬toothache	
	pcatch	¬ pcatch	pcatch	¬ pcatch
cavity	0.108	0.012	0.072	0.008
¬ cavity	0.016	0.064	0.144	0.576

P(cavity) = 0.108 + 0.012 + 0.072 + 0.008
= 0.2

Probabilistic Inference

	toothache		¬toothache	
	pcatch	¬ pcatch	pcatch	¬ pcatch
cavity	0.108	0.012	0.072	0.008
¬ cavity	0.016	0.064	0.144	0.576

Marginalization: $P(c) = \Sigma_t \Sigma_{pc} P(c \wedge t \wedge pc)$
using the conventions that c = cavity or ¬ cavity and that Σ_t is the sum over t = {toothache, ¬ toothache}

Conditional Probability

- P(A^B) = P(A|B) P(B)
  = P(B|A) P(A)
  P(A|B) is the posterior probability of A given B

	toothache		¬toothache	
	pcatch	¬ pcatch	pcatch	¬ pcatch
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

P(cavity|toothache) = P(cavity^toothache)/P(toothache)
= (0.108+0.012)/(0.108+0.012+0.016+0.064) = 0.6

Interpretation: After observing Toothache, the patient is no longer an "average" one, and the prior probabilities of Cavity is no longer valid

P(cavity|toothache) is calculated by keeping the ratios of the probabilities of the 4 cases unchanged, and normalizing their sum to 1

	toothache		¬toothache	
	pcatch	¬ pcatch	pcatch	¬ pcatch
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

P(cavity|toothache) = P(cavity^toothache)/P(toothache)
= (0.108+0.012)/(0.108+0.012+0.016+0.064) = 0.6
P(¬ cavity|toothache)=P(¬ cavity^toothache)/P(toothache)
= (0.016+0.064)/(0.108+0.012+0.016+0.064) = 0.4

P(C|toochache) = α P(C ^ toothache)
= α Σ_{pc} P(C ^ toothache ^ pc)
= α [(0.108, 0.016) + (0.012, 0.064)]
= α (0.12, 0.08) = (0.6, 0.4)

normalization constant

Conditional Probability

- P(A^B) = P(A|B) P(B)
  = P(B|A) P(A)
- P(A^B^C) = P(A|B,C) P(B^C)
  = P(A|B,C) P(B|C) P(C)
- P(Cavity) = $\Sigma_t \Sigma_{pc}$ P(Cavity^t^pc)
  = $\Sigma_t \Sigma_{pc}$ P(Cavity|t,pc) P(t^pc)
- P(c) = $\Sigma_t \Sigma_{pc}$ P(c^t^pc)
  = $\Sigma_t \Sigma_{pc}$ P(c|t,pc)P(t^pc)

Independence

- Two random variables A and B are independent if
  P(A^B) = P(A) P(B)
  hence if P(A|B) = P(A)
- Two random variables A and B are independent given C, if
  P(A^B|C) = P(A|C) P(B|C)
  hence if P(A|B,C) = P(A|C)

Issues

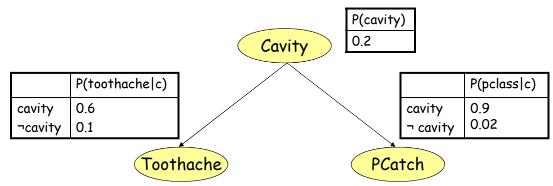
- If a state is described by n propositions, then a belief state contains 2^n states (possibly, some have probability 0)
- → Modeling difficulty: many numbers must be entered in the first place
- → Computational issue: memory size and time

	toothache		¬ toothache	
	pcatch	¬pcatch	pcatch	¬ pcatch
cavity	0.108	0.012	0.072	0.008
¬ cavity	0.016	0.064	0.144	0.576

- toothache and pcatch are independent given cavity (or ¬ cavity), but this relation is hidden in the numbers ! [Verify this]
- Bayesian networks explicitly represent independence among propositions to reduce the number of probabilities defining a belief state

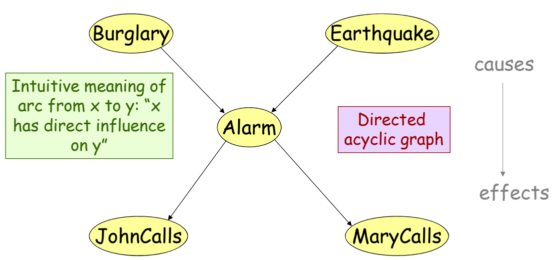
Bayesian Network

- Notice that Cavity is the "cause" of both Toothache and PCatch, and represent the causality links explicitly
- Give the prior probability distribution of Cavity
- Give the conditional probability tables of Toothache and PCatch

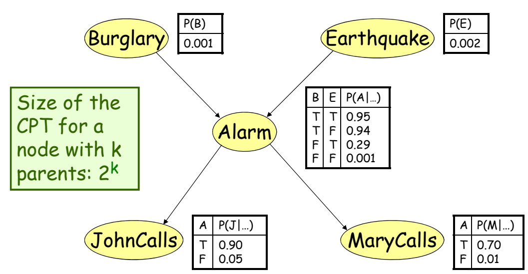


5 probabilities, instead of 7

A More Complex BN

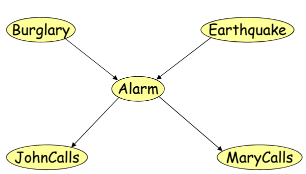


A More Complex BN



10 probabilities, instead of 31

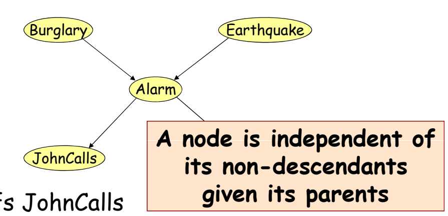
What does the BN encode?



Each of the beliefs JohnCalls and MaryCalls is independent of Burglary and Earthquake given Alarm or ¬Alarm

For example, John does not observe any burglaries directly

What does the BN encode?

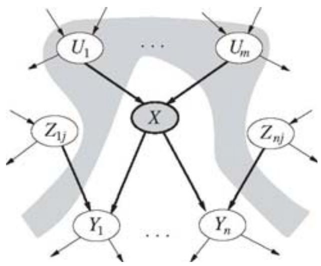


A node is independent of its non-descendants given its parents

The beliefs JohnCalls and MaryCalls are independent given Alarm or ¬Alarm

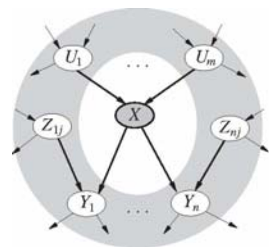
For instance, the reasons why John and Mary may not call if there is an alarm are unrelated

Conditional Independence of non-descendents



A node X is conditionally independent of its non-descendents (e.g., the Zijs) given its parents (the Uis shown in the gray area).

Markov Blanket



A node X is conditionally independent of all other nodes in the network, given its parents, chlidren, and chlidren's parents.

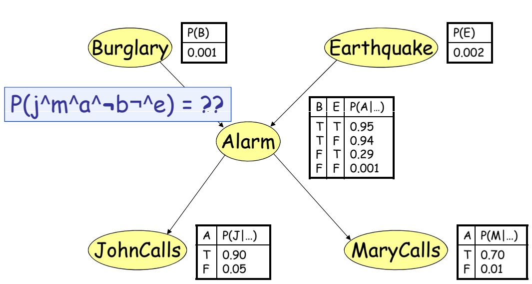
Locally Structured World

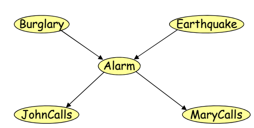
- A world is locally structured (or sparse) if each of its components interacts directly with relatively few other components
- In a sparse world, the CPTs are small and the BN contains many fewer probabilities than the full joint distribution
- If the # of entries in each CPT is bounded, i.e., O(1), then the # of probabilities in a BN is linear in n – the # of propositions – instead of 2^n for the joint distribution

But does a BN represent a belief state?

In other words, can we compute the full joint distribution of the propositions from it?

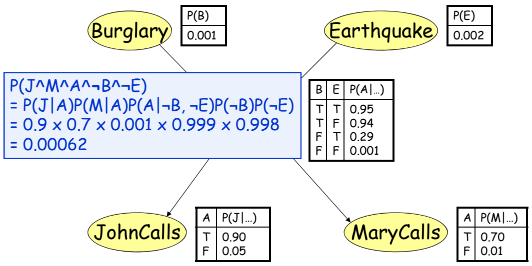
Calculation of Joint Probability



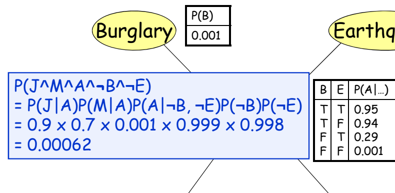


- P(J^M^A^¬B^¬E)
  = P(J^M|A, ¬B, ¬E) * P(A^¬B^¬E)
  = P(J|A, ¬B, ¬E) * P(M|A, ¬B, ¬E) * P(A^¬B^¬E)
  (J and M are independent given A)
- P(J|A, ¬B, ¬E) = P(J|A)
  (J and ¬B^¬E are independent given A)
- P(M|A, ¬B, ¬E) = P(M|A)
- P(A^¬B^¬E) = P(A|¬B, ¬E) * P(¬B|¬E) * P(¬E)
  = P(A|¬B, ¬E) * P(¬B) * P(¬E)
  (¬B and ¬E are independent)
- P(J^M^A^¬B^¬E) = P(J|A)P(M|A)P(A|¬B, ¬E)P(¬B)P(¬E)

Calculation of Joint Probability



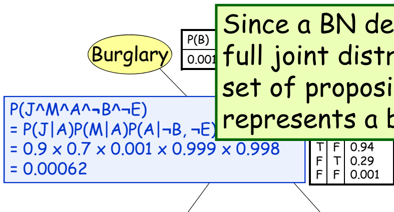
Calculation of Joint Probability



$P(x_1 \wedge x_2 \wedge \ldots \wedge x_n) = \Pi_{i=1,\ldots,n} P(x_i | parents(X_i))$

→ full joint distribution table

Calculation of Joint Probability

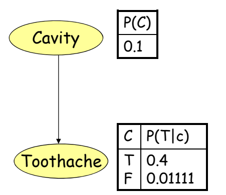


Since a BN defines the full joint distribution of a set of propositions, it represents a belief state

$P(x_1 \wedge x_2 \wedge \ldots \wedge x_n) = \Pi_{i=1,\ldots,n} P(x_i | parents(X_i))$

→ full joint distribution table

Querying the BN



- The BN gives P(t|c)
- What about P(c|t)?
- P(cavity|t)
  = P(cavity ^ t)/P(t)
  = P(t|cavity) P(cavity) / P(t)
  [Bayes' rule]
- P(c|t) = α P(t|c) P(c)
- *Querying a BN is just applying the trivial Bayes' rule on a larger scale*

Exact Inference in Bayesian Networks

- Let's generalize that last example a little – suppose we are given that JohnCalls and MaryCalls are both true, what is the probability distribution for Burglary?
- P(Burglary | JohnCalls = true, MaryCalls=true)
- Look back at using full joint distribution for this purpose – summing over hidden variables.

Inference by enumeration (example in the text book) – figure 14.8

$\mathbf{P}(X \mid \mathbf{e}) = \alpha\, \mathbf{P}(X, \mathbf{e}) = \alpha \sum_y \mathbf{P}(X, \mathbf{e}, \mathbf{y})$

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

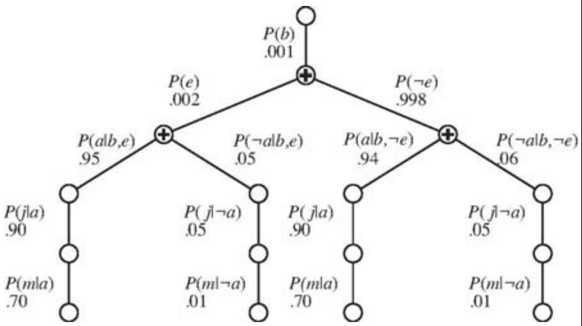
$P(b| j,m) = \alpha \sum_e \sum_a P(b)P(e)P(a|be)P(j|a)P(m|a)$

$P(b| j,m) = \alpha P(b)\sum_e P(e)\sum_a P(a|be)P(j|a)P(m|a)$

$\mathbf{P}(B| j,m) = \alpha$ <0.00059224, 0.0014919>

$\mathbf{P}(B| j,m) \approx$ <0.284, 0.716>

Enumeration-Tree Calculation



Inference by enumeration (another way of looking at it) – figure 14.8

P(X | **e**) = α **P** (X, **e**) = α $\sum_y$ **P**(X, **e**, **y**)

P(B| j,m) = α**P**(B,j,m) = α $\sum_e \sum_a$ **P**(B,e,a,j,m)

P(b| j,m) = **P**(B,e,a,j,m) +
P(B,e,¬a,j,m) +
P(B,¬e,a,j,m) +
P(B,¬e,¬a,j,m)

P(B| j,m) = α <0.00059224, 0.0014919>

P(B| j,m) ≈ <0.284, 0.716>

Constructing Bayesian networks

- 1. Choose an ordering of variables $X_1, \ldots, X_n$ *such that root causes are first in the order, then the variables that they influence, and so forth.*
- 2. For *i* = 1 to *n*
  - add X_i to the network
  - select parents from $X_1, \ldots, X_{i-1}$ such that
    $\mathbf{P}(X_i \mid Parents(X_i)) = \mathbf{P}(X_i \mid X_1, \ldots X_{i-1})$
  - *Note:the parents of a node are all of the nodes that influence it. In this way, each node is conditionally independent of its predecessors in the order, given its parents.*

This choice of parents guarantees:

$\mathbf{P}(X_1, \ldots, X_n) = \pi_{i=1} \mathbf{P}(X_i \mid X_1, \ldots, X_{i-1})$ (chain rule)
$= \pi_{i=1} \mathbf{P}(X_i \mid Parents(X_i))$ (by construction)

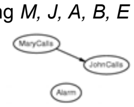
Example – How important is the ordering?

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

P*(J | M) =* ***P****(J)?*

Example

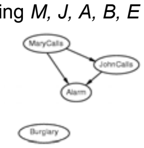
- Suppose we choose the ordering *M, J, A, B, E*



P*(J | M) =* ***P****(J)?* **No**
P*(A | J, M) =* ***P****(A | J)*? ***P****(A | J, M) =* ***P****(A)*?

Example

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



P*(J | M) =* ***P****(J)?* **No**
P*(A | J, M) =* ***P****(A | J)*? ***P****(A | J, M) =* ***P****(A)*? **No**
P*(B | A, J, M) =* ***P****(B | A)*?
P*(B | A, J, M) =* ***P****(B)*?

Example

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

***P**(J | M) = **P**(J)?* **No**
***P**(A | J, M) = **P**(A | J)*? ***P**(A | J, M) = **P**(A)*? **No**
***P**(B | A, J, M) = **P**(B | A)*? **Yes**
***P**(B | A, J, M) = **P**(B)*? **No**
***P**(E | B, A ,J, M) = **P**(E | A)*?
***P**(E | B, A, J, M) = **P**(E | A, B)*?

Example

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

***P**(J | M) = **P**(J)?* **No**
***P**(A | J, M) = **P**(A | J)*? ***P**(A | J, M) = **P**(A)*? **No**
***P**(B | A, J, M) = **P**(B | A)*? **Yes**
***P**(B | A, J, M) = **P**(B)*? **No**
***P**(E | B, A ,J, M) = **P**(E | A)*? **No**
***P**(E | B, A, J, M) = **P**(E | A, B)*? **Yes**

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

Summary

- Bayesian networks provide a natural representation for (causally induced) conditional independence
- Topology + CPTs = compact representation of joint distribution
- Generally easy for domain experts to construct