CS 161 Fundamentals of Artificial Intelligence
Lecture 12

Bayesian Networks

Quanquan Gu

Department of Computer Science
UCLA

March 14, 2023

Teaching/Course Evaluation

- The online evaluations will close 8:00 AM Saturday, March 18.
- Your feedbacks are valuable and more than welcome!
- Students who have done the online evaluation will receive 1 extra point in the final grade:)

Outline

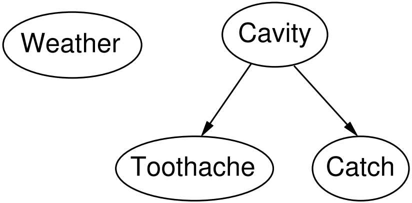
- Syntax
- Semantics
- Parameterized distributions

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 $\approx$ "directly influences")
  - a conditional distribution for each node given its parents:
    $\mathbf{P}(X_i|Parents(X_i))$
- In the simplest case, conditional distribution is represented by 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:

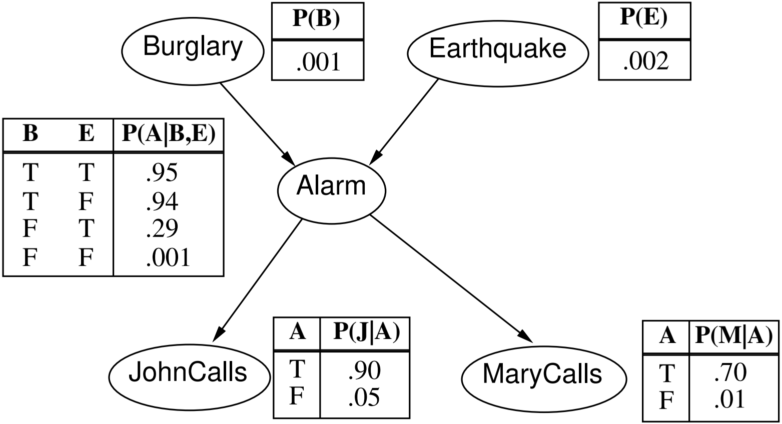


- $Weather$ is independent of the other variables
- $Toothache$ and $Catch$ are conditionally independent given $Cavity$

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: $Burglar$, $Earthquake$, $Alarm$, $JohnCalls$, $MaryCalls$
- 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

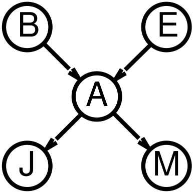
P(E)
.002

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

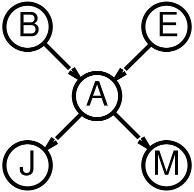
A	P(M|A)
T	.70
F	.01

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. $2^5 - 1 = 31$)

Global semantics



Global semantics defines the full joint distribution as the product of the local conditional distributions:

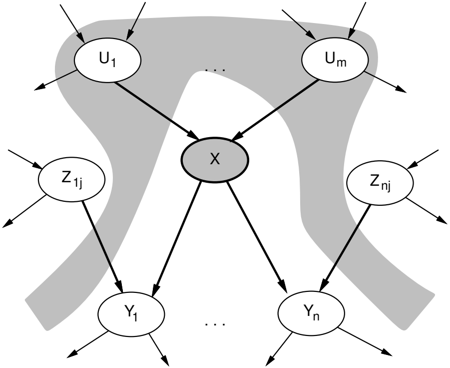
$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | parents(X_i))$$

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

$$\begin{aligned} &= P(j|a)P(m|a)P(a|\neg b, \neg e)P(\neg b)P(\neg e) \\ &= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998 \\ &\approx 0.00063 \end{aligned}$$

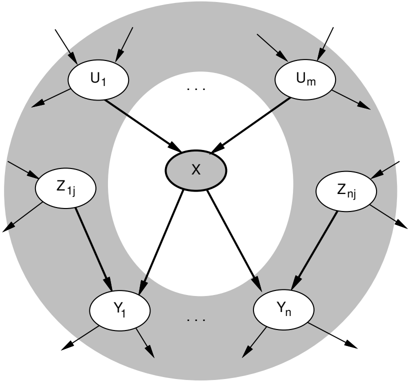
Local semantics

Local semantics (a.k.a., non-descendant property, local Markov property): each node is conditionally independent of its nondescendants given its parents



Markov blanket

Each node is conditionally independent of all others given its **Markov blanket**: parents + children + children's parents



Constructing Bayesian networks

• Need a method such that a series of locally testable assertions of conditional independence guarantees the required global semantics

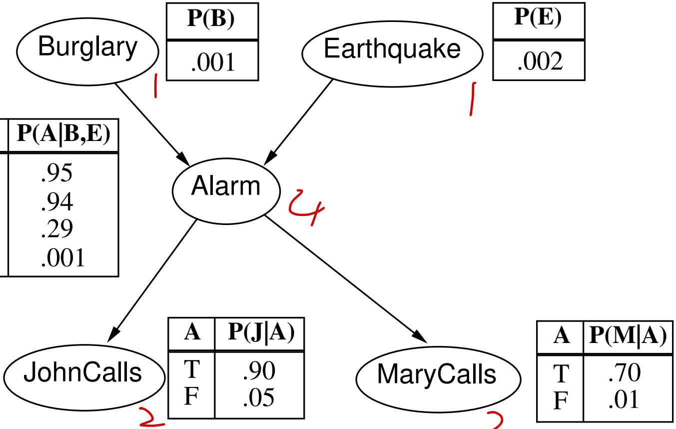
1. Choose an ordering of variables $X_1, \ldots, X_n$
2. For $i = 1$ to n
   add X_i to the network
   select parents from $X_1, \ldots, X_{i-1}$ such that
   $\mathbf{P}(X_i|Parents(X_i)) = \mathbf{P}(X_i|X_1,\ \ldots,\ X_{i-1})$

• This choice of parents guarantees the global semantics:

$$\begin{aligned}\mathbf{P}(X_1, \ldots, X_n) &= \prod_{i=1}^{n} \mathbf{P}(X_i|X_1,\ \ldots,\ X_{i-1}) \quad \text{(chain rule)} \\ &= \prod_{i=1}^{n} \mathbf{P}(X_i|Parents(X_i)) \quad \text{(by construction)}\end{aligned}$$

• The complexity of the network depends on the ordering of variables!

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

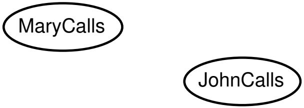
A	P(M|A)
T	.70
F	.01

Let's see what will happen if the variable ordering is no good enough...

1+1+4+2+2 = 10 independant parameters.

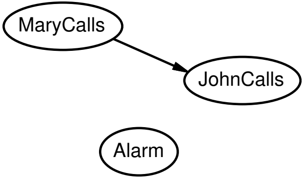
Example

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



Example

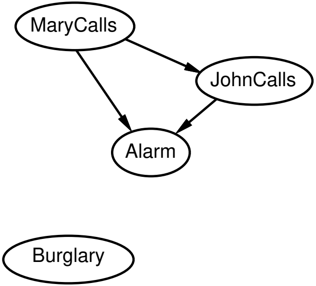
Suppose we choose the ordering M, J, A, B, E



$P(J|M) = P(J)$?? No! If Mary calls, that probably means the alarm has gone off, which of course would make it more likely that John calls. Therefore, JohnCalls needs MaryCalls as a parent.

Example

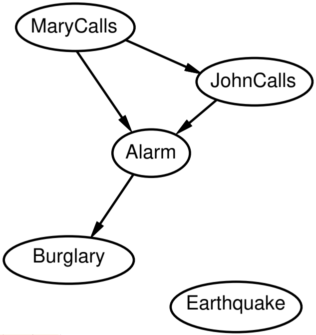
Suppose we choose the ordering M, J, A, B, E



$P(A|J,M) = P(A|J)$?? $P(A|J,M) = P(A)$? No! if both John and Mary call, it is more likely that the alarm has gone off than if just one or neither calls, so we need both MaryCalls and JohnCalls as parents.

Example

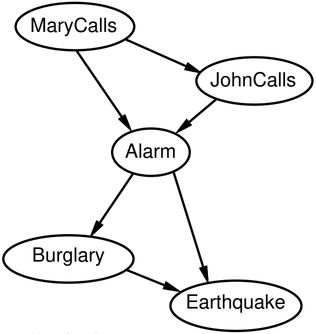
Suppose we choose the ordering M, J, A, B, E



$P(B|A, J, M) = P(B|A)$?? Yes! Alarm or not is the only thing we need to know about burglary!

Example

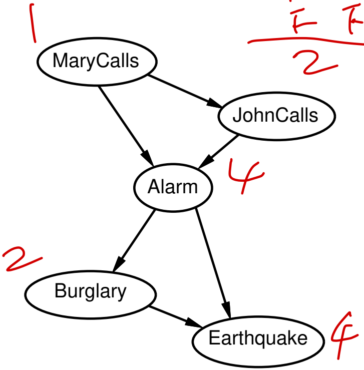
Suppose we choose the ordering M, J, A, B, E



$\underline{P(E|B, A, J, M) = P(E|A)}$?? No!
$\underline{P(E|B, A, J, M) = P(E|A, B)}$?? Yes!
(1) If the alarm is on, it is more likely that there has been an earthquake. (2) But if we know that there has been a burglary, then that explains the alarm, and the probability of an earthquake would be only slightly above normal. (3) Hence, we need both Alarm and Burglary as parents.

Example contd.

M	J	A
T	T	·
T	F	·
F	T	·
F	F	·
2		



- Deciding conditional independence is hard in noncausal directions
- Assessing conditional probabilities is hard in noncausal directions.
- Network is less compact: $1+2+4+2+4=13$ numbers needed

Compact conditional distributions

- CPT grows exponentially with number of parents

Solution: **canonical distributions** that are defined compactly

- **Deterministic** nodes are the simplest case of canonical distribution:

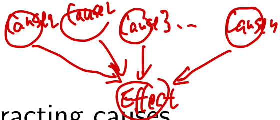
$X = f(Parents(X))$ for some function f, a deterministic node has its value specified exactly by the values of its parents, with no uncertainty.

e.g., Boolean functions

$NorthAmerican \Leftrightarrow Canadian \vee US \vee Mexican$

NorthAmerican = Canadian ∨ US ∨ Mexican
= f(Canadian, US, Mexican)

Compact conditional distributions contd.



- **Noisy-OR** distributions model multiple noninteracting causes

Definitions: We have nodes $x_1, ..., x_n$, each of them is assigned to be true or false

- Assumptions: for any node x_i,

1) Parents of x_i include all causes to x_i (if we miss some, we can add **leak node** that covers "miscellaneous causes.")

2) Independent failure probability q_i for each parent alone

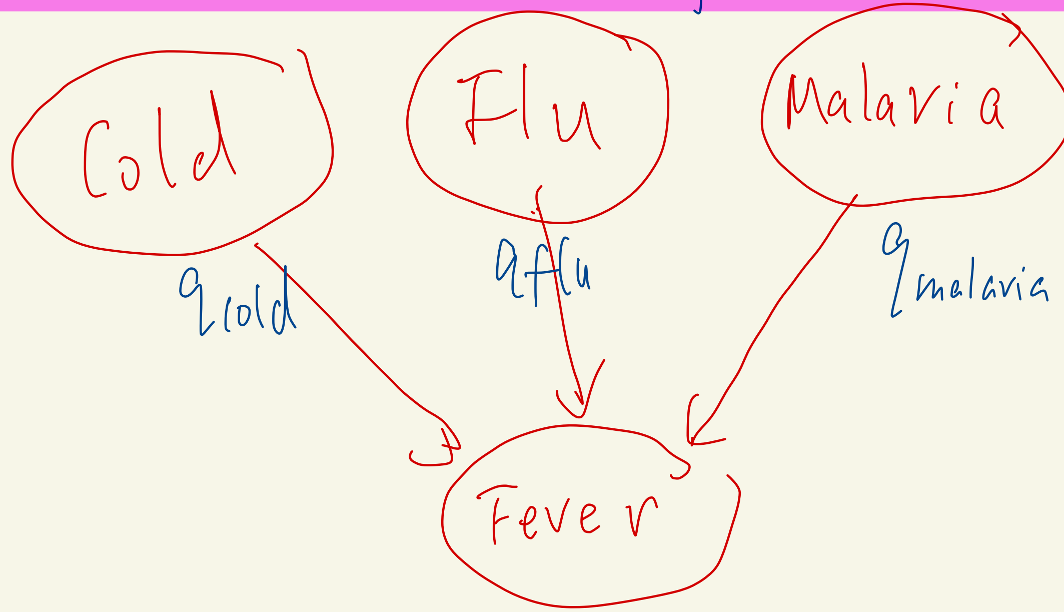
$\Rightarrow P(x_i|parents(X_i)) = 1 - \prod_{j:X_j=true, X_j \in parents(X_i)} q_j$

- Example: we have four nodes $Fever, Cold, Flu, Malaria$, where the parents of $Fever$ are $Cold, Flu, Malaria$. We also have

$$q_{cold} = P(\neg fever|cold, \neg flu, \neg malaria) = 0.6$$
$$q_{flu} = P(\neg fever|\neg cold, flu, \neg malaria) = 0.2$$
$$q_{malaria} = P(\neg fever|\neg cold, \neg flu, malaria) = 0.1$$

Def.

$$\boxed{\begin{aligned} P(X_i \mid \text{Parent}(X_i)) &= 1 - P(\neg X_i \mid \text{Parent}(X_i)) \\ &= 1 - \prod_{X_j \in \text{Parent}(X_i) \wedge X_j = \text{True}} q_j \end{aligned}}$$



Compact conditional distributions contd.

We now can compute the whole CPT table following noisy-or assumptions using only $q_{cold}, q_{flu}, q_{malaria}$:

$Cold$	Flu	$Malaria$	$P(Fever)$	$P(\neg Fever)$
F	F	F	$\mathbf{0.0}$	1.0
F	F	T	0.9	$\mathbf{0.1}$
F	T	F	0.8	$\mathbf{0.2}$
F	T	T	0.98	$0.02 = 0.2 \times 0.1$
T	F	F	0.4	$\mathbf{0.6}$
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	$0.012 = 0.6 \times 0.2 \times 0.1$

Number of parameters **linear** in number of parents

Summary

- Bayesian nets provide a natural representation for (causally induced) conditional independence
- Topology + CPTs = compact representation of joint distribution
- Canonical distributions (e.g., noisy-OR) = compact representation of CPTs

Thank you for taking this class!

Acknowledgment

The slides are adapted from Stuart Russell.