Last update: May 11, 2010

Bayesian networks

CMSC 421: Chapter 14.1–4

Outline

◇ Syntax

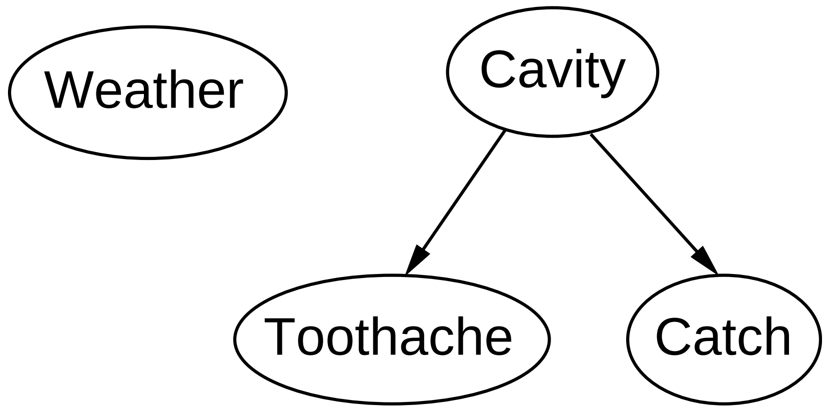
◇ Semantics

◇ Parameterized distributions

Bayesian networks

Graphical network that encodes conditional independence assertions:

- ◇ a set of nodes, one per variable
- ◇ a directed, acyclic graph (link $\approx$ "directly influences")
- ◇ a conditional distribution $\mathbf{P}(X_i|Parents(X_i))$ for each node X_i



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

For each node X_i, $\mathbf{P}(X_i|Parents(X_i))$ is represented as a *conditional probability table* (CPT); we'll have examples later

Example

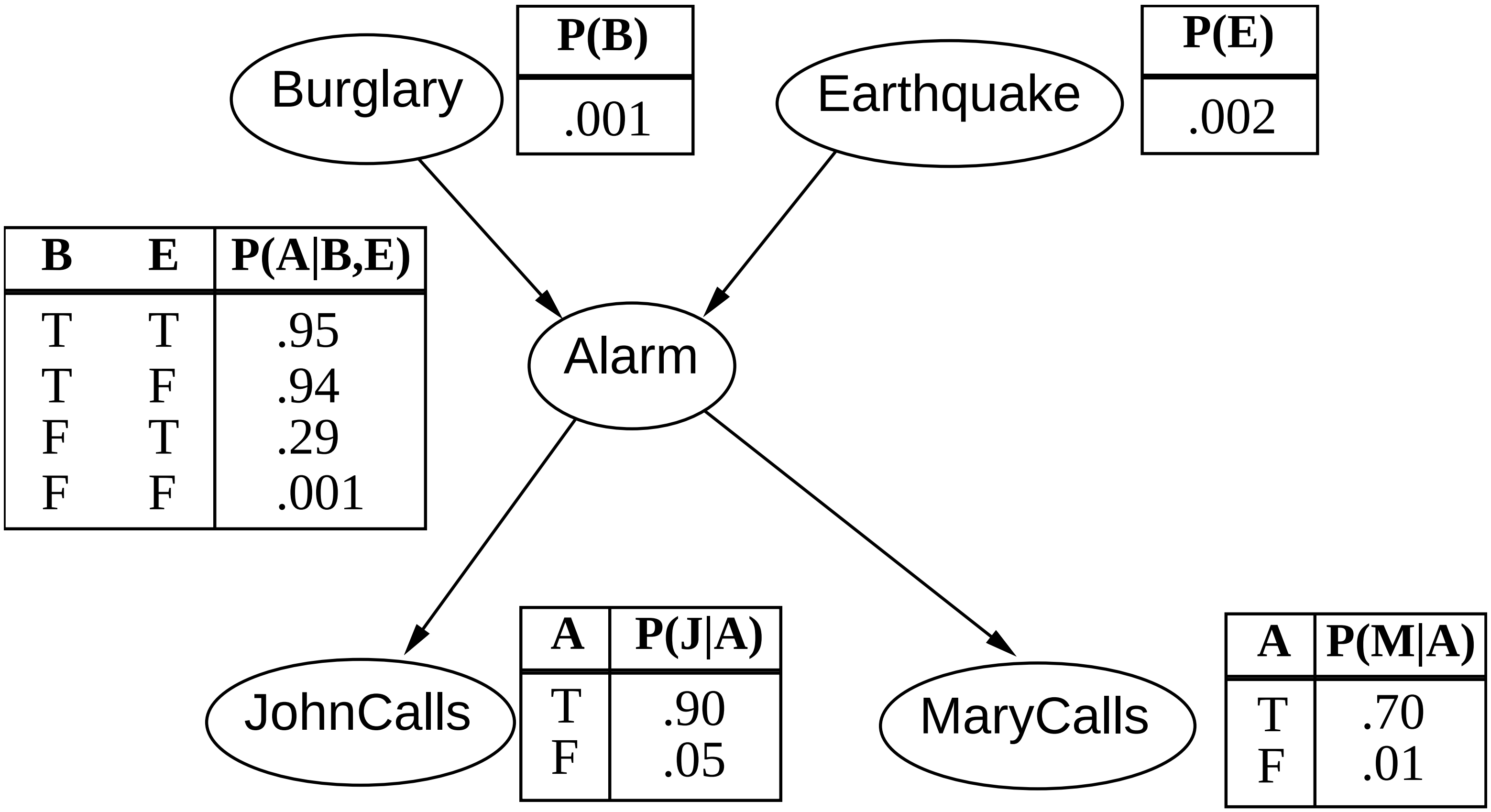
Example from Judea Pearl at UCLA:

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, continued



P(B)
.001

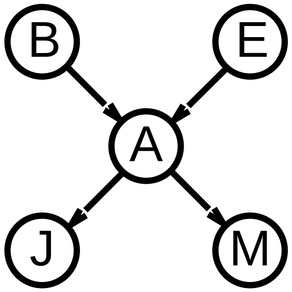
P(E)
.002

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



For a Boolean node X_i with k Boolean parents, the CPT has 2^k rows, one for each combination of parent values

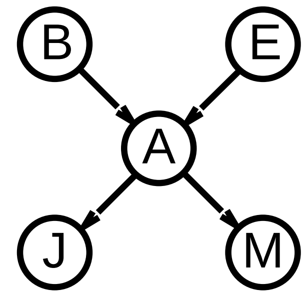
Each row requires one number p for $X_i = true$
(the number for $X_i = false$ is just $1 - p$)

If there are n variables and if
each variable has no more than k parents,
the complete network requires no more than $n \cdot 2^k$ numbers

I.e., grows linearly with n, vs. $O(2^n)$ for the full joint distribution

How many numbers for the burglary net?

Compactness



For a Boolean node X_i with k Boolean parents, the CPT has 2^k rows, one for each combination of parent values

Each row requires one number p for $X_i = true$
(the number for $X_i = false$ is just $1 - p$)

If there are n variables and if
each variable has no more than k parents,
the complete network requires no more than $n \cdot 2^k$ numbers

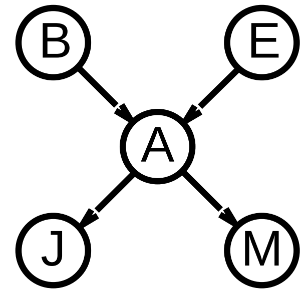
I.e., grows linearly with n, vs. $O(2^n)$ for the full joint distribution

How many numbers for the burglary net?
$1 + 1 + 4 + 2 + 2 = 10$ numbers
(vs. $2^5 - 1 = 31$ for the joint distribution)

Semantics of Bayesian nets

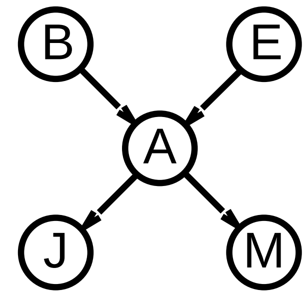
In general, *semantics* = "what things mean."
Here, we're interested in what a Bayesian net means.

We'll look at *global* and *local* semantics



Global semantics

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



If $X_1, \ldots, X_n$ are all of the random variables,
then by combining the chain rule and conditional
independence, we get

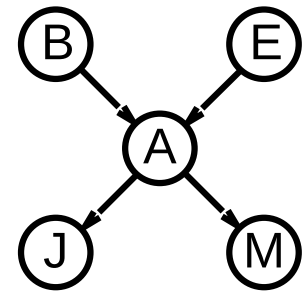
$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | parents(X_i))$$

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

$$=$$

Global semantics

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



If $X_1, \ldots, X_n$ are all of the random variables,
then by combining the chain rule and conditional
independence, we get

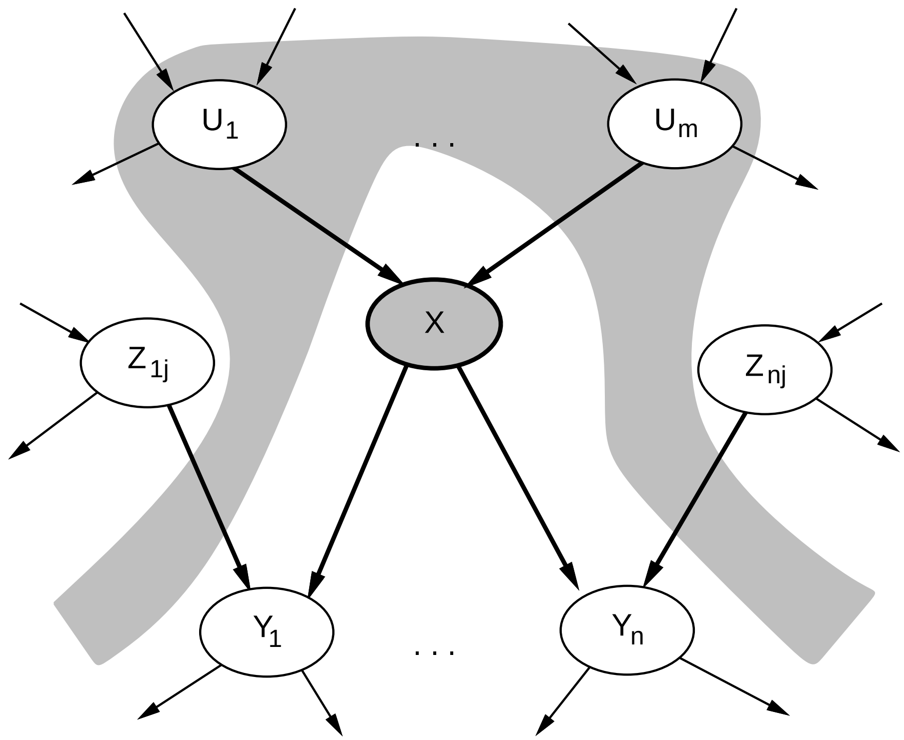
$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | parents(X_i))$$

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

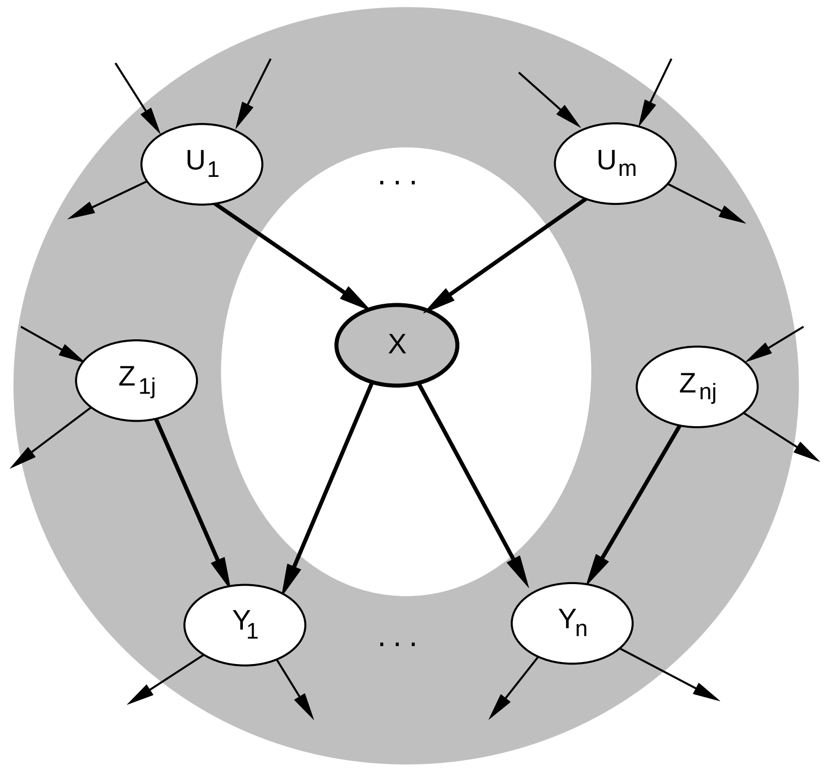
Local semantics: each node is conditionally independent of its nondescendants given its parents



Theorem: Local semantics $\Leftrightarrow$ global semantics

Markov blanket

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



Constructing Bayesian networks

Given a set of random variables

1. Choose an ordering $X_1, \ldots, X_n$
   In principle, *any* ordering will work

2. For $i = 1$ to n, add X_i to the network as follows:

   For $Parents(X_i)$, choose a subset of $\{X_1, \ldots, X_{i-1}\}$ such that X_i is conditionally independent of the other nodes in $\{X_1, \ldots, X_{i-1}\}$,

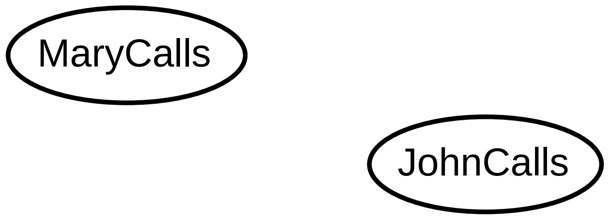
   i.e., $\mathbf{P}(X_i|Parents(X_i)) = \mathbf{P}(X_i|X_1, \ldots, X_{i-1})$

This choice of parents guarantees the global semantics:

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

Example

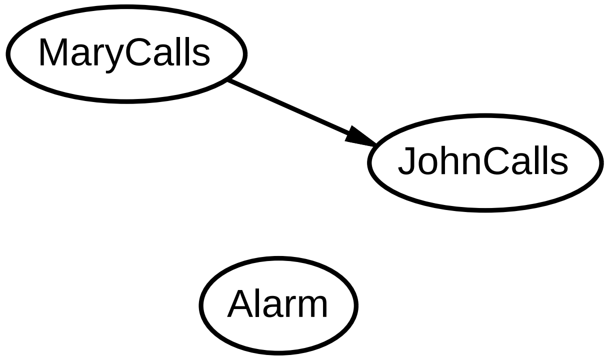
Suppose we choose the ordering M, J, A, B, E



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

Example

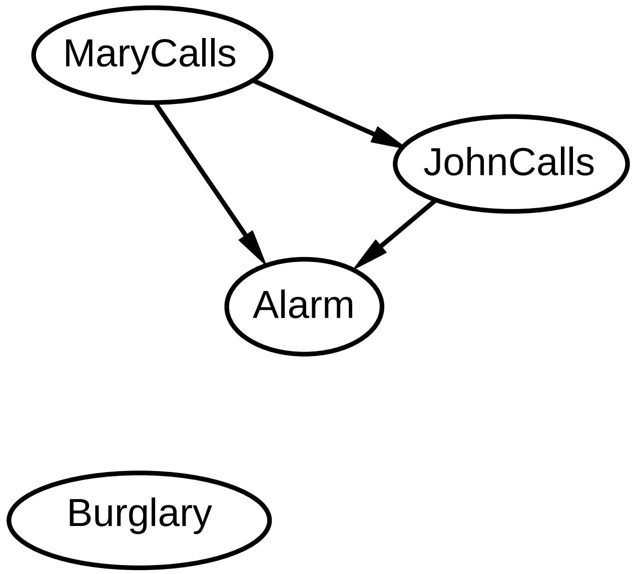
Suppose we choose the ordering M, J, A, B, E



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

Example

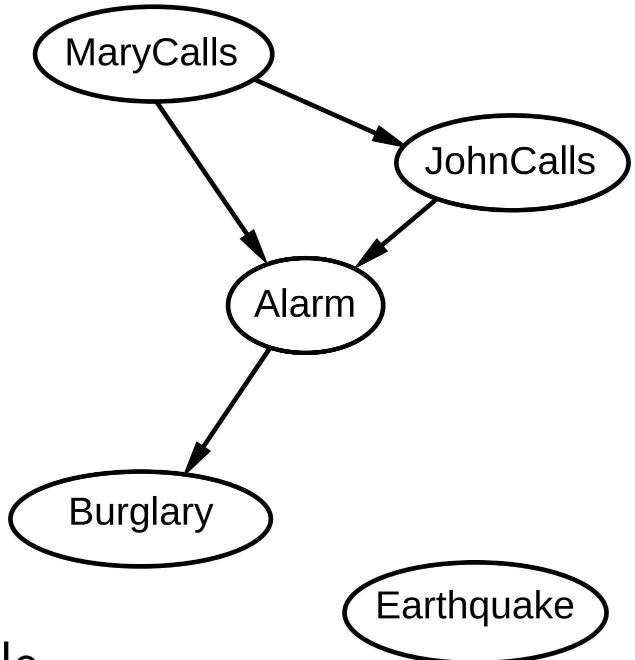
Suppose we choose the ordering M, J, A, B, E



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

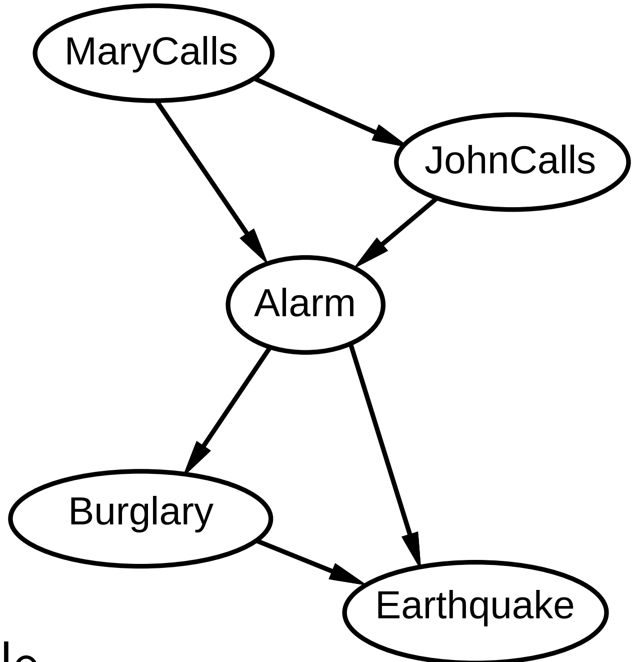
Suppose we choose the ordering M, J, A, B, E



$P(J|M) = P(J)$? 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
$P(E|B,A,J,M) = P(E|A)$?
$P(E|B,A,J,M) = P(E|A,B)$?

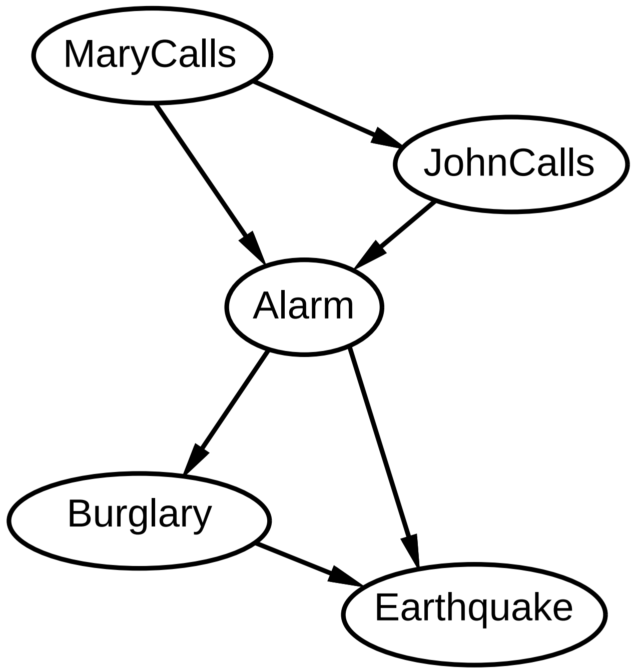
Example

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, continued



Deciding conditional independence is hard in noncausal directions

Assessing conditional probabilities is hard in noncausal directions

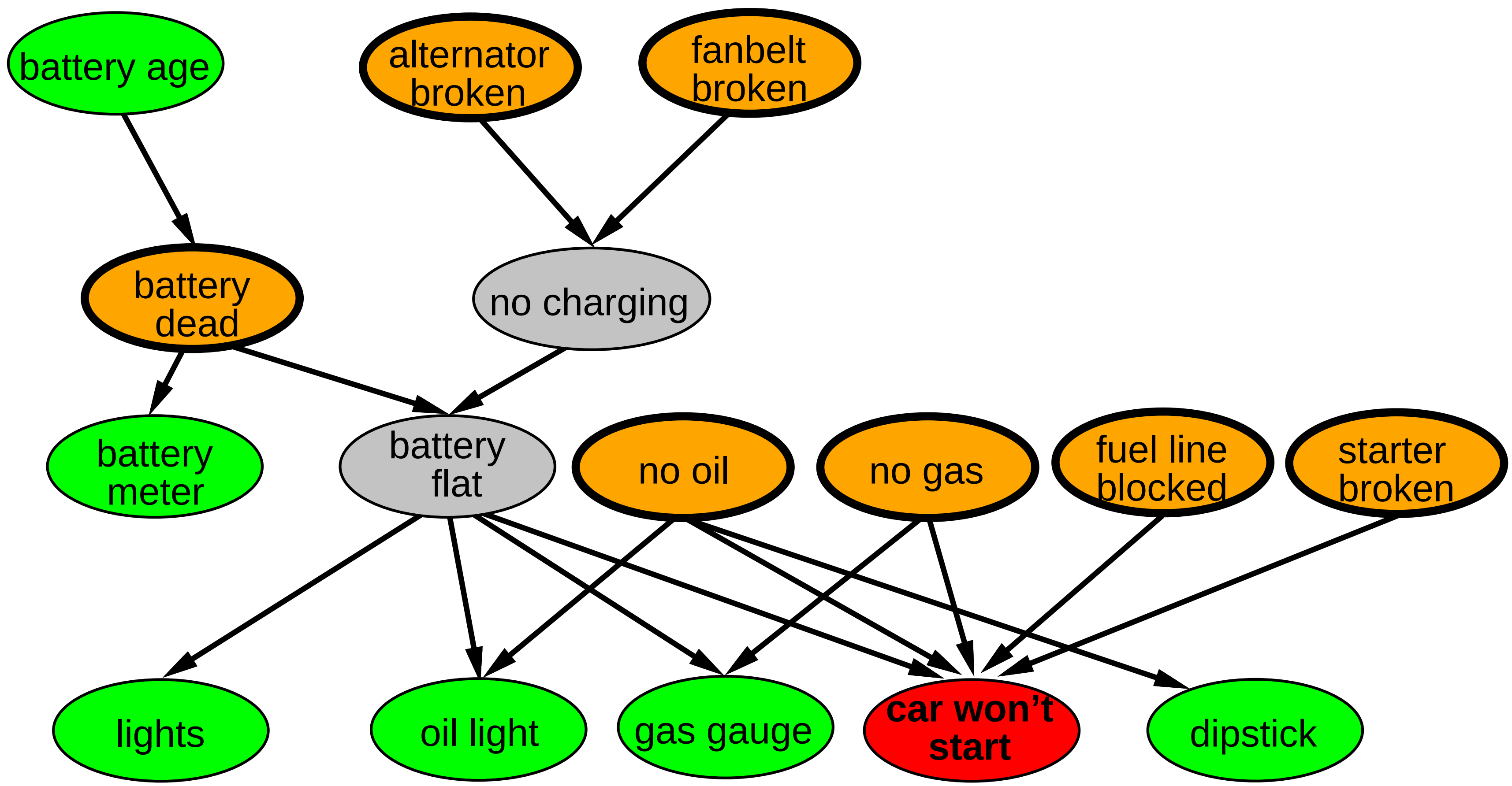
Network is less compact: $1 + 2 + 4 + 2 + 4 = 13$ numbers needed

Example: Car diagnosis

Initial evidence: car won't start

Testable variables (green), "broken, so fix it" variables (orange)

Hidden variables (gray) ensure sparse structure, reduce parameters



Compact conditional distributions

Problem: CPT grows exponentially with number of parents.

Can overcome this if the causes don't interact: use a *Noisy-OR* distribution
1) Parents $U_1 \ldots U_k$ include all causes (can add *leak node*)
2) Independent failure probability q_i for each cause U_i by itself
$\Rightarrow P(\neg X|U_1 \ldots U_j, \neg U_{j+1} \ldots \neg U_k) = \prod_{i=1}^{j} q_i$
Number of parameters **linear** in number of parents

$Cold$	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	$0.012 = 0.6 \times 0.2 \times 0.1$

Compact conditional distributions, continued

Problem: CPT becomes infinite with continuous-valued parent or child

Solution: *canonical* distributions that are defined compactly
canonical: conforming to orthodox or well-established rules or patterns
(e.g., something for which we can write an equation)

Deterministic nodes are the simplest case:
$X = f(Parents(X))$ for some function f

Examples:

◇ Boolean functions
$NorthAmerican \;\Leftrightarrow\; Canadian \vee US \vee Mexican$

◇ Numerical relationships among continuous variables

$$\frac{\partial Level}{\partial t} = \text{ inflow + precipitation - outflow - evaporation}$$

The book discusses this in detail, but I'll skip that part.

Inference tasks

Simple queries: compute posterior marginal distribution $\mathbf{P}(X_i|\mathbf{E}=\mathbf{e})$

e.g., $P(NoGas|Gauge=empty, Lights=on, Starts=false)$

Conjunctive queries: $\mathbf{P}(X_i, X_j|\mathbf{E}=\mathbf{e}) = \mathbf{P}(X_i|\mathbf{E}=\mathbf{e})\mathbf{P}(X_j|X_i, \mathbf{E}=\mathbf{e})$

Optimal decisions: decision networks include utility information;
probabilistic inference required for $P(outcome|action, evidence)$

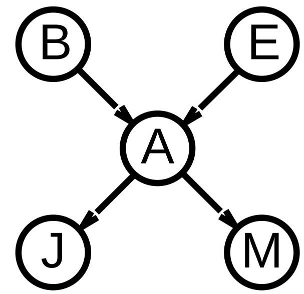
Value of information: which evidence to seek next?

Sensitivity analysis: which probability values are most critical?

Explanation: why do I need a new starter motor?

Inference by enumeration

Simple query on the burglary network:



$\mathbf{P}(B|j,m)$
$= \mathbf{P}(B,j,m)/P(j,m)$ (def. of cond. probability)
$= \alpha\mathbf{P}(B,j,m)$ (normalization constant)
$= \alpha\Sigma_e\Sigma_a\mathbf{P}(B,e,a,j,m)$ (sum over hidden variables)

[notation: $\Sigma_a\mathbf{P}(\ldots,a,\ldots)$ means $\mathbf{P}(\ldots,\neg a,\ldots) + \mathbf{P}(\ldots,a,\ldots)$]

Can sum out variables from the joint distribution without actually constructing its explicit representation:

Joint probabilities are products of probabilities in the network:
$\mathbf{P}(B|j,m)$
$= \alpha\Sigma_e\Sigma_a\mathbf{P}(B,e,a,j,m)$
$= \alpha\Sigma_e\Sigma_a\mathbf{P}(B)P(e)\mathbf{P}(a|B,e)P(j|a)P(m|a)$ (conditional independence)
$= \alpha\mathbf{P}(B)\Sigma_eP(e)\Sigma_a\mathbf{P}(a|B,e)P(j|a)P(m|a)$ (move term outside of Σ)

Recursive depth-first enumeration: $O(n)$ space, $O(d^n)$ time

Enumeration algorithm

function ENUMERATION-ASK(X, $\mathbf{e}$, bn) **returns** a distribution over X
  inputs: X, the query variable
    $\mathbf{e}$, observed values for variables $\mathbf{E}$
    bn, a Bayesian network with variables $\{X\} \cup \mathbf{E} \cup \mathbf{Y}$

  $\mathbf{Q}(X) \leftarrow$ a distribution over X, initially empty
  for each value x_i of X **do**
    extend $\mathbf{e}$ with value x_i for X
    $\mathbf{Q}(x_i) \leftarrow$ ENUMERATE-ALL(VARS[bn], $\mathbf{e}$)
  return NORMALIZE($\mathbf{Q}(X)$)

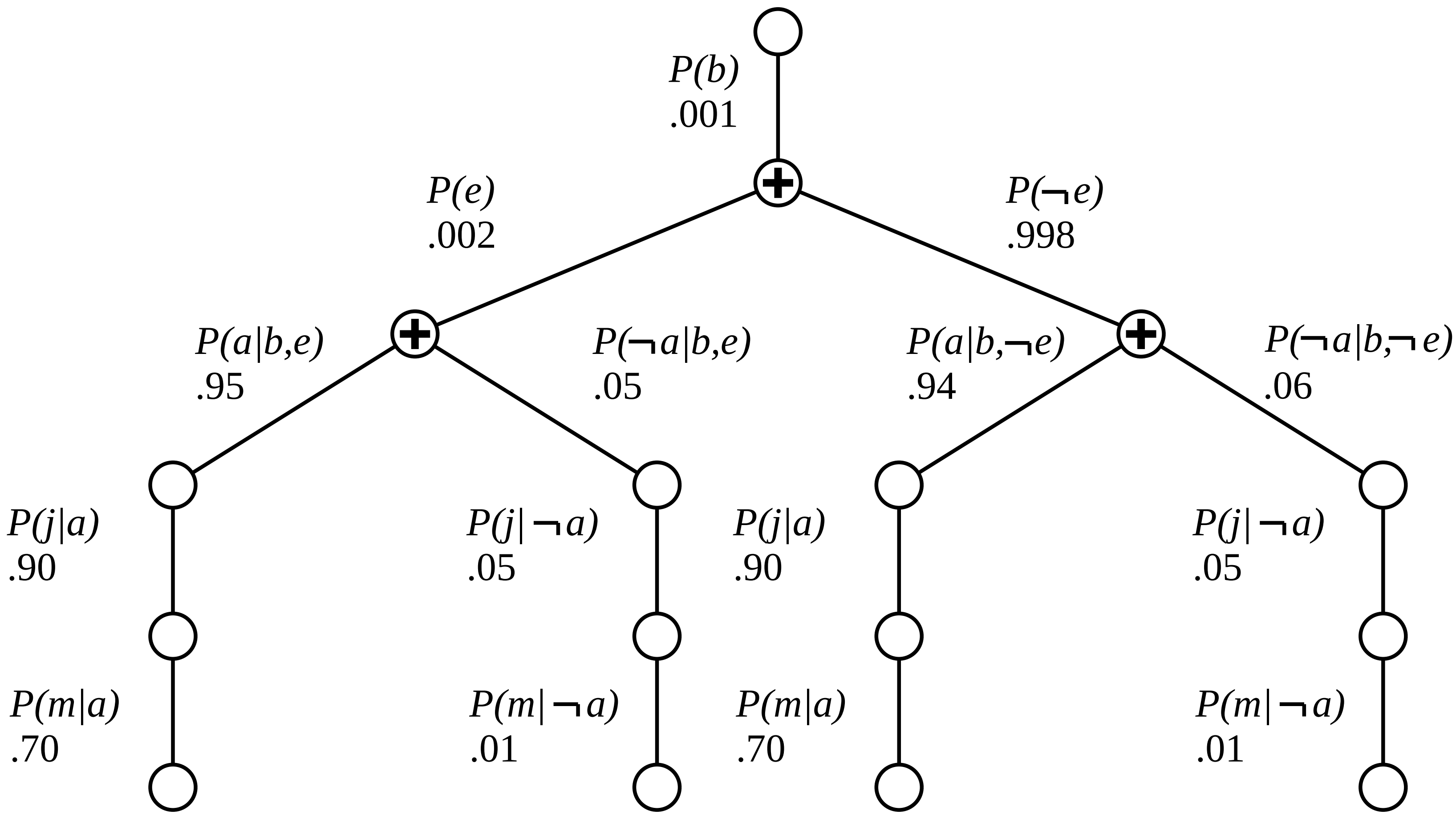
---

function ENUMERATE-ALL($vars$, $\mathbf{e}$) **returns** a real number
  if EMPTY?($vars$) **then return** 1.0
  $Y \leftarrow$ FIRST($vars$)
  if Y has value y in $\mathbf{e}$
    then return $P(y \mid Pa(Y)) \times$ ENUMERATE-ALL(REST($vars$), $\mathbf{e}$)
    else return $\Sigma_y\ P(y \mid Pa(Y)) \times$ ENUMERATE-ALL(REST($vars$), $\mathbf{e}_y$)
      where $\mathbf{e}_y$ is $\mathbf{e}$ extended with $Y = y$

Inefficient due to repeated computation

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

computes $P(j|a)P(m|a)$ for each value of e



Inference by variable elimination

Variable elimination: carry out summations right-to-left,
storing intermediate results (*factors*) to avoid recomputation

$$
\begin{aligned}
&\mathbf{P}(B|j,m) \\
&\quad = \alpha\, \underbrace{\mathbf{P}(B)}_{B} \textstyle\sum_e \underbrace{P(e)}_{E} \sum_a \underbrace{\mathbf{P}(a|B,e)}_{A} \underbrace{P(j|a)}_{J} \underbrace{P(m|a)}_{M} \\
&\quad = \alpha \mathbf{P}(B) \textstyle\sum_e P(e) \sum_a \mathbf{P}(a|B,e) P(j|a) f_M(a) \\
&\quad = \alpha \mathbf{P}(B) \textstyle\sum_e P(e) \sum_a \mathbf{P}(a|B,e) f_J(a) f_M(a) \\
&\quad = \alpha \mathbf{P}(B) \textstyle\sum_e P(e) \sum_a f_A(a,b,e) f_J(a) f_M(a) \\
&\quad = \alpha \mathbf{P}(B) \textstyle\sum_e P(e) f_{\bar{A}JM}(b,e) \text{ (sum out } A) \\
&\quad = \alpha \mathbf{P}(B) f_{\bar{E}\bar{A}JM}(b) \text{ (sum out } E) \\
&\quad = \alpha f_B(b) \times f_{\bar{E}\bar{A}JM}(b)
\end{aligned}
$$

Summing out a variable from a product of factors:
  move any constant factors outside the summation
  add up submatrices in pointwise product of remaining factors

Less complicated than it looks. Algorithmically, it's just caching

Irrelevant variables

Consider the query $P(JohnCalls|Burglary = true)$

$$P(J|b) = \alpha P(b) \sum_e P(e) \sum_a P(a|b,e) P(J|a) \sum_m P(m|a)$$

Sum over m is identically 1; M is **irrelevant** to the query

Thm 1: Y is irrelevant unless $Y \in Ancestors(\{X\} \cup \mathbf{E})$

Here, $X = JohnCalls$, $\mathbf{E} = \{Burglary\}$, and
$Ancestors(\{X\} \cup \mathbf{E}) = \{Alarm, Earthquake\}$
so M is irrelevant

Irrelevant variables contd.

Defn: moral graph of Bayes net: marry all parents and drop arrows

Defn: $\mathbf{U}$ is m-separated from $\mathbf{V}$ by $\mathbf{W}$ iff separated by $\mathbf{W}$ in the moral graph

Thm 2: Y is irrelevant if m-separated from X by $\mathbf{E}$

For $P(JohnCalls|Alarm = true)$, both $Burglary$ and $Earthquake$ are irrelevant

Complexity of exact inference

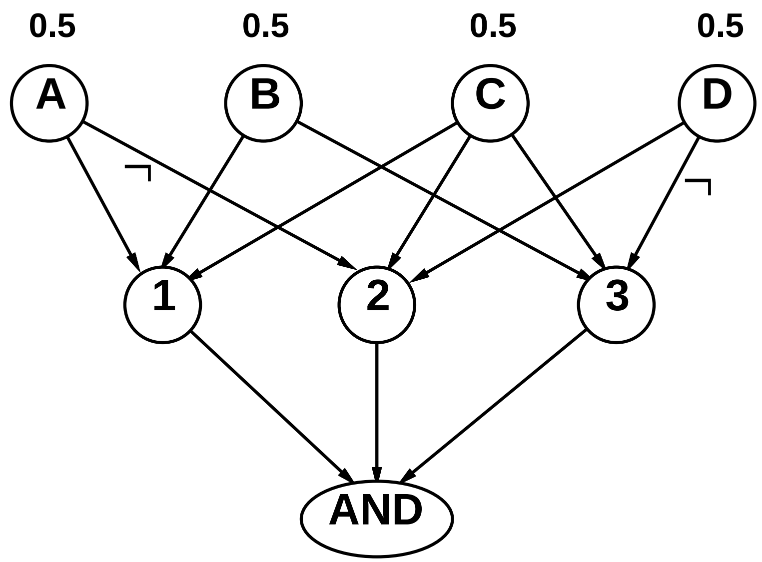
Singly connected networks (or *polytrees*):
- any two nodes are connected by at most one (undirected) path
- time and space cost of variable elimination are $O(d^k n)$

Multiply connected networks:
- exponential time and space in the worst case
- as hard as **counting** the number of models of a propositional formula

1. A v B v C

2. C v D v ¬A

3. B v C v ¬D



Summary

Bayes nets provide a natural representation for (causally induced) conditional independence

Topology + CPTs = compact representation of joint distribution

Generally easy for (non)experts to construct

Canonical distributions (e.g., noisy-OR) = compact representation of CPTs

Continuous variables $\Rightarrow$ parameterized distributions (e.g., linear Gaussian)