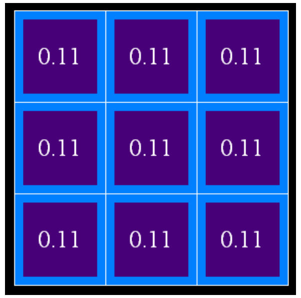
Today – Part 1: Uncertainty

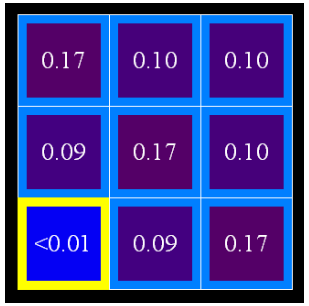
- **Only a degree of belief**
- Use **probability theory**
  - Assign to each sentence a numerical degree of belief.
  - Not degree of truth!
  - Summarizing the uncertainty that comes from laziness, ignorance.

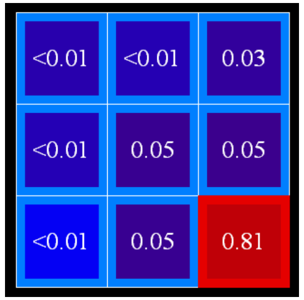
- Random variables
- Joint and marginal distributions
- Conditional distributions
- Product rule, chain rule, Bayes' rule
- Inference
- Independence, conditional independence

Uncertainty

- General situation:
  - **Observed variables (evidence)**: Agent knows certain things about the state of the world (e.g., sensor readings or symptoms)
  - **Unobserved variables**: Agent needs to reason about other aspects (e.g. where an object is or what disease is present)
  - **Model**: Agent knows something about how the known variables relate to the unknown variables
- Probabilistic reasoning gives us a framework for managing our beliefs and knowledge

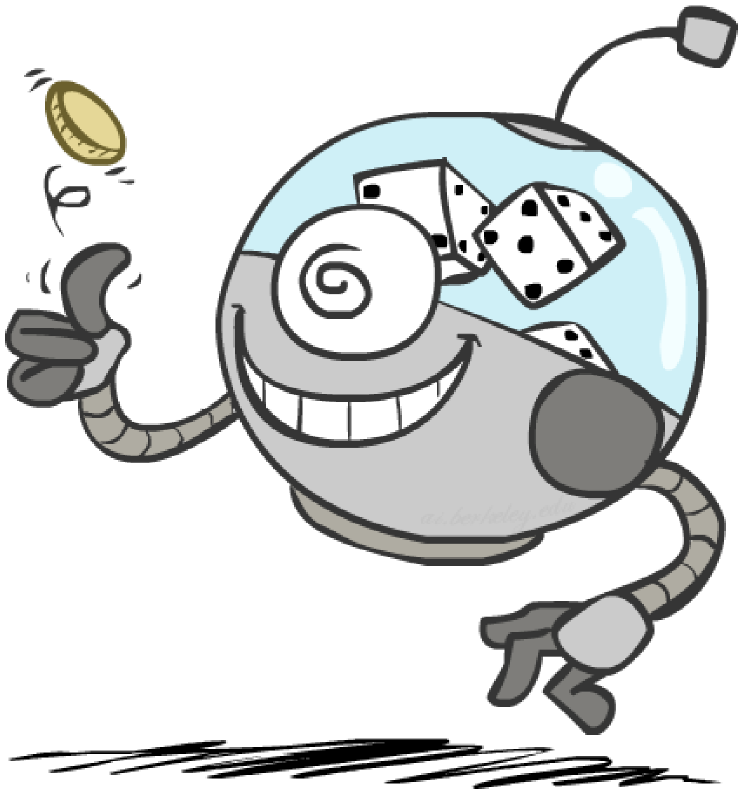






Random Variables

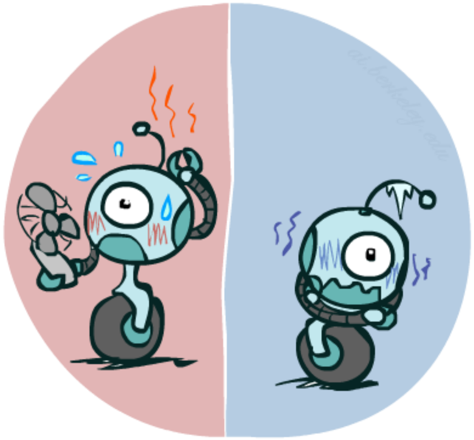
- A random variable is some aspect of the world about which we (may) have uncertainty
  - R = Is it raining?
  - T = Is it hot or cold?
  - D = How long will it take to drive to work?
  - L = Where is the ghost?
- We denote random variables with capital letters
- Like variables in a CSP, random variables have domains
  - R in {true, false}   (often write as {+r, -r})
  - T in {hot, cold}
  - D in [0, ∞)
  - L in possible locations, maybe {(0,0), (0,1), …}



Probability Distributions

- Associate a probability with each value

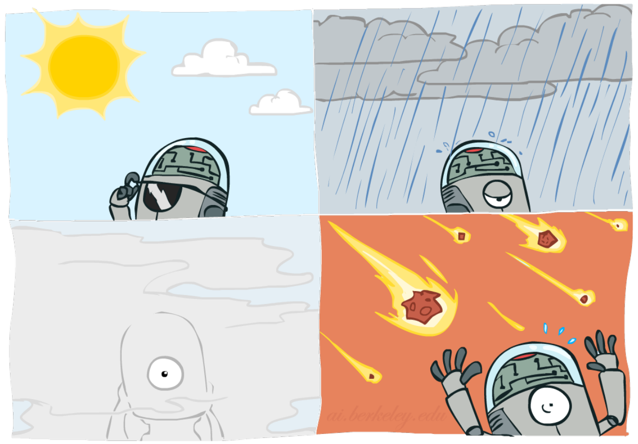
  - Temperature:



$P(T)$

T	P
hot	0.5
cold	0.5

  - Weather:



$P(W)$

W	P
sun	0.6
rain	0.1
fog	0.3
meteor	0.0

Probability Distributions

- Unobserved random variables have distributions

$P(T)$

T	P
hot	0.5
cold	0.5

$P(W)$

W	P
sun	0.6
rain	0.1
fog	0.3
meteor	0.0

- A distribution is a TABLE of probabilities of values

- A probability (lower case value) is a single number

$$P(W = rain) = 0.1$$

- Must have: $\forall x \;\; P(X = x) \geq 0$ and $\sum_x P(X = x) = 1$

Shorthand notation:

$$P(hot) = P(T = hot),$$
$$P(cold) = P(T = cold),$$
$$P(rain) = P(W = rain),$$
$$\ldots$$

OK *if* all domain entries are unique

Joint Distributions

- A *joint distribution* over a set of random variables: $X_1, X_2, \ldots X_n$ specifies a real number for each assignment (or *outcome*):

$$P(X_1 = x_1, X_2 = x_2, \ldots X_n = x_n)$$

$$P(x_1, x_2, \ldots x_n)$$

$$P(x_1, x_2, \ldots x_n) \geq 0$$

$$\sum_{(x_1, x_2, \ldots x_n)} P(x_1, x_2, \ldots x_n) = 1$$

Size of distribution if n variables with domain sizes d?

  - For all but the smallest distributions, impractical to write out!

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

Prior probability

Prior or unconditional probabilities of propositions

e.g., $P(Cavity\!=\!true) = 0.1$ and $P(Weather\!=\!sunny) = 0.72$

correspond to belief prior to arrival of any (new) evidence

Probability distribution gives values for all possible assignments:

$\mathbf{P}(Weather) = \langle 0.72, 0.1, 0.08, 0.1 \rangle$ (normalized, i.e., sums to 1)

Joint probability distribution for a set of r.v.s gives the probability of every atomic event on those r.v.s (i.e., every sample point)

$\mathbf{P}(Weather, Cavity) =$ a 4×2 matrix of values:

$Weather =$	$sunny$	$rain$	$cloudy$	$snow$
$Cavity\!=\!true$	0.144	0.02	0.016	0.02
$Cavity\!=\!false$	0.576	0.08	0.064	0.08

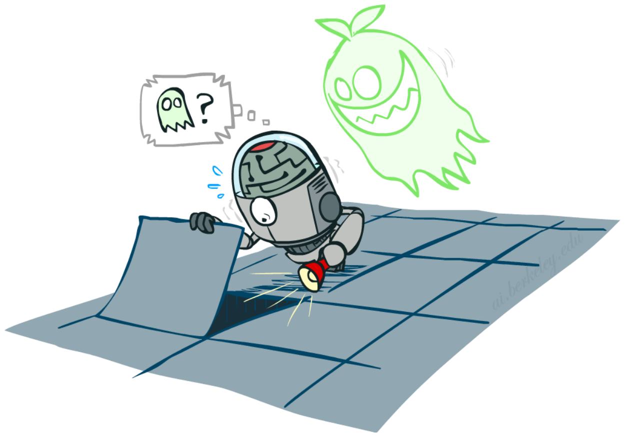
Every question about a domain can be answered by the joint distribution because every event is a sum of sample points

Probabilistic Models

- A probabilistic model is a joint distribution over a set of random variables
- Probabilistic models:
  - (Random) variables with domains
  - Assignments are called *outcomes*
  - Joint distributions: say whether assignments (outcomes) are likely
  - *Normalized:* sum to 1.0
  - Ideally: only certain variables directly interact
- Constraint satisfaction problems:
  - Variables with domains
  - Constraints: state whether assignments are possible
  - Ideally: only certain variables directly interact

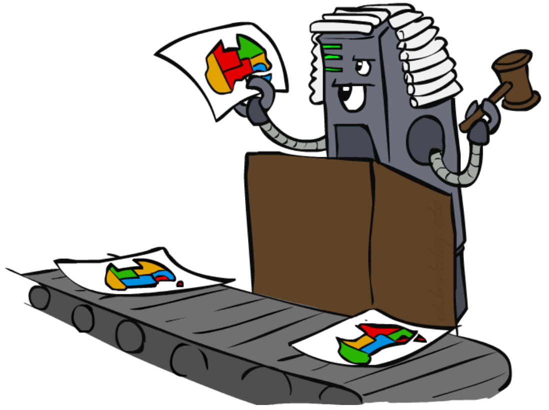
Distribution over T,W

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3



Constraint over T,W

T	W	P
hot	sun	T
hot	rain	F
cold	sun	F
cold	rain	T



Events

- An *event* is a set E of outcomes

$$P(E) = \sum_{(x_1 \ldots x_n) \in E} P(x_1 \ldots x_n)$$

- From a joint distribution, we can calculate the probability of any event
  - Probability that it's hot AND sunny?
  - Probability that it's hot?
  - Probability that it's hot OR sunny?
- Typically, the events we care about are *partial assignments*, like P(T=hot)

$$P(T, W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

Quiz: Events

- P(+x, +y) ?

- P(+x) ?

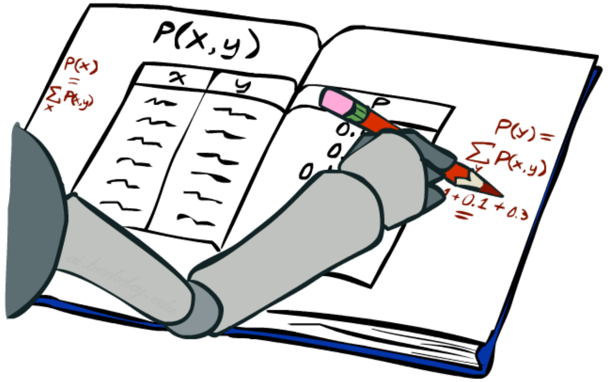
- P(-y OR +x) ?

$$P(X, Y)$$

X	Y	P
+x	+y	0.2
+x	-y	0.3
-x	+y	0.4
-x	-y	0.1

Marginal Distributions

- Marginal distributions are sub-tables which eliminate variables
- Marginalization (summing out): Combine collapsed rows by adding



$$P(T,W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P(t) = \sum_s P(t,s)$$

$$P(s) = \sum_t P(t,s)$$

$$P(T)$$

T	P
hot	0.5
cold	0.5

$$P(W)$$

W	P
sun	0.6
rain	0.4

$$P(X_1 = x_1) = \sum_{x_2} P(X_1 = x_1, X_2 = x_2)$$

Quiz: Marginal Distributions

$P(X, Y)$

X	Y	P
+x	+y	0.2
+x	-y	0.3
-x	+y	0.4
-x	-y	0.1

$$P(x) = \sum_y P(x, y)$$

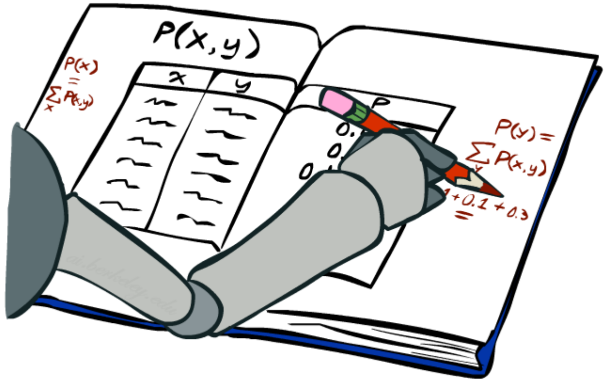
$$P(y) = \sum_x P(x, y)$$

$P(X)$

X	P
+x	
-x	

$P(Y)$

Y	P
+y	
-y	



Conditional probability

Conditional or posterior probabilities

e.g., $P(cavity|toothache) = 0.8$

i.e., **given that** $toothache$ **is all I know**

NOT "if $toothache$ then 80% chance of $cavity$"

(Notation for conditional distributions:

$\mathbf{P}(Cavity|Toothache) =$ 2-element vector of 2-element vectors)

If we know more, e.g., $cavity$ is also given, then we have

$P(cavity|toothache, cavity) = 1$

Note: the less specific belief **remains valid** after more evidence arrives, but is not always **useful**

New evidence may be irrelevant, allowing simplification, e.g.,

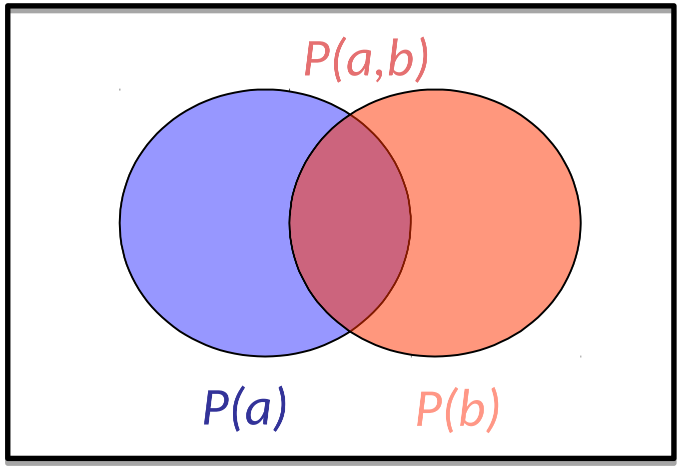
$P(cavity|toothache, 49ersWin) = P(cavity|toothache) = 0.8$

This kind of inference, sanctioned by domain knowledge, is crucial

Conditional Probabilities

- A simple relation between joint and conditional probabilities
  - In fact, this is taken as the *definition* of a conditional probability

$$P(a|b) = \frac{P(a,b)}{P(b)}$$



$$P(T,W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P(W=s|T=c) = \frac{P(W=s,T=c)}{P(T=c)} = \frac{0.2}{0.5} = 0.4$$

$$= P(W=s,T=c) + P(W=r,T=c)$$
$$= 0.2 + 0.3 \quad = 0.5$$

Quiz: Conditional Probabilities

$$P(X, Y)$$

X	Y	P
+x	+y	0.2
+x	-y	0.3
-x	+y	0.4
-x	-y	0.1

- P(+x | +y) ?

- P(-x | +y) ?

- P(-y | +x) ?

Conditional Distributions

- Conditional distributions are probability distributions over some variables given fixed values of others

Conditional Distributions

$P(W|T)$

$P(W|T = hot)$

W	P
sun	0.8
rain	0.2

$P(W|T = cold)$

W	P
sun	0.4
rain	0.6

Joint Distribution

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

Normalization Trick

$$P(T, W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P(W|T = c)$$

W	P
sun	
rain	

Normalization Trick

$$P(T,W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P(W=s|T=c) = \frac{P(W=s,T=c)}{P(T=c)}$$
$$= \frac{P(W=s,T=c)}{P(W=s,T=c)+P(W=r,T=c)}$$
$$= \frac{0.2}{0.2+0.3} = 0.4$$

$$P(W=r|T=c) = \frac{P(W=r,T=c)}{P(T=c)}$$
$$= \frac{P(W=r,T=c)}{P(W=s,T=c)+P(W=r,T=c)}$$
$$= \frac{0.3}{0.2+0.3} = 0.6$$

$$P(W|T=c)$$

W	P
sun	0.4
rain	0.6

Normalization Trick

$$
\begin{aligned}
P(W=s|T=c) &= \frac{P(W=s,T=c)}{P(T=c)} \\
&= \frac{P(W=s,T=c)}{P(W=s,T=c)+P(W=r,T=c)} \\
&= \frac{0.2}{0.2+0.3} = 0.4
\end{aligned}
$$

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

SELECT the joint probabilities matching the evidence

$P(c, W)$

T	W	P
cold	sun	0.2
cold	rain	0.3

NORMALIZE the selection (make it sum to one)

$P(W|T=c)$

W	P
sun	0.4
rain	0.6

$$
\begin{aligned}
P(W=r|T=c) &= \frac{P(W=r,T=c)}{P(T=c)} \\
&= \frac{P(W=r,T=c)}{P(W=s,T=c)+P(W=r,T=c)} \\
&= \frac{0.3}{0.2+0.3} = 0.6
\end{aligned}
$$

Quiz: Normalization Trick

- P(X | Y=-y) ?

$$P(X, Y)$$

X	Y	P
+x	+y	0.2
+x	-y	0.3
-x	+y	0.4
-x	-y	0.1

SELECT the joint probabilities matching the evidence →

NORMALIZE the selection (make it sum to one) →

Probabilistic Inference

- Probabilistic inference: compute a desired probability from other known probabilities (e.g. conditional from joint)

- We generally compute conditional probabilities
  - P(on time | no reported accidents) = 0.90
  - These represent the agent's *beliefs* given the evidence

- Probabilities change with new evidence:
  - P(on time | no accidents, 5 a.m.) = 0.95
  - P(on time | no accidents, 5 a.m., raining) = 0.80
  - Observing new evidence causes *beliefs to be updated*

Inference by enumeration

Start with the joint distribution:

	toothache		$\neg$ *toothache*	
	catch	$\neg$ *catch*	*catch*	$\neg$ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
$\neg$ *cavity*	**.016**	**.064**	**.144**	**.576**

For any proposition ϕ, sum the atomic events where it is true:

$P(\phi) = \Sigma_{\omega:\omega\models\phi} P(\omega)$

Inference by enumeration

Start with the joint distribution:

	$toothache$		$\neg\, toothache$	
	$catch$	$\neg\, catch$	$catch$	$\neg\, catch$
$cavity$	**.108**	**.012**	**.072**	**.008**
$\neg\, cavity$	**.016**	**.064**	**.144**	**.576**

For any proposition ϕ, sum the atomic events where it is true:
$P(\phi) = \Sigma_{\omega:\omega \models \phi} P(\omega)$

$P(toothache) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$

Inference by enumeration

Start with the joint distribution:

	toothache		$\neg$ *toothache*	
	catch	$\neg$ *catch*	*catch*	$\neg$ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
$\neg$ *cavity*	**.016**	**.064**	**.144**	**.576**

For any proposition ϕ, sum the atomic events where it is true:

$$P(\phi) = \Sigma_{\omega:\omega \models \phi} P(\omega)$$

$$P(cavity \lor toothache) = 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28$$

Inference by enumeration

Start with the joint distribution:

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ *cavity*	**.016**	**.064**	**.144**	**.576**

Can also compute conditional probabilities:

$$
\begin{aligned}
P(\neg cavity | toothache) &= \frac{P(\neg cavity \land toothache)}{P(toothache)} \\
&= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4
\end{aligned}
$$

Normalization

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ *cavity*	**.016**	**.064**	**.144**	**.576**

Denominator can be viewed as a normalization constant α

$$\mathbf{P}(Cavity|toothache) = \alpha\,\mathbf{P}(Cavity, toothache)$$
$$= \alpha\,[\mathbf{P}(Cavity, toothache, catch) + \mathbf{P}(Cavity, toothache, \neg catch)]$$
$$= \alpha\,[\langle 0.108, 0.016\rangle + \langle 0.012, 0.064\rangle]$$
$$= \alpha\,\langle 0.12, 0.08\rangle = \langle 0.6, 0.4\rangle$$

General idea: compute distribution on query variable
by fixing evidence variables and summing over hidden variables

Inference by enumeration, contd.

Let $\mathbf{X}$ be all the variables. Typically, we want
the posterior joint distribution of the query variables $\mathbf{Y}$
given specific values $\mathbf{e}$ for the evidence variables $\mathbf{E}$

Let the hidden variables be $\mathbf{H} = \mathbf{X} - \mathbf{Y} - \mathbf{E}$

Then the required summation of joint entries is done by summing out the hidden variables:

$$\mathbf{P}(\mathbf{Y}|\mathbf{E}=\mathbf{e}) = \alpha\mathbf{P}(\mathbf{Y},\mathbf{E}=\mathbf{e}) = \alpha\Sigma_{\mathbf{h}}\mathbf{P}(\mathbf{Y},\mathbf{E}=\mathbf{e},\mathbf{H}=\mathbf{h})$$

The terms in the summation are joint entries because $\mathbf{Y}$, $\mathbf{E}$, and $\mathbf{H}$ together exhaust the set of random variables

Obvious problems:
1) Worst-case time complexity $O(d^n)$ where d is the largest arity
2) Space complexity $O(d^n)$ to store the joint distribution
3) How to find the numbers for $O(d^n)$ entries???

Inference by Enumeration

- P(W)?
- P(W | winter)?
- P(W | winter, hot)?

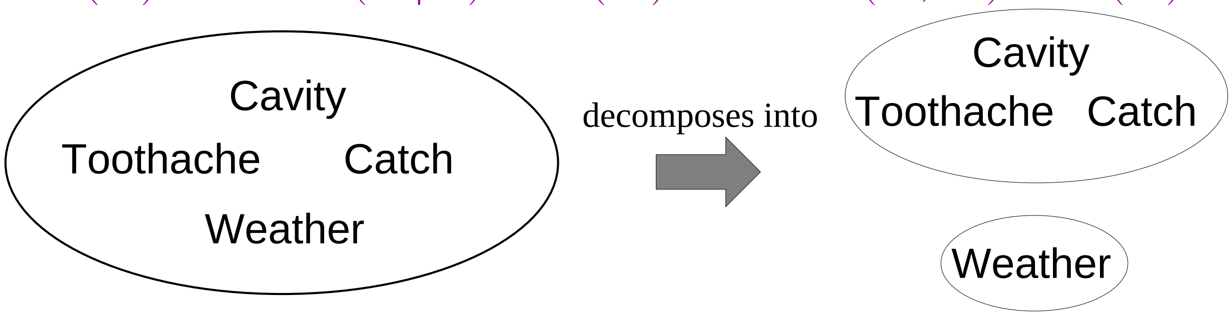
S	T	W	P
summer	hot	sun	0.30
summer	hot	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	hot	sun	0.10
winter	hot	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.20

Inference by Enumeration

- Obvious problems:
  - Worst-case time complexity $O(d^n)$
  - Space complexity $O(d^n)$ to store the joint distribution

Independence

A and B are independent iff
$\mathbf{P}(A|B) = \mathbf{P}(A)$ or $\mathbf{P}(B|A) = \mathbf{P}(B)$ or $\mathbf{P}(A,B) = \mathbf{P}(A)\mathbf{P}(B)$



$$\mathbf{P}(Toothache, Catch, Cavity, Weather)$$
$$= \mathbf{P}(Toothache, Catch, Cavity)\mathbf{P}(Weather)$$

32 entries reduced to 12; for n independent biased coins, $2^n \rightarrow n$

Absolute independence powerful but rare

Dentistry is a large field with hundreds of variables, none of which are independent. What to do?

Conditional independence

$\mathbf{P}(Toothache, Cavity, Catch)$ has $2^3 - 1 = 7$ independent entries

If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:

(1) $P(catch|toothache, cavity) = P(catch|cavity)$

The same independence holds if I haven't got a cavity:

(2) $P(catch|toothache, \neg cavity) = P(catch|\neg cavity)$

$Catch$ is conditionally independent of $Toothache$ given $Cavity$:

$\mathbf{P}(Catch|Toothache, Cavity) = \mathbf{P}(Catch|Cavity)$

Equivalent statements:

$\mathbf{P}(Toothache|Catch, Cavity) = \mathbf{P}(Toothache|Cavity)$

$\mathbf{P}(Toothache, Catch|Cavity) = \mathbf{P}(Toothache|Cavity)\mathbf{P}(Catch|Cavity)$

Conditional independence contd.

Write out full joint distribution using chain rule:

$$\mathbf{P}(Toothache, Catch, Cavity)$$
$$= \mathbf{P}(Toothache|Catch, Cavity)\mathbf{P}(Catch, Cavity)$$
$$= \mathbf{P}(Toothache|Catch, Cavity)\mathbf{P}(Catch|Cavity)\mathbf{P}(Cavity)$$
$$= \mathbf{P}(Toothache|Cavity)\mathbf{P}(Catch|Cavity)\mathbf{P}(Cavity)$$

I.e., $2 + 2 + 1 = 5$ independent numbers (equations 1 and 2 remove 2)

In most cases, the use of conditional independence reduces the size of the representation of the joint distribution from exponential in n to linear in n.

Conditional independence is our most basic and robust form of knowledge about uncertain environments.

Bayes' Rule

Product rule $P(a \wedge b) = P(a|b)P(b) = P(b|a)P(a)$

$\Rightarrow$ Bayes' rule $P(a|b) = \dfrac{P(b|a)P(a)}{P(b)}$

or in distribution form

$$\mathbf{P}(Y|X) = \frac{\mathbf{P}(X|Y)\mathbf{P}(Y)}{\mathbf{P}(X)} = \alpha\mathbf{P}(X|Y)\mathbf{P}(Y)$$

Useful for assessing diagnostic probability from causal probability:

$$P(Cause|Effect) = \frac{P(Effect|Cause)P(Cause)}{P(Effect)}$$

E.g., let M be meningitis, S be stiff neck:

$$P(m|s) = \frac{P(s|m)P(m)}{P(s)} = \frac{0.8 \times 0.0001}{0.1} = 0.0008$$

Note: posterior probability of meningitis still very small!

Quiz: Bayes' Rule

- Given:

$P(W)$

R	P
sun	0.8
rain	0.2

$P(D|W)$

D	W	P
wet	sun	0.1
dry	sun	0.9
wet	rain	0.7
dry	rain	0.3

- What is P(W | dry) ?

Bayes' Rule and conditional independence

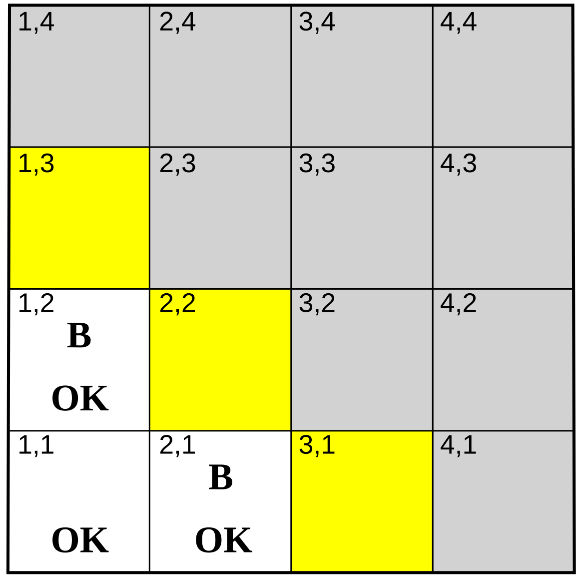
$$\mathbf{P}(Cavity|toothache \wedge catch)$$
$$= \alpha\, \mathbf{P}(toothache \wedge catch|Cavity)\mathbf{P}(Cavity)$$
$$= \alpha\, \mathbf{P}(toothache|Cavity)\mathbf{P}(catch|Cavity)\mathbf{P}(Cavity)$$

This is an example of a naive Bayes model:

$$\mathbf{P}(Cause, Effect_1, \ldots, Effect_n) = \mathbf{P}(Cause)\Pi_i \mathbf{P}(Effect_i|Cause)$$

Total number of parameters is **linear** in n

Wumpus World



$P_{ij} = true$ iff $[i, j]$ contains a pit

$B_{ij} = true$ iff $[i, j]$ is breezy

Include only $B_{1,1}, B_{1,2}, B_{2,1}$ in the probability model

Specifying the probability model

The full joint distribution is $\mathbf{P}(P_{1,1}, \ldots, P_{4,4}, B_{1,1}, B_{1,2}, B_{2,1})$

Apply product rule: $\mathbf{P}(B_{1,1}, B_{1,2}, B_{2,1} \,|\, P_{1,1}, \ldots, P_{4,4})\mathbf{P}(P_{1,1}, \ldots, P_{4,4})$

(Do it this way to get $P(Effect|Cause)$.)

First term: 1 if pits are adjacent to breezes, 0 otherwise

Second term: pits are placed randomly, probability 0.2 per square:

$$\mathbf{P}(P_{1,1}, \ldots, P_{4,4}) = \prod_{i,j=1,1}^{4,4} \mathbf{P}(P_{i,j}) = 0.2^n \times 0.8^{16-n}$$

for n pits.

Observations and query

We know the following facts:

$b = \neg b_{1,1} \wedge b_{1,2} \wedge b_{2,1}$
$known = \neg p_{1,1} \wedge \neg p_{1,2} \wedge \neg p_{2,1}$

Query is $\mathbf{P}(P_{1,3}|known, b)$

Define $Unknown = P_{ij}$s other than $P_{1,3}$ and $Known$

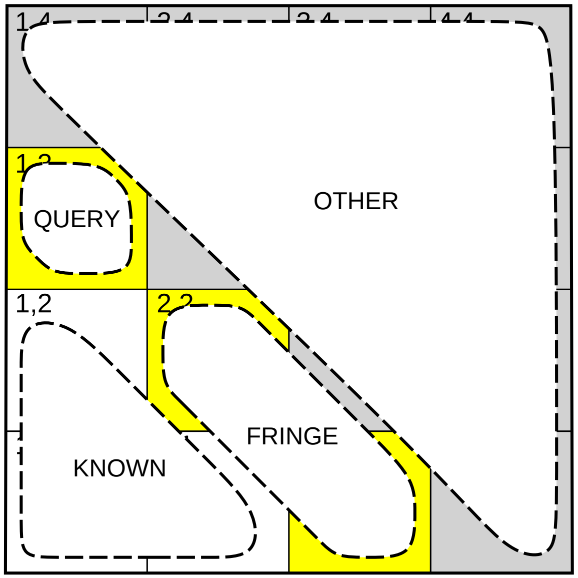
For inference by enumeration, we have

$$\mathbf{P}(P_{1,3}|known, b) = \alpha \Sigma_{unknown} \mathbf{P}(P_{1,3}, unknown, known, b)$$

Grows exponentially with number of squares!

Using conditional independence

Basic insight: observations are conditionally independent of other hidden squares given neighbouring hidden squares



Define $Unknown = Fringe \cup Other$

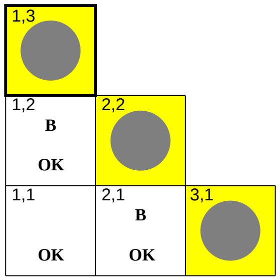
$\mathbf{P}(b|P_{1,3}, Known, Unknown) = \mathbf{P}(b|P_{1,3}, Known, Fringe)$

Manipulate query into a form where we can use this!

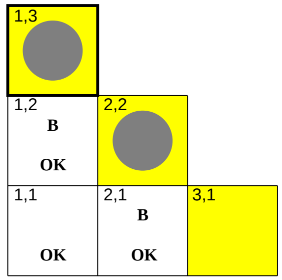
Using conditional independence contd.

$$
\begin{aligned}
\mathbf{P}(P_{1,3}|known, b) &= \alpha \sum_{unknown} \mathbf{P}(P_{1,3}, unknown, known, b) \\
&= \alpha \sum_{unknown} \mathbf{P}(b|P_{1,3}, known, unknown)\mathbf{P}(P_{1,3}, known, unknown) \\
&= \alpha \sum_{fringe} \sum_{other} \mathbf{P}(b|known, P_{1,3}, fringe, other)\mathbf{P}(P_{1,3}, known, fringe, other) \\
&= \alpha \sum_{fringe} \sum_{other} \mathbf{P}(b|known, P_{1,3}, fringe)\mathbf{P}(P_{1,3}, known, fringe, other) \\
&= \alpha \sum_{fringe} \mathbf{P}(b|known, P_{1,3}, fringe) \sum_{other} \mathbf{P}(P_{1,3}, known, fringe, other) \\
&= \alpha \sum_{fringe} \mathbf{P}(b|known, P_{1,3}, fringe) \sum_{other} \mathbf{P}(P_{1,3})P(known)P(fringe)P(other) \\
&= \alpha \, P(known)\mathbf{P}(P_{1,3}) \sum_{fringe} \mathbf{P}(b|known, P_{1,3}, fringe)P(fringe) \sum_{other} P(other) \\
&= \alpha' \, \mathbf{P}(P_{1,3}) \sum_{fringe} \mathbf{P}(b|known, P_{1,3}, fringe)P(fringe)
\end{aligned}
$$

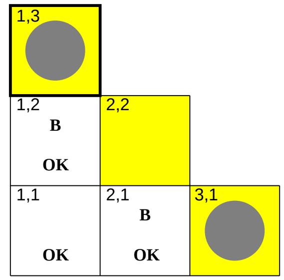
Using conditional independence contd.



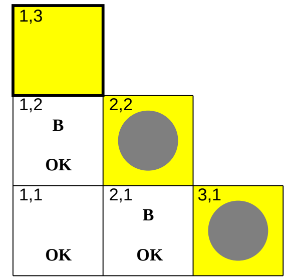
0.2 x 0.2 = 0.04



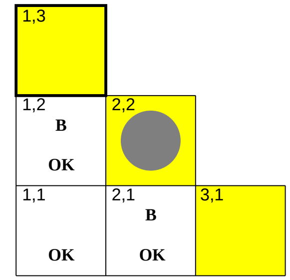
0.2 x 0.8 = 0.16



0.8 x 0.2 = 0.16



0.2 x 0.2 = 0.04



0.2 x 0.8 = 0.16

$$\begin{aligned}\mathbf{P}(P_{1,3}|known, b) &= \alpha' \langle 0.2(0.04 + 0.16 + 0.16),\ 0.8(0.04 + 0.16) \rangle \\ &\approx \langle 0.31, 0.69 \rangle\end{aligned}$$

$$\mathbf{P}(P_{2,2}|known, b) \approx \langle 0.86, 0.14 \rangle$$

Today – Part 2: Probabilistic Reasoning

- Bayesian networks
  - Systematic way to represent the independence and conditional independence relationships.

Outline

- Syntax
- Semantics
- Parameterized distributions

Bayesian networks

A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions

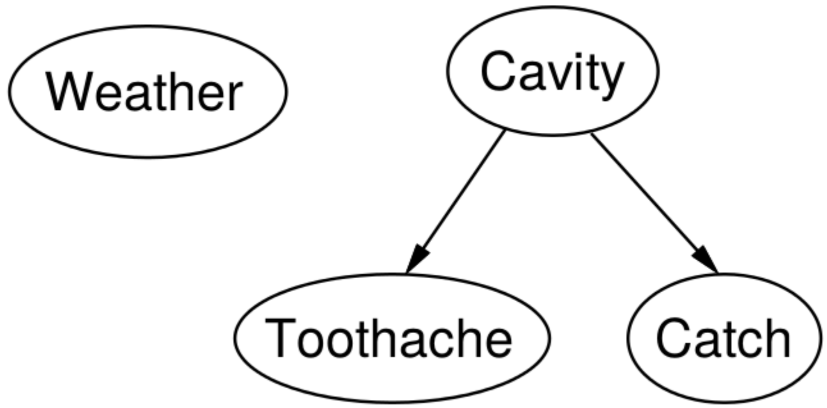
Syntax:

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 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 $Cavity$

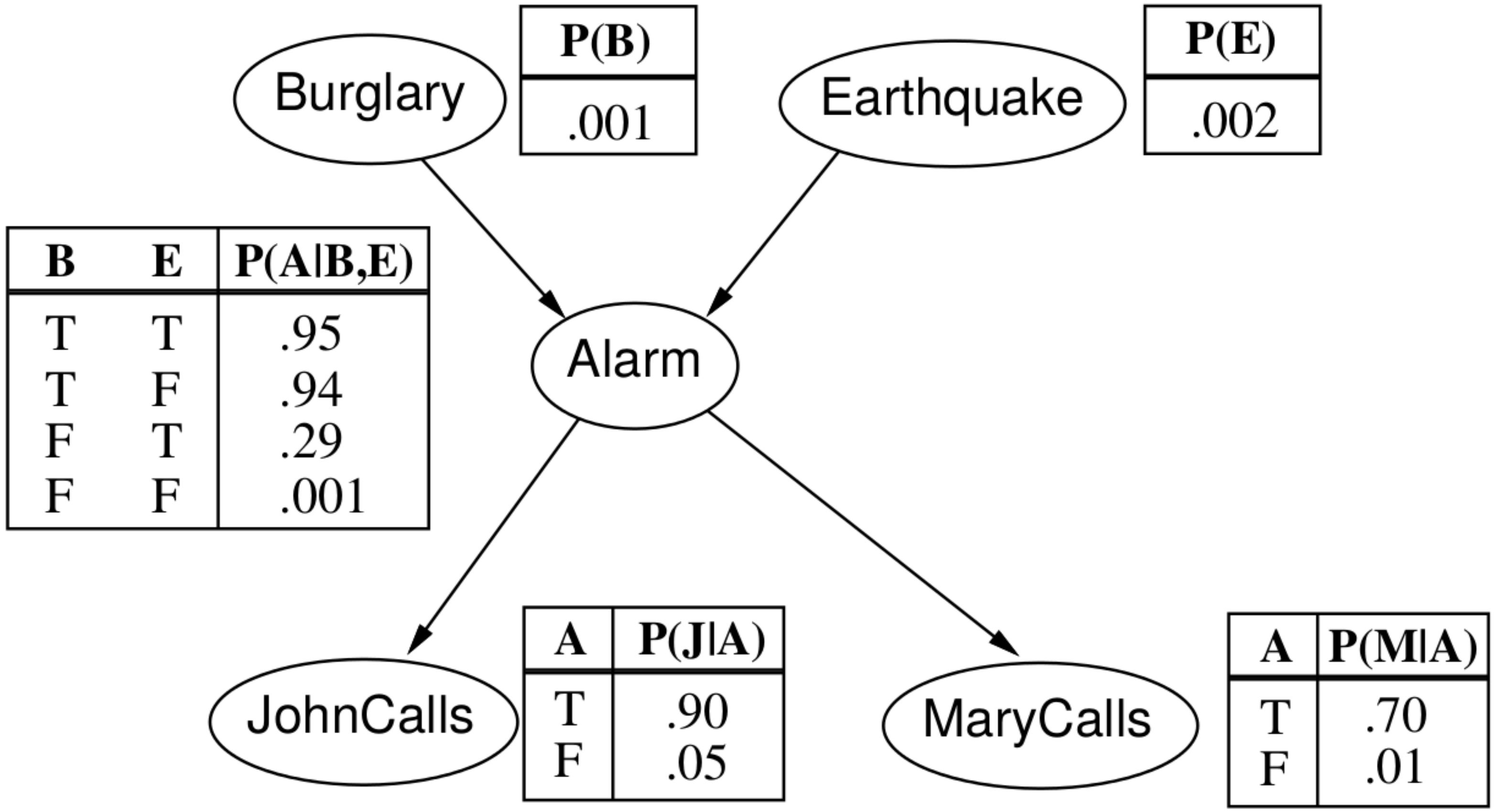
Example

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: $Burglar$, $Earthquake$, $Alarm$, $JohnCalls$, $MaryCalls$
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 -- Mary likes loud music so sometimes misses
- The alarm can cause John to call -- always calls, but sometimes confuses with telephone ringing,

Example contd.



P(B)
.001

P(E)
.002

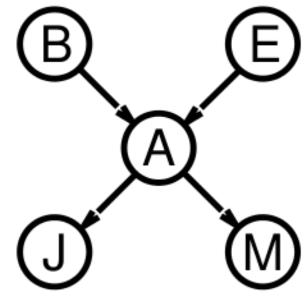
B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

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
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 $1 - p$)

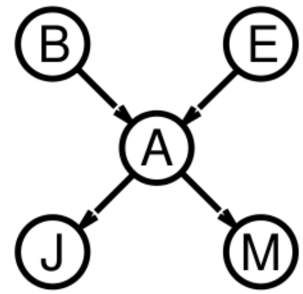
If each variable has no more than k parents,
the complete network requires $O(\quad)$ numbers

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

For burglary net, $1 + 1 + 4 + 2 + 2 = 10$ numbers (vs. $2^5 - 1 = 31$)

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 $1 - p$)

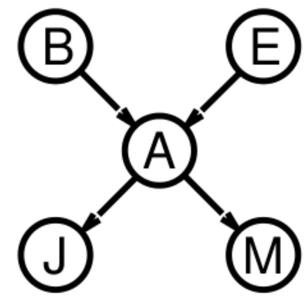
If each variable has no more than k parents, the complete network requires $O(n \cdot 2^k)$ numbers

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:



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

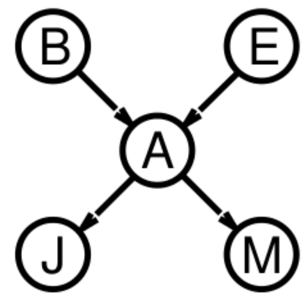
e.g., $P(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$

$=$

recall chain rule and conditional independence.

Global semantics

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



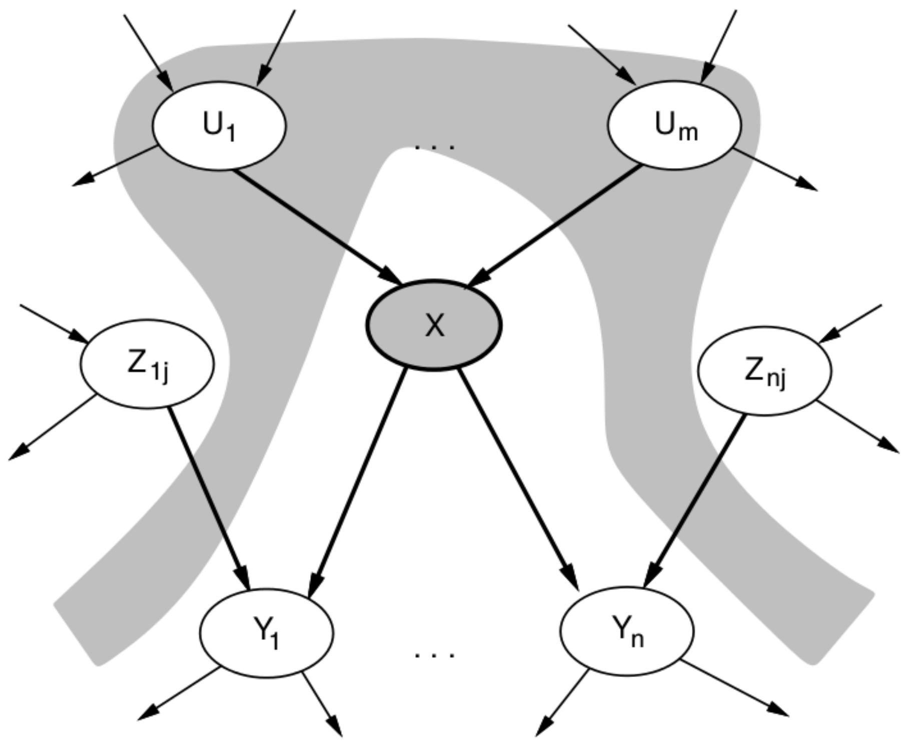
$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | parents(X_i))$$

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

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

Local semantics

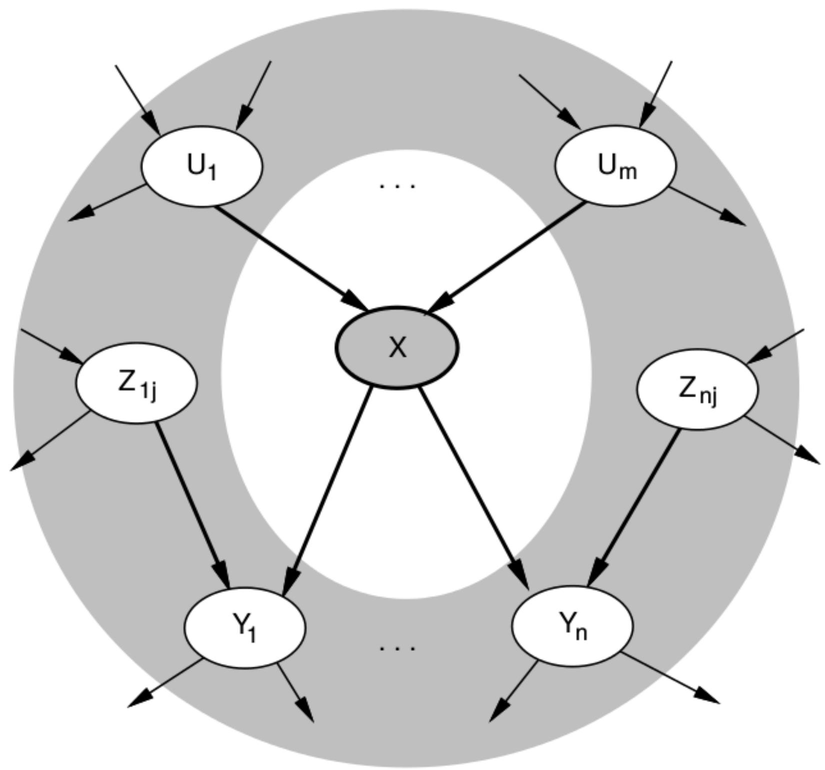
Local semantics: each node is conditionally independent of its nondescendants given its parents



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

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

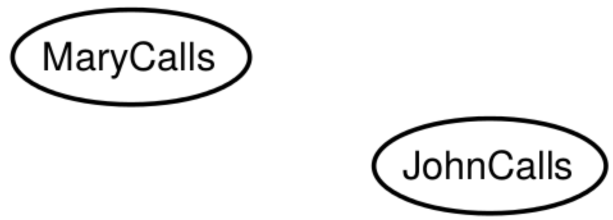
1. Choose an ordering of variables $X_1, \ldots, X_n$
2. For $i = 1$ to n
   add X_i to the network
   select parents from $X_1, \ldots, X_{i-1}$ such that
   $\mathbf{P}(X_i | Parents(X_i)) = \mathbf{P}(X_i | X_1,\ \ldots,\ X_{i-1})$

This choice of parents guarantees the global semantics:

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

Example

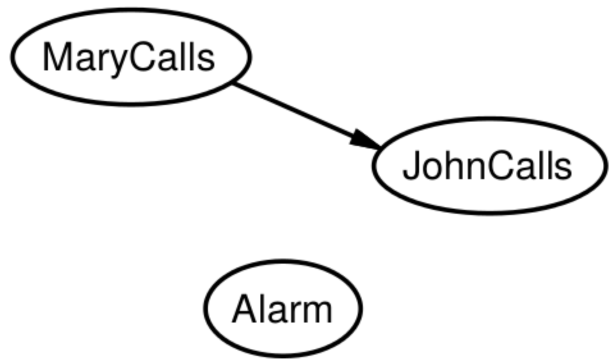
Suppose we choose the ordering M, J, A, B, E



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

Example

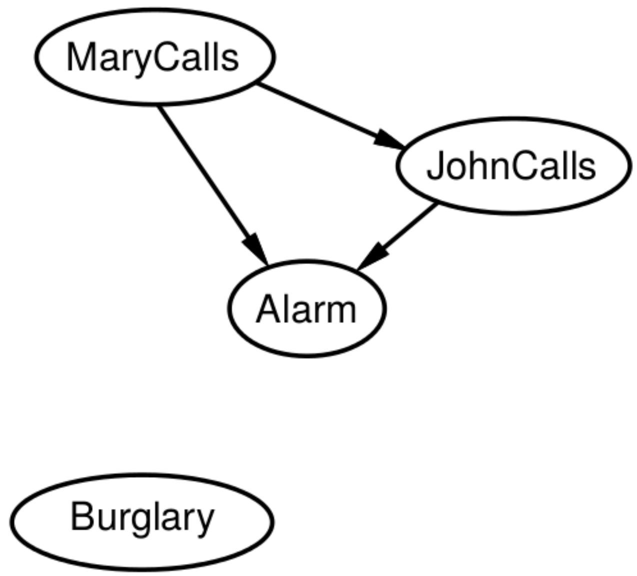
Suppose we choose the ordering M, J, A, B, E



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

Example

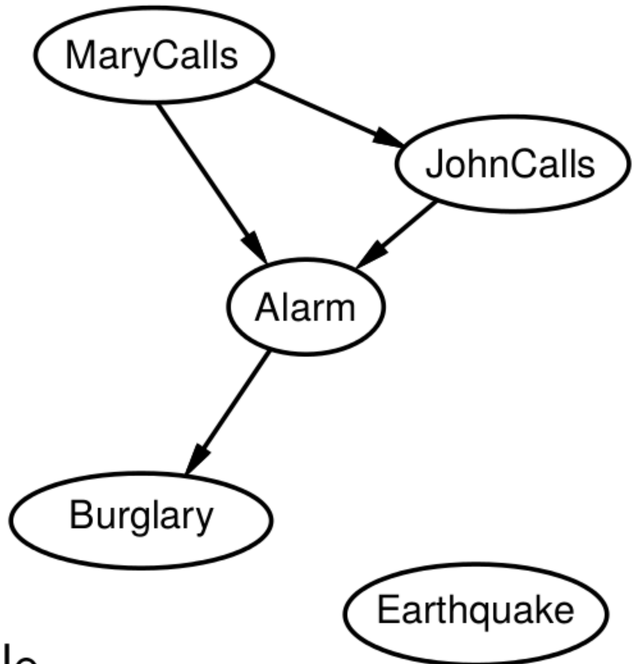
Suppose we choose the ordering M, J, A, B, E



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

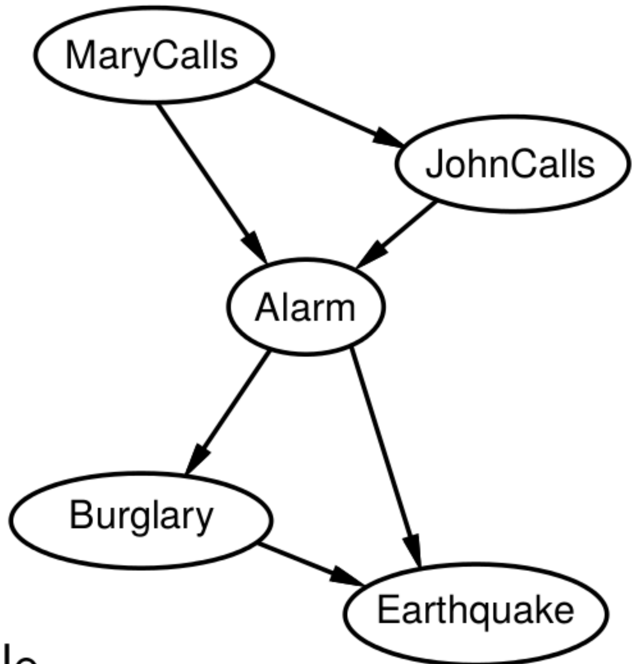
Suppose we choose the ordering M, J, A, B, E



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

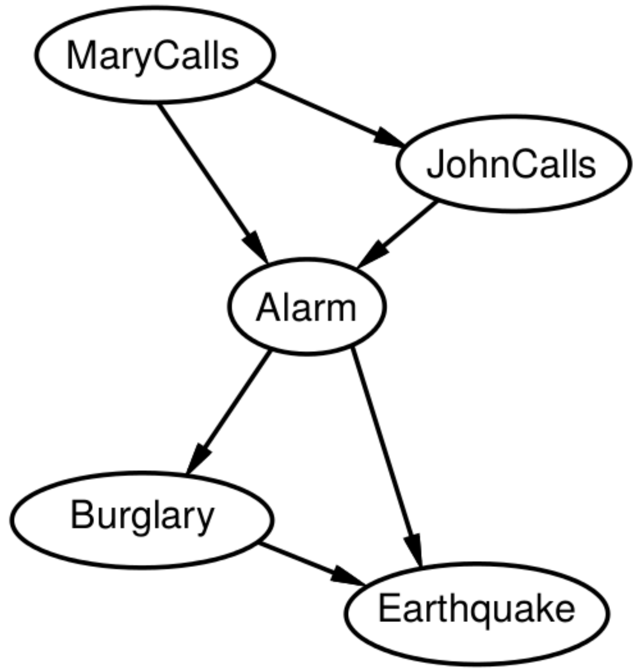
$P(B|A,J,M) = P(B|A)$? Yes

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

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

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

Example contd.



Deciding conditional independence is hard in noncausal directions

(Causal models and conditional independence seem hardwired for humans!)

Assessing conditional probabilities is hard in noncausal directions

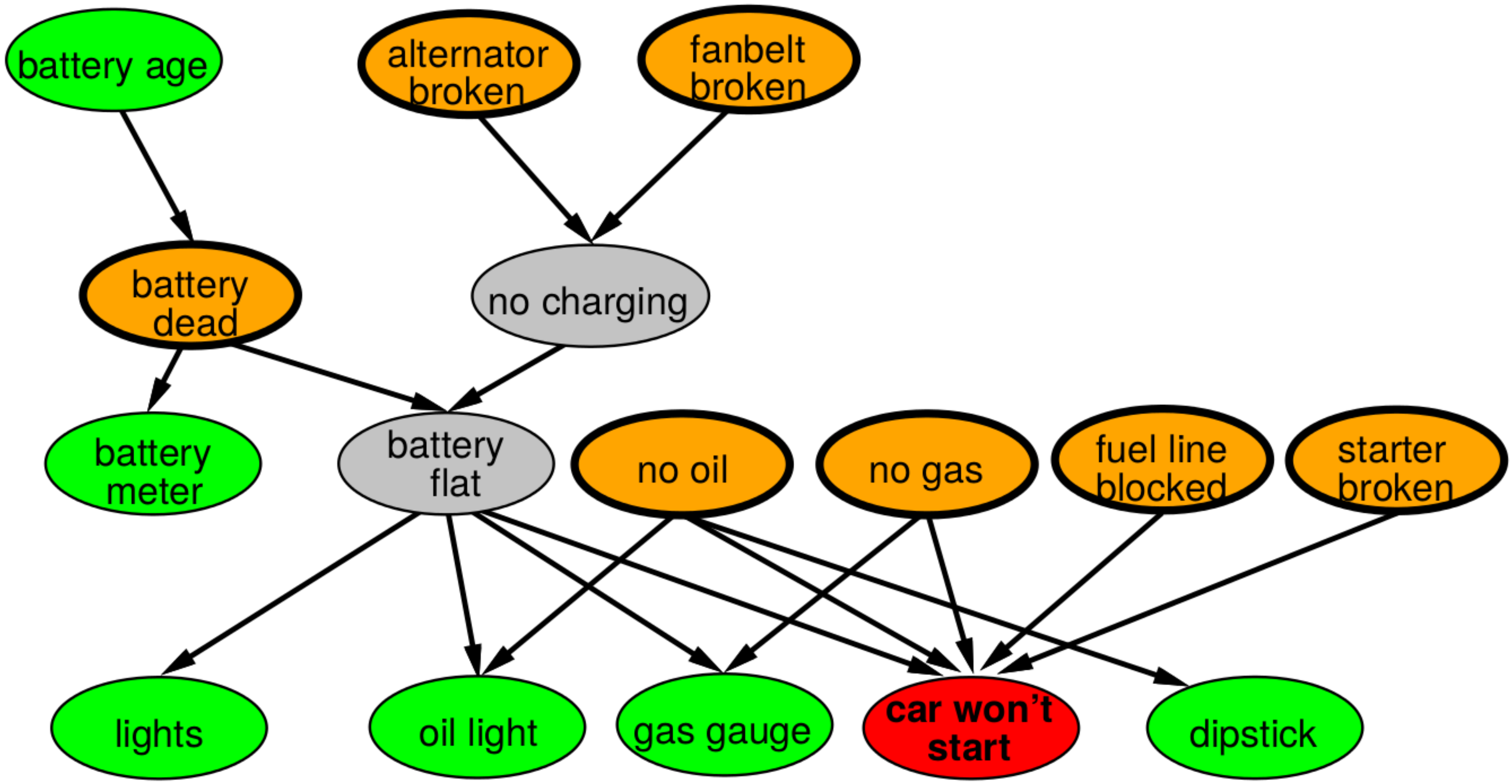
Network is less compact: $1 + 2 + 4 + 2 + 4 = 13$ numbers needed

Example: Car diagnosis

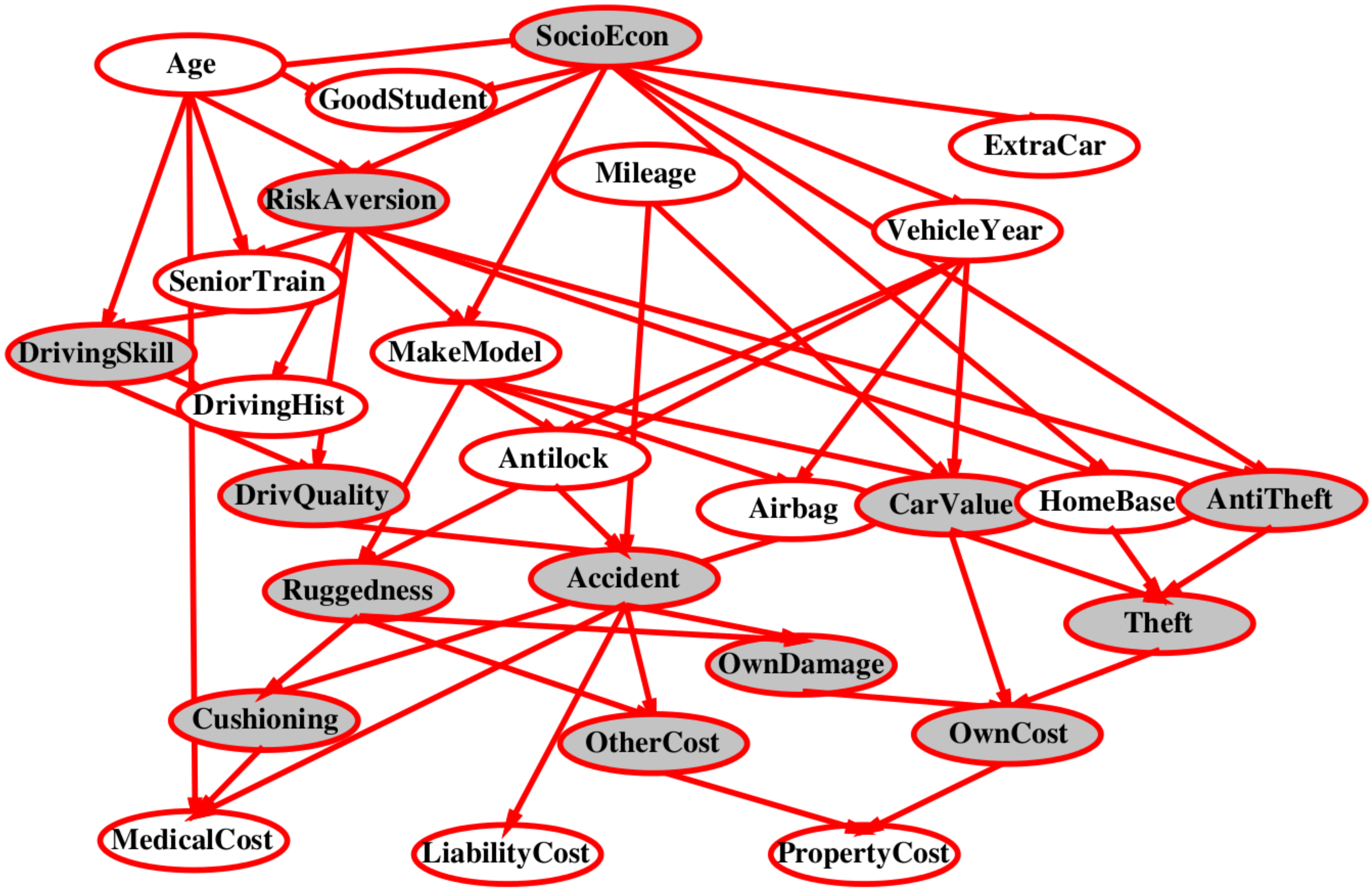
Initial evidence: car won't start

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

Hidden variables (gray) ensure sparse structure, reduce parameters



Example: Car insurance

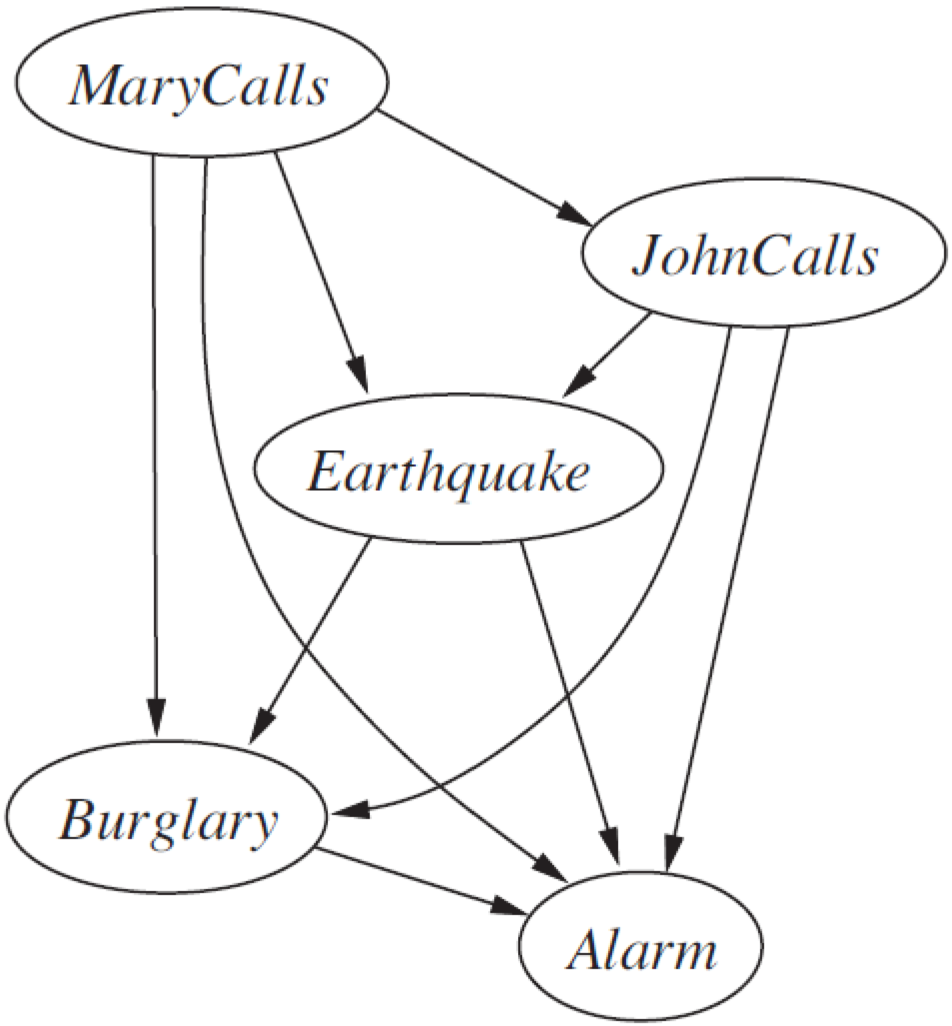


Constructing Bayesian Networks

- Quiz: MaryCalls, JohnCalls, Earthquake, Burglary, Alarm

Constructing Bayesian Networks

- MaryCalls, JohnCalls, Earthquake, Burglary, Alarm



Variable elimination algorithm

$$\mathbf{P}(B \mid j, m) = \alpha \underbrace{\mathbf{P}(B)}_{\mathbf{f}_1(B)} \sum_e \underbrace{P(e)}_{\mathbf{f}_2(E)} \sum_a \underbrace{\mathbf{P}(a \mid B, e)}_{\mathbf{f}_3(A,B,E)} \underbrace{P(j \mid a)}_{\mathbf{f}_4(A)} \underbrace{P(m \mid a)}_{\mathbf{f}_5(A)}$$

- Annotate each part of expression with the name of the corresponding **factor**
- Each factor is a matrix indexed by the values of its argument variables
- f_4(A) and f_5(A) corresponding to P(j | a) and P(m| a) depend just on A because J and M are fixed by the query.

$$\mathbf{f}_4(A) = \begin{pmatrix} P(j \mid a) \\ P(j \mid \neg a) \end{pmatrix} = \begin{pmatrix} 0.90 \\ 0.05 \end{pmatrix} \qquad \mathbf{f}_5(A) = \begin{pmatrix} P(m \mid a) \\ P(m \mid \neg a) \end{pmatrix} = \begin{pmatrix} 0.70 \\ 0.01 \end{pmatrix}$$

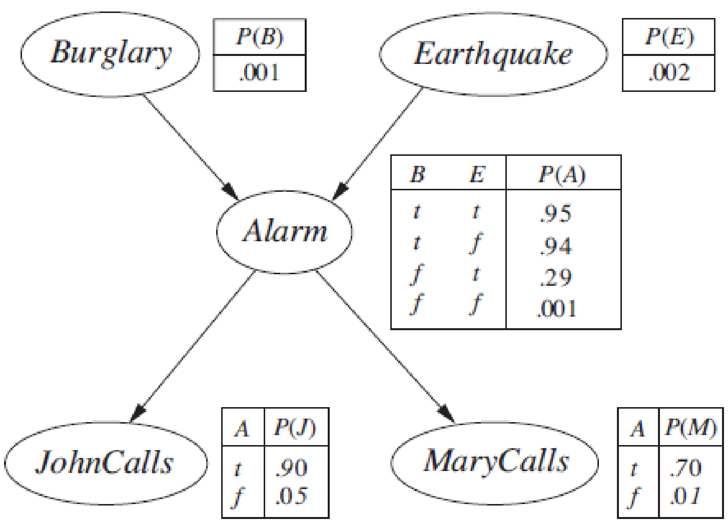
Variable elimination algorithm

$$\mathbf{P}(B \mid j, m) = \alpha \underbrace{\mathbf{P}(B)}_{\mathbf{f}_1(B)} \sum_e \underbrace{P(e)}_{\mathbf{f}_2(E)} \sum_a \underbrace{\mathbf{P}(a \mid B, e)}_{\mathbf{f}_3(A,B,E)} \underbrace{P(j \mid a)}_{\mathbf{f}_4(A)} \underbrace{P(m \mid a)}_{\mathbf{f}_5(A)}$$

$$\mathbf{f}_4(A) = \begin{pmatrix} P(j \mid a) \\ P(j \mid \neg a) \end{pmatrix} = \begin{pmatrix} 0.90 \\ 0.05 \end{pmatrix} \qquad \mathbf{f}_5(A) = \begin{pmatrix} P(m \mid a) \\ P(m \mid \neg a) \end{pmatrix} = \begin{pmatrix} 0.70 \\ 0.01 \end{pmatrix}$$

- f_3(A,B,E) will be a 2×2×2 matrix

A	B	E	p(A|B,E)
T	T	T	0.95
T	T	F	0.94
T	F	T	0.29
T	F	F	0.01
F	T	T	0.05
F	T	F	0.06
F	F	T	0.71
F	F	F	0.999



Variable elimination algorithm

$$\mathbf{P}(B \mid j, m) = \alpha \underbrace{\mathbf{P}(B)}_{\mathbf{f}_1(B)} \sum_e \underbrace{P(e)}_{\mathbf{f}_2(E)} \sum_a \underbrace{\mathbf{P}(a \mid B, e)}_{\mathbf{f}_3(A,B,E)} \underbrace{P(j \mid a)}_{\mathbf{f}_4(A)} \underbrace{P(m \mid a)}_{\mathbf{f}_5(A)}$$

$$\mathbf{P}(B \mid j, m) = \alpha \, \mathbf{f}_1(B) \times \sum_e \mathbf{f}_2(E) \times \sum_a \mathbf{f}_3(A, B, E) \times \mathbf{f}_4(A) \times \mathbf{f}_5(A)$$

- “x” operator is not ordinary matrix multiplication but instead the **pointwise product** operation
- The pointwise product of two factors f_1 and f_2 yields a new factor f whose variables are the union of the variables in f_1 and f_2 and whose elements are given by the product of the corresponding elements in the two factors.

Variable elimination algorithm

$$\mathbf{P}(B \mid j, m) = \alpha\, \mathbf{f}_1(B) \times \sum_e \mathbf{f}_2(E) \times \sum_a \mathbf{f}_3(A, B, E) \times \mathbf{f}_4(A) \times \mathbf{f}_5(A)$$

- Pointwise product operation:

A	B	$\mathbf{f}_1(A,B)$	B	C	$\mathbf{f}_2(B,C)$	A	B	C	$\mathbf{f}_3(A,B,C)$
T	T	.3	T	T	.2	T	T	T	$.3 \times .2 = .06$
T	F	.7	T	F	.8	T	T	F	$.3 \times .8 = .24$
F	T	.9	F	T	.6	T	F	T	$.7 \times .6 = .42$
F	F	.1	F	F	.4	T	F	F	$.7 \times .4 = .28$
						F	T	T	$.9 \times .2 = .18$
						F	T	F	$.9 \times .8 = .72$
						F	F	T	$.1 \times .6 = .06$
						F	F	F	$.1 \times .4 = .04$

Figure 14.10 Illustrating pointwise multiplication: $\mathbf{f}_1(A,B) \times \mathbf{f}_2(B,C) = \mathbf{f}_3(A,B,C)$.

Variable elimination algorithm

$$\mathbf{P}(B \mid j, m) = \alpha\, \mathbf{f}_1(B) \times \sum_e \mathbf{f}_2(E) \times \sum_a \mathbf{f}_3(A, B, E) \times \mathbf{f}_4(A) \times \mathbf{f}_5(A)$$

- Summation operation: Summing out a variable from a product of factors is done by adding up the submatrices formed by fixing the variable to each of its values in turn

$$\begin{aligned}\mathbf{f}(B, C) &= \sum_a \mathbf{f}_3(A, B, C) = \mathbf{f}_3(a, B, C) + \mathbf{f}_3(\neg a, B, C) \\ &= \begin{pmatrix} .06 & .24 \\ .42 & .28 \end{pmatrix} + \begin{pmatrix} .18 & .72 \\ .06 & .04 \end{pmatrix} = \begin{pmatrix} .24 & .96 \\ .48 & .32 \end{pmatrix}.\end{aligned}$$

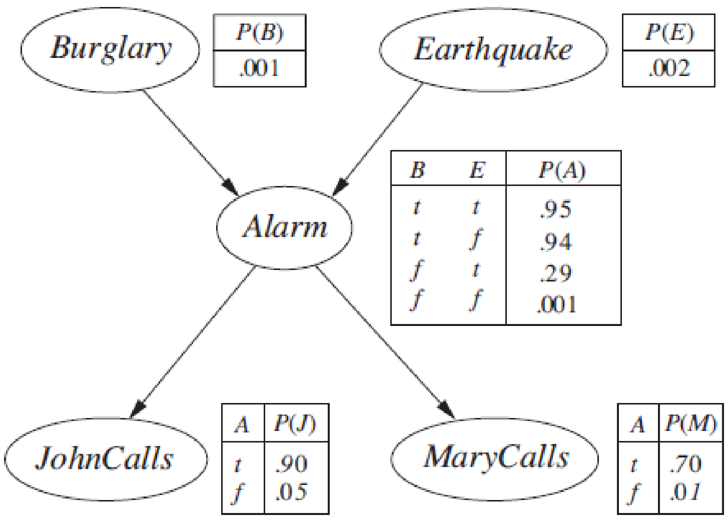
Variable elimination algorithm

$$\mathbf{P}(B \mid j, m) = \alpha \underbrace{\mathbf{P}(B)}_{\mathbf{f}_1(B)} \sum_e \underbrace{P(e)}_{\mathbf{f}_2(E)} \sum_a \underbrace{\mathbf{P}(a \mid B, e)}_{\mathbf{f}_3(A,B,E)} \underbrace{P(j \mid a)}_{\mathbf{f}_4(A)} \underbrace{P(m \mid a)}_{\mathbf{f}_5(A)}$$

$$\mathbf{P}(B \mid j, m) = \alpha\, \mathbf{f}_1(B) \times \sum_e \mathbf{f}_2(E) \times \sum_a \mathbf{f}_3(A, B, E) \times \mathbf{f}_4(A) \times \mathbf{f}_5(A)$$

$$\mathbf{f}_4(A) = \begin{pmatrix} P(j \mid a) \\ P(j \mid \neg a) \end{pmatrix} = \begin{pmatrix} 0.90 \\ 0.05 \end{pmatrix} \qquad \mathbf{f}_5(A) = \begin{pmatrix} P(m \mid a) \\ P(m \mid \neg a) \end{pmatrix} = \begin{pmatrix} 0.70 \\ 0.01 \end{pmatrix}$$

A	B	E	f_3(A,B,E)
T	T	T	0.95
T	T	F	0.94
T	F	T	0.29
T	F	F	0.01
F	T	T	0.05
F	T	F	0.06
F	F	T	0.71
F	F	F	0.999



Variable elimination algorithm

$$\mathbf{P}(B \mid j, m) = \alpha\, \mathbf{f}_1(B) \times \sum_e \mathbf{f}_2(E) \times \sum_a \mathbf{f}_3(A, B, E) \times \mathbf{f}_4(A) \times \mathbf{f}_5(A)$$

- The trick to notice is that any factor that does not depend on the variable to be summed out can be moved outside the summation.

$$\sum_e \mathbf{f}_2(E) \times \mathbf{f}_3(A, B, E) \times \mathbf{f}_4(A) \times \mathbf{f}_5(A) = \mathbf{f}_4(A) \times \mathbf{f}_5(A) \times \sum_e \mathbf{f}_2(E) \times \mathbf{f}_3(A, B, E)$$

- Different orderings cause different intermediate factors to be generated during the calculation

$$\mathbf{P}(B \mid j, m) = \alpha\, \mathbf{f}_1(B) \times \sum_a \mathbf{f}_4(A) \times \mathbf{f}_5(A) \times \sum_e \mathbf{f}_2(E) \times \mathbf{f}_3(A, B, E)$$

- Idea: Eliminate whichever variable minimizes the size of the next factor to be constructed.