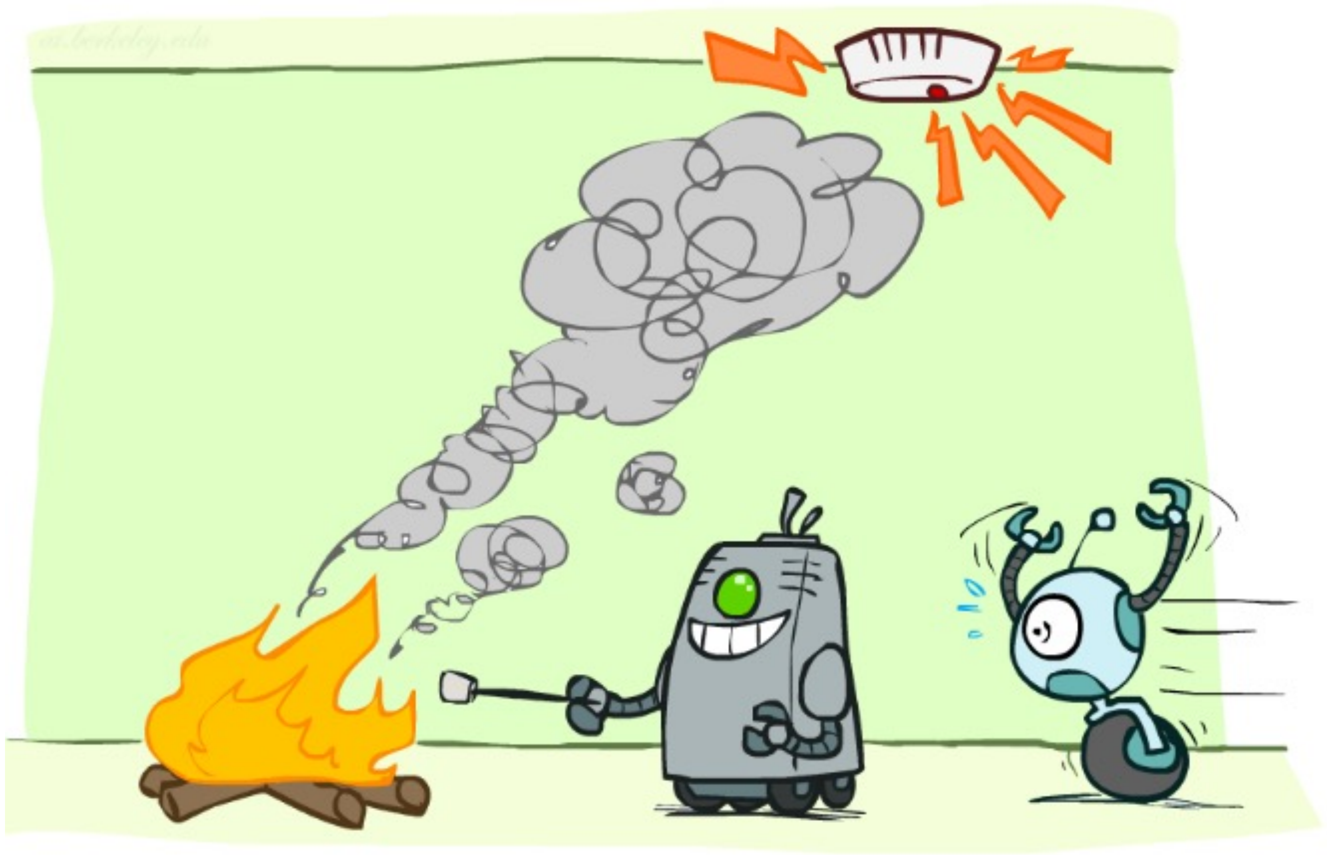
Reminder: elementary probability

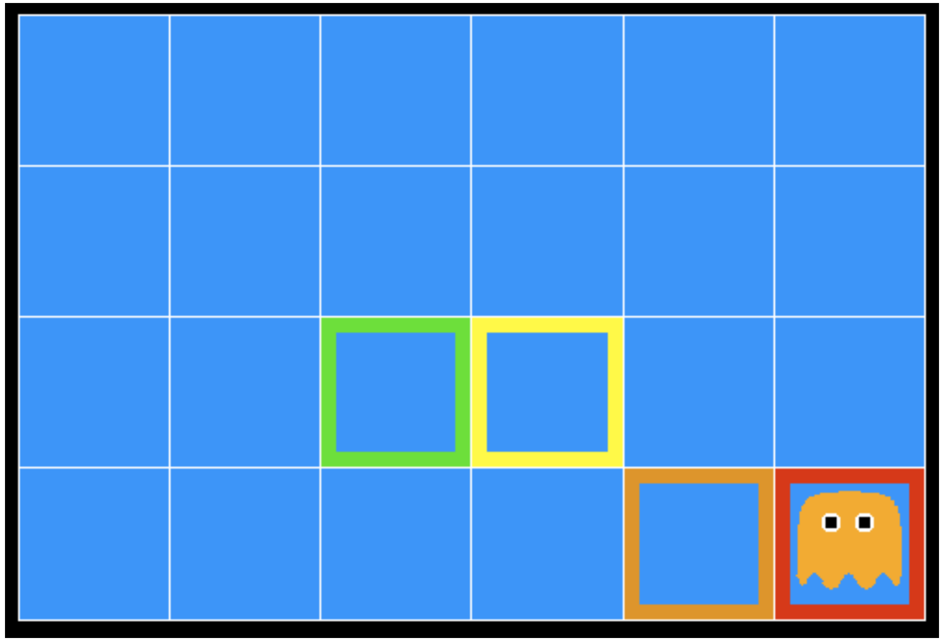
- Basic laws: $0 \leq P(\omega) \leq 1$ $\quad \sum_{\omega \in \Omega} P(\omega) = 1$
- Events: subsets of Ω: $P(A) = \sum_{\omega \in A} P(\omega)$
- Random variable $X(\omega)$ has a value in each ω
  - Distribution $P(X)$ gives probability for each possible value x
  - Joint distribution $P(X,Y)$ gives total probability for each combination x,y
- Summing out/marginalization: $P(X=x) = \sum_y P(X=x,Y=y)$
- Conditional probability: $P(X|Y) = \mathrm{P}(X,Y)/\mathrm{P}(Y)$
- Product rule: $P(X|Y)P(Y) = \mathrm{P}(X,Y) = P(Y|X)P(X)$
  - Generalize to chain rule: $P(X_1,..,X_n) = \prod_i P(X_i \mid X_1,..,X_{i-1})$

Conditional Independence



Ghostbusters

- A ghost is in the grid somewhere
- Sensor readings tell how close a square is to the ghost
  - On the ghost: usually red
  - 1 or 2 away: mostly orange
  - 3 or 4 away: typically yellow
  - 5+ away: often green
- Click on squares until confident of location, then "***bust***"



Video of Demo Ghostbusters with Probability

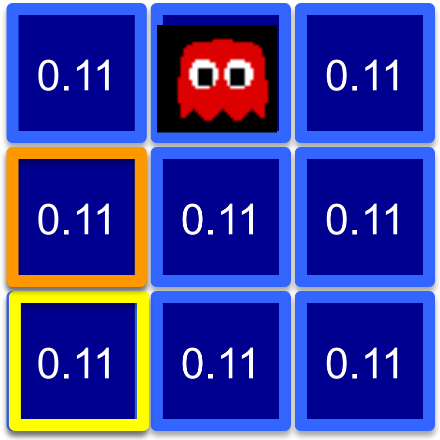
Ghostbusters model

- Variables and ranges:
  - G (ghost location) in {(1,1),…,(3,3)}
  - $C_{x,y}$ (color measured at square x,y) in {red,orange,yellow,green}

0.11	0.11	0.11
0.11	0.11	0.11
0.11	0.11	0.11

- Ghostbuster physics:
  - ***Uniform prior distribution*** over ghost location: $P(G)$
  - ***Sensor model***: $P(C_{x,y} \mid G)$ (depends only on distance to G)
    - E.g. $P(C_{1,1} = \text{yellow} \mid G = (1,1)\ ) = 0.1$

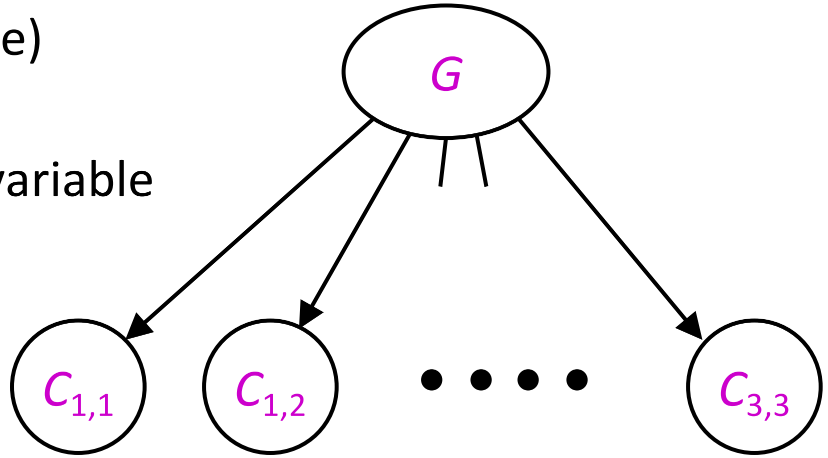
Ghostbusters model, contd.



- $P(G, C_{1,1}, \ldots C_{3,3})$ has $9 \times 4^9 = 2{,}359{,}296$ entries!!!
- Ghostbuster independence:
  - Are $C_{1,1}$ and $C_{1,2}$ independent?
    - E.g., does $P(C_{1,1} = \text{yellow}) = P(C_{1,1} = \text{yellow} \mid C_{1,2} = \text{orange})$?
- Ghostbuster physics again:
  - $P(C_{x,y} \mid G)$ ***depends only on distance to G***
    - So $P(C_{1,1} = \text{yellow} \mid G = (2,3)) = P(C_{1,1} = \text{yellow} \mid G = (2,3), C_{1,2} = \text{orange})$
    - I.e., $C_{1,1}$ is ***conditionally independent*** of $C_{1,2}$ ***given*** G

Ghostbusters model, contd.

- Apply the chain rule to decompose the joint probability model:
- $P(G, C_{1,1}, \ldots C_{3,3}) = P(G)\, P(C_{1,1} \mid G)\, P(C_{1,2} \mid G, C_{1,1})\, P(C_{1,3} \mid G, C_{1,1}, C_{1,2}) \ldots P(C_{3,3} \mid G, C_{1,1}, \ldots, C_{3,2})$
- Now simplify using conditional independence:
- $P(G, C_{1,1}, \ldots C_{3,3}) = P(G)\, P(C_{1,1} \mid G)\, P(C_{1,2} \mid G)\, P(C_{1,3} \mid G) \ldots P(C_{3,3} \mid G)$
- I.e., conditional independence properties of ghostbuster physics simplify the probability model from ***exponential*** to ***quadratic*** in the number of squares
- This is called a ***Naïve Bayes*** model:
  - One discrete query variable (often called the ***class*** or ***category*** variable)
  - All other variables are (potentially) evidence variables
  - Evidence variables are all conditionally independent given the query variable



Conditional Independence

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

- X is conditionally independent of Y given Z if and only if:

$$\forall x,y,z \qquad P(x \mid y, z) = P(x \mid z)$$

  or, equivalently, if and only if

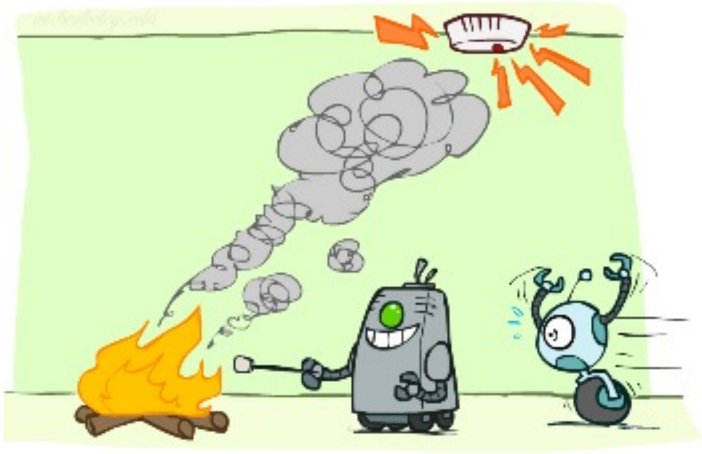
$$\forall x,y,z \qquad P(x, y \mid z) = P(x \mid z)\, P(y \mid z)$$

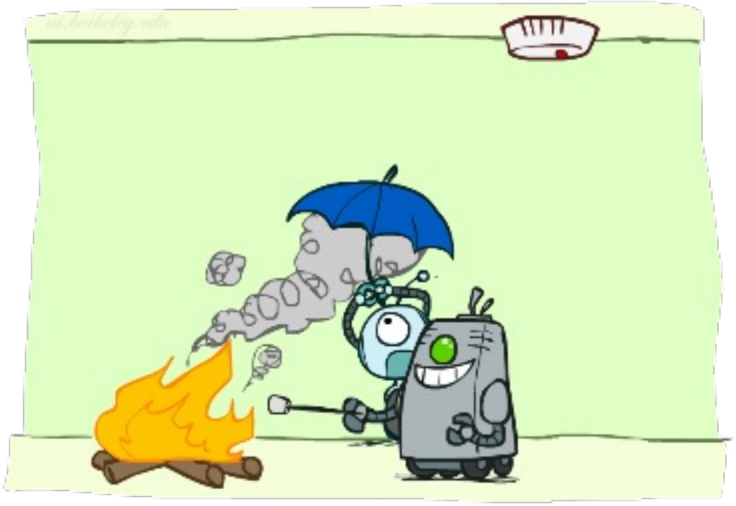
Conditional Independence

- What about this domain:
  - Traffic
  - Umbrella
  - Raining

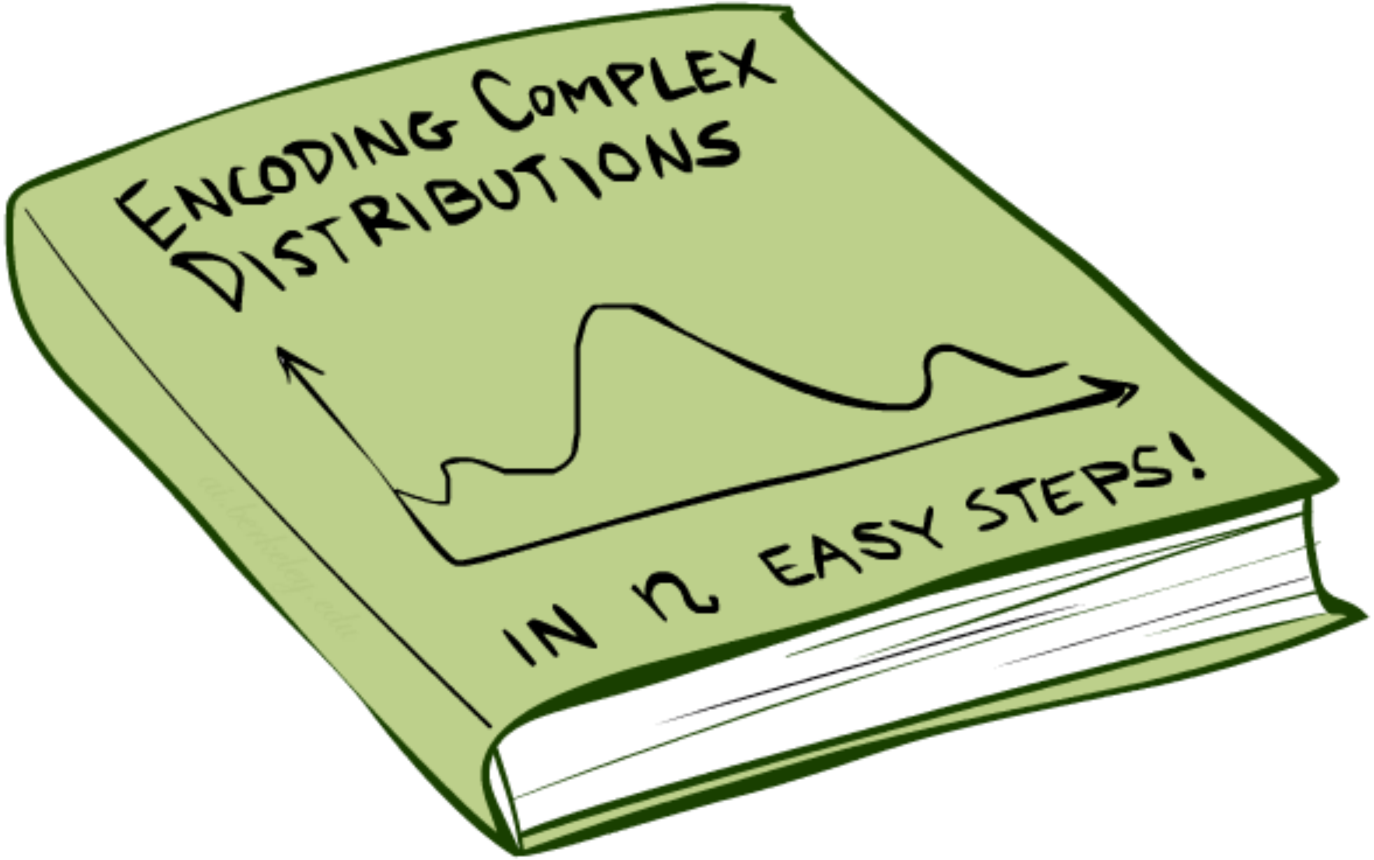
Conditional Independence

- What about this domain:
  - Fire
  - Smoke
  - Alarm



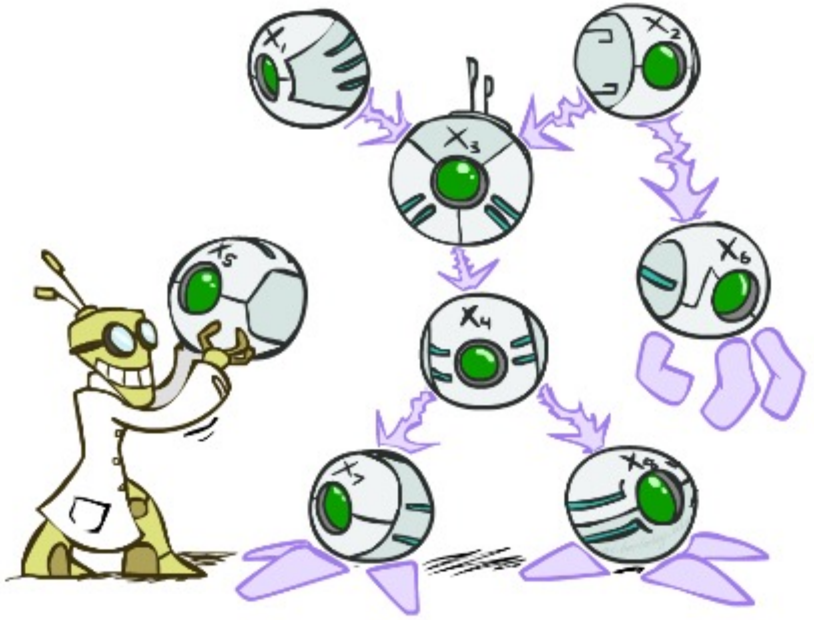


Bayes Nets: Big Picture



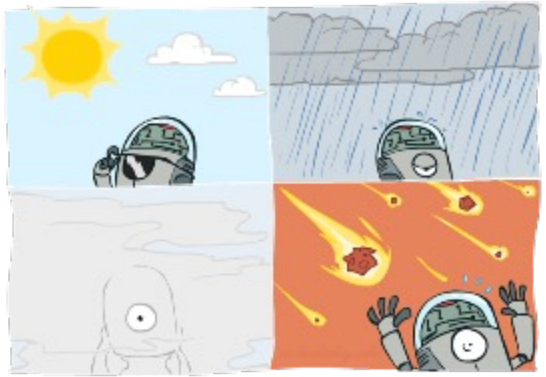
Bayes Nets: Big Picture

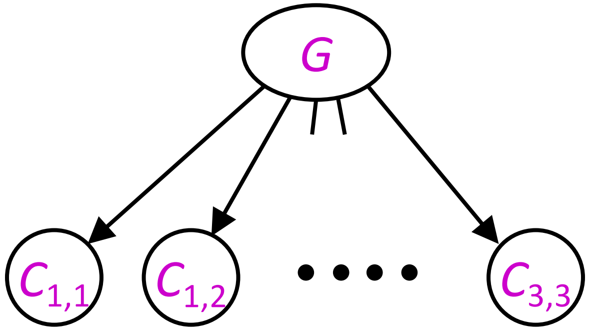
- Bayes nets: a technique for describing complex joint distributions (models) using simple, conditional distributions
  - A subset of the general class of graphical models
- Use local causality/conditional independence:
  - the world is composed of many variables,
  - each interacting locally with a few others
- Outline
  - Representation
  - Exact inference
  - Approximate inference

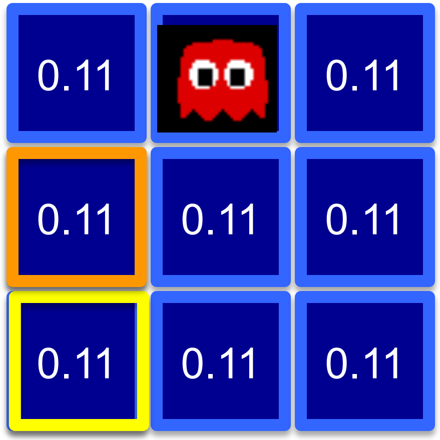


Graphical Model Notation

- Nodes: variables (with domains)
  - Can be assigned (observed) or unassigned (unobserved)
- Arcs: interactions
  - Indicate "direct influence" between variables
  - Formally: absence of arc encodes conditional independence (more later)

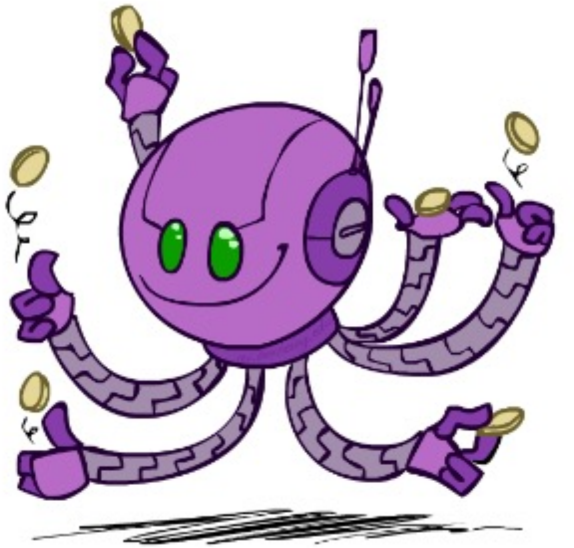






Example: Coin Flips

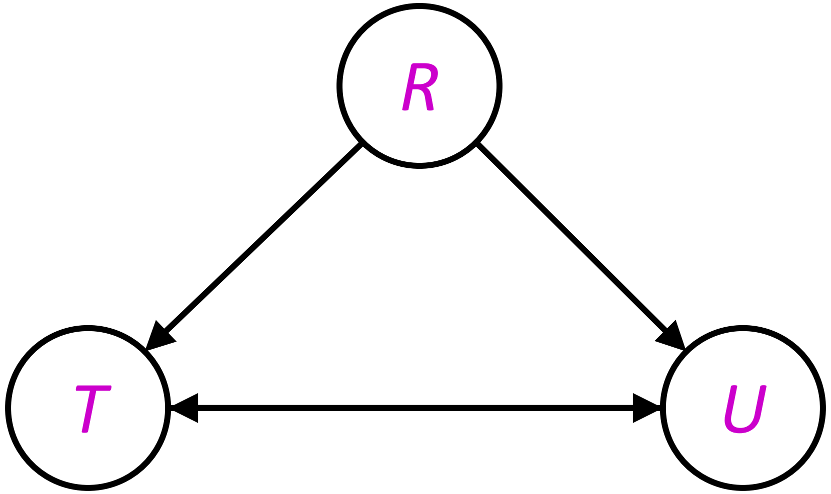
- *n* independent coin flips



- No interactions between variables: absolute independence

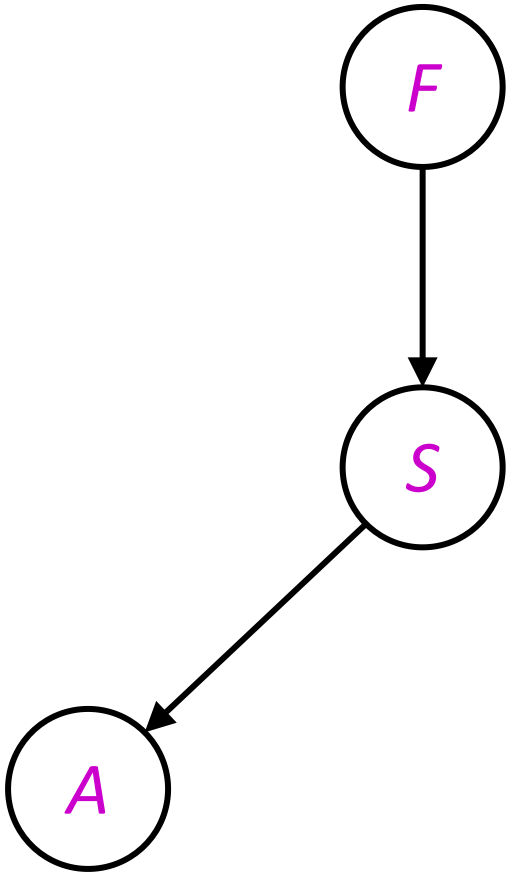
Example: Traffic

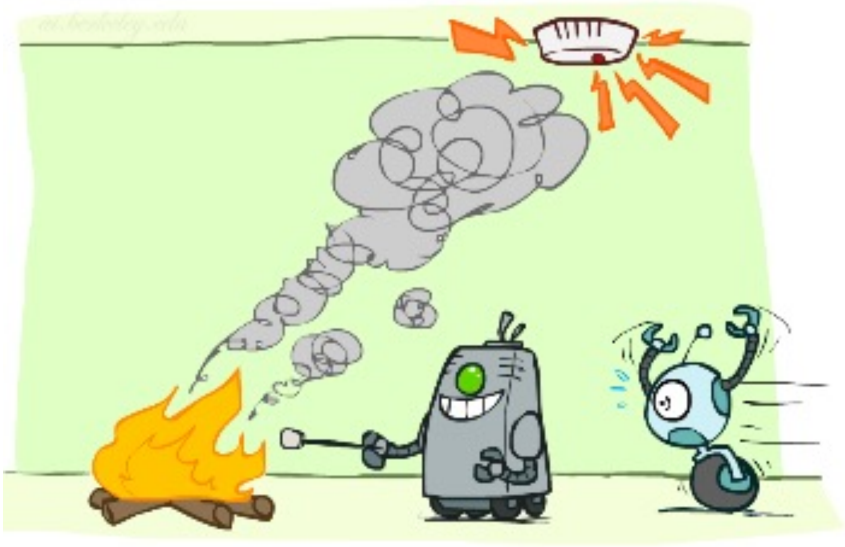
- Variables:
  - T: There is traffic
  - U: I'm holding my umbrella
  - R: It rains



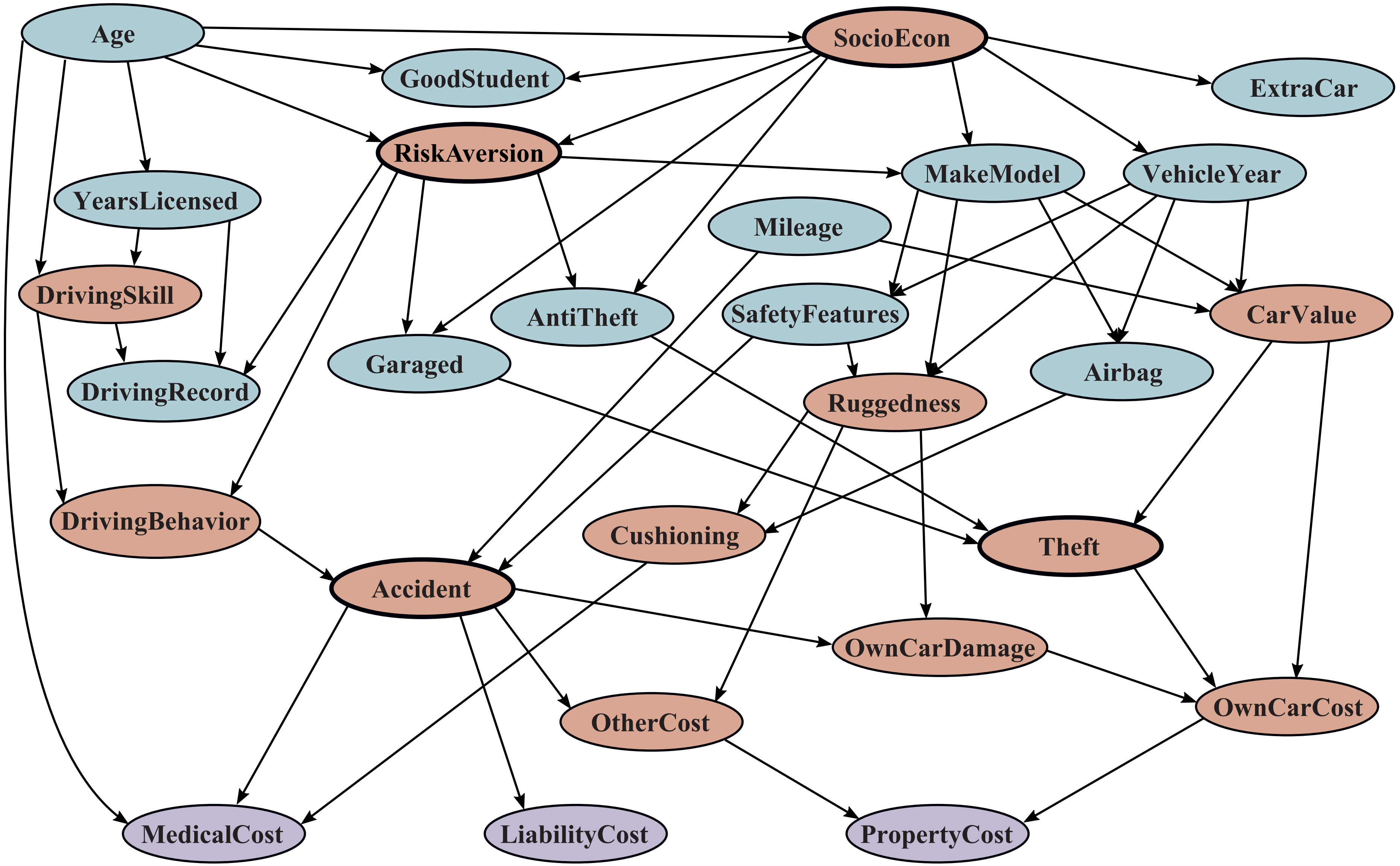
Example: Smoke alarm

- Variables:
  - F: There is fire
  - S: There is smoke
  - A: Alarm sounds

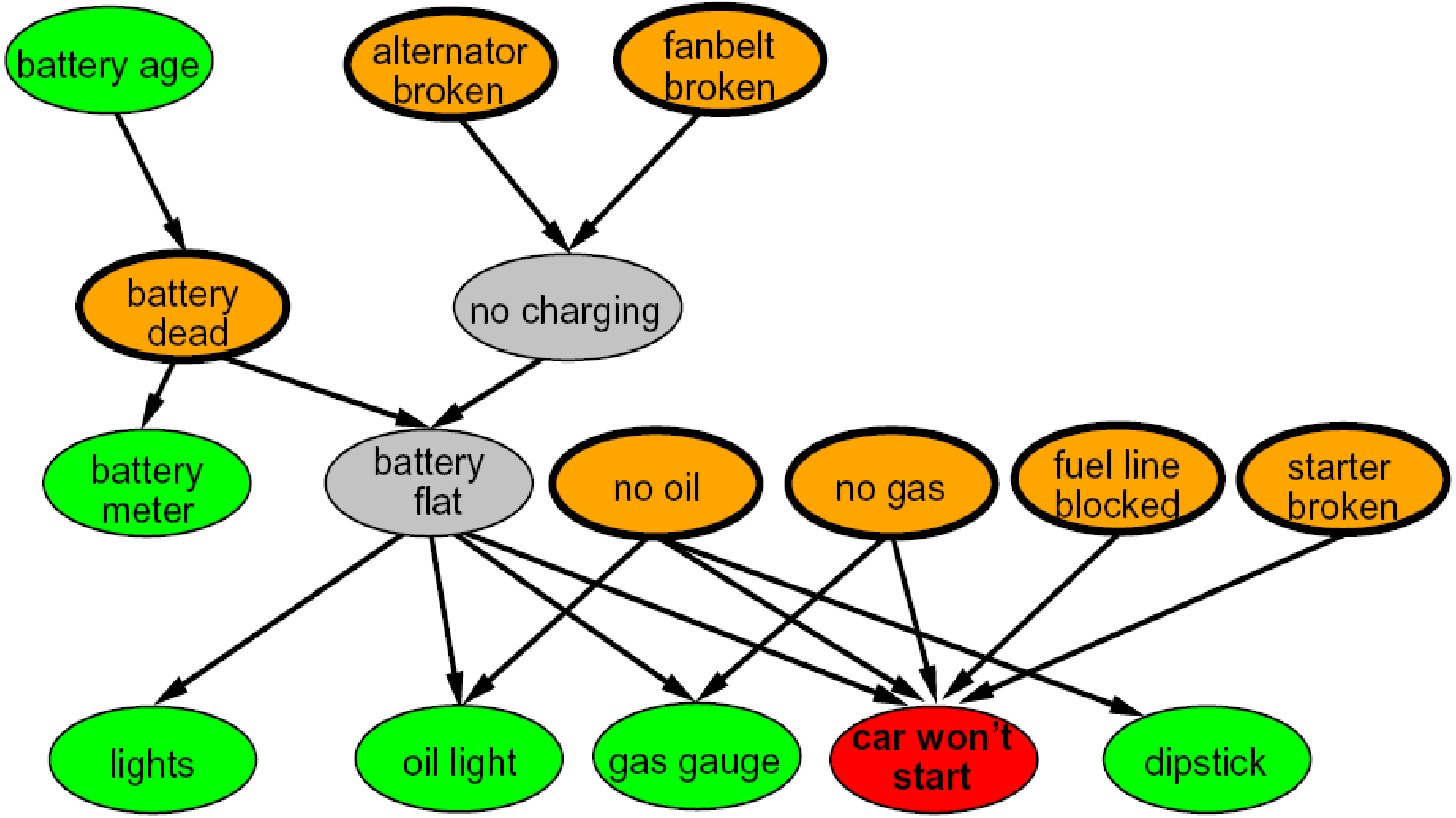




Example Bayes' Net: Car Insurance



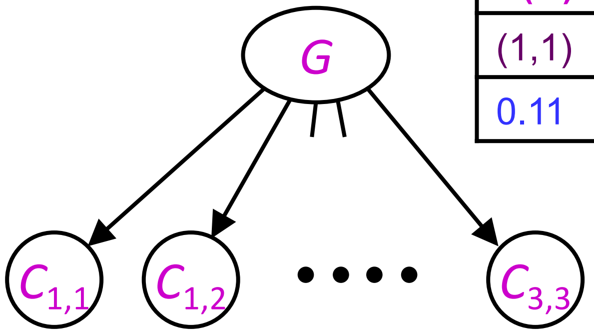
Example Bayes' Net: Car Won't Start



Bayes Net Syntax and Semantics

Bayes Net Syntax

- A set of nodes, one per variable X_i
- A directed, acyclic graph
- A conditional distribution for each node given its ***parent variables*** in the graph
  - ***CPT*** (conditional probability table); each row is a distribution for child given values of its parents



P(*G*)			
(1,1)	(1,2)	(1,3)	…
0.11	0.11	0.11	…

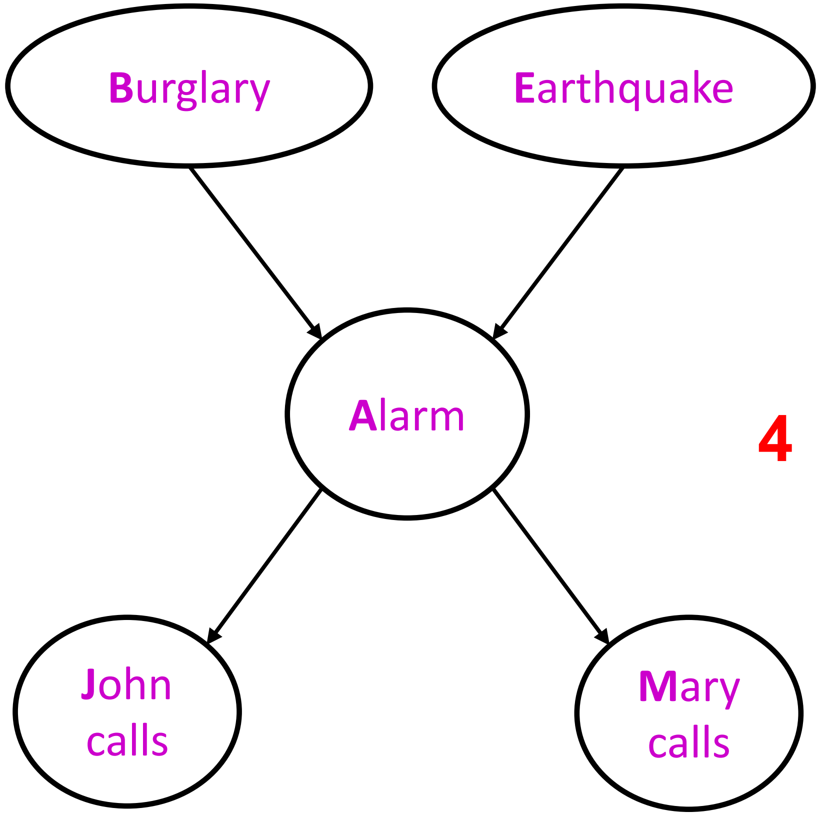
G	**P(*$C_{1,1}$* | *G*)**			
	g	y	o	r
(1,1)	0.01	0.1	0.3	0.59
(1,2)	0.1	0.3	0.5	0.1
(1,3)	0.3	0.5	0.19	0.01
…				

Bayes net = Topology (graph) + Local Conditional Probabilities

Example: Alarm Network

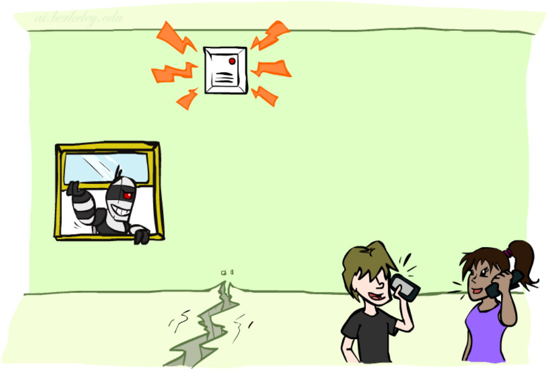
P(B)	
true	false
0.001	0.999

1



1

P(E)	
true	false
0.002	0.998



4

B	E	P(A|B,E)	
		true	false
true	true	0.95	0.05
true	false	0.94	0.06
false	true	0.29	0.71
false	false	0.001	0.999

A	P(J|A)	
	true	false
true	0.9	0.1
false	0.05	0.95

2

2

A	P(M|A)	
	true	false
true	0.7	0.3
false	0.01	0.99

Number of ***free parameters*** in each CPT:

Parent range sizes $d_1,\ldots,d_k$

Child range size d

Each table row must sum to 1

$(d-1)\ \Pi_i\ d_i$

General formula for sparse BNs

- Suppose
  - n variables
  - Maximum range size is d
  - Maximum number of parents is k
- Full joint distribution has size $O(d^n)$
- Bayes net has size $O(n \cdot d^k)$
  - Linear scaling with n as long as causal structure is local

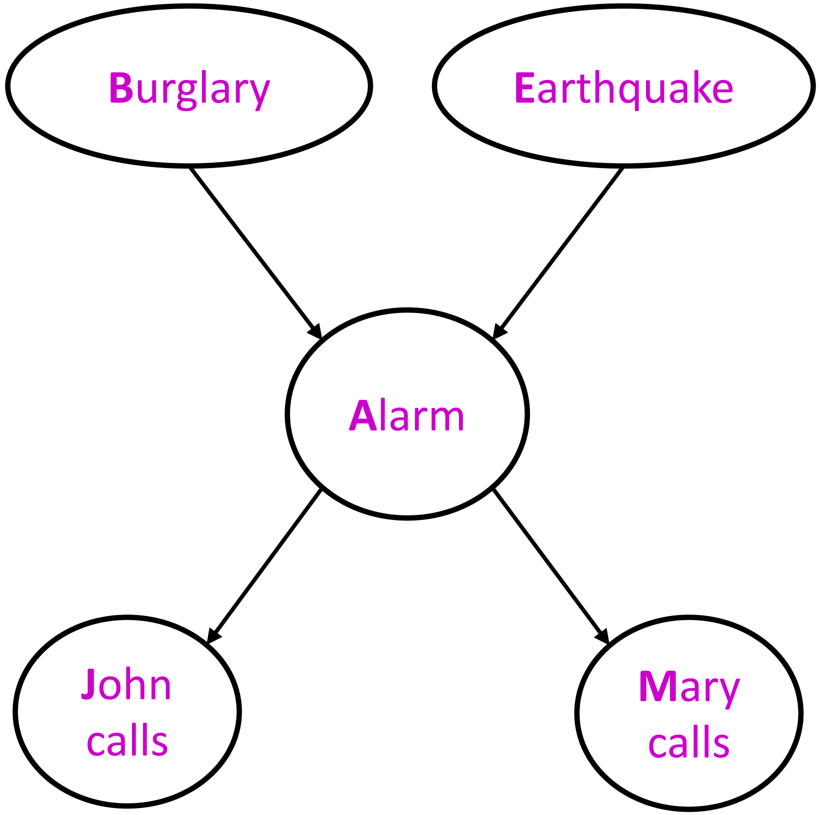
Bayes net global semantics

- Bayes nets encode joint distributions as product of conditional distributions on each variable:

$$P(X_1,..,X_n) = \prod_i P(X_i \mid Parents(X_i))$$

Example

P(B)	
true	false
0.001	0.999



P(E)	
true	false
0.002	0.998

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

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

$$= .001 \times .998 \times .94 \times .1 \times .3 = .000028$$

B	E	P(A|B,E)	
		true	false
true	true	0.95	0.05
true	false	0.94	0.06
false	true	0.29	0.71
false	false	0.001	0.999

A	P(J|A)	
	true	false
true	0.9	0.1
false	0.05	0.95

A	P(M|A)	
	true	false
true	0.7	0.3
false	0.01	0.99

Conditional independence in BNs

- Compare the Bayes net global semantics

$$P(X_1,..,X_n) = \prod_i P(X_i \mid Parents(X_i))$$

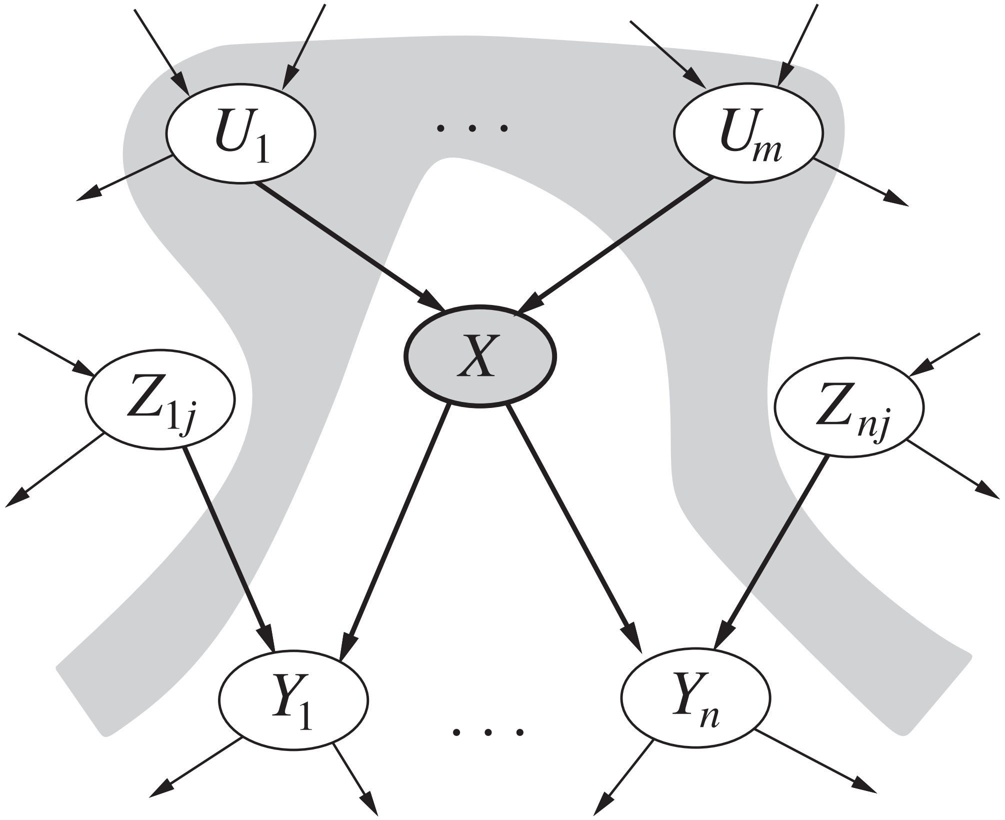
  with the chain rule identity

$$P(X_1,..,X_n) = \prod_i P(X_i \mid X_1,\ldots,X_{i-1})$$

- Assume (without loss of generality) that $X_1,..,X_n$ sorted in topological order according to the graph (i.e., parents before children), so $Parents(X_i) \subseteq X_1,\ldots,X_{i-1}$
- So the Bayes net asserts conditional independences $P(X_i \mid X_1,\ldots,X_{i-1}) = P(X_i \mid Parents(X_i))$
  - To ensure these are valid, choose parents for node X_i that “shield” it from other predecessors

Conditional independence semantics

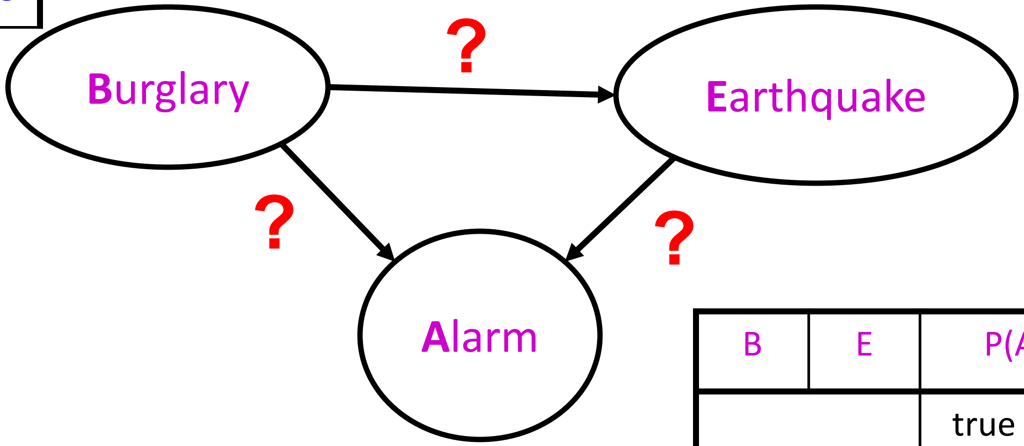
- ***Every variable is conditionally independent of its non-descendants given its parents***
- Conditional independence semantics <=> global semantics



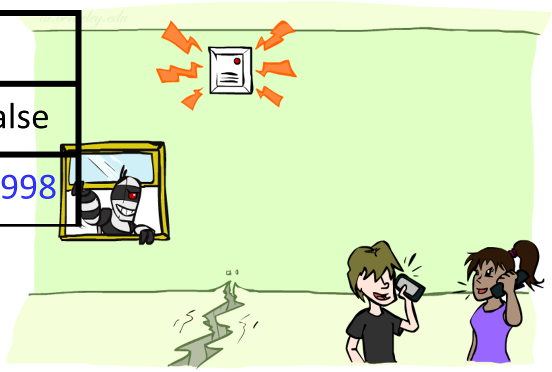
Example: Burglary

- Burglary
- Earthquake
- Alarm

P(B)	
true	false
0.001	0.999



P(E)	
true	false
0.002	0.998

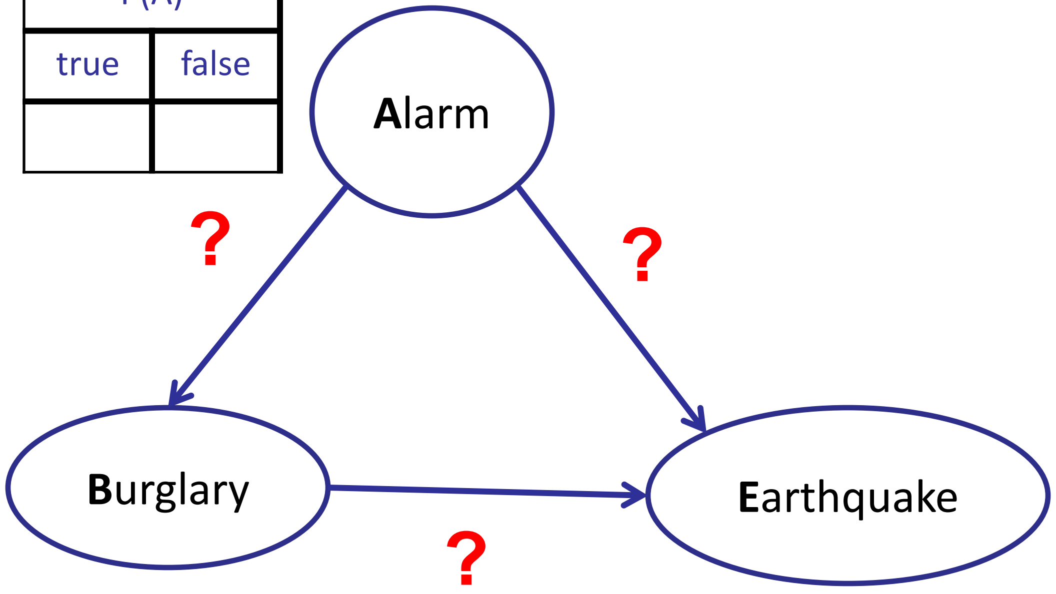


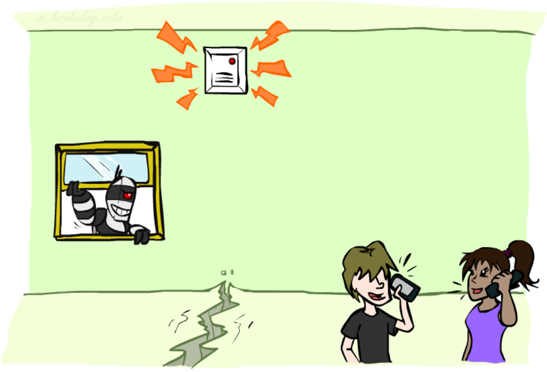
B	E	P(A|B,E)	
		true	false
true	true	0.95	0.05
true	false	0.94	0.06
false	true	0.29	0.71
false	false	0.001	0.999

Example: Burglary

- Alarm
- Burglary
- Earthquake

P(A)	
true	false



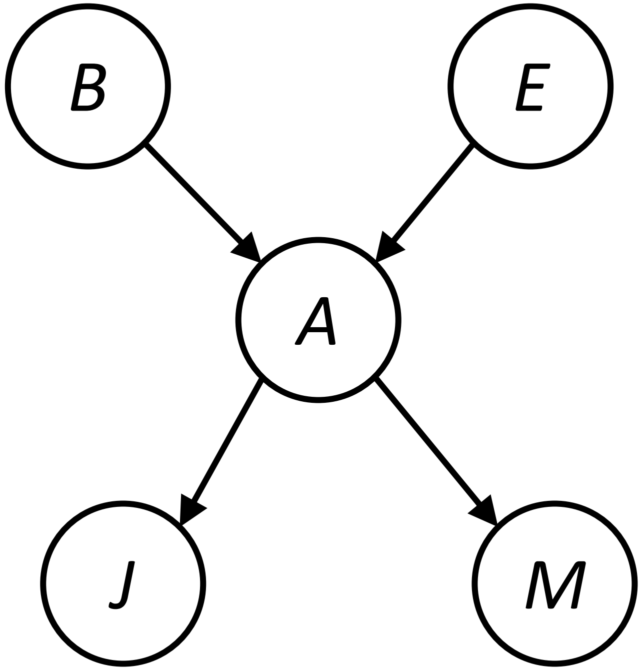


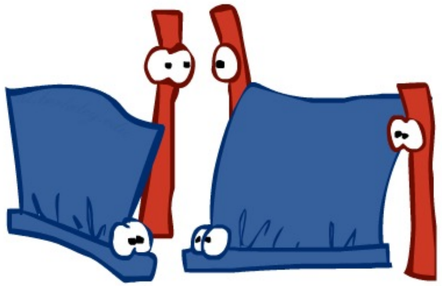
A	P(B|A)	
	true	false
true	?	
false		

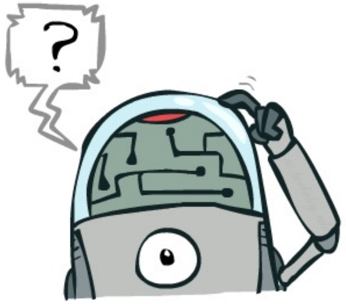
A	B	P(E|A,B)	
		true	false
true	true	?	
true	false		
false	true		
false	false		

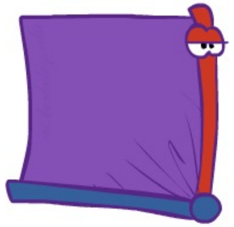
Inference by Enumeration in Bayes Net

- Reminder of inference by enumeration:
  - Any probability of interest can be computed by summing entries from the joint distribution: $P(\mathbf{Q} \mid \mathbf{e}) = \alpha \sum_{h} P(\mathbf{Q}, \mathbf{h}, \mathbf{e})$
  - Entries from the joint distribution can be obtained from a BN by multiplying the corresponding conditional probabilities
- $P(B \mid j, m) = \alpha \sum_{e,a} P(B, e, a, j, m)$
- $= \alpha \sum_{e,a} P(B)\, P(e)\, P(a \mid B,e)\, P(j \mid a)\, P(m \mid a)$
- So inference in Bayes nets means computing sums of products of numbers: sounds easy!!
- Problem: sums of ***exponentially many*** products!









Can we do better?

- Consider **uwy + uwz + uxy + uxz + vwy + vwz + vxy +vxz**
  - 16 multiplies, 7 adds
  - Lots of repeated subexpressions!
- Rewrite as **(u+v)(w+x)(y+z)**
  - 2 multiplies, 3 adds
- $\sum_{e,a} P(B)\, P(e)\, P(a|B,e)\, P(j|a)\, P(m|a)$
- $= P(B)P(e)P(a|B,e)P(j|a)P(m|a) + P(B)P(\neg e)P(a|B,\neg e)P(j|a)P(m|a)$

  $+ P(B)P(e)P(\neg a|B,e)P(j|\neg a)P(m|\neg a) + P(B)P(\neg e)P(\neg a|B,\neg e)P(j|\neg a)P(m|\neg a)$

  Lots of repeated subexpressions!

Summary

- Independence and conditional independence are important forms of probabilistic knowledge
- Bayes net encode joint distributions efficiently by taking advantage of conditional independence
  - Global joint probability = product of local conditionals
- Exact inference = sums of products of conditional probabilities from the network