CSE 575
Statistical Machine Learning

Lecture 15

YooJung Choi

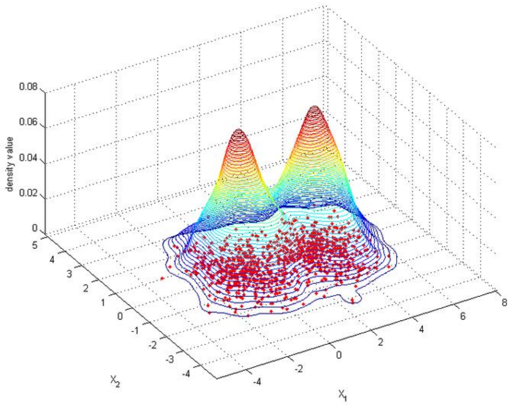
Fall 2022

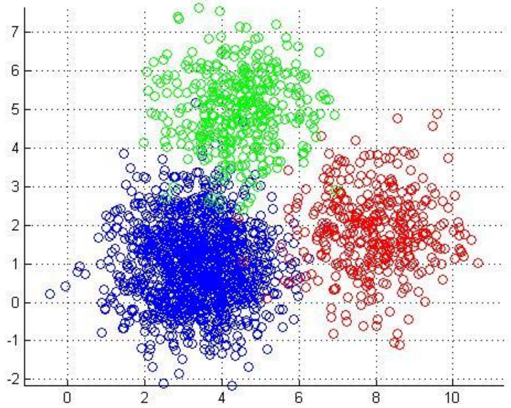
- Homework 2 due tonight (11:59pm, Oct 24)
- Homework 3 due Friday, Nov 4
- Midterm 2 in class on Wednesday, Nov 9
- Final project presentations starting Monday, Nov 14
- Final project report due Dec 7
- Do not wait until the last minute to work on your project!

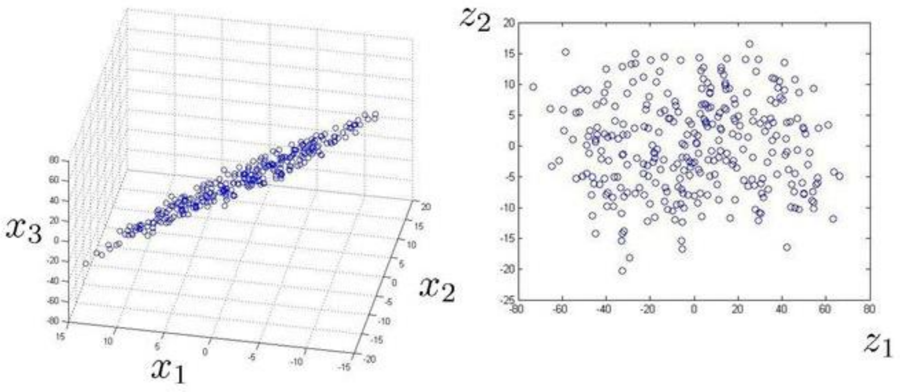
Unsupervised learning

Unlabeled data – goal is to find underlying properties or patterns in data

- Density (and distribution) estimation
- Clustering
- Dimensionality reduction

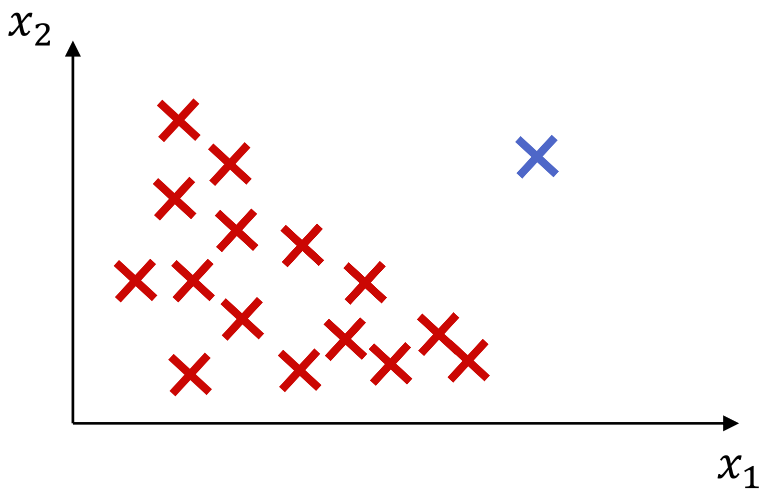






Density estimation

- Model the distribution that generated this data
  - I.e. for each input x, what is p(x)?
- Anomaly detection: is p(x) very small?
- Generating samples
- Probabilistic reasoning



Joint distribution table

- Consider a distribution of D discrete variables, each having m values

X_1	X_2	$\cdots$	X_D	$P(X_1, X_2, \dots X_D)$
1	1		1	
1	1		2	
$\vdots$	$\vdots$		$\vdots$	
1	1		m	
$\vdots$	$\vdots$			
m	m		1	
$\vdots$	$\vdots$		$\vdots$	
m	m	$\cdots$	m	

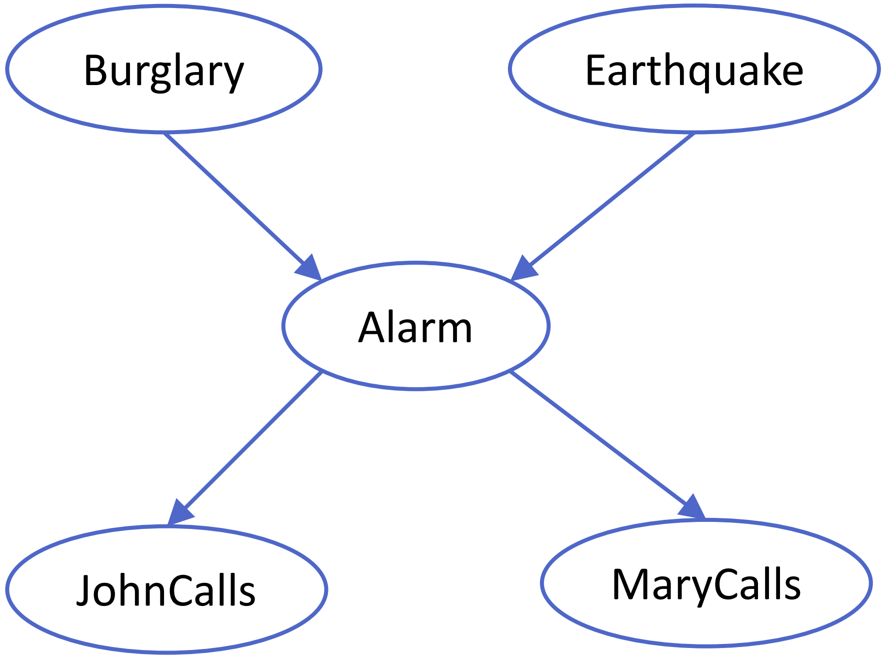
Joint probability table has m^D row

$m^D - 1$ independent parameters to specify the distribution!

Probabilistic graphical models

- Using a graphical structure to more concisely represent distributions
- Graph encodes certain properties of the generative model
- Bayesian networks: directed PGM
- Naïve Bayes: special case of Bayesian networks

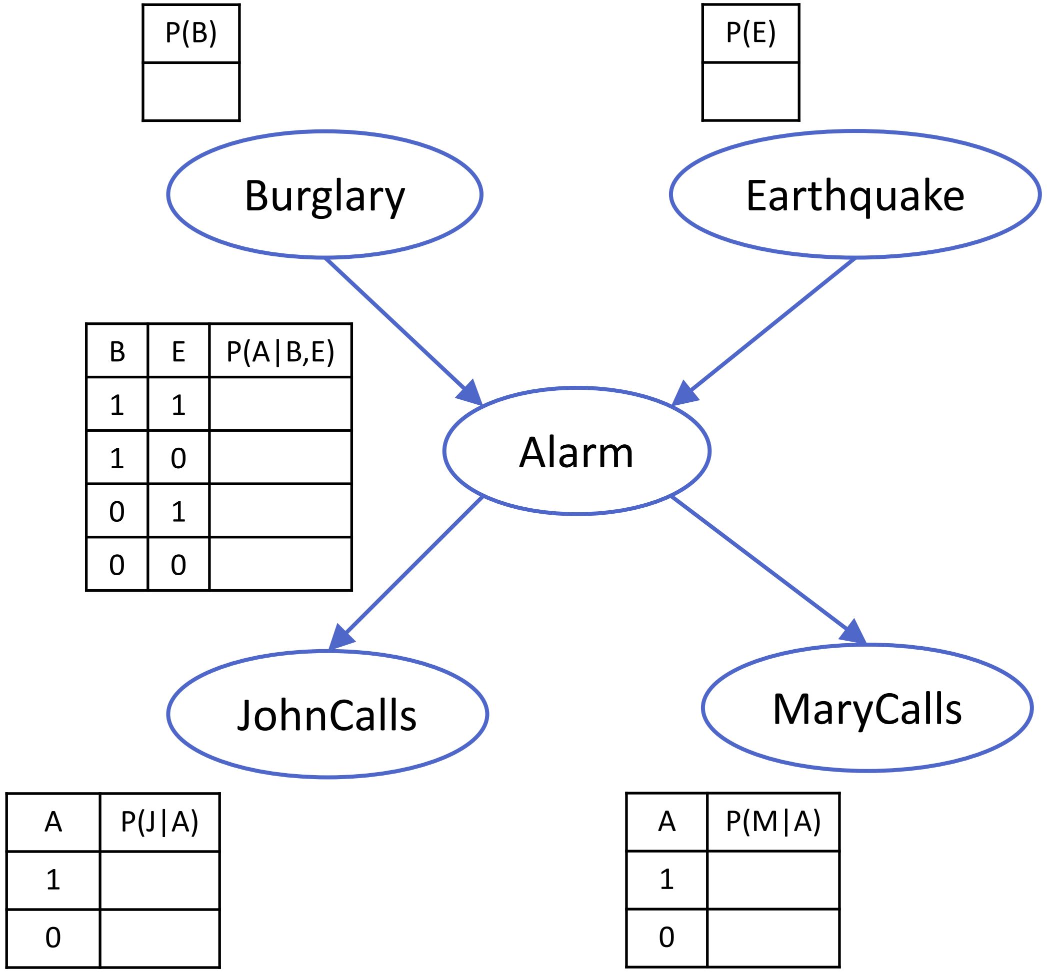
Bayesian network structure



Intuition: an arrow X -> Y means that X has a *direct influence* on Y

- Whether the alarm goes off depends directly on burglary and earthquake
- Whether John and Mary call depends only on the alarm

Bayesian network



P(B)

P(E)

B	E	P(A|B,E)
1	1	
1	0	
0	1	
0	0	

A	P(J|A)
1	
0	

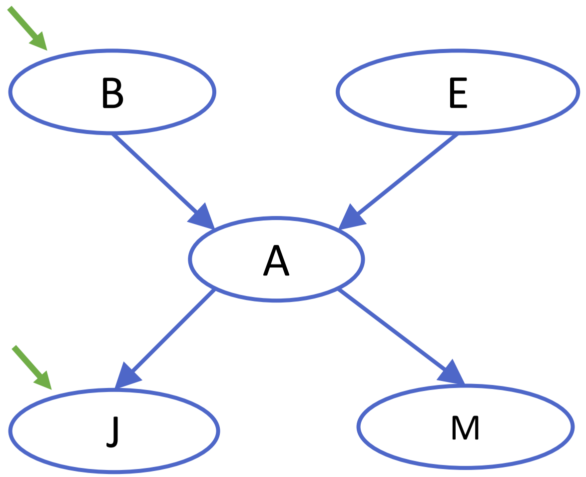
A	P(M|A)
1	
0	

A *Bayesian network* consists of:

1. A directed acyclic graph (DAG) whose nodes correspond to random variables
2. A conditional probability distribution associated with each node

Independence in the graph

- Parents(V) = {N: there is an edge N -> V}
- Descendants(V) = {N: there is a directed path from V to N}
- Non-descendants(V): variables other than V, Parents(V), Descendants(V)
- *Markovian assumption*: Every variable V is conditionally independent of Non-descendants(V) given Parents(V)



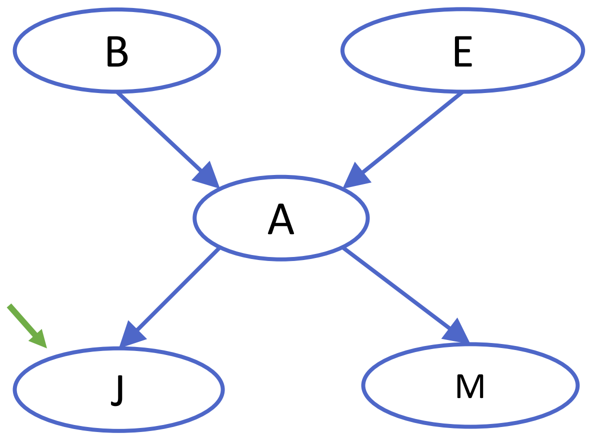
Parents(B) = {}   Non-descendants(B) = {E}

Burglary is independent of Earthquake: $B \perp E$

Parents(J) = {A}   Non-descendants(J) = {B, E, M} $J \perp B,E,M \mid A$

Independence in the graph

- Markov blanket: Parents, children, and children's parents
- A node is independent of all other nodes in the graph, given its Markov blanket

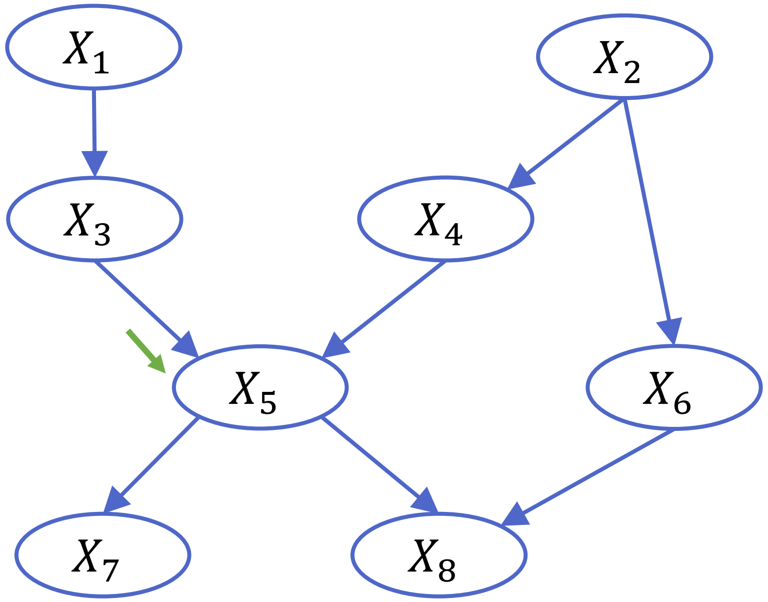


Markov-blanket(J) = {A}

$J \perp B,E,M \mid A$

Independence in the graph

- Markov blanket: Parents, children, and children's parents
- A node is independent of all other nodes in the graph, given its Markov blanket



Parents(X_5) = $\{X_3, X_4\}$

Non-descendents(X_5) = $\{X_1, X_2, X_6\}$

$X_5 \perp X_1, X_2, X_6 \mid X_3, X_4$

Markov-blanket(X_5) = $\{X_3, X_4, X_6, X_7, X_8\}$

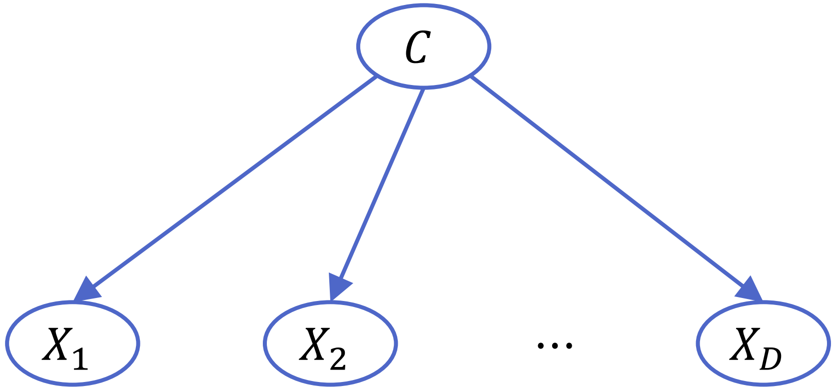
$X_5 \perp X_1, X_2 \mid X_3, X_4, X_6, X_7, X_8$

Omitted: algorithm to derive and check many more conditional independencies implied by the graph (e.g. d-separation)

Recall: naïve Bayes

- Naïve Bayes assumption: *features are independent given class*

$$X_i \perp X_j \mid C \quad \forall\, i \neq j$$



C	$P(X_1 \mid C)$
1	
0	

C	$P(X_2 \mid C)$
1	
0	

C	$P(X_D \mid C)$
1	
0	

Parents(X_1) = {C}

Non-descendants(X_1) = {X_2, …, X_D}

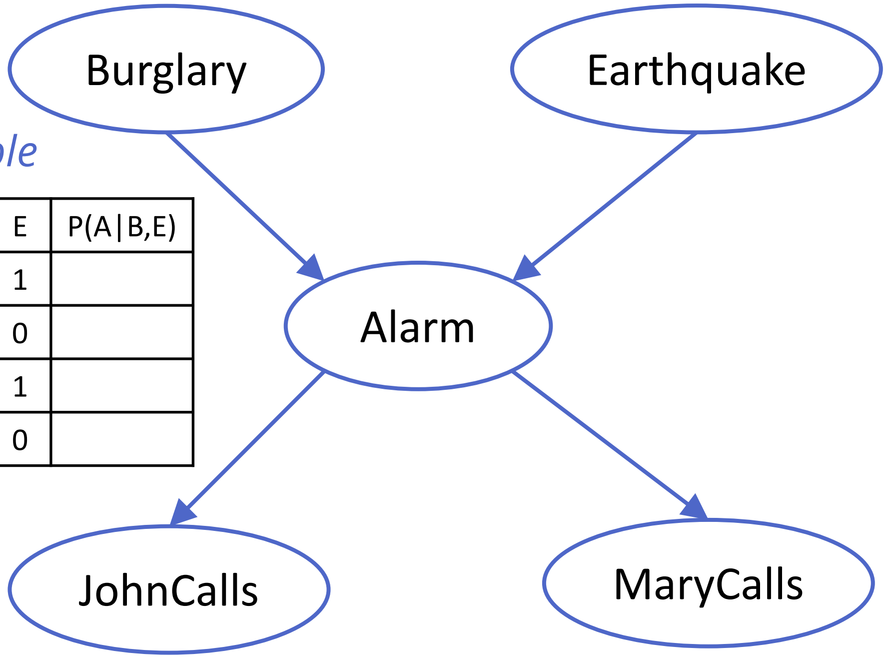
Markovian assumption given by this DAG:

$$X_1 \perp X_2, \ldots, X_D \mid C$$

Bayesian network parameters

P(B)

P(E)



Conditional probability table (CPT)

B	E	P(A|B,E)
1	1	
1	0	
0	1	
0	0	

A	P(J|A)
1	
0	

A	P(M|A)
1	
0	

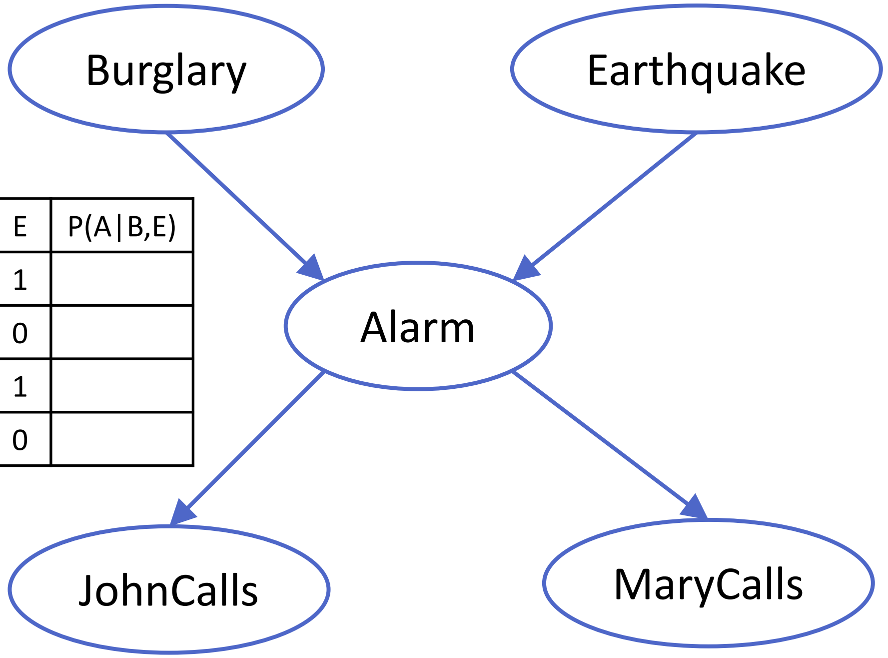
Each node V associated with a conditional probability distribution P(V|Parents(V))

B	E	A	P(A|B,E)
1	1	1	θ_1
1	1	0	$1-\theta_1$
1	0	1	θ_3
1	0	0	$1-\theta_3$
0	1	1	θ_5
0	1	0	$1-\theta_5$
0	0	1	θ_7
0	0	0	$1-\theta_7$

How many parameters do we need to fully specify this table?

4 independent parameters

Bayesian network parameters



P(B)

P(E)

B	E	P(A|B,E)
1	1	
1	0	
0	1	
0	0	

A	P(J|A)
1	
0	

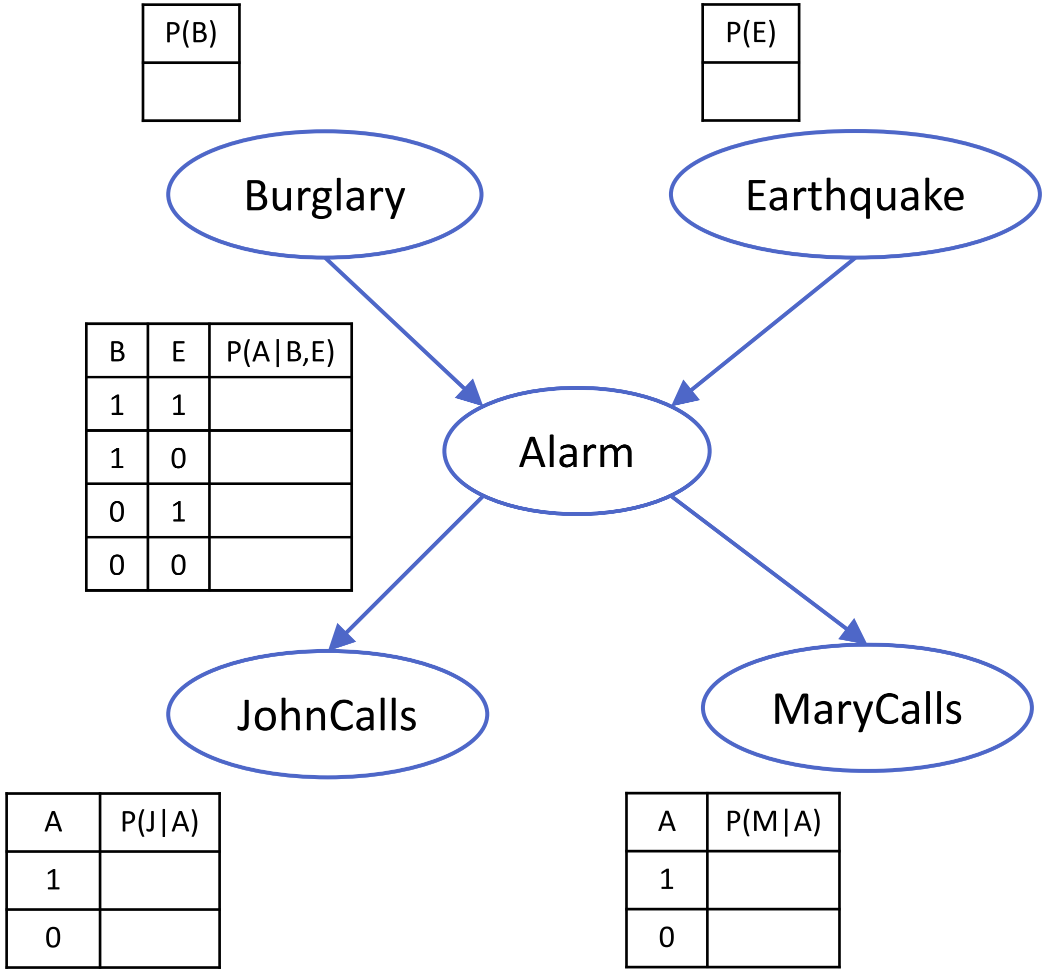
A	P(M|A)
1	
0	

Each node V associated with a conditional probability distribution P(V|Parents(V))

- If every variable has m values and at most k parents: each CPT size is bounded by $O\left(m^{k+1}\right)$
- If there are d variables: total number of BN parameters bounded by $O\left(d \cdot m^{k+1}\right)$

v.s. $O\left(m^{d}\right)$ for a full joint probability table

Representing the joint distribution



P(B)

P(E)

B	E	P(A|B,E)
1	1	
1	0	
0	1	
0	0	

A	P(J|A)
1	
0	

A	P(M|A)
1	
0	

Product rule: $P(X, Y) = P(X|Y)P(Y)$

Chain rule of probability:

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

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

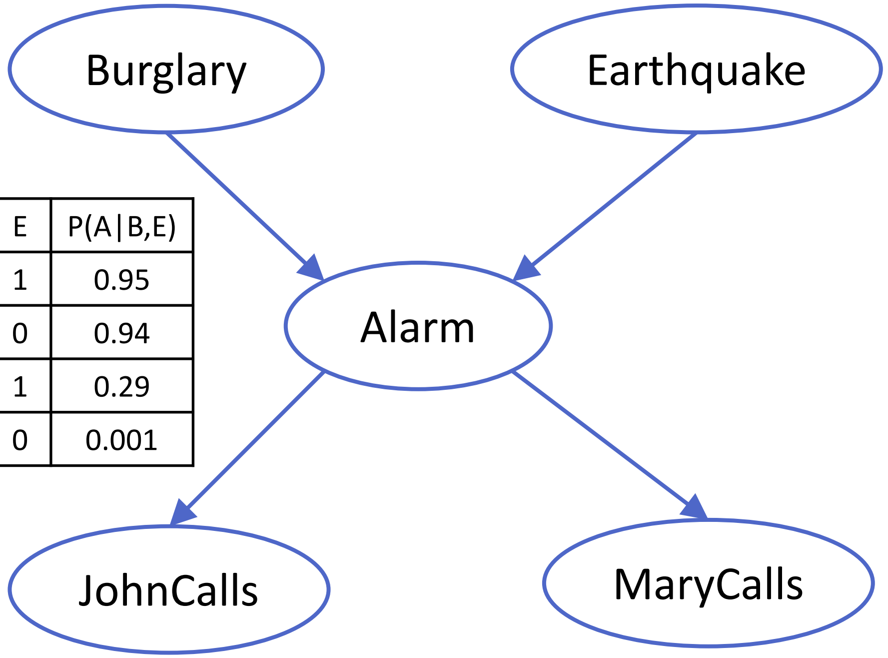
$$= P(J|M, A, B, E)P(M|A, B, E)P(A|B, E)P(B, E)$$

$$= P(J|\cancel{M}, A, \cancel{B}, \cancel{E})P(M|A, \cancel{B}, \cancel{E})P(A|B, E)P(B|\cancel{E})P(E)$$

Using the Markovian assumptions

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

Representing the joint distribution



P(B)
0.01

P(E)
0.02

B	E	P(A|B,E)
1	1	0.95
1	0	0.94
0	1	0.29
0	0	0.001

A	P(J|A)
1	0.90
0	0.10

A	P(M|A)
1	0.70
0	0.01

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

$$P(J = 1, M = 0, A = 1, B = 0, E = 1)$$

$$= P(J = 1|A = 1) \times P(M = 0|A = 1)$$

$$\times P(A = 1|B = 0, E = 1) \times P(B = 0) \times P(E = 1)$$

$$= \ 0.90 \times (1 - 0.70) \times 0.29 \times (1 - 0.01) \times 0.02$$

In general, $P(x_1, \ldots, x_D) = \prod_{i=1}^{D} \underline{P(x_i|\text{Parents}(X_i))}$

Given by the conditional probability tables

Representing the joint distribution

Quick validity check: Joint probabilities for all possible combinations of values of $X_1, \dots, X_D$ should sum to 1

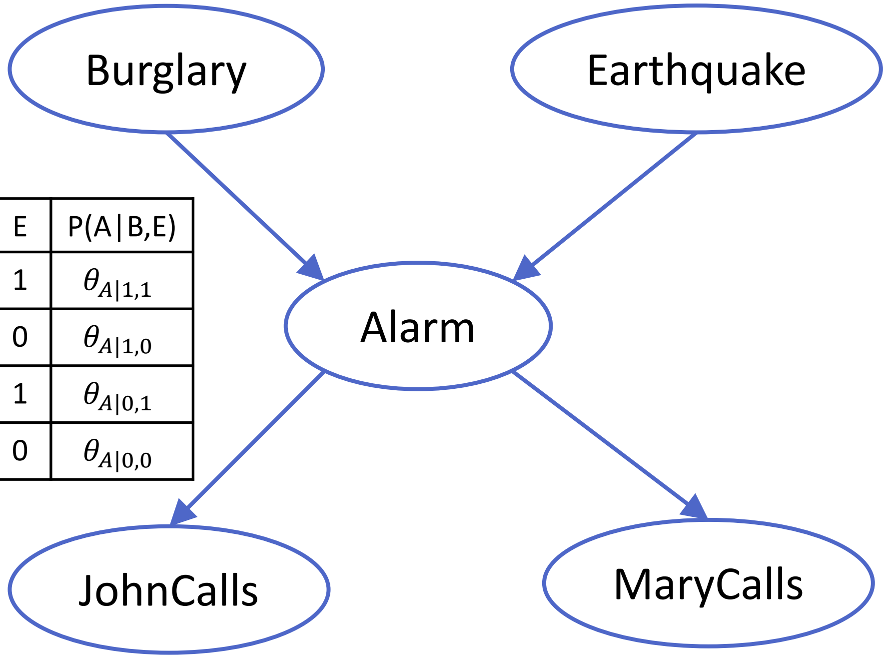
$$\sum_{x_1,\dots,x_D} P(x_1, \dots, x_D) = \sum_{x_1,\dots,x_D} \prod_{i=1}^{D} P(x_i | \text{Parents}(X_i))$$

$$= \prod_{i=1}^{D} \sum_{x_i} P(x_i | \text{Parents}(X_i)) = \prod_{i=1}^{D} 1 = 1$$

Parameter learning

P(B)
θ_B

P(E)
θ_E



B	E	P(A|B,E)
1	1	$\theta_{A\|1,1}$
1	0	$\theta_{A\|1,0}$
0	1	$\theta_{A\|0,1}$
0	0	$\theta_{A\|0,0}$

A	P(J|A)
1	$\theta_{J\|1}$
0	$\theta_{J\|0}$

A	P(M|A)
1	$\theta_{M\|1}$
0	$\theta_{M\|0}$

$$\theta_{v|\boldsymbol{u}} = P(v|\text{Parents}(V) = \boldsymbol{u})$$

Maximum likelihood parameters: $\theta_{v|\boldsymbol{u}}^{\text{MLE}} = \frac{\#\{v,\boldsymbol{u}\}}{\#\{\boldsymbol{u}\}}$

B	E	A	J	M
1	0	0	1	0
0	1	1	1	1
1	1	1	0	0
1	0	1	1	0
1	1	0	1	0
0	1	0	0	1
0	1	0	0	0
1	1	1	1	1
0	0	0	1	0

$$\theta_{J|1} = \frac{\#\{J=1, A=1\}}{\#\{A=1\}} = \frac{3}{4}$$

$$\theta_{A|0,1} = \frac{\#\{A=1, B=0, E=1\}}{\#\{B=0, E=1\}}$$

$$= \frac{1}{3}$$

More on learning…

Parameter learning for Bayesian networks

- Closed-form MLE solutions *from complete data*
- Next: learn from data with missing values (*incomplete data*) via *expectation-maximization*

Structure learning

- Likelihood of a structure G given data $\mathcal{D}$: $L(G|\mathcal{D}) = L(\theta^{MLE}|\mathcal{D})$

  where θ^{MLE} are the maximum-likelihood parameters for structure G and data $\mathcal{D}$
- Trivial solution for a maximum-likelihood structure? Complete (fully-connected) DAG!
- Popular approaches: heuristic-based search starting from a simple graph, adding edges