CS 161
Discussion 9
Bayesian Network

Bayesian Network

Motivations:

- Tool for modeling uncertainties
- Capture our perception of causality

Bayesian Network

For each family of models, we care about:

1. Representation 2. Inference 3. Learning

- Representation
  - Model conditional dependency (causation)
  - Represent joint probability over a set of random variables
- Inference
  - Typical Tasks
    - Conditional Probability Query P(X1, X2, ...|E1 ,E2,...)
    - Marginalize one or a set variables P(X1, X2, …), P(X1, X2), …
  - Algorithms: Variable Elimination
    - Operations:
      - Summing out a variable
      - Factor multiplication

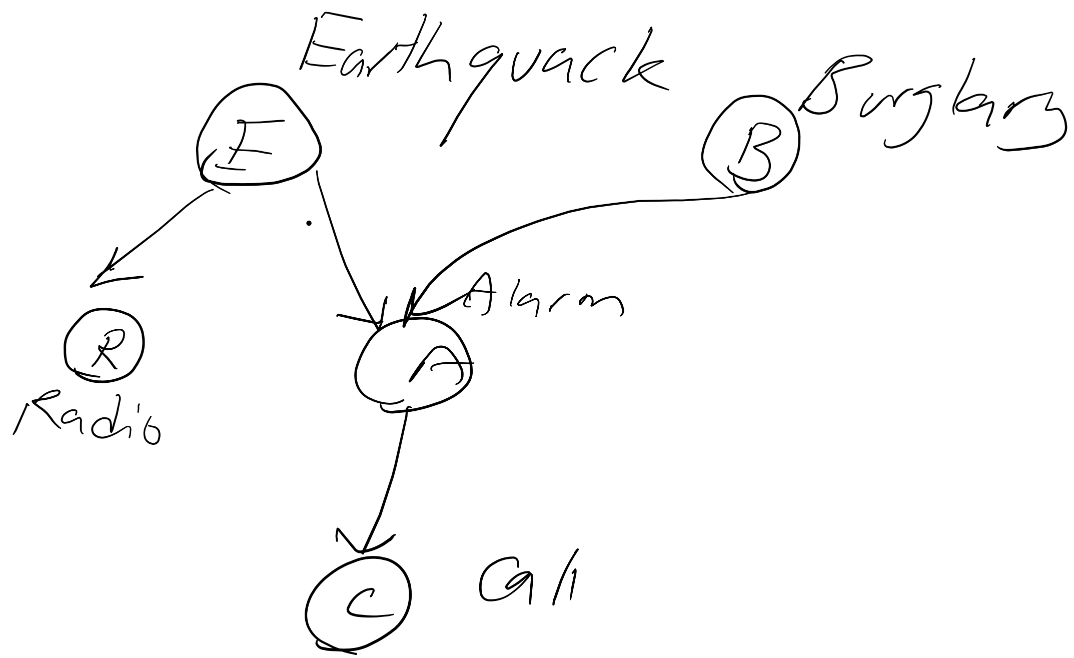
Bayesian Network

For each family of models, we care about:

1. Representation
2. Inference
3. Learning

Bayesian Network Representation

- Representation
  - Model conditional dependency (causation)
  - Represent joint probability over a set of random variables
  - For example, P(E, A, B, C, R);



Bayesian Network Representation

Goal: Represent joint probability over a set of random variables

- Facilitate probability computation

Component:

(1) Graph Structure: a Directed Acyclic Graph (DAG)

- Nodes: random variables (events)
- Edges: $y \rightarrow x$ means y causes/influences x

(2) Local Probability Model

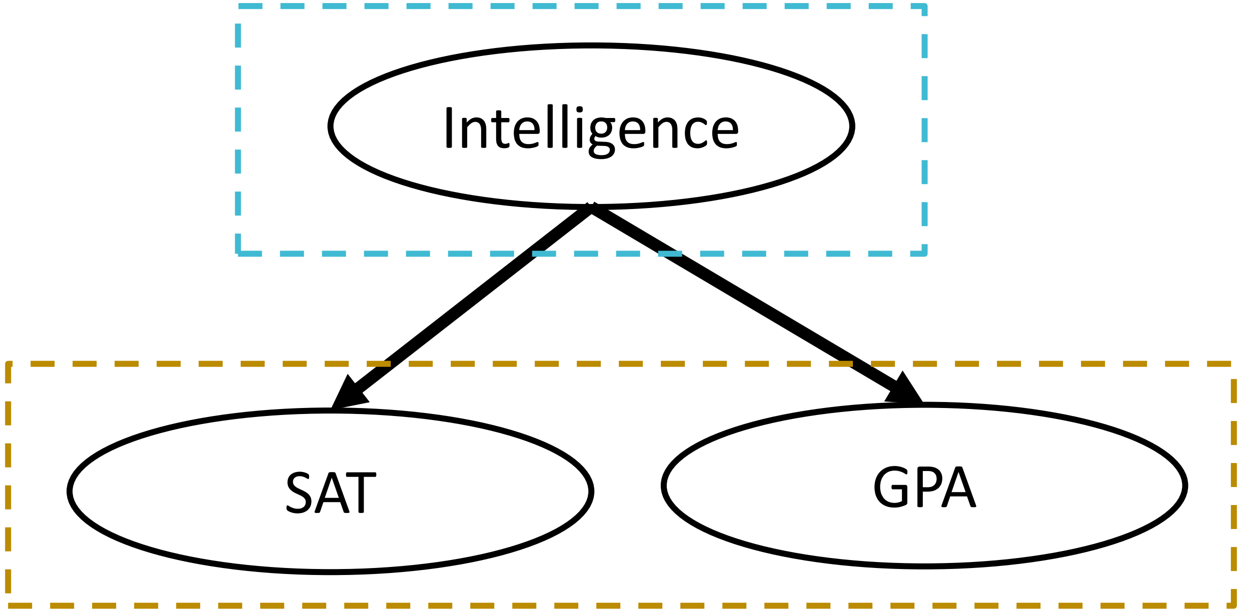
- Represent the dependence of each variable on its parents
- $y_1, y_2, \ldots, y_k \rightarrow x$: conditional probability $p(x|y_1, y_2, \ldots, y_k)$
- Root variables: marginal probability

Student Example – Graph Structure

- A company wants to hire an intelligent student.
  - But intelligence cannot be directly measured.
  - But the company may have access to the student's SAT and GPA score.
- Based on the observable evidence (SAT and GPA), company can try to infer whether this student is intelligent or not.

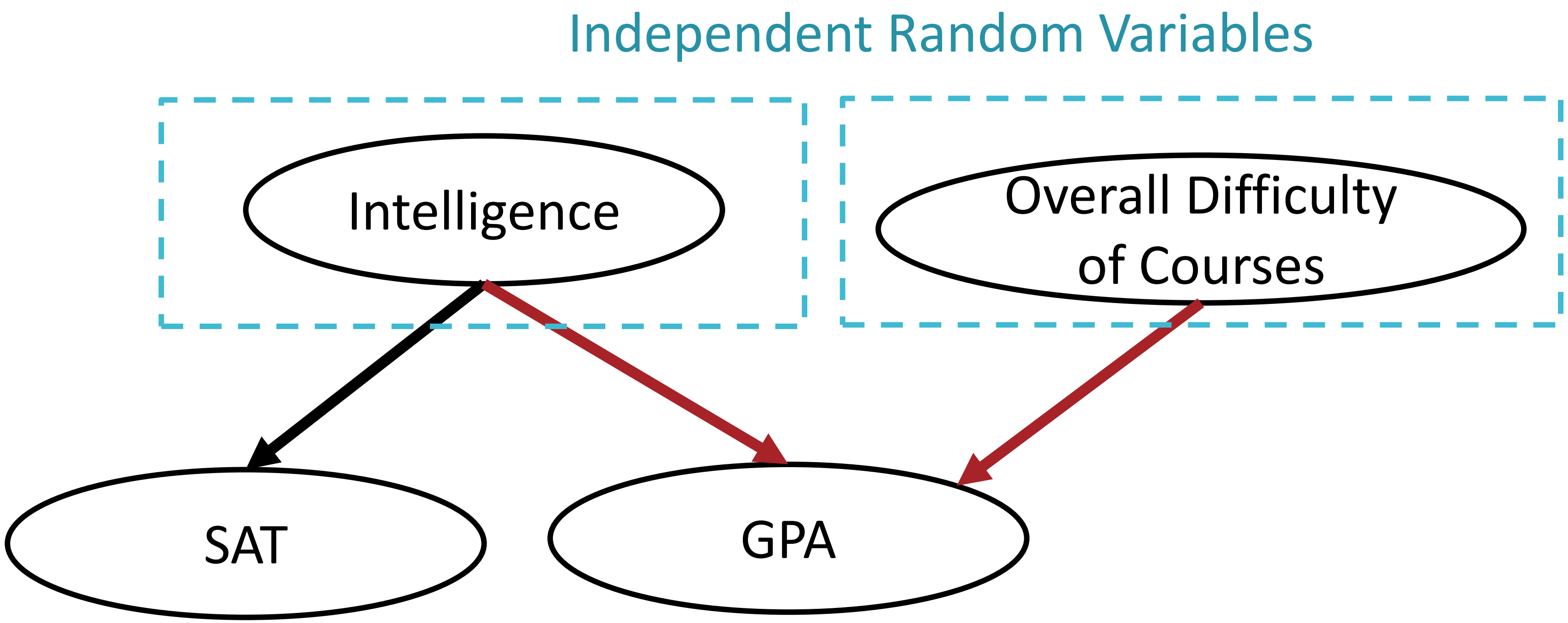
Student Example – Graph Structure

- Cannot be directly measured
- Latent factor of its children



Can be directly observed

Student Example – Graph Structure



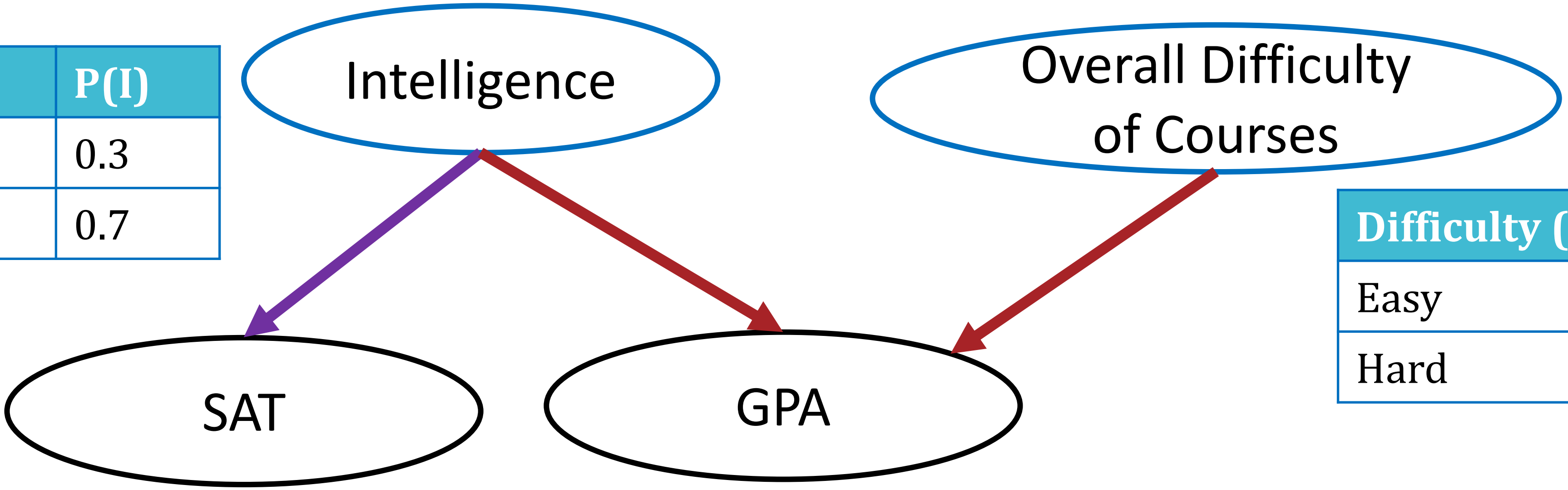
P(GPA | Intelligence, Difficulty of Courses)

Student Example – A Full Bayesian Network

Component: (1) Graph Structure (DAG) (2) Local probability model (CPDs)

Conditional Probability Distribution

Intelligence (I)	P(I)
True	0.3
False	0.7



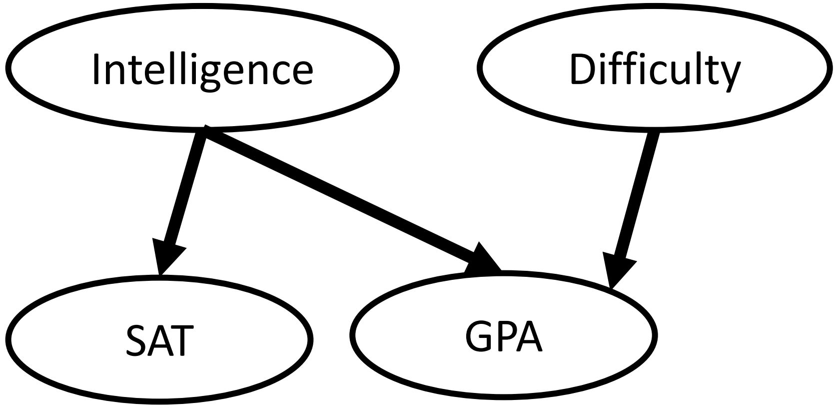
Difficulty (D)	P(D)
Easy	0.6
Hard	0.4

I	SAT	P(SAT|I)
True	High	0.7
True	Low	0.3
False	High	0.4
False	Low	0.6

I	D	GPA	P(GPA|I,D)
True	Easy	High	0.8
True	Easy	Low	0.2
True	Hard	High	0.6
…	…	…	…

Topological Semantics

- BN satisfies **local Markov property**:
  - A node is conditionally independent of its non-descendants given its parents.
- A BN encodes a set of (conditional) independence assumptions (Markovian assumptions)

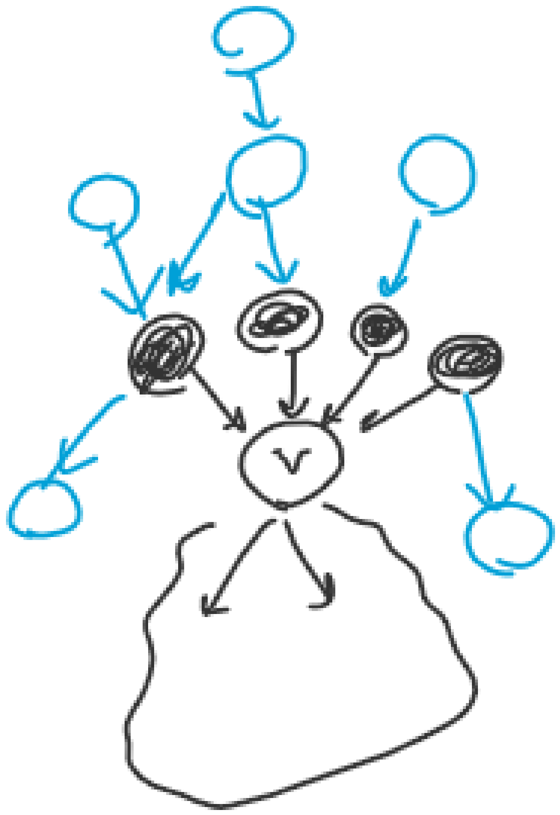


Markovian Assumptions

Intelligence $\perp$ Difficulty
SAT $\perp$ GPA | Intelligence
SAT $\perp$ Difficulty | Intelligence
GPA $\perp$ SAT | Intelligence, Difficulty

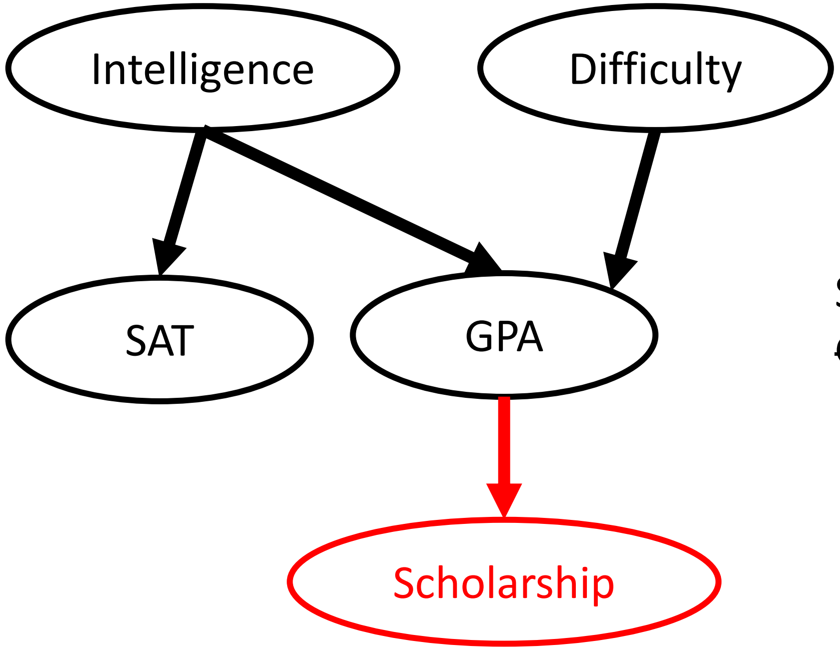
Topological Semantics

Given Parents(V), then V becomes independent of its NonDescendants



Topological Semantics

- BN satisfies **local Markov property:**
  - A node is conditionally independent of its non-descendants given its parents.

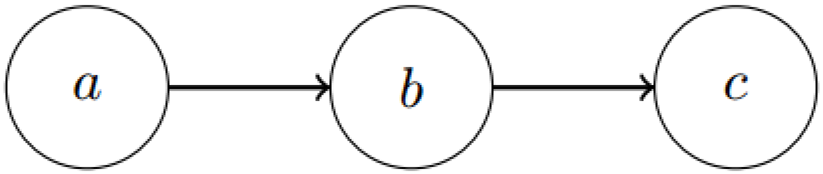


SAT ⊥ Scholarship | Intelligence

~~GPA ⊥ Scholarship | Intelligence, Difficulty~~

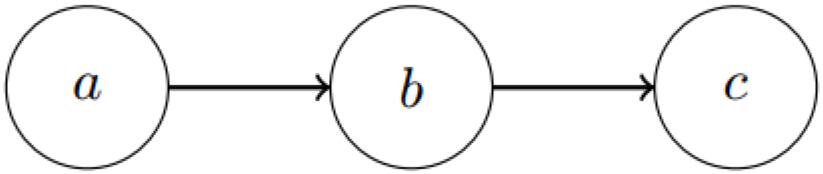
Exercise – conditional independency

Give the topological semantics encoded in the BN.



Exercise – conditional independency

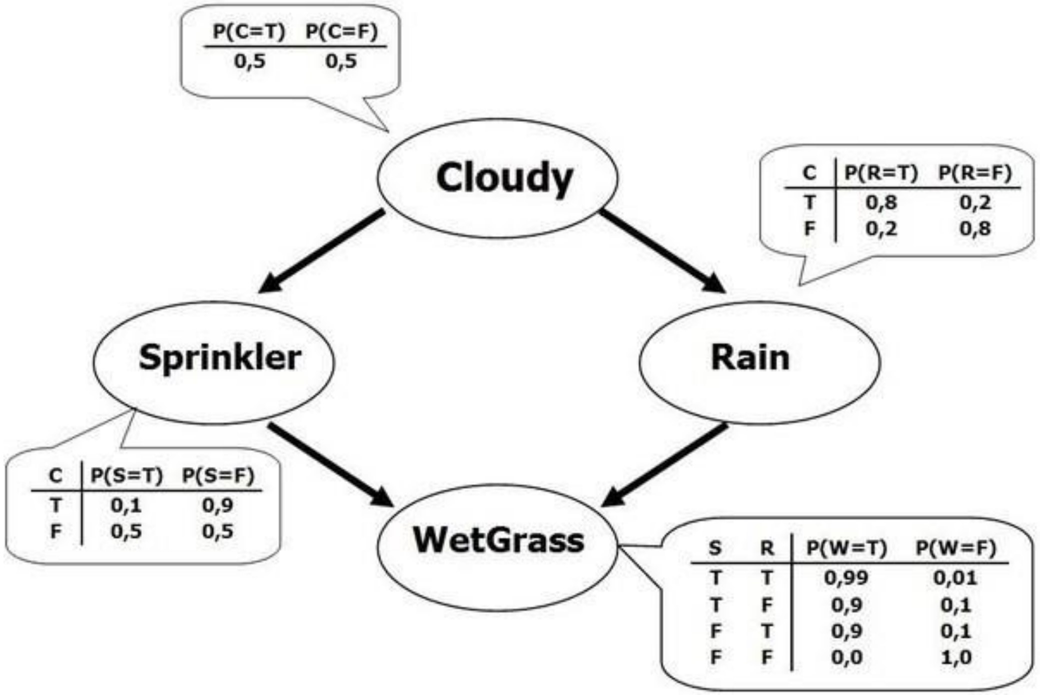
Give the topological semantics encoded in the BN.



$c \perp a \mid b$

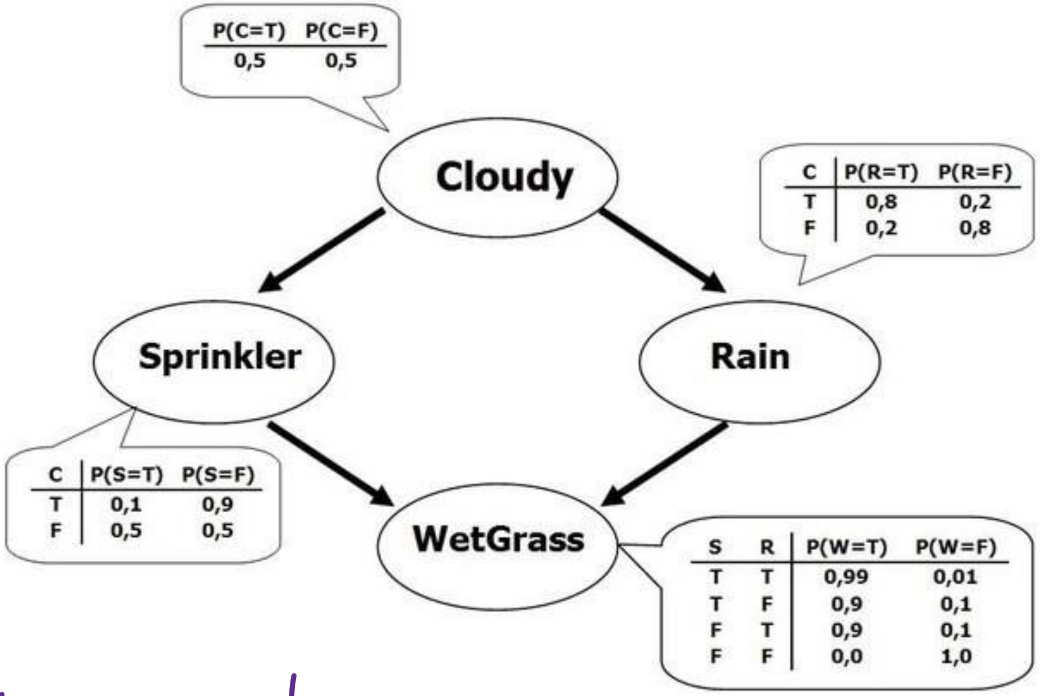
Exercise – conditional independency

- Given Cloudy, what variables is Sprinkler conditional independent of?



Exercise – conditional independency

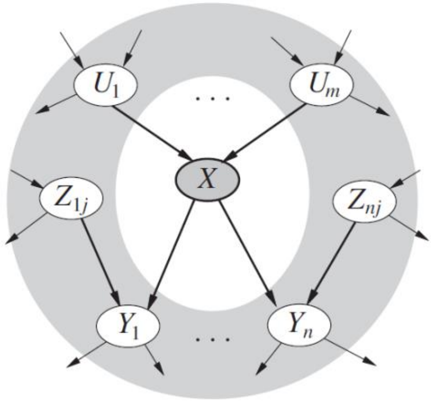
- Given Cloudy, what variables is Sprinkler conditional independent of?



sprinkler $\perp$ rain | cloudy

Markov Blanket

- Markov Blanket
  - The node's parents, children and children's parents
  - The node is conditionally independent of all other nodes given this Markov Blanket



Joint Probability – Chain Rule for BN

Bayesian network models the following **joint probability**

$$P(X_1, X_2, \dots, X_N) = \prod_{i=1}^{N} P(X_i | \text{parents of } X_i),$$

Why?

Joint Probability – Chain Rule for BN

Bayesian network models the following **joint probability**

$$P(X_1, X_2, \dots, X_N) = \prod_{i=1}^{N} P(X_i|\text{parents of } X_i)$$

Why?

- Without loss of generality, assume $X_1, X_2, \dots, X_N$ is a topological ordering
- Chain rule

$$P(X_1, X_2, \dots, X_N) = \prod_{i=1}^{N} P(X_i|X_1, X_2, \dots, X_{i-1})$$

- $P(X_i|X_1, X_2, \dots, X_n) = P(X_i|\text{parents of } X_i)$
    - Topological ordering => parents are in $X_1, X_2, \dots, X_{i-1}$
    - Local Markov property => given parents,
        - independent of other variables in $X_1, X_2, \dots, X_{i-1}$

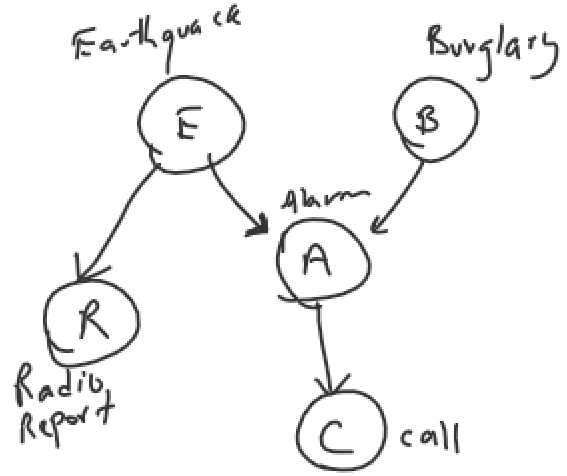
Joint Probability

$$P(X_1, X_2, \dots, X_N) = \prod_{i=1}^{N} P(X_i|\text{parents of } X_i)$$

- We call the above equation ***chain rule for Bayesian networks***.
- If P and G satisfy the above equation, we say **P factorizes according to G**
- $P(X_i|\text{parents of } X_i)$: a **factor**

Joint Probability

$$P(X_1, X_2, \ldots, X_N) = \prod_{i=1}^{N} P(X_i \mid \text{parents of } X_i)$$



Probability of
Alarm, no burglary, Call,
no earthquake, Radio?

$$p(a, \bar{b}, c, \bar{e}, r)$$

$$= p(\bar{e}) \cdot p(\bar{b}) \cdot p(a \mid \bar{b}, \bar{e})$$

$$\cdot\ p(r \mid \bar{e}) \cdot p(c \mid a)$$

Inference

Task: Probability Query

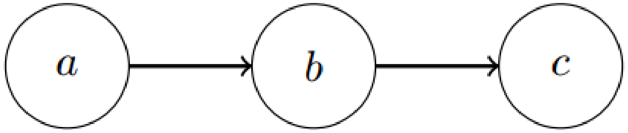
Given a Bayesian Network, we know what's the joint probability of all random variables.

Now we want to compute some other probability!

- **Conditional probability query:** Compute
$$P(Y|E=e)$$
  - E: *Evidence*
    - A subset of random variables with known (instantiated) values e
  - Y: Query variables
    - A subset of random variables (values unknown)

- **Marginalize one or a set of variable**
$$P(X_1, X_2)$$

Example – Inference (A Really Simple One)



a	$\Pr(a)$
1	1/2
0	1/2

a	b	$\Pr(b \mid a)$
1	1	1/8
1	0	7/8
0	1	1/4
0	0	3/4

b	c	$\Pr(c \mid b)$
1	1	4/5
1	0	1/5
0	1	1/4
0	0	3/4

compute Pr(a=T|b=T)

Example – Inference (A Really Simple One)

$$a \rightarrow b \rightarrow c$$

a	$\Pr(a)$
1	1/2
0	1/2

a	b	$\Pr(b \mid a)$
1	1	1/8
1	0	7/8
0	1	1/4
0	0	3/4

b	c	$\Pr(c \mid b)$
1	1	4/5
1	0	1/5
0	1	1/4
0	0	3/4

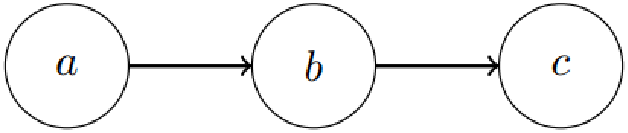
compute Pr(a=T|b=T)

$$= \frac{Pr(a=1,\ b=1)}{Pr(b=1)}$$

$$= \frac{Pr(b=1 \mid a=1)\, Pr(a=1)}{Pr(b=1,\ a=1) + Pr(b=1,\ a=0)}$$

$$= \frac{Pr(b=1 \mid a=1)\, Pr(a=1)}{Pr(b=1 \mid a=1) P(a=1) + Pr(b=1 \mid a=0)\, Pr(a=0)}$$

Example – Inference (A Really Simple One)



a	$\Pr(a)$
1	1/2
0	1/2

a	b	$\Pr(b \mid a)$
1	1	1/8
1	0	7/8
0	1	1/4
0	0	3/4

b	c	$\Pr(c \mid b)$
1	1	4/5
1	0	1/5
0	1	1/4
0	0	3/4

compute Pr(a=T|b=T)

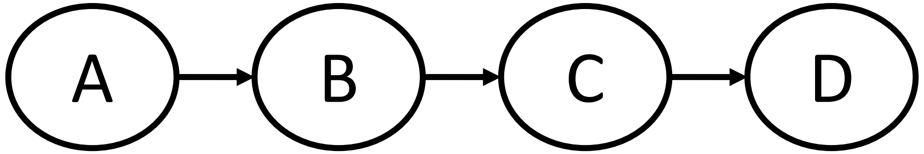
$$\begin{aligned}
\Pr(a = \texttt{true} \mid b = \texttt{true}) &= \frac{Pr(a = \texttt{true}, b = \texttt{true})}{Pr(b = \texttt{true})} \\
&= \frac{\frac{1}{2} \cdot \frac{1}{8}}{\frac{1}{2} \cdot \frac{1}{8} + \frac{1}{2} \cdot \frac{1}{4}} \\
&= \frac{\frac{1}{16}}{\frac{1}{16} + \frac{1}{8}} \\
&= \frac{1}{3}
\end{aligned}$$

Variable Elimination

- Dynamic Programming
- **Sum out one variable at a time**
- Basic computation step: manipulation of factors
- Cache intermediate results to improve efficiency

Let's start from a simple example and move to complex ones.

Example - Try to Compute Some Probability



- **Goal: Compute $P(D)$ ----** *Seems very easy!*

Example - Try to Compute Some Probability

$$A \rightarrow B \rightarrow C \rightarrow D$$

$$P(D) = \sum_C \sum_B \sum_A P(A, B, C, D)$$

$$P(D) = \overbrace{P(D|c)P(c)}^{\text{C=True}} + \overbrace{P(D|\bar{c})P(\bar{c})}^{\text{C=False}}$$

written as

$$= \sum_C P(D|C)P(C)$$

$$= \sum_C P(D|C) \sum_B P(C|B)P(B)$$

similarly

$$= \sum_C P(D|C) \sum_B P(C|B) \sum_A P(B|A)P(A)$$

$$= \sum_C \sum_B \sum_A P(A, B, C, D)$$

Example - Try to Compute Some Probability

$$A \rightarrow B \rightarrow C \rightarrow D$$

$$P(D) = \sum_{C} \sum_{B} \sum_{A} P(A, B, C, D)$$

$$\begin{aligned}
P(D) &= P(D|c)P(c) + P(D|\bar{c})P(\bar{c}) \\
&= \sum_{C} P(D|C)P(C) \\
&= \sum_{C} P(D|C) \sum_{B} P(C|B)P(B) \\
&= \sum_{C} P(D|C) \sum_{B} P(C|B) \sum_{A} P(B|A)P(A) \\
&= \sum_{C} \sum_{B} \sum_{A} P(A, B, C, D)
\end{aligned}$$

Example - Try to Compute Some Probability

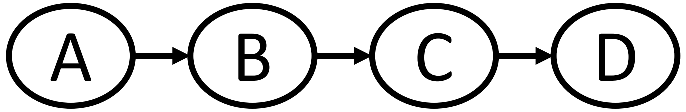
$$A \rightarrow B \rightarrow C \rightarrow D$$

- Goal: Compute $P(D)$

$$P(D) = \sum_C \sum_B \sum_A P(A, B, C, D)$$

Sum out extra variables

Example - Try to Compute Some Probability



- What if we want to compute $P(C)$? Does this equation hold?

$$P(C) = \sum_{D} \sum_{B} \sum_{A} P(A, B, C, D)$$

Variable Elimination

$$A \to B \to C \to D$$

- It's not efficient to P(A,B,C,D) for all possibilities of (A,B,C,D) ! (Why?)
- In practice, we first write out $\sum_C \sum_B \sum_A P(A,B,C,D)$ and then **push in the summations** as follows

$$P(D) = \sum_C P(D|C) \sum_B P(C|B) \sum_A P(B|A)P(A)$$

How to efficiently compute it???

Summing Out a Variable (Factor Marginalization)

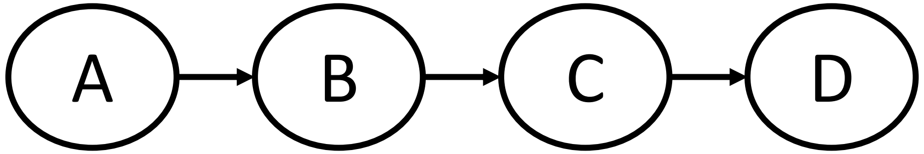
- $\boldsymbol{X}$: a set of variables
- Y: one variable. $Y \notin X$
- $\phi(\boldsymbol{X}, Y)$: a factor
  - $\phi: Val(\boldsymbol{X}) \longmapsto \mathbb{R}$
  - $\text{Scope}(\phi) = \{\boldsymbol{X}, Y\}$

- **Sum out** of Y in ψ (marginalize Y in ϕ):

$$\psi(\boldsymbol{X}) = \sum_{Y} \phi(\boldsymbol{X}, Y)$$

The result is a new factor without Y.

Factors

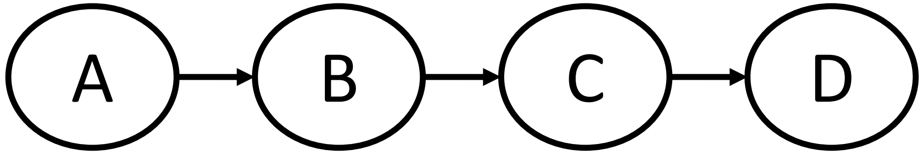


$$P(D) = \sum_C P(D|C) \sum_B P(C|B) \sum_A \underbrace{P(B|A)}_{\phi_2(A,B)} \underbrace{P(A)}_{\phi_1(A)}$$

A	$\phi_1(A)$
True	0.4
False	0.6

A	B	$\phi_2(A,B)$
True	True	0.3
True	False	0.7
False	True	0.5
False	False	0.5

Factor Multiplication



A	$\phi_1(A)$
True	0.4
False	0.6

$$P(D) = \sum_C P(D|C) \sum_B P(C|B) \sum_A \underbrace{P(B|A)}_{\phi_2(A,B)} \underbrace{P(A)}_{\phi_1(A)}$$

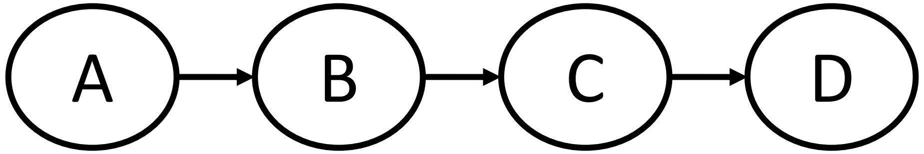
A	B	$\phi_2(A,B)$
True	True	0.3
True	False	0.7
False	True	0.2
False	False	0.8

Factor Multiplication

$\Rightarrow$ intermediate result $\varphi_1(A,B)$

A	B	$\varphi_1(A,B)$
True	True	0.4*0.3=0.12
True	False	0.4*0.7=0.28
False	True	0.6*0.2=0.12
False	False	0.6*0.8=0.48

Summing Out a Variable



$$P(D) = \sum_{C} P(D|C) \sum_{B} P(C|B) \boxed{\sum_{A} \underbrace{P(B|A)P(A)}_{\varphi_1(A,B)}}$$

A	B	$\phi_2(A,B)$
True	True	0.3
True	False	0.7
False	True	0.2
False	False	0.8

$\Rightarrow \psi_1(B)$

A	$\phi_1(A)$
True	0.4
False	0.6

Summing Out a Variable

$A \rightarrow B \rightarrow C \rightarrow D$

Intermediate Result $\varphi_1(A, B)$

A	B	$\varphi_1(A, B)$
True	True	0.4*0.3=0.12
True	False	0.4*0.7=0.28
False	True	0.6*0.2=0.12
False	False	0.6*0.8=0.48

Summing out A

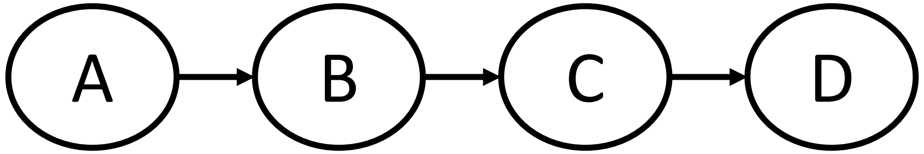
$$P(D) = \sum_C P(D|C) \sum_B P(C|B) \sum_A \underbrace{P(B|A)P(A)}_{\varphi_1(A,B)}$$

$\Rightarrow$ New factor $\psi_1(B)$

new factor without *A*

B	$\psi_1(B)$
True	0.12+0.12=0.24
False	0.28+0.48=0.76

Variable Elimination



$$P(D) = \sum_{C} \underset{\phi_4(C,D)}{P(D|C)} \sum_{B} \underset{\phi_3(B,C)}{P(C|B)} \sum_{A} \underset{\phi_2(A,B)}{P(B|A)} \underset{\phi_1(A)}{P(A)}$$

$\Rightarrow \psi_1(B)$

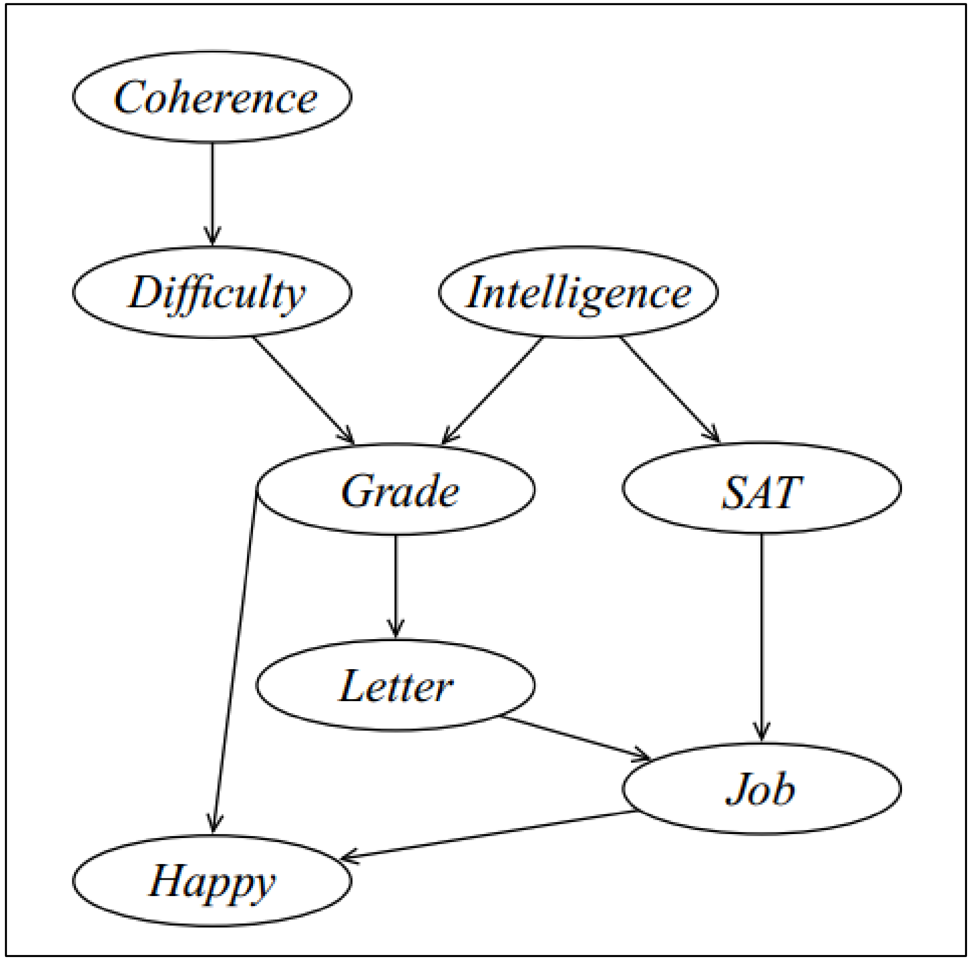
$\Rightarrow \psi_2\ (C)$

$\Rightarrow \psi_3\ (D)$

Result!

Example – Marginalize a Variable

Given a Bayesian network as follows:



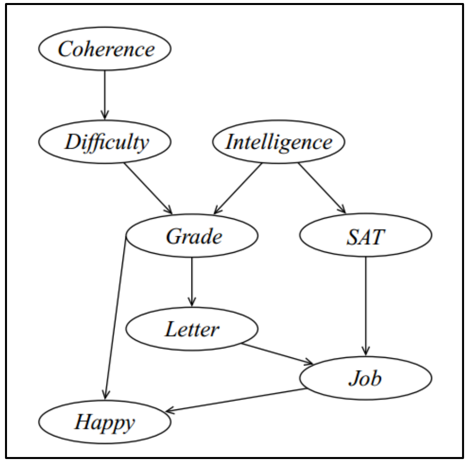
Random Variables:
C,D,I,G,S,L,J,H

Objective:
Compute $P(J)$

Let us demonstrate the procedure on a nontrivial example. Consider the network demonstrated in figure 9.8, which is an extension of our Student *network. The chain rule for this network asserts that*

$$\begin{aligned} P(C,D,I,G,S,L,J,H) &= P(C)P(D \mid C)P(I)P(G \mid I,D)P(S \mid I) \\ &\quad P(L \mid G)P(J \mid L,S)P(H \mid G,J) \\ &= \phi_C(C)\phi_D(D,C)\phi_I(I)\phi_G(G,I,D)\phi_S(S,I) \\ &\quad \phi_L(L,G)\phi_J(J,L,S)\phi_H(H,G,J). \end{aligned}$$

We will now apply the VE algorithm to compute $P(J)$. We will use the elimination ordering: C, D, I, H, G, S, L:



1. *Eliminating C: We compute the factors*

$$\begin{aligned} \psi_1(C,D) &= \phi_C(C) \cdot \phi_D(D,C) \\ \tau_1(D) &= \sum_C \psi_1. \end{aligned}$$

2. *Eliminating D: Note that we have already eliminated one of the original factors that involve D — $\phi_D(D,C) = P(D \mid C)$. On the other hand, we introduced the factor $\tau_1(D)$ that involves*

D. Hence, we now compute:

$$\begin{aligned}\psi_2(G,I,D) &= \phi_G(G,I,D)\cdot\tau_1(D)\\ \tau_2(G,I) &= \sum_D \psi_2(G,I,D).\end{aligned}$$

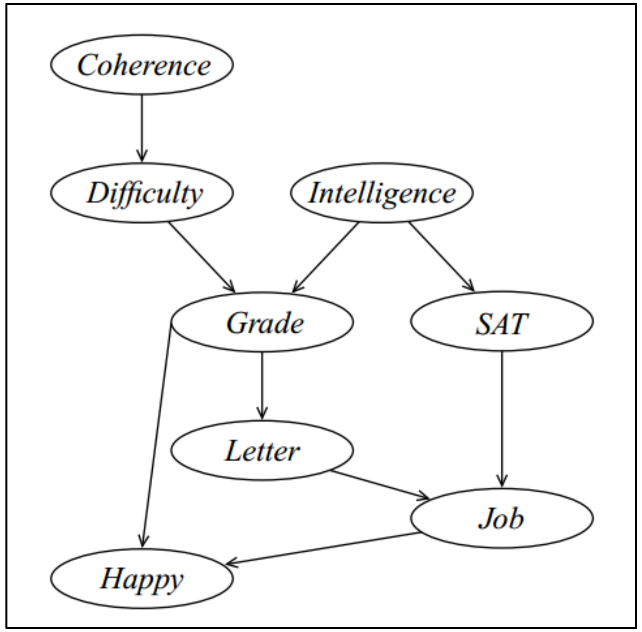
3. *Eliminating I: We compute the factors*

$$\begin{aligned}\psi_3(G,I,S) &= \phi_I(I)\cdot\phi_S(S,I)\cdot\tau_2(G,I)\\ \tau_3(G,S) &= \sum_I \psi_3(G,I,S).\end{aligned}$$

4. *Eliminating H: We compute the factors*

$$\begin{aligned}\psi_4(G,J,H) &= \phi_H(H,G,J)\\ \tau_4(G,J) &= \sum_H \psi_4(G,J,H).\end{aligned}$$

Note that $\tau_4 \equiv 1$ (all of its entries are exactly 1): we are simply computing $\sum_H P(H \mid G,J)$, which is a probability distribution for every G,J, and hence sums to 1. A naive execution of this algorithm will end up generating this factor, which has no value. Generating it has no impact on the final answer, but it does complicate the algorithm. In particular, the existence of this factor complicates our computation in the next step.



5. *Eliminating G: We compute the factors*

$$\begin{aligned}\psi_5(G,J,L,S) &= \tau_4(G,J)\cdot\tau_3(G,S)\cdot\phi_L(L,G)\\ \tau_5(J,L,S) &= \sum_G \psi_5(G,J,L,S).\end{aligned}$$

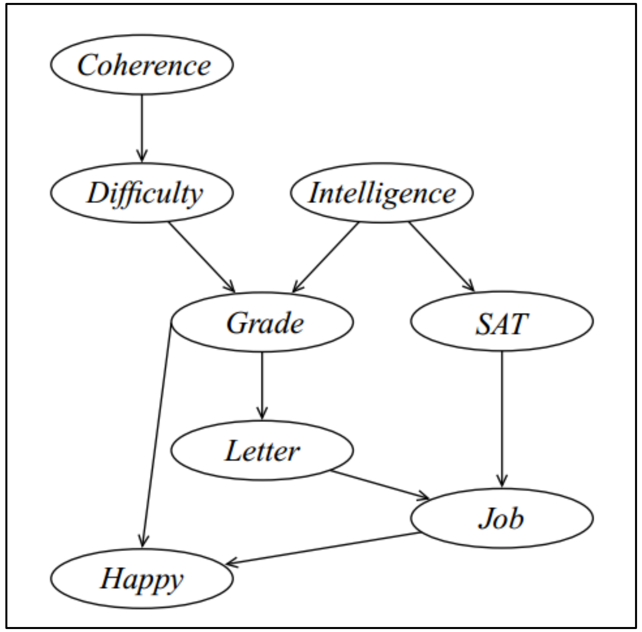
Note that, without the factor $\tau_4(G,J)$, the results of this step would not have involved J.

6. *Eliminating S: We compute the factors*

$$\begin{aligned}\psi_6(J, L, S) &= \tau_5(J, L, S) \cdot \phi_J(J, L, S) \\ \tau_6(J, L) &= \sum_S \psi_6(J, L, S).\end{aligned}$$

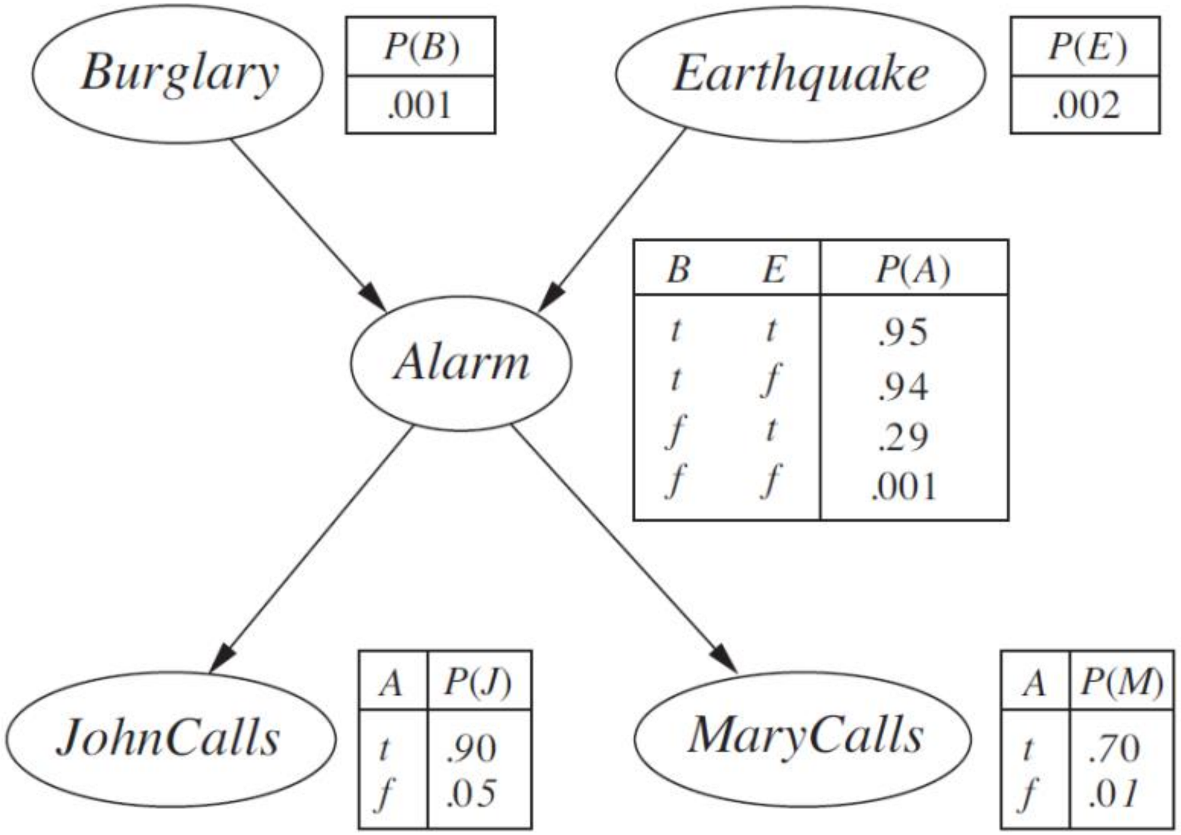
7. *Eliminating L: We compute the factors*

$$\begin{aligned}\psi_7(J, L) &= \tau_6(J, L) \\ \tau_7(J) &= \sum_L \psi_7(J, L).\end{aligned}$$



Koller, Daphne, and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Exercise – Conditional Probability Query



Compute $\mathbf{P}(Burglary \mid JohnCalls = true, MaryCalls = true)$.

Exercise – Conditional Probability Query

VE order: Earthquake (E), Alarm (A)

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

$$= \phi_B(B)\,\phi_E(E)\,\phi_A(A,B,E)\,\phi_J(A,J)\,\phi_M(A,M)$$

$$\tau_1(A,B) = \sum_e \phi_E(e)\,\phi_A(A,B,e)$$

B	τ(A)
1	0.94
0	0.001

$$\tau_2(B) = \sum_a \tau_1(a,B)\,\phi_J(a,J=1)\,\phi_M(a,M=1)$$

τ(B)
0.592

$$P(B|J=1,M=1) = \tau_2(B)\,P(B)$$

P(B|J=1, M=1)
0.00592