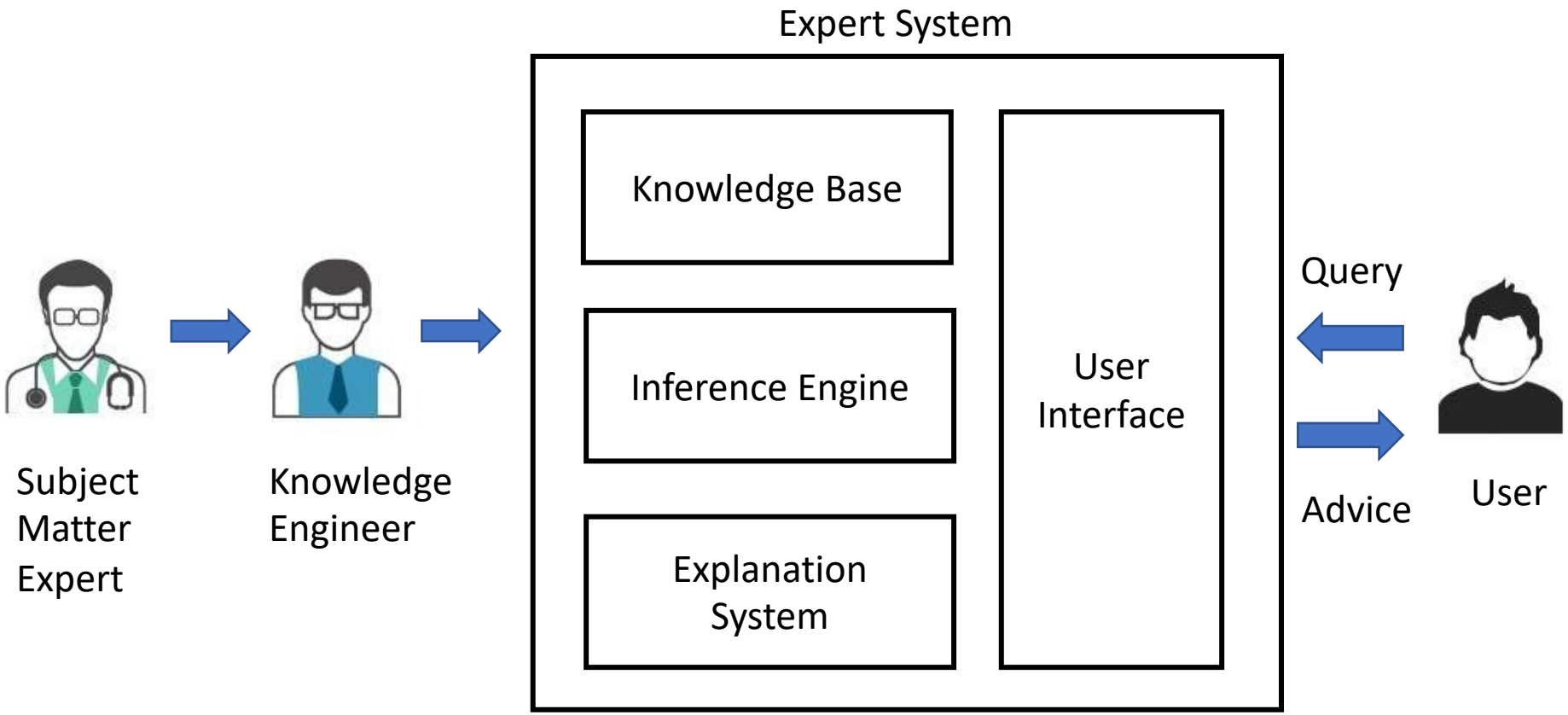
IT5005 Artificial Intelligence

Sirigina Rajendra Prasad
AY2025/2026: Semester 1

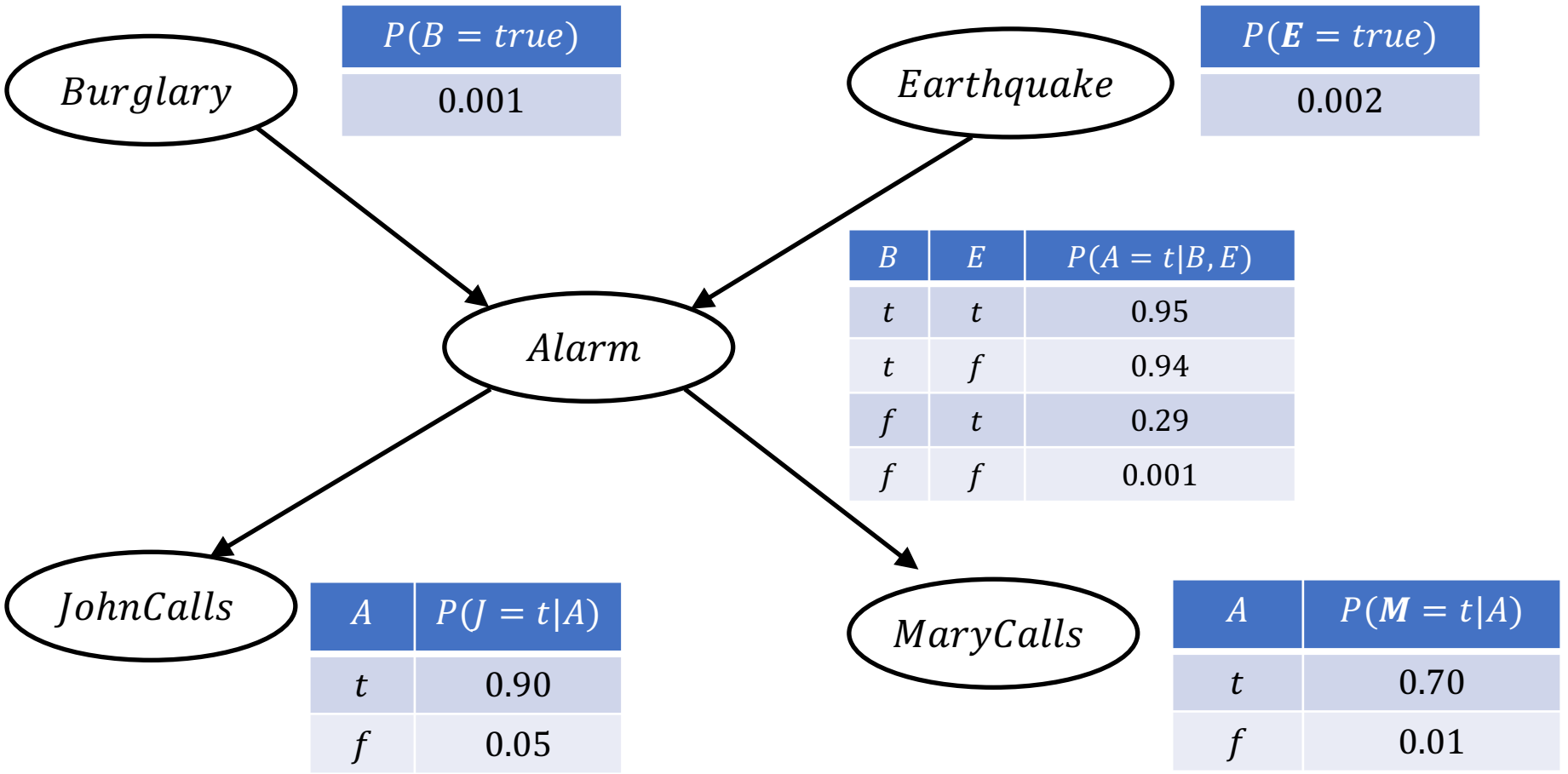
Probabilistic Reasoning

Expert Systems: Probabilistic Inference



Today: Partial Observability and Uncertainty

- Representation of Domain Knowledge as a Bayesian Network



$P(B = true)$
0.001

$P(\boldsymbol{E} = true)$
0.002

B	E	$P(A = t \mid B, E)$
t	t	0.95
t	f	0.94
f	t	0.29
f	f	0.001

A	$P(J = t \mid A)$
t	0.90
f	0.05

A	$P(\boldsymbol{M} = t \mid A)$
t	0.70
f	0.01

Reasoning:

What is the probability that *Earthquake* = True, given *MaryCalls* = True?

So far: Logic-based Systems

- KB contains rules like

IF	the season is autumn
AND	the sky is cloudy
AND	the forecast is drizzle
THEN	the advice is 'take an umbrella'

- What if the advice is:
  - 'take an umbrella with probability 0.8'

Why Probabilistic Inference?

- Logic-based systems are unsuitable, if:
  - It is difficult to define all antecedents
    - Due to large number of rules
  - Complete theory is absent
    - Due to lack of domain knowledge
  - Availability of only partial evidence (observability)
    - Due to lack of sensors

Probabilistic Inference

- Relies on *beliefs*
  - Probabilities of beliefs about query, given some evidence
  - Known as Bayesian (subjective) view

- Prior belief:
  - The belief about the query (x) prior to the observation of the evidence (y)
  - Eg: $P(x)$

- Posterior belief:
  - The belief about the query (x) after the observation of the evidence (y)
  - Eg: $P(x|y)$

Probabilistic Inference

- Knowledge Base:
  - Bayesian Network
- Queries
  - Probability of an event of interest
- Inference:
  - Exact Inference
  - Approximate Inference

Probability Prelims

Preliminaries

- Random Variables (RVs)
- Probability Distributions
  - Joint Probability Distributions
  - Conditional Probability Distributions
- Bayes Rule

Random Variables and Values

- Types of Random Variables:
  - Boolean
    - $Alarm$: $\{alarm, \neg alarm\}$
  - Categorical
    - $Weather$: $\{rainy, sunny, cloudy, snow\}$
    - $Gender$: $\{male, female\}$
  - Discrete
    - $Dice$: $\{1,2,3,4,5,6\}$
  - Continuous (Real numbers)
    - Temperature T: {30}
- An environment can be represented by multiple random variables

Notation:
Variable: Upper case alphabet (Eg: $Alarm$)
Value: Lower case alphabet (Eg: $alarm$)

Boolean Random Variable

- Range: $true\ (1)/false(0)$

Propositional Random Variables

- Example
  - $Alarm = \ true$
  - $Toothache \ = \ true$

- Alternative Representation
  - $Alarm \ = \ true \ (or) \ Alarm \ = \ t \ (or) \ Alarm \ = \ alarm \ (or) \ Alarm \ = a$
  - $Alarm \ = false \ (or) \ Alarm \ = f \ (or) \ Alarm \ = \neg alarm \ (or) \ Alarm \ = \neg a$

- Each proposition is associated with a probability
  - $P(alarm) \ = \ 0.4$ and $P(\neg alarm) \ = \ 0.6$

Multi-Valued Random Variable

- Range:
  - a set of tokens
- Example:
  - $range\{Weather\} = \{sunny, snow, rainy, cloud\}$
  - $range\{Gender\} = \{male, female\}$
  - $range\{Age\} = \{juvenile, teen, adult\}$
  - Proposition
    - $Weather = sunny$
  - Probability
    - $P(Weather = sunny) = P(sunny) = 0.4$

Categorical Random Variables

Discrete Random Variable

- Range:
  - a set of integers (finite/infinite)
- Example:
  - $range\{Die\} = \{1,2,3,4,5,6\}$
  - $range\{Total\} = \{1,2, \ldots., 12\}$ (sum of two dice)
  - Proposition
    - $Total = 11$
  - Probability
    - $P(Total = 11) = P(11) = 0.2$

Continuous Random Variables

- Range: set of real numbers
- Example:
  - $PulseRate = 80$ in bpm (can take any value greater than 0)

- Cannot assign probabilities to real valued random variables
  - $P(PulseRate = 80) = 0$

- Instead we use probability density function and assign probabilities to a range of values

Probability Distribution

- Indicates probabilities of all values assigned to a discrete random variable
- Example:
  - $dom\{Weather\} = \{sunny, snow, rainy, cloud\}$
  - $\boldsymbol{P}(Weather) = [0.6, 0.1, 0.29, 0.01]$
  - It means:
    - $P(Weather = sunny) = P(sunny) = 0.6$
    - $P(Weather = snow) = P(snow) = 0.1$
    - $P(Weather = rainy) = P(rainy) = 0.29$
    - $P(Weather = cloud) = P(cloud) = 0.01$

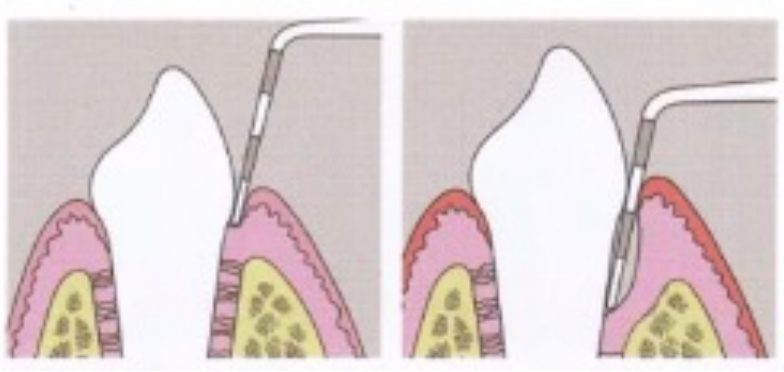
Prior or Unconditional Probabilities

also called Probability Mass Function (PMF)

Joint Probability Distributions

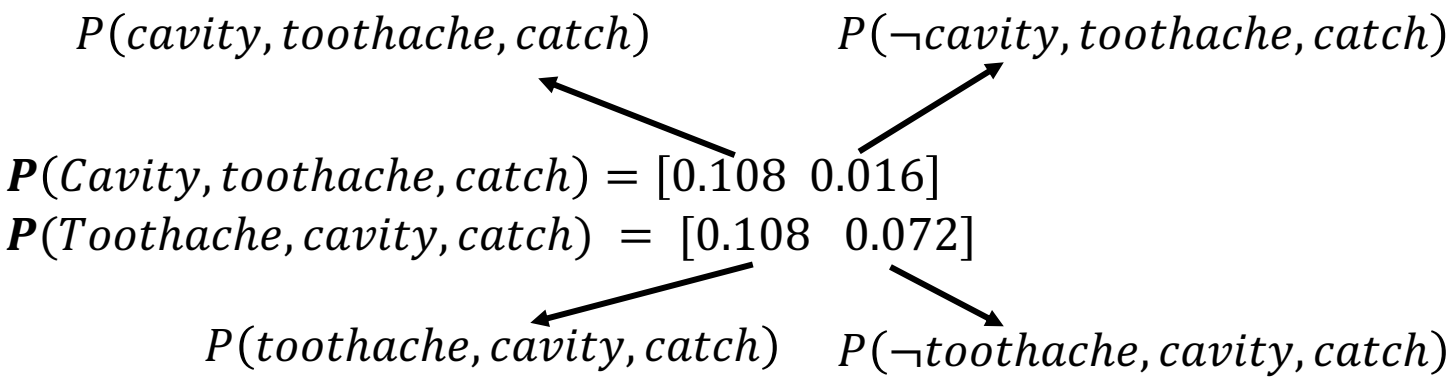
- Defined for a set of random variables
  - Indicates probabilities for all possible assignments of values to all variables
  - **Example**:

$ToothAche$	$Cavity$	$Catch$	$\Pr(ToothAche, Cavity, Catch)$
t	t	t	0.108
t	t	f	0.012
t	f	t	0.016
t	f	f	0.064
f	t	t	0.072
f	t	f	0.008
f	f	t	0.144
f	f	f	0.576



$Catch = true$ means dentist's steel probe catches in teeth (Fig. 13.3 AIMA4e)

Joint Probability Distributions



$ToothAche$	$Cavity$	$Catch$	$\Pr(ToothAche, Cavity, Catch)$
t	t	t	0.108
t	t	f	0.012
t	f	t	0.016
t	f	f	0.064
f	t	t	0.072
f	t	f	0.008
f	f	t	0.144
f	f	f	0.576

Bayes Rule

- $P(a, b) = P(a|b)P(b) = P(b|a)P(a)$

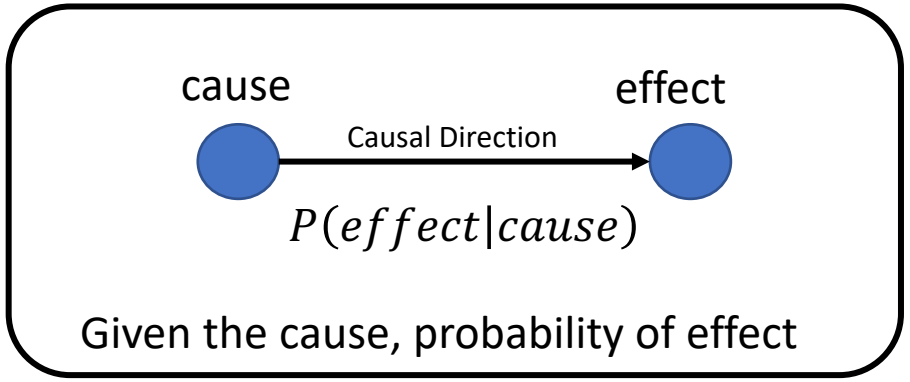
- $P(b|a) = \frac{P(a|b)P(b)}{P(a)}$

Bayes Rule Interpretation

Example:

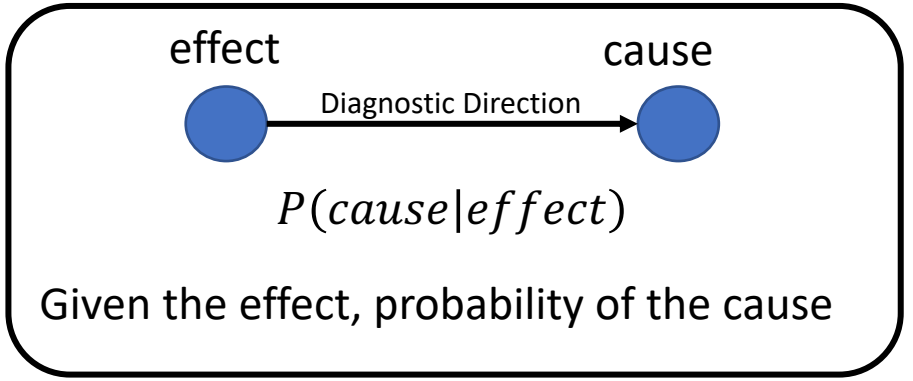
Cause: $Disease$
Effect: $Symptoms$

Causal Direction: $P(symptoms|disease)$
Causal Reasoning



Causal knowledge
(easy to acquire from domain experts)

Diagnostic Direction: $P(disease|symptoms)$
Evidential Reasoning



Diagnostic knowledge

Causal Knowledge for Diagnosis

Joint Probability

Posterior Probabilities

Prior Probabilities or Facts

- $P(cause, effect) = P(cause|effect)P(effect)$
  $= P(effect|cause)P(cause)$

- Diagnosis using causal knowledge
  - $P(cause|effect) = \frac{P(effect|cause)P(cause)}{P(effect)}$
  - Example:
    - $P(disease|symptoms) = \frac{P(symptoms|disease)P(disease)}{P(symptoms)}$

Causal Knowledge for Diagnosis

- Example:
  - $Meningitis = \{m, \neg m\}$
  - $StiffNeck = \{s, \neg s\}$
  - $P(s|m) = 0.7$ → Causal knowledge
  - $P(m) = 1/50000$
  - $P(s) = 0.01$

  } Prior Probabilities or Facts

  - Diagnosis:
    - $P(m|s) = \frac{P(S|m)P(m)}{P(s)} = \frac{0.7*1/50000}{0.01} \approx \frac{1}{700}$

Knowledge of stiff neck increased the probability of Meningitis from $\frac{1}{50000}$ to $\frac{1}{700}$

Examples

X	Y	$P(X,Y)$
t	t	0.2
t	f	0.4
f	t	0.3
f	f	0.1

What is $P(X = t)$?

What is $P(Y = f)$?

What is $P(X = t|Y = f)$?

Probabilistic Inference

- Inference with joint distribution
  - Query the probabilities of a variable

- Queries
  - Without evidence (no observation)
    - Probabilities of events
  - With evidence (after observing some variables from environment)
    - Conditional probabilities of events

Query without Evidence

- How to infer the probability of a variable without evidence?

- Split the variables into two subsets
  - Query Variable (Y)
  - Unknown (unobserved) variables ($\boldsymbol{Z}$)

Y: Random Variable
$\boldsymbol{Z}$: Random Vector
(each element is a Random Variable)
$\boldsymbol{z}$: An assignment of Random Vector

- The joint distribution: $\boldsymbol{P}(Y, \boldsymbol{Z})$

- Probability of Query Variable:
  - $\boldsymbol{P}(Y) = \sum_{\boldsymbol{z} \in \boldsymbol{Z}} \boldsymbol{P}(Y, \boldsymbol{z} = \boldsymbol{Z})$ ⟶ Marginalization

Sum over all possible combinations of possible values of set of variables of $\boldsymbol{Z}$

Query without Evidence: Example

$ToothAche$	$Cavity$	$Catch$	$\Pr(ToothAche, Cavity, Catch)$
t	t	t	0.108
t	t	f	0.012
t	f	t	0.016
t	f	f	0.064
f	t	t	0.072
f	t	f	0.008
f	f	t	0.144
f	f	f	0.576

Query 1:

What is the probability of $Cavity$?

$$\Pr(Cavity = true) = ?$$
$$\Pr(Cavity = false) = ?$$

Query 2:

What is the probability of $cavity$?

$$\Pr(Cavity = true) = ?$$

Query without Evidence: Example

Query 1:
What is the probability of $Cavity$?

Query Variable: $\boldsymbol{Y} = [Cavity]$

Unobserved Variables: $\boldsymbol{Z} = [Catch, Toothache]$

$ToothAche$	$Cavity$	$Catch$	$\Pr(ToothAche, Cavity, Catch)$
t	t	t	0.108
t	t	f	0.012
t	f	t	0.016
t	f	f	0.064
f	t	t	0.072
f	t	f	0.008
f	f	t	0.144
f	f	f	0.576

$\boldsymbol{P}(Cavity) = [P(cavity), P(\neg cavity)]$

$\boldsymbol{P}(Cavity) = \sum_{\boldsymbol{z} \in \{Catch, Toothache\}} \boldsymbol{P}(Cavity, \boldsymbol{z})$

Values

$= \boldsymbol{P}(Cavity, toothache, catch) + \boldsymbol{P}(Cavity, \neg toothache, catch) + \boldsymbol{P}(Cavity, toothache, \neg catch) + \boldsymbol{P}(Cavity, \neg toothache, \neg catch)$

= [0.108 0.016]]+[0.072 0.144]+[0.012 0.064+[0.008 0.576]

= [0.2 0.8]

$P(cavity) = 0.2$

$P(\neg cavity) = 0.8$

Query with Evidence

- Probability of a variable (Y) given some evidence ($\boldsymbol{e}$)
  - $\boldsymbol{P}(Y|\boldsymbol{e})$ (also called *Posterior* Probabilities)

- Evidence ($\boldsymbol{e}$) can be partial
  - Only few variables are known

- Three categories of variables:
  - Query Variable (Y)
  - Evidence Variables ($\boldsymbol{E}$) (also known as observed/known variables)
  - Unobserved Variables ($\boldsymbol{Z}$) (also known as hidden/unknown/latent variables)

Query with Evidence

- How to calculate $\boldsymbol{P}(Y|\boldsymbol{e})$?
- We have: $\boldsymbol{P}(Y, \boldsymbol{E}, \boldsymbol{Z})$
- Given: $\boldsymbol{E} = \boldsymbol{e}$
- $\boldsymbol{P}(Y|\boldsymbol{e}) = \alpha \boldsymbol{P}(Y, \boldsymbol{e}) = \alpha \sum_{\boldsymbol{z}} \boldsymbol{P}(Y, \boldsymbol{e}, \boldsymbol{z})$

Y: Query Variable
$\boldsymbol{E}$: Evidence Variables
$\boldsymbol{Z}$: Unobserved Variables

Normalization Constant

Marginalization of Unobserved variables

Query with evidence: Example

- Query:
  - Given $toothcahe$, what is the probability of $Cavity$?
- Variables:
  - Query ($\boldsymbol{Y}$): $Cavity$
  - Evidence ($\boldsymbol{e}$): $toothache$
  - Unknown ($\boldsymbol{Z}$): $Catch$

$ToothAche$	$Cavity$	$Catch$	$\Pr(ToothAche, Cavity, Catch)$
t	t	t	0.108
t	t	f	0.012
t	f	t	0.016
t	f	f	0.064
f	t	t	0.072
f	t	f	0.008
f	f	t	0.144
f	f	f	0.576

- $\boldsymbol{P(Y|e)} = \alpha \boldsymbol{P(Y,e)} = \alpha \sum_{\boldsymbol{z}} \boldsymbol{P(Y,e,z)}$

Normalization constant

- We need to find $P(cavity|toothache)$ and $P(\neg cavity|toothache)$

Query with evidence: Example

- $\boldsymbol{P}(Y|\boldsymbol{e}) = \alpha \boldsymbol{P}(Y, \boldsymbol{e}) = \alpha \sum_{\boldsymbol{z} \in \boldsymbol{Z}} \boldsymbol{P}(Y, \boldsymbol{e}, \boldsymbol{z})$

- $\boldsymbol{P}(Cavity|toothache) = \alpha\, \boldsymbol{P}(Cavity, toothache)$

Three Boolean random variables: $Toothache$, $Catch$, and $Cavity$

	$toothache$		$\neg toothache$	
	$catch$	$\neg catch$	$catch$	$\neg catch$
$cavity$	0.108	0.012	0.072	0.008
$\neg cavity$	0.016	0.064	0.144	0.576

- $\boldsymbol{P}(Cavity, toothache) = \sum_{z \in \{Catch\}} \boldsymbol{P}(Cavity, toothache, z)$
  - $P(cavity, toothache) = \sum_{z \in \{Catch\}} P(cavity, toothache, z)$ = 0.108+0.012 = 0.12
  - $P(\neg cavity, toothache) = \sum_{z \in \{Catch\}} P(\neg cavity, toothache, z)$ = 0.016+0.064 = 0.08

Marginalization of $Catch$

- $\boldsymbol{P}(Cavity|toothache) = \sum_{z \in \{Catch\}} \boldsymbol{P}(Cavity, toothache, z) = \alpha[0.12 \quad 0.08] = [0.6 \quad 0.4]$

Select α such that values sum to one

$$\alpha = \frac{1}{0.12 + 0.08} = \frac{1}{0.2} = 5$$

Why Bayesian Networks?

Issues with Joint Probability Distributions

- Number of Parameters : d^n
  - n: number of random variables
  - d: size of domain
    - $d = 2$ for Boolean Random Variables

- Direct evaluation doesn't scale well with the number of variables
  - Acquisition of these parameters is also challenging

Independence to the Rescue

- If variables X and Y are independent, i.e., if $X \perp Y$:
  - $\boldsymbol{P}(X|Y) = \boldsymbol{P}(X)$
  - $\boldsymbol{P}(Y|X) = \boldsymbol{P}(Y)$
  - $\boldsymbol{P}(X,Y) = \boldsymbol{P}(X)\boldsymbol{P}(Y)$ -> factoring
- If X and Y are conditionally independent given a variable Z, i.e., if $X \perp Y | Z$:
  - $\boldsymbol{P}(X|Y) \neq \boldsymbol{P}(X)$
  - $\boldsymbol{P}(X|Y,Z) = \boldsymbol{P}(X|Z)$
  - $\boldsymbol{P}(Y|X,Z) = \boldsymbol{P}(Y|Z)$
  - $\boldsymbol{P}(X,Y|Z) = \boldsymbol{P}(X|Z)\boldsymbol{P}(Y|Z)$

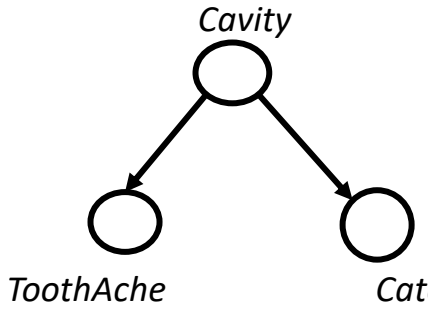
- Independence enables factoring joint distribution into joint distribution on subsets

Conditional Independence: Example

$ToothAche$	$Cavity$	$Catch$	$\Pr(ToothAche, Cavity, Catch)$
t	t	t	0.108
t	t	f	0.012
t	f	t	0.016
t	f	f	0.064
f	t	t	0.072
f	t	f	0.008
f	f	t	0.144
f	f	f	0.576

# of parameters = 8

$$ToothAche \perp Catch \mid Cavity$$



$\Pr(Cavity = t)$
0.2

$Cavity$	$\Pr(ToothAche=t \mid Cavity)$
t	0.6
f	0.1

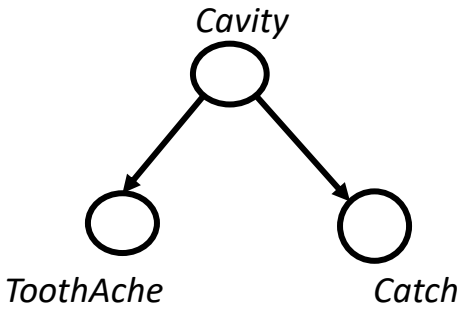
$Cavity$	$\Pr(Catch=t \mid Cavity)$
t	0.9
f	0.2

$$\boldsymbol{P}(ToothAche, Cavity, Catch) = \boldsymbol{P}(Cavity)\boldsymbol{P}(ToothAche|Cavity)\boldsymbol{P}(Catch|Cavity)$$

# of parameters = 5

Conditional Independence: Example

$ToothAche \perp Catch \mid Cavity$



$ToothAche$	$Cavity$	$Catch$	$\Pr(ToothAche, Cavity, Catch)$
t	t	t	0.108
t	t	f	0.012
t	f	t	0.016
t	f	f	0.064
f	t	t	0.072
f	t	f	0.008
f	f	t	0.144
f	f	f	0.576

$Cavity$	$\Pr(ToothAche=t \mid Cavity)$
t	0.6
f	0.1

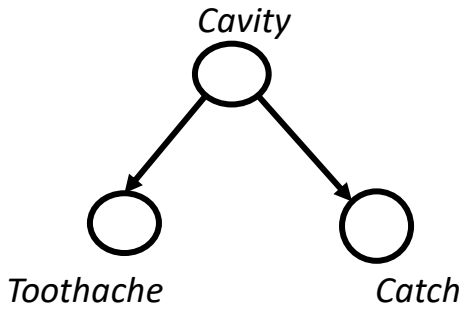
$$\Pr(ToothAche \mid Cavity= t) = \alpha[0.108 + 0.012 \quad 0.072 + 0.008]$$
$$= \alpha[0.12 \quad 0.08]$$

$$\alpha = \frac{1}{0.12+0.08} = \frac{1}{0.2} = 5$$

$$\Pr(ToothAche \mid Cavity= t) = 5[0.12 \quad 0.08] = [0.6 \quad 0.4]$$

$$\Pr(ToothAche{=}t \mid Cavity= t) = \Pr(ToothAche{=}t, Catch = t \mid Cavity = t) + \Pr(ToothAche{=}t, Catch = f \mid Cavity = t)$$
$$= \frac{\Pr(ToothAche{=}t, Catch = t, Cavity = t) + \Pr(ToothAche{=}t, Catch = f, Cavity = t)}{\Pr(Cavity = t)}$$

Bayesian Networks



- Utilizes local structure and conditional independence

- Helps to represent full joint distribution with less number of parameters

Semantics of Bayesian Network

Bayesian Network (BN)

- It is a Directed Acyclic Graph (DAG)
  - Each node is a random variable
    - Can be Boolean/numerical/mixed
  - Nodes are connected by directed edges
    - Edge is from Parent to Child
  - Each node is described with a conditional probability distribution
    - Conditional probability distribution depends only on parents

Example BN

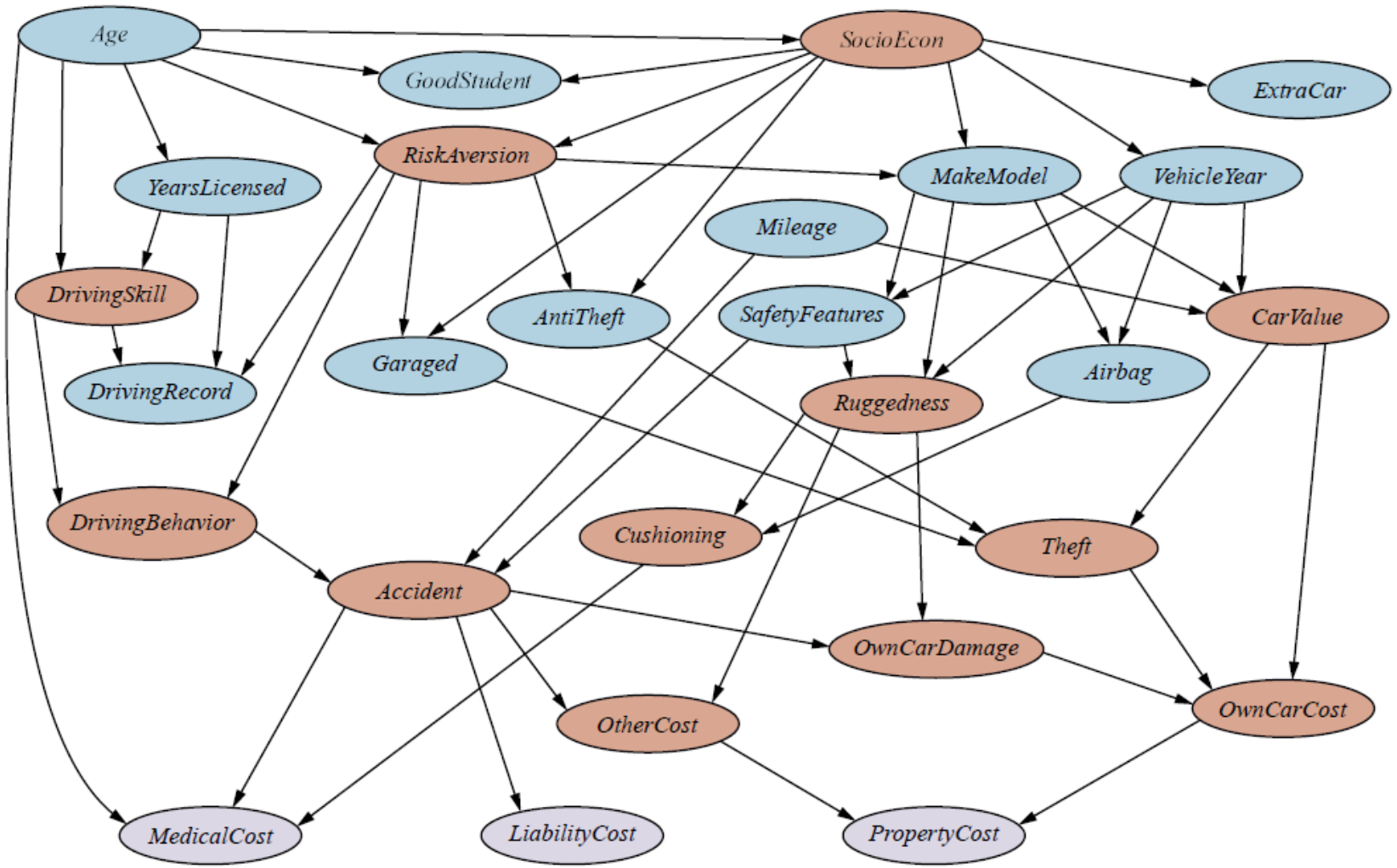


Figure 13.9 A Bayesian network for evaluating car insurance applications.

Brown: Hidden/Unobserved Variables
Lavender: Claims to be paid out
Light Blue: Input Information (obtained from user)

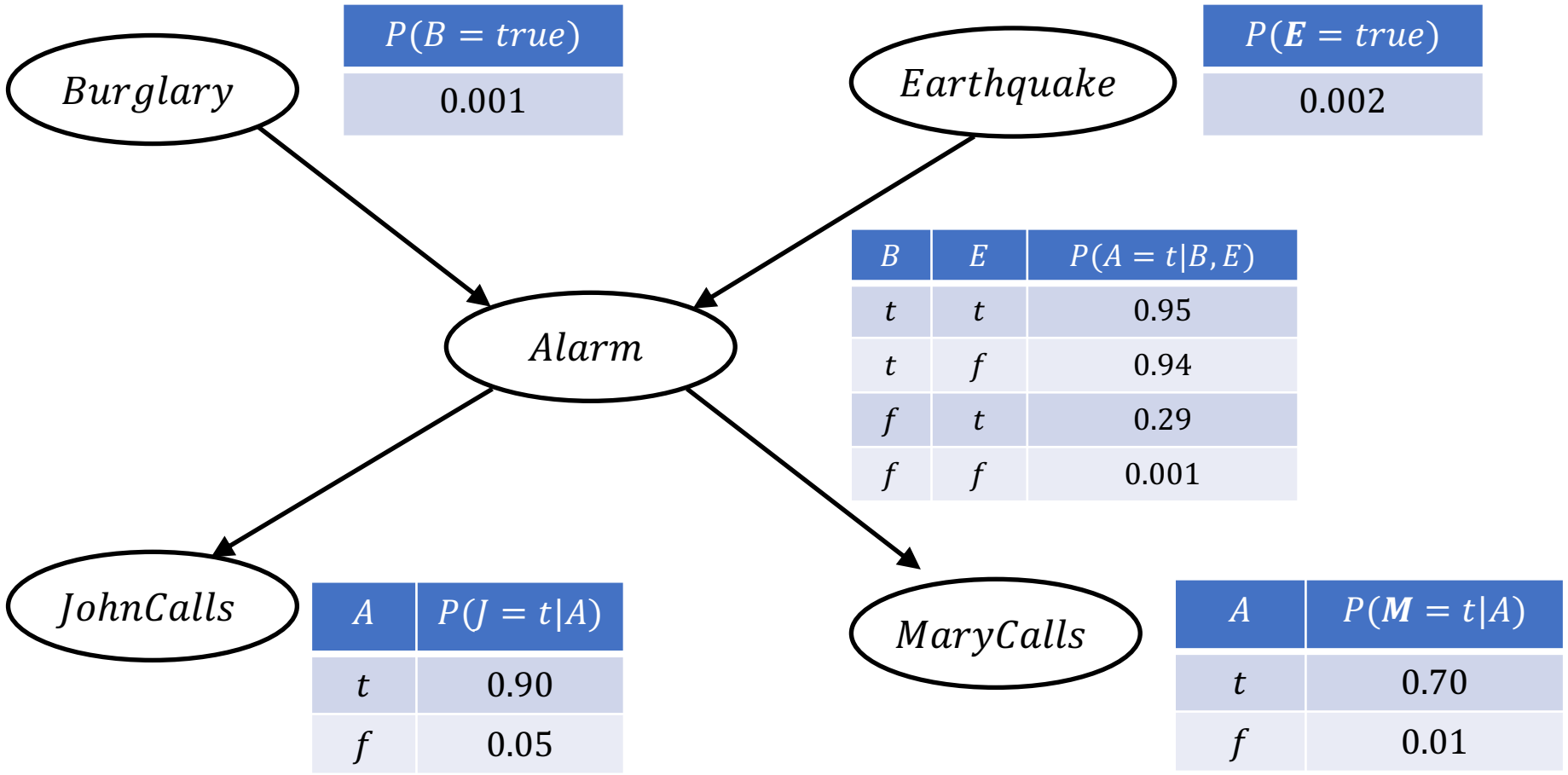
Guidelines to Construct Bayesian Network

- Identify variables $X_1, X_2,\ldots, X_n$
- Choose an ordering of variables
  - Causes must precede effects
- For $i = 1$ to n
  - Add node X_i to network
  - Select minimal set of parents from $X_1, X_2,\ldots,X_{i-1}$ such that
    - $P(X_i|Parents(X_i)) = P(X_i|X_1, \ldots, X_{i-1})$
  - Insert a link from every parent to X_i
  - Write down the CPTs $P(X_i|Parents(X_i))$ for each node
- Joint Probability Distribution for Bayesian Network
  - $P(X_1, X_2, \ldots Xn) = \prod_i P(Xi|Parents(Xi))$

Example

Variable Order: B, E, A, J, M

$\boldsymbol{B}$: $Burglary$
$\boldsymbol{E}$: $Earthquake$
$\boldsymbol{J}$: $Johncalls$
$\boldsymbol{M}$: $Marycalls$
$\boldsymbol{A}$: $Alarm$



$P(B = true)$
0.001

$P(\boldsymbol{E} = true)$
0.002

B	E	$P(A = t \mid B, E)$
t	t	0.95
t	f	0.94
f	t	0.29
f	f	0.001

A	$P(J = t \mid A)$
t	0.90
f	0.05

A	$P(\boldsymbol{M} = t \mid A)$
t	0.70
f	0.01

Total Number of Parameters: 20
Total parameters in JPT: 32 (= 2^5)

CPT: Conditional Probability Table

How BN reduces parameters?

- Let n be the total number of variables.
  - Total number of values: d^n
  - Example:
    - 30 Boolean variables require 2^{30} values
- If each variable has at most k parents:
  - Number of entries in CPT of each variable: d^k
  - Total number of values are upper bounded by nd^k
  - Example:
    - If each variable is influenced by only 5 other variables: 30 $2^5 = 960$ values
- Naïve Bayes net:
  - All variables are independent , we only need nd values
  - Example:
    - 30 Boolean variables need only 30 values

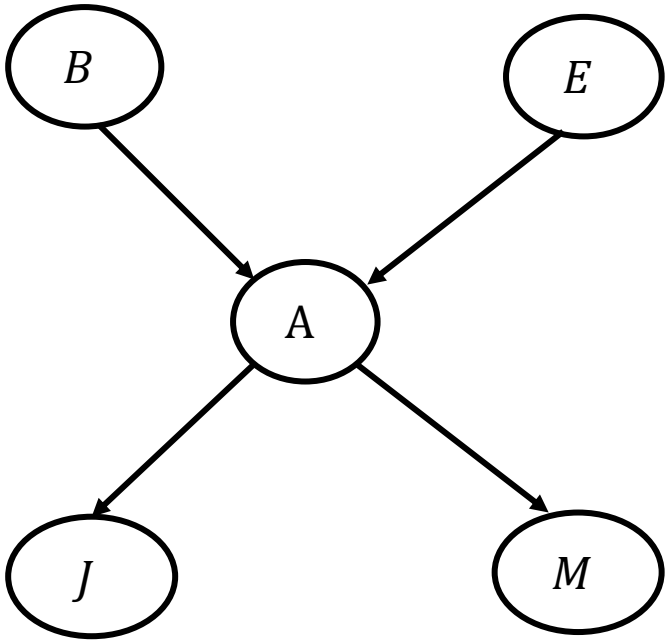
How to construct the Bayesian Network?

- Hand-crafted
  - Build the network
  - Fill the parameters

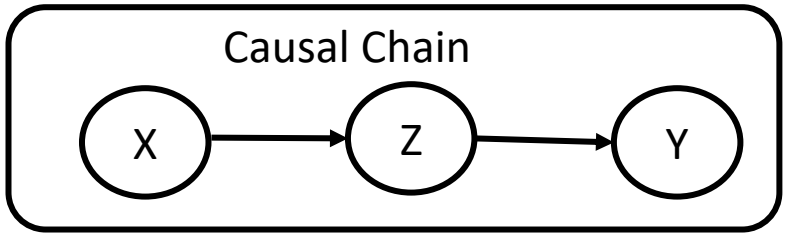
Assume graph and parameters are known
Focus is on inference

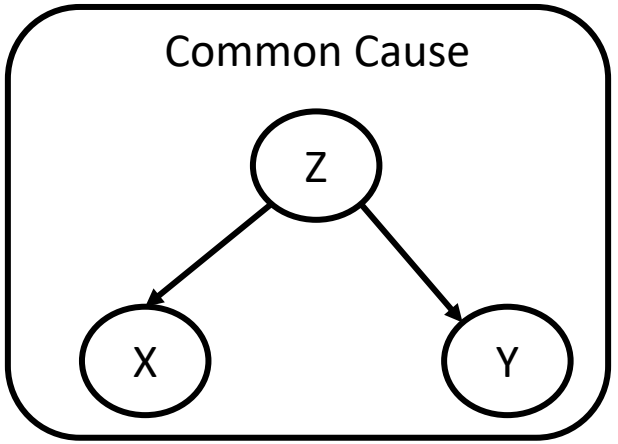
- Data-driven
  - Learn the structure and parameters from data
  - Structure learning from data is still an active research area

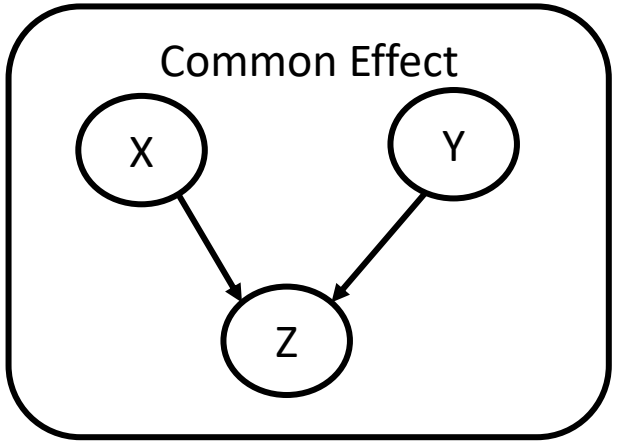
Components of Bayesian Network



$B: Burglary$
$E: Earthquake$
$A: Alarm$
$J: JohnCalls$
$M: MaryCalls$

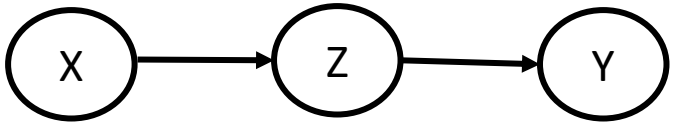






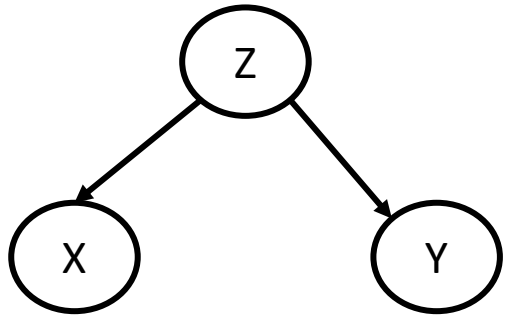
Components of Bayesian Network

Causal Chain



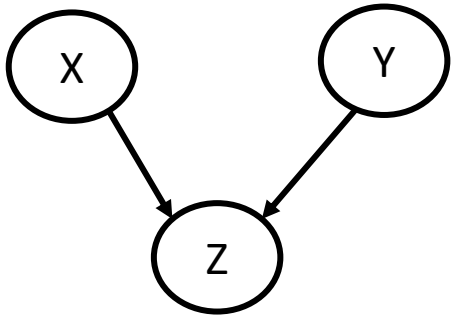
$$P(X,Y,Z) = P(X)P(Z|X)P(Y|Z)$$

Common Cause



$$P(X,Y,Z) = P(Z)P(X|Z)P(Y|Z)$$

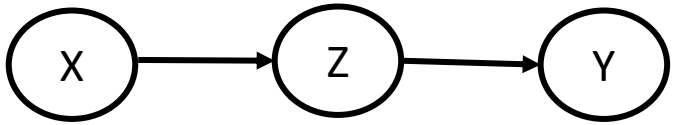
Common Effect



$$P(X,Y,Z) = P(X)P(Y)P(Z|X,Y)$$

Dependencies in Causal Chains

Causal Chain: Indirect Causal Effect



$$P(X, Y, Z) = P(X)P(Z|X)P(Y|Z)$$

Does knowledge of X influence the belief about Y?
Yes, X causes Y indirectly via Z

$$P(Y|X) \neq P(Y)$$

$$X \not\perp Y$$

d-separation in Causal Chains

Assume evidence on Z, i. e. , Z is known

$$P(Y|X,Z) = \frac{P(X,Y,Z)}{P(X,Z)} = \frac{P(X)P(Z|X)P(Y|Z)}{P(Z|X)P(X)} = P(Y|Z)$$

X is independent of Y given Z

(or)

$X \perp Y \mid Z$

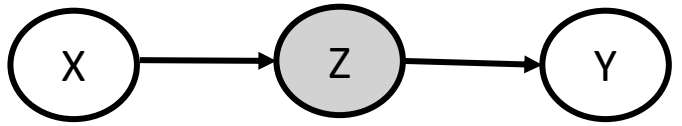
(or)

$P(Y|X,Z) = P(Y|Z)$

(or)

X and Y are d-separated by Z

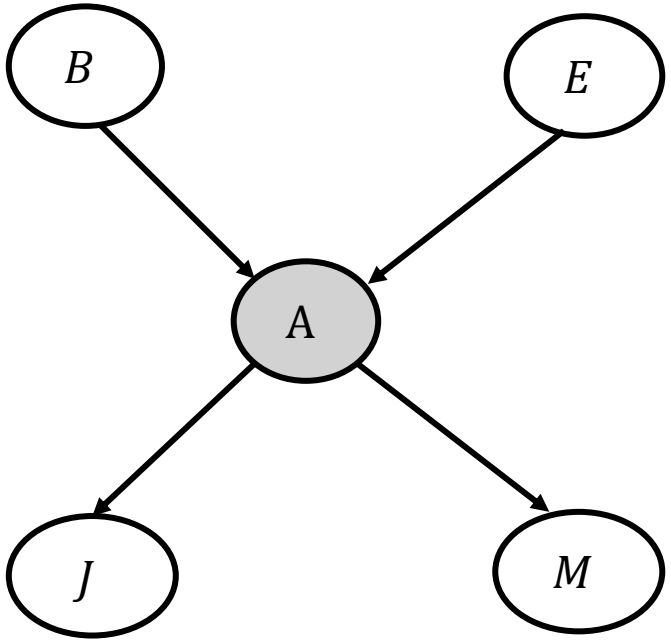
Causal Chain: Indirect Causal Effect



$$P(X,Y,Z) = P(X)P(Z|X)P(Y|Z)$$

- Evidence Z blocks influence of X on Y, i.e., knowledge of X does not influence our belief about Y

d-separation in Causal Chains: Example



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

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

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

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

- Without status of $Alarm, JohnCalls$ and $Burglary$ are dependent
  - Knowledge of $Burglary$ influences belief about $JohnCalls$
- With the knowledge of $Alarm, JohnCalls$ and $Burglary$ are independent, i.e.,
  - Knowledge of $Burglary$ does not influence our belief about $JohnCalls$

Dependencies in Common Cause

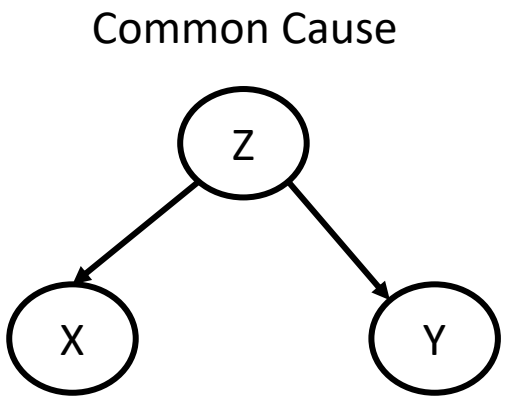
- Does knowledge of X influence the belief about Y?
  Yes, X causes Y indirectly via Z

$$P(Y|X) \neq P(Y)$$

$$X \not\perp Y$$

- X causes Y indirectly via Z

$$P(Y|X,Z) = P(Y|Z)$$
$$P(Z|X,Y) \neq P(Z|Y)$$



If we don't know the common cause, correlation between X and Y leads to dependencies between X and Y

d-separation in Common Cause

Assumption: Z is known

$$P(X,Y|Z) = \frac{P(X,Y,Z)}{P(Z)} = \frac{P(X|Z)P(Y|Z)P(Z)}{P(Z)} = P(X|Z)P(Y|Z)$$

X is independent of Y given Z

(or)

$X \perp Y \mid Z$

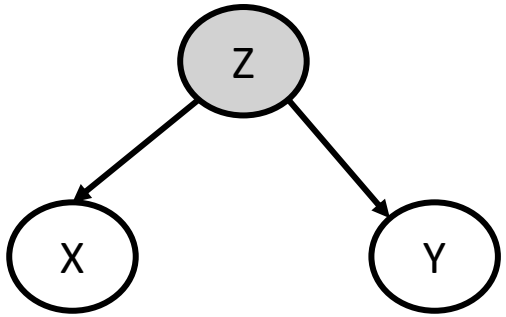
(or)

$P(Y|X,Z) = P(Y|Z)$

(or)

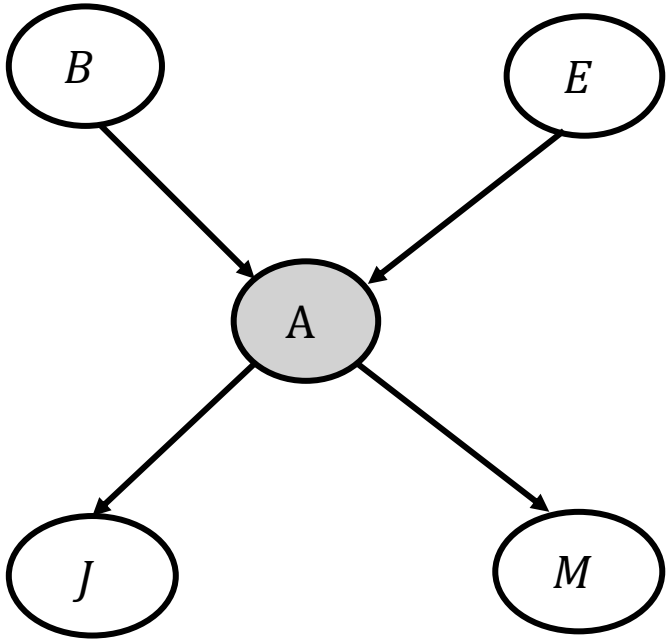
X and Y are d-separated by Z

Common Cause



Evidence Z blocks the influence of knowledge of X on belief about Y
i.e., knowledge of X does not influence our belief about Y

d-separation in Common Cause



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

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

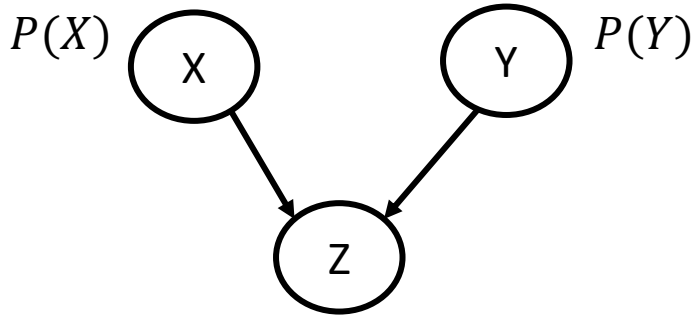
- Without the status of $Alarm$, the knowledge of the event $JohnCalls$, increases our belief about the event $MaryCalls$ and vice-versa
- $JohnCalls$ and $MaryCalls$ are dependent without $Alarm$ status
- Knowledge of $Alarm$ makes $JohnCalls$ and $MaryCalls$ independent events. $Alarm$ explained the reason for $MaryCalls$ and $JohnCalls$

Dependencies in Common Effect

Z is unknown

$$P(X,Y) = \sum_{Z} P(X)P(Y)P(z|X,Y)$$
$$= P(X)P(Y)\ \textstyle\sum_z P(z|X,Y)$$
$$= P(X)P(Y)$$

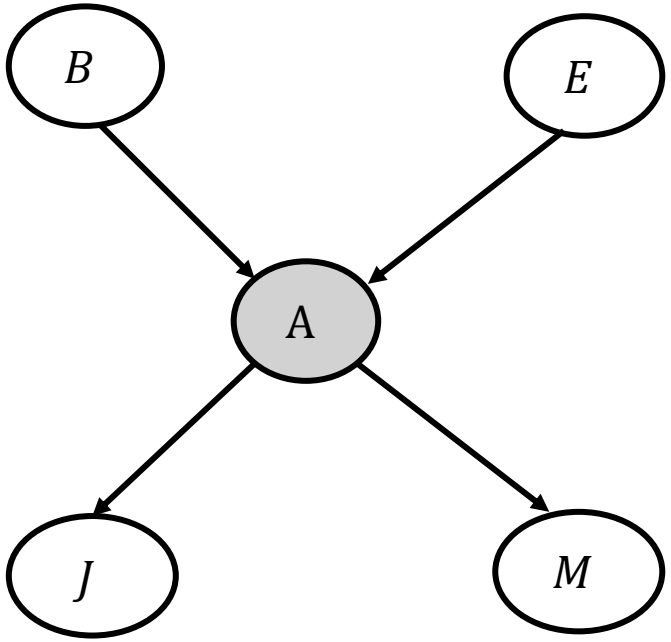
⟶ Marginalize unknown (Z)



If Z is unknown, we are not sure about the events X and Y
Knowledge about one event does not reduce uncertainty about another event

If Z is known, then knowledge of X reduces the uncertainty about Y

Dependencies in Common Effect

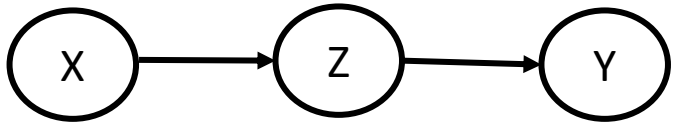


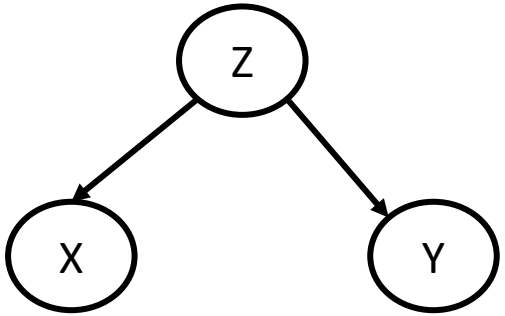
- $Burglary$ and $Earthquake$ are independent events
- If you know $Alarm$ status, then the knowledge of $Burglary$ decreases the uncertainty about $Earthquake$
- So $Burglary$ and $Earthquake$ are not independent given $Alarm$

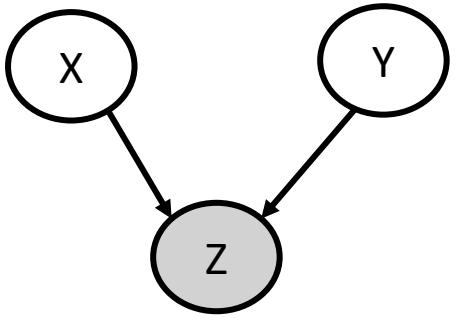
We can exploit the properties of these structures to further simplify the BN

d-Separation: Summary

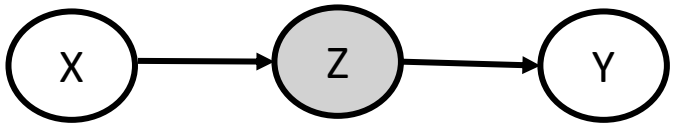
X and Y are dependent

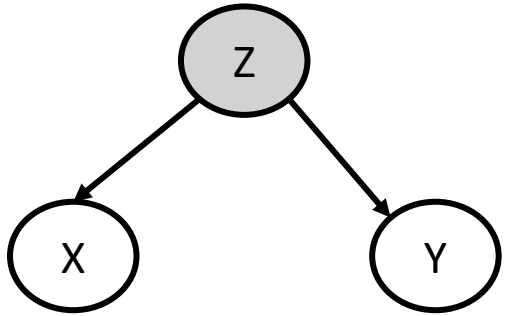


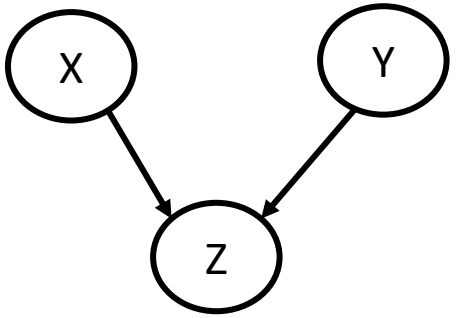




X and Y are independent





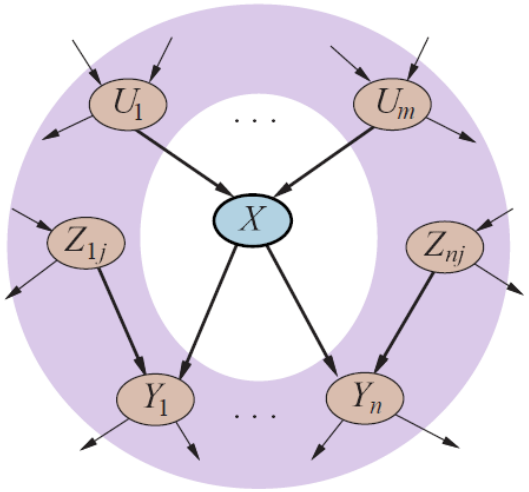


Z is known (given)

Z is unknown

Conditional Independence in BNs

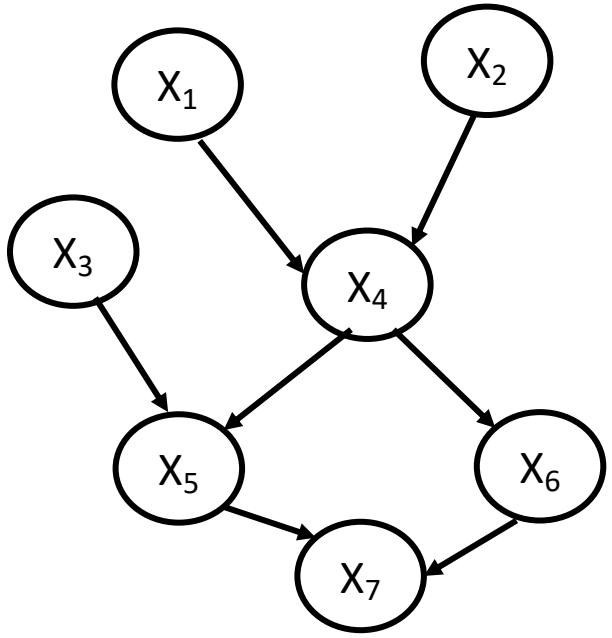
- Markov Blanket (MB):
  - A variable is independent of all other nodes in the network given its
    - Parents
    - Children
    - Children's Parents
- MB provides d-separation



If X is the class variable then Markov Blanket can be used for feature selection

More informative than traditional correlation based feature selection

Example: d-separation



Is X_1 independent of X_2?

Is X_1 independent of X_2 given X_4?

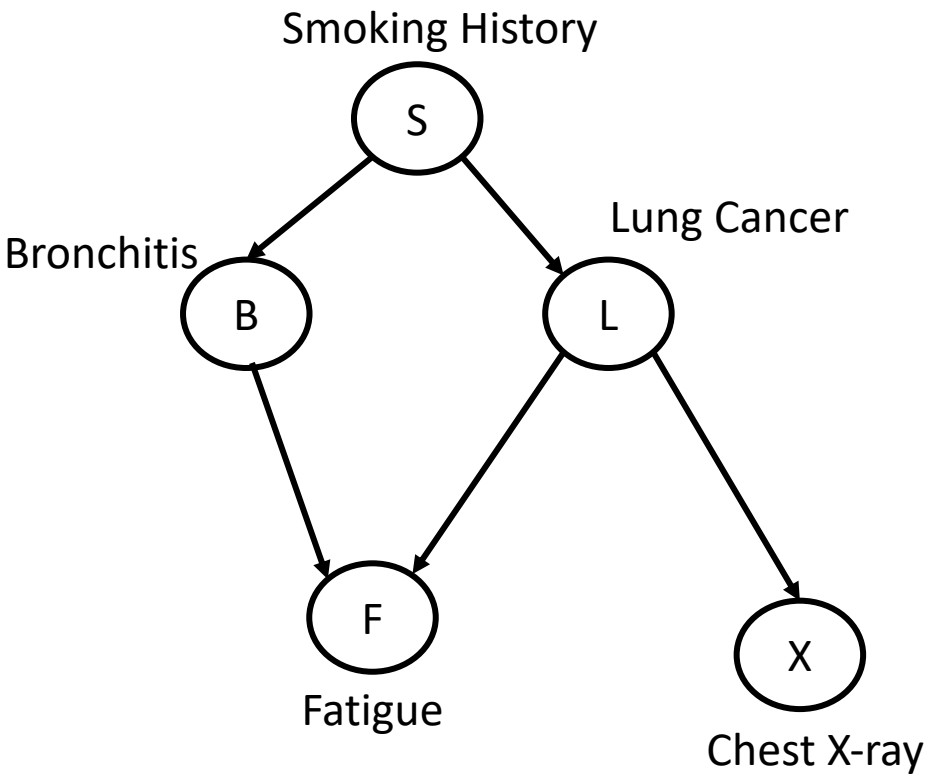
Is X_3 independent of X_7 given X_5?

Procedure to identify d-Separation

- Given a set of nodes $\boldsymbol{Z}$, are the set of nodes $\boldsymbol{X}$ conditionally independent of set of nodes $\boldsymbol{Y}$**?**
  1. Consider the ancestral subgraph consisting of $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$
  2. Construct Moral Graph, i.e., add links between any unlinked pair of nodes that share a common child
  3. Replace directed links by undirected links
  4. If $\boldsymbol{Z}$ blocks all paths between $\boldsymbol{X}$ and $\boldsymbol{Y}$ in the resulting graph, then $\boldsymbol{Z}$ d-separates $\boldsymbol{X}$ and $\boldsymbol{Y}$

AIMA4e: Section 13.2.1

Examples

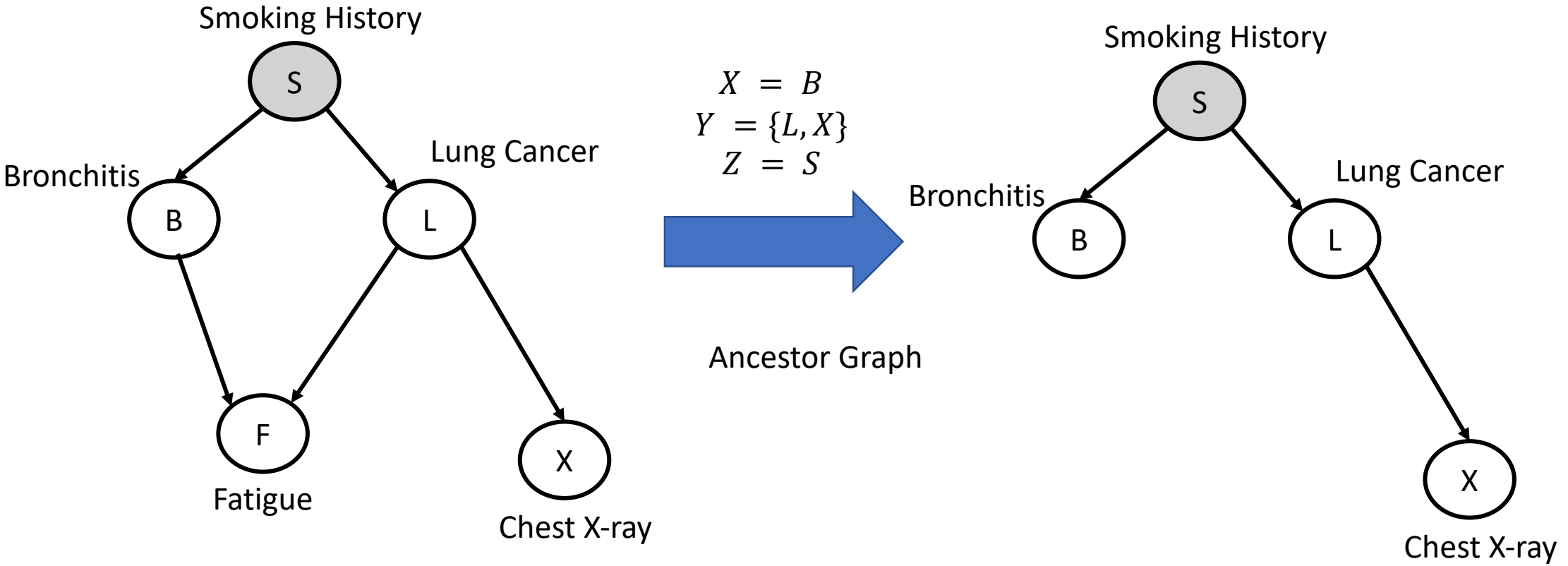


Ex: Query-$\boldsymbol{P}(B|S)$

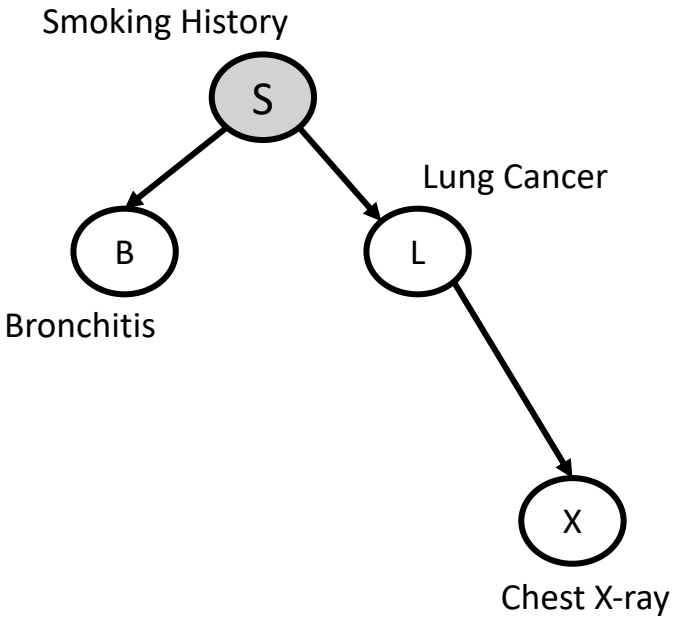
1. $B \perp \{L, X\} \mid S$
2. $F \perp \{S, X\} \mid \{B, L\}$

Example 1: $X = B, Y = \{L, X\}, Z = \{S\}$

- Step 1: Ancestor graph is a subnetwork consisting of
  - Nodes under consideration
  - Parents, Parents of parents, etc. of nodes under consideration

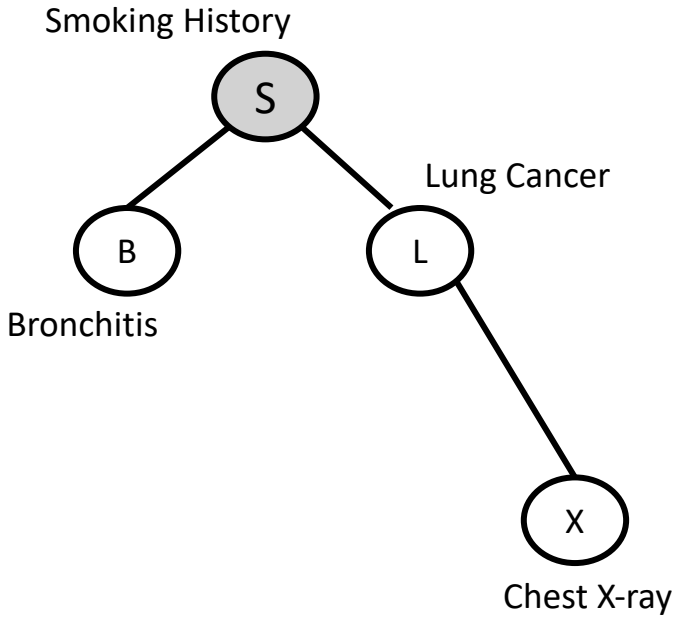


Example 1: $X = B, Y = \{L, X\}, Z = \{S\}$



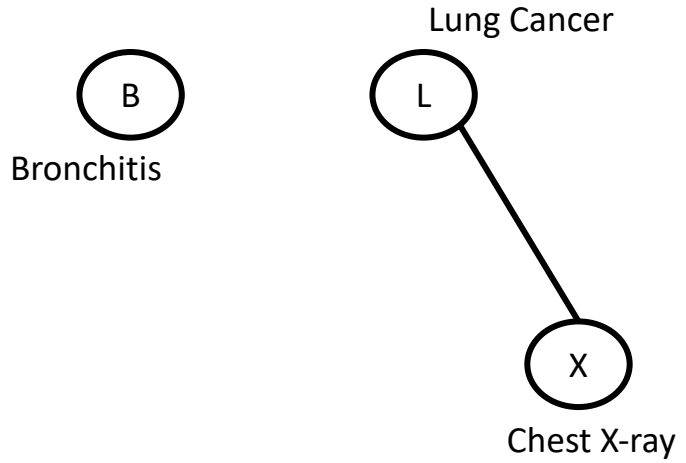
No Common Child
No Moral Graph

Step 2: Moral Graph



Step 3: Undirected Graph

$X = B$
$Y = \{L, X\}$
$Z = S$

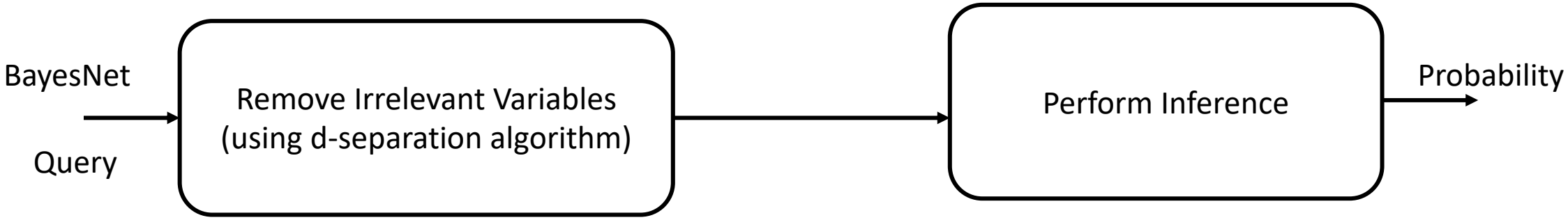


$B \perp \{L, X\} \mid S$

Step 4: Removal of Evidence

Inference on BN

Workflow of Probabilistic Inference



Variables:

Query Variable
Hidden Variable
Evidence Variable
Irrelevant Variable

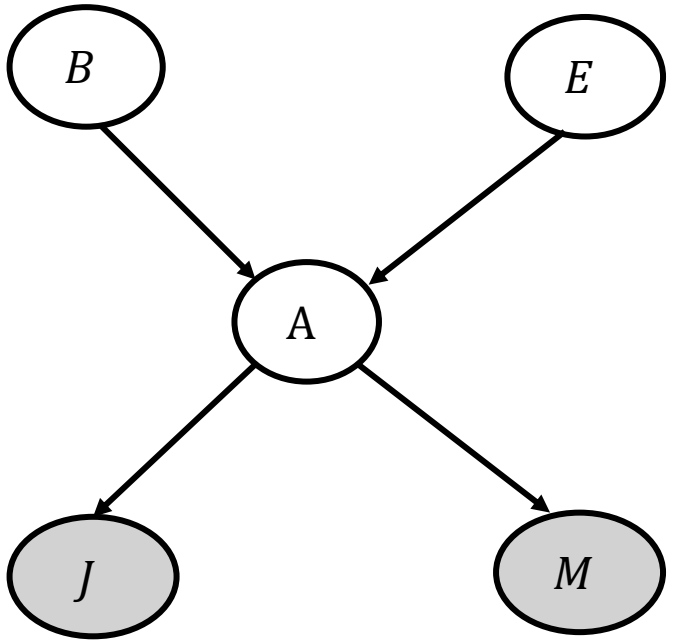
Inference in Bayesian Networks

- Exact Methods
  - Enumeration Method
  - Variable Elimination Method

- Approximate Methods
  - Sampling Techniques
    - Based on stochastic simulation

Enumeration Method

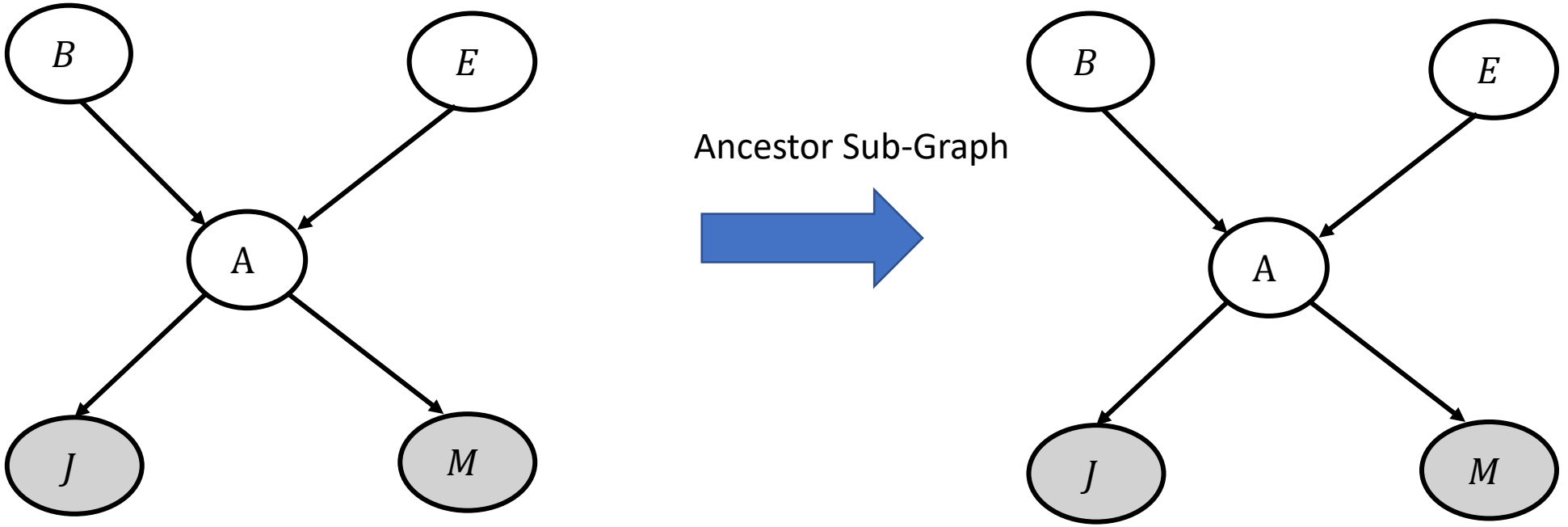
- **Example**:
  - Query Variable: B
  - Evidences: j, m
  - Hidden Variables: E, A



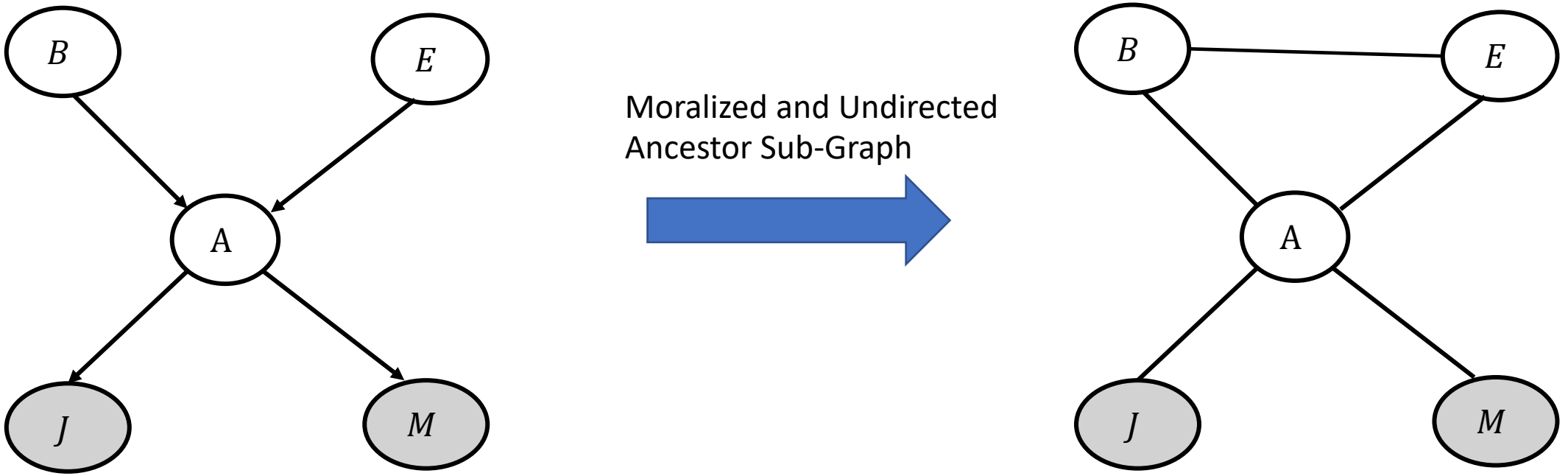
- **Query: $\boldsymbol{P}(B|j,m)$**

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

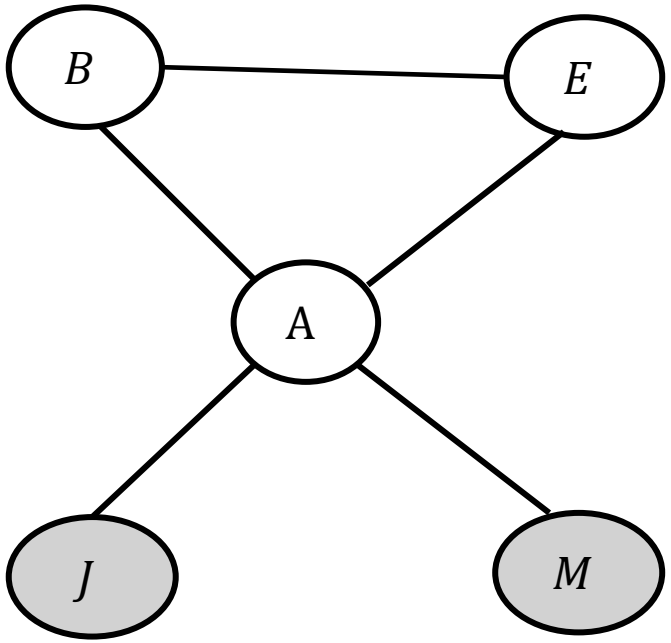
Identify Irrelevant Variables



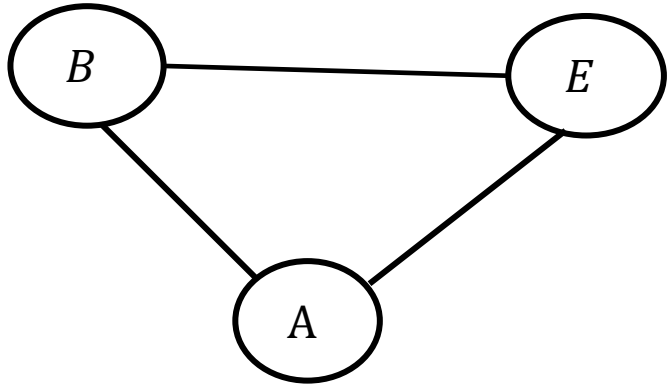
Identify Irrelevant Variables



Identify Irrelevant Variables



Remove evidence variables



E and A are relevant variables for the query B

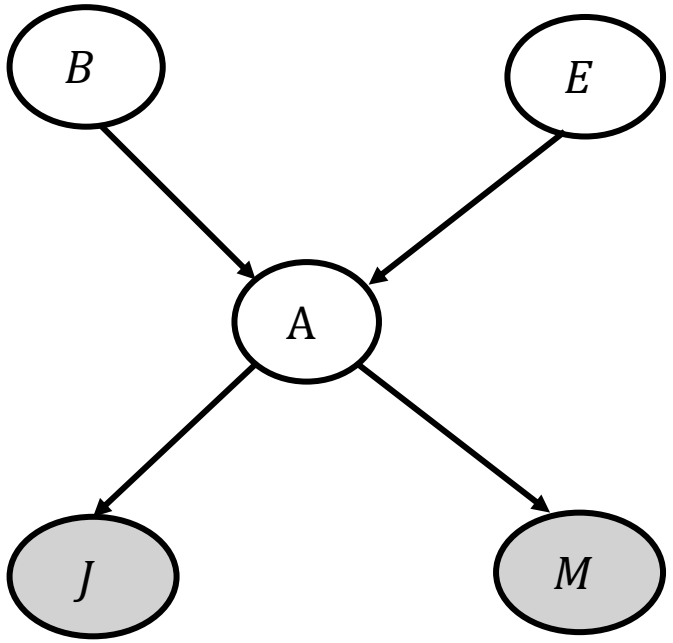
Additional Material

(only for your reading)

Enumeration Method

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

$$\begin{aligned}
\boldsymbol{P}(B|j,m) &= \alpha \boldsymbol{P}(B,j,m) \\
&= \alpha \sum_e \sum_a \boldsymbol{P}(B,j,m,e,a) \\
&= \alpha \sum_e \sum_a \boldsymbol{P}(B) P(e) P(a|B) P(j|a) P(m|a) \\
&= \alpha P(B) \sum_e P(e) \sum_a P(a|b,e) P(j|a) P(m|a)
\end{aligned}$$



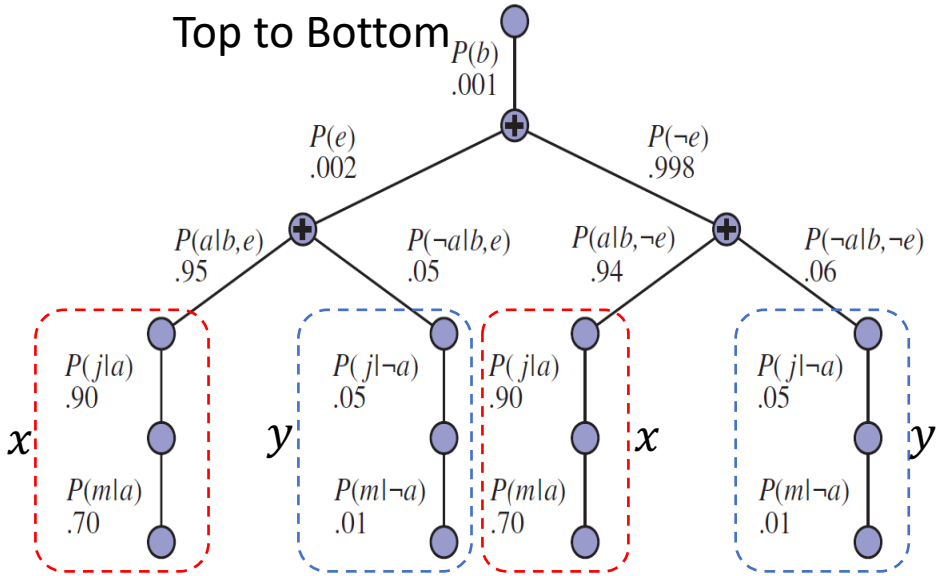
Enumeration Method

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

$$= \alpha(P(b)\, P(e) P(a|b,e) x + P(b) P(e) P(\neg a|b,e) y +$$
$$P(b)\, P(\neg e) P(a|b,\neg e) x + P(b)\, P(\neg e) P(\neg a|b,\neg e) y)$$

Left to Right

Top to Bottom



Similar to depth-first search algorithm

Requires less memory

But repeated computations

Alternative: Bottom-Up Approach (Memoization in Dynamic Programming) - Variable Elimination Method

ENUMERATION-ASK Algorithm

function ENUMERATION-ASK(X, **e**, bn) **returns** a distribution over X
  inputs: X, the query variable
    e, observed values for variables **E**
    bn, a Bayes net with variables $vars$

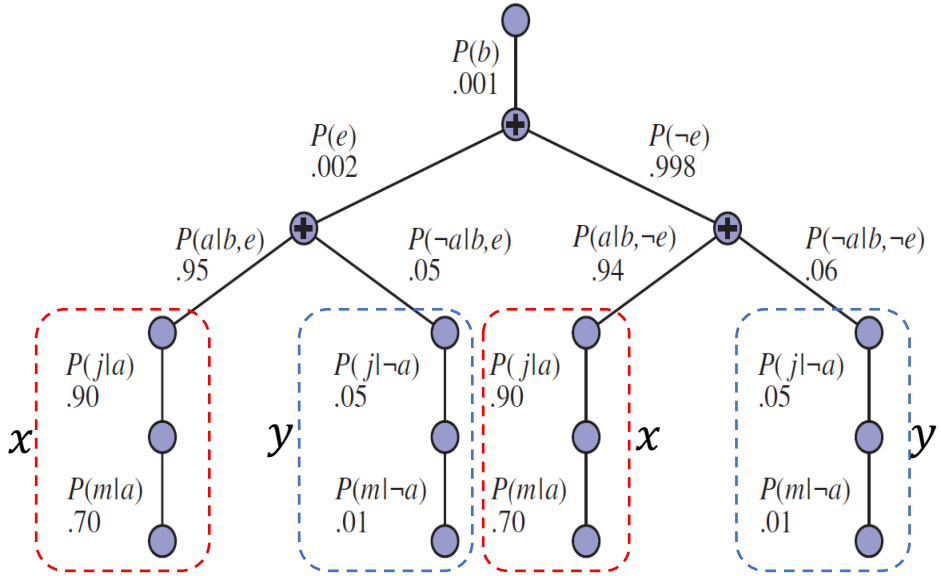
  $\mathbf{Q}(X) \leftarrow$ a distribution over X, initially empty
  for each value x_i of X **do**
    $\mathbf{Q}(x_i) \leftarrow$ ENUMERATE-ALL($vars$, $\mathbf{e}_{x_i}$)
      where $\mathbf{e}_{x_i}$ is **e** extended with $X = x_i$
  return NORMALIZE($\mathbf{Q}(X)$)

function ENUMERATE-ALL($vars$, **e**) **returns** a real number
  if EMPTY?($vars$) **then return** 1.0
  $V \leftarrow$ FIRST($vars$)
  if V is an evidence variable with value v in **e**
    then return $P(v \mid parents(V)) \times$ ENUMERATE-ALL(REST($vars$), **e**)
    else return $\sum_v P(v \mid parents(V)) \times$ ENUMERATE-ALL(REST($vars$), $\mathbf{e}_v$)
      where $\mathbf{e}_v$ is **e** extended with $V = v$

marginalization

Store Intermediate Results

- Bottom-up approach in dynamic programming
- Store x and y and reuse them



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

Start from right side

- Called *Variable Elimination*
  - Variables are marginalized (eliminated) from right side
  - Results stored as factors

Stochastic Simulation

- Relies on generation of samples from probability distribution
- Sampling Methods
  - Direct Sampling / Prior Sampling
  - Rejection Sampling

Sample Generation

- How to generate samples from a given distribution

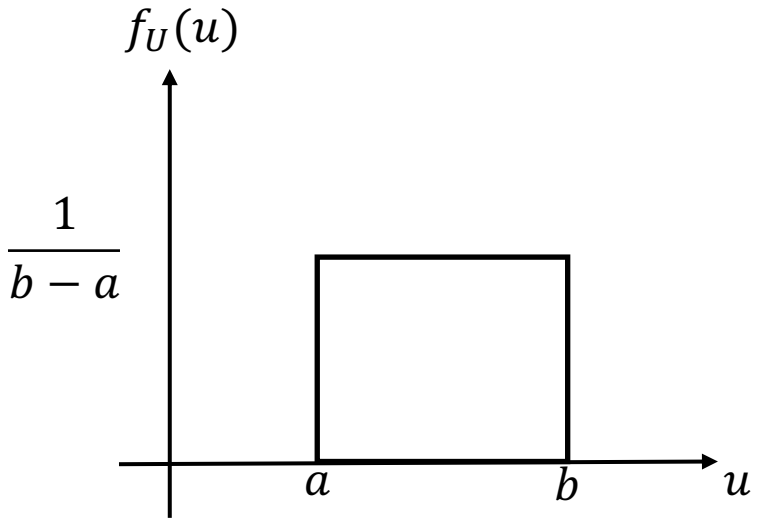
Inverse transform sampling

From Wikipedia, the free encyclopedia

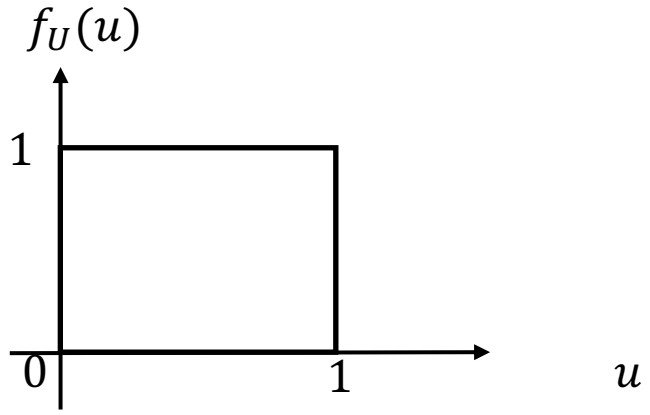
Inverse transform sampling (also known as **inversion sampling**, the **inverse probability integral transform**, the **inverse transformation method**, **Smirnov transform**, or the **golden rule**[1]) is a basic method for pseudo-random number sampling, i.e., for generating sample numbers at random from any probability distribution given its cumulative distribution function.

https://en.wikipedia.org/wiki/Inverse_transform_sampling

Uniform Distribution



Special Case ($a = 0, b = 1$)



$$F_U(u) = \Pr(U \leq u)$$

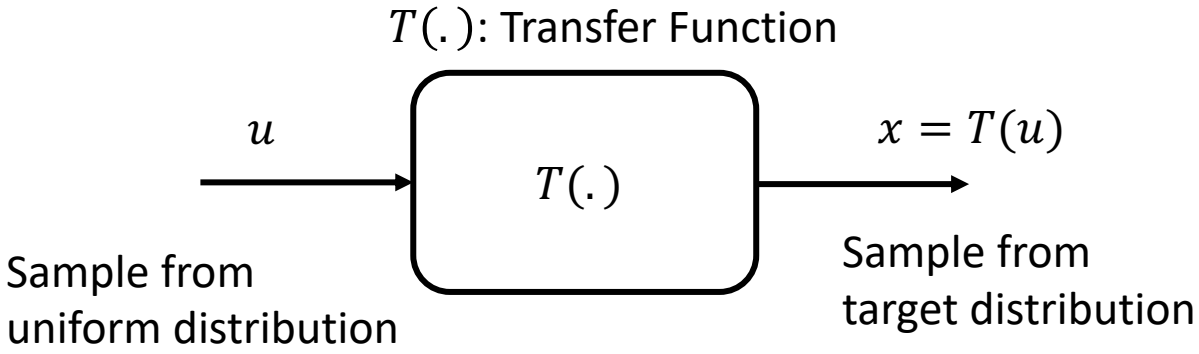
$$= \int_{-\infty}^{u} f_U(u)\, du$$

$$F_U(u) = \Pr(U \leq u)$$

$$= \begin{cases} u & 0 \leq u \leq 1 \\ 1 & u > 1 \\ 0 & u < 0 \end{cases}$$

$f_X(x)$: Probability Density Function
$F_X(x)$: Cumulative Distribution Function

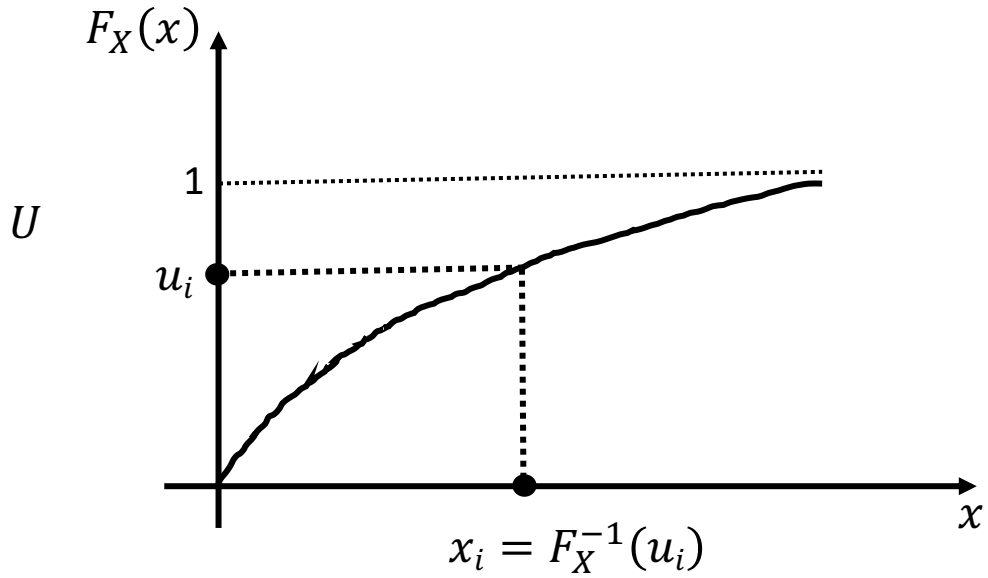
Inverse Transform Sampling



$$
\begin{aligned}
F_X(x) &= \Pr(X \le x) \\
&= \Pr(T(U) \le x) \\
&= \Pr\left(U \le T^{-1}(x)\right) \\
&= T^{-1}(x)
\end{aligned}
$$

$\Rightarrow T(u) = F_X^{-1}(u), u \in [0,1]$

Inverse Transform Sampling

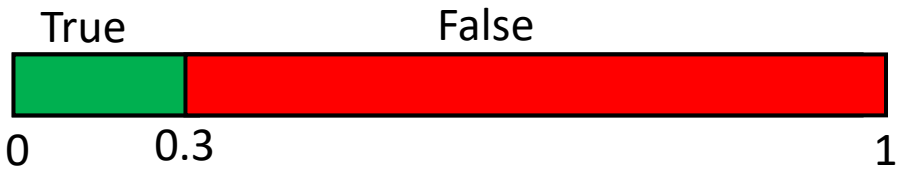


Inverse Transform Sampling: Example

Distribution of Variable X

Value	Probability
True (T)	0.3
False (F)	0.7

CDF of X: $F_X(x)$



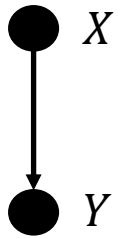

```
>>> import random
>>> random.uniform(0,1)
0.2894913983065621
```

Sampling BayesNet

- Prior Sampling

- Rejection Sampling

X	$\Pr(X)$
t	0.5
f	0.5



$\boldsymbol{X}$	$\Pr(Y = \boldsymbol{t}\|\boldsymbol{X})$
t	0.1
f	0.5

Prior Sampling

- Don’t take evidence into account while sampling
  - Hence the name prior sampling

$$\hat{P}(X|e) = \frac{N_{PS}(X,e)}{N_{PS}(e)}$$

$$\approx \frac{P(X,e)}{P(e)}$$

$$= P(X|e)$$

$$P(x_1, x_2, \ldots, x_n) = \frac{N_{PS}(x_1, x_2, \ldots, x_n)}{N}$$

$N_{PS}(x_1, x_2, \ldots, x_n)$: number of times the event $(x_1, x_2, \ldots, x_n)$ occurs
N: Total number of samples

Prior Sampling

- Don’t take evidence into account while sampling
  - Hence the name prior sampling

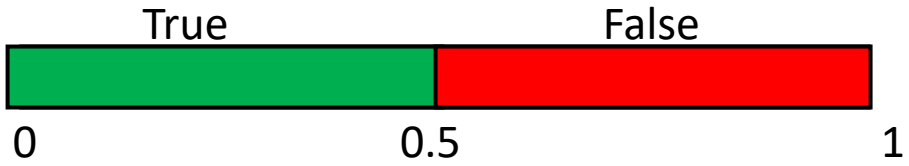
function PRIOR-SAMPLE(bn) **returns** an event sampled from the prior specified by bn
  inputs: bn, a Bayesian network specifying joint distribution $\mathbf{P}(X_1, \ldots, X_n)$

  $\mathbf{x} \leftarrow$ an event with n elements
  for each variable X_i **in** $X_1, \ldots, X_n$ **do**
    $\mathbf{x}[i] \leftarrow$ a random sample from $\mathbf{P}(X_i \mid parents(X_i))$
  return x

Figure 13.16 A sampling algorithm that generates events from a Bayesian network. Each variable is sampled according to the conditional distribution given the values already sampled for the variable’s parents.

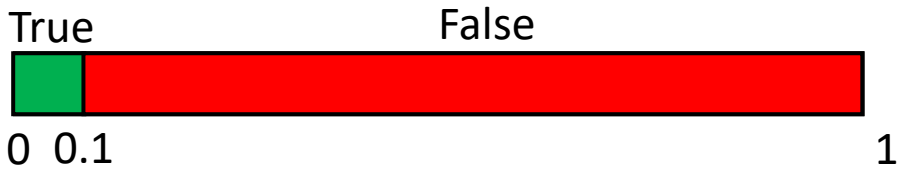
Prior Sampling: Example

1. Sample X according to $P(X)$

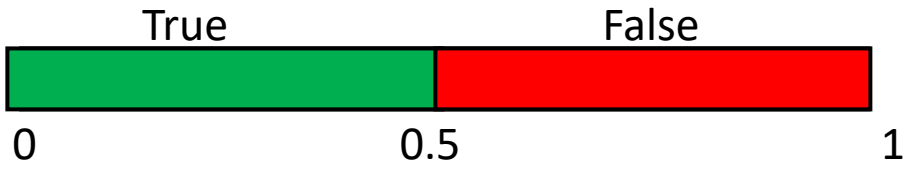


2. Sample Y according to $P(Y|X = x)$

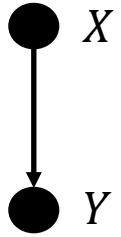
a. If $X = true$ in previous sample



a. If $X = false$ in previous sample



Query:
$P(Y|X = true)$



X	$\Pr(X)$
t	0.5
f	0.5

X	$\Pr(Y = t\|X)$
t	0.1
f	0.5

Rejection Sampling

- Useful for queries given some evidence
- Rejects samples that do not match evidence
- Saves time and space when evidence is at ancestral nodes of query

Rejection Sampling

function $REJECT - SAMPLING(bn, \boldsymbol{e}, X, N)$ **returns** an estimate $\boldsymbol{P}(X|e)$

inputs: X, the query variable

$\boldsymbol{e}$, the observed variables E

bn, a Bayesian network

N, the number of samples to be generated

local variables: $\boldsymbol{C}$, a vector counts for each of value of X, initially zero

for $j = 1$ to N **do**

reject = false

for all variables x_i **in** bn **do**:

sample x_i from $P(x_i|Parents(x_i))$

if x_i is consistent with evidence:

continue

else:

reject = true

break

if not reject:

$\boldsymbol{C}[j] = \boldsymbol{C}[j] + 1$ where x_j is the value of X in $\boldsymbol{x}$

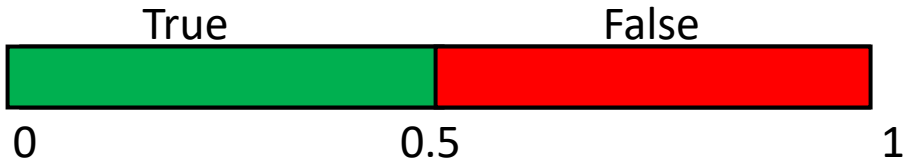
return $NORMALIZE(\boldsymbol{C})$

Algorithm in Figure 13.7 samples all variables and reject if evidence does not match

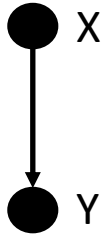
⟸ Saves memory as it does not save rejected samples

Rejection Sampling: Example

1. Sample X according to $P(X)$



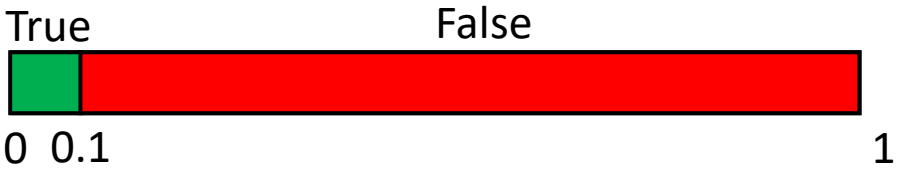
Query:
$P(Y|X = true)$



X	Pr(X)
t	0.5
f	0.5

2. Sample Y according to $P(Y|X = x)$

a. If $X = true$ in previous sample



X	Pr(Y|x)
t	0.1
f	0.5

a. If $X = false$ in previous sample, reject the sample and start sampling X again

Issues with Rejection Sampling

- Number of rejections increases with increase in evidence
- Results in less number of samples
- Less number of samples implies less confidence in estimation

Conclusion

- Partial observation leads to uncertainty
- Probabilistic reasoning is intractable with large number of variables
- Bayesian networks
  - Enforces structure
  - Reduces number of parameters
  - Reduces complexity of inference