

CS 461: Machine Learning Principles

Class 20: Nov. 14

Bayesian Network

Instructor: Diana Kim

Outline

1. Bayesian Network :
DAG (**D**irected **A**cyclic **G**raph) & CPT (**C**onditional **P**robability **T**able)
2. Inference using Bayesian Network:
Direct Inference: Variable Elimination
Indirect Inference: MCMC (Markov Chain Monte Carlo)
3. Learning Bayesian Network:
when a structure is given and all data points are observed
when a structure is given and partial data points are observed (EM)

Bayes Net defines a joint density by a structure and parameters

- structure: a DAG encodes conditional independence relations among R.Vs.
- parameters: the conditional densities. (CPT: Conditional Probability Table)

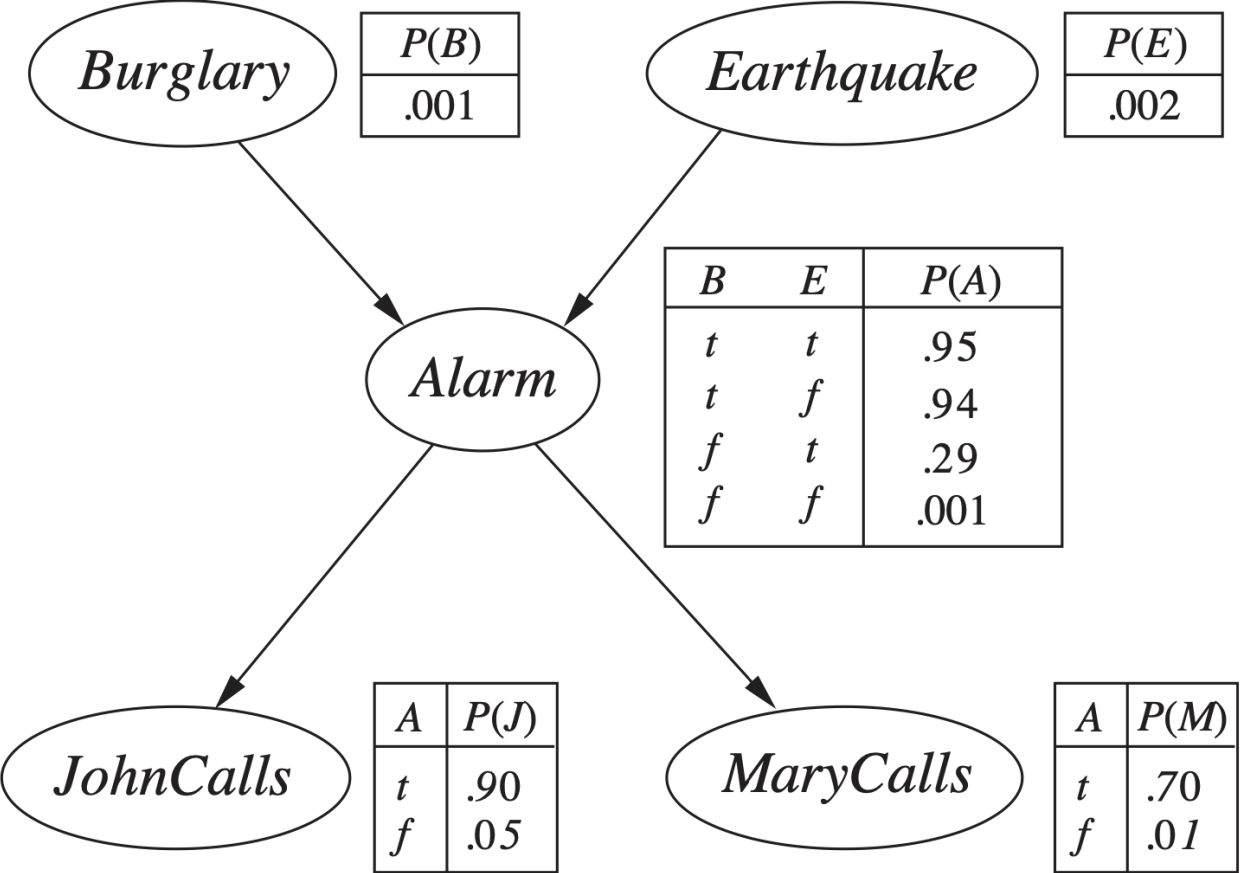


Figure 14.2 From the book “AI: A Modern Approach”

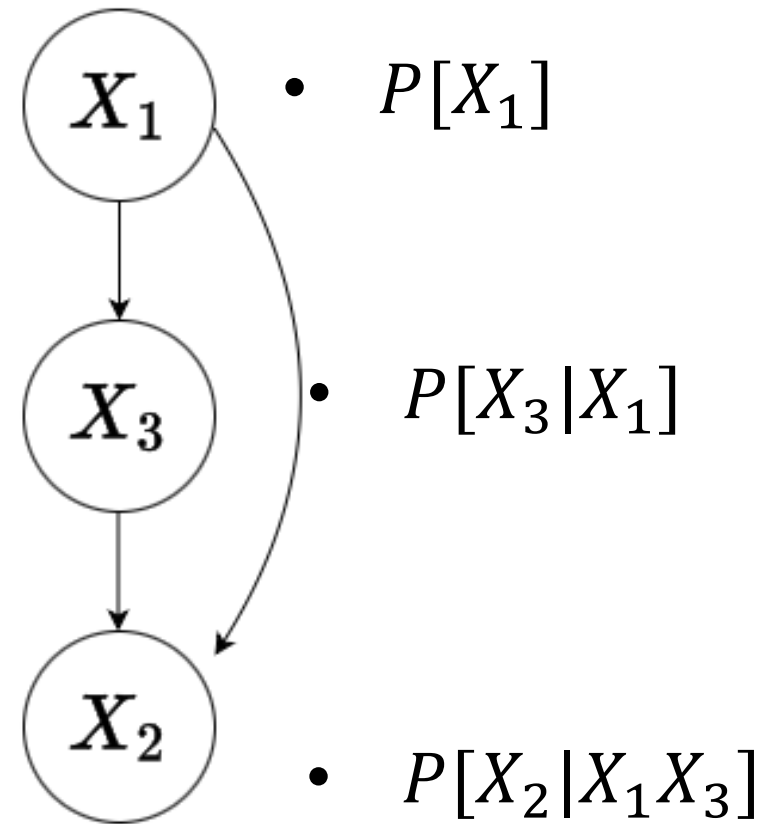
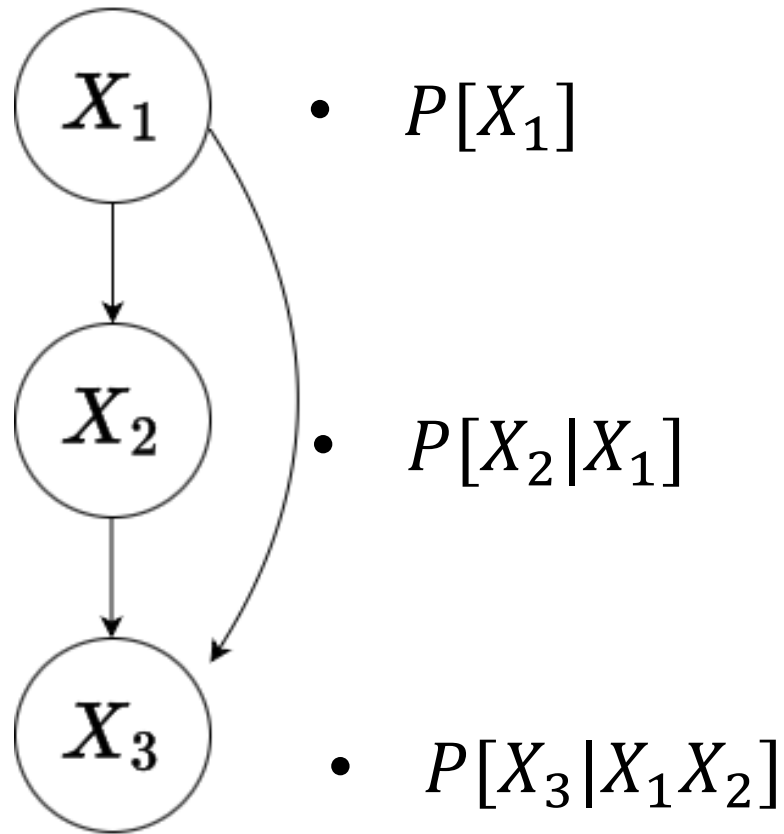
A joint density can be represented in various Bayes nets.
(depending on orders)
But a Bayes net represents a joint probabilistic density.

Suppose we three random variable X_1, X_2, X_3 .

The same joint density $P(X_1, X_2, X_3)$ can be represented by $3 \times 2 \times 1 = 6$ possibles ways but all are the same $P(X_1, X_2, X_3)$.

- $P(X_1) P(X_2 | X_1) P(X_3 | X_1, X_2)$
- $P(X_1) P(X_3 | X_1) P(X_2 | X_1, X_3)$
- ...

A joint density can be represented
by different Bayes net: different DAG and CPT.



In a Bayesian network,
the lack of edges indicates conditional independence in the forward direction.

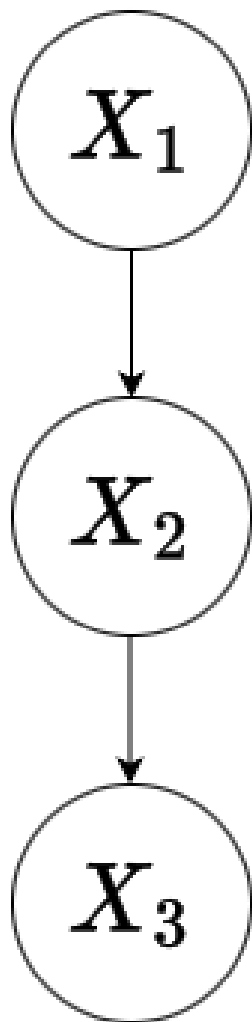
- $P(X_1) P(X_2|X_1) P(X_3|X_2)$
- $[X_3 \perp X_1 \mid X_2]$
: X_3 and X_1 are conditionally independent given X_2 .

- $P(X_1) P(X_3|X_1) P(X_2|X_1)$

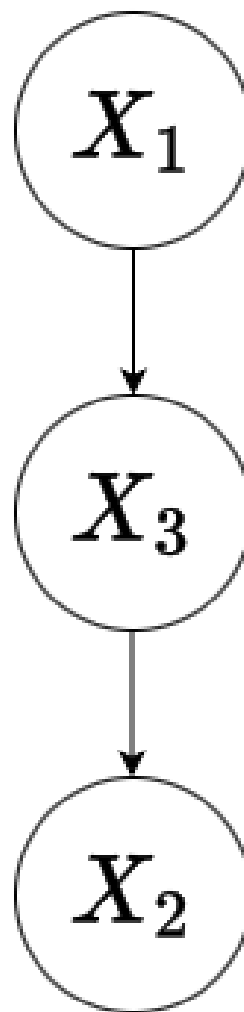
$$\begin{aligned} P(X_1, X_3|X_2) &= \frac{P(X_1, X_2, X_3)}{P(X_2)} \\ &= \frac{P(X_1, X_2, X_3)}{P(X_2)} \\ &= \frac{P(X_1)P(X_2|X_1)P(X_3|X_2)}{P(X_2)} \\ &= \frac{P(X_1, X_2)P(X_3|X_2)}{P(X_2)} = P(X_1|X_2)P(X_3|X_2) \end{aligned}$$

Are they same densities?

- $P(X_1) P(X_2 | X_1) P(X_3 | X_2)$



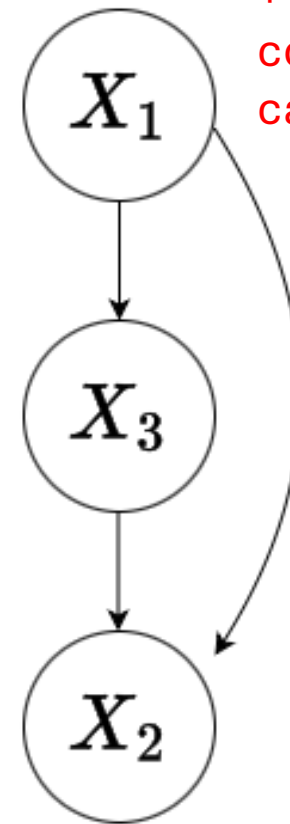
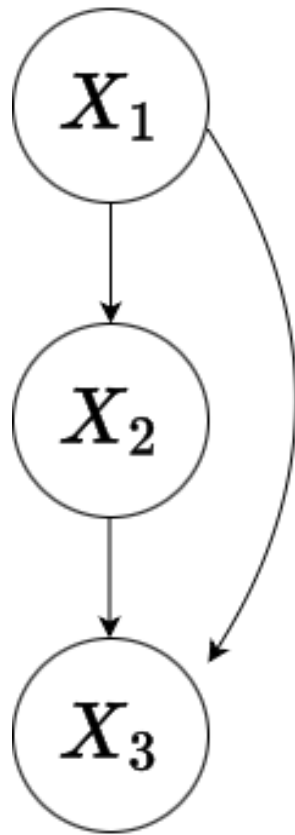
- $P(X_1) P(X_3 | X_1) P(X_2 | X_1)$



Q: The arrows in a Bayesian network has a causal meaning?

+ no, direct arrows between R.Vs indicate statistical correlations among them
but does not guarantee the relationship of causal and effect.

A joint density can be represented by several Bayes Nets.
There will be a contradiction if the arrows necessarily imply causation.
However, sometimes they can encode the relation of causal and effect.



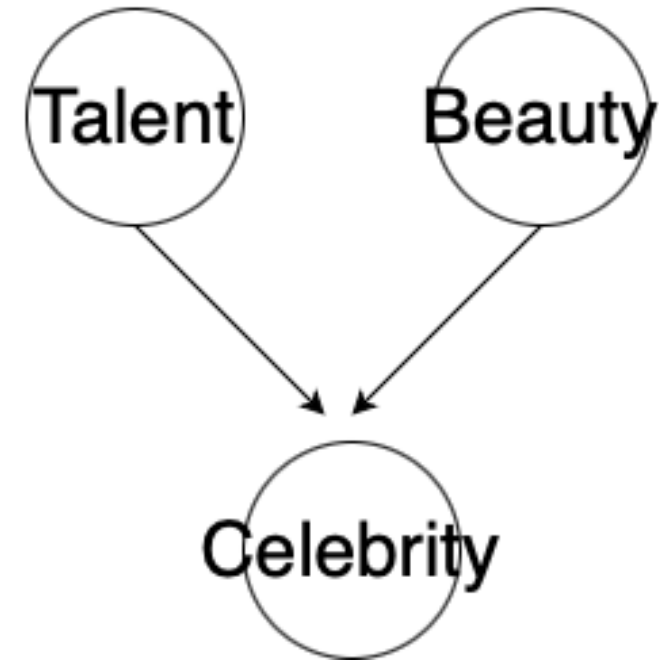
+ when the structure of a Bayes Net is constructed based prior knowledge about causal and effect relationship.

It is true that we can create the structure based on prior knowledge about causal and effect but the arrow no guarantees about causation.

The Examples of Bayes net that encodes Cause and Effect



From the book “The book of Why” by Judea Peral

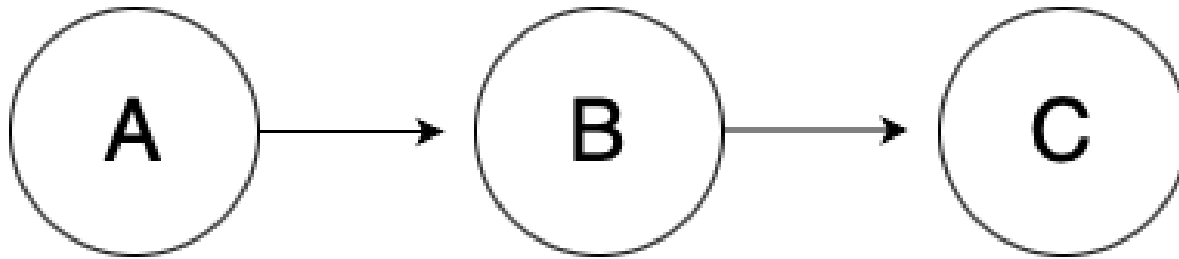


We can design a Bayes net
based on prior knowledge about causal relationship

Elementary Building Blocks for Bayes Net & their Conditional Independence

Three Elementary Building Blocks
with three R.Vs and two links (Junctions)

[1] **Chain Structure**: the middle node of chain breaks into two.



B mediator screen off information about A from C: $A \perp B \mid C$

[1] **Chain Structure**: the middle node of chain breaks into two.



The fire by itself does not set off an alarm.

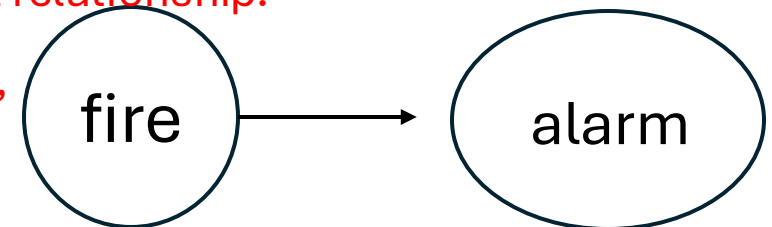
Q: is it possible? fire is a cause for the alarm?

+ we can draw a direct arrow between fire and alarm.

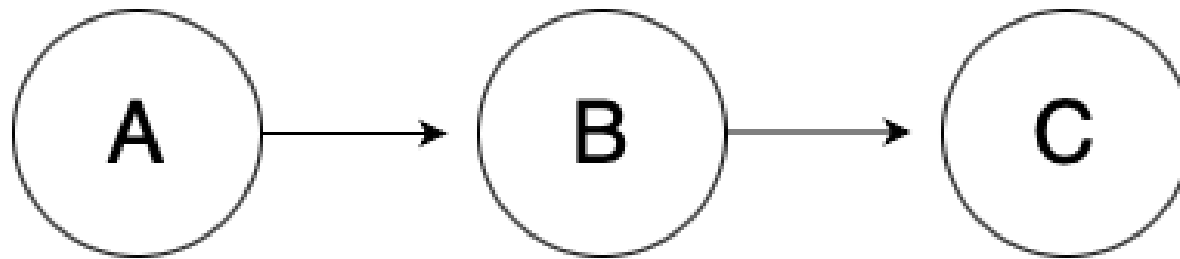
However, the direct arrow does not necessarily imply causal and effect relationship.

Bayesian Network cannot be a ground to say “A causes B”

But, Bayesian Network can be a ground to say “C is not the cause of D.”

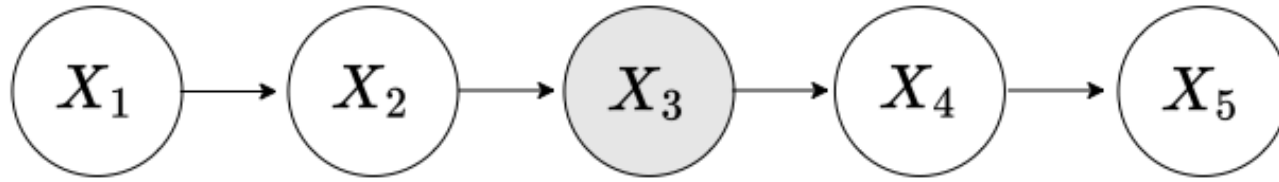


Q: a possible conclusion about causation from the chain structure ?



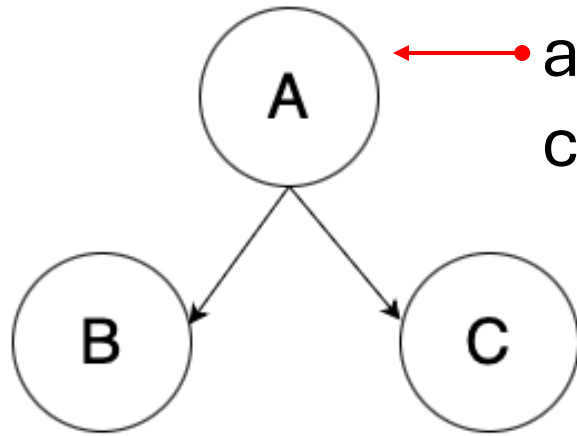
+ A and C is conditionally independent given B \leftrightarrow
“A is not the cause of C.”

Extended Chain Structure: the middle node of chain breaks into two.



$$\begin{aligned} P(X_1, X_5 | X_3) &= \frac{P(X_1, X_3, X_5)}{P(X_3)} \\ &= \frac{\sum_{X_2, X_4} P(X_1, X_2, \dots, X_5)}{P(X_3)} \\ &= \frac{\sum_{X_2, X_4} P(X_1, X_2, X_3) P(X_4, X_5 | X_1, X_2, X_3)}{P(X_3)} \\ &= \sum_{X_2, X_4} P(X_1, X_2 | X_3) P(X_4, X_5 | X_3) \\ &= P(X_1 | X_3) P(X_5 | X_3) \end{aligned}$$

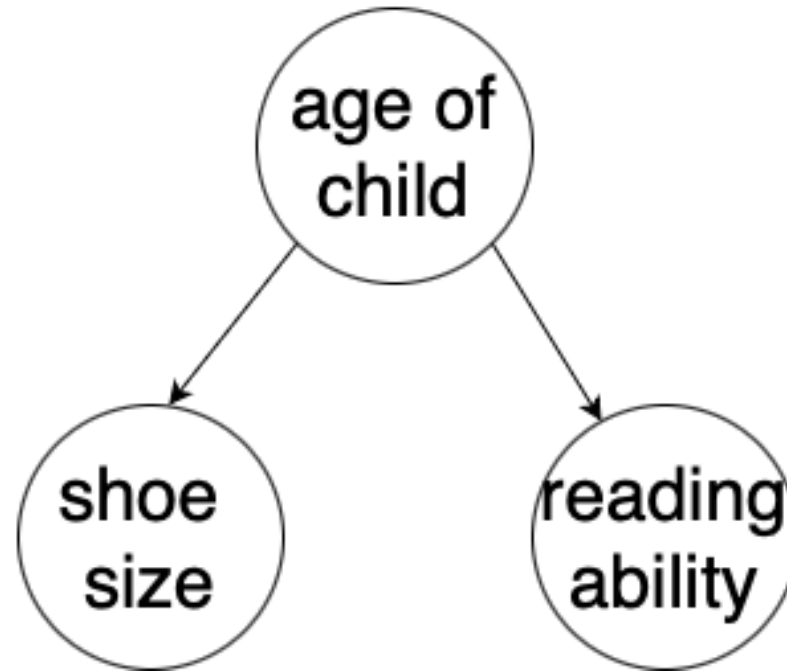
[2] Tent (Fork) Structure: root node separates its children



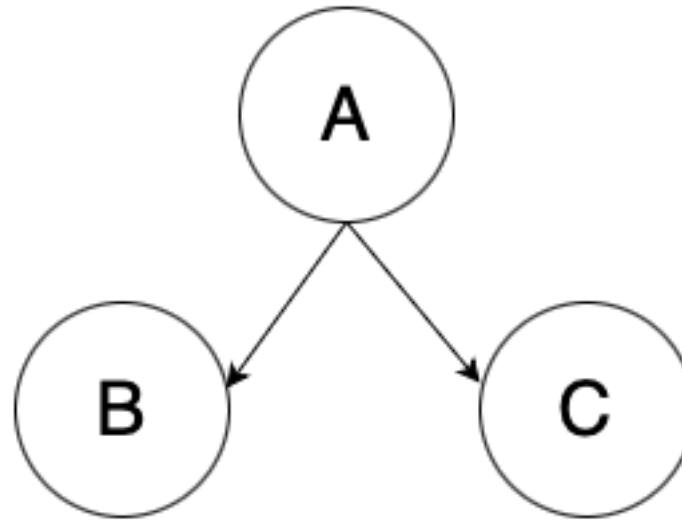
- The cofounder A makes B and C are statistically correlated.
- $B \perp C \mid A$

$$\begin{aligned} P[B, C|A] &= \frac{P[ABC]}{P[A]} \\ &= \frac{P[AB]P[C|A]}{P[A]} \\ &= P[B|A] \cdot P[C|A] \end{aligned}$$

[2] Tent (Fork) Structure: root node separates its children

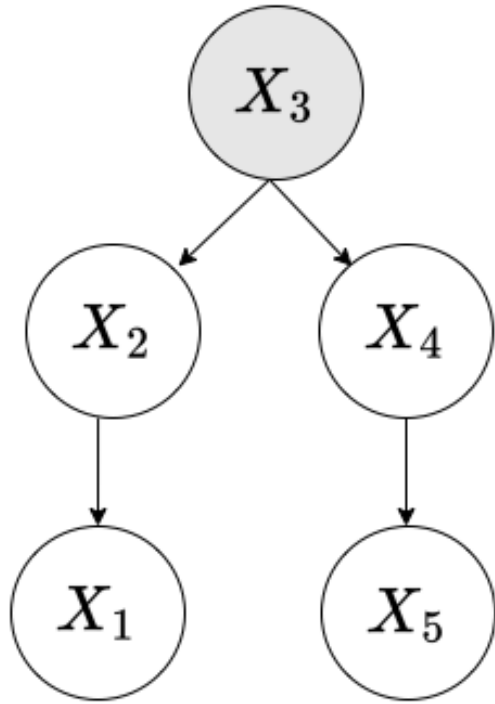


Q: a possible conclusion about causation from the fork structure ?



+ B and C is conditionally independent given A \leftrightarrow
“B is not the cause of C.”

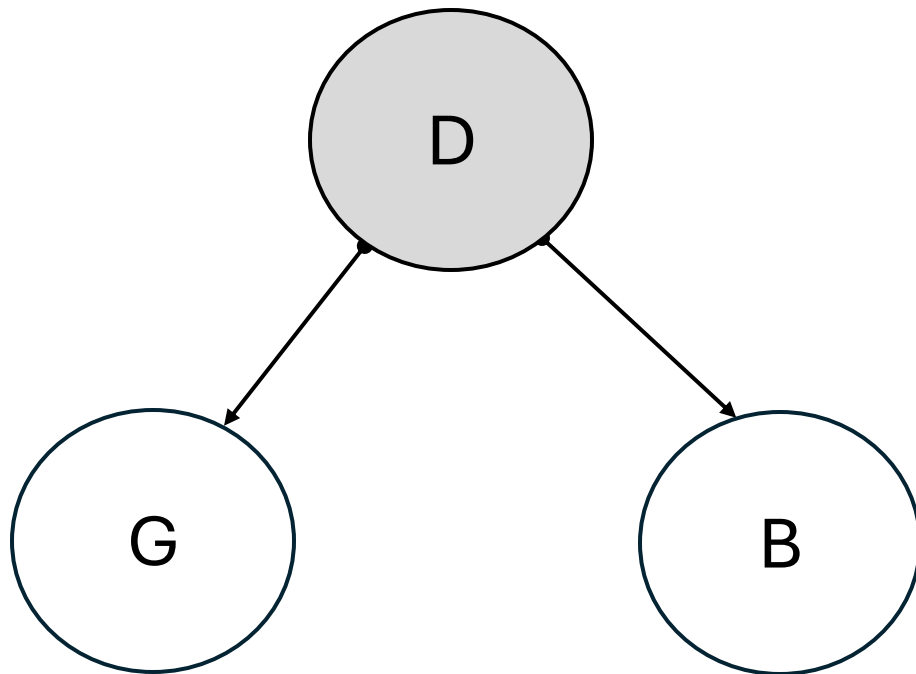
Extended Tent Structure: root node separates its children



$$\begin{aligned} P(X_1, X_5 | X_3) &= \frac{P(X_1, X_3, X_5)}{P(X_3)} \\ &= \frac{\sum_{X_2, X_4} P(X_1, X_2, \dots, X_5)}{P(X_3)} \\ &= \frac{\sum_{X_2, X_4} P(X_1, X_2, X_3) P(X_4, X_5 | X_1, X_2, X_3)}{P(X_3)} \\ &= \sum_{X_2, X_4} P(X_1, X_2 | X_3) P(X_4, X_5 | X_3) \\ &= P(X_1 | X_3) P(X_5 | X_3) \end{aligned}$$

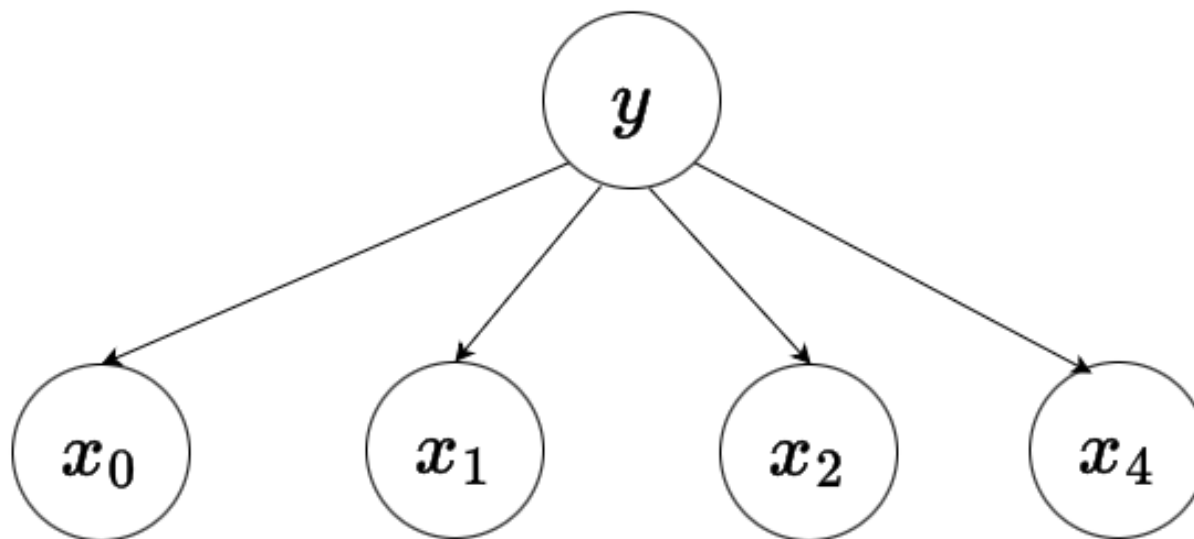
[Example of Tent Structure]

In the hw1, the feature of Glucose (G) and blood pressure (B) were conditionally independent given Diabetes (D). The conditional independence was originated from the assumption of causal relationship between D and the feature of G and B.



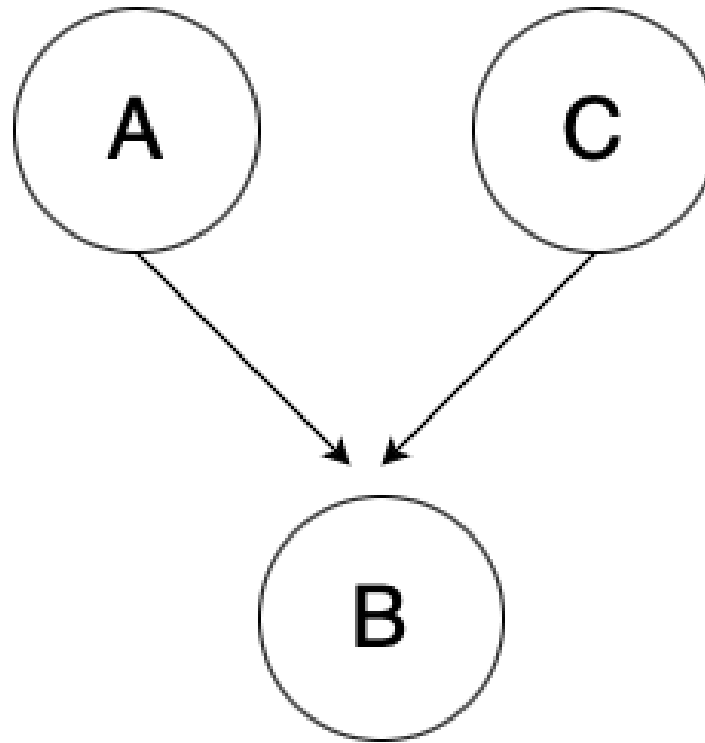
- When D information observed, G and B becomes independent.
$$P(G, B|D) = P(G|D) P(B|D)$$

Conditional Independence Assumption for Naïve Bayes



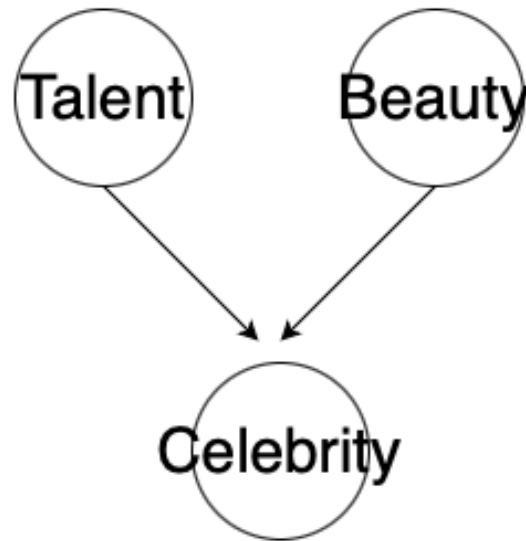
[3] V-Structure (collider):

conditioning on a common child/ (or a descendant of the child) or at the bottom of a v-structure makes its parents become dependent.



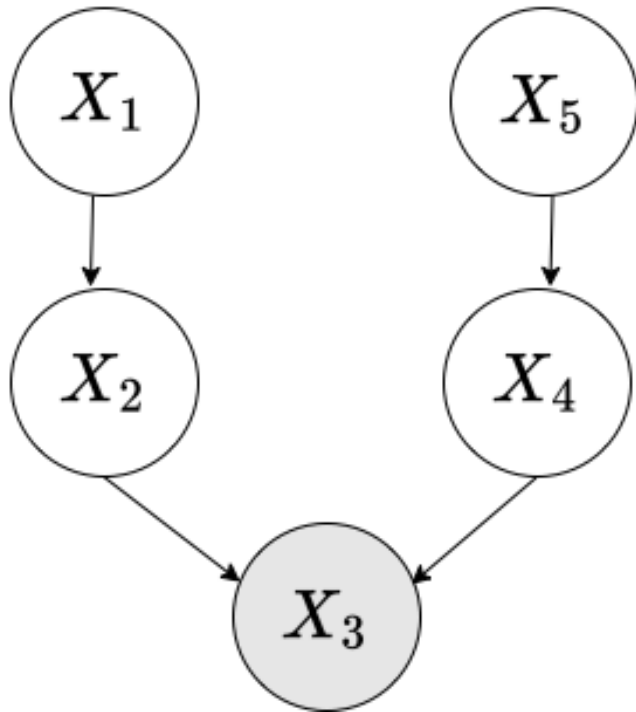
[3] V-Structure (collider):

conditioning on a common child/ (or a descendant of the child) or at the bottom of a v-structure makes its parents become dependent.



Both talent and beauty contribute to an actor's success, but beauty and talent are completely uncorrelated. If a celebrity is a particularly good actor that explains away the person's success, the person does not need to be more beautiful than the average person.

[3] V-Structure: conditioning on a common child at the bottom of a v-structure makes its parents become dependent.



$$\begin{aligned} P(X_1, X_5 | X_3) &= \frac{P(X_1, X_3, X_5)}{P(X_3)} \\ &= \frac{\sum_{X_2, X_4} P(X_1, X_2, \dots, X_5)}{P(X_3)} \\ &= \frac{\sum_{X_2, X_4} P(X_1, X_2) P(X_4, X_5) P(X_3 | X_2, X_4)}{P(X_3)} \\ &\neq P(X_1 | X_3) P(X_5 | X_3) \end{aligned}$$

[D separation for a path]

- An undirected path P is **d-separated** by a set of observation nodes E iff at least one of the conditions hold.
 - (1) it contains a **chain structure** but the middle node is observed.
 - (2) it contains a **tent structure** but the co-founder is observed.
 - (3) it contains a **V-structure** but the common child/descendent is not observed.

[D separation between two nodes]

The node A is **d-separated** from B by a set of observation E
 \Leftrightarrow each (all) undirected path from A to B is d-separated.

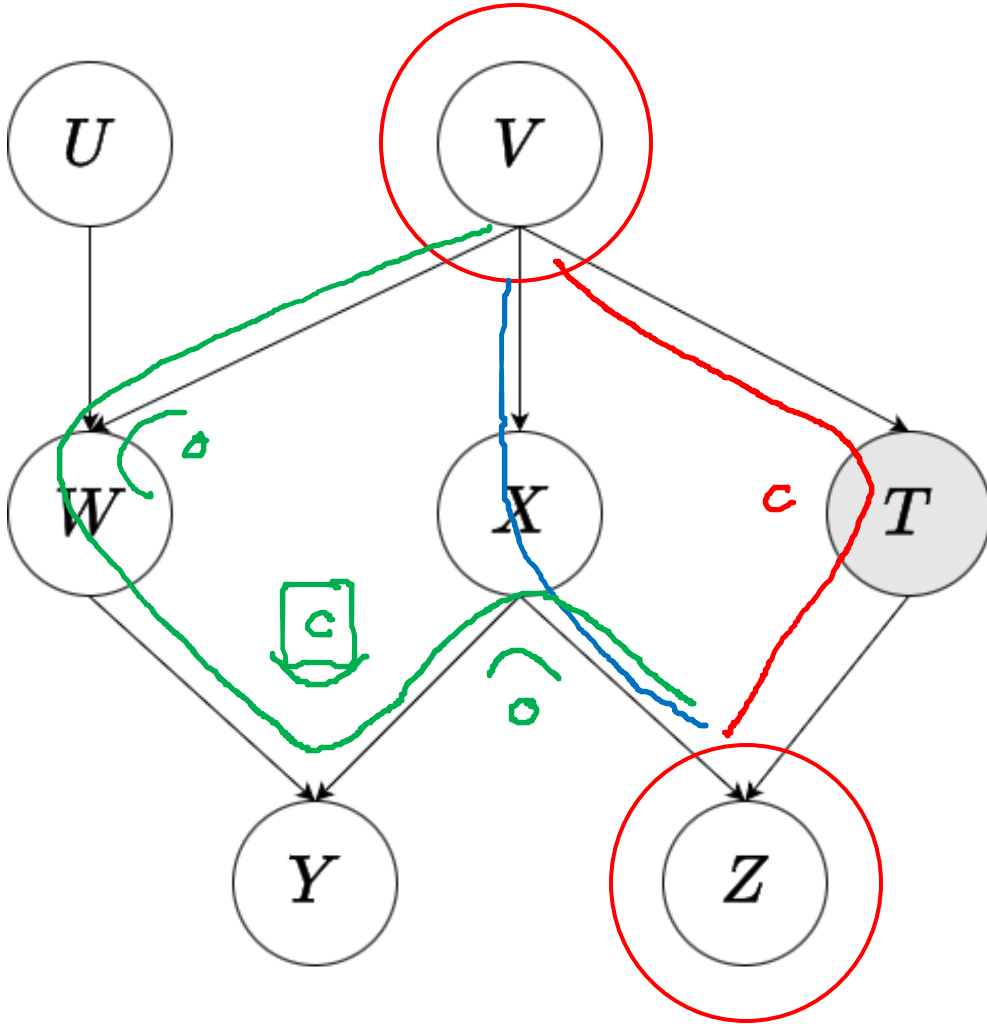
$$\Leftrightarrow [A \perp B | E]$$

Q: What if we find a path between A and B that is not d-separated?

+ then A and B is not d-separated

In order to be d-separated, all undirected paths from A to B must be d-separated.

Ex1: D-Separation)

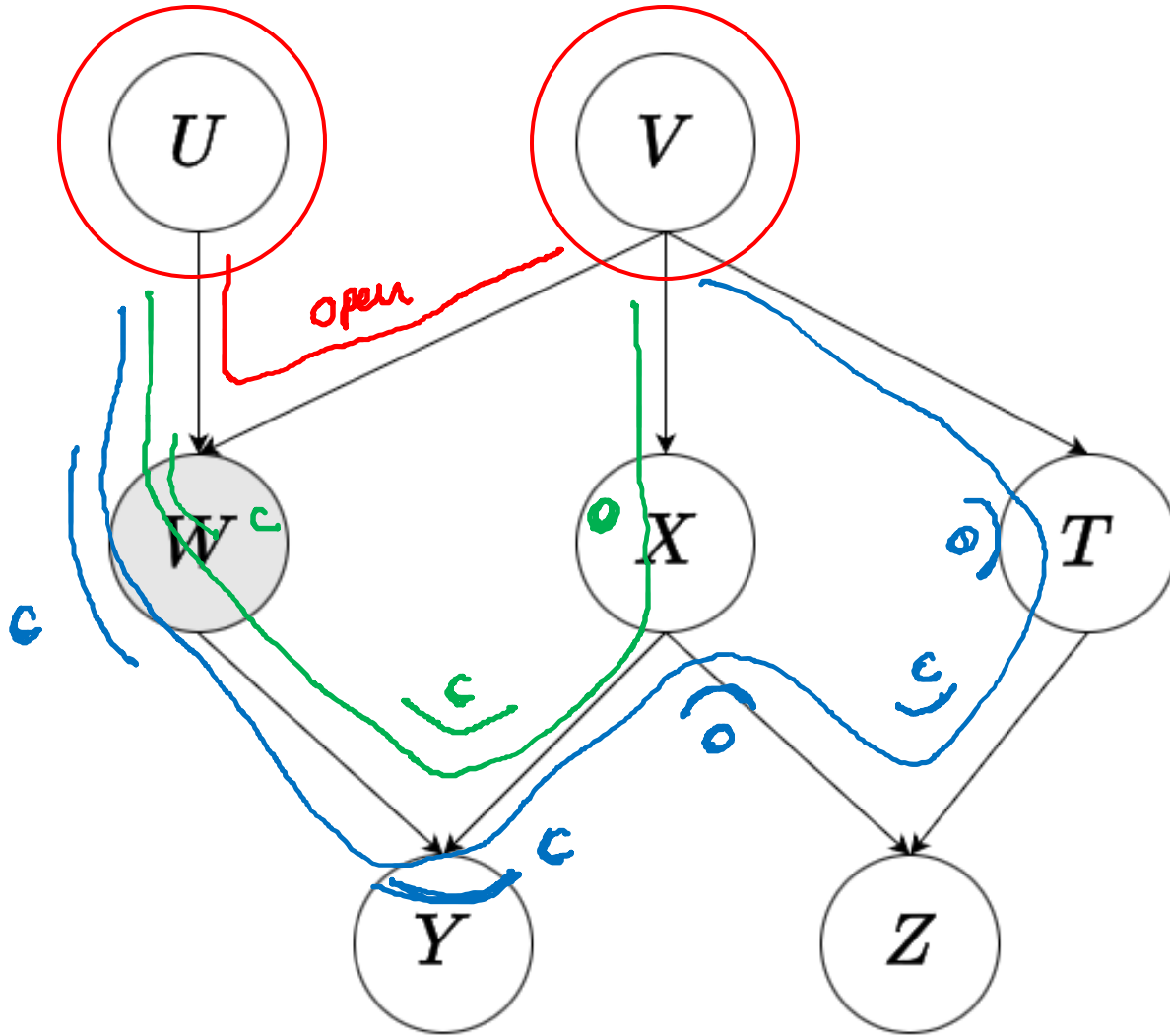


Q: $[V \perp Z | T]$? Not D-separated

$\text{—} : \text{closed}$
 $\text{—} : \text{closed}$
 $\text{—} : \text{open}$

} open

Ex2: D-Separation)

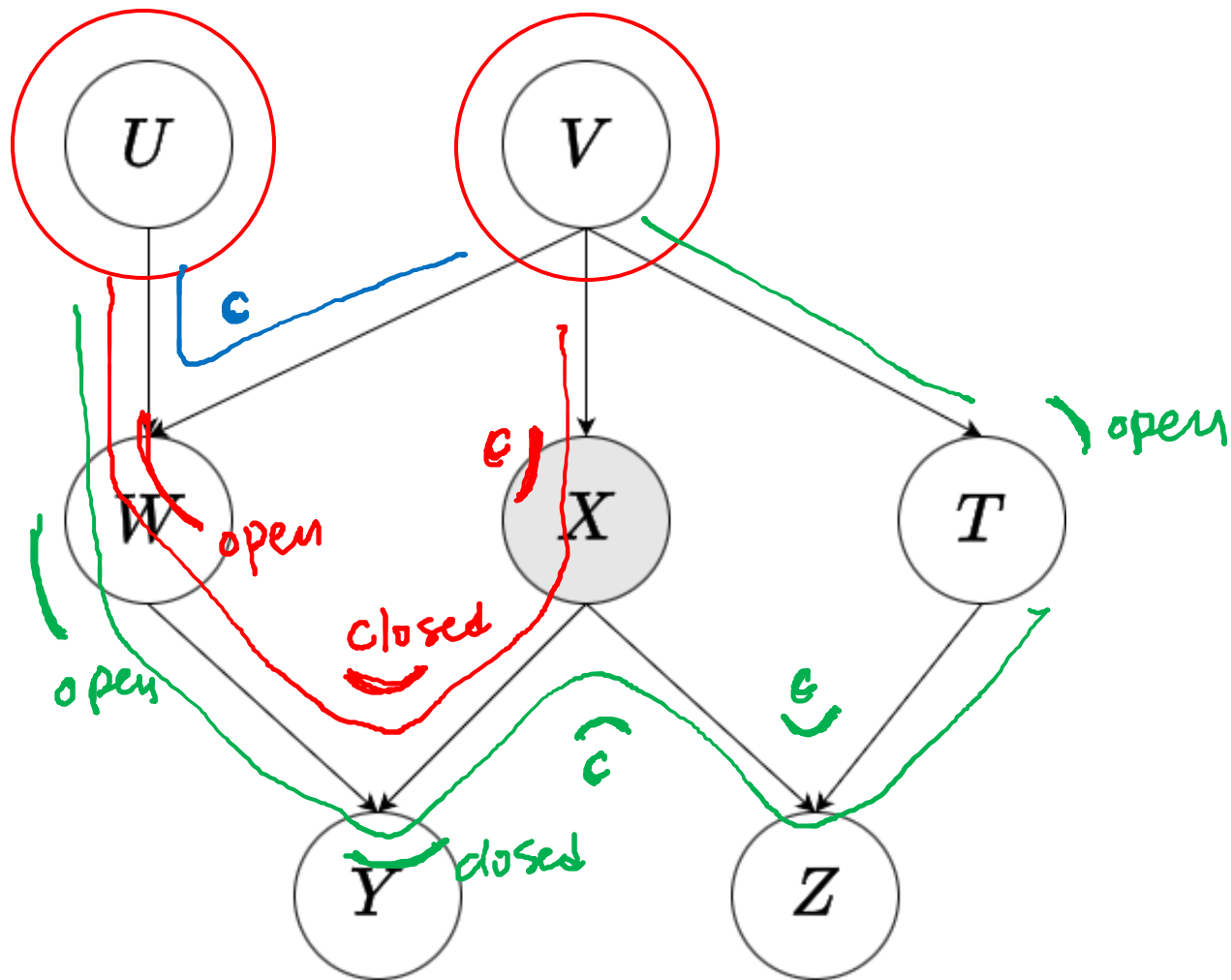


Q: $[U \perp V | W]$? *Not D-separated*

$\text{---} : \text{closed}$
 $\text{---} : \text{closed}$
 $\text{---} : \text{open}$

open

Ex3: D-Separation)

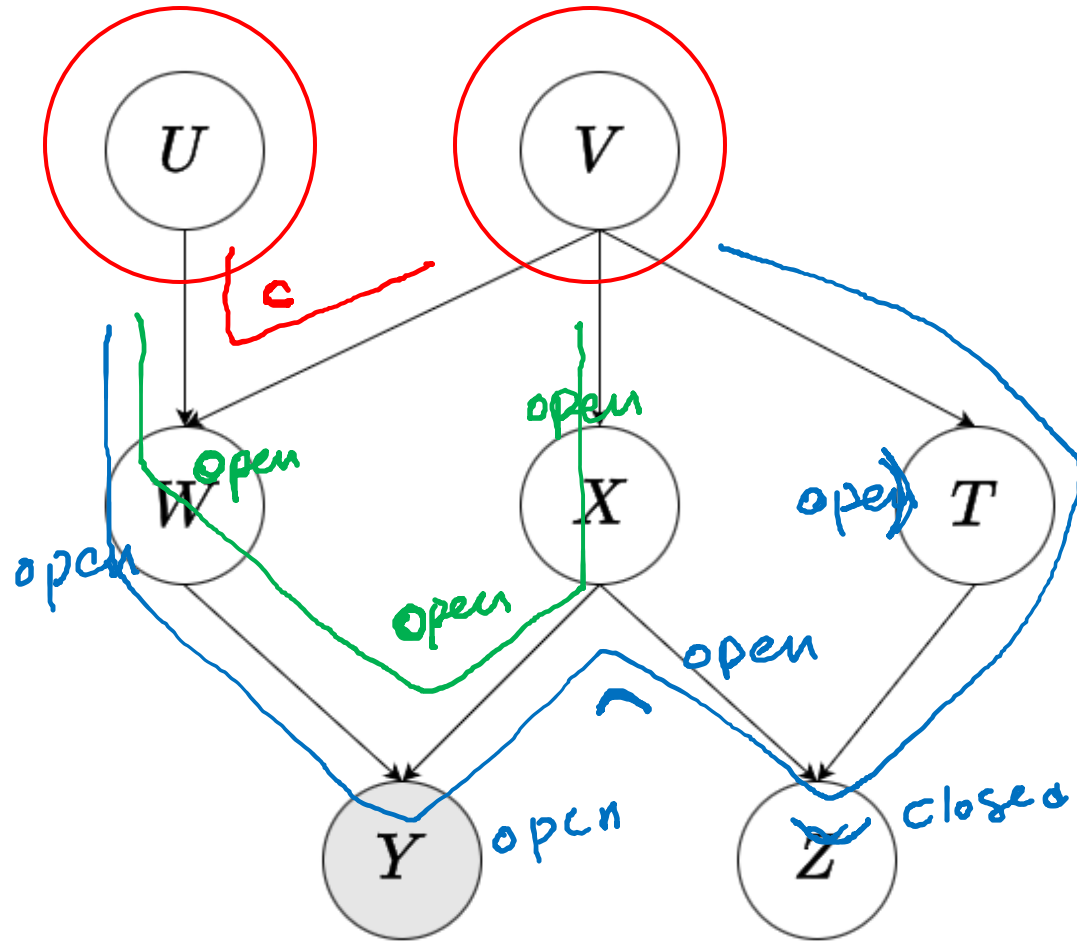


$Q: [U \perp V | X]?$ *D-separated*

— : closed
 — : closed
 — : closed

} closed

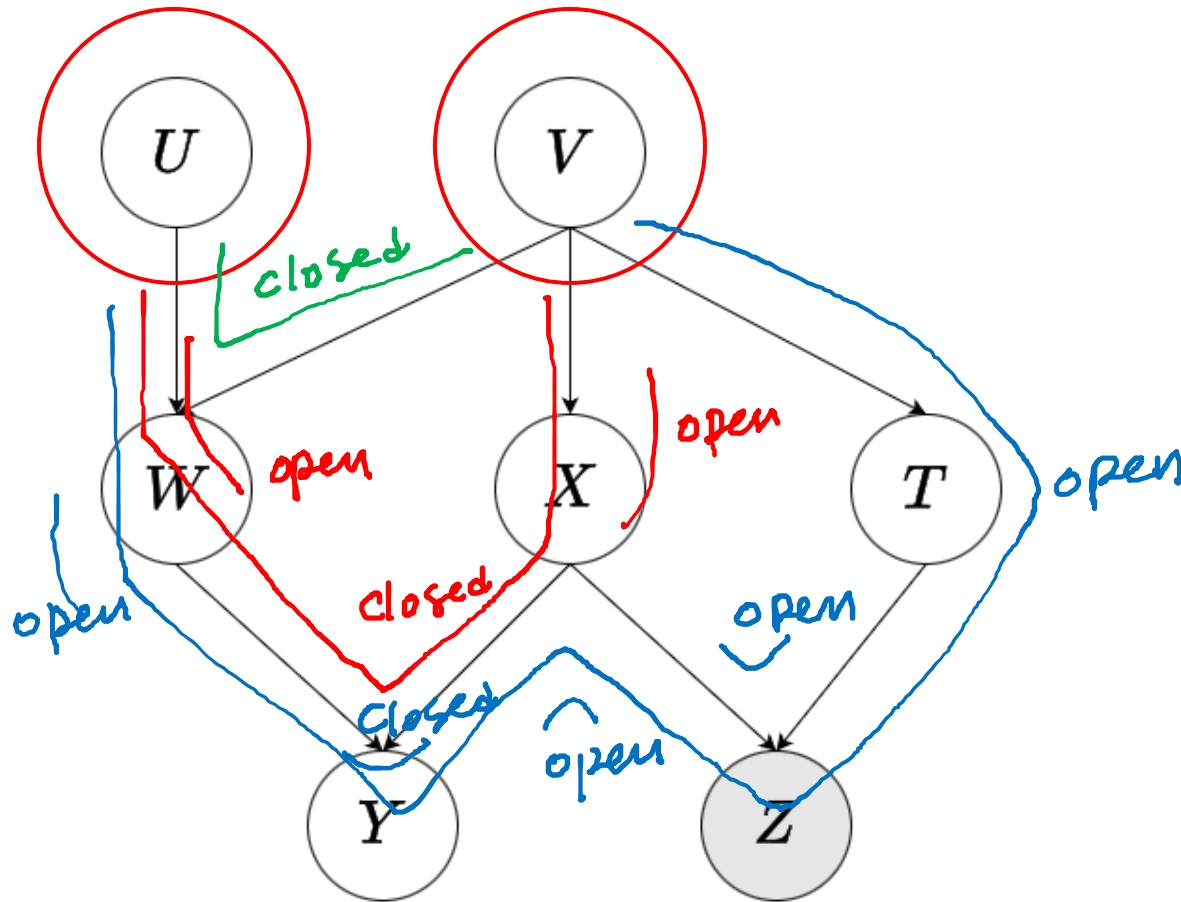
Ex4: D-Separation)



Q: $[U \perp V | Y]$? Not D-separated

— closed
— open
— closed } open

Ex5: D-Separation)



Q: $[U \perp V | Z]$? *D-separated*

— : closed
— : closed
— : closed } *closed*

How D-separation matter?

How $A \perp B \mid C$ matter? + A and B are not in a causal and effect relationship.

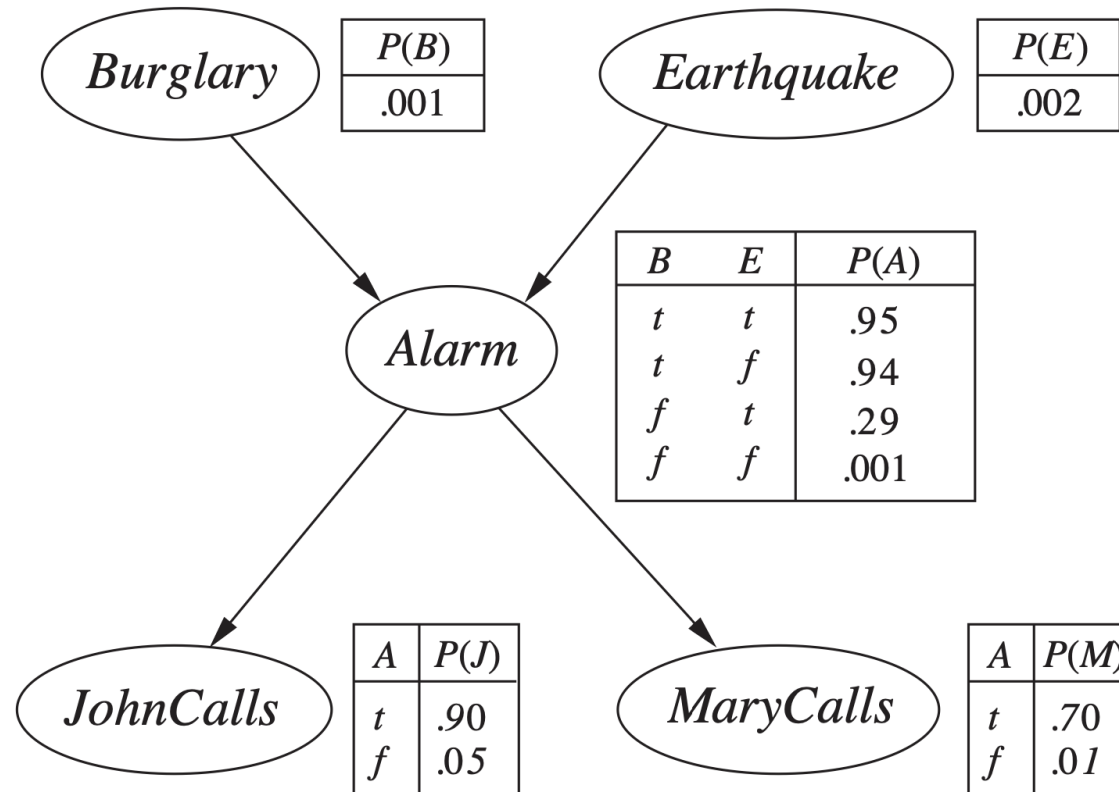
What conclusion can be drive from it about causal relationship ?

In the previous slide,
Depending on the order of variable introduction,
the same density can be represented by various ways.

This Bayes Net encodes the joint density:

$P[\text{Burglary}] P[\text{Earthquake}] P[\text{Alarm} | \text{Burglary}, \text{Earthquake}]$

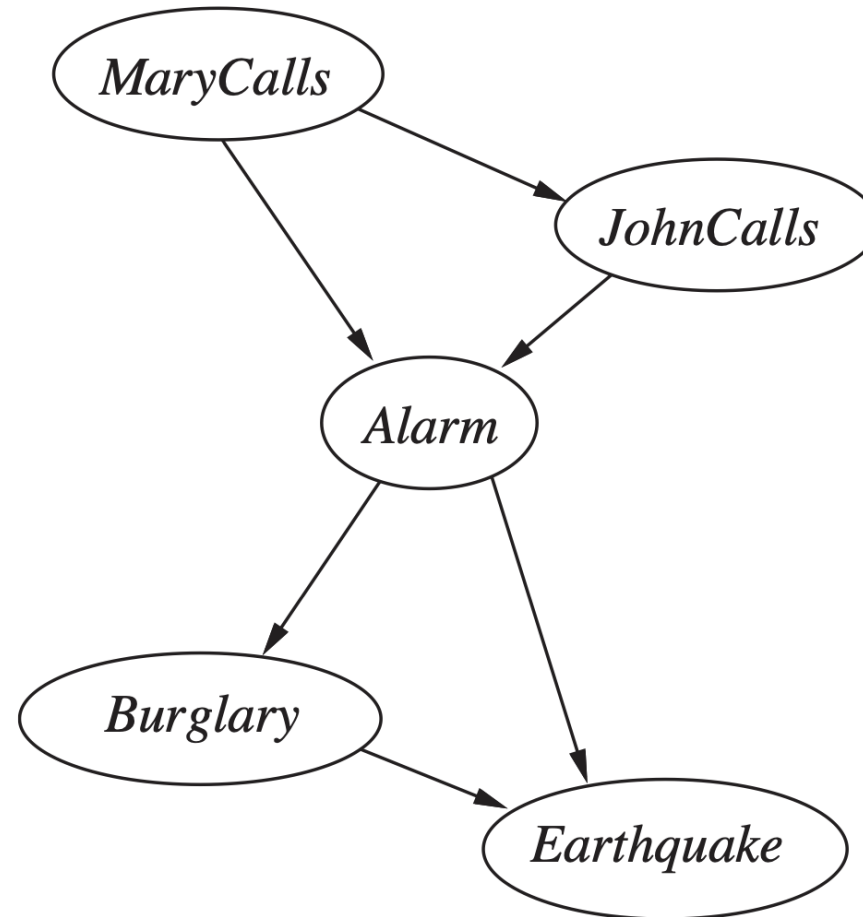
$P[\text{JohnCalls} | \text{Alarm}] P[\text{MaryCalls} | \text{Alarm}]$



How the structure will be changed as the order of
 $\text{MaryCalls} \rightarrow \text{JohnCalls} \rightarrow \text{Alarm} \rightarrow \text{Burglary} \rightarrow \text{Earthquake}$?

Bayesian network based on new order

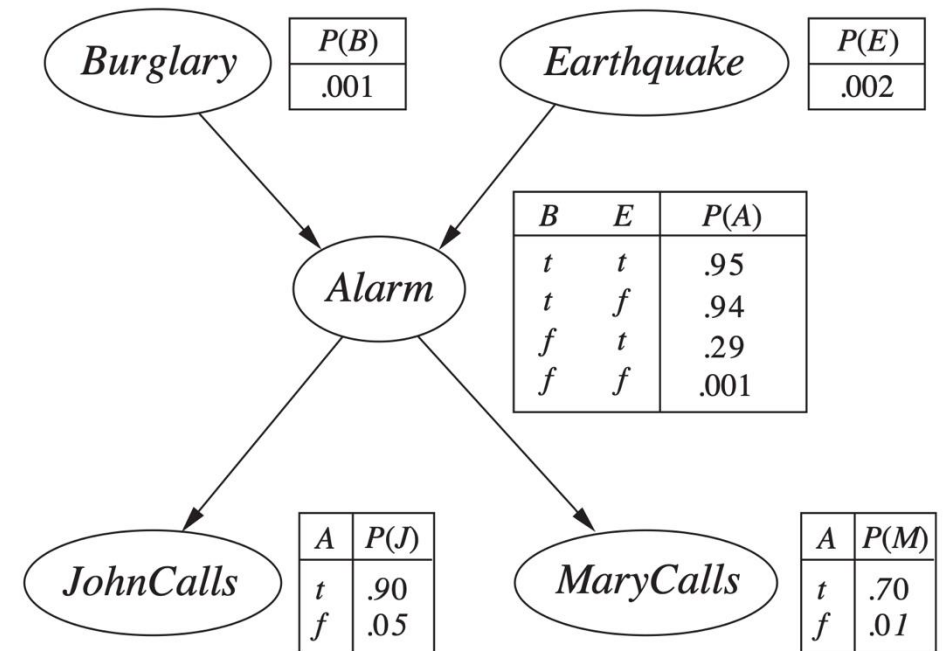
: $MaryCalls \rightarrow JohnCalls \rightarrow Alarm \rightarrow Burglary \rightarrow Earthquake$?



The same density information can be represented by the new order of MaryCalls \rightarrow JohnCalls \rightarrow Alarm \rightarrow Burglary \rightarrow Earthquake

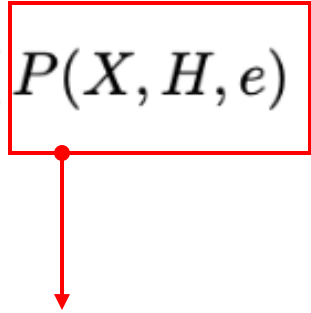
For the new structure, the following conditional independence are derived from the original structure.

- Burglary \perp MaryCalls and JohnCalls | Alarm:
- Earthquake \perp MaryCalls, JohnCalls | Alarm
- Earthquake $\not\perp$ Burglary | Alarm



Inference: Querying Posterior Probability
how to compute $P[\text{Burglary} \mid \text{MaryCalls and JohnCalls}]?$

The general format of query

$$P(X|e) = \frac{P(X, e)}{P(e)} = \frac{\sum_H P(X, H, e)}{P(e)} = \alpha \sum_H P(X, H, e)$$


- X query variable
- H non-query variable
- E evidence variable

Bayes Net provides the joint density information,
so any type of query can be answered!

Given $P(X_1, X_2, X_3)$

$$P(X_1 | X_2 = x_2)? = \frac{P(X_1, X_2 = x_2)}{P(X_2 = x_2)} = \alpha \sum_{X_3} P(X_1, X_2 = x_2, X_3)?$$

Variable Elimination Method (Exact Inference)

In the last class,
we compared the computational complexities for the two cases:

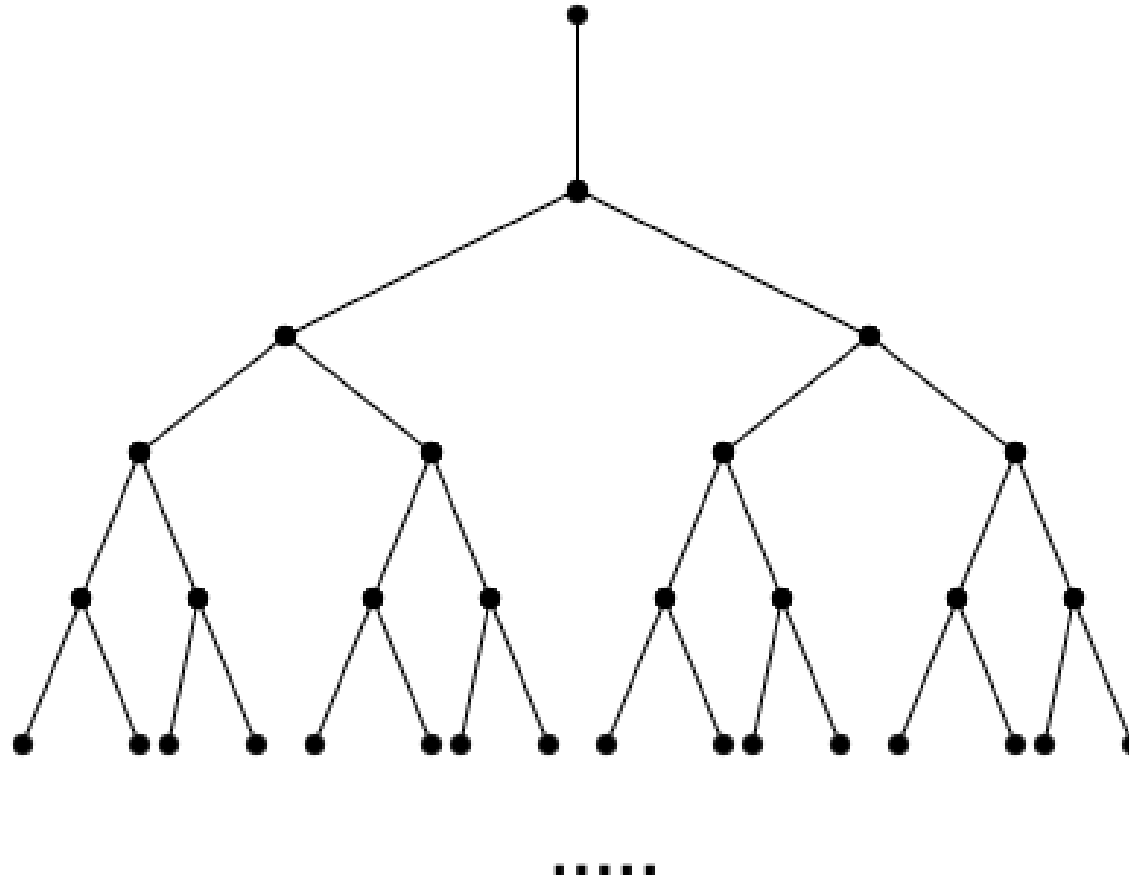
- without CI

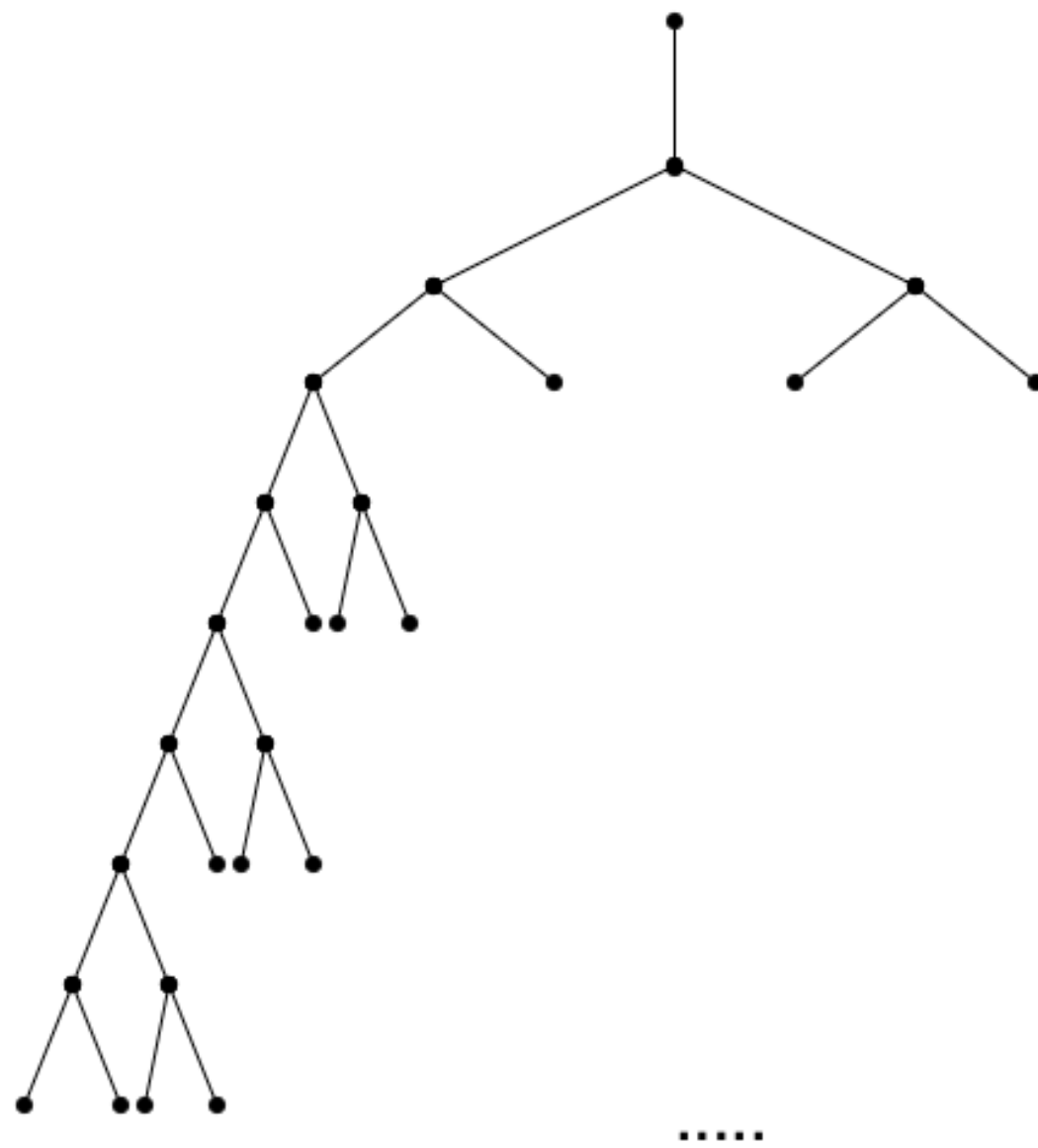
$$P(X_{36}) = \sum_{X_1} \sum_{X_2} \dots \sum_{X_{35}} \sum_{X_{37}} \dots \sum_{X_{1000}} P(X_1, X_2, \dots, X_{35}, X_{36}, X_{37}, \dots, X_{1000})$$

- with CI

$$\begin{aligned} P(X_{36}) &= \sum_{X_1} \sum_{X_2} \dots \sum_{X_{35}} \sum_{X_{37}} \dots \sum_{X_{1000}} P(X_1)P(X_2|X_1), \dots, P(X_{36}|X_{35})P(X_{37}|X_{36}), \dots, P(X_{1000}|X_{999}) \\ &= \sum_{X_1} P(X_1) \sum_{X_2} P(X_2|X_1) \dots \sum_{X_{35}} P(X_{35}|X_{34})P(X_{36}|X_{35}) \sum_{X_{37}} P(X_{37}|X_{36}) \dots \sum_{X_{1000}} P(X_{1000}|X_{999}) \end{aligned}$$

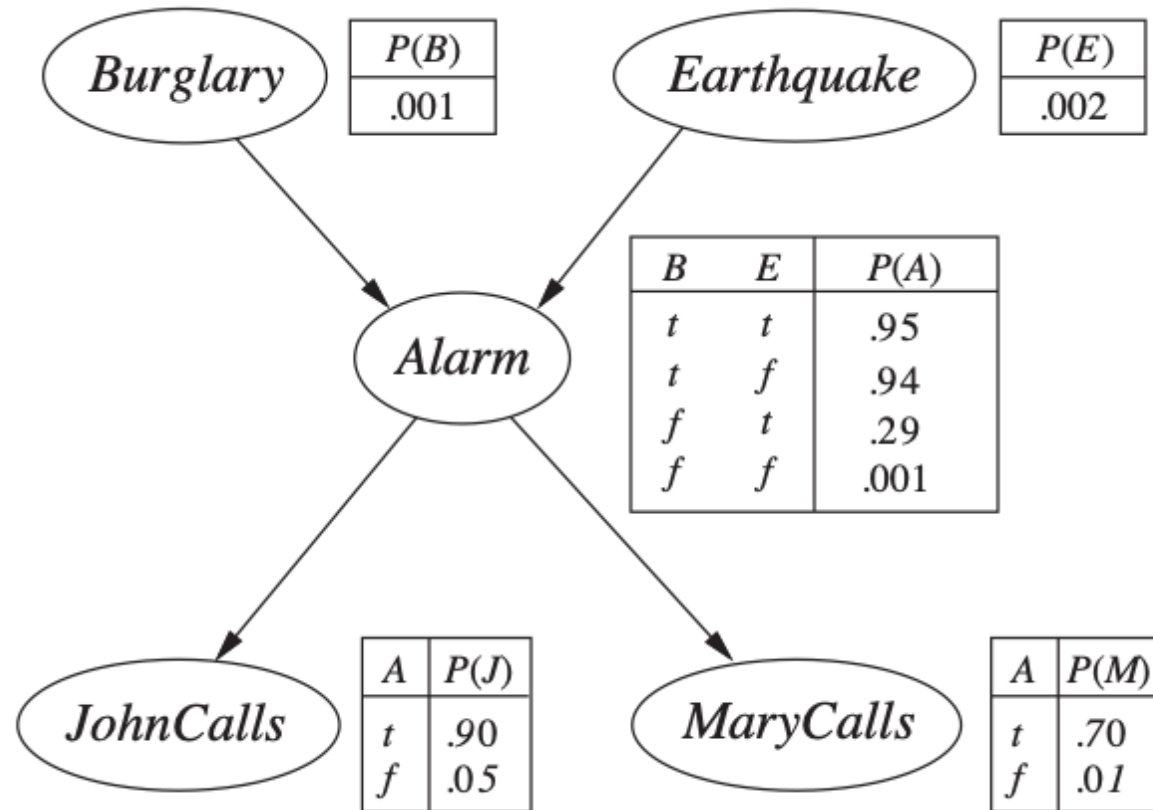
- without CI: $O(K^N)$ operations
- with CI (Markov Chain Structure): $O(2^N)$





Q: $P[B = + \text{ and } M = +]$?

How can we compute this?



$$P[B|J = +, M = +] = \frac{P[B, J = +, M = +]}{P[J = +, M = +]} = \alpha \cdot P[B, J = +, M = +]$$

$$\alpha \cdot P[B, J = +, M = +] = \alpha \cdot \left[\frac{P[B = +, J = +, M = +]}{P[B = -, J = +, M = +]} \right]$$

we need to compute this $\alpha P[B, J = +, M = +]$.

$$\alpha (P[B = +, J = +, M = +] + P[B = -, J = +, M = +]) = 1$$

How to compute


$$P[B, J = +, M = +] = \left| \frac{P[B = +, J = +, M = +]}{P[B = -, J = +, M = +]} \right|?$$

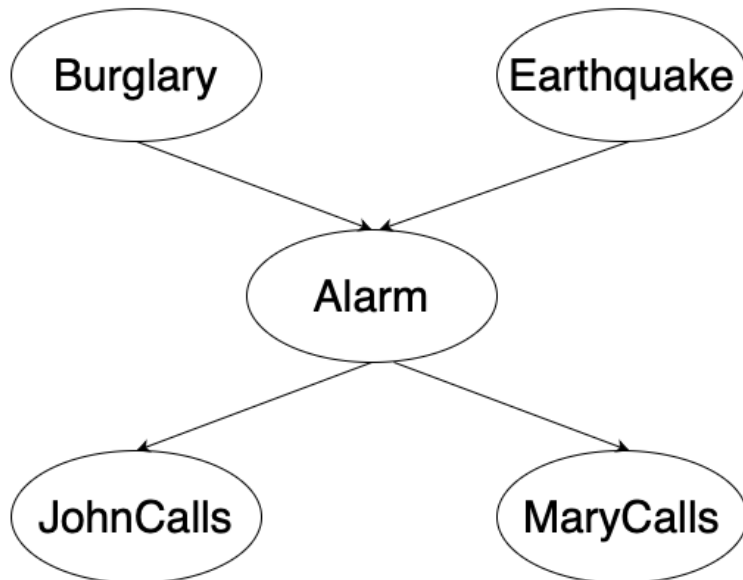
By marginalization,

$$P[B, J = +, M = +] = \sum_E \sum_A P[B, E, A, J = +, M = +]$$

By using the CI structure defined in Bayes Net

$$\begin{aligned}
 P[B, J = +, M = +] &= \sum_E \sum_A P[B, E, A, J = +, M = +] \\
 &= \underbrace{P[B]}_{f_1(B)} \sum_E \underbrace{P[E]}_{f_2(E)} \sum_A \underbrace{P[A|B, E]}_{f_3(A, B, E)} \underbrace{P[J = +|A]}_{f_4(A)} \underbrace{P[M = +|A]}_{f_5(A)}
 \end{aligned}$$





We need to compute from right to left direction.

By Conditional Independence defined in Bayes Net

$$\begin{aligned}
 P[B, J = +, M = +] &= \sum_E \sum_A P[B, E, A, J = +, M = +] \\
 &= \underbrace{P[B]}_{f_1(B)} \sum_E \underbrace{P[E]}_{f_2(E)} \sum_A \underbrace{P[A|B, E]}_{f_3(A, B, E)} \underbrace{P[J = +|A]}_{f_4(A)} \underbrace{P[M = +|A]}_{f_5(A)} \\
 &\hspace{15em} \underbrace{\hspace{10em}}_{f_6(A)}
 \end{aligned}$$

$$f_6[A] = \begin{bmatrix} f_6[A = +] \\ f_6[A = -] \end{bmatrix} = \begin{bmatrix} P[J = +|A = +] \\ P[J = +|A = -] \end{bmatrix} \begin{bmatrix} P[M = +|A = +] \\ P[M = +|A = -] \end{bmatrix}$$

By Conditional Independence defined in Bayes Net

$$\begin{aligned}
 P[B, J = +, M = +] &= \sum_E \sum_A P[B, E, A, J = +, M = +] \\
 &= \underbrace{P[B]}_{f_1(B)} \sum_E \underbrace{P[E]}_{f_2(E)} \sum_A \underbrace{P[A|B, E]}_{f_3(A, B, E)} \underbrace{P[J = +|A]}_{f_4(A)} \underbrace{P[M = +|A]}_{f_5(A)} \\
 &\quad \underbrace{\hspace{10em}}_{f_6(A)} \\
 &\quad \underbrace{\hspace{10em}}_{f_7(BE)} = \sum_A f_3(A, B, E) f_6(A)
 \end{aligned}$$

$$\begin{aligned}
 f_7(BE) &= \sum_A P[A|BE] f_6[A] \\
 &= f_6(A+) \begin{bmatrix} P[A+|B+E+] & P[A+|B+E-] \\ P[A+|B-E+] & P[A+|B-E-] \end{bmatrix} + f_6(A-) \begin{bmatrix} P[A-|B+E+] & P[A-|B+E-] \\ P[A-|B-E+] & P[A-|B-E-] \end{bmatrix}
 \end{aligned}$$

By Conditional Independence defined in Bayes Net

$$\begin{aligned}
 P[B, J = +, M = +] &= \sum_E \sum_A P[B, E, A, J = +, M = +] \\
 &= \underbrace{P[B]}_{f_1(B)} \underbrace{\sum_E P[E]}_{f_2(E)} \underbrace{\sum_A P[A|B, E] P[J = +|A] P[M = +|A]}_{f_7(BE)} \\
 &\quad \underbrace{P[A|B, E]}_{f_3(A, B, E)} \underbrace{P[J = +|A]}_{f_4(A)} \underbrace{P[M = +|A]}_{f_5(A)}
 \end{aligned}$$

$f_8(B) = \sum_E f_2(E) f_7(BE)$

$$f_8(B) = P(E+)f_7(B, E+) + P(E-)f_7(B, E-)$$

By Conditional Independence defined in Bayes Net

$$\begin{aligned}
 P[B, J = +, M = +] &= \sum_E \sum_A P[B, E, A, J = +, M = +] \\
 &= \underbrace{P[B]}_{f_1(B)} \sum_E \underbrace{P[E]}_{f_2(E)} \sum_A \underbrace{P[A|B, E]}_{f_3(A, B, E)} \underbrace{P[J = +|A]}_{f_4(A)} \underbrace{P[M = +|A]}_{f_5(A)} \\
 &\quad \underbrace{\hspace{10em}}_{f_8(B) = \sum_E f_2(E) f_7(BE)} \\
 &\quad \underbrace{\hspace{10em}}_{f_9(B) = P[B] f_8(B)}
 \end{aligned}$$

$$\textcircled{a} \quad P(B | J=j, M=m)$$

$$= \alpha \cdot P(B, J=j, M=m) = \alpha \cdot \sum_e \sum_a P(B, E, A, J=j, M=m)$$

$$= \alpha \cdot P(B) \cdot \sum_e P(E) \cdot \sum_a P(A|E, B) \cdot P(J=j|A) \cdot P(M=m|A)$$

By factors

$$= \alpha \cdot f_1(B) \cdot \sum_e f_2(E) \cdot \sum_a f_3(A, E, B) \cdot f_4(A) \cdot f_5(A)$$

$$\textcircled{*} f_6(E, B) = f_3(a \in B) f_4(a) f_5(a)$$

$$\textcircled{1} + f_3(\neg a \in B) f_4(\neg a) f_5(\neg a)$$

$$= \begin{bmatrix} 0.95 & 0.94 \\ 0.29 & 0.001 \end{bmatrix} \times 0.9 \times 0.7 + \begin{bmatrix} 1-0.95 & 1-0.94 \\ 1-0.29 & 1-0.001 \end{bmatrix} 0.05 \times 0.01$$

$$= \begin{bmatrix} 0.5985 & 0.5922 \\ 0.1831 & 0.0011 \end{bmatrix}$$

(11) counts operation

$$② \quad (*) f_7(B) = f_2(e) \cdot f_6(e, B) + f_2(\sim e) \cdot f_6(\sim e, B)$$

$$= 0.002 \begin{bmatrix} 0.5985 \\ 0.1831 \end{bmatrix} + [1-0.002] \begin{bmatrix} 0.5922 \\ 0.0011 \end{bmatrix} = \begin{bmatrix} 0.5922 \\ 0.0015 \end{bmatrix} \quad (5)$$

$$③ \quad (*) f_8(B) = f_1(B) + f_7(B)$$

$$\begin{bmatrix} 0.001 \\ 1-0.001 \end{bmatrix} \begin{bmatrix} 0.5922 \\ 0.0015 \end{bmatrix} = \begin{bmatrix} 0.000592 \\ 0.0015 \end{bmatrix} \quad (2)$$

$$= \alpha \begin{bmatrix} 0.000592 \\ 0.0015 \end{bmatrix}$$

By Normalization

$$= \begin{bmatrix} 0.28... \\ 0.71... \end{bmatrix}$$

$$\alpha = 416.1905$$

Normalization

counts : 4.