Faculty of SET / School of Computer and Mathematical Sciences

COMP SCI 3007/7059/7659
Artificial Intelligence
Bayesian Network

Acknowledgement of Country

We acknowledge and pay our respects to the Kaurna people, the traditional custodians whose ancestral lands we gather on.

We acknowledge the deep feelings of attachment and relationship of the Kaurna people to the country and we respect and value their past, present and ongoing connection to the land and cultural beliefs.

Bayesian Network

AIMA C13.4 – 14.3

Bayesian Network

- One kind of Probabilistic Graphic Models

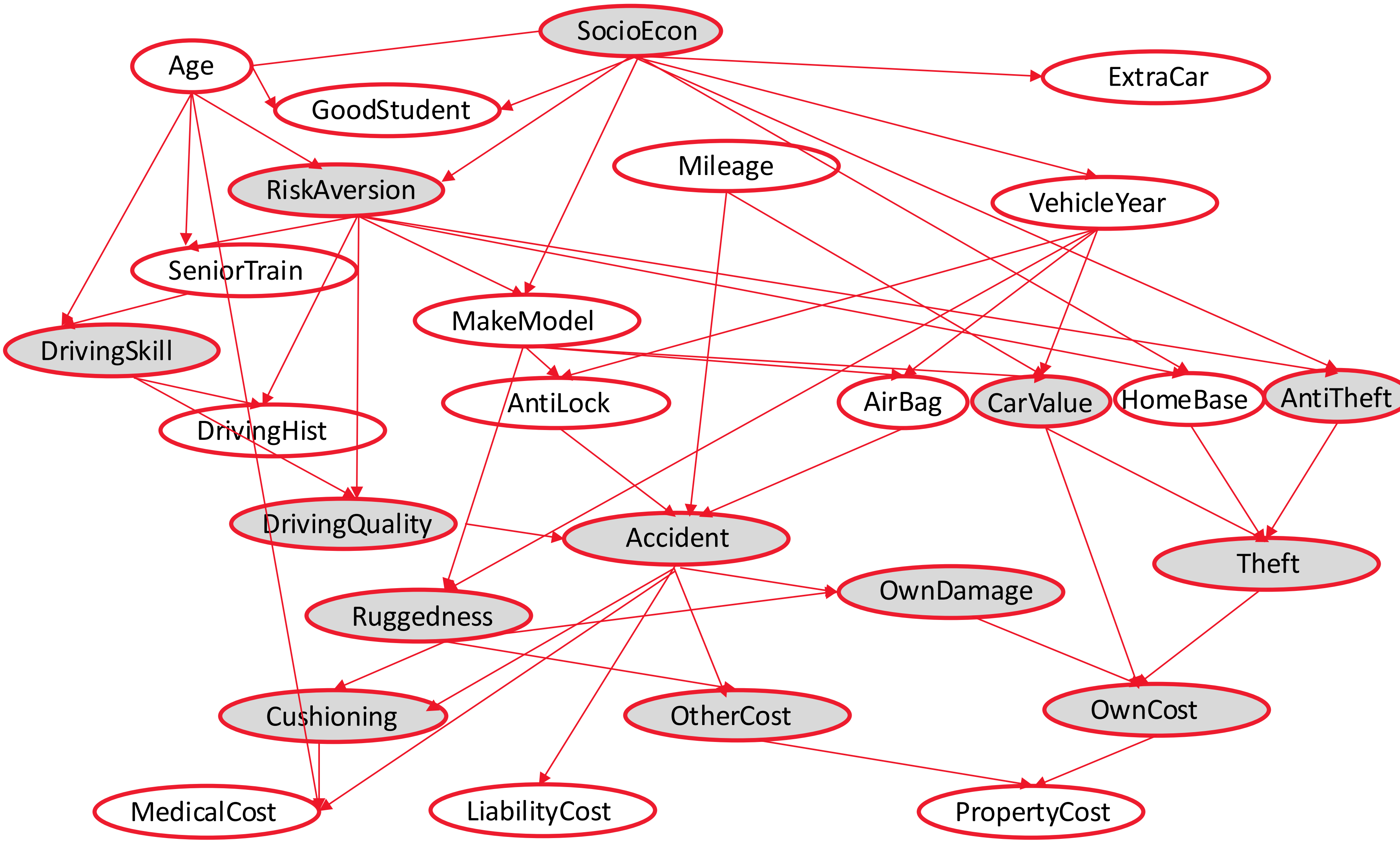
- Represent probability model with a graph

ChatGPT/Wikipedia: A Bayesian network—also called a belief network or a probabilistic directed acyclic graph (DAG)—is a graphical model that represents a set of random variables and their conditional dependencies via a DAG.

Example Application

The potential customer is trying to insure his Mercedes Benz. It is his 3rd car. He was involved in 2 previous minor accidents. The car has airbags and anti-lock braking system. He is 38 years old, married and has 2 kids, and makes $120,000 annually. Is he a risky driver?

Example: car insurance risk assessment



Concept

In the last lecture we saw how to do some simple inference in a set of three variables. Here we introduce two important ideas, and then show how they can be encoded in a *graphical model* or Bayesian network.

- Bayes' rule

- Independence (and conditional independence)

Bayes' Rules

- It is convenient to build statistic model by using causal relationship: $P(effect|cause)$

- Real world requirement $P(cause|effect)$

Bayes' rules: make connection between them!

Bayes' Rules

From product rule we can write

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

Rearranging yields Bayes'rule:

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} = \frac{P(Y|X)P(X)}{\sum_X P(Y|X)P(X)} = \alpha P(Y|X)P(X)$$

Again $\alpha = \frac{1}{P(Y)}$ can be treated as a normalizing constant.

The names of the various components are:

$$\underbrace{P(X|Y)}_{\text{posterior}} = \frac{\overbrace{P(Y|X)}^{\text{likelihood}}\overbrace{P(X)}^{\text{prior}}}{\underbrace{P(Y)}_{\text{evidence}}}$$

Bayes' Rules

The difference between

$P(effect|cause)$ and $P(cause|effect)$?

likelihood posterior

Example

A doctor knows that ***meningitis*** (brain infection, event m) causes the patient to have a ***stiff neck*** (event s) 50% of the time. i.e., $P(s|m) = 0.5$ (likelihood)

She also knows some facts: At any given time, the probability that someone has ***meningitis*** is 1/50,000, i.e., $P(m)$ =1/50,000 (prior)
and the probability that a patient has ***stiff neck*** is 1/20, i.e., $P(s) = 1/20$ (evidence)

A patient visits her with a ***stiff neck***. He is concerned that he might have ***meningitis***.
Performing statistical inference on the meningitis proposition using Bayes' rule yields

$$P(m|s) = \frac{P(s|m)P(m)}{P(s)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

which is very small! This is because the $P(s) \gg P(m)$.

Observe the possibility of large discrepancies between the causal $P(Effect|Cause)$ and diagnostic $P(Cause|Effect)$ probabilities.

Independence

Another concept central to probability and statistics is **independence**.

Formally, random variables A and B are statistically independent if and only if

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

With **Independence**, we can simplify the joint distribution.

Independence

Example:

The variables $Toothache$, $Catch$, and $Cavity$ and independent from the variable $Weather$, i.e.,

$$P(Toothache, Catch, Cavity, Weather)$$

$$= P(Toothache, Catch, Cavity)P(Weather)$$

2*2*2*4-1 = 31
independent entries

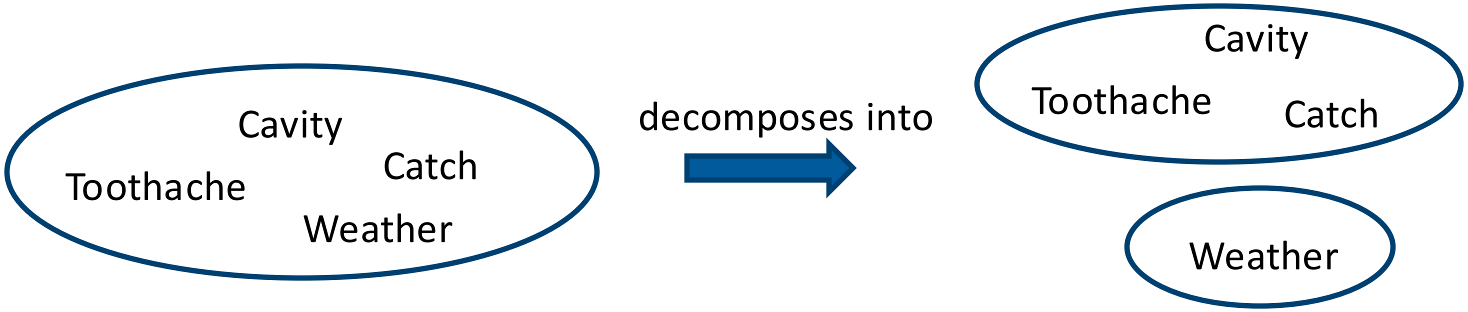
(2*2*2-1)+(4-1) = 10
independent entries

$Weather =$	$sunny$	$rain$	$cloudy$	$snow$
$Cavity = true$	0.144	0.02	0.016	0.02
$Cavity = false$	0.576	0.08	0.064	0.08

$$P(Cavity, Weather) = P(Cavity)P(Weather)$$

Independence

This can be graphically represented as:



This notion of independence is sometimes called absolute independence (we shall see different type of independence later).

It is worth nothing that absolute independence is **powerful** (useful for simplifying statistical inference) but **rare**, e.g., dentistry is a large field with hundreds of variables, none of which are independent.

Conditional Independence

Formally, two random variables X and Y are conditionally independent given a third variable Z if and only if:

$$P(X, Y|Z) = P(X|Z)P(Y|Z) \qquad \text{Eq. (1)}$$

As a comparison, independence is:

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

Equivalently we can write

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

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

Use Eq. (1) => $P(X, Y|Z) = P(X|Z)\cancel{P(Y|Z)} = P(X|Y, Z)\cancel{P(Y|Z)}$

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

Conditional Independence

- Absolute independence -> conditional independence?

No

- Conditional independence -> absolute independence?

No

Example

	toothache		*¬toothache*	
	catch	*¬catch*	*catch*	*¬catch*
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576
Marginalize Cavity	0.124	0.076	0.216	0.584

Now, variables $Catch$ and $Toothache$ are not independent: If the probe catches in the tooth, it probably has cavity and that probably causes toothache.

$$P(Toothache|Catch) \neq P(Toothache)$$

P(Toothache| catch) = <0.62, 0.38>
P(Toothache|¬catch) = <0.27, 0.73>

P(Toothache) = <0.2, 0.8>

However, the two variables are independent, given the presence or absence of cavity.

- Each of toothache and catch is directly caused by the cavity, but neither affects the other: toothache depends on the state of the nerves in the tooth, whereas the probe's accuracy depends on the dentist' skill, to which the toothache is irrelevant.

$$P(Toothache|Catch,\ Cavity) = P(Toothache|Cavity)$$

$$P(Catch|Toothache, Cavity) = P(Catch|Cavity)$$

	toothache		*¬toothache*	
	catch	*¬catch*	*catch*	*¬catch*
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

$$P(Toothache|Catch,\ Cavity) = P(Toothache|Cavity)$$

P(Toothache| catch, cavity) = <0.108, 0.072>/(0.108+0.072)=<0.6, 0.4>
P(Toothache| ¬catch, cavity) = <0.6, 0.4>
P(Toothache| catch, ¬cavity) = <0.1, 0.9>
P(Toothache| ¬catch, ¬cavity) = <0.1, 0.9>

P(Toothache| cavity) = <0.6, 0.4>
P(Toothache|¬cavity) = <0.1, 0.9>

$$P(Catch|Toothache, Cavity) = P(Catch|Cavity)$$

Try it yourself!

Simplification due to conditional independence

The joint probability table of $P(Toothache, Cavity, Catch)$ has 8 entries (see previous lecture notes). However, only **7** of these are independent since the entries must sum to 1.

If we write out the full joint distribution using chain rule and then apply conditional independence:

$$P(Catch, Toothache, Cavity)$$

$$= P(Catch|Toothache, Cavity)P(Toothache, Cavity)$$

$$= P(Catch|Toothache, Cavity)P(Toothache|Cavity)P(Cavity)$$ Chain rule

$$= P(Catch|Cavity)P(Toothache|Cavity)P(Cavity)$$ Conditional independence

Assuming conditional independence on two of the variables allows us to reduce the number of independent entries from **7 to 5:**

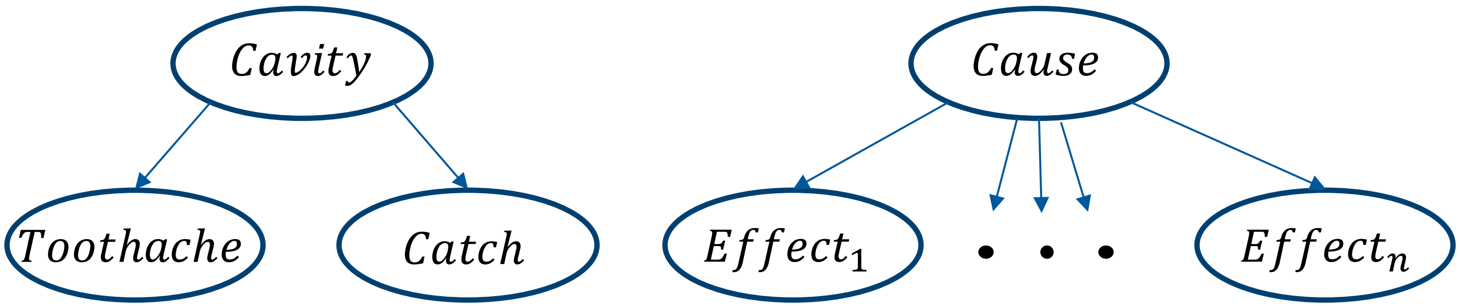
- 1 for $P(Cavity)$
- 2 for $P(Toothache|Cavity)$

  P(toothache|cavity)+P(¬toothache|cavity) =1, P(toothache|-cavity)+P(¬toothache|¬cavity) =1
- 2 for $P(Catch|Cavity)$

  P(catch|cavity)+P(¬catch|cavity) =1, P(catch|¬cavity)+P(¬catch|¬cavity) =1

Naive Bayes

This is an example of a naïve Bayes model:

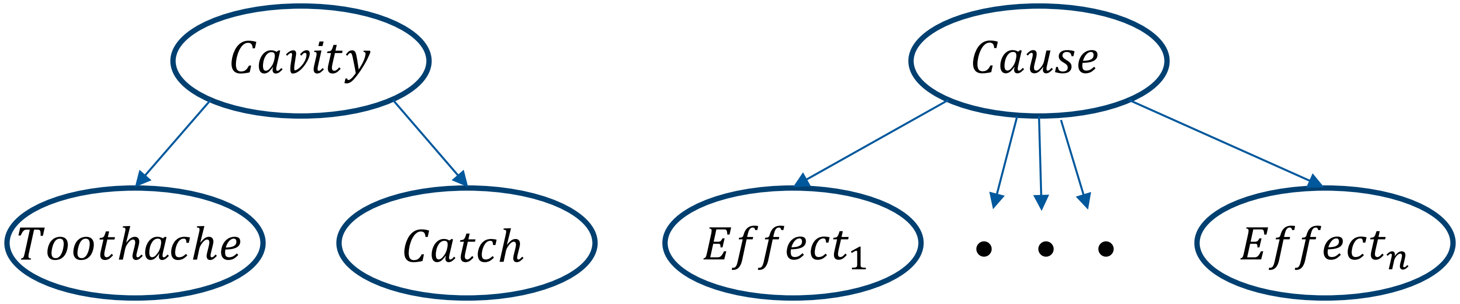


"Naïve" = strong assumption of conditional independence. It is often used as a simplifying assumption in cases where the effect variables are not necessarily conditionally independent given the cause variable.

However, naïve Bayes models can work well in cases where the conditional dependencies between effect variables are weak (this occurs in a surprisingly large number of real-life applications).

Naive Bayes

This is an example of a naïve Bayes model:



The full joint probability:

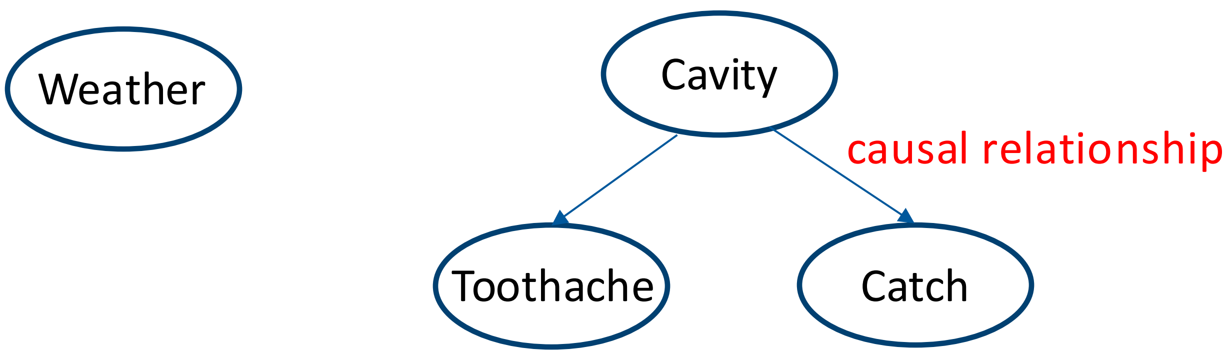
$$P(Casue, Effect_1, \ldots, Effect_n) = P(Cause) \prod_i P(Effect_i | Cause)$$

Bayesian Network

A Bayesian network comprises of the following:

- A set of nodes, one per variable.
- A directed, acyclic graph. This means if you start from a node and follow the arrows there is no way of getting back to the original node.

Example:



Bayesian Network

A Bayesian Network reflects a simple **conditional independence** statement. Namely that each variable is **independent of its nondescendents** in the graph given the state of its parents.

Each node is associated with a conditional probability

$$P\left(X_i \middle| \{X_j\}\right) = P(X_i|Parents(X_i))$$

Once you know the values of X_i's parent nodes, you don't gain any additional information about X_i by looking at any other nodes in the network. This is called local Markov property.

Example: Burglar problem

An inference problem

I'm at work, neighbor John called to say my alarm is ringing, but neighbor Mary didn't call. Sometimes the alarm is set off by minor earthquakes. Is there a burglar?

Variables

$Burglar, Earthquake, Alarm, JohnCalls, MaryCalls$

Network topology reflects "causal" knowledge:

- A burglar can set the alarm off
- An earthquake can set the alarm off
- The alarm can cause Mary to call
- The alarm can cause John to call

Example: Burglar problem

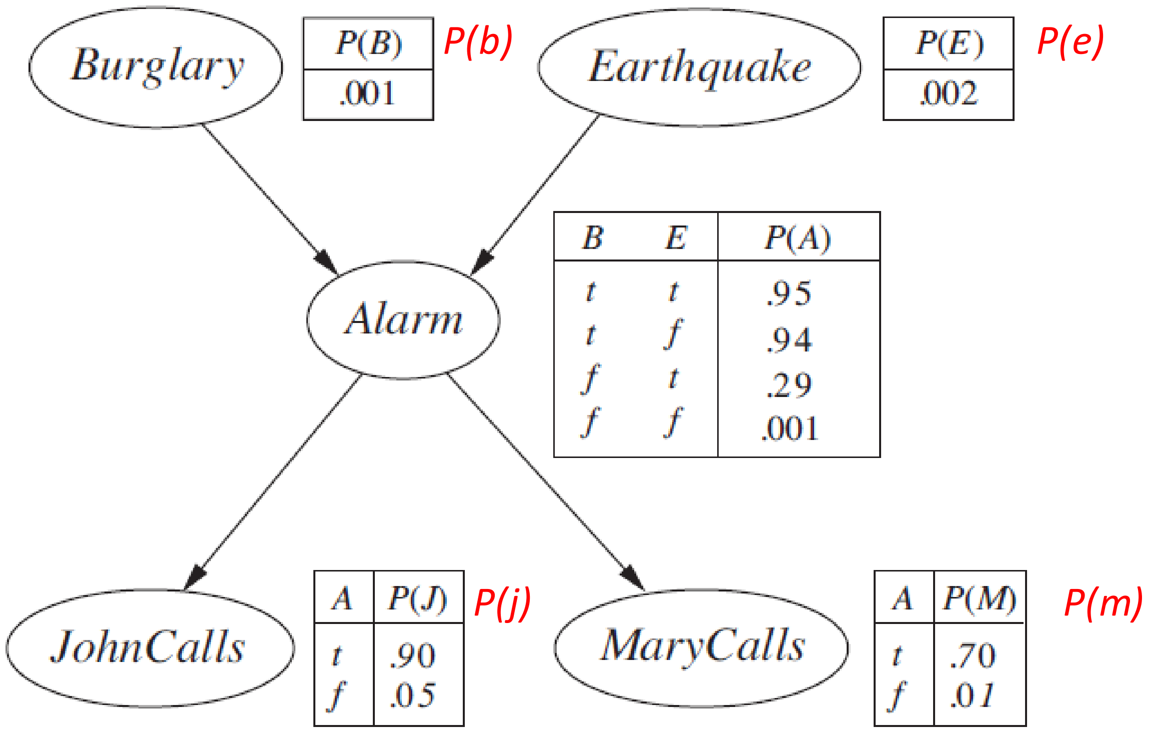
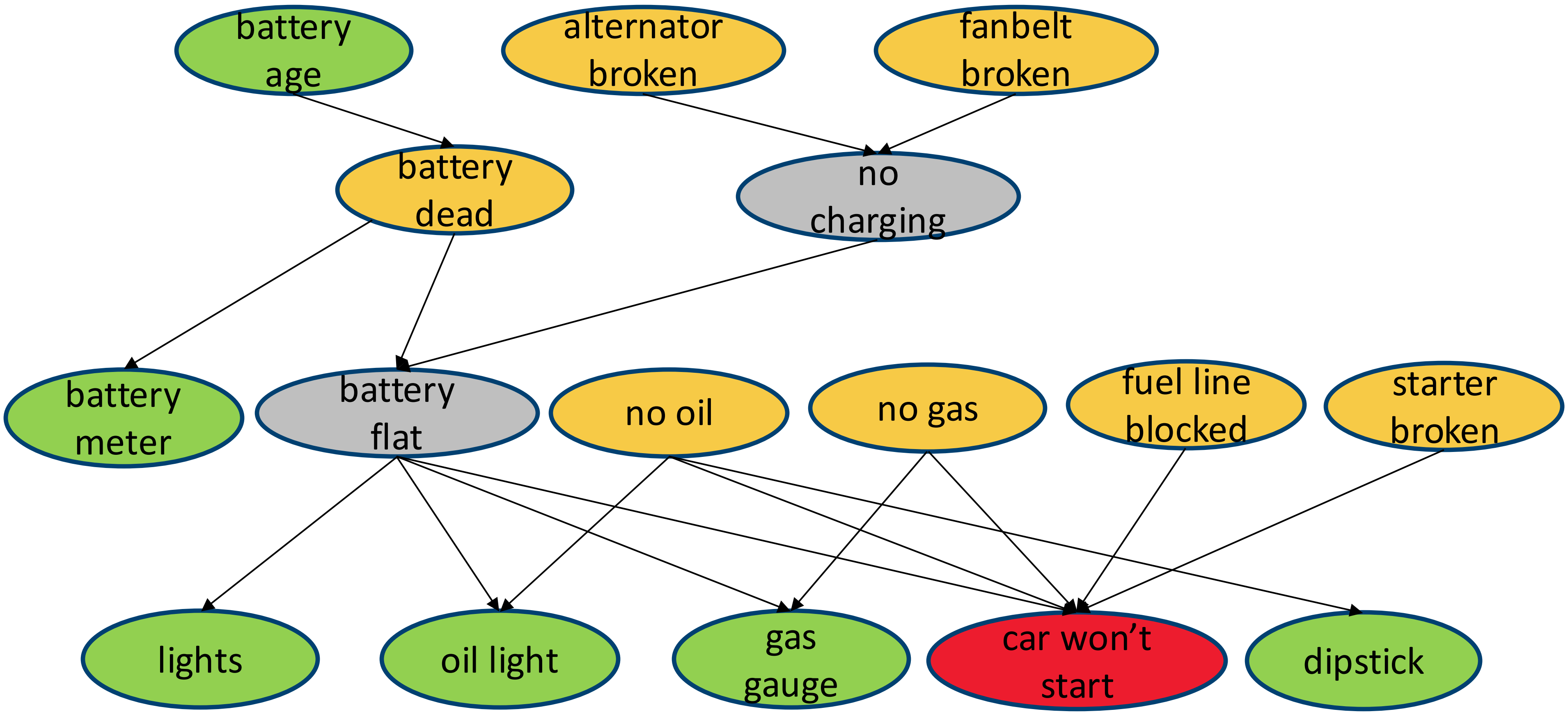


Figure 14.2 A typical Bayesian network, showing both the topology and the conditional probability tables (CPTs). In the CPTs, the letters B, E, A, J, and M stand for *Burglary*, *Earthquake*, *Alarm*, *JohnCalls*, and *MaryCalls*, respectively.

Example: Car start problem



Global Semantics

The global semantics of a network define a joint distribution of all variables as the product of local conditional distributions.

The joint distribution defined by a Bayesian Network with variables $X_1, \ldots, X_n$ is:

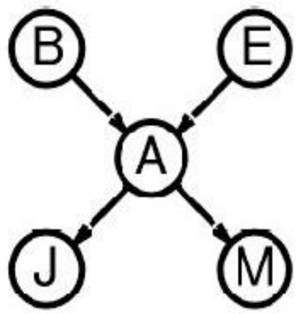
$$P(X_1, \ldots, X_n) = P(X_1|Parents(X_1)) \times P(X_2|Parents(X_2)) \times \cdots \times P(X_n|Parents(X_n))$$

$$= \prod_{i=1}^{n} P(X_i|Parents(X_i))$$

where $Parents(X_i)$ are parents of X_i as specified by the particular Bayesian Network.

Example

For the Burglar Alarm network,



The joint probability distribution of all variables as specified by the network is

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

Given evidence (i.e., observed values) for all the variables, we use the global semantics to obtain the joint probability of the obtained evidence.

Example

Let's say the observed values are John and Mary called, the alarm is ringing, there is no burglary and no earthquake.

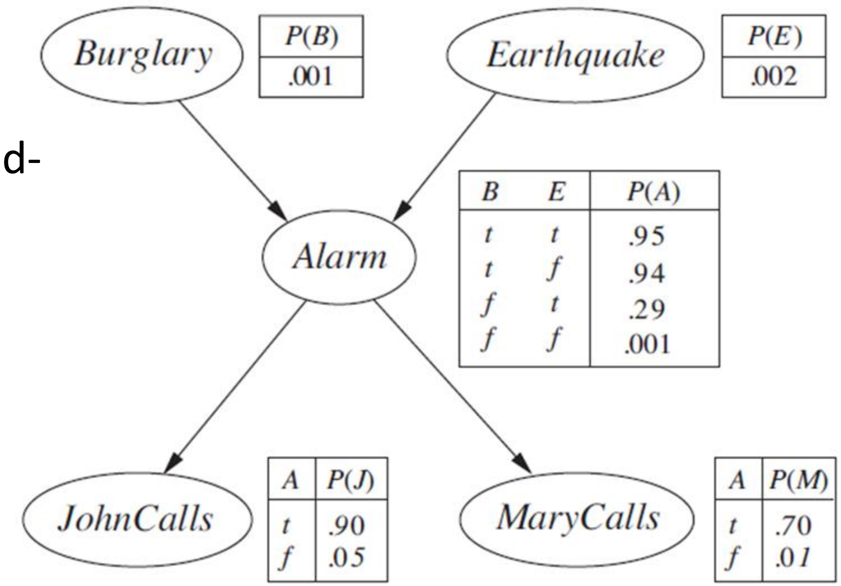
The joint probability of this is

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

$$= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998$$

$$\approx 0.00063$$

For each of the component on the rights-hand-side, simply read off the corresponding CPTs

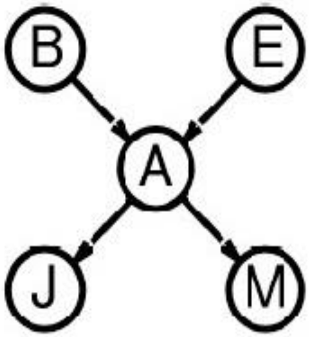


Compactness

The conditional independence assumptions encoded in a Bayesian Network defines a simplified joint distribution of the variables.

For the Burglar Alarm problem where there are **5 Boolean variables**, without conditional independence assumption we need to specify $\mathbf{2^5 - 1 = 31}$ independent numbers to define the joint distribution.

Utilizing the corresponding Bayesian Network, we require only $\mathbf{1 + 1 + 4 + 2 + 2 = 10}$ independent numbers.



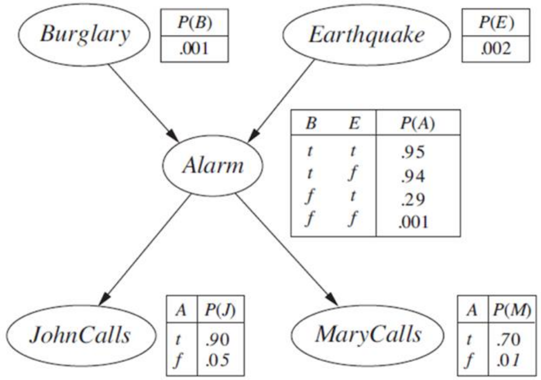
Compactness

More generally, a CPT for a Boolean X_i with k Boolean parents has 2^k rows for the combinations of parent values.

Each row requires one number p for $X_i = True$ (the number for $X_i = False$ is just $1 - p$).

If each variable has no more than k parents, the complete network requires $O(n \cdot 2^k)$ independent numbers, where n is the total number of variables.

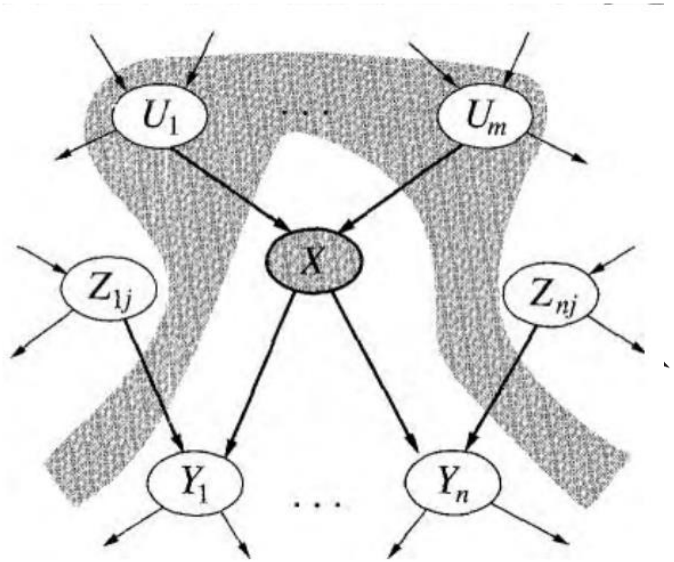
This implies that the required numbers grow linearly with n, versus $O(2^n)$ for the full joint distribution.



Local Semantics

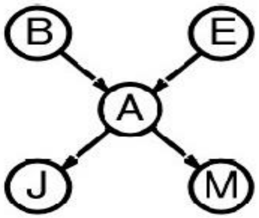
Conditional independence assumptions can simply be “read off” the network topology.

Local semantics: each node is conditionally independent of its non-descendants given its parents.



(a)

Example



Variable J is not independent of variable M, i.e.,

$$P(J,M) \neq P(J)P(M)$$

Intuitively, if John calls, Mary will probably call as well since both would have heard the alarm. The reverse is also true.

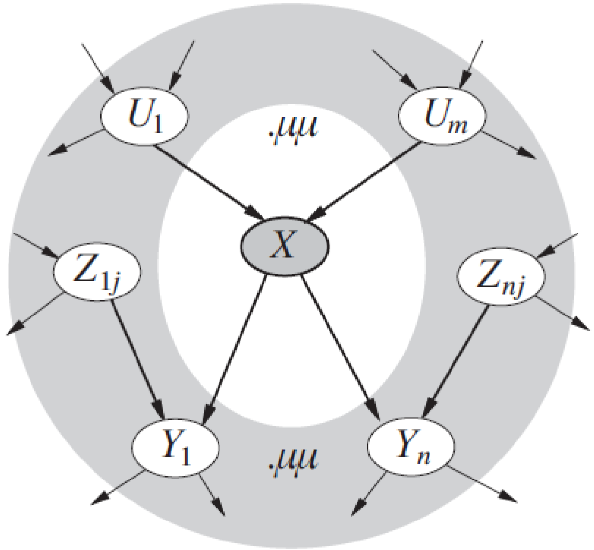
However, J is conditionally independent of M given A, since A is the only parent of J and M is a non-descendent of J. So

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

If we know the alarm did ring, the fact that John calls has no bearing on the probability that Mary calls.

Markov blanket

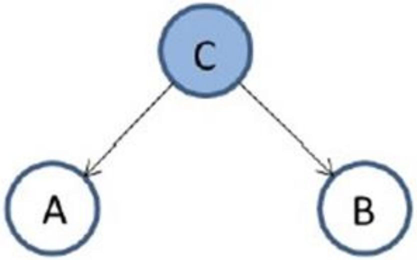
A more specific way to state the local semantics: A node is conditionally independent of all others given its parents, children, and children's <u>other</u> parents -- i.e., given the Markov blanket of the node.

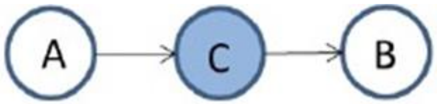


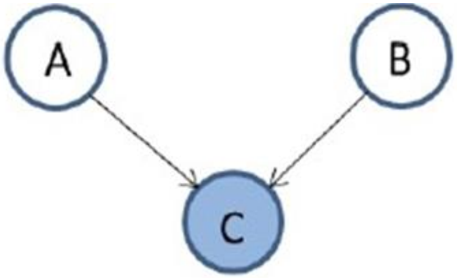
Why do we need to consider the children's parents?

Local Semantics

Consider the possible arrangement of a triplet of nodes in a directed acyclic graph





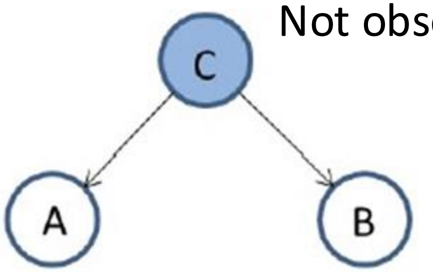


Fork

Chain

Inverted fork

Case 1, Fork (Tail-to-Tail)



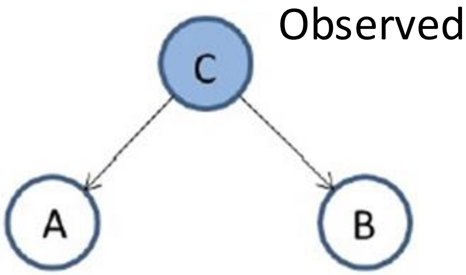
$$P(A, B, C) = P(A|C)P(B|C)P(C)$$

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

In general, this does not factorize into the product $P(A)P(B)$, so

$$A \not\perp\!\!\!\perp B$$

Case 1, Fork (Tail-to-Tail)



$$P(A,B|C) = P(A,B,C)/P(C)$$

Product rule

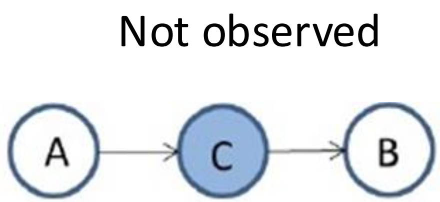
$$= P(A|C)P(B|C)P(C)/P(C)$$

$$= P(A|C)P(B|C)$$

A and B are conditionally independent given C

$$A \perp\!\!\!\perp B|C$$

Case 2, Chain (Head-to-Tail)



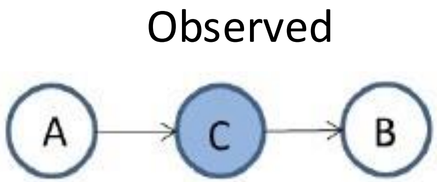
$$P(A,B,C) = P(A)P(C|A)P(B|C)$$

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

In general, this does not factorize into the product $P(A)P(B)$, so

$$A \not\perp\!\!\!\perp B$$

Case 2, Chain (Head-to-Tail)



$$P(A, B|C) = P(A, B, C)/P(C)$$

$$= P(B|C)P(C|A)P(A)/P(C)$$

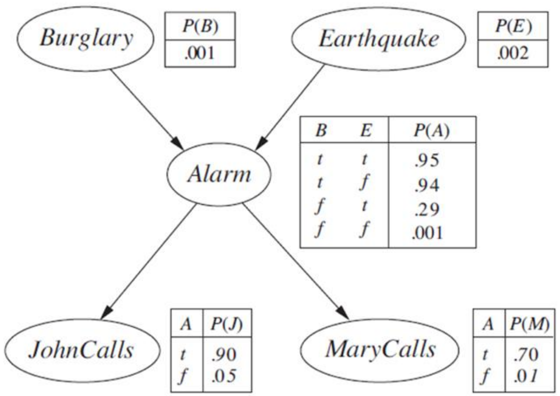
$$= P(B|C)P(A|C)$$

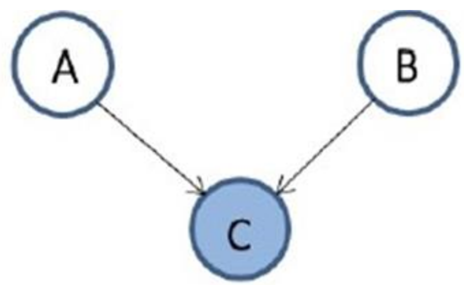
Similar to case 1, A and B are not (unconditionally) independent, but they are given C:

$$A \perp\!\!\!\perp B|C$$

Case 3, Inverted Fork

(Head-to-Head, Collider, or V-structure)





Not observed

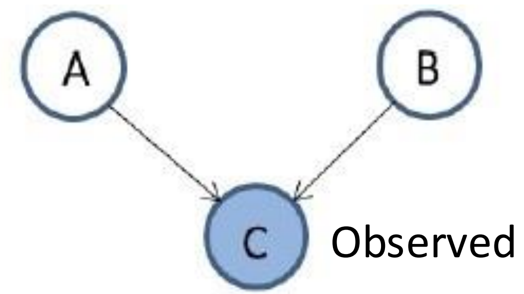
0.001*0.002*(0.95+0.05)+
0.001*0.998*1+
0.999*0.002*1+
0.999*0.998*1 = 1

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

$$P(A,B) = \sum_C P(A)P(B)P(C|A,B) = P(A)P(B)$$

so $A \perp\!\!\!\perp B$

Case 3, Inverted Fork (Head-to-Head)



$$P(A, B|C) = P(A, B, C)/P(C)$$

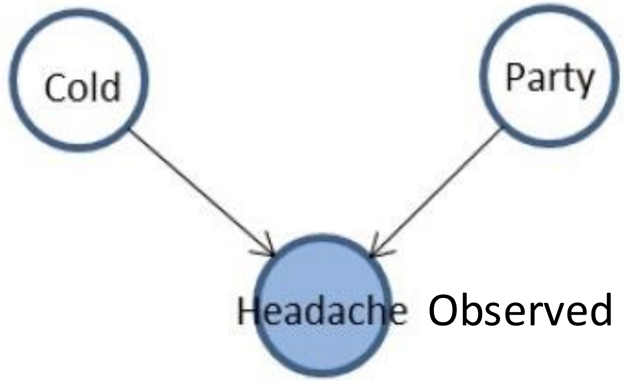
$$= P(C|A, B)P(A)P(B)/P(C)$$

so $A \not\perp\!\!\!\perp B|C$

Case 3 The Explaining-Away

Head-to-head, multiple possible causes, same effect

$P(Cold) = \langle 0.2, 0.8 \rangle$



$P(Party) = \langle 0.1, 0.9 \rangle$

Observed

$$P(headache|party, cold) = 0.95$$

$$P(headache|\neg party, cold) = 0.7$$

$$P(headache|\neg party, \neg cold) = 0.1$$

$$P(headache|party, \neg cold) = 0.8$$

Case 3 The Explaining-Away

This is sufficient information for us to write down the full joint probability because $P(H, P, C) = P(H \mid P, C)P(P)P(C)$:

	party		¬ party	
	cold	¬ cold	cold	¬ cold
headache	0.019	0.056	0.144	0.072
¬ headache	0.001	0.024	0.036	0.648

Before any observations $P(cold) = 0.2$. Now suppose we observe that the person has a headache. What is the probability of having a cold now?

Recall the general rule we used in last lecture

$$P(Cold|headache) = \alpha \sum_{Party} P(Cold, headache, Party)$$

$$= \alpha\langle 0.019 + 0.144, 0.056 + 0.072\rangle = \langle 0.56, 0.44\rangle$$

<c,h,p + c,h,¬p, ¬c,h,p + ¬c,h,-p>

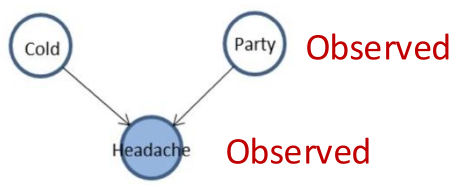
$P(\neg cold|headache)$

$P(cold|headache)$

Case 3 The Explaining-Away

So, the probability of having a cold has increased from 0.2 to 0.56 by knowing the headache (as we would expect intuitively). Now suppose further that we observe that the person went to a party last night.

$$P(cold|headache, party) = \frac{P(cold, headache, party)}{\sum_{Cold} P(Cold, headache, party)}$$



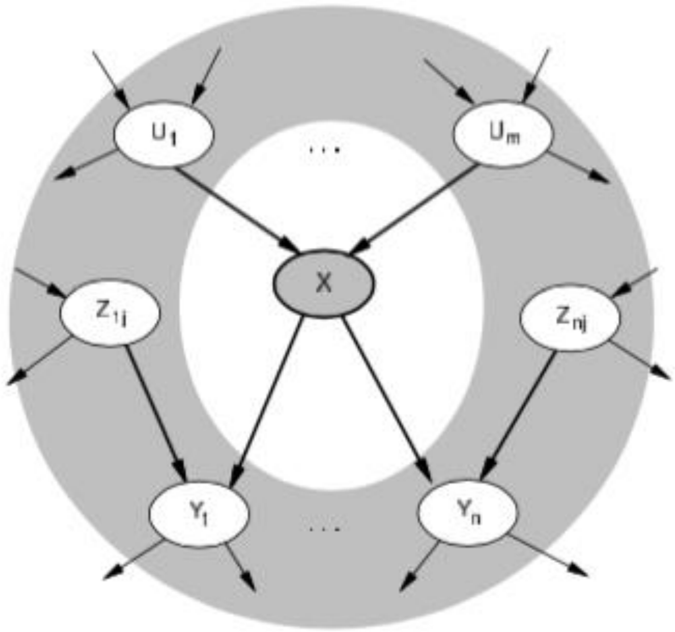
$$= 0.019/(0.019 + 0.056) = 0.25$$

Which is significantly less than $P(cold|headache) = 0.56$. The observation that she went to a party last night (so many well have a hangover) *explains away* the cold as a cause.

P(¬cold|headache, party) = 0.75!!

Markov blanket

A node is conditionally independent of all others given its parents, children, and children's <u>other</u> parents -- i.e., given the Markov blanket of the node. Why do we need to consider the children's other parents?? Because of the explaining away effect.



Back to inference problem

So far, we have learnt how to obtain the joint probability according to a Bayesian network given the value of all variables.

However, the sort of problems we wish to solve are statistical inference problems, i.e., we have a query variable, some evidence variables, and some unobserved variables, i.e., we want to compute

$$P(X|e) = \alpha \sum_{\forall Y} P(X, e, Y)$$

Example: "I'm at work, neighbor John called to say my alarm is ringing, but neighbor Mary didn't call. Sometimes it's set off by minor earthquakes. Is there a burglar?" $P(burglar|jCall, \neg mCall)$

How to accomplish this using Bayesian Networks? We shall study this in the next lecture.