Bayesian Networks

3007/7059 Artificial Intelligence

School of Computer Science
The University of Adelaide

Inference

To perform inference is to make conclusions on the basis of evidence and reasoning.

We make inferences on a daily basis:

- In the morning I turn the ignition key and the car doesn't start. The fuel gauge says there is fuel in the tank. The car was serviced 3 weeks ago. Is the battery dead?
- I plug in the network cable but I can't go online. The router seems to have assigned my laptop an IP address. I paid the latest broadband bill last week. Is my ethernet card is still working?

The ability to perform inference also has wide practical and commercial applications.

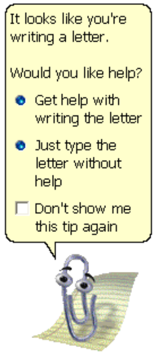
Insurance risk assessment

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 with 2 kids and makes $120,000 annually. Is he a risky driver?

"I know you haven't had an accident in 15 years. That's why we're raising your rates - because you're about due one."

Microsoft Office Assistant

The user started her document with what looks like an address which is then followed by today's date. She then types the words "Dear Sir" in a new line. In the line that follows the words "Subject" and "Complaint" appear, and the whole line is underlined. Is the user typing a letter?



It looks like you're writing a letter.

Would you like help?

● Get help with writing the letter

● Just type the letter without help

□ Don't show me this tip again

Concepts

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

- Independence (and conditional independence)
- Bayes rule

Basic probability and statistics again: Independence

Another concept central to probability and statistics is independence.

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

$$P(A|B) = P(A), \text{ or } P(B|A) = P(B), \text{ or } P(A,B) = P(A)P(B)$$

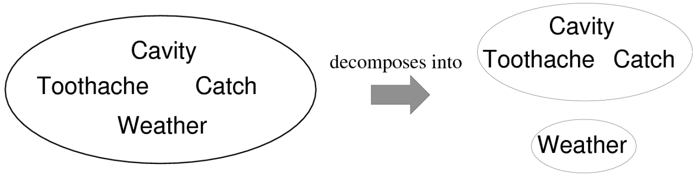
Example:

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

$$\begin{aligned} & P(Toothache, Catch, Cavity, Weather) \\ = & P(Toothache, Catch, Cavity)P(Weather) \end{aligned}$$

Basic probability and statistics again: Independence (cont.)

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 noting 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.

Bayes' rule

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 normalising constant.

The names of the various components are

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

Bayes' rule (cont.)

In certain areas (e.g. medical diagnostics) it is more natural to express our domain knowledge in terms of causal relationships (rather than a full joint probability distribution of the variables).

Given observations of the effects, Bayes' rule allows us to convert our knowledge expressed in terms of causal relationships into diagnostic probabilities:

$$P(Cause|Effect) = \frac{\overbrace{P(Effect|Cause)}^{\text{Causal relationship}} P(Cause)}{P(Effect)}$$

Of course, for this to be applicable we also require knowledge of $P(Cause)$ and $P(Effect)$.

Bayes' rule (cont.)

Bayes rule interepreted for robot sensing:

$$P(\mathbf{x}|\mathbf{z}) = \frac{\overbrace{P(\mathbf{z}|\mathbf{x})}^{\text{Sensor model, likelihood}} \overbrace{P(\mathbf{x})}^{\text{Prior}}}{P(\mathbf{z})}$$

Example

A doctor knows that meningitis causes the patient to have a stiff neck 50% of the time.

She also knows some facts: At any given time the probability that someone has meningitis is 1/50000, and the probability that a patient has stiff neck is 1/20.

A patient visits her with a stiff neck (indicated by event s). He is concerned that he might have meningitis (event m).

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.

Combining evidences using Bayes' rule

Sometimes we have more than one piece of evidence. For example, what can a dentist conclude about cavity if the steel probe catches the aching tooth of a patient? Via Bayes' rule,

$$P(Cavity|catch, toothache) = \alpha P(catch, toothache|Cavity)P(Cavity)$$

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

The 2 variables are independent, however, given the presence or absence of cavity. Each 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.

Combining evidences using Bayes' rule (cont.)

This can be expressed as

$$P(toothache, catch|Cavity) = P(toothache|Cavity)P(catch|Cavity)$$

This implies that variables $Toothache$ and $Catch$ are conditionally independent given $Cavity$.

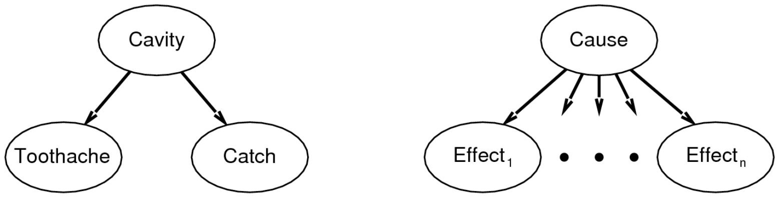
Plugging this back into Bayes' rule yields

$$\begin{aligned} & P(Cavity|toothache, catch) \\ = \; & \alpha P(toothache, catch|Cavity)P(Cavity) \\ = \; & \alpha \underbrace{P(toothache|Cavity)}_{\text{causal}} \underbrace{P(catch|Cavity)}_{\text{causal}} P(Cavity) \end{aligned}$$

Observe now that the inference makes use of two causal relationships.

Combining evidences using Bayes' rule (cont.)

This is an example of a naive Bayes model:



"Naive" because it is often used as a simplifying assumption in cases where the effect variables are not conditionally independent given the cause variable.

However, naive 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).

Conditional independence

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

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

Equivalently we can write

$$P(X|Y, Z) = P(X|Z) \;\; \text{and} \;\; P(Y|X, Z) = P(Y|Z)$$

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:

$$\begin{aligned} & P(Catch, Toothache, Cavity) \\ = \; & P(Catch|Toothache, Cavity)P(Toothache, Cavity) \\ = \; & P(Catch|Toothache, Cavity)P(Toothache|Cavity)P(Cavity) \\ = \; & P(Catch|Cavity)P(Toothache|Cavity)P(Cavity) \end{aligned}$$

Assuming conditional independence on 2 of the variables allows us to reduce the number of independent entries from 7 to 5:

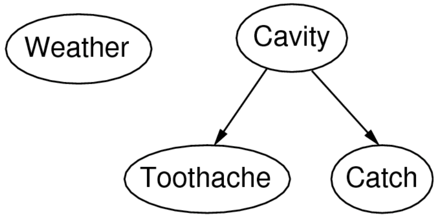
- 1 for $P(Cavity)$
- 2 for $P(Toothache|Cavity)$
- 2 for $P(Catch|Cavity)$

Bayesian Networks

Bayesian Networks are simple, graphical notations for conditional independence assertions.

Since a Bayesian Network is built upon a set of conditional independence assumptions, it also provides a compact specification of joint distribution of the variables involved.

Example:



$Weather$ is independent of the other variables.

$Toothache$ and $Catch$ are conditionally independent given $Cavity$.

Bayesian Networks (cont.)

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.
- A conditional distribution for each node given its parents:

$$P(X_i|Parents(X_i))$$

In the simplest case, each conditional distribution is represented as a conditional probability table (CPT) giving the distribution over X_i for each combination of parent values.

Example: Burglar alarm

An inference problem

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

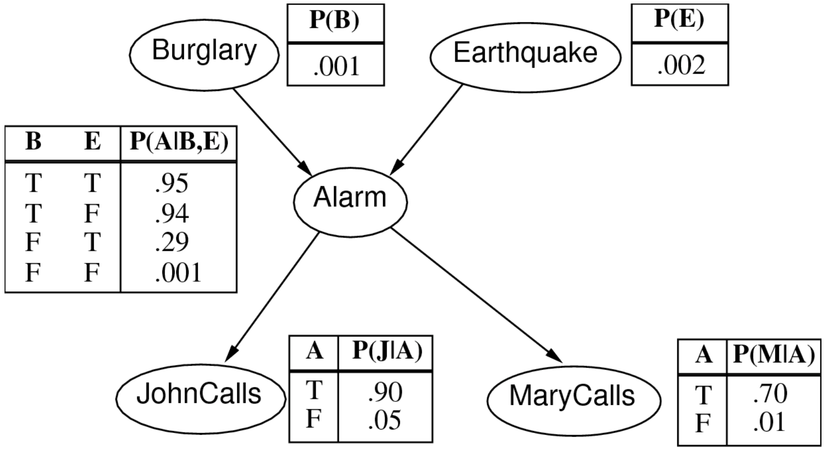
Variables

$Burglar$, $Earthquake$, $Alarm$, $JohnCalls$, $MaryCalls$

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 alarm (cont.)



P(B)
.001

P(E)
.002

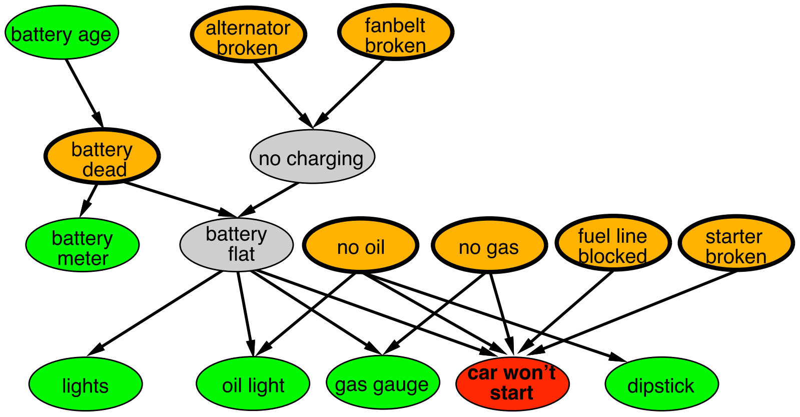
B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J|A)
T	.90
F	.05

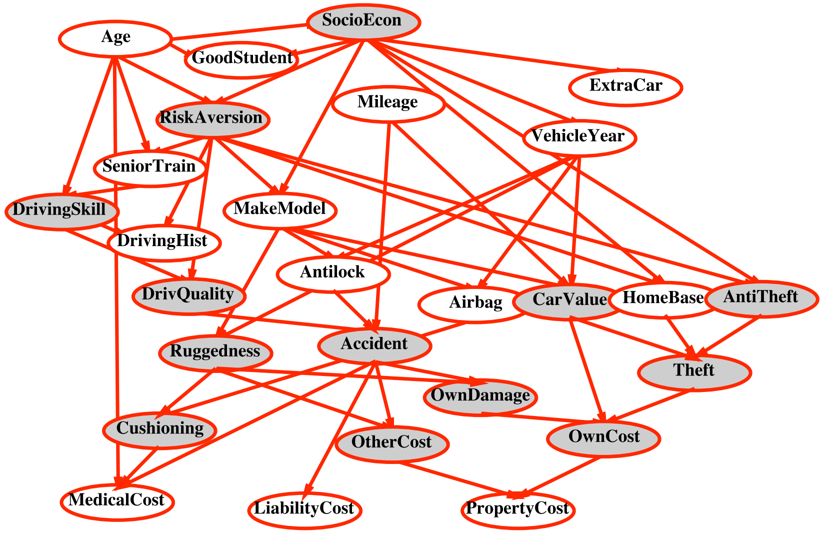
A	P(M|A)
T	.70
F	.01

Legend: $B = Burglar$, $E = Earthquake$, $A = Alarm$,
$M = MaryCalls$, $J = JohnCalls$.

Example: Car start problem



Example: Car insurance risk assessment



Global semantics

A Bayesian Network encodes our knowledge on how the variables interact. This is specified in terms of conditional independence assertions.

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

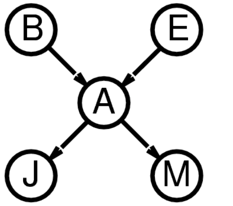
The joint distribution defined by a Bayesian Network with variables $X_1, \dots, X_n$ is:

$$\begin{aligned} P(X_1, \dots, X_n) &= P(X_1|Parents(X_1)) \times P(X_2|Parents(X_2)) \\ &\quad \times \cdots \times P(X_n|Parents(X_n)) \\ &= \prod_{i=1}^{n} P(X_i|Parents(X_i)) \end{aligned}$$

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 obtaining the evidences.

Example (cont.)

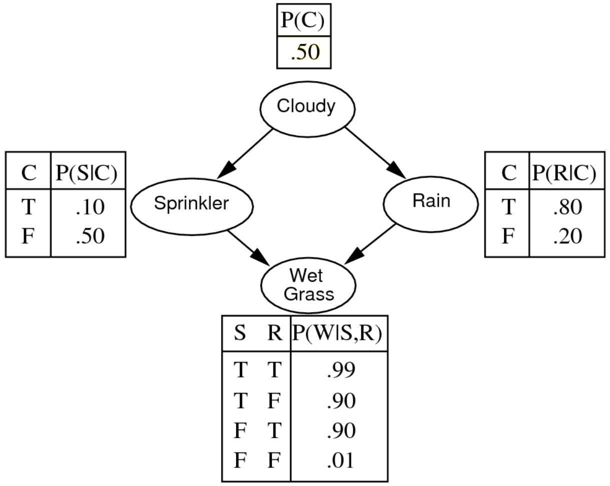
Lets say the observed values are John and Mary called, the alarm is ringing, there is no burglary and no earthquake.

The joint probability of this is

$$\begin{aligned} P(j, m, a, \neg b, \neg e) &= P(j|a)P(m|a)P(a|\neg b, \neg e)P(\neg b)P(\neg e) \\ &= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998 \\ &\approx 0.00063 \end{aligned}$$

(for each of the component on the right-hand-side, simply read off the correspoding CPTs)

Another example



P(C)
.50

C	P(S|C)
T	.10
F	.50

C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

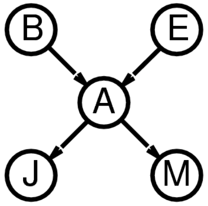
Write the joint distribution as specified by the Bayesian Network. Then obtain the joint probability of $Cloudy = false$, $Sprinkler = True$, $Rain = False$ and $WetGrass = True$.

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 assumptions we need to specify $2^5 - 1 = 31$ independent numbers to define the joint distribution.

Utilising the corresponding Bayesian Network, we require only $1 + 1 + 4 + 2 + 2 = 10$ independent numbers.



Compactness (cont.)

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 $1 - p$).

For example, the CPT for $Alarm(A)$ in the Burglar Alarm network with parents $Burglary(B)$ and $Earthquake(E)$:

B	E	$P(a\|B,E)$	$P(\neg a\|B,E)$
T	T	0.95	$1 - 0.95$
T	F	0.94	$1 - 0.94$
F	T	0.29	$1 - 0.29$
F	F	0.001	$1 - 0.001$

Recall the notation: Lowercase letters (e.g. a, $\neg a$) represent instantiations (values) of Boolean random variables (e.g. A).

Compactness (cont.)

We need $4 = 2^2$ rows for the CPT $P(Alarm|Parents(Alarm))$ and thus 4 independent probability values. Note that the probabilities of each column do not sum to 1.

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.

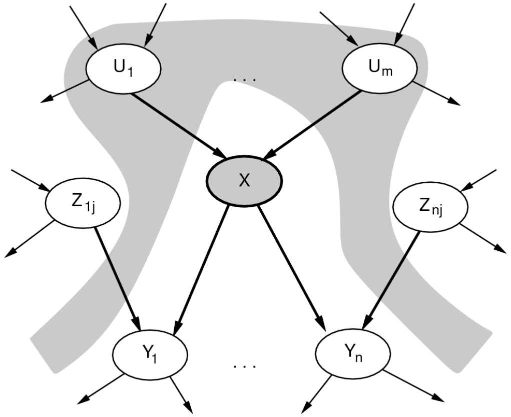
This implies that the required numbers grow linearly with n, versus $O(2^n)$ for the full joint distribution.

Example: How many independent numbers are required to specify the "Car start problem" Bayesian Network? How many do we need if we discard all conditional independence assumptions? Assume all variables are Boolean.

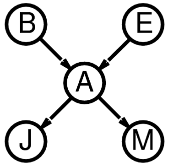
Local semantics

It is not surprising that the conditional independence assumptions can simply be "read off" the network topology.

The local semantics: each node is conditionally independent of its nondescendants given its parents.



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 true.

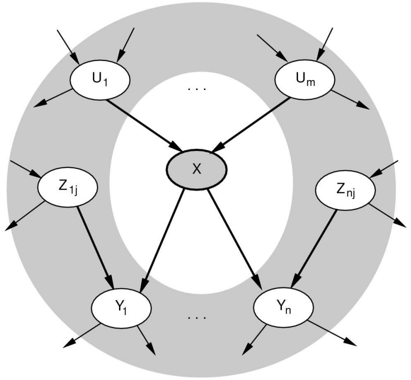
However J is conditionally independent of M given A, since A is the only parent of J and M is a non-descendent of J. So

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

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

Markov blanket

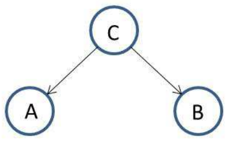
A more specific way to state the local semantics: A node is conditionally independent of all others given its parents, children, and children's parents— i.e., given the Markov blanket of the node.



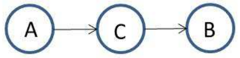
Why do we need to consider the children's parents??

Local semantics

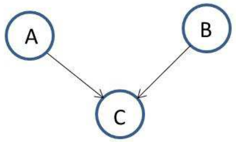
Consider the possible arrangement of a triplet of nodes in a directed acyclic graph



Fork

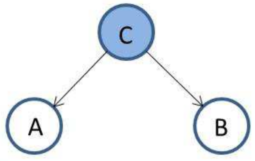


Chain



Inverted fork

Triplet semantics: Case 1 Fork (tail-to-tail)

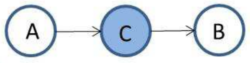


$$\begin{aligned} P(A,B|C) &= P(A,B,C)/P(C) \\ &= P(A|C)P(B|C)P(C)/P(C) \\ &= P(A|C)P(B|C) \end{aligned}$$

A and B are not (unconditionally) independent, but A and B are *conditionally independent given* C

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

Triplet semantics: Case 2 Head-to-tail

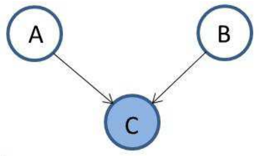


$$\begin{aligned} P(A,B|C) &= P(A,B,C)/P(C) \\ &= P(B|C)P(C|A)P(A)/P(C) \\ &= P(B|C)P(A|C) \end{aligned}$$

As for case 1, A and B are not (unconditionally) independent, but:

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

Triplet semantics: Case 3 Inverted fork

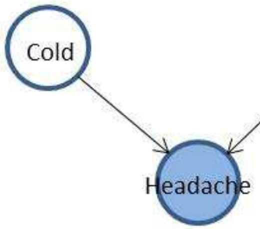


$$\begin{aligned} P(A,B|C) &= P(A,B,C)/P(C) \\ &= P(C|A,B)P(A)P(B)/P(C) \end{aligned}$$

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

Case 3: Explaining away

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



$$\begin{array}{rclrcl} & & & P(\text{headache}|\text{party}, \text{cold}) & = & 0.95 \\ P(\text{Party}) & = & \langle 0.1, 0.9\rangle & P(\text{headache}|\neg\text{party}, \text{cold}) & = & 0.7 \\ P(\text{Cold}) & = & \langle 0.2, 0.8\rangle & P(\text{headache}|\neg\text{party}, \neg\text{cold}) & = & 0.1 \\ & & & P(\text{headache}|\text{party}, \neg\text{cold}) & = & 0.8 \end{array}$$

Case 3: Explaining away (cont.)

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

	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 $P(\text{cold}) = 0.2$. Now suppose we observe that the person has a headache. What is the probability of having a cold now?

$$\begin{aligned} P(cold|headache) &= \alpha \sum_{Party} P(cold, headache, Party) \\ &= \alpha\langle 0.019 + 0.144, 0.056 + 0.072\rangle = \langle 0.56, 0.44\rangle \end{aligned}$$

Case 3: Explaining away (cont.)

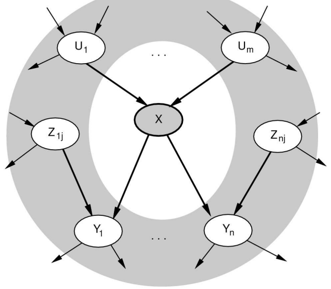
so the probability of having a cold has increased (as we would expect intuitively). Now suppose further that we observe that the person went to a party last night.

$$\begin{aligned} P(cold|headache, party) &= \frac{P(cold, headache, party)}{\sum_{Cold} P(Cold, headache, party)} \\ &= 0.019/(0.019 + 0.056) = 0.25 \end{aligned}$$

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

Markov blanket revisited

A node is conditionally independent of all others given its parents, children, and children's parents— i.e., given the Markov blanket of the node. Why do we need to consider the children's parents?? Because of the explainng away effect.



More general concept still known as *D-separation* (see Bishop and other texts).

Back to the 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, neighbour John called to say my alarm is ringing, but neighbour Mary didn’t call. Sometimes it’s set off by minor earthquakes. Is there a burglar?”

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