Bayesian Networks

Today we'll introduce Bayesian Networks.

This material is covered in chapters 13 and 14. Chapter 13 gives basic background on probability and Chapter 14 talks about Bayesian Networks. This includes methods for exact reasoning in Bayes Nets as well as approximate reasoning.

Bayesian Networks

Ultimately, our goal is to model the relationship between events and to use this model to make informed guesses about events. For example, consider this story:

> "If Craig woke up too early (i.e. if E is true), he probably needs coffee (C); if Craig needs coffee, he's likely angry (A). If he is angry, he has an increased chance of bursting a brain vessel (B). If he bursts a brain vessel, Craig is quite likely to be hospitalized (H)."

E – Craig woke too early   A – Craig is angry   H – Craig hospitalized
C – Craig needs coffee   B – Craig burst a blood vessel

But before we get to this place, let's quickly remember ….

Notation

1. P(A) = Pr(A)

2. P(A and B) = P(A,B) = $P(A \cap B)$ = $P(A \wedge B)$

3. P(A or B) = $P(A \cup B)$ = P(AVB)

4. We sometimes use P(A|B) to indicate P(A=true|B=true)

5. We sometimes use P(~A) to indicate P(A=false)

6. Events can be thought of as feature (or variable) vectors; each feature can take on values, i.e. {A=a, B=b,…, N=n}. Call P(A=a) the probability that variable A takes on the value a.

Axioms of Probability?

Axioms of Probability?

1. $P(U) = 1$

2. $P(A) \in [0,1]$

3. $P(\{\}) = 0$

4. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

NB: if $A \cap B = \{\}$ *then* $P(A \cup B) = P(A) + P(B)$

Conditional Probability?

Conditional Probability?

$$P(B|A) = P(B \cap A)/P(A)$$

Independence?

Independence?

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

Conditional Independence?

Conditional Independence?

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

Chain Rule?

Chain Rule?

$$P(A_1 \wedge A_2 \wedge \ldots \wedge A_n) =$$
$$P(A_1 | A_2 \wedge \ldots \wedge A_n) * P(A_2 | A_3 \wedge \ldots \wedge A_n)$$
$$* \ldots * P(A_{n-1} | A_n) * P(A_n)$$

Marginalizing (Summing Out) a variable?

Given P(A,B)…

P(A) =

Marginalizing (Summing Out) a variable?

Given P(A,B)…

$P(A) = P(A|B=b_1)P(B=b_1) + P(A|B=b_2)P(B=b_2) + \ldots + P(A|B=b_k)P(B=b_k)$

Bayes Rule?

Bayes Rule?

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

$$\begin{aligned} P(Y|X) &= P(Y,X)/Pr(X) \\ &= P(Y,X)/P(X) * P(Y)/P(Y) \\ &= P(Y,X)/P(Y) * P(Y)/P(X) \\ &= P(X|Y)P(Y)/P(X) \end{aligned}$$

The benefits of independence

If A and B are independent, that means we can break apart P(A,B) as follows:

$$P(A,B) = P(B) * P(A)$$

Why is this so?

$$P(B|A) = P(B) \qquad \text{(def'n of independence)}$$
$$P(A,B)/P(A) = P(B)$$
$$P(A,B) = P(B) * P(A)$$

The benefits of independence

If B and C are conditionally independent given A that means we can break apart P(B,C|A) as follows:

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

Why?

P(B|C,A) = P(B|A) (def'n of conditional independence)
P(B,C,A)/P(C,A) = P(B,A)/P(A)
P(B,C,A)/P(A) = P(C,A)/P(A) * P(B,A)/P(A)
P(B,C|A) = P(B|A) * P(C|A) .

The benefits of independence

Conditional independence also allows us to ignore certain pieces of information during our computations, as

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

The benefits of independence

- Given independence (and binary variables), our representation of joints and inference requires O(n) variables instead of O(2^n)!
- Unfortunately, complete mutual independence is very rare. Most realistic domains do not exhibit this property.
- But, most domains *do* exhibit a fair amount of conditional independence. We can exploit conditional independence for representation and inference.
- **Bayesian networks** do just this.

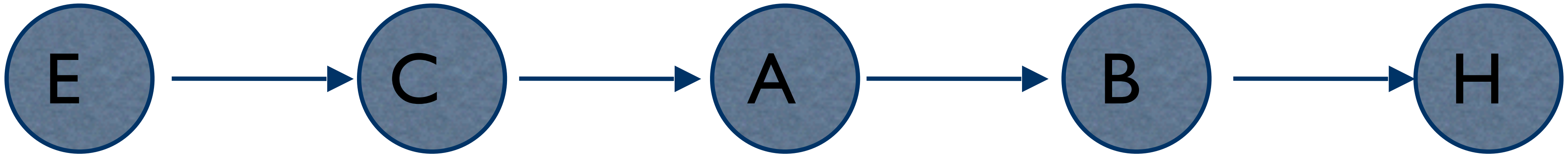
Now back to our story

“If Craig woke up too early (E is true), Craig probably needs coffee (C); if Craig needs coffee, he's likely angry (A). If he is angry, he has an increased chance of bursting a brain vessel (B). If he bursts a brain vessel, Craig is quite likely to be hospitalized (H).”

E – Craig woke too early   A – Craig is angry   H – Craig hospitalized

C – Craig needs coffee   B – Craig burst a blood vessel

Modelling independence



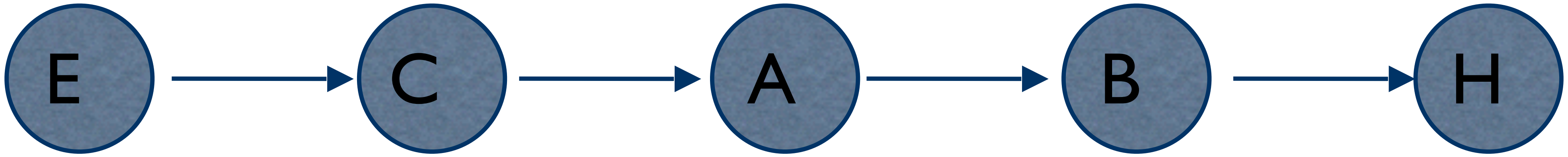
If you knew E, C, A, or B, your assessment of P(H) would change.

- E.g., if any of these are seen to be true, you would increase P(H) and decrease P(~H).
- This means H is not independent of E, or C, or A, or B.

If you knew B, you'd be in good shape to evaluate P(H). You would not need to know the values of E, C, or A. The influence these factors have on H is mediated by B.

- Craig doesn't get sent to the hospital because he's angry, he gets sent because he's had an aneurysm.
- So H is independent of E, and C, and A given B

Modelling independence



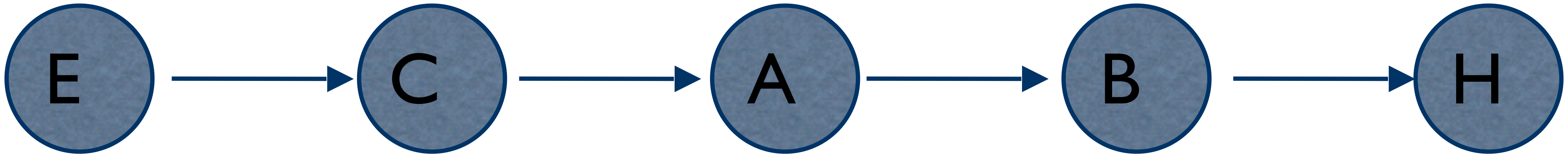
Similarly:

- B is independent of E, and C, given A
- A is independent of E, given C

This means that:

- P(H | B, {A,C,E}) = P(H|B)
  - i.e., for any subset of {A,C,E}, this relation holds
- P(B | A, {C,E}) = P(B | A)
- P(A | C, {E}) = P(A | C)
- P(C | E) and P(E) don't "simplify"

Modelling independence



By the chain rule (for any instantiation of H…E):

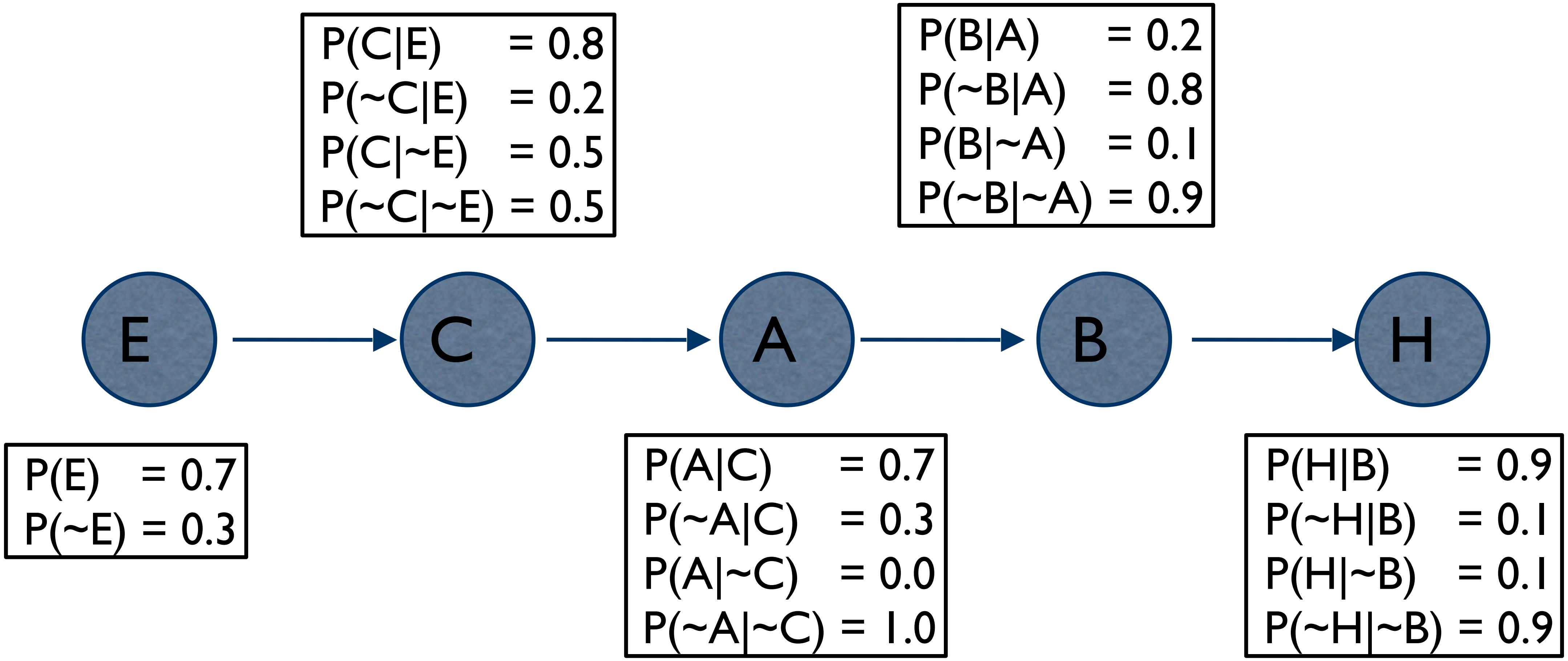
$$P(H,B,A,C,E) = P(H|B,A,C,E)\ P(B|A,C,E)\ P(A|C,E)\ P(C|E)\ P(E)$$

By our independence assumptions:

$$P(H,B,A,C,E) = P(H|B)\ P(B|A)\ P(A|C)\ P(C|E)\ P(E)$$

So we can specify the full joint by specifying five local conditional distributions (joints): P(H|B); P(B|A); P(A|C); P(C|E); and P(E)

Adding the numbers



P(E)	= 0.7
P(~E)	= 0.3

P(C|E)	= 0.8
P(~C|E)	= 0.2
P(C|~E)	= 0.5
P(~C|~E)	= 0.5

P(A|C)	= 0.7
P(~A|C)	= 0.3
P(A|~C)	= 0.0
P(~A|~C)	= 1.0

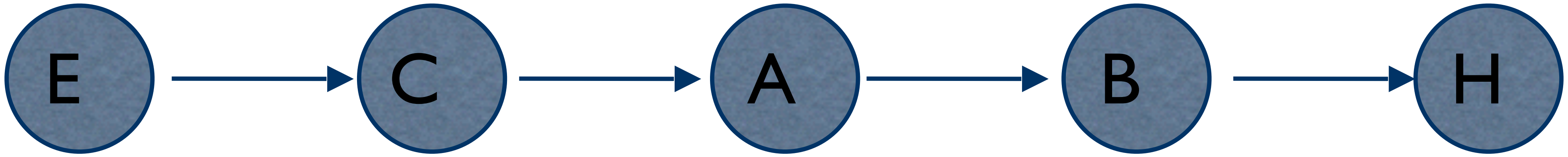
P(B|A)	= 0.2
P(~B|A)	= 0.8
P(B|~A)	= 0.1
P(~B|~A)	= 0.9

P(H|B)	= 0.9
P(~H|B)	= 0.1
P(H|~B)	= 0.1
P(~H|~B)	= 0.9

Specifying the joint requires only 9 parameters (if we note that half of these are "1 minus" the others), instead of 31 for explicit representation

- That means inference is linear in the number of variables instead of exponential!
- Moreover, inference is linear generally if dependence has a chain structure

Inference with a BN



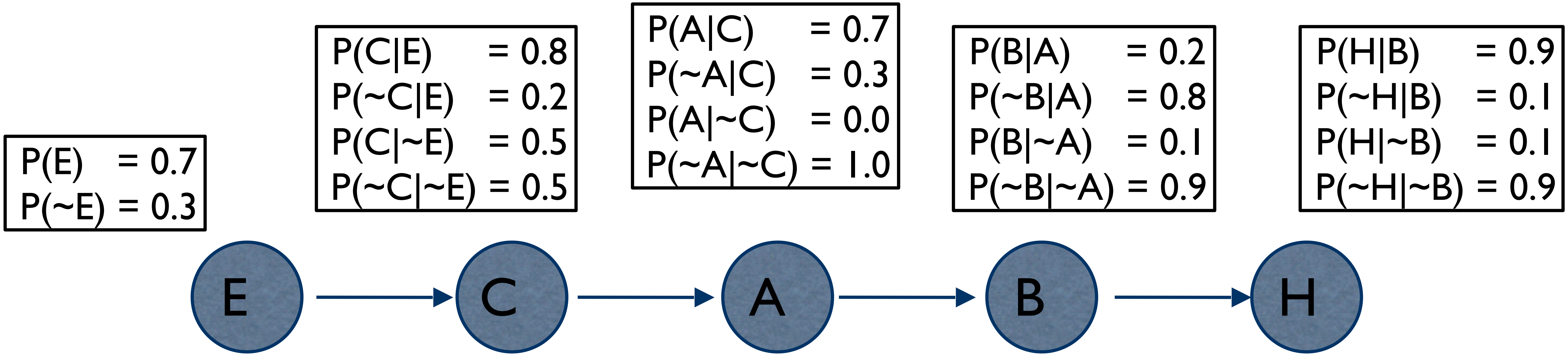
Want to know P(A)? Proceed as follows:

$$P(A=true) = \sum_{c_i \text{ in Dom}(C)} P(A=true|C = c_i) * P(C = c_i)$$

$$= \sum_{c_i \text{ in Dom}(C)} P(A=true|C = c_i) \sum_{e_i \text{ in Dom}(E)} P(C = c_i|E = e_i) * P(E = e_i)$$

These are all terms specified in our local distributions!

Making an inference



$P(E) = 0.7$
$P(\sim E) = 0.3$

$P(C|E) = 0.8$
$P(\sim C|E) = 0.2$
$P(C|\sim E) = 0.5$
$P(\sim C|\sim E) = 0.5$

$P(A|C) = 0.7$
$P(\sim A|C) = 0.3$
$P(A|\sim C) = 0.0$
$P(\sim A|\sim C) = 1.0$

$P(B|A) = 0.2$
$P(\sim B|A) = 0.8$
$P(B|\sim A) = 0.1$
$P(\sim B|\sim A) = 0.9$

$P(H|B) = 0.9$
$P(\sim H|B) = 0.1$
$P(H|\sim B) = 0.1$
$P(\sim H|\sim B) = 0.9$

Computing P(A) in more concrete terms:

$P(C) = P(C|E)P(E) + P(C|\sim E)P(\sim E) = 0.8 * 0.7 + 0.5 * 0.3 = 0.78$

$P(\sim C) = P(\sim C|E)P(E) + P(\sim C|\sim E)P(\sim E) = 0.22$

$P(\sim C) = 1 - P(C)$, as well

$P(A) = P(A|C)P(C) + P(A|\sim C)P(\sim C) = 0.7 * 0.78 + 0.0 * 0.22 = 0.546$

$P(\sim A) = 1 – P(A) = 0.454$

Bayesian Networks

- The structure we just described is a Bayesian Network (BN). A BN is a graphical representation of the direct dependencies over a set of variables, together with a set of conditional probability tables (local conditional joint distributions) that quantify the strength of those influences.
- Bayesian Networks generalize the above ideas in very interesting ways, leading to effective means of representation and inference under uncertainty.

Bayesian Networks

A BN over variables $\{X_1, X_2, \ldots, X_n\}$ consists of:

a directed acyclic graph (DAG) whose nodes are the variables

a set of conditional probability tables (CPTs) that specify $P(X_i | \text{Parents}(X_i))$ for each X_i

Key notions (see text for definitions, all are intuitive):

parents of a node: Parents(X_i)

children of node

descendants of a node

ancestors of a node

family of a node consists of X_i and its parents

CPTs are defined over families in the BN

Another Bayesian Network

M: McIlraith gives the lecture

S: It is sunny out

L: The lecturer arrives late

Assume that all instructors may arrive late in bad weather. Some instructors may be more likely to be late than others.

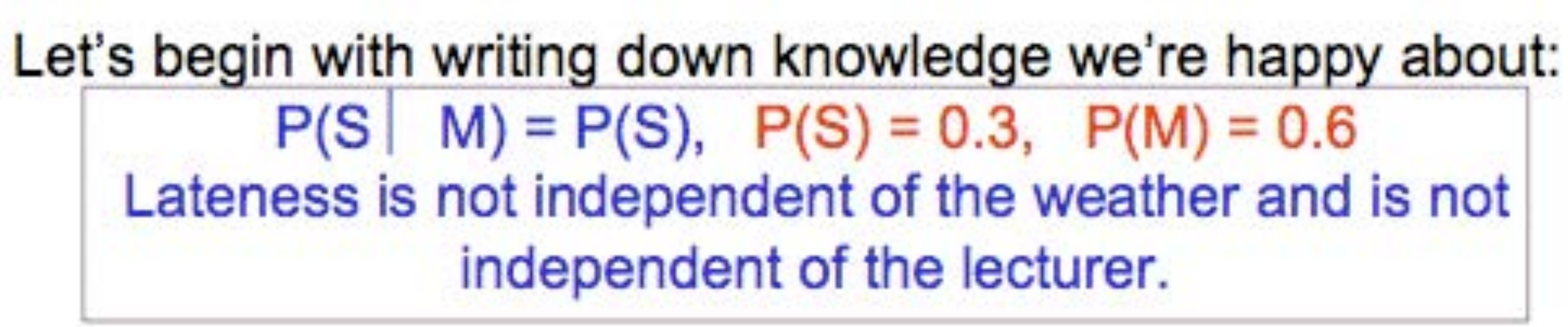
Another Bayesian Network

M: McIlraith gives the lecture

S: It is sunny out

L: The lecturer arrives late

Assume that all instructors may arrive late in bad weather. Some instructors may be more likely to be late than others.



Let's begin with writing down knowledge we're happy about:

$P(S \mid M) = P(S)$, $P(S) = 0.3$, $P(M) = 0.6$

Lateness is not independent of the weather and is not independent of the lecturer.

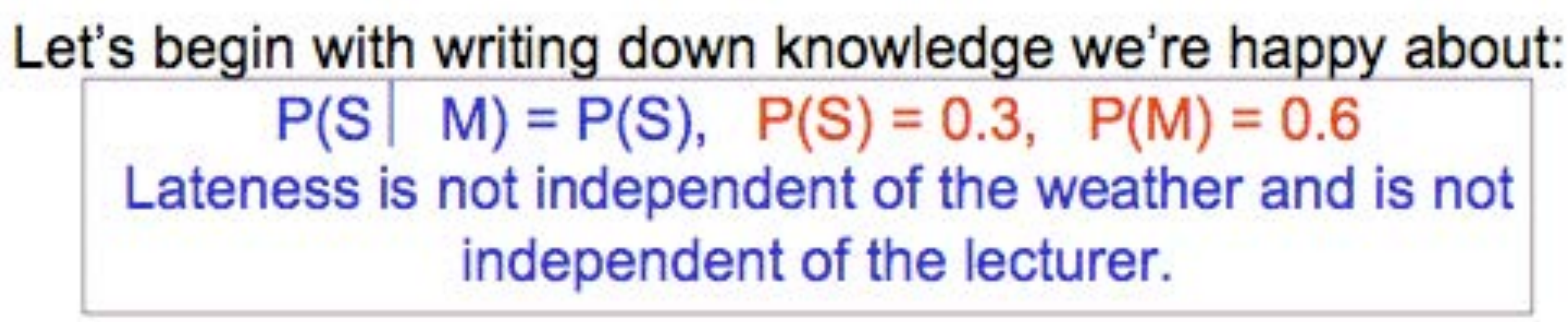
Another Bayesian Network

M: McIlraith gives the lecture

S: It is sunny out

L: The lecturer arrives late

Assume that all instructors may arrive late in bad weather. Some instructors may be more likely to be late than others.



Let's begin with writing down knowledge we're happy about:

$P(S| M) = P(S)$, $P(S) = 0.3$, $P(M) = 0.6$

Lateness is not independent of the weather and is not independent of the lecturer.

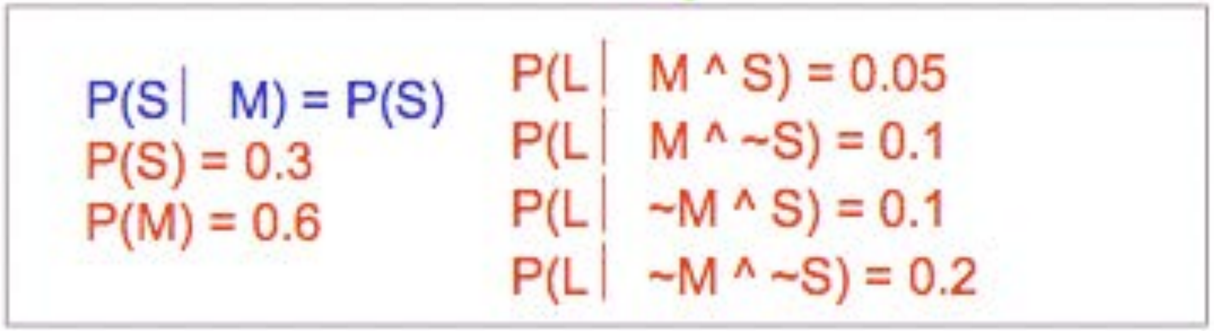
We need to formulate $P(L|S,M)$ for all of the values of S and M.

Another Bayesian Network

M: McIlraith gives the lecture

S: It is sunny out

L: The lecturer arrives late



P(S | M) = P(S)
P(S) = 0.3
P(M) = 0.6

P(L | M ^ S) = 0.05
P(L | M ^ ~S) = 0.1
P(L | ~M ^ S) = 0.1
P(L | ~M ^ ~S) = 0.2

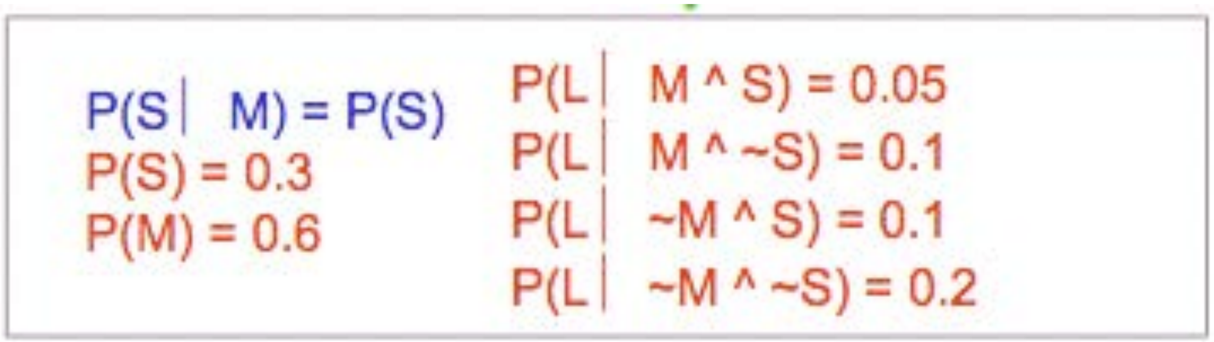
Because of conditional independence, we only need 6 values in the joint instead of 7. Conditional independence leads to computational savings!

Another Bayesian Network

M: McIlraith gives the lecture

S: It is sunny out

L: The lecturer arrives late



$P(S \mid M) = P(S)$

$P(S) = 0.3$

$P(M) = 0.6$

$P(L \mid M \wedge S) = 0.05$

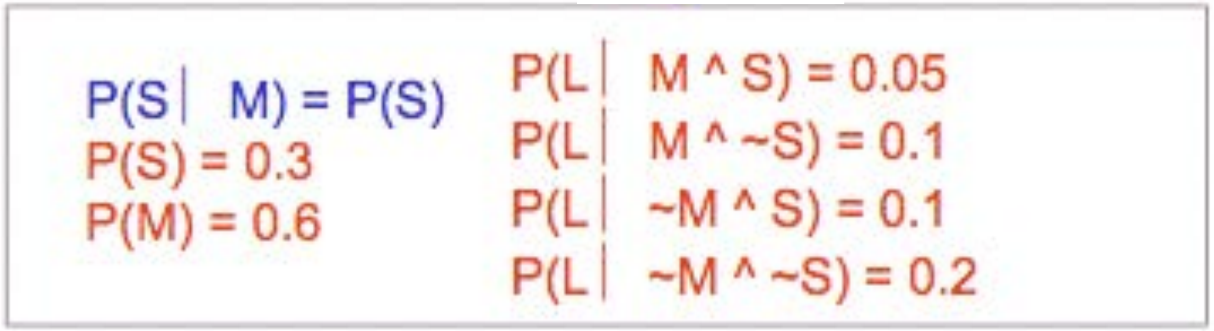
$P(L \mid M \wedge \sim S) = 0.1$

$P(L \mid \sim M \wedge S) = 0.1$

$P(L \mid \sim M \wedge \sim S) = 0.2$

How can we calculate P(L,M,S)? Or P(L,~M,S)?

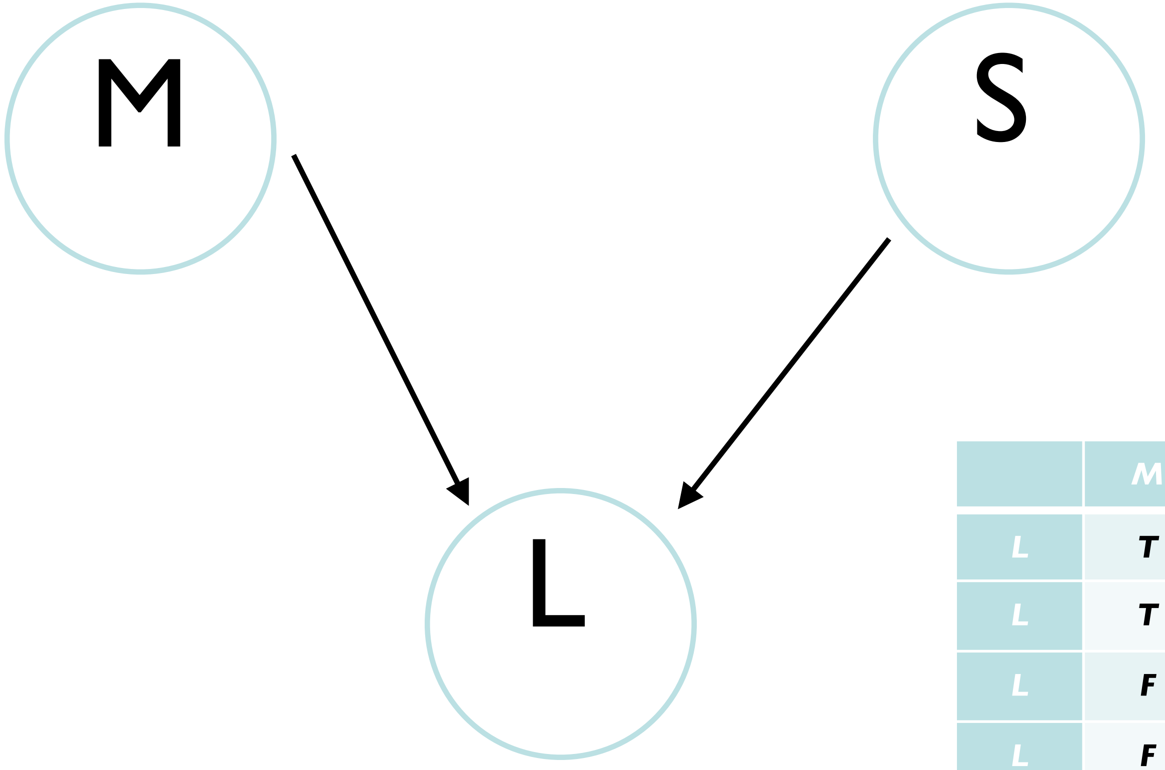
Drawing the Network



$P(S | M) = P(S)$
$P(S) = 0.3$
$P(M) = 0.6$

$P(L | M \wedge S) = 0.05$
$P(L | M \wedge \sim S) = 0.1$
$P(L | \sim M \wedge S) = 0.1$
$P(L | \sim M \wedge \sim S) = 0.2$

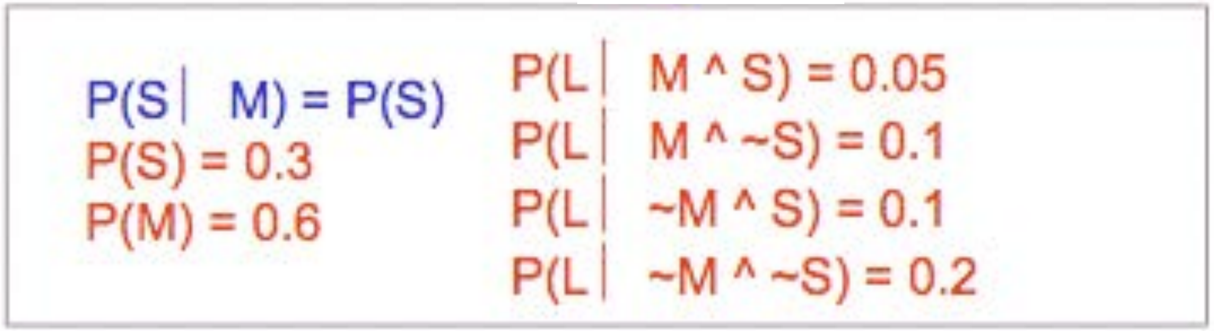
S	T	0.6



S	T	0.3

	M	S	P(L=T|M,S)
L	T	T	0.05
L	T	F	0.1
L	F	T	0.1
L	F	F	0.2

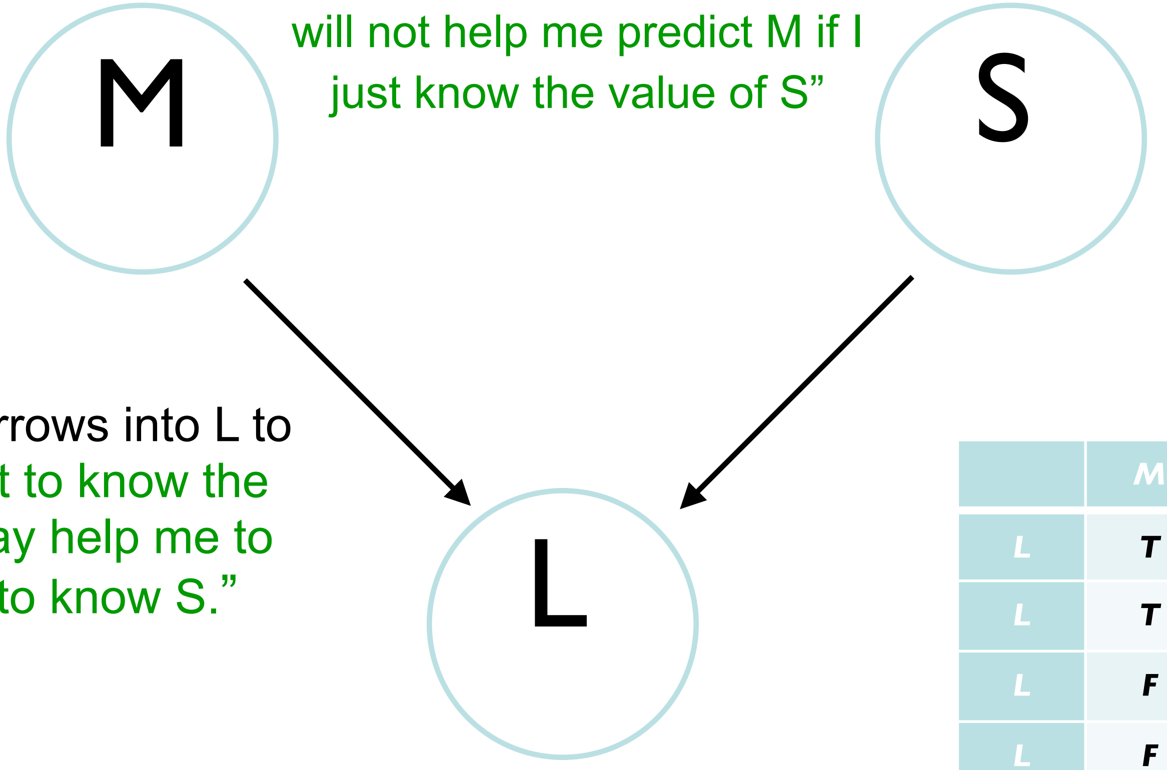
Drawing the Network



P(S| M) = P(S)
P(S) = 0.3
P(M) = 0.6

P(L| M ^ S) = 0.05
P(L| M ^ ~S) = 0.1
P(L| ~M ^ S) = 0.1
P(L| ~M ^ ~S) = 0.2

S	T	0.6



Read the absence of an arrow between S and M to mean "It will not help me predict M if I just know the value of S"

S	T	0.3

Read the two arrows into L to mean "If I want to know the value of L it may help me to know M and to know S."

	M	S	P(L=T|M,S)
L	T	T	0.05
L	T	F	0.1
L	F	T	0.1
L	F	F	0.2

Back to the network

Now let's suppose we have these three events:

M: McIlraith gives the lecture

L: The lecturer arrives late

R : The lecturer concerns Reasoning with Bayes' Nets

And we know:

- Allin has a higher chance of being late than McIlraith.
- Allin has a higher chance of giving lectures about reasoning with BNs

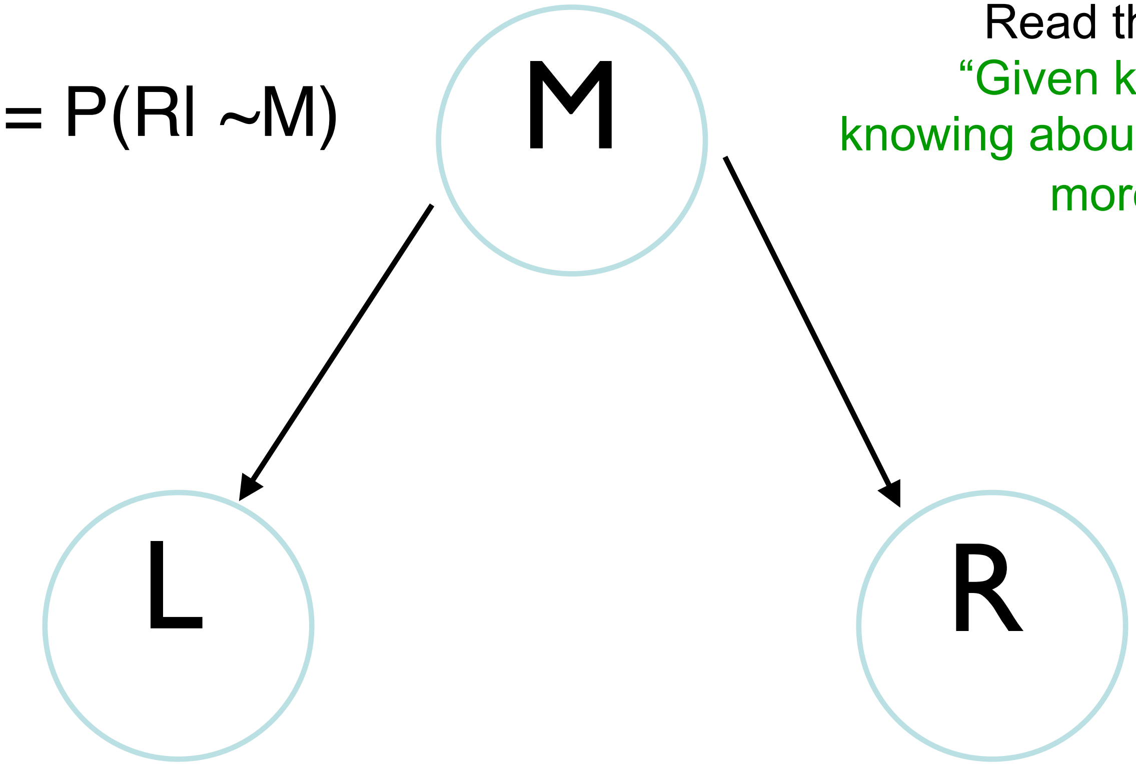
What kind of independences exist in our graph?

Back to the network

Once you know who the lecturer is, then whether they arrive late doesn't affect whether the lecture concerns Reasoning with Bayes' Nets, i.e.:

$P(R|M,L) = P(R|M)$

$P(R| \sim M,L) = P(R| \sim M)$

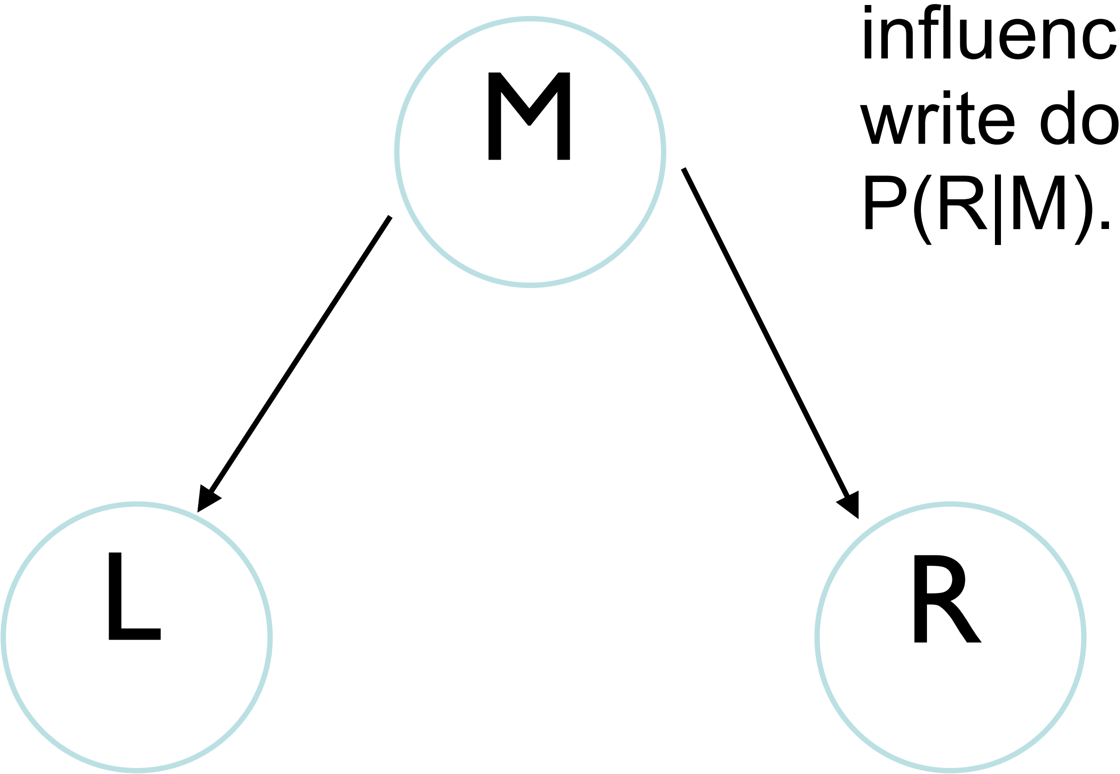


Read this diagram as
"Given knowledge of M, knowing about L won't tell anything more about R."

The network reflects conditional independences

To specify CPTs, we first write down P(M). Then, because L and R are only directly influenced by M, we write down P(L|M) and P(R|M).

M	T	0.6



	M	P(L=T|M)
L	T	0.085
L	F	0.17

	M	P(R=T|M)
R	T	0.3
R	F	0.6

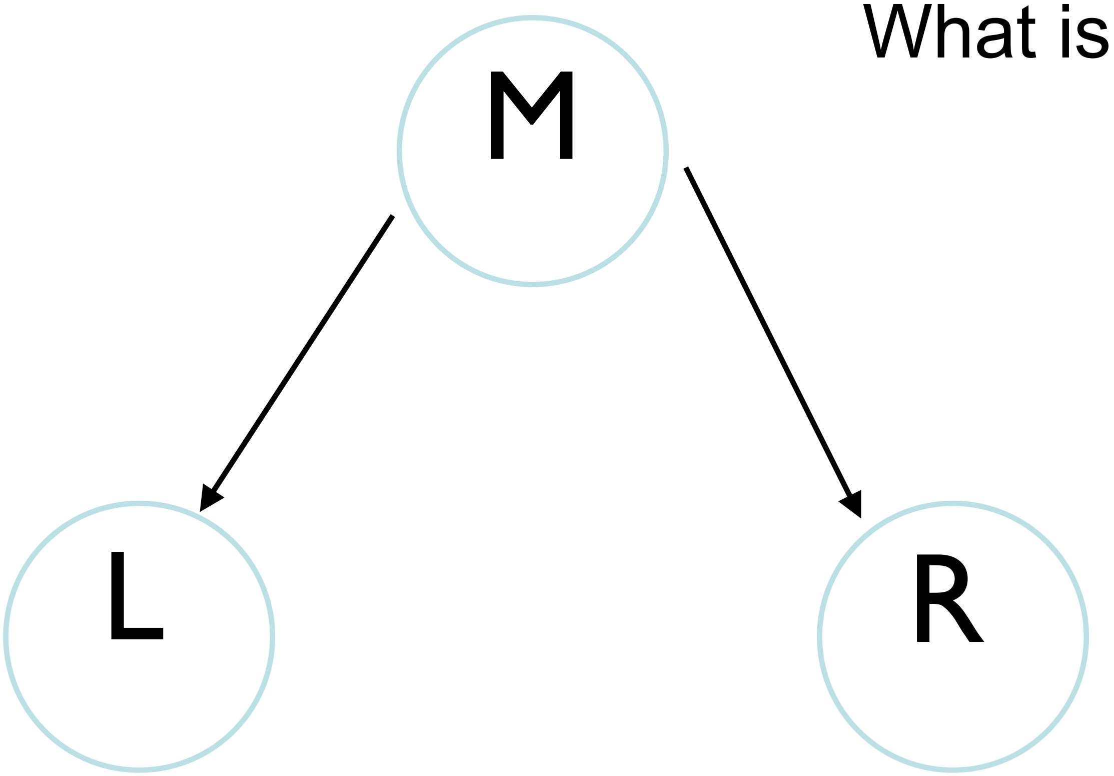
R is conditionally independent of L given M (and vice versa)

The network reflects conditional independences

What is P(L,R,M)?

What is P(~L,R,M)?

M	T	0.6



	M	$P(L=T \mid M)$
L	T	0.085
L	F	0.17

	M	$P(R=T \mid M)$
R	T	0.3
R	F	0.6

R is conditionally independent of L given M (and vice versa)

Building a Bayes Net

M: McIlraith gives the lecture

L: The lecturer arrives late

R : The lecturer concerns Reasoning with Bayes' Nets

S: It is sunny out

T: The lecture starts by 1:15 (or 3:15 if you're in the later section)

- T is only directly influenced by L (i.e. T is conditionally independent of R,M,S given L)
- L is only directly influenced by M and S (i.e. L is conditionally independent of R given M & S)
- R is only directly influenced by M (i.e. R is conditionally independent of L,S, given M)
- M and S are independent

Building a Bayes Net

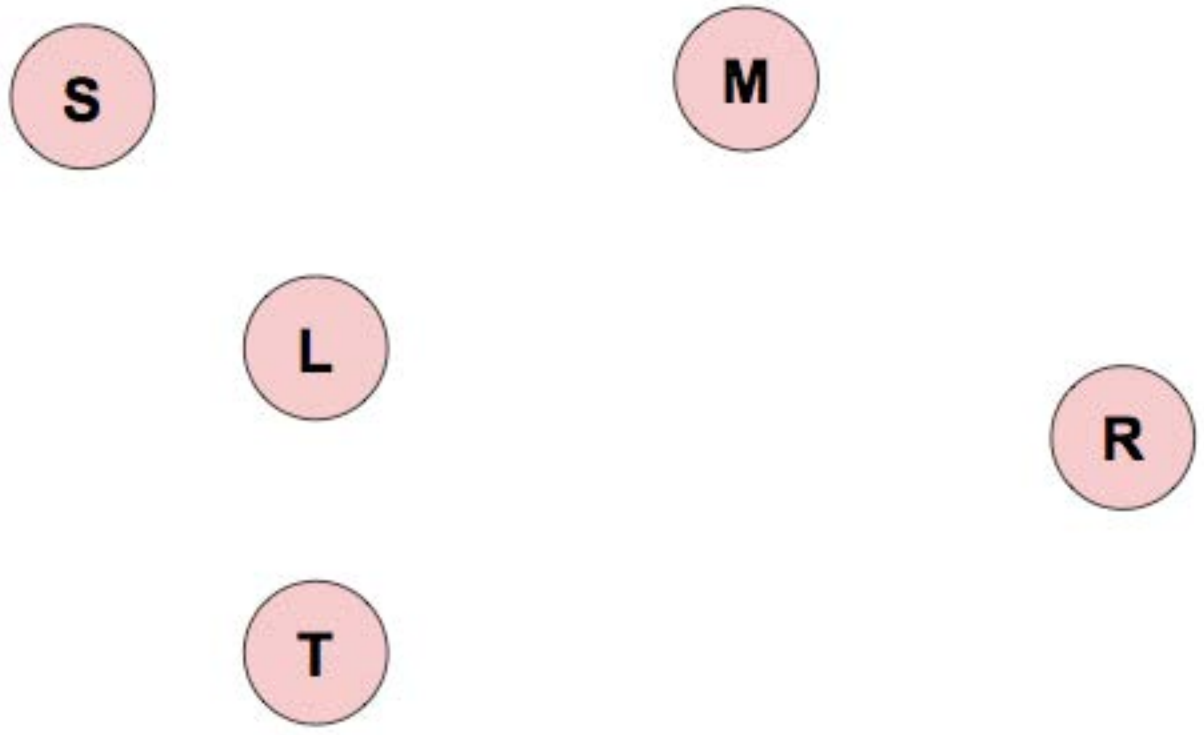
M: McIlraith gives the lecture

L: The lecturer arrives late

R : The lecturer concerns Reasoning with Bayes' Nets

S: It is sunny out

T: The lecture starts by 1:15 (or 3:15)



Step One: Add variables

Building a Bayes Net

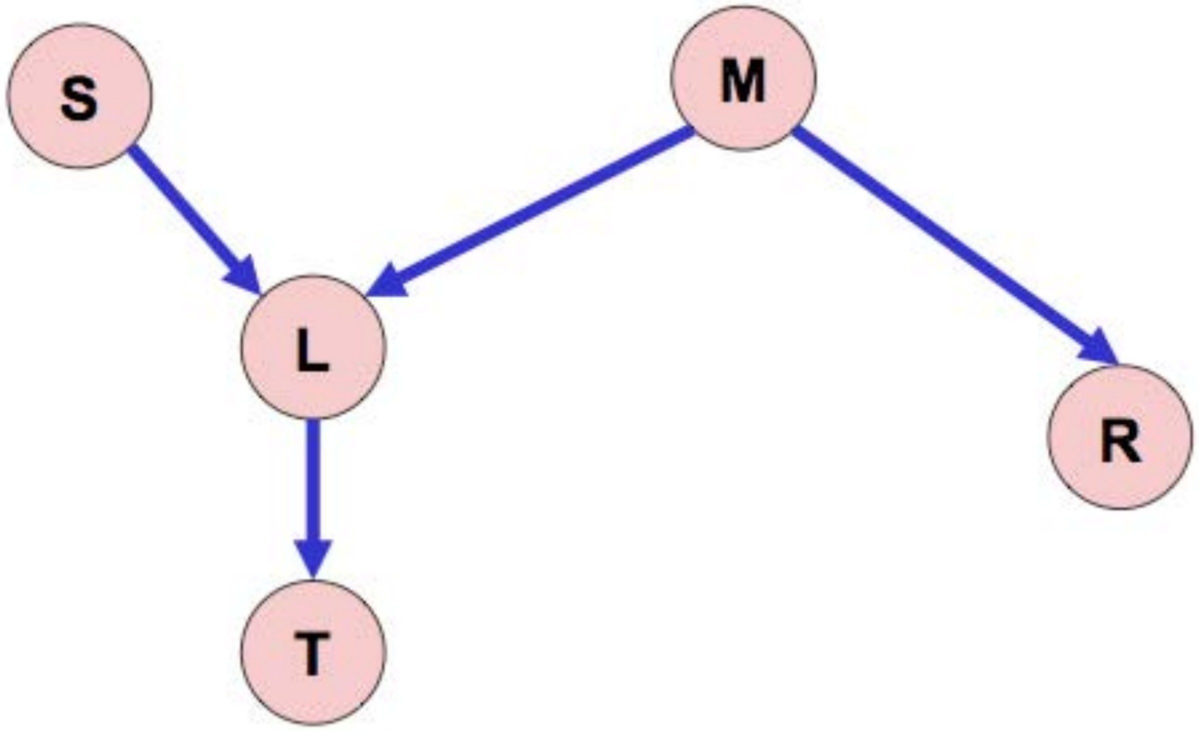
M: McIlraith gives the lecture

L: The lecturer arrives late

R : The lecturer concerns Reasoning with Bayes' Nets

S: It is sunny out

T: The lecture starts by 1:15 (or 3:15)

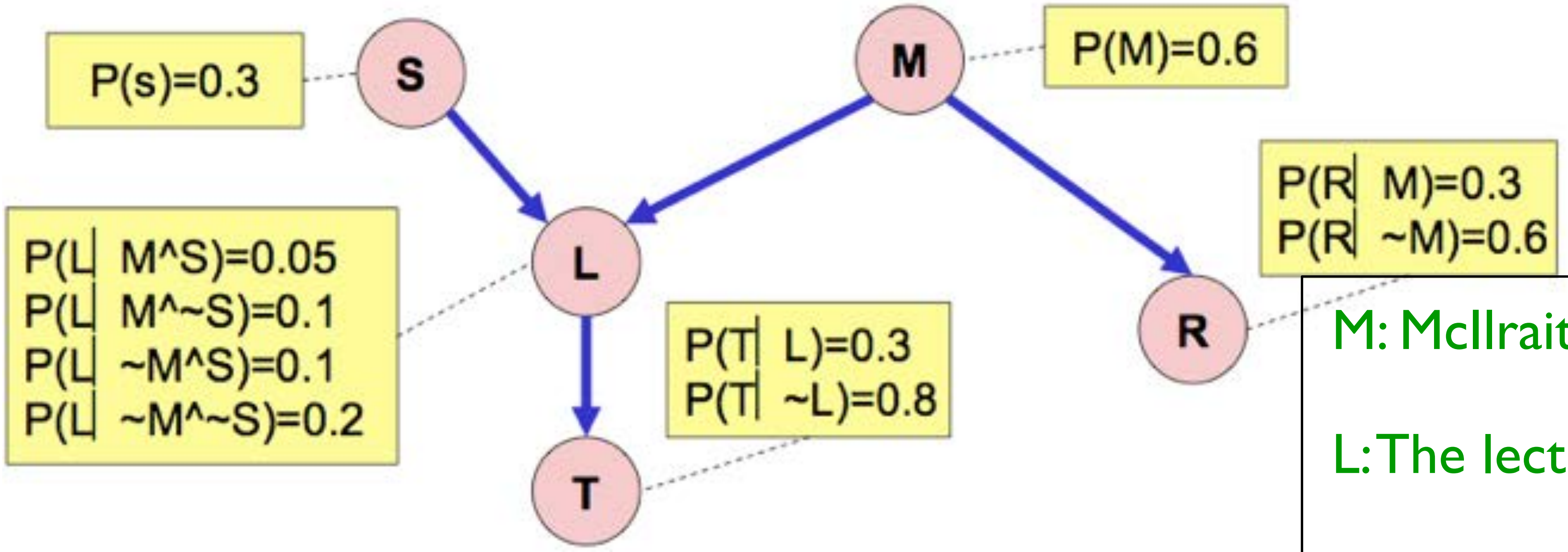


Step Two: add links.

The link structure must be acyclic.

If you assign node Y the parents $X_1, X_2, \ldots, X_n$, you are promising that, given $\{X_1, X_2, \ldots, X_n\}$, Y is conditionally independent of any other variable that's not a descendent of Y

Building a Bayes Net



M: McIlraith gives the lecture

L: The lecturer arrives late

R : The lecturer concerns Reasoning with Bayes' Nets

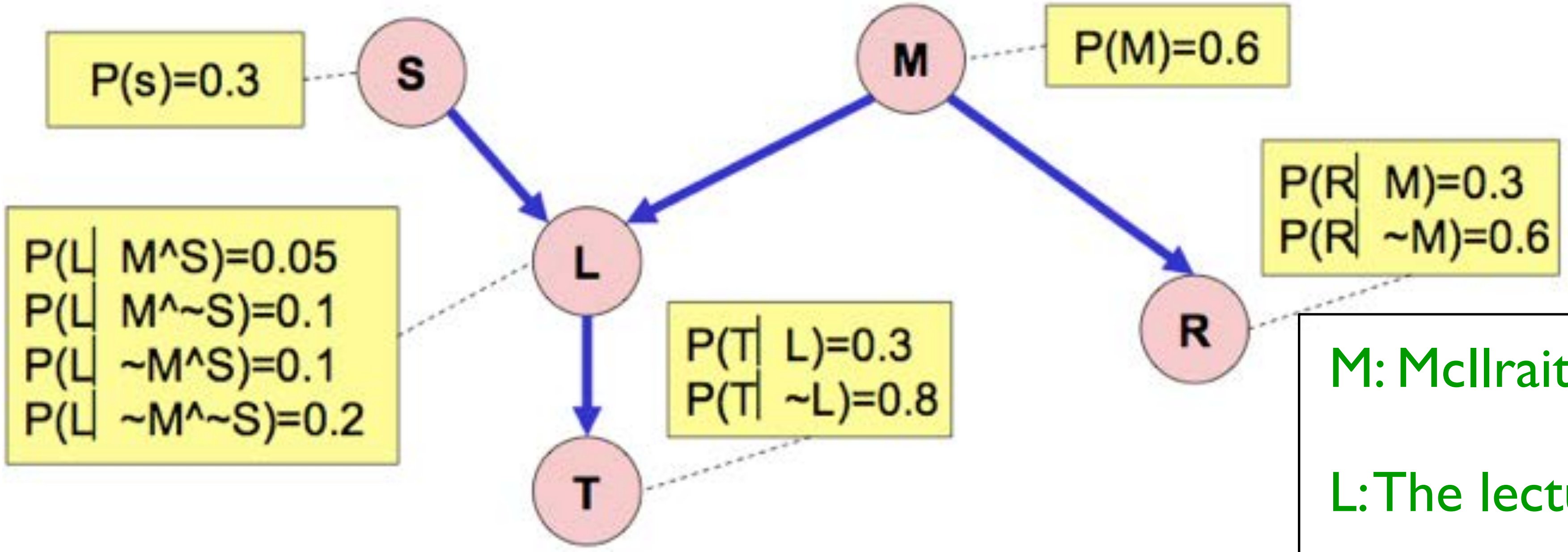
S: It is sunny out

T: The lecture starts by 1:15 (or 3:15)

Step Three: add a conditional probability table (CPT) for each node.

The table for X must define P(X|Parents) and for all combinations of the possible parent values.

Building a Bayes Net



M: McIlraith gives the lecture

L: The lecturer arrives late

R : The lecturer concerns Reasoning with Bayes' Nets

S: It is sunny out

T: The lecture starts by 1:15 (or 3:15)

You can deduce many probability relations from a Bayes Net.

Note that variables that are not directly connected may still be correlated.

Building a Bayes Net

It is always possible to construct a Bayes net to represent any distribution over the variables X_1, X_2,…, X_n, using any ordering of the variables.

Take any ordering of the variables (say, the order given). From the chain rule we obtain.

$$\Pr(X_1,\ldots,X_n) = \Pr(X_n|X_1,\ldots,X_{n-1})\Pr(X_{n-1}|X_1,\ldots,X_{n-2})\ldots\Pr(X_1)$$

Now for each X_i go through its conditioning set $X_1,\ldots,X_{i-1}$, and iteratively remove all variables X_j such that X_i is conditionally independent of X_j given the remaining variables. Do this until no more variables can be removed.

The final product specifies a Bayes net.

Causal Intuitions

- The BN can be constructed using an arbitrary ordering of the variables.
- However, some orderings will yield BN's with very large parent sets. This requires exponential space, and (as we will see later) exponential time to perform inference.
- Empirically, and conceptually, a good way to construct a BN is to use an ordering based on causality. This often yields a more natural and compact BN.

Causal Intuitions

Malaria, the flu and a cold all "cause" aches. So use the ordering that places causes before effects. Variables are Malaria (M), Flu (F), Cold (C), Aches (A):

$$P(M,F,C,A) = P(A|M,F,C)\ P(C|M,F)\ P(F|M)\ Pr(M)$$

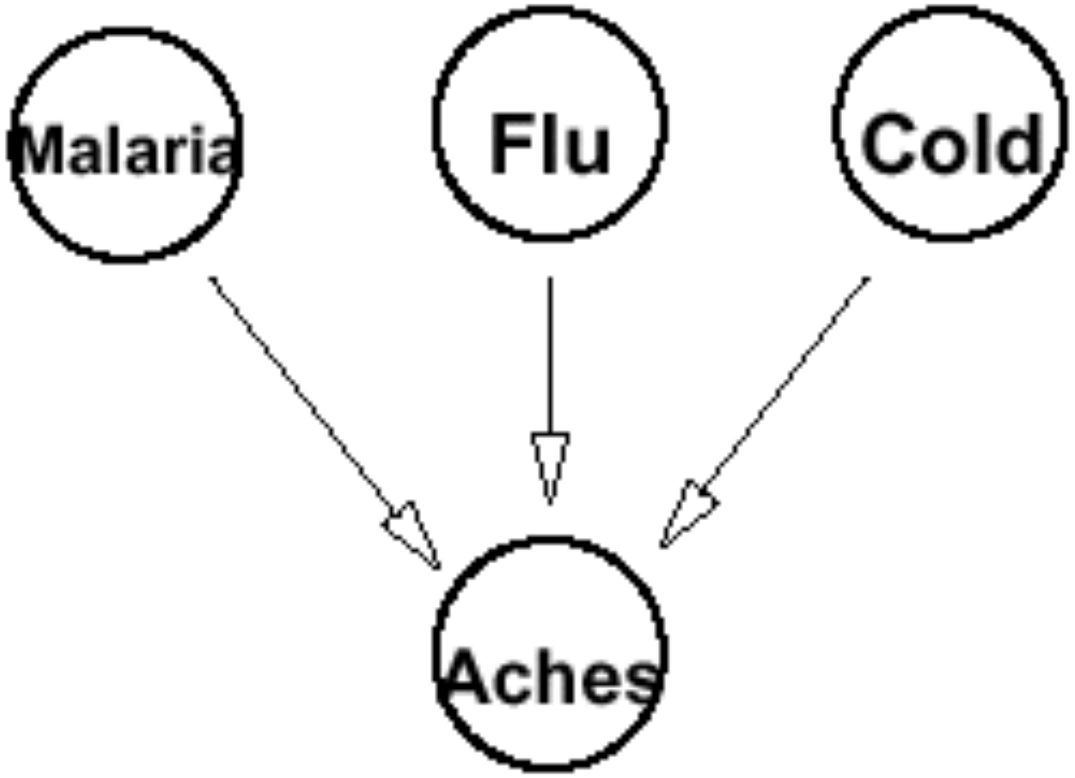
Each of these disease affects the probability of aches, so the first conditional probability does not change.

It is however reasonable to assume that these diseases are independent of each other: having or not having one does not change the probability of having the others. So $P(C|M,F) = P(C)$ and $P(F|M) = P(F)$

Causal Intuitions

This yields a fairly simple Bayes net.

We only need one big CPT, involving the family of “Aches”.



Causal Intuitions

Suppose we build the BN for distribution P using the opposite ordering, i.e., we use ordering Aches, Cold, Flu, Malaria

$$P(A,C,F,M) = P(M|A,C,F)\ P(F|A,C)\ P(C|A)\ P(A)$$

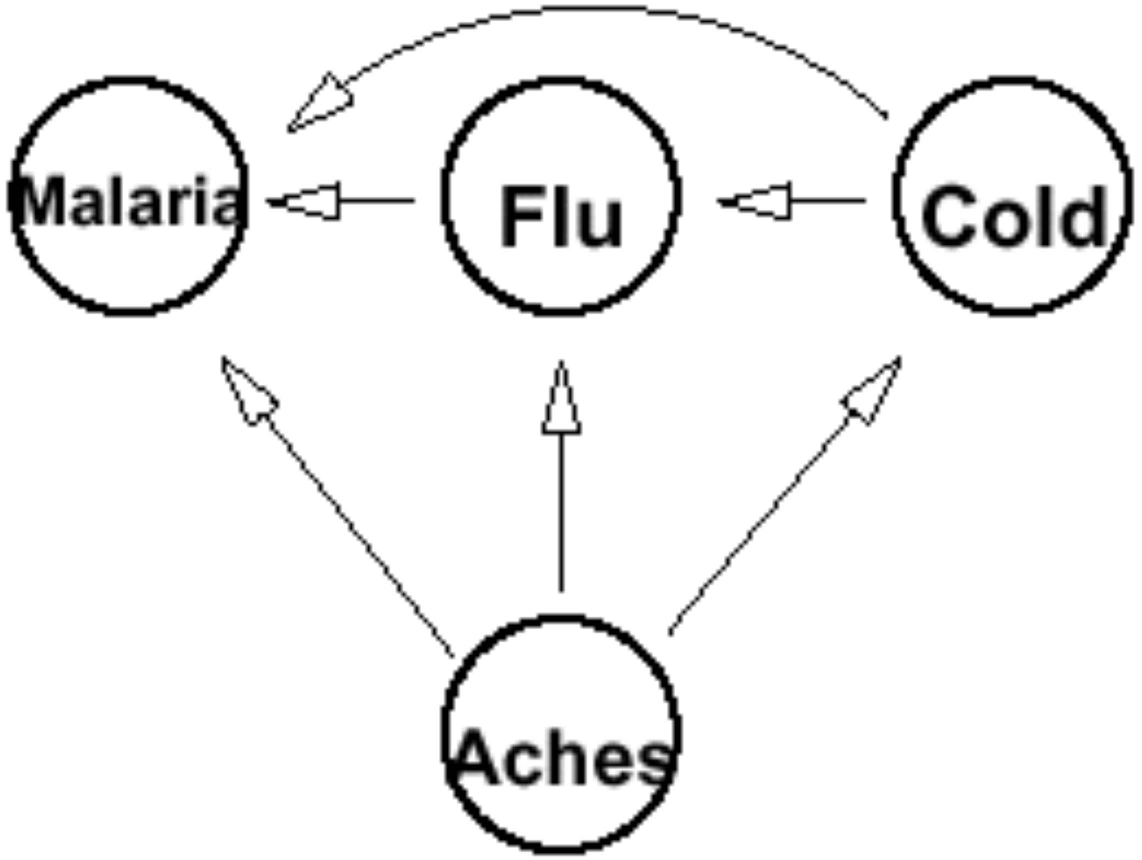
We can't reduce P(M|A,C,F). The probability of Malaria is clearly affected by knowing Aches. What about knowing Aches and Cold, or Aches and Cold and Flu?

Probability of Malaria is affected by both of these additional pieces of knowledge.

Knowing Cold and Flu lowers the probability that Aches are related to Malaria since they "explain away" the Aches!

Causal Intuitions

We obtain a much more complex Bayes net. In fact, we obtain no savings over explicitly representing the full joint distribution (i.e., representing the probability of every atomic event).



Bayes Net Example

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

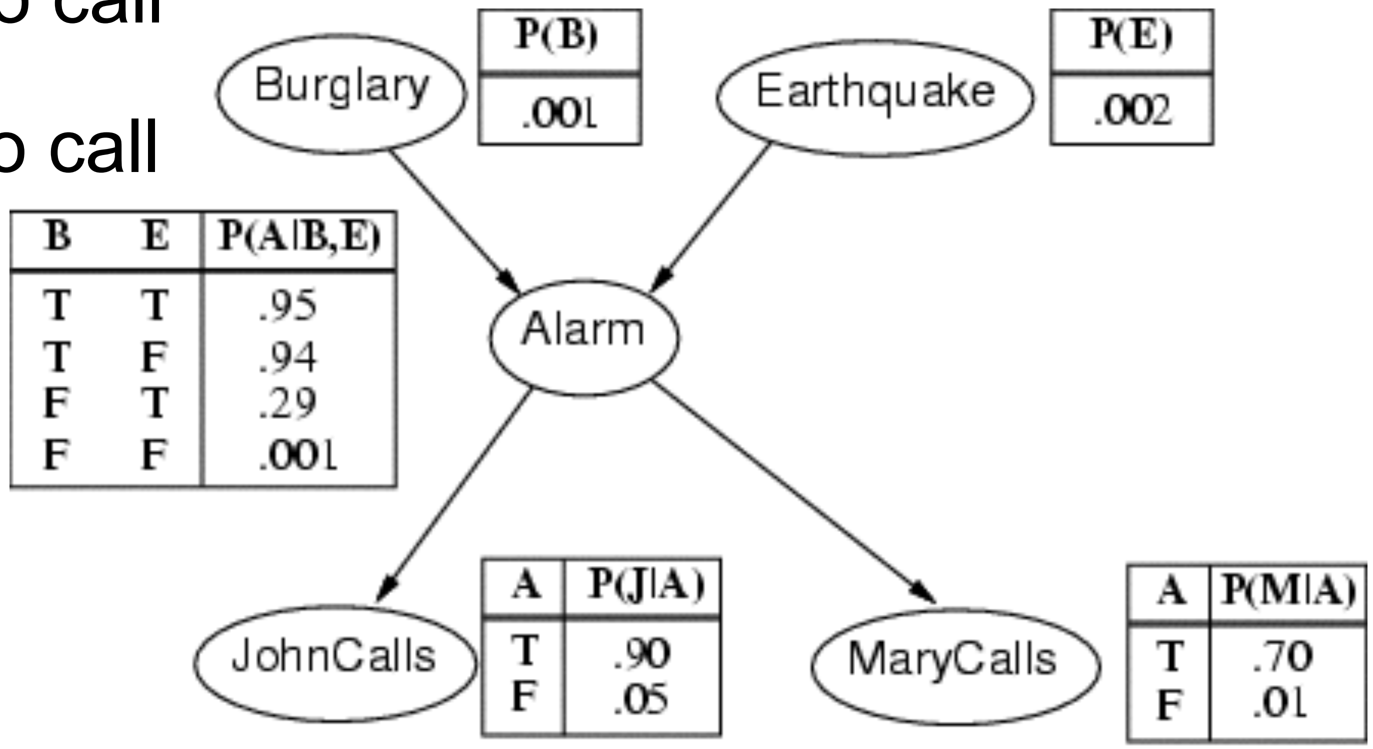
Variables: *Burglary*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*

The 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

Burglary Example

- 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



P(B)
.001

P(E)
.002

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

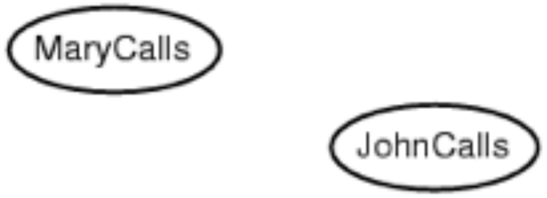
A	P(J|A)
T	.90
F	.05

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

# of Params:1 + 1 + 4 + 2 + 2 = 10 (vs. 2^5-1 = 31)

Burglary Example

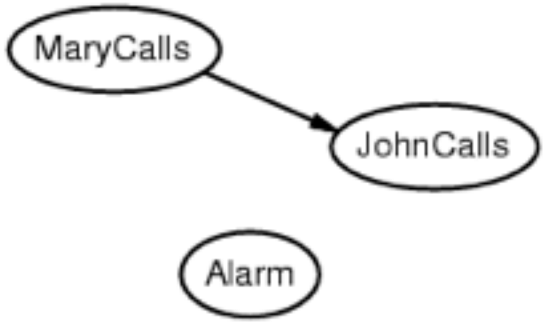
Suppose we choose the ordering M, J, A, B, E



P(J | M) = P(J)?

Burglary Example

Suppose we choose the ordering M, J, A, B, E

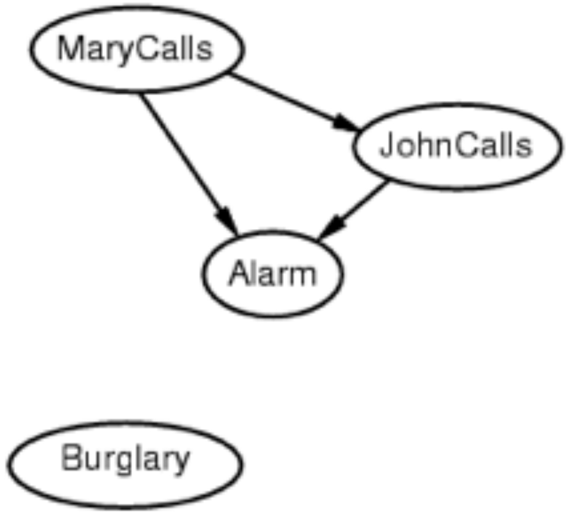


P(J | M) = P(J)? No

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

Burglary Example

Suppose we choose the ordering M, J, A, B, E



P(J | M) = P(J)? No

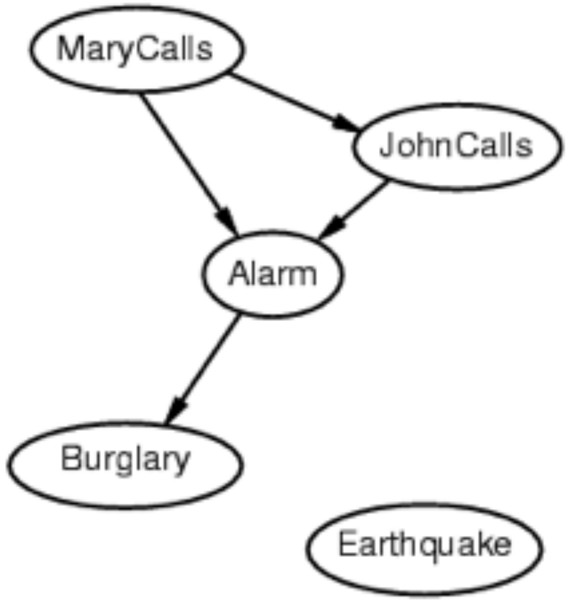
P(A | J, M) = P(A | J)? P(A | J, M) = P(A)? No

P(B | A, J, M) = P(B | A)?

P(B | A, J, M) = P(B)?

Burglary Example

Suppose we choose the ordering M, J, A, B, E



P(J | M) = P(J)? No

P(A | J, M) = P(A | J)? P(A | J, M) = P(A)? No

P(B | A, J, M) = P(B | A)? Yes

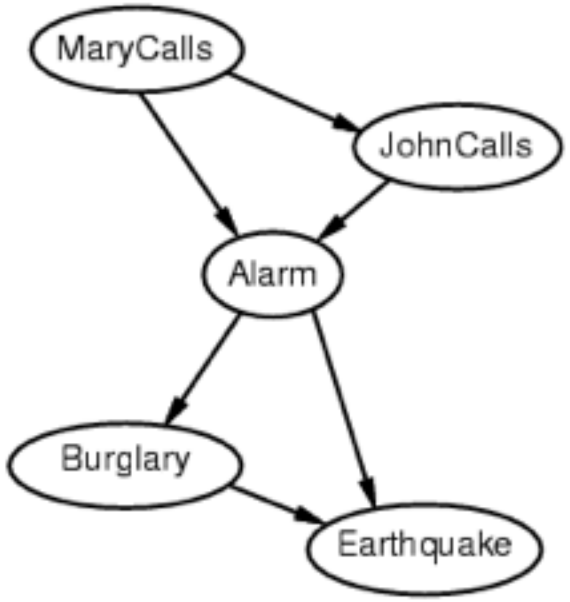
P(B | A, J, M) = P(B)? No

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

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

Burglary Example

Suppose we choose the ordering M, J, A, B, E



P(J | M) = P(J)? No

P(A | J, M) = P(A | J)? P(A | J, M) = P(A)? No

P(B | A, J, M) = P(B | A)? Yes

P(B | A, J, M) = P(B)? No

P(E | B, A ,J, M) = P(E | A)? No

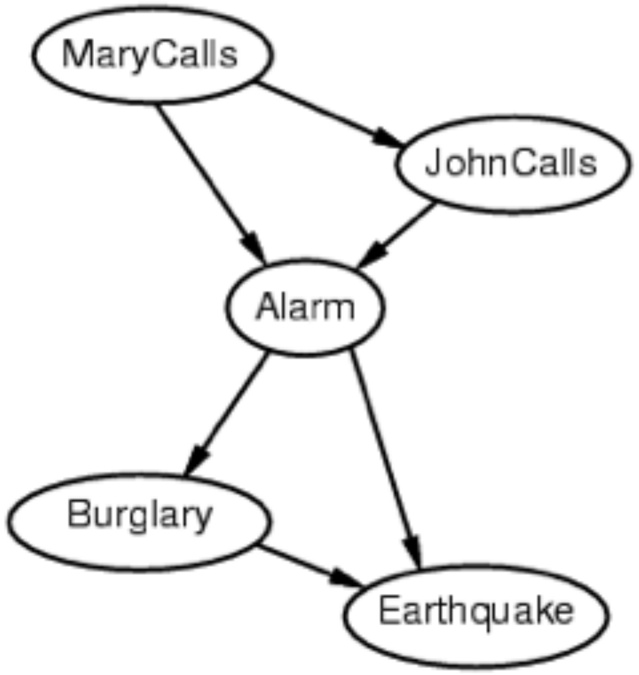
P(E | B, A, J, M) = P(E | A, B)? Yes

Burglary Example

Deciding conditional independence **is hard** in non-causal directions!

(Causal models and conditional independence seem hardwired in humans!)

Network is **less compact**: 1 + 2 + 4 + 2 + 4 = 13 numbers needed


MaryCalls
JohnCalls
Alarm
Burglary
Earthquake