



# Bayesian networks

Introduction to Artificial Intelligence  
CSCI561a  
Hadi Moradi

AIMA Chapter 14

(some slides are from Prof. A. Moore, S. Davis from CMU and Min-Yen Kan and may be updated)



# Outline

- Syntax
- Semantics



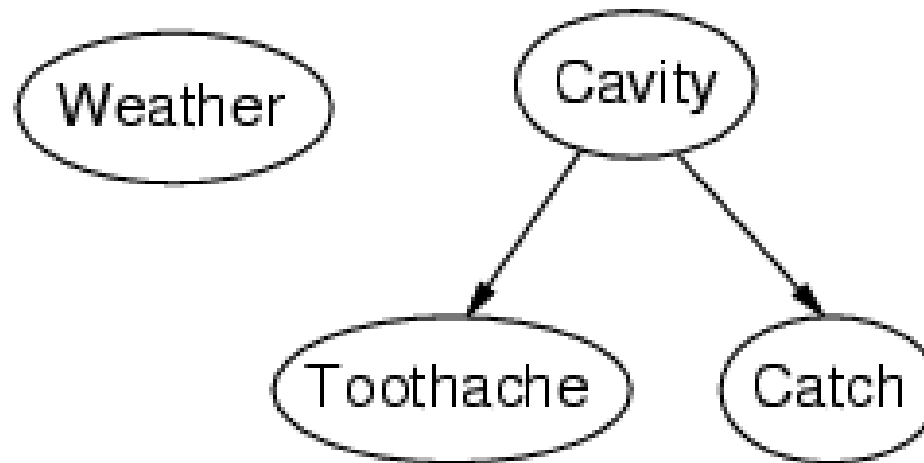
# Bayesian networks

- A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions
- Syntax:
  - a set of nodes, one per variable
  - a directed, acyclic graph (link  $\approx$  "directly influences")
  - a conditional distribution for each node given its parents:
$$P(X_i \mid \text{Parents}(X_i))$$
- In the simplest case, conditional distribution represented as a **conditional probability table** (CPT) giving the distribution over  $X_i$  for each combination of parent values



# Example

- Topology of network encodes conditional independence assertions:



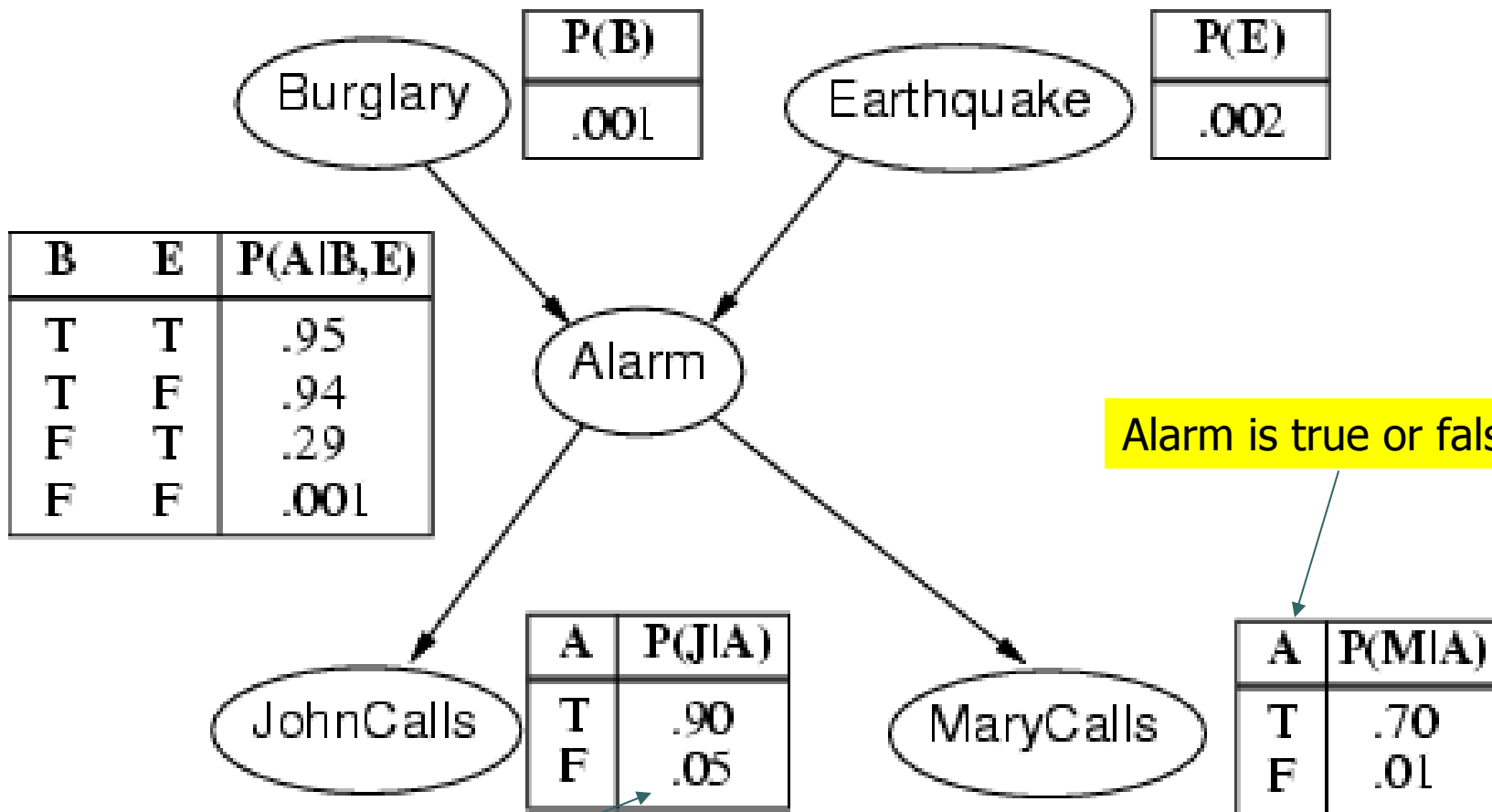
- *Weather* is independent of the other variables
- *Toothache* and *Catch* are conditionally independent given *Cavity*



# Example

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



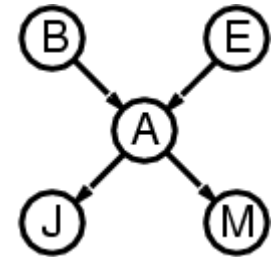
Alarm is true or false

Each row should sum up to one (not showing the opposite case)



# Compactness

- A CPT for 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  $O(n \cdot 2^k)$  numbers
  - I.e., grows linearly with  $n$ , vs.  $O(2^n)$  for the full joint distribution
- For burglary net,  $1 + 1 + 4 + 2 + 2 = 10$  numbers (vs.  $2^5 - 1 = 31$ ) (Previous slide)

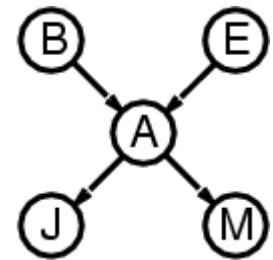




# Semantics

The full joint distribution is defined as the product of the local conditional distributions:

$$P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i | \text{Parents}(X_i))$$



e.g.,  $P(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$   
 $= P(j | a) P(m | a) P(a | \neg b, \neg e) P(\neg b) P(\neg e)$   
 $= 0.90 \times 0.70 \times 0.001 \times 0.999 \times 0.998 = 0.00062$



# Constructing Bayesian networks

- Choose an ordering of variables  $X_1, \dots, X_n$
- For  $i = 1$  to  $n$ 
  - add  $X_i$  to the network
  - select parents from  $X_1, \dots, X_{i-1}$  such that

$$P(X_i | \text{Parents}(X_i)) = P(X_i | X_1, \dots, X_{i-1})$$

This choice of parents guarantees:

$$\begin{aligned} P(X_1, \dots, X_n) &= \prod_{i=1}^n P(X_i | X_1, \dots, X_{i-1}) && \text{(chain rule)} \\ &= \prod_{i=1}^n P(X_i | \text{Parents}(X_i)) && \text{(by construction)} \end{aligned}$$

Note: Parent of  $X_i$  is among  $X_1$  to  $X_{i-1}$ .

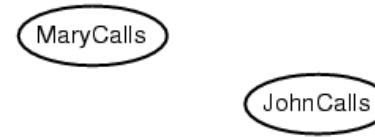


# Example

For  $i = 1$  to  $n$

- add  $X_i$  to the network
- select parents from  $X_1, \dots, X_{i-1}$  such that  $P(X_i | \text{Parents}(X_i)) = P(X_i | X_1, \dots, X_{i-1})$

- o Suppose we choose the ordering  $M, J, A, B, E$



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

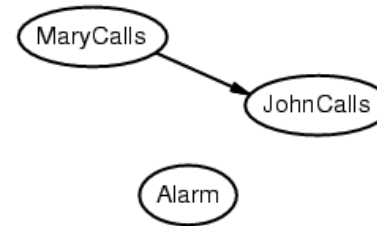


# Example

For  $i = 1$  to  $n$

- add  $X_i$  to the network
- select parents from  $X_1, \dots, X_{i-1}$  such that  $P(X_i | \text{Parents}(X_i)) = P(X_i | X_1, \dots, X_{i-1})$

- o Suppose we choose the ordering  $M, J, A, B, E$



$P(J | M) = P(J)$ ? No: If Marry calls, then John may call

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

Chain rule:

$$P(x_n, x_{n-1}, \dots, x_1) = P(x_n | x_{n-1}, \dots, x_1) P(x_{n-1} | x_{n-2}, \dots, x_1) \dots P(x_2 | x_1) P(x_1)$$

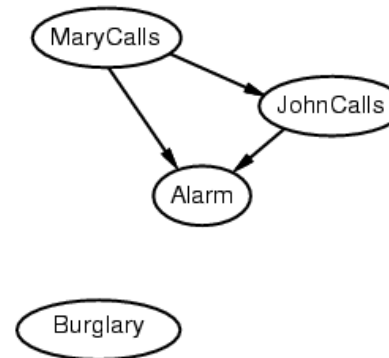


# Example

For  $i = 1$  to  $n$

- add  $X_i$  to the network
- select parents from  $X_1, \dots, X_{i-1}$  such that  $P(X_i | \text{Parents}(X_i)) = P(X_i | X_1, \dots, X_{i-1})$

- o 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**

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

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

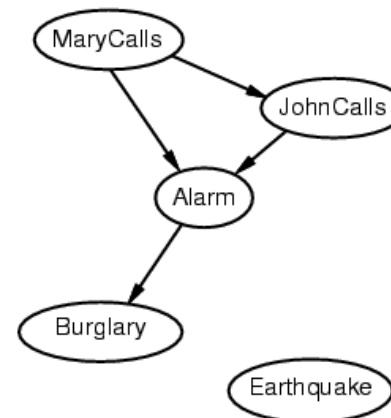


# Example

For  $i = 1$  to  $n$

- add  $X_i$  to the network
- select parents from  $X_1, \dots, X_{i-1}$  such that  $P(X_i | \text{Parents}(X_i)) = P(X_i | X_1, \dots, X_{i-1})$

- o 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**

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

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

Alarm state is enough to tell us about Burglary.

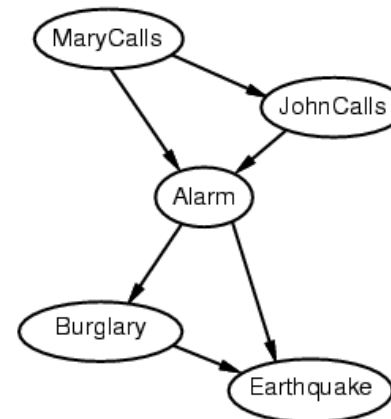


# Example

For  $i = 1$  to  $n$

- add  $X_i$  to the network
- select parents from  $X_1, \dots, X_{i-1}$  such that  $P(X_i | \text{Parents}(X_i)) = P(X_i | X_1, \dots, X_{i-1})$

- o 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**

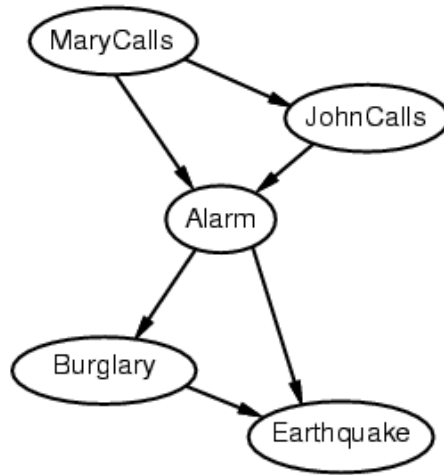
$P(B | A, J, M) = P(B)$ ? **No**

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

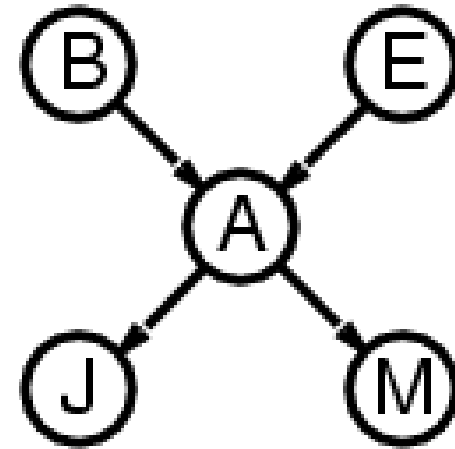
$P(E | B, A, J, M) = P(E | A, B)$ ? **Yes**



# Example contd.



New Bayesian Network



Original Bayesian Network

- Deciding conditional independence is hard in non-causal directions
  - Causal models and conditional independence seem hardwired for humans!
- Network is less compact:  $1 + 2 + 4 + 2 + 4 = 13$  numbers needed

- ● ● | Example (Previous lecture): Assume five variables

T: The lecture started by 10:35

L: The lecturer arrives late

R: The lecture concerns robots

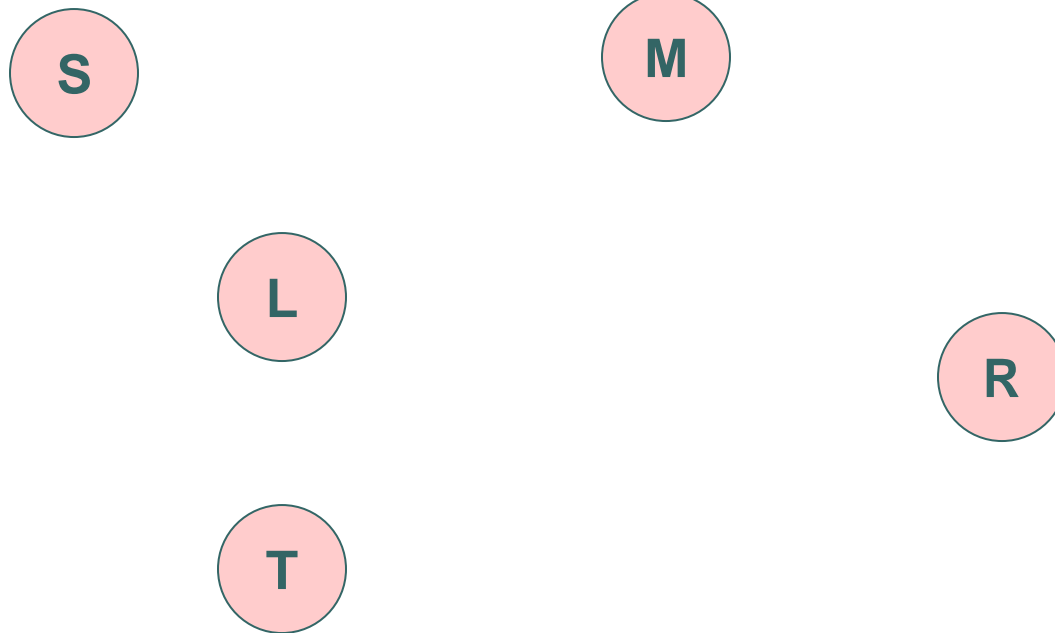
M: The lecturer is Manuela

S: It is sunny

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

# Making a Bayes net

T: The lecture started by 10:35  
L: The lecturer arrives late  
R: The lecture concerns robots  
M: The lecturer is Manuela  
S: It is sunny

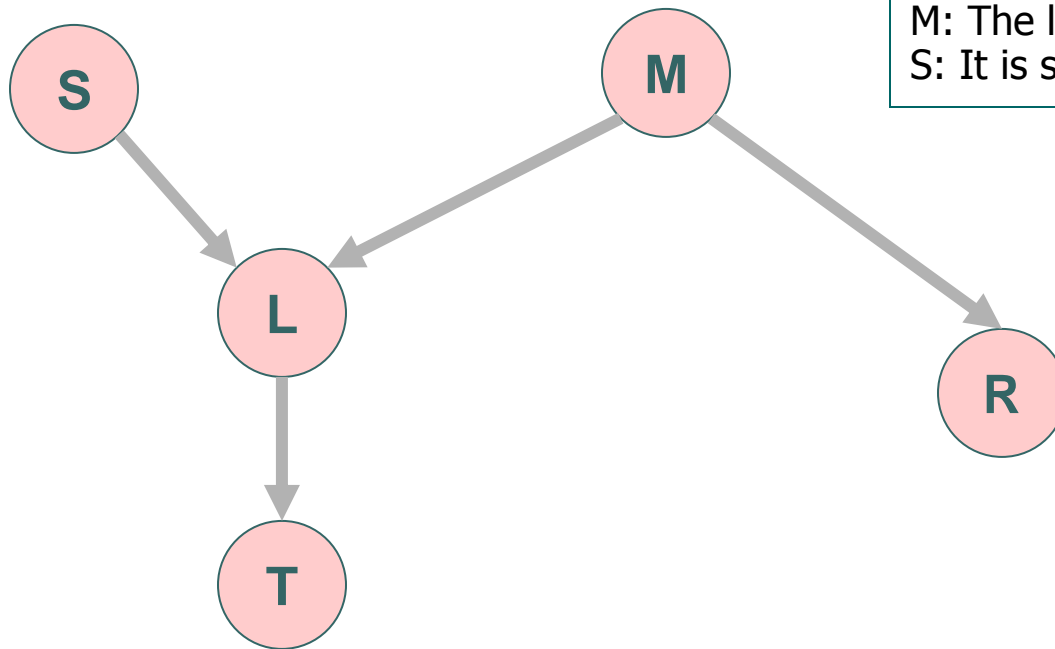


Step One: add variables.

- Just choose the variables you'd like to be included in the net.

# Making a Bayes net

T: The lecture started by 10:35  
L: The lecturer arrives late  
R: The lecture concerns robots  
M: The lecturer is Manuela  
S: It is sunny

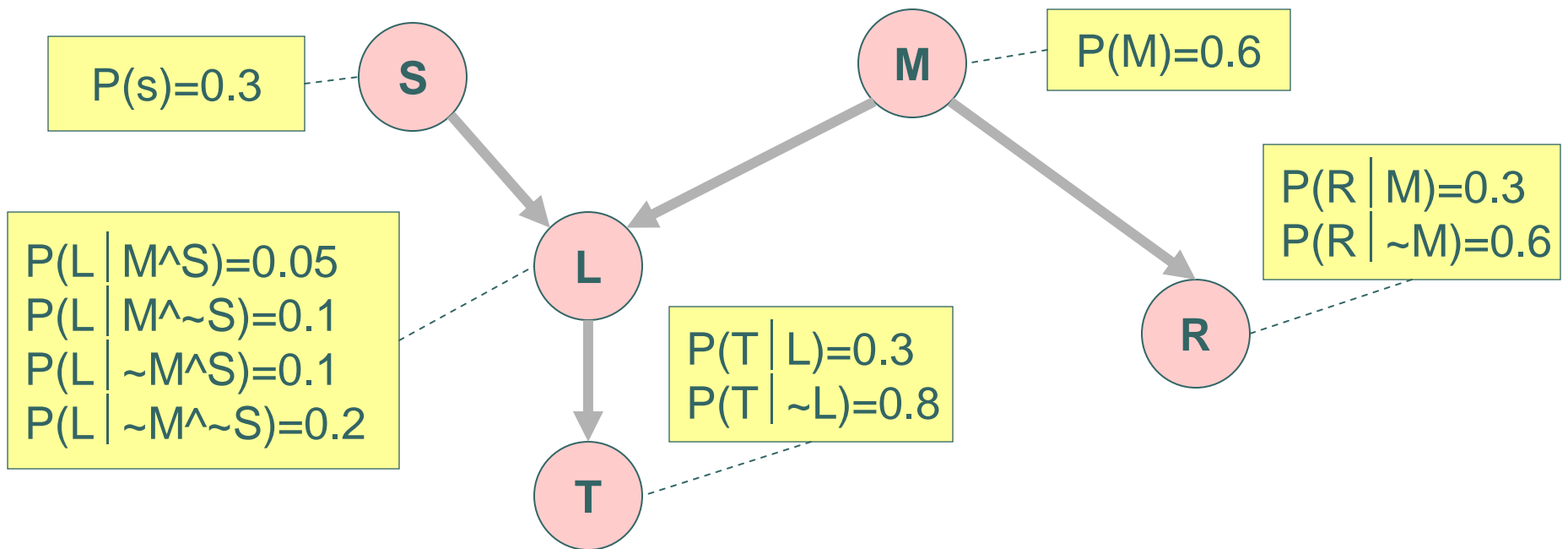


Step Two: add links.

- The link structure must be acyclic.
- If node  $X$  is given parents  $Q_1, Q_2, \dots, Q_n$  you are promising that any variable that's a non-descendent of  $X$  is conditionally independent of  $X$  given  $\{Q_1, Q_2, \dots, Q_n\}$

# Making a Bayes net

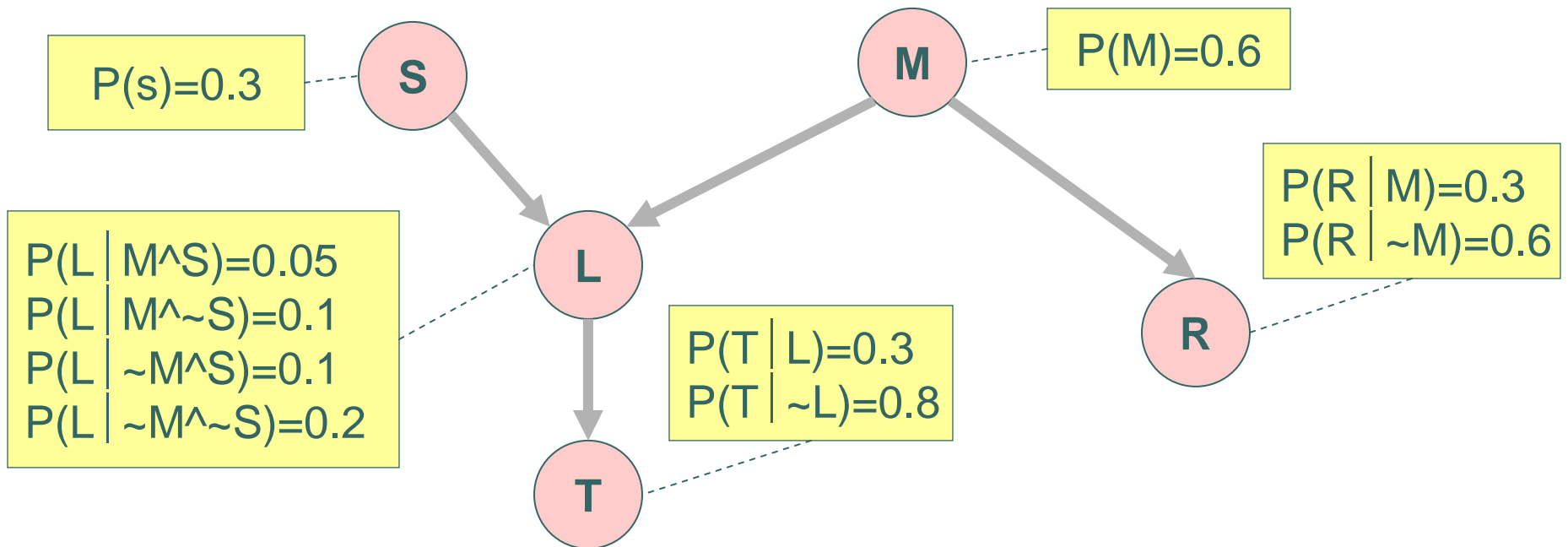
T: The lecture started by 10:35  
L: The lecturer arrives late  
R: The lecture concerns robots  
M: The lecturer is Manuela  
S: It is sunny



Step Three: add a probability table for each node.

- The table for node X must list  $P(X|\text{Parent Values})$  for each possible combination of parent values

# Making a Bayes net



- Two unconnected variables may still be correlated
- Each node is conditionally independent of all non-descendants in the tree, given its parents.
- You can deduce many other conditional independence relations from a Bayes net.



# Do It Yourslef: Example Bayes Net Building

Suppose we're building a nuclear power station.  
There are the following random variables:

GRL : Gauge Reads Low.

CTL : Core temperature is low.

FG : Gauge is faulty.

FA : Alarm is faulty

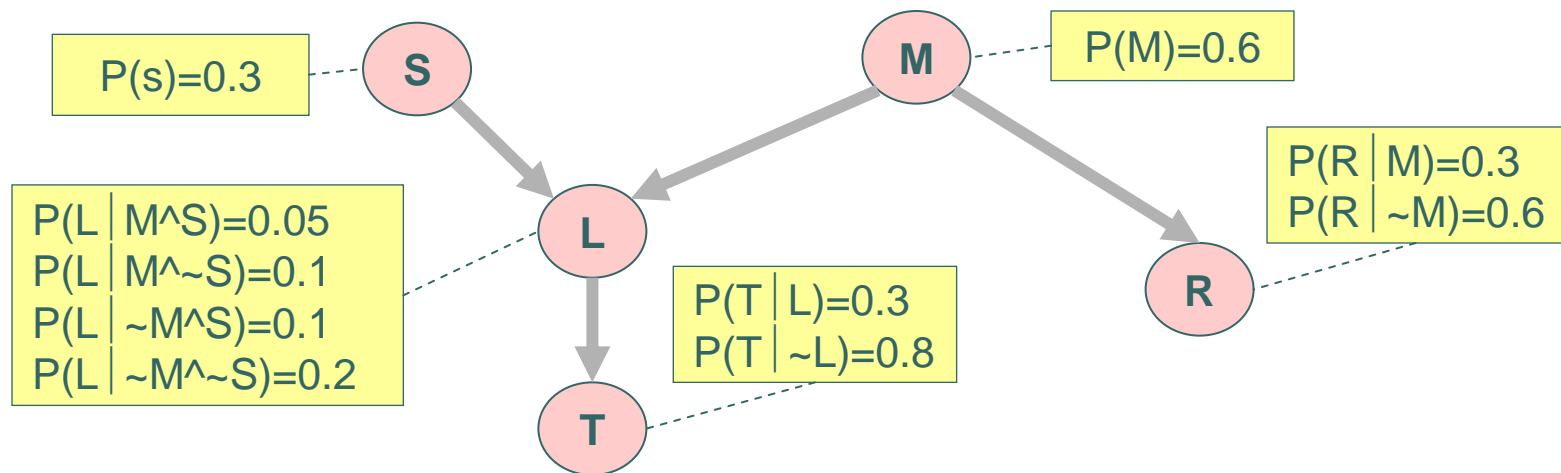
AS : Alarm sounds

- If alarm working properly, the alarm is meant to sound if the gauge stops reading a low temp.
- If gauge working properly, the gauge is meant to read the temp of the core.

# Computing a Joint Entry

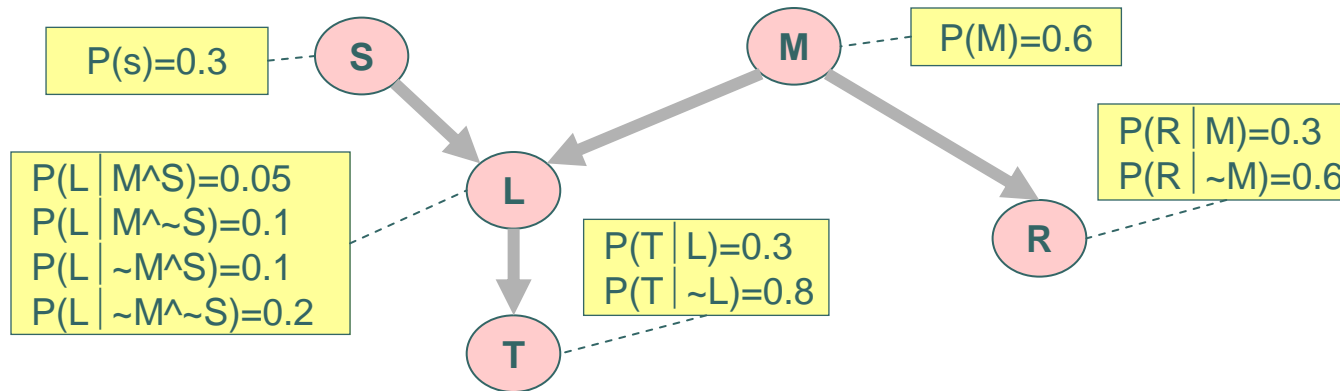
How to compute an entry in a joint distribution?

E.G: What is  $P(S \wedge \sim M \wedge L \wedge \sim R \wedge T)$ ?





# Computing with Bayes Net



$$\begin{aligned}
 &P(T \wedge \sim R \wedge L \wedge \sim M \wedge S) = \\
 &P(T \mid \sim R \wedge L \wedge \sim M \wedge S) * P(\sim R \wedge L \wedge \sim M \wedge S) = \\
 &P(T \mid L) * P(\sim R \wedge L \wedge \sim M \wedge S) = \\
 &P(T \mid L) * P(\sim R \mid L \wedge \sim M \wedge S) * P(L \wedge \sim M \wedge S) = \\
 &P(T \mid L) * P(\sim R \mid \sim M) * P(L \wedge \sim M \wedge S) = \\
 &P(T \mid L) * P(\sim R \mid \sim M) * P(L \mid \sim M \wedge S) * P(\sim M \wedge S) = \\
 &P(T \mid L) * P(\sim R \mid \sim M) * P(L \mid \sim M \wedge S) * P(\sim M \mid S) * P(S) = \\
 &P(T \mid L) * P(\sim R \mid \sim M) * P(L \mid \sim M \wedge S) * P(\sim M) * P(S).
 \end{aligned}$$



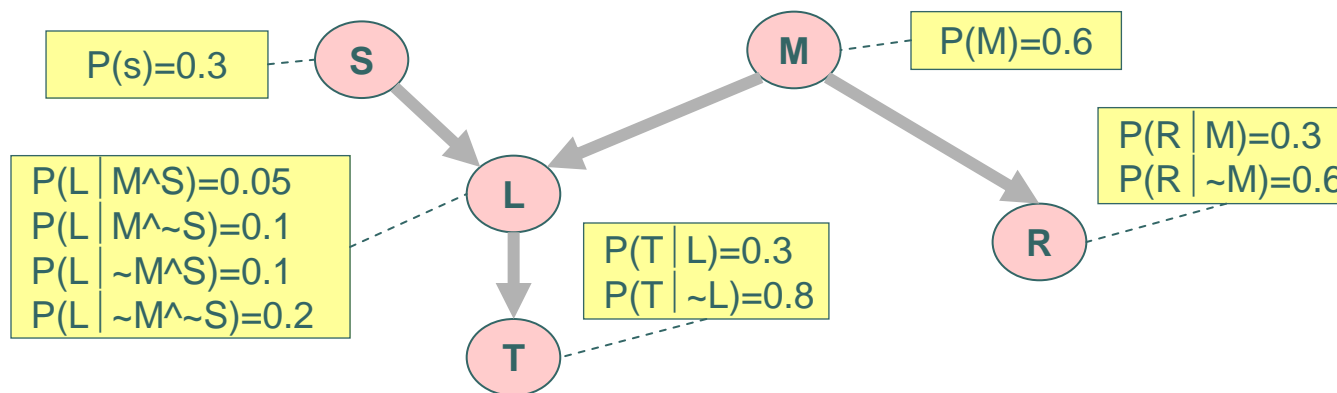
# The general case

$$\begin{aligned}
 &P(X_1=x_1 \wedge X_2=x_2 \wedge \dots \wedge X_{n-1}=x_{n-1} \wedge X_n=x_n) = \\
 &P(X_n=x_n \wedge X_{n-1}=x_{n-1} \wedge \dots \wedge X_2=x_2 \wedge X_1=x_1) = \\
 &P(X_n=x_n \mid X_{n-1}=x_{n-1} \wedge \dots \wedge X_2=x_2 \wedge X_1=x_1) * P(X_{n-1}=x_{n-1} \wedge \dots \wedge X_2=x_2 \wedge X_1=x_1) = \\
 &P(X_n=x_n \mid X_{n-1}=x_{n-1} \wedge \dots \wedge X_2=x_2 \wedge X_1=x_1) * P(X_{n-1}=x_{n-1} \mid \dots \wedge X_2=x_2 \wedge X_1=x_1) * \\
 &P(X_{n-2}=x_{n-2} \wedge \dots \wedge X_2=x_2 \wedge X_1=x_1) = \\
 &\quad \vdots \\
 &\quad \vdots \\
 = &\prod_{i=1}^n P((X_i = x_i) \mid ((X_{i-1} = x_{i-1}) \wedge \dots \wedge (X_1 = x_1))) \\
 = &\prod_{i=1}^n P((X_i = x_i) \mid \text{Assignments of Parents}(X_i))
 \end{aligned}$$

So any entry in joint pdf table can be computed. And so **any conditional probability** can be computed.

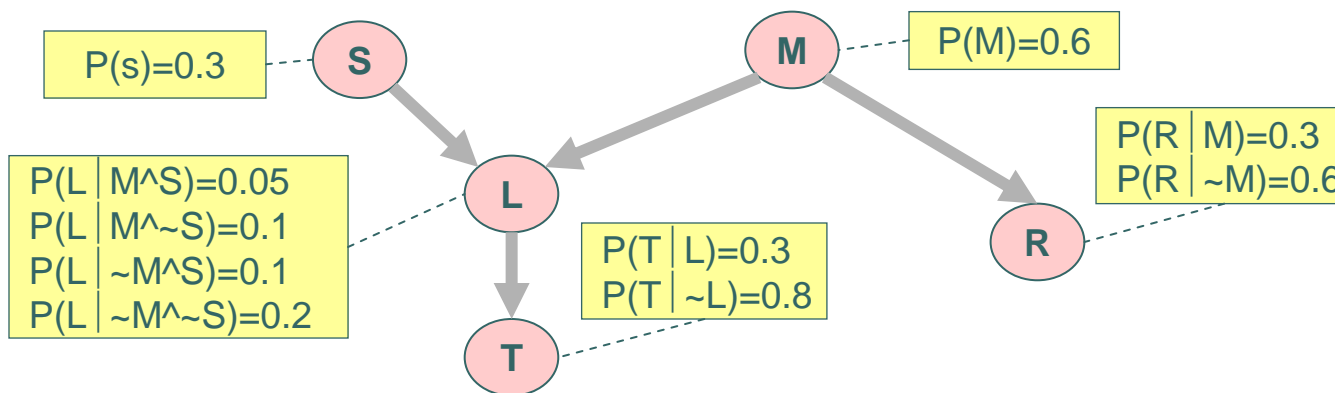
# Where are we now?

- We have a methodology for building Bayes nets.
- We don't require exponential storage to hold our probability table. Only exponential in the maximum number of parents of any node.
- We can compute probabilities of any given assignment of truth values to the variables. And we can do it in time linear with the number of nodes.
- So we can also compute answers to any questions.



E.G. What could we do to compute  $P(R | T, \sim S)$ ?

# A simple example (1)



E.G. What could we do to compute  $P(R | L)$ ?

Note: The normalizing factor ( $\alpha$ ) would be ignored for ease of presentation

# A simple example (1)

What is the probability of having a lecture about robotics if the lecturer is late? i.e.  
 $P(R|L)$

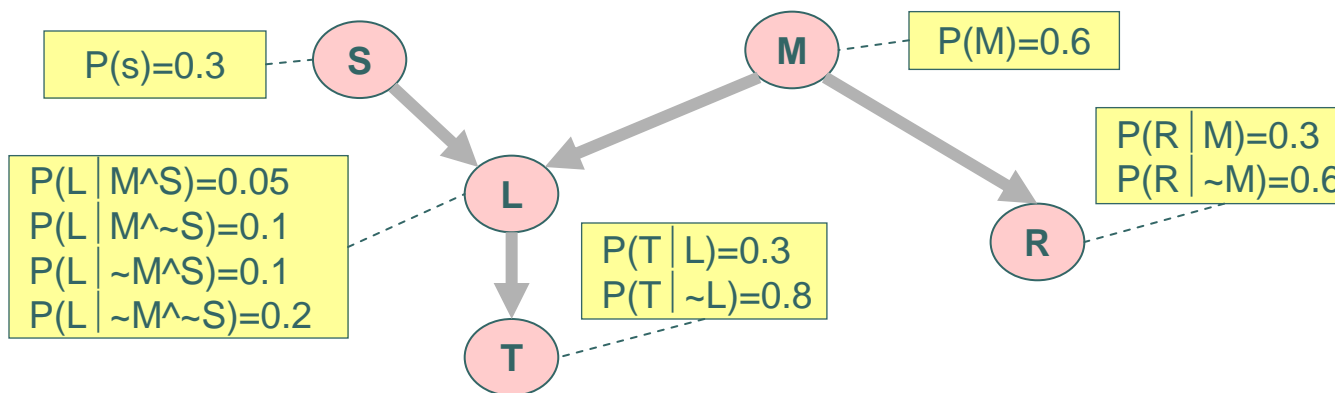
$$P(R|L) = P(R, M|L) + P(R, \sim M|L)$$

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

$$\begin{aligned} P(L|M) &= P(L, S|M) + P(L, \sim S|M) = P(L|S, M) * P(M, S) + P(L|\sim S, M) * P(M, \sim S) \\ &= P(L|S, M) * P(M) * P(S) + P(L|\sim S, M) * P(M) * P(\sim S) \\ &= 0.05 * 0.3 * 0.6 + 0.1 * 0.6 * 0.7 = 0.009 + 0.042 = 0.051 \end{aligned}$$

$$P(R, M|L) = 0.3 * 0.051 * 0.6 = 0.00918$$

Note: The normalizing factor ( $\alpha$ ) would be ignored for ease of presentation



E.G. What could we do to compute  $P(R | L)$ ?

# A simple example (2)

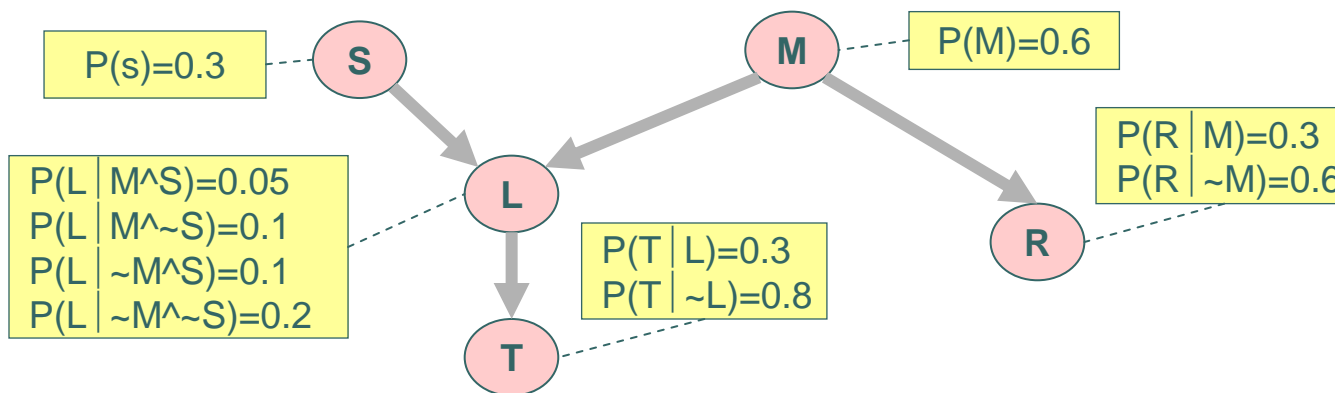
What is the probability of having a lecture about robotics if the lecturer is late? i.e.  
 $P(R|L)$

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

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

$$\begin{aligned} P(L|\sim M) &= P(L, S|\sim M) + P(L, \sim S|\sim M) = P(L|S, \sim M) * P(\sim M, S) + P(L|\sim S, \sim M) * P(\sim M, \sim S) \\ &= P(L|S, \sim M) * P(\sim M) * P(S) + P(L|\sim S, \sim M) * P(\sim M) * P(\sim S) \\ &= 0.1 * 0.4 * 0.3 + 0.2 * 0.4 * 0.7 = 0.012 + 0.056 = 0.068 \end{aligned}$$

$$P(R, \sim M|L) = 0.6 * 0.068 * 0.4 = 0.0022032$$



Note: The normalizing factor ( $\alpha$ ) would be ignored for ease of presentation

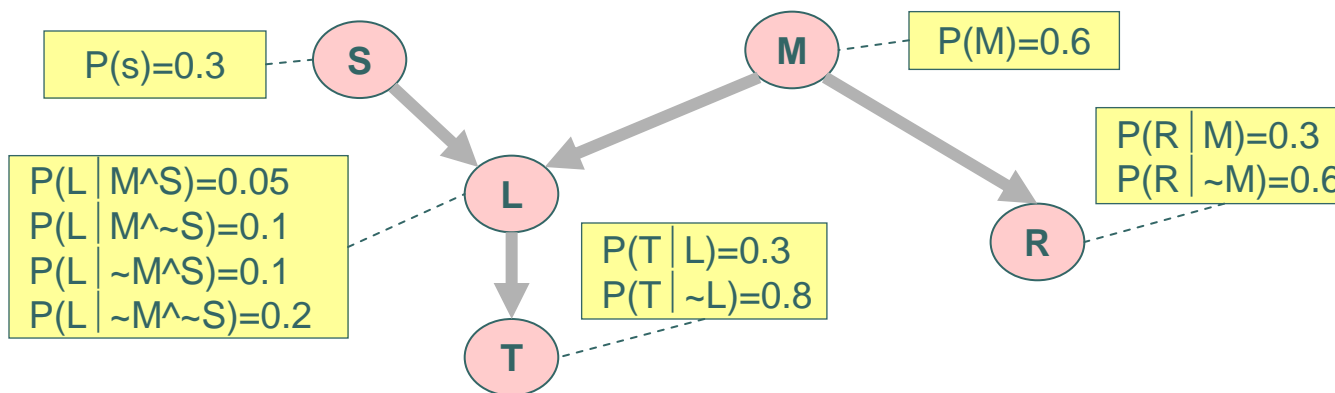
E.G. What could we do to compute  $P(R | L)$ ?

# A simple example (3)

What is the probability of having a lecture about robotics if the lecturer is late? i.e.  $P(R|L)$

$$P(R|L) = P(R, M|L) + P(R, \sim M|L) = 0.00918 + 0.0022032 = 0.0113832$$

You should note that this is not normalized. You need to calculate  $P(\sim R|L)$  too to normalized it to one.



E.G. What could we do to compute  $P(R | L)$ ?

# Where are we now?

Step 1: Compute  $P(R \wedge T \wedge \sim S)$

Step 2: Compute  $P(\sim R \wedge T \wedge \sim S)$

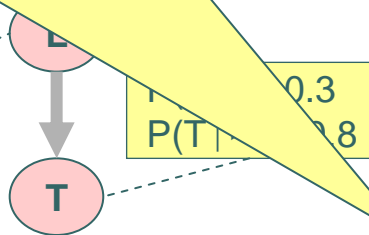
Step 3: Return

$$P(R \wedge T \wedge \sim S)$$

---


$$P(R \wedge T \wedge \sim S) + P(\sim R \wedge T \wedge \sim S)$$

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



$$P(M) = 0.6$$

$P(R \mid M)$	$=0.3$
$P(R \mid \sim M)$	$=0.6$

technology for building Bayes nets.

exponential storage to hold our  
likely exponential in the maximum  
of any node.

probabilities of any given assignment  
of variables. And we can do it in  
number of nodes.

provide answers to any questions.

E.G. What could we do to compute  $P(R \mid T, \sim S)$ ?

# Where are we now?

Step 1: Compute  $P(R \wedge T \wedge \sim S)$

Sum of all the rows in the Joint that match  $R \wedge T \wedge \sim S$

Step 2: Compute  $P(\sim R \wedge T \wedge \sim S)$

Technology for building Bayes nets.

Step 3: Return

Sum of all the rows in the Joint that match  $\sim R \wedge T \wedge \sim S$

$$P(R \wedge T \wedge \sim S)$$

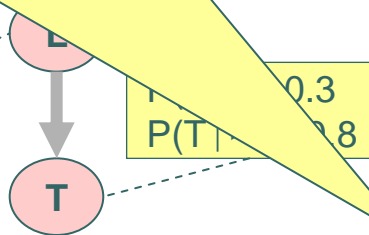
one  
ly e  
any node.

$$P(R \wedge T \wedge \sim S) + P(\sim R \wedge T \wedge \sim S)$$

probabilities of any given assignment of variables. And we can do it in a number of nodes.

Give answers to any questions.

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



$$P(M) = 0.6$$

$P(R   M)$	$= 0.3$
$P(R   \sim M)$	$= 0.6$

E.G. What could we do to compute  $P(R | T, \sim S)$ ?

# Where are we now?

Step 1: Compute  $P(R \wedge T \wedge \sim S)$

Step 2: Compute  $P(\sim R \wedge T \wedge \sim S)$

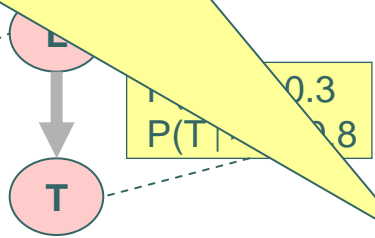
Step 3: Return

$$P(R \wedge T \wedge \sim S)$$

---


$$P(R \wedge T \wedge \sim S) + P(\sim R \wedge T \wedge \sim S)$$

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



$P(R \mid \sim M) = 0.6$

4 joint computes

Sum of all the rows in the Joint that match  $R \wedge T \wedge \sim S$

Technology for building Bayes nets.

Sum of all the rows in the Joint that match  $\sim R \wedge T \wedge \sim S$

one  
ly e  
any node.

4 joint computes

Each of these obtained by the "computing a joint probability entry" method of the earlier slides

E.G. What could we do to compute  $P(R \mid T, \sim S)$ ?



# The good news

We can do inference. We can compute any conditional probability:

$P(\text{Some variable} \mid \text{Some other variable values})$

$$P(E_1 \mid E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{joint entries matching } E_1 \text{ and } E_2} P(\text{joint entry})}{\sum_{\text{joint entries matching } E_2} P(\text{joint entry})}$$



# The good news

$$P(E_1 | E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{joint entries matching } E_1 \text{ and } E_2} P(\text{joint entry})}{\sum_{\text{joint entries matching } E_2} P(\text{joint entry})}$$

Suppose you have  $m$  binary-valued variables in your Bayes Net and expression  $E_2$  mentions  $k$  variables.

How much work is the above computation?



# The sad, bad news

Conditional probabilities by enumerating all matching entries in the joint are expensive:

**Exponential in the number of variables.**



# The sad, bad news

Conditional probabilities by enumerating all matching entries in the joint are expensive:

**Exponential in the number of variables.**

But perhaps there are faster ways of querying Bayes nets?

- In fact, if I ever ask you to manually do a Bayes Net inference, you'll find there are often many tricks to save you time.
- So we've just got to program our computer to do those tricks too, right?

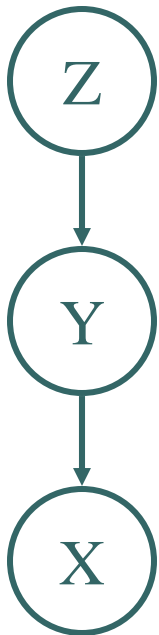
**Sadder and worse news:**

**General querying of Bayes nets is NP-complete.**

# What Independencies does a Bayes Net Model?



## Example:



Given  $Y$ , does learning the value of  $Z$  tell us nothing new about  $X$ ?

I.e., is  $P(X|Y, Z)$  equal to  $P(X | Y)$ ?

Yes. Since we know the value of all of  $X$ 's parents (namely,  $Y$ ), and  $Z$  is not a descendant of  $X$ ,  $X$  is conditionally independent of  $Z$ .

Also, since independence is symmetric,  
 $P(Z|Y, X) = P(Z|Y)$ .



# Quick proof that independence is symmetric

- Assume:  $P(X/Y, Z) = P(X/Y)$
- Then:

$$P(Z | X, Y) = \frac{P(X, Y | Z)P(Z)}{P(X, Y)} \quad (\text{Bayes's Rule})$$

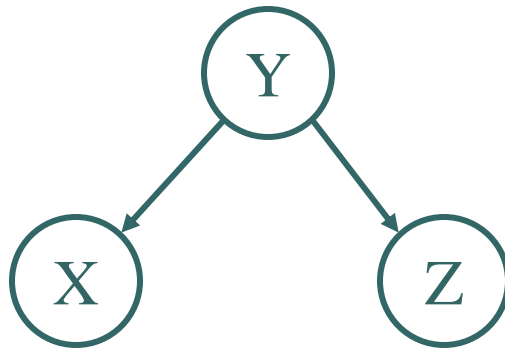
$$= \frac{P(Y | Z)P(X | Y, Z)P(Z)}{P(X | Y)P(Y)} \quad (\text{Chain Rule})$$

$$= \frac{P(Y | Z)P(X | Y)P(Z)}{P(X | Y)P(Y)} \quad (\text{By Assumption})$$

$$= \frac{P(Y | Z)P(Z)}{P(Y)} = P(Z | Y) \quad (\text{Bayes's Rule})$$

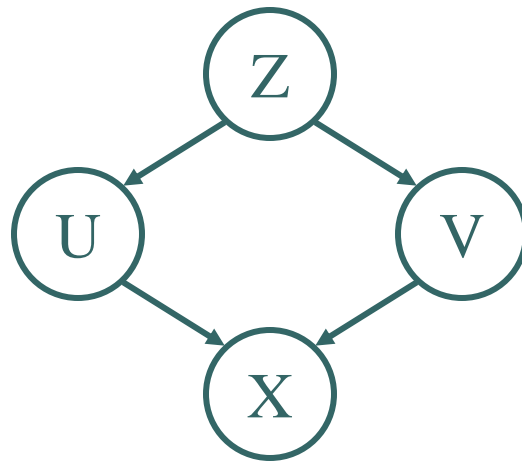
# What Independencies does a Bayes Net Model?

- Let  $I\langle X, Y, Z \rangle$  represent  $X$  and  $Z$  being conditionally independent given  $Y$ .



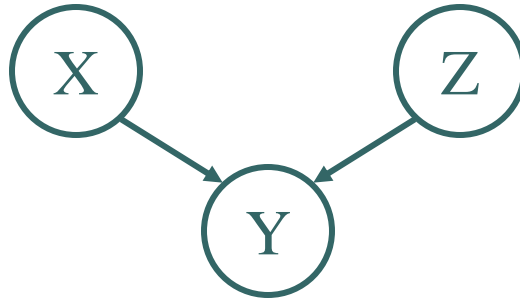
- $I\langle X, Y, Z \rangle$ ? Yes, just as in previous example: All  $X$ 's parents given, and  $Z$  is not a descendant.

# What Independencies does a Bayes Net Model?



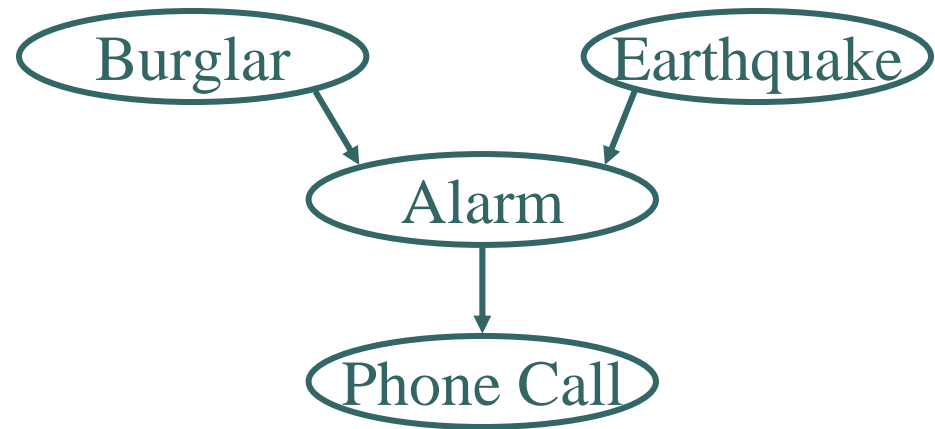
- $I\langle X, \{U\}, Z \rangle$ ? No.
- $I\langle X, \{U, V\}, Z \rangle$ ? Yes.
- Maybe  $I\langle X, S, Z \rangle$  iff  $S$  acts a cutset between  $X$  and  $Z$  in an undirected version of the graph...?

## Things get a little more confusing



- X has no parents, so we know all its parents' values trivially
- Z is not a descendant of X
- So,  $I\langle X, \{\}, Z \rangle$ , even though there's a undirected path from X to Z through an unknown variable Y.
- What if we do know the value of Y, though? Or one of its descendants?

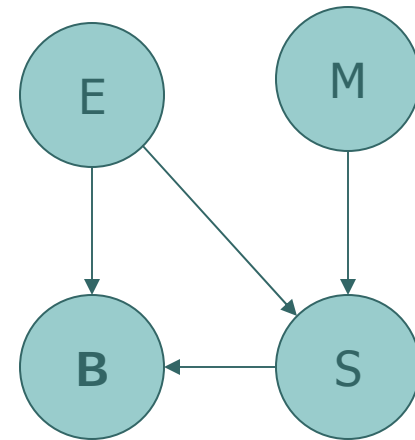
# The “Burglar Alarm” example



- One of the neighbors call for alarm.
- Suppose you learn that there was a medium-sized earthquake in your neighborhood. Oh, whew! Probably not a burglar after all.
- Earthquake “explains away” the hypothetical burglar.
- But then it must **not** be the case that  $I\langle \text{Burglar}, \{\text{Phone Call}\}, \text{Earthquake} \rangle$ , even though  $I\langle \text{Burglar}, \{\}, \text{Earthquake} \rangle$ !

# ● ● ● | Another Example

- Credit scoring problem:
  - B: Bank account (credit/debt)
  - E: Employment status (emp/unemp)
  - M: Money management (good/bad)
  - S: Spending habit (spend/thrift)



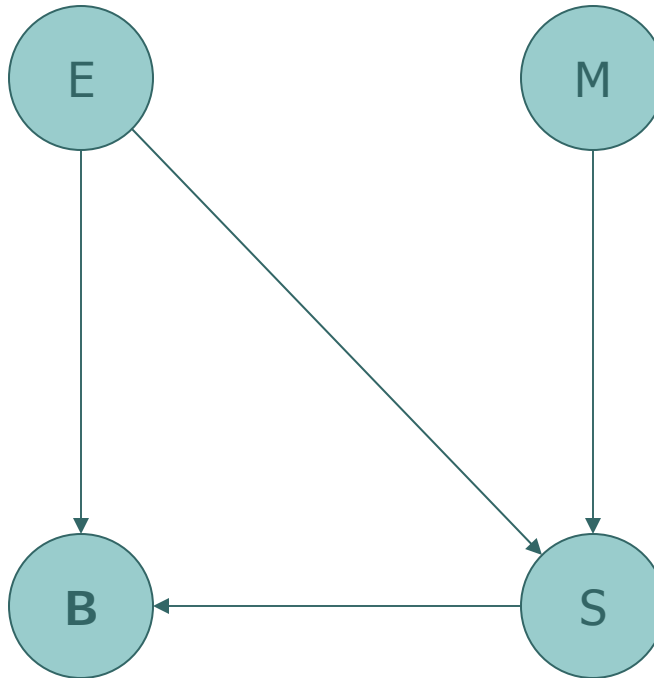


# Another Example (2)

E=Employed  
M=Good money management  
S= Spend  
B=credit

$$P(E)=0.9$$

$$P(M)=0.6$$



$$\begin{aligned} P(B \mid E \wedge S) &= 0.6 \\ P(B \mid E \wedge \sim S) &= 1.0 \\ P(B \mid \sim E \wedge S) &= 0.2 \\ P(B \mid \sim E \wedge \sim S) &= 0.6 \end{aligned}$$

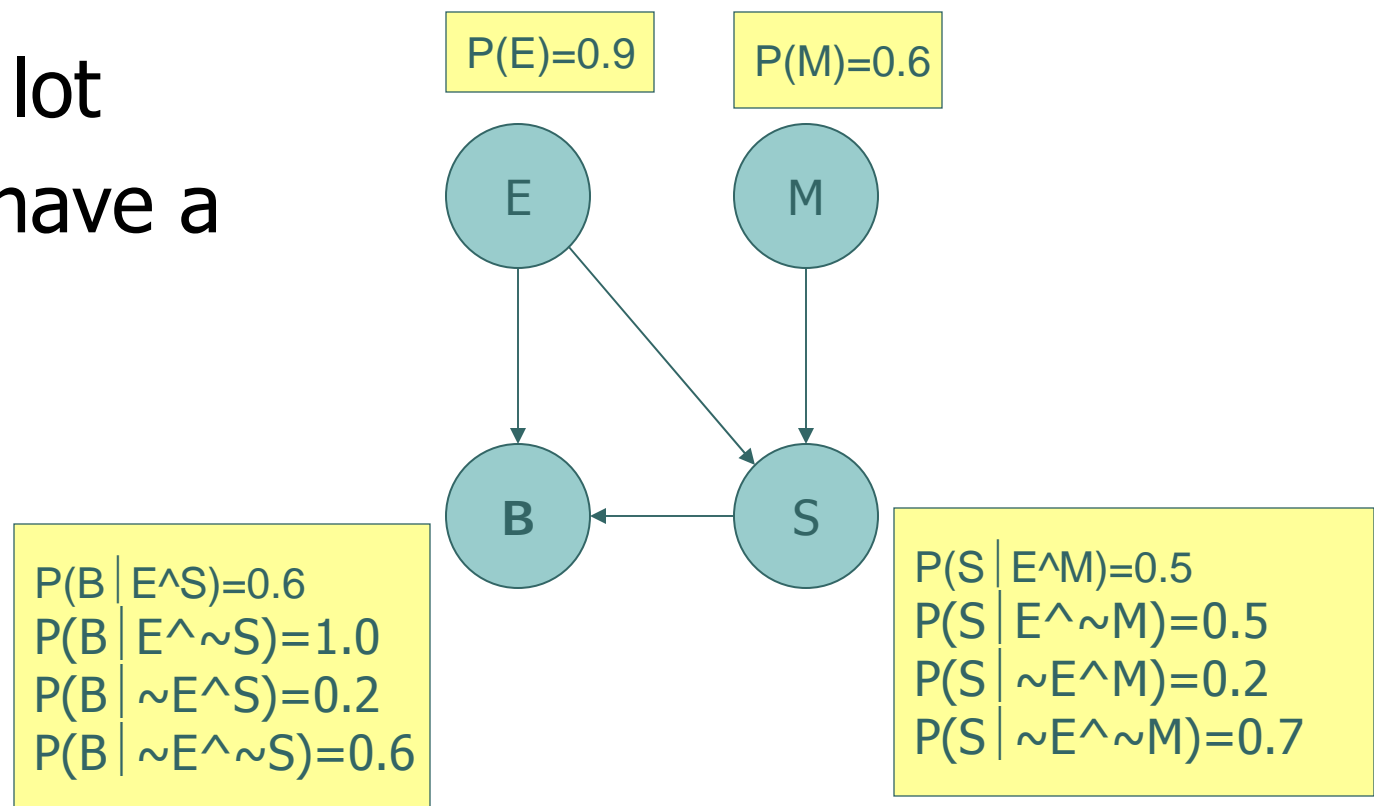
$$\begin{aligned} P(S \mid E \wedge M) &= 0.5 \\ P(S \mid E \wedge \sim M) &= 0.5 \\ P(S \mid \sim E \wedge M) &= 0.2 \\ P(S \mid \sim E \wedge \sim M) &= 0.7 \end{aligned}$$



# Another Example (3)

- A person is in debt and spends a lot
- Does he have a job?

E=Employed  
M=Good money management  
S= Spend  
B=credit





# Summary

- Bayesian networks provide a natural representation for (causally induced) conditional independence
- Topology + CPTs = compact representation of joint distribution
- Generally easy for domain experts to construct