MONASH University
Information Technology

MONASH University
Information Technology

FIT5047: Intelligent Systems

Bayesian Networks: Representation and Inference Chapters 14.1-4, 16.1-2, 16.5

Some slides are adapted from Stuart Russell, Andrew Moore, or Dan Klein

Bayesian Reasoning: Learning Objectives

- **Bayesian AI**
- **Bayesian networks**
- **Decision networks**

MONASH University
Information Technology

FIT5047 – Intelligent Systems

Bayesian AI

Assumptions about the Environment

- **Fully / partially observable**
- **Known**
- **Single / multi agent**
- **Stochastic**
- **Sequential / episodic**
- **Static**
- **Discrete / continuous**

Bayesian Conception of an AI

- **An autonomous agent that**
  - has a utility structure (preferences)
  - can learn about its world and the relationship (probabilities) between its actions and future states
  - maximizes its expected utility
- **The techniques used to learn about the world are mainly statistical**
  ➔Machine learning

Bayesian Decision Theory

- **Frank Ramsey (1926)**
- **Decision making under uncertainty – what action to take when the state of the world is unknown**
- **Bayesian answer –**
  Find the utility of each possible outcome (action-state pair), and take the action that maximizes the *expected utility*

Bayesian Decision Theory – Example

Action	Rain (p=0.4)	Shine (1-p=0.6)
Take umbrella	60	-10
Leave umbrella	-100	50

Expected utilities:

- E(Take umbrella) = $60 \times 0.4 + (-10) \times 0.6 = 18$
- E(Leave umbrella) = $-100 \times 0.4 + 50 \times 0.6 = -10$

MONASH University
Information Technology

FIT5047 – Intelligent Systems

Bayesian Networks

Bayesian Networks (BNs) – Overview

- **Introduction to BNs**
  - Nodes, structure and probabilities
  - Reasoning with BNs
  - Understanding BNs
- **Extensions of BNs**
  - Decision Networks
  - (Dynamic Bayesian Networks (DBNs))

Conditional Independence (reminder)

- **X and Y are independent if**

$$\forall x, y \; \Pr(x, y) = \Pr(x)\Pr(y)$$

$$\dashrightarrow \quad X \perp\!\!\!\perp Y$$

- **X and Y are conditionally independent given Z**

$$\forall x, y, z \; \Pr(x, y|z) = \Pr(x|z)\Pr(y|z)$$

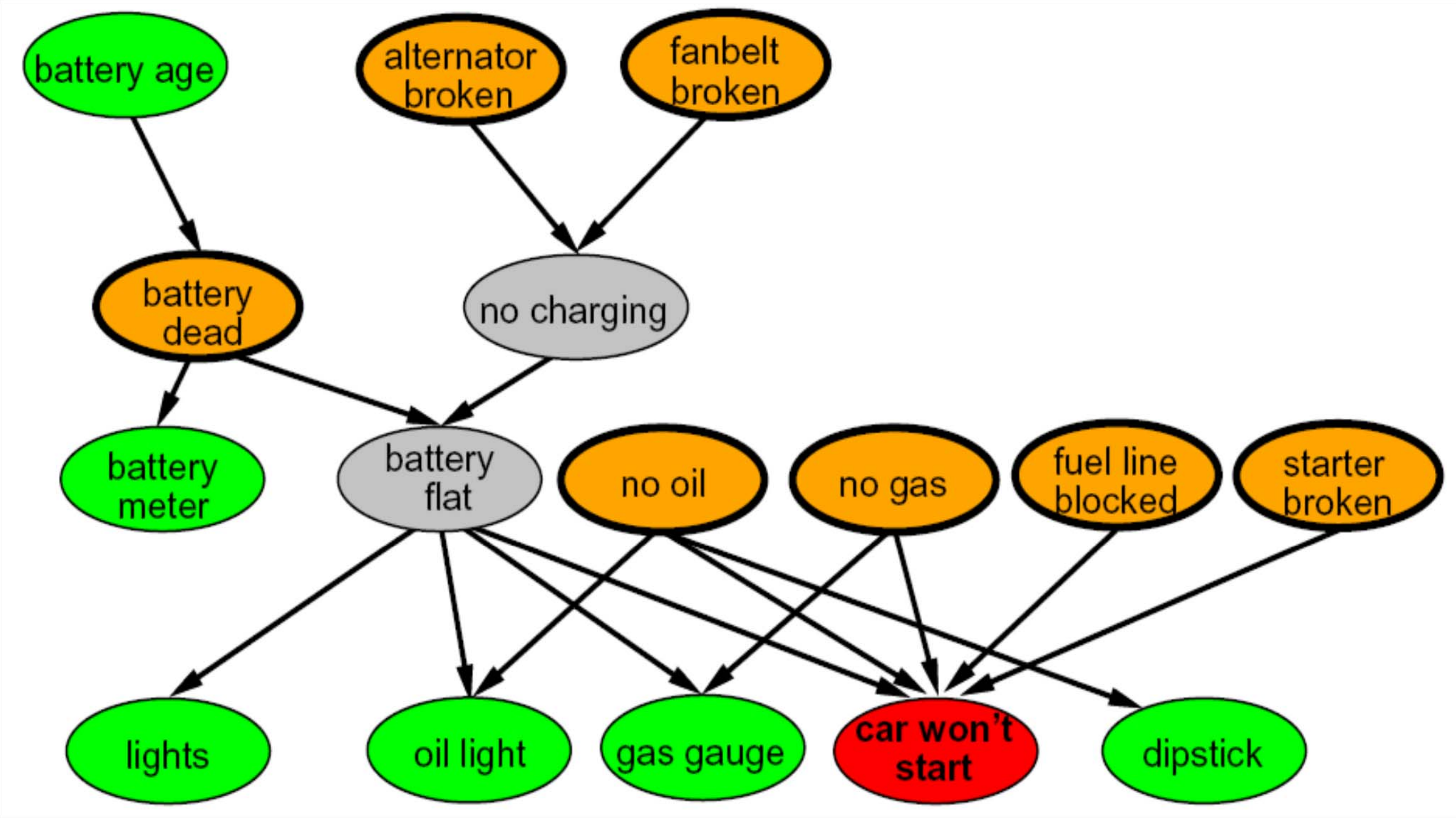
$$\dashrightarrow \quad X \perp\!\!\!\perp Y | Z$$

- **(Conditional) independence is a property of a distribution**

Bayesian Networks: The Big Picture

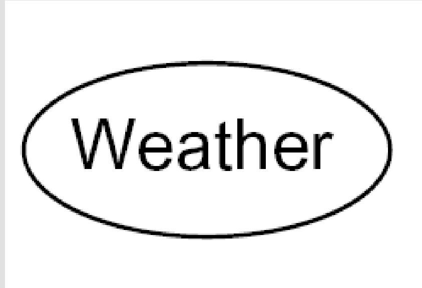
- **Two problems with using full joint distribution tables as our probabilistic models:**
  - Unless there are only a few variables, the joint is too big to represent explicitly
  - Hard to learn (estimate) anything empirically about more than a few variables at a time
- **Bayes nets (aka graphical models): a technique for describing complex joint distributions (models) using simple, local distributions (conditional probabilities)**
  - Describe how variables interact locally
    - > Local interactions chain together to give global, indirect interactions

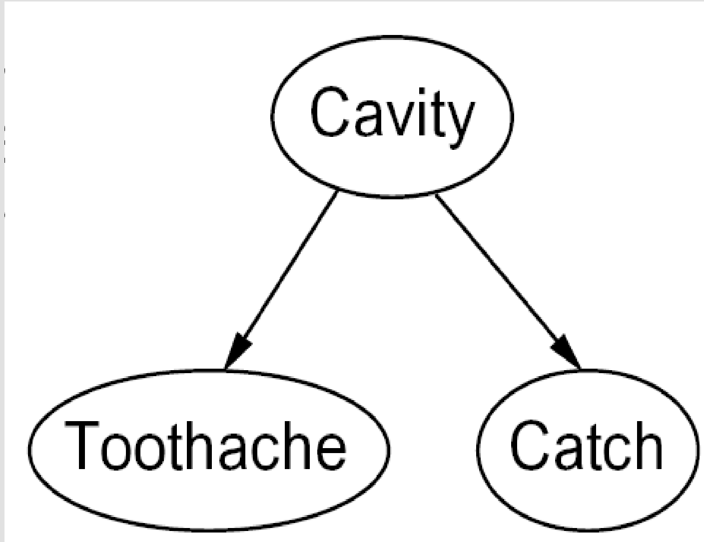
Example Bayesian Network: Car



Graphical Model – Notation

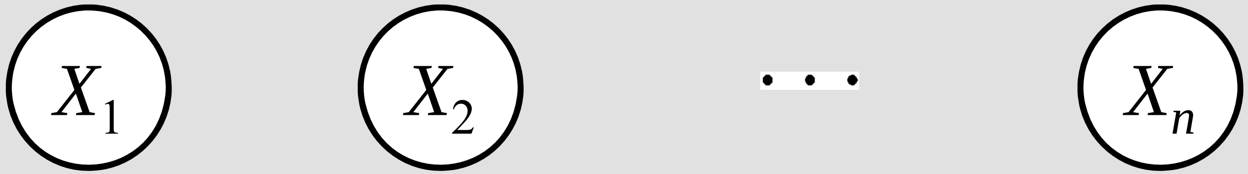
- **Nodes: variables (with domains)**
  - Can be assigned (observed) or unassigned (unobserved)
- **Arcs: interactions**
  - Indicate "direct influence" between variables
  - Formally: encode conditional independence
- **For now, imagine that arrows mean direct causation**





Example: Coin Flips (I)

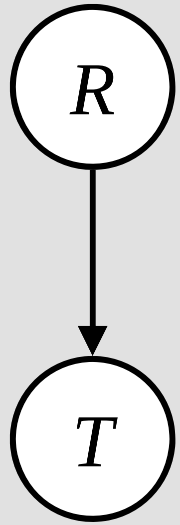
- **N independent coin flips**



- **No interactions between variables: absolute independence**

Example: Traffic (I)

- **Variables:**
  - R: It rains
  - T: There is traffic



- **Model 1: independence**

- **Model 2: rain causes traffic**

- **Why is model 2 better?**

Bayesian Networks – Definition (I)

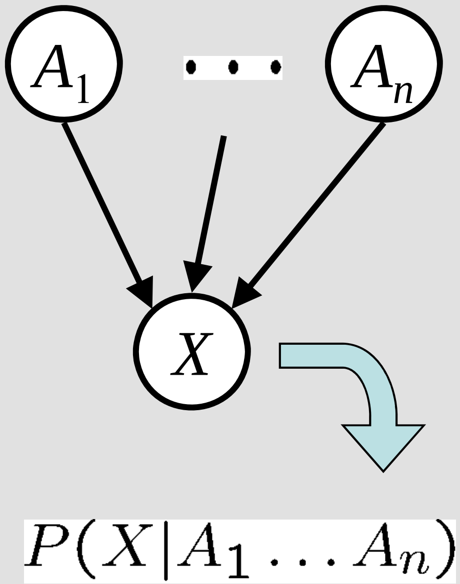
- **A data structure that represents the dependence between random variables**
- **A Bayesian Network is a directed acyclic graph (DAG) in which the following holds:**
  1. A set of random variables makes up the nodes in the network
  2. A set of directed links connects pairs of nodes
  3. Each node has a probability distribution that quantifies the effects of its ***parent nodes***
     > Discrete nodes have ***Conditional Probability Tables*** (***CPTs***)
- **Gives a concise specification of the *joint probability distribution* of the variables**

Bayesian Networks – Definition (II)

- **The probability distribution for each node X is a collection of distributions over X, one for each combination of its parents' values**

$$\mathbf{Pr}(X|a_1, \dots, a_n)$$

  - described by a ***Conditional Probability Table (CPT)***
  - describes a "noisy" causal process



Bayesian network = Topology (graph) + Local Conditional Probabilities

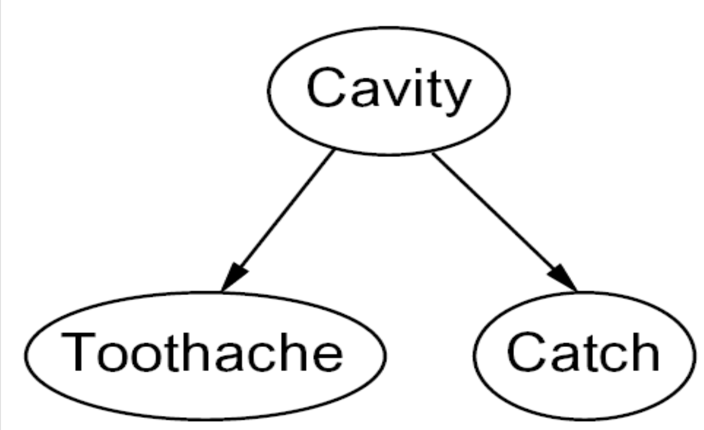
Probabilities in BNs

- **Bayes nets implicitly encode joint distributions**
  - As a product of local conditional distributions
- **To see what probability a BN gives to a full assignment, multiply all the relevant conditionals**

$$\Pr(x_1, x_2, \dots, x_n) = \prod_{i=1}^{n} \Pr(x_i \mid \text{parents}(X_i))$$

- **This lets us reconstruct any entry of the full joint distribution**
  - But not every BN can represent every joint distribution

Building the Joint Distribution – Example



Pr(+cavity, +catch, ¬toothache) = ?

= Pr(¬toothache|+cavity,+catch) Pr(+catch|+cavity) Pr(+cavity)

= Pr(¬toothache|+cavity) Pr(+catch|+cavity) Pr(+cavity)

Conditional independence

Example: Coin Flips (II)

X_1 X_2 … X_n

$\Pr(X_1)$

h	0.5
t	0.5

$\Pr(X_2)$

h	0.5
t	0.5

…

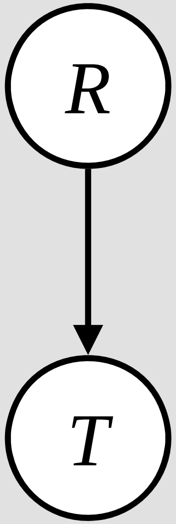
$\Pr(X_n)$

h	0.5
t	0.5

$$\Pr(h, t, t, h) = 0.5 \text{ x } 0.5 \text{ x } 0.5 \text{ x } 0.5$$

Only distributions whose variables are independent can be represented by a Bayes net with no arcs

Example: Traffic (II)



$\Pr(R)$

+r	1/4
¬r	3/4

$\Pr(T|R)$

+r	+t	3/4
+r	¬t	1/4
¬r	+t	1/2
¬r	¬t	1/2

$=$ $\Pr(T|R)$

	+t	¬t
+r	3/4	1/4
¬r	1/2	1/2

$$\Pr(+r, \neg t) = ?$$

$$\Pr(+r, \neg t) = \Pr(\neg t | +r)\Pr(+r) = \frac{1}{4} \times \frac{1}{4} = \frac{1}{16}$$

Example – Lung Cancer Diagnosis

A patient has been suffering from shortness of breath (called dyspnoea) and visits the doctor, worried that he has lung cancer.

The doctor knows that other relevant information includes whether or not the patient is a smoker (increasing the chances of cancer) and what sort of air pollution he has been exposed to. A positive Xray would indicate lung cancer.

Nodes and Values

Q: What do the nodes represent and what values can they take?

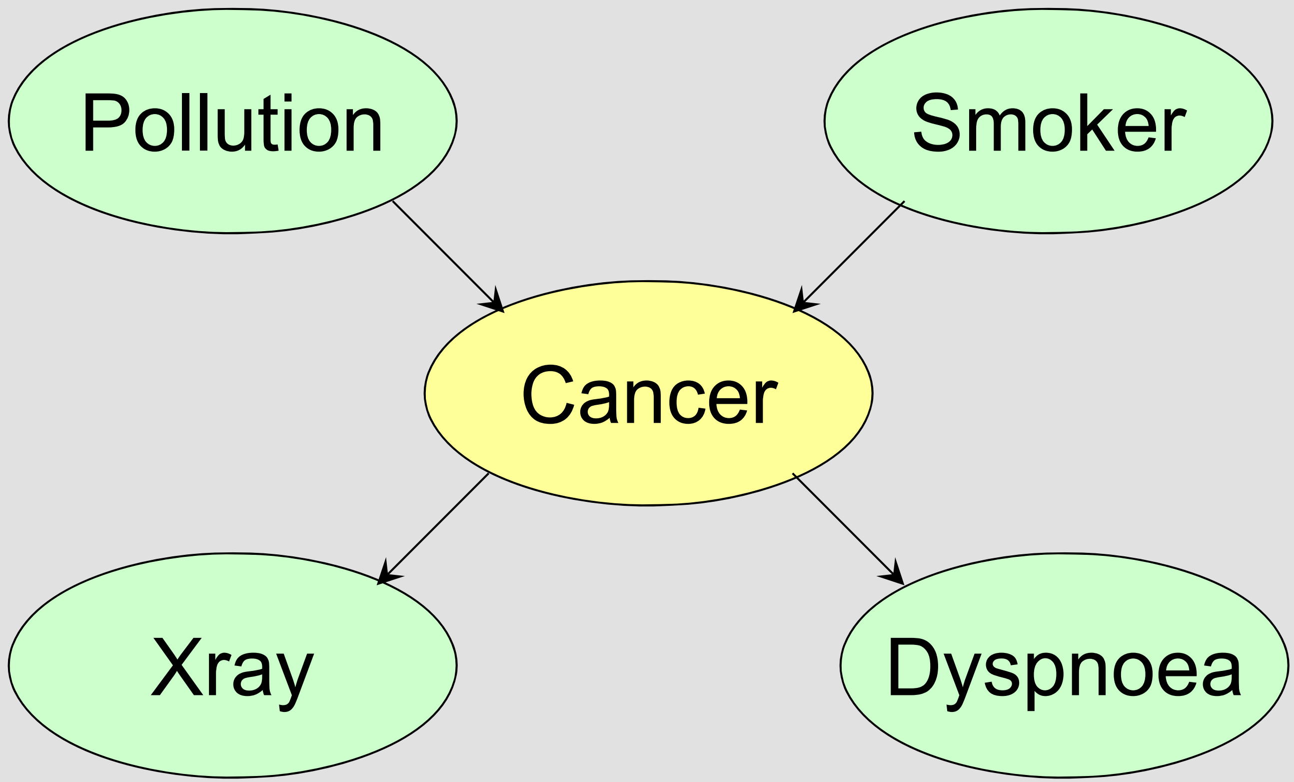
A: Nodes can be discrete or continuous

- **Binary values**
  - Boolean nodes (special case)
    Example: *Cancer* node represents proposition *"the patient has cancer"*
- **Ordered values**
  - Example: *Pollution* node with values *low, medium, high*
- **Integral values**
  - Example: *Age* with possible values 1-120

Lung Cancer Example: Nodes and Values

Node name	Type	Values
Pollution	Binary	{*low*,*high*}
Smoker	Boolean	{T,F}
Cancer	Boolean	{T,F}
Dyspnoea	Boolean	{T,F}
Xray	Binary	{*pos*,*neg*}

Lung Cancer Example: Network Structure



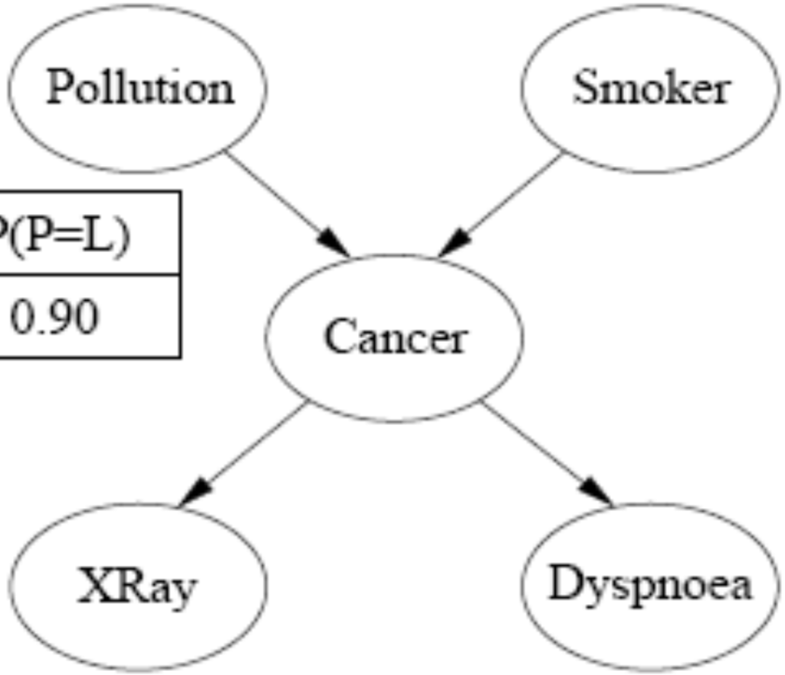
Conditional Probability Tables (CPTs)

After specifying topology, we must specify the CPT for each discrete node

- **Each row contains the conditional probability of each node value for each possible combination of values in its parent nodes**
- **Each row must sum to 1**
- **A CPT for a Boolean variable with n Boolean parents contains 2^{n+1} probabilities**
- **A node with no parents has one row (its prior probabilities)**

MONASH University
Information Technology

Lung Cancer Example: CPTs



P(P=L)
0.90

P(S=T)
0.30

P	S	P(C=T|P,S)
H	T	0.05
H	F	0.02
L	T	0.03
L	F	0.001

C	P(X=pos|C)
T	0.90
F	0.20

C	P(D=T|C)
T	0.65
F	0.30

The value for =F is the complement

Understanding Bayesian Networks

- **Understand how to construct a network**
  - A (more compact) representation of the joint probability distribution, which encodes a collection of conditional independence statements
- **Understand how to design inference procedures**
  - Encode a collection of conditional independence statements
  - Apply the ***Markov property***
    - > There are no direct dependencies in the system being modeled which are not already explicitly shown via arcs
    - > Example: smoking can influence dyspnoea only through causing cancer

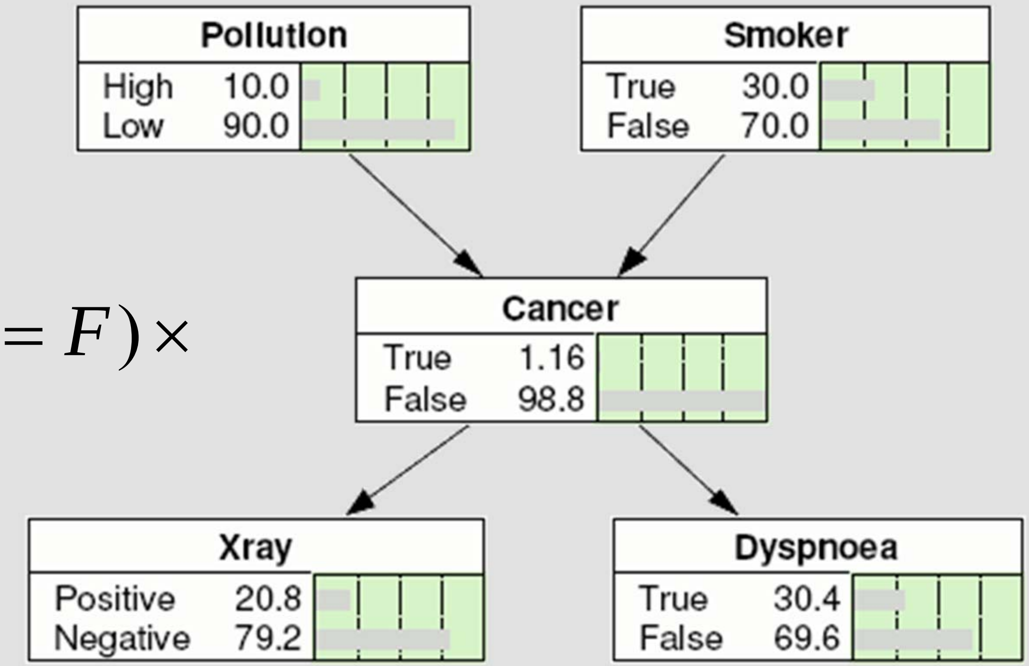
Representing Joint Probability Distribution: Example

$$\Pr(P = low \wedge S = F \wedge C = T \wedge X = pos \wedge D = T) =$$

$$\Pr(P = low) \times$$
$$\Pr(S = F) \times$$

$$\Pr(C = T \mid P = low, S = F) \times$$

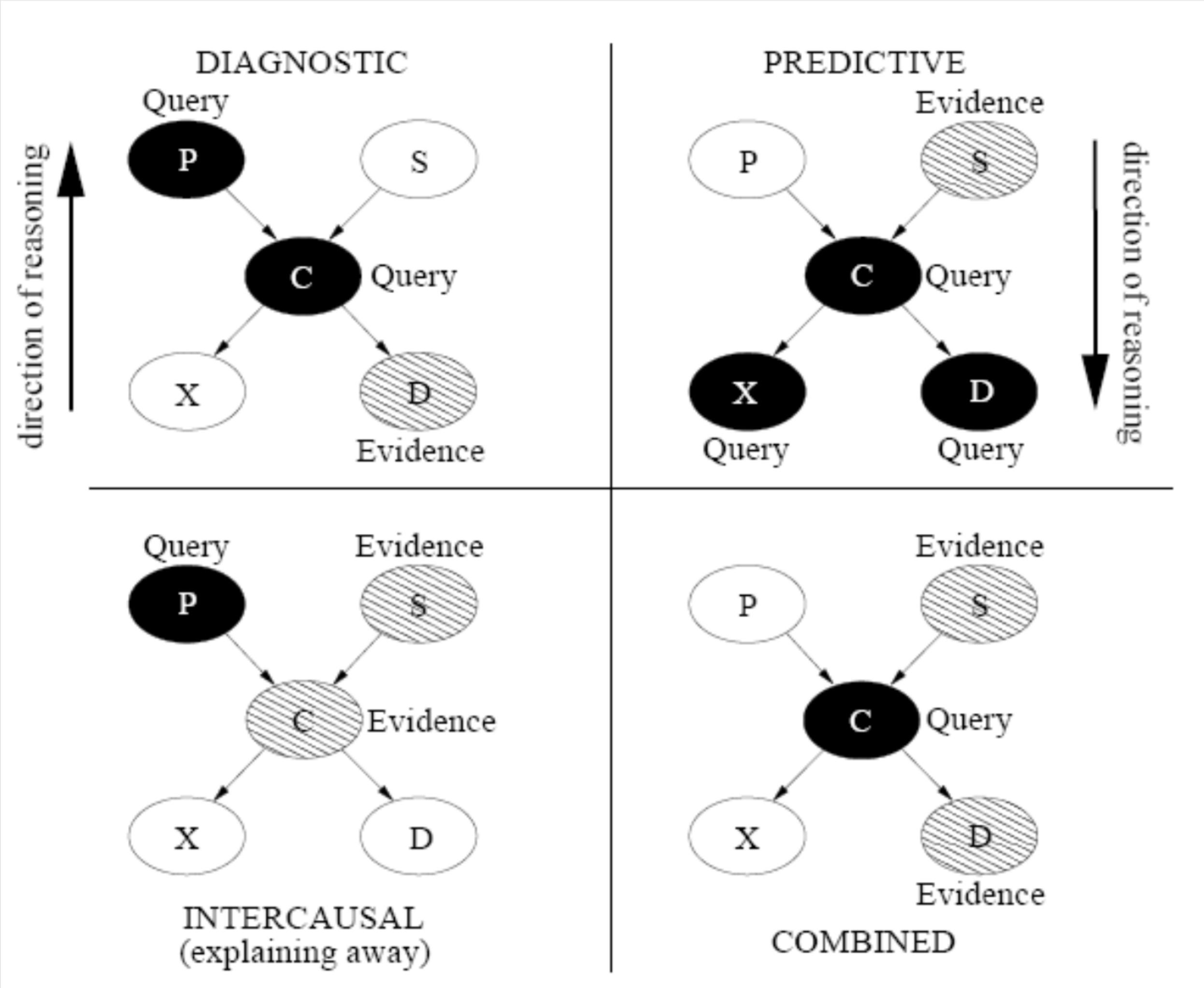
$$\Pr(X = pos \mid C = T) \times$$
$$\Pr(D = T \mid C = T)$$



Reasoning with Bayesian Networks

- **Basic task for any probabilistic inference system:**
  Compute the posterior probability distribution for a set of *query variables*, given new information about some *evidence variables*
- **Also called *conditioning* or *belief updating* or *inference***

Types of Reasoning



Example – Earthquake (Pearl 1988)

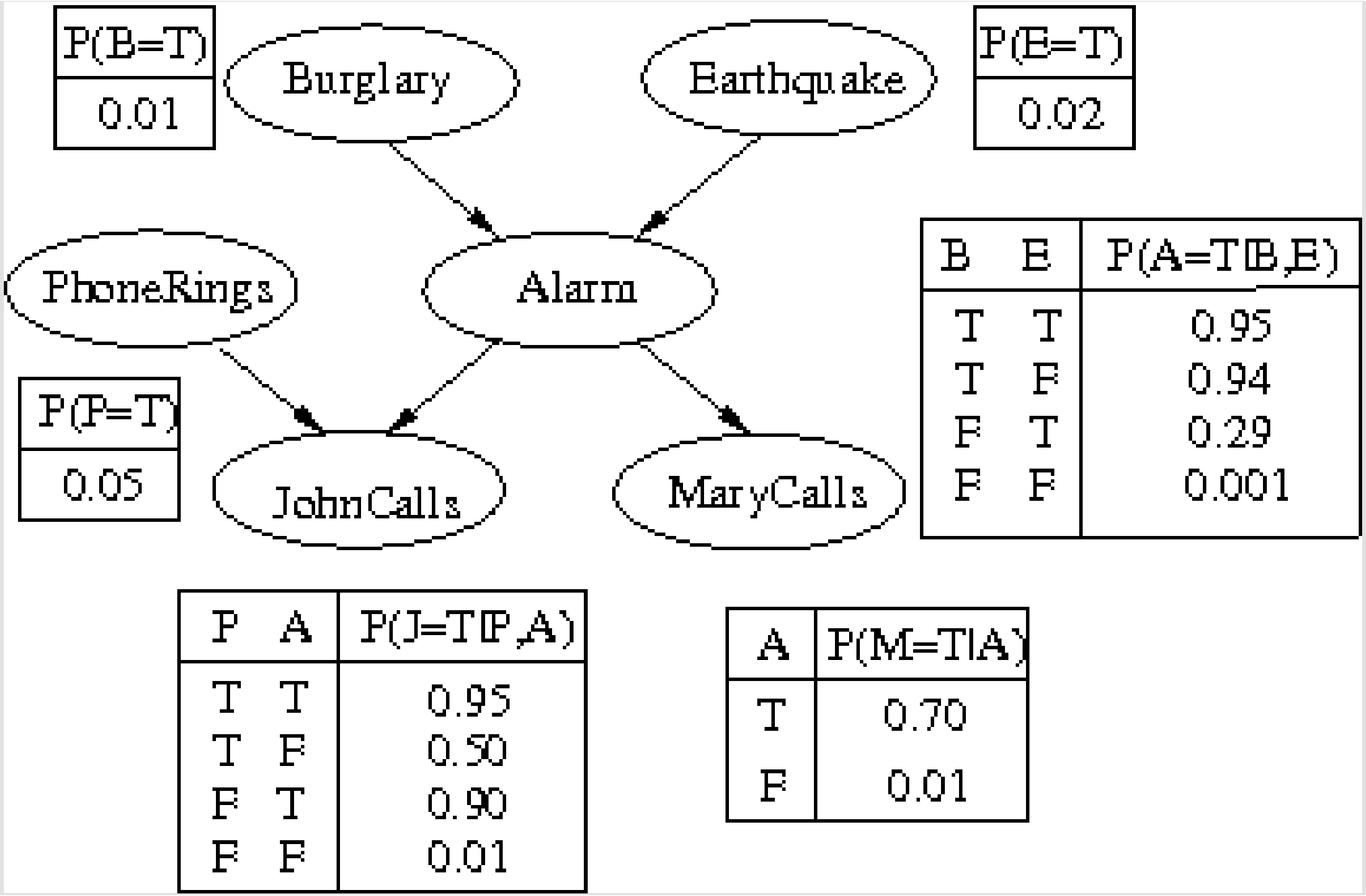
You have a new burglar alarm installed. It reliably detects burglary, but also responds to minor earthquakes. Two neighbours, John and Mary, promise to call the police when they hear the alarm. John always calls when he hears the alarm, but sometimes confuses the alarm with the phone ringing and calls then also. On the other hand, Mary likes loud music and sometimes doesn't hear the alarm.

Given evidence about who has and hasn't called, you'd like to estimate the probability of a burglary.

Earthquake Example: Nodes and Values

Node name	Type	Values
B: Burglary	Boolean	{T,F}
A: Alarm (goes off)	Boolean	{T,F}
M: Mary calls	Boolean	{T,F}
J: John calls	Boolean	{T,F}
P: Phone rings	Boolean	{T,F}
E: Earthquake	Boolean	{T,F}

BN for Earthquake Example



P(B=T)
0.01

P(E=T)
0.02

P(P=T)
0.05

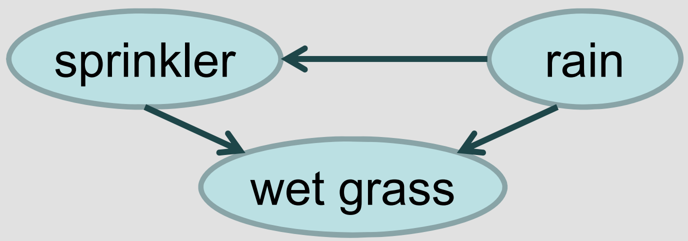
B	E	P(A=T|B,E)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

P	A	P(J=T|P,A)
T	T	0.95
T	F	0.50
F	T	0.90
F	F	0.01

A	P(M=T|A)
T	0.70
F	0.01

Causality?

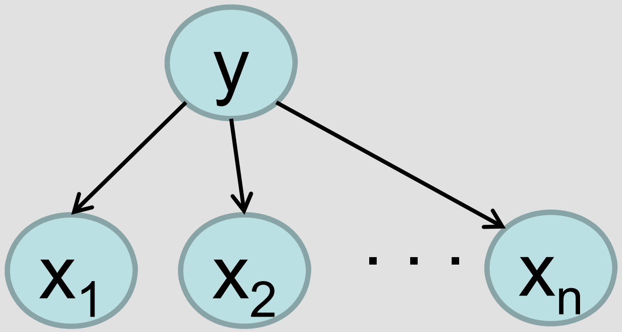
- **When Bayesian networks reflect causal patterns:**
  - Often simpler (nodes have fewer parents)
  - Often easier to think about
  - Often easier to elicit from experts
- **BNs need not actually be causal, but it is good practice**



  - Arrows reflect correlation, not causation
- **What do the arrows really mean?**
  - Topology may happen to encode causal structure
  - Topology really encodes conditional independence

Example: Naïve Bayes

- **Imagine we have one cause *y* and several effects *x*:**



$$\Pr(y, x_1, x_2, \ldots, x_n)$$
$$= \Pr(y)\Pr(x_1|y)\Pr(x_2|y) \ldots \Pr(x_n|y)$$

- **This is a *naïve Bayes* model**

Size of a Bayes Net

- **How big is a joint distribution over N Boolean variables?**

  2^N

- **How big is an N-node net if each node has up to k parents?**

  $O(N \times 2^{k+1})$

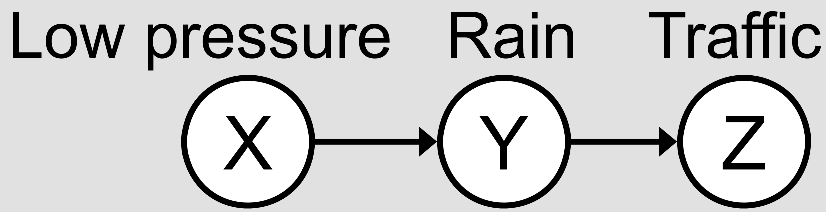
- **Both give the power to calculate $Pr(X_1,X_2,\ldots,X_N)$, but BNs give huge space savings!**
- **Also easier to elicit local CPTs**

Conditional Independence and BN Structure

- **The relationship between conditional independence and BN structure is important for understanding how BNs work**
- **Factors that affect conditional independence**
  + Causal chains
  + Common causes
  − Common effects

Causal Chains

- **A causal chain of events**



$$\mathbf{Pr}(x, y, z) = \mathbf{Pr}(x)\,\mathbf{Pr}(y|x)\,\mathbf{Pr}(z|y)$$

- **Is Z independent of X given Y?** ***Yes!***

$$\mathbf{Pr}(z|x, y) = \frac{\mathbf{Pr}(x, y, z)}{\mathbf{Pr}(x, y)} = \frac{\cancel{\mathbf{Pr}(x)}\cancel{\mathbf{Pr}(y|x)}\mathbf{Pr}(z|y)}{\cancel{\mathbf{Pr}(x)}\cancel{\mathbf{Pr}(y|x)}}$$

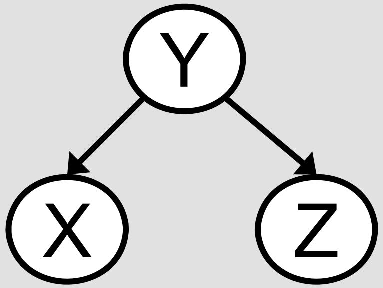
$$= \mathbf{Pr}(z|y)$$

- **Evidence along the chain “blocks” the influence**

Common Cause

- **Two effects of the same cause**
  - Are X and Z independent? ***No***
  - Are X and Z independent given Y? ***Yes!***

Y: Project due
X: Newsgroup busy
Z: Lab full



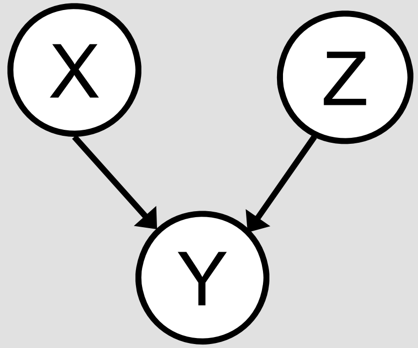
$$\Pr(z|x,y) = \frac{\Pr(x,y,z)}{\Pr(x,y)} = \frac{\Pr(\cancel{y})\Pr(\cancel{x|}y)\Pr(z|y)}{\Pr(\cancel{y})\Pr(\cancel{x|y})} = \Pr(z|y)$$

- **Observing the cause blocks influence between the effects**

Common Effect

- **Two causes of one effect (v-structures)**
- **Are X and Z independent?**
  - **Yes**: the ballgame and the rain cause traffic, but they are not correlated
- **Are X and Z independent given Y?**
  - **No:** seeing traffic puts the rain and the ballgame in competition as explanation
- **This is different from the other cases**
  - Observing an effect activates the influence between possible causes

X: Rain
Z: Ballgame
Y: Traffic



The General Case

- **Any complex example can be analyzed using these three canonical cases**
- **General question: in a given BN, are two variables independent (given evidence)?**
- **Solution: analyze the graph**

Direction-dependent Separation

- **Graphical criterion of conditional independence**
- **We can determine whether a set of nodes X is independent of another set Y, given a set of evidence nodes E, via the Markov property**
  – If a set of nodes X and a set of nodes Y are ***d-separated*** by evidence E, then X and Y are conditionally independent (via the Markov property)
- **D-separation is a property of the evidence**
  – One can say that evidence E d-separates two nodes
    > This happens when **all the paths** between these nodes are **blocked**

D-separation – Path

- ***Path (Undirected path):*** **A path between two *sets of nodes* X and Y is any sequence of nodes between a member of X and a member of Y such that every adjacent pair of nodes is connected by an arc (regardless of direction), and no node appears in the sequence twice**

D-separation – Blocked Path

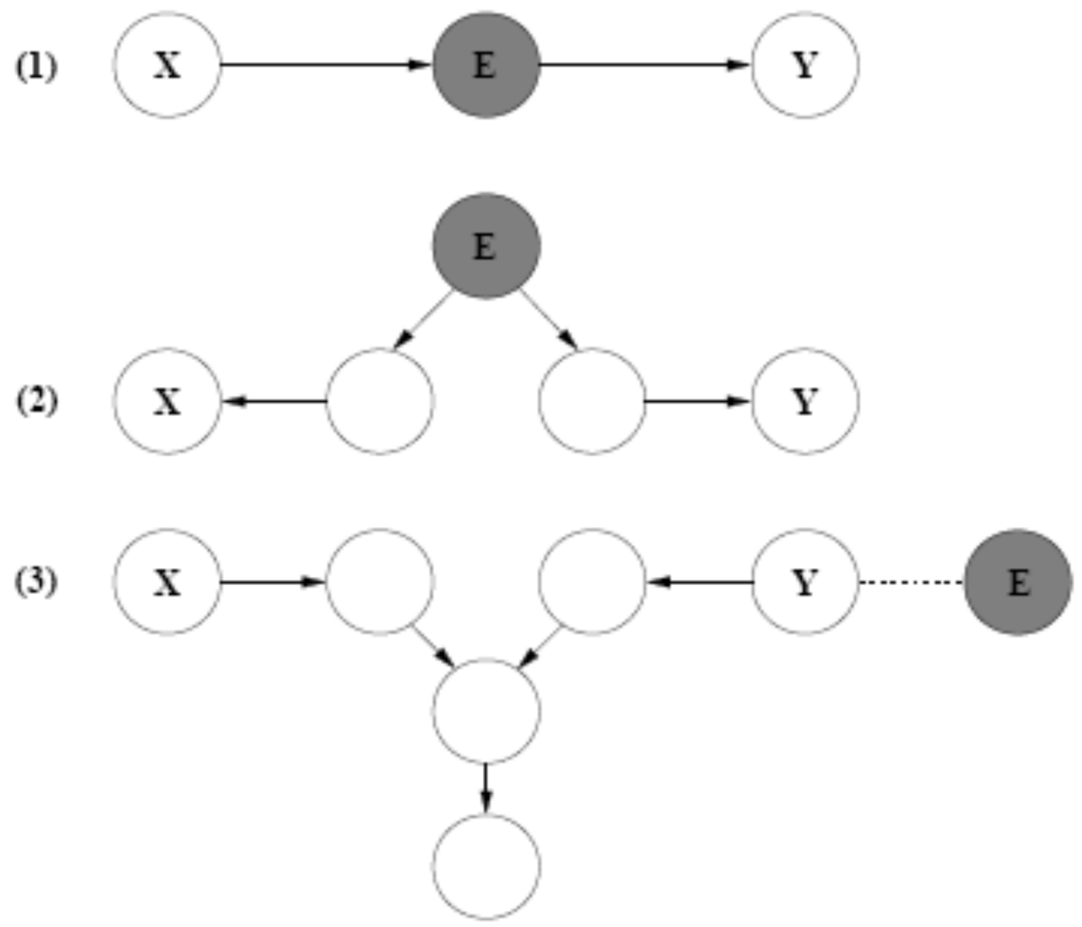
- ***Blocked path*: A path is *blocked* given a set of nodes E, if there is a node Z on the path for which at least one of three conditions holds:**
  1. Z is in **E** and Z has one arrow on the path leading in and one arrow out (**chain**)
  2. Z is in **E** and Z has both path arrows leading out (**common cause**)
  3. Neither Z nor any descendant of Z is in **E**, and both path arrows lead into Z (**common effect**)
- **A set of nodes E *d-separates* two sets of nodes X and Y, if <u>every</u> undirected path from a node in X to a node in Y is *blocked* given E**

Determining D-separation

Chain

Common cause

Common effect

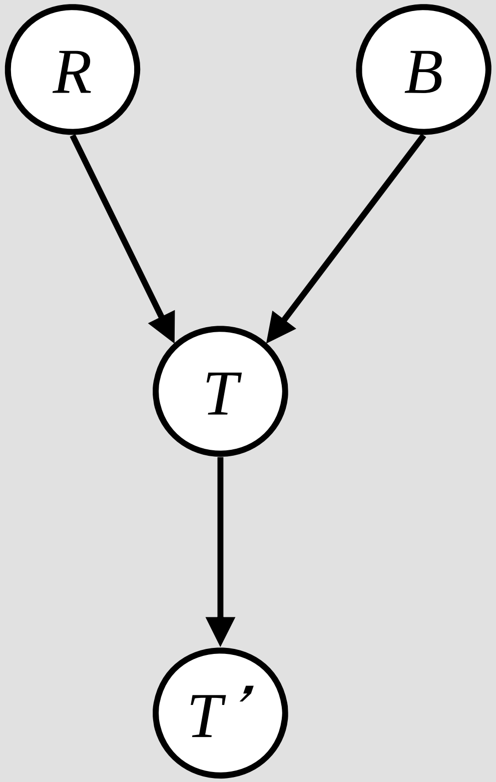


D-separation – Example (I)

$R \perp\!\!\!\perp B$

$R \perp\!\!\!\perp B | T$

$R \perp\!\!\!\perp B | T'$



D-separation – Example (II)

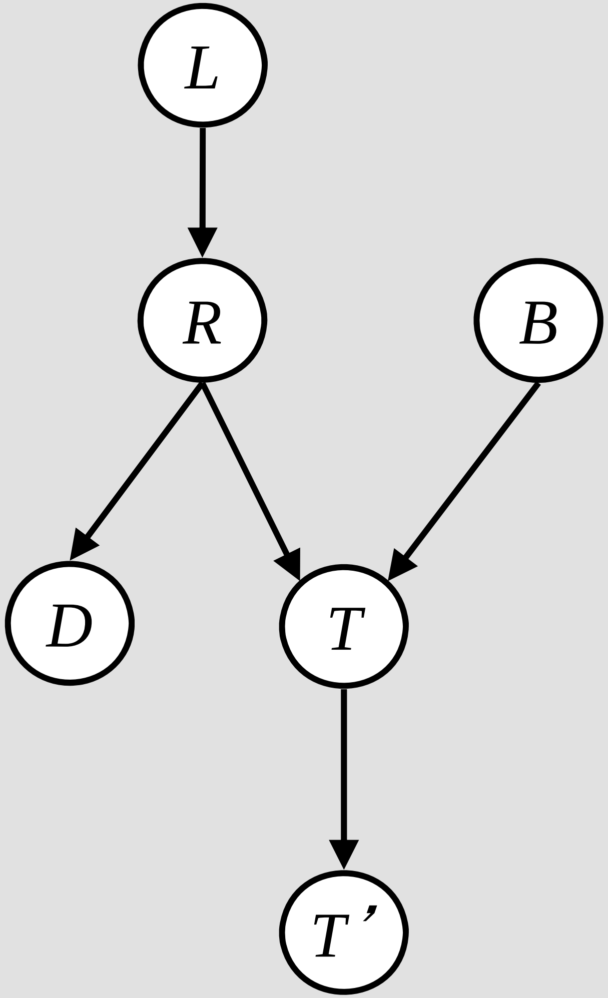
$L \perp\!\!\!\perp T' | T$

$L \perp\!\!\!\perp B$

$L \perp\!\!\!\perp B | T$

$L \perp\!\!\!\perp B | T'$

$L \perp\!\!\!\perp B | T, R$



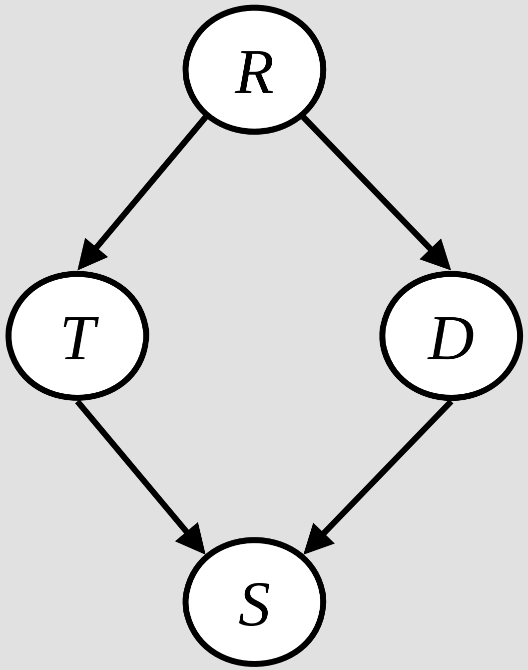
D-separation – Example (III)

- **Variables:**
  - R: Raining
  - T: Traffic
  - D: Roof drips
  - S: I' m sad
- **Questions:**

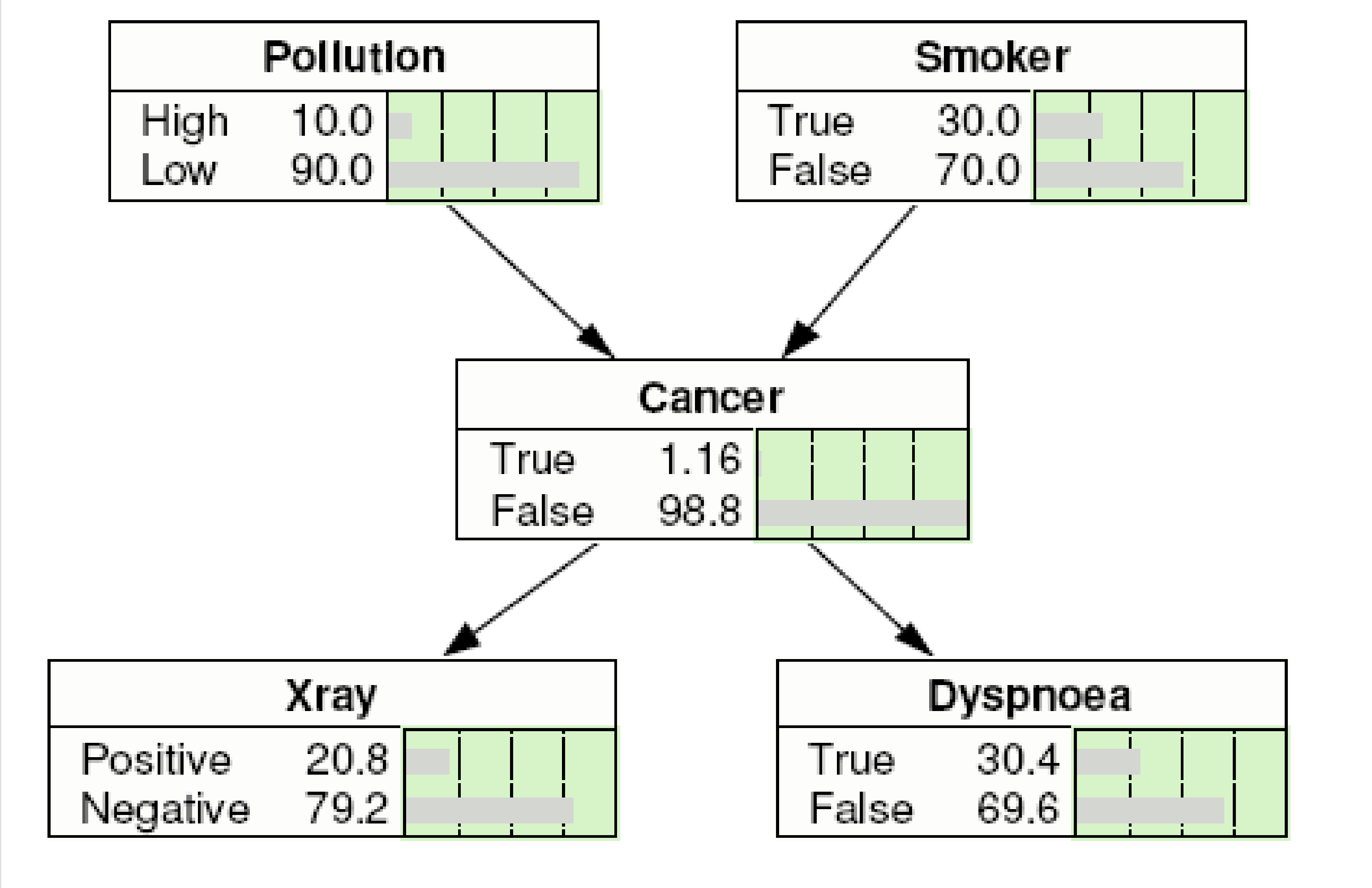
$$T \perp\!\!\!\perp D$$

$$T \perp\!\!\!\perp D | R$$

$$T \perp\!\!\!\perp D | R, S$$



Reasoning with Numbers – Using Netica Software



MONASH University
Information Technology

FIT5047 – Intelligent Systems

Decision Networks

Decision Networks

- **Extension of BNs to support making decisions**
- **Utility theory represents preferences between different outcomes of various plans**
- **Decision theory =**
  Utility theory + Probability theory

Expected Utility

$$EU(A \mid E) = \sum_i \Pr(O_i \mid E, A) U(O_i \mid A)$$

- E = available evidence
- A = a non-deterministic action
- O_i = a possible outcome state
- U = utility

Decision Networks

A Decision network represents information about

- the agent's current state
- its possible actions
- the state that will result from the agent's action
- the utility of that state

Also called, *Influence Diagrams* (Howard & Matheson, 1981)

Types of Nodes

- ***Chance nodes*** **– (ovals) random variables**
  - Have an associated CPT
  - Parents can be decision nodes and other chance nodes
- ***Decision nodes*** **– (rectangles) points where the decision maker has a choice of actions**
  - The table is the decision with the highest computed EU for each combination of evidence in the ***information link*** parents
- ***Utility nodes (Value nodes)*** **– (diamonds) the agent's utility function**
  - The table represents a multi-attribute utility function
  - Parents are variables describing the outcome states that directly affect utility

Types of Links

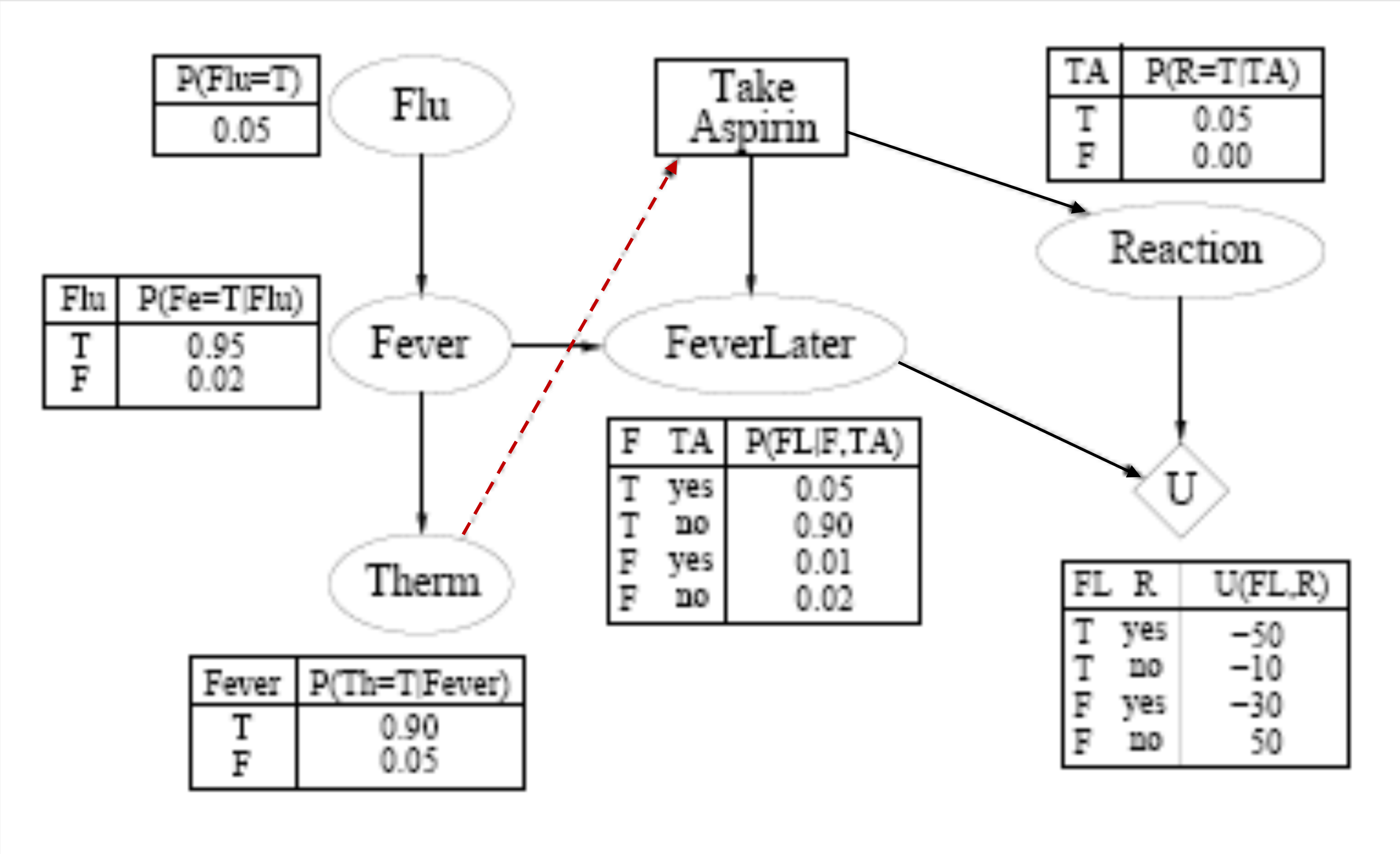
- ***Informational Links*** **– indicate when a chance node needs to be observed before a decision is made**
  - – Any link entering a decision node is an informational link
- ***Conditioning links*** **– indicate the variables on which the probability assignment to a chance node will be conditioned**

Fever Problem Description

Suppose that you know that a fever can be caused by the flu. You can use a thermometer, which is fairly reliable, to test whether or not you have a fever.

Suppose you also know that if you take aspirin it will almost certainly lower a fever to normal. Some people (about 5% of the population) have a negative reaction to aspirin. You'll be happy to get rid of your fever, so long as you don't suffer an adverse reaction if you take aspirin.

Fever Decision Network



P(Flu=T)
0.05

Flu	P(Fe=T|Flu)
T	0.95
F	0.02

Fever	P(Th=T|Fever)
T	0.90
F	0.05

TA	P(R=T|TA)
T	0.05
F	0.00

F	TA	P(FL|F,TA)
T	yes	0.05
T	no	0.90
F	yes	0.01
F	no	0.02

FL	R	U(FL,R)
T	yes	−50
T	no	−10
F	yes	−30
F	no	50

Fever Decision Table

Evidence	*Bel(Flu=T)*	*EU(TA=yes)*	*EU(TA=no)*	Decision
None	0.046	45.27	45.29	no
Th=F	0.018	45.40	48.41	no
Th=T	0.273	44.12	19.13	yes
Th=T & Re=T	0.033	-30.32	0	no

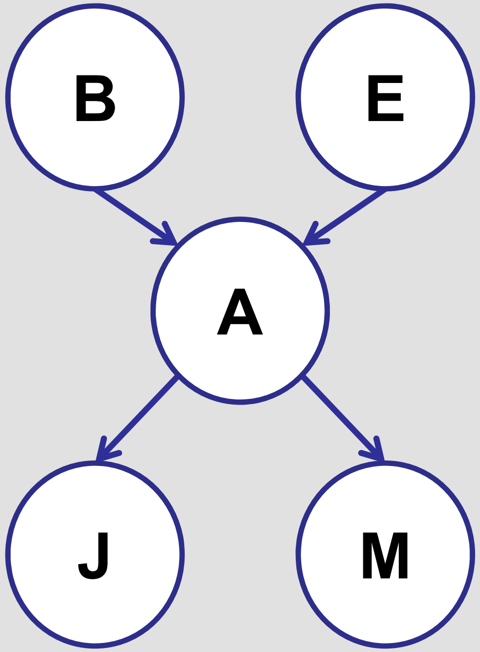
MONASH University
Information Technology

FIT5047 – Intelligent Systems

Exact Inference in Bayesian Networks

Inference

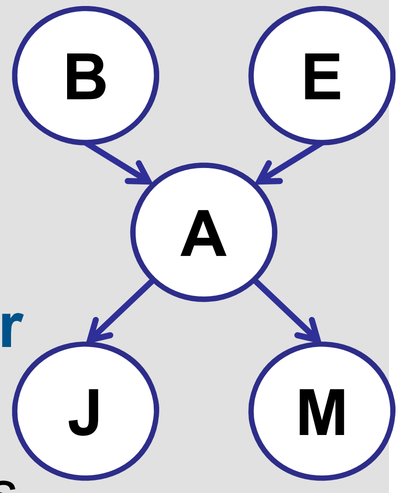
- **Calculating some useful quantity from a joint probability distribution**
- **Posterior probability of a query variable Q**
  $\Pr(Q|E_1 = e_1, \dots, E_k = e_k)$
- **Most likely explanation of evidence**
  $\operatorname{argmax}_q \Pr(Q = q|\, E_1 = e_1, \dots, E_k = e_k)$



Inference by Enumeration

- **Given unlimited time, inference in BNs is easy**
- **Recipe:**
  - State the **marginal** probabilities you need
  - Figure out ALL the **joint** probabilities you need
  - Calculate and combine them
- **Example:**

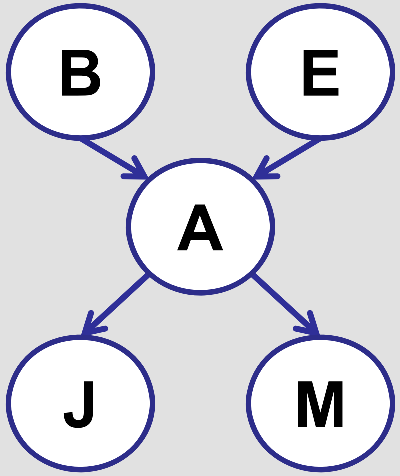
$$\Pr(+b \mid +j,+m) = \frac{\Pr(+b,+j,+m)}{\Pr(+j,+m)}$$



- **There is no need to calculate the denominator**
  - Compute the numerator for $\Pr(\neg b|+j,+m)$
  - Replace the denominator with α in both equations
  - $\Pr(+b|+j,+m) + \Pr(\neg b|+j,+m) =$
    $\alpha \Pr(+b,+j,+m) + \alpha \Pr(\neg b,+j,+m) = 1$

Enumeration – Example (I)

- **In this simple method, we only need the BN to synthesize the joint entries from the CPTs**

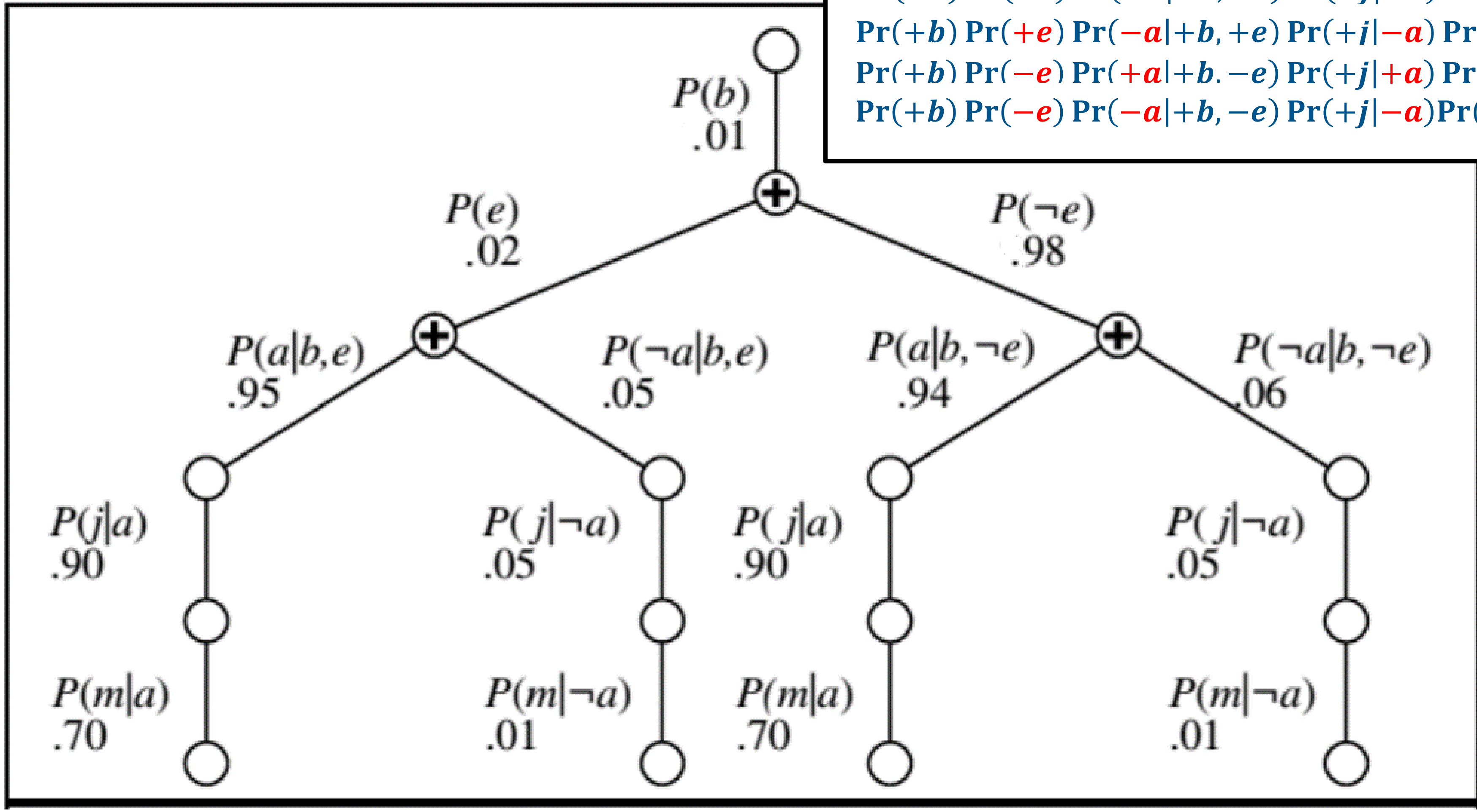


$$\Pr(+b, +j, +m) =$$

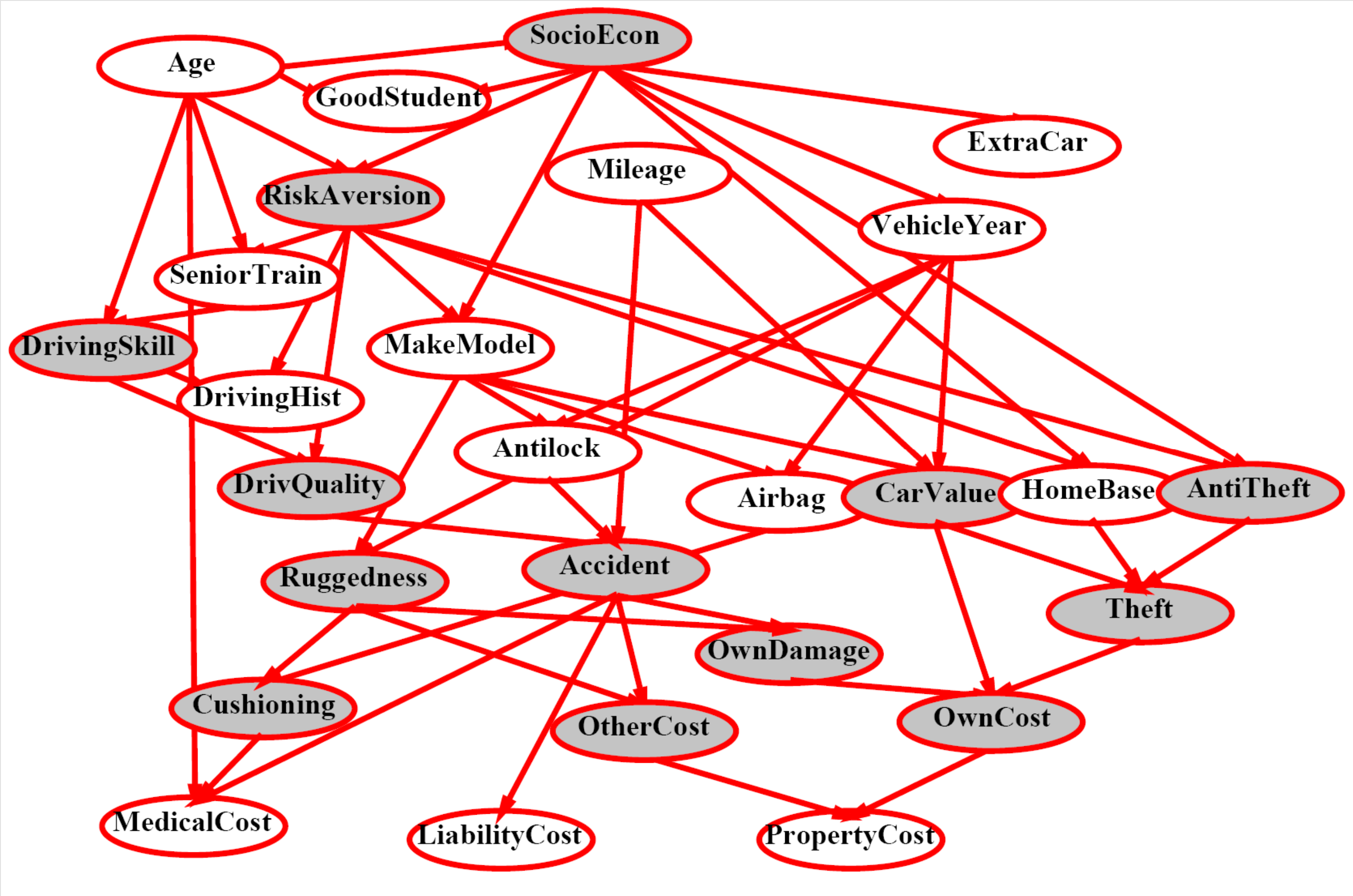
$$\Pr(+b)\Pr(+e)\Pr(+a|+b,+e)\Pr(+j|+a)\Pr(+m|+a) +$$
$$\Pr(+b)\Pr(+e)\Pr(-a|+b,+e)\Pr(+j|-a)\Pr(+m|-a) +$$
$$\Pr(+b)\Pr(-e)\Pr(+a|+b,-e)\Pr(+j|+a)\Pr(+m|+a) +$$
$$\Pr(+b)\Pr(-e)\Pr(-a|+b,-e)\Pr(+j|-a)\Pr(+m|-a)$$

Enumeration – Example (II)

$$
\begin{aligned}
\Pr(+b,+j,+m) = {} & \\
& \Pr(+b)\Pr(+e)\Pr(+a|+b,+e)\Pr(+j|+a)\Pr(+m|+a) \\
& \Pr(+b)\Pr(+e)\Pr(-a|+b,+e)\Pr(+j|-a)\Pr(+m|-a) \\
& \Pr(+b)\Pr(-e)\Pr(+a|+b,-e)\Pr(+j|+a)\Pr(+m|+a) \\
& \Pr(+b)\Pr(-e)\Pr(-a|+b,-e)\Pr(+j|-a)\Pr(+m|-a)
\end{aligned}
$$



Inference by Enumeration?



Variable Elimination (VE)

- **Why is inference by enumeration so slow?**
  - You join up the whole joint distribution before you sum out the hidden variables
  - You end up repeating a lot of work!
- **Idea: interleave joining and marginalizing**
  - Called "Variable Elimination"
  - Still NP-hard, but usually much faster than inference by enumeration

Factors – Joint Distributions

- **Joint distribution: Pr(X,Y)**
  - Entries Pr(x,y) for all x, y
  - Sums to 1

Pr(T,W)

T	W	Pr
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

- **Selected joint: Pr(x,Y)**
  - A slice of the joint distribution
  - Entries Pr(x,y) for a fixed x, all y
  - Sums to Pr(x)

Pr(cold,W)

T	W	Pr
cold	sun	0.2
cold	rain	0.3

Factors – Conditionals

- **Single conditional: Pr(Y | x)**
  – Entries Pr(y | x) for a fixed x, all y
  – Sums to 1

Pr(W|cold)

T	W	Pr
cold	sun	0.4
cold	rain	0.6

Pr(sun|cold)
Pr(rain|cold)

- **Family of conditionals: Pr(Y | X)**
  – Entries Pr(y | x) for all x, y
  – Sums to |X|

Pr(W|T)

T	W	Pr
hot	sun	0.8
hot	rain	0.2
cold	sun	0.4
cold	rain	0.6

Pr(W|hot) Pr(W|cold)

- **Specific family: Pr(y | X)**
  – Entries Pr(y | x) for a fixed y, all x
  – Sums to … who knows!

Pr(rain|T)

T	W	Pr
hot	rain	0.2
cold	rain	0.6

Pr(rain|hot)
Pr(rain|cold)

Factors in General

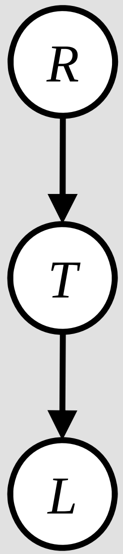
When we write $Pr(Y_1 \ldots Y_N \mid X_1 \ldots X_M)$

- **It is a *factor* – a multi-dimensional array**
- **Its values are all values $Pr(y_1 \ldots y_N \mid x_1 \ldots x_M)$**
- **Any value assigned to X or Y is a dimension that is missing from the array**

Example: Traffic Domain

- **Random Variables**
  - R: Raining
  - T: Traffic
  - L: Late for class

- **Query: Pr(L)**



Pr(R)

+r	0.1
-r	0.9

Pr(T|R)

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

Pr(L|T)

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

Variable Elimination (VE) – Outline

- **Track factors**
  - Initial factors are local CPTs (one per node)

Pr(R)

+r	0.1
-r	0.9

Pr(T|R)

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

Pr(L|T)

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

- **Select any known values**
  - E.g., if we know $L=+l$, the initial factors are

Pr(R)

+r	0.1
-r	0.9

Pr(T|R)

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

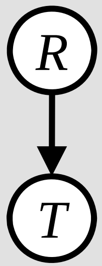
Pr($+l$ |T)

+t	+l	0.3
-t	+l	0.1

- **VE: Alternately join factors and eliminate variables**

Operation 1: Join Factors

- **First basic operation: joining factors**
  - Get all factors over the joining variable
  - Build a new factor over the union of the variables
  - Just like a database join
- **Example: Join on R**



Pr(R)

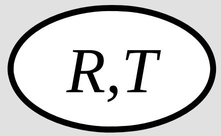
+r	0.1
-r	0.9

Pr(T|R)

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

Pr(R,T)

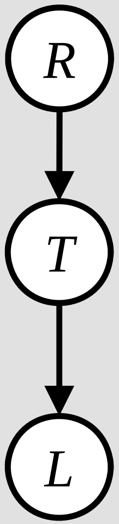
+r	+t	0.08
+r	-t	0.02
-r	+t	0.09
-r	-t	0.81



- **Computation for each entry: *pointwise product***

$$\forall r, t \qquad \Pr(r,t) = \Pr(r) \cdot \Pr(t|r)$$

Pr(L) Example: Multiple Joins



Pr(R)

+r	0.1
-r	0.9

Pr(T|R)

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

Pr(L|T)

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

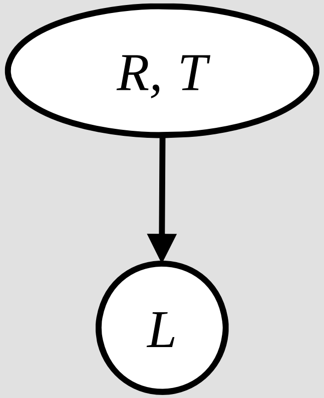
Join R

Pr(R,T)

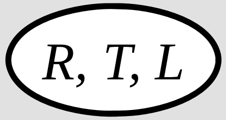
+r	+t	0.08
+r	-t	0.02
-r	+t	0.09
-r	-t	0.81

Pr(L|T)

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9



Join T



Pr(R,T,L)

+r	+t	+l	0.024
+r	+t	-l	0.056
+r	-t	+l	0.002
+r	-t	-l	0.018
-r	+t	+l	0.027
-r	+t	-l	0.063
-r	-t	+l	0.081
-r	-t	-l	0.729

Operation 2: Eliminate

- **Second basic operation: marginalize**
  - Take a factor and sum out a variable
    - > Shrinks a factor to a smaller one
    - > A projection operation
- **Example:**

Pr(R,T)

+r	+t	0.08
+r	-t	0.02
-r	+t	0.09
-r	-t	0.81

Sum out R →

Pr(T)

+t	0.17
-t	0.83

MONASH University
Information Technology

Pr(L) Example: Multiple Eliminations

R, T, L → T, L → L

Pr(R,T,L)

+r	+t	+l	0.024
+r	+t	-l	0.056
+r	-t	+l	0.002
+r	-t	-l	0.018
-r	+t	+l	0.027
-r	+t	-l	0.063
-r	-t	+l	0.081
-r	-t	-l	0.729

Sum out R

Pr(T,L)

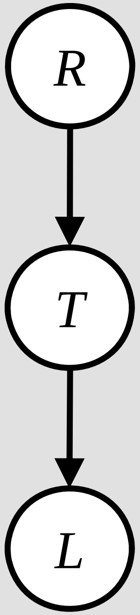
+t	+l	0.051
+t	-l	0.119
-t	+l	0.083
-t	-l	0.747

Sum out T

Pr(L)

+l	0.134
-l	0.886

Pr(L) Example: Marginalizing Early (I)



Pr(R)

+r	0.1
-r	0.9

Pr(T|R)

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

Pr(L|T)

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

Join R

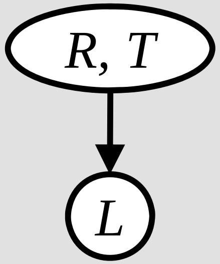
Pr(R,T)

+r	+t	0.08
+r	-t	0.02
-r	+t	0.09
-r	-t	0.81

Pr(L|T)

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

Sum out R

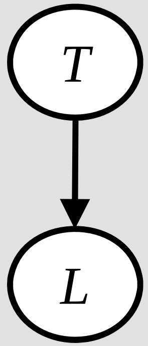


Pr(T)

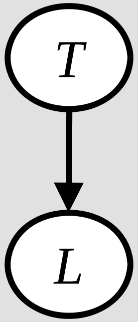
+t	0.17
-t	0.83

Pr(L|T)

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9



Pr(L) Example: Marginalizing Early (II)



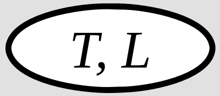
Pr(T)

+t	0.17
-t	0.83

Pr(L|T)

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

Join T



Pr(T,L)

+t	+l	0.051
+t	-l	0.119
-t	+l	0.083
-t	-l	0.747

Sum out T

Pr(L)

+l	0.134
-l	0.886

Evidence (I)

- **If no evidence, use the initial factors**

Pr(R)

+r	0.1
-r	0.9

Pr(T|R)

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

Pr(L|T)

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

- **If evidence, start with factors that select that evidence**
  - Given +r, we can compute Pr(L|+r) → the initial factors become:

Pr(+r)

+r	0.1

Pr(T|+r)

+r	+t	0.8
+r	-t	0.2

Pr(L|T)

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

- **Eliminate all variables other than query + evidence**

Evidence (II)

1. Perform a selected join of query and evidence

2. Normalize it

Example: Pr(L | +r)

1. Pr(+r)

+r	0.1

Pr(T|+r)

+r	+t	0.8
+r	-t	0.2

Join +r →

Pr(+r,T)

+r	+t	0.08
+r	-t	0.02

Pr(+r,T)

+r	+t	0.08
+r	-t	0.02

Pr(L|T)

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

Join T →

Pr(+r,T,L)

+r	+t	+l	0.024
+r	+t	-l	0.056
+r	-t	+l	0.002
+r	-t	-l	0.018

Sum out T →

Pr(+r,L)

+r	+l	0.026
+r	-l	0.074

Normalize →

Pr(L|+r)

+l	0.26
-l	0.74

General Variable Elimination

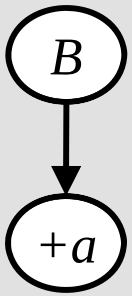
Query: $\mathbf{Pr}(Q|E_1 = e_1, \dots, E_k = e_k)$

1. **Start with initial factors:**
   - Local CPTs (but instantiated by evidence)
2. **While there are hidden variables (not Q or evidence):**
   a. Select a hidden variable H
   b. Join all factors mentioning H
   c. Eliminate (sum out) H
3. **Join all remaining factors and normalize**

Variable Elimination and Bayes Rule

Pr(B)

B	Pr
+b	0.1
¬b	0.9



Pr(B|+a)= ?

Bayes rule: $\Pr(B|+a) = \dfrac{\Pr(+a, B)}{\Pr(+a)}$

1. Start / Select

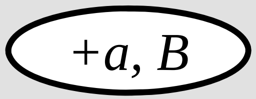
Pr(A|B)→Pr(+a|B)

B	A	Pr
+b	+a	0.8
~~b~~	~~¬a~~	~~0.2~~
¬b	+a	0.1
~~¬b~~	~~¬a~~	~~0.9~~

2. Join on B

Pr(+a,B)

A	B	Pr
+a	+b	0.08
+a	¬b	0.09



3. Normalize

Pr(B|+a)

A	B	Pr
+a	+b	8/17
+a	¬b	9/17

Summary: Bayesian Networks

- **Bayes nets compactly encode joint distributions**
- **BNs are a natural way to represent conditional independence information**
  - **qualitative**: links between nodes – independencies of distributions can be deduced from a BN graph structure by D-separation
  - **quantitative**: conditional probability tables (CPTs)
- **BN inference**
  - computes the probability of query variables given evidence variables
  - is flexible – we can enter evidence about any node and update beliefs in other nodes
  - using variable elimination is better than using enumeration

Reading and Software

- **Reading**
  - Russell, S. and Norvig, P. (2010), *Artificial Intelligence – A Modern Approach* (3rd ed), Prentice Hall
    - > Chapters 14.1-14.4.1, 16.1-2, 16.5
  - Korb, K. and Nicholson, A. (2010), *Bayesian Artificial Intelligence* (2nd ed), Chapman and Hall
    - > Chapters 1, 2, 3.1-3.3 and 4.1-4.4
- **Software**
  - Netica – http://www.norsys.com/

Next Lecture Topic

- **Lecture Topic 6**
  - Machine Learning