MONASH University
Information Technology

MONASH University
Information Technology

FIT5047: Intelligent Systems

Bayesian Networks I: Representation and Reasoning

Russell & Norvig Chapters 14.1-3

Korb & Nicholson, 2011: Chapter 2

Some slides are adapted from those of Stuart Russell, Andrew Moore, Dan Klein & Kevin Korb

Bayesian Reasoning: Learning Objectives

- **Bayesian AI**
- **Bayesian 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) – not covered in FIT5047
  - Object-oriented Bayesian networks – not covered in FIT5047)

Conditional Independence (reminder)

- **X and Y are independent if**
  $$\forall x, y \ \Pr(x, y) = \Pr(x)\Pr(y)$$
  $$\dashrightarrow \quad X \perp\!\!\!\perp Y$$

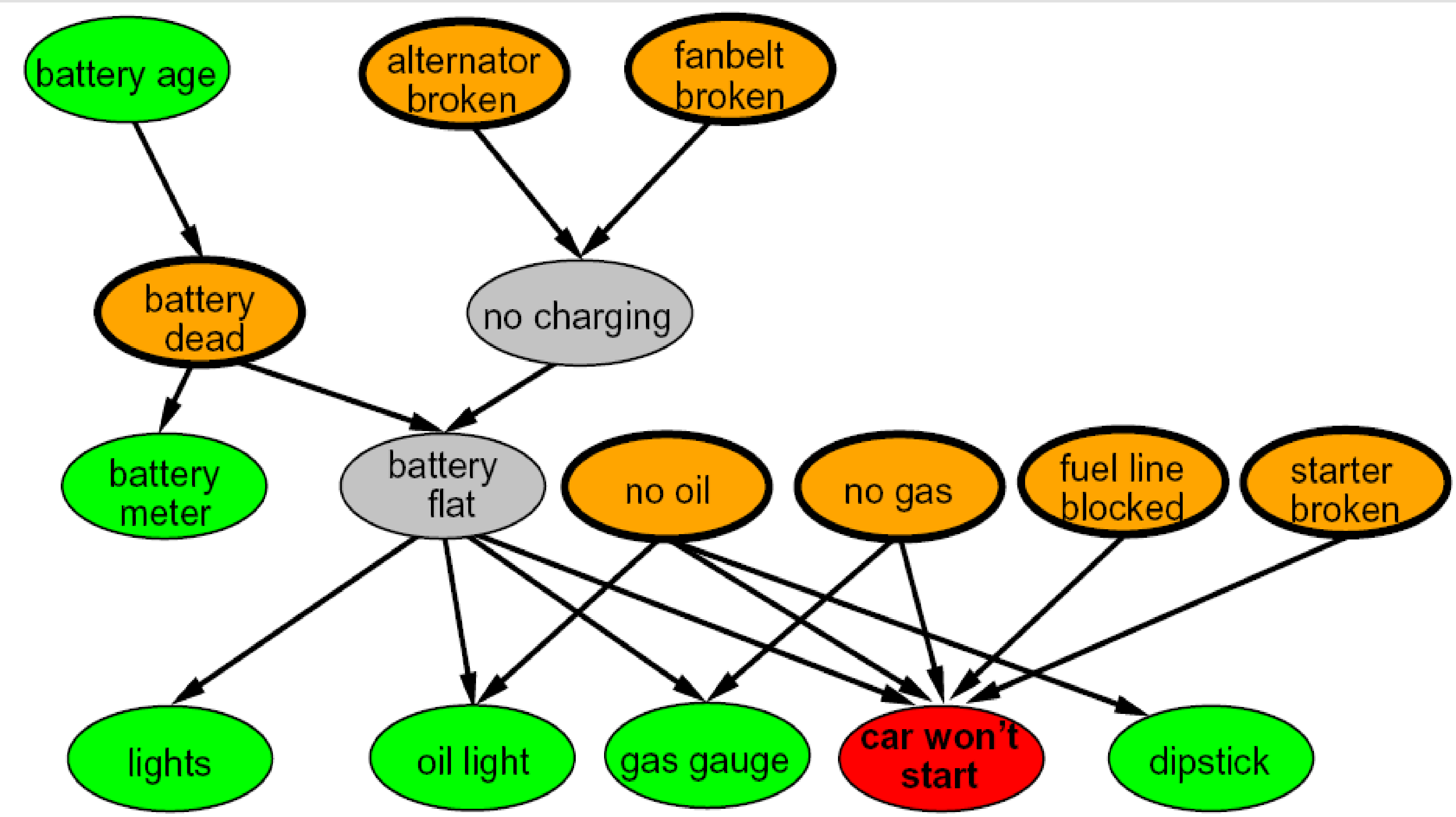
- **X and Y are conditionally independent given Z**
  $$\forall x, y, z \ \Pr(x, y|z) = \Pr(x|z)\Pr(y|z)$$
  $$\dashrightarrow \quad X \perp\!\!\!\perp Y | Z$$

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

Bayesian Networks: The Big Picture

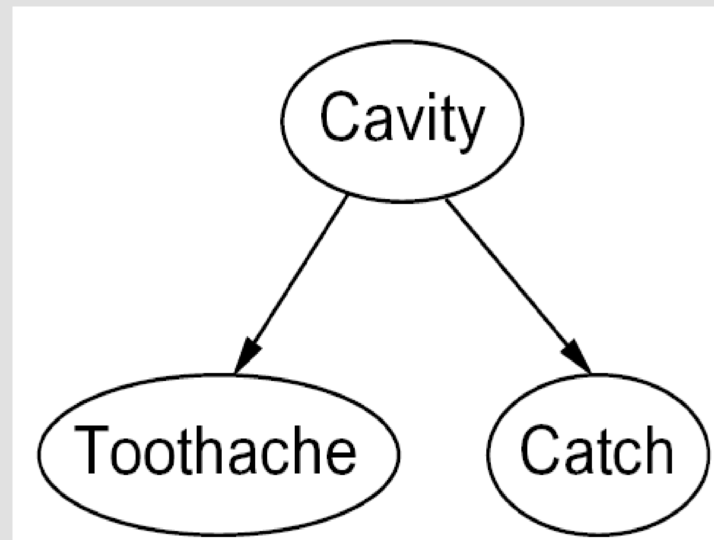
- **Two problems with using full joint distribution tables as our probabilistic models:**
  - Unless there are only a few variables, the joint is too big to represent explicitly
  - Hard to learn (estimate) anything empirically about more than a few variables at a time
- **Bayes nets (aka graphical models): 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



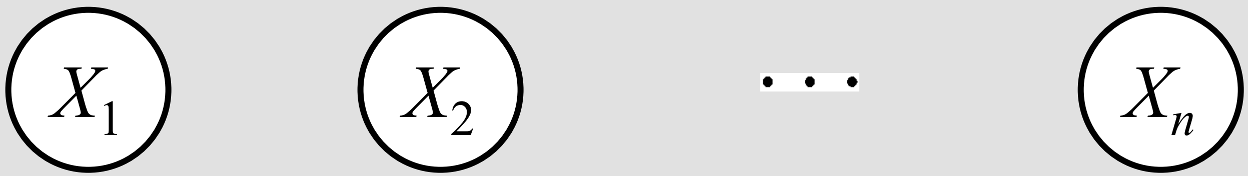
Graphical Model – Notation

- **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)

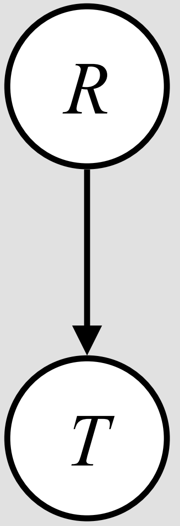
- **N independent coin flips**



- **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)

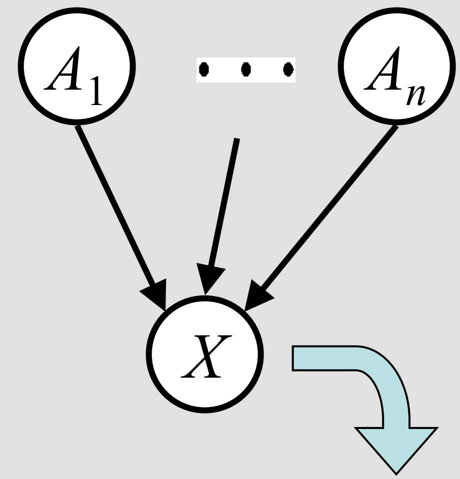
- **A data structure that represents the dependence between random variables**
- **A Bayesian Network is a directed acyclic graph (DAG) in which the following holds:**
  1. A set of random variables makes up the nodes in the network
  2. A set of directed links connects pairs of nodes
  3. Each node 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**

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

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



$P(X|A_1 \ldots A_n)$

Bayesian network = Topology (graph) + Local Conditional Probabilities

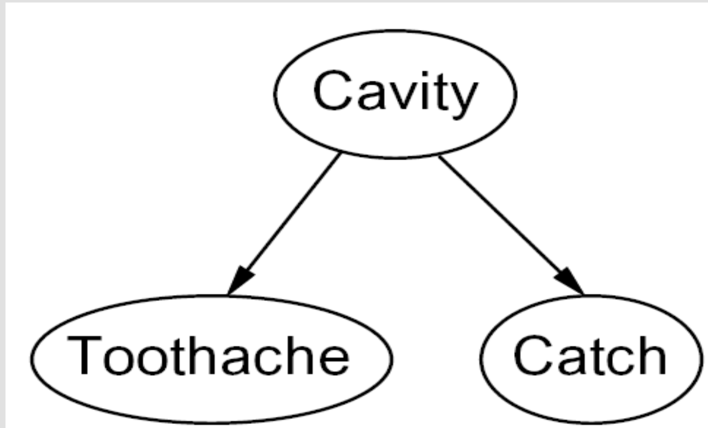
Probabilities in BNs

- **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**

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

- **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



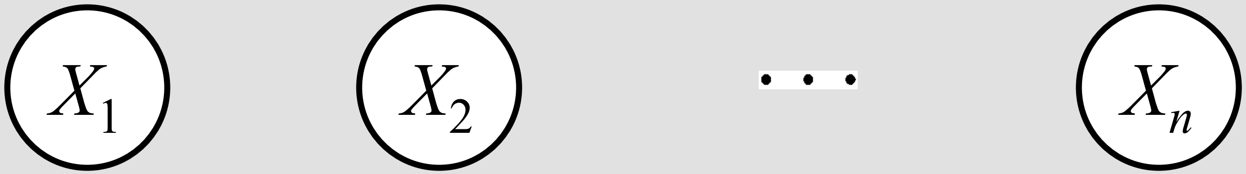
$$\Pr(+cavity, +catch, \neg toothache) = ?$$

$$= \Pr(\neg toothache | +cavity, +catch) \Pr(+catch | +cavity) \Pr(+cavity)$$

$$= \Pr(\neg toothache | +cavity) \Pr(+catch | +cavity) \Pr(+cavity)$$

Conditional independence

Example: Coin Flips (II)



$\Pr(X_1)$

h	0.5
t	0.5

$\Pr(X_2)$

h	0.5
t	0.5

…

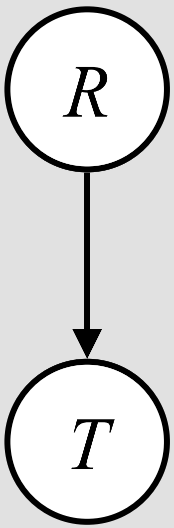
$\Pr(X_n)$

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.

(This example BN is taken from Korb&Nicholson, 2011 Section 2.2; it is a variant of the classic "Asia" BN, Lauritzen & Spiegelhalter, 1988, used in Neapolitan 1990, Netica tutorials, and so on)

Nodes and Values

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

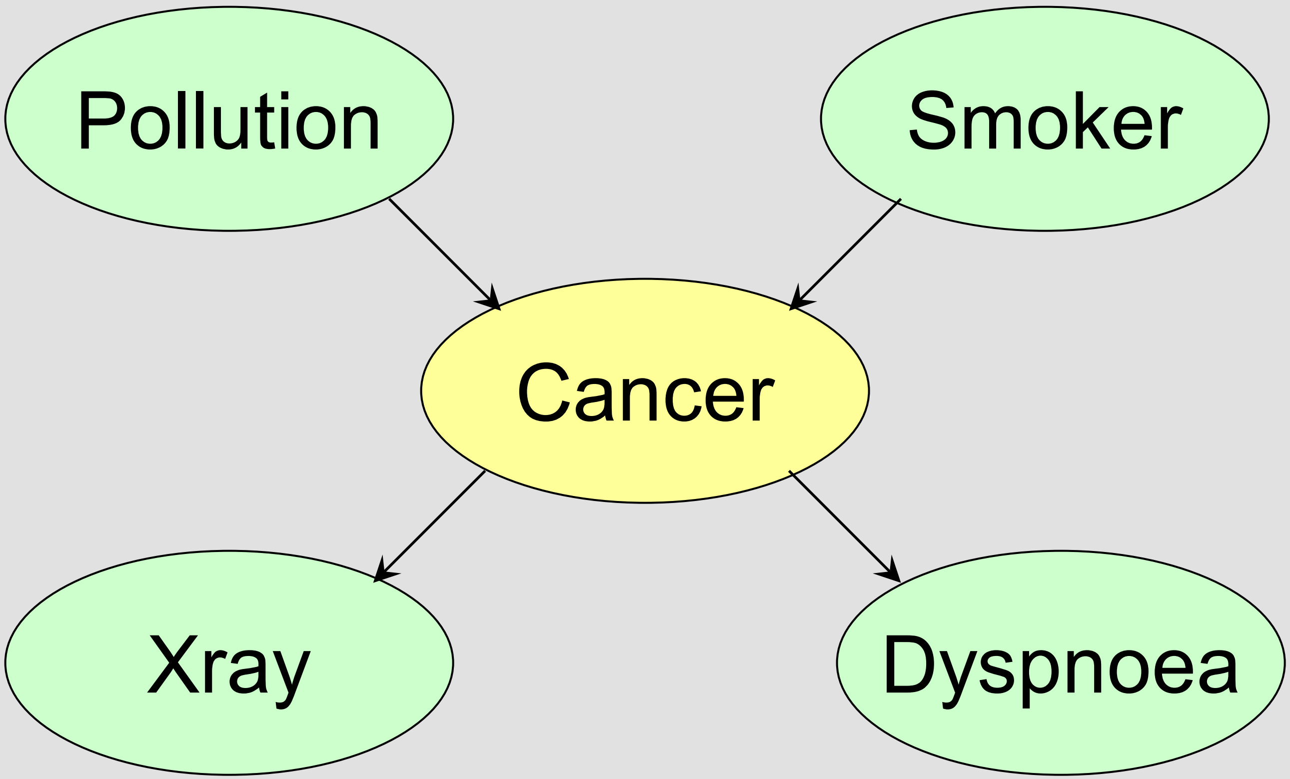
A: Nodes can be discrete or continuous

- **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

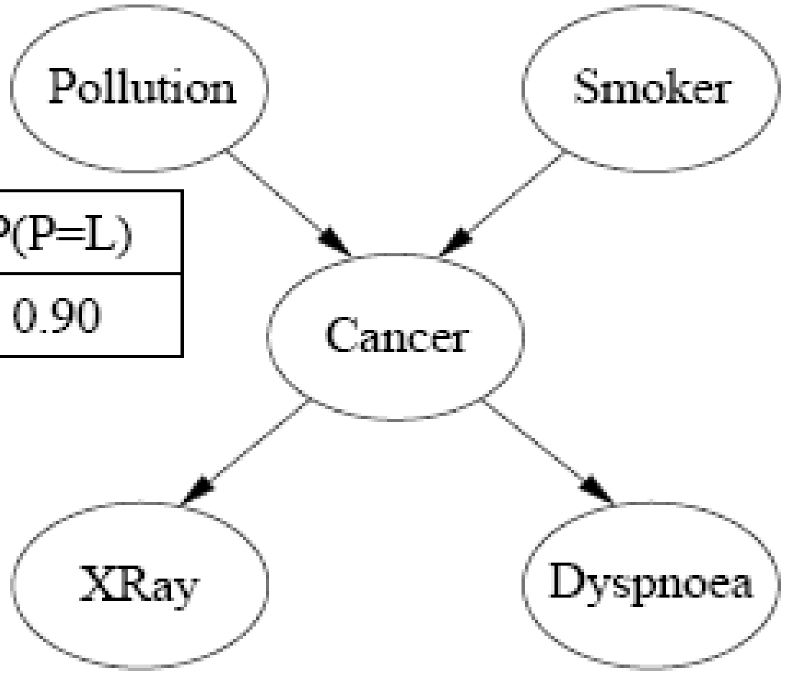


Conditional Probability Tables (CPTs)

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)**

Lung Cancer Example: CPTs



P(S=T)
0.30

The value for =F is the complement

P(P=L)
0.90

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

K&N Figure 2.1

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

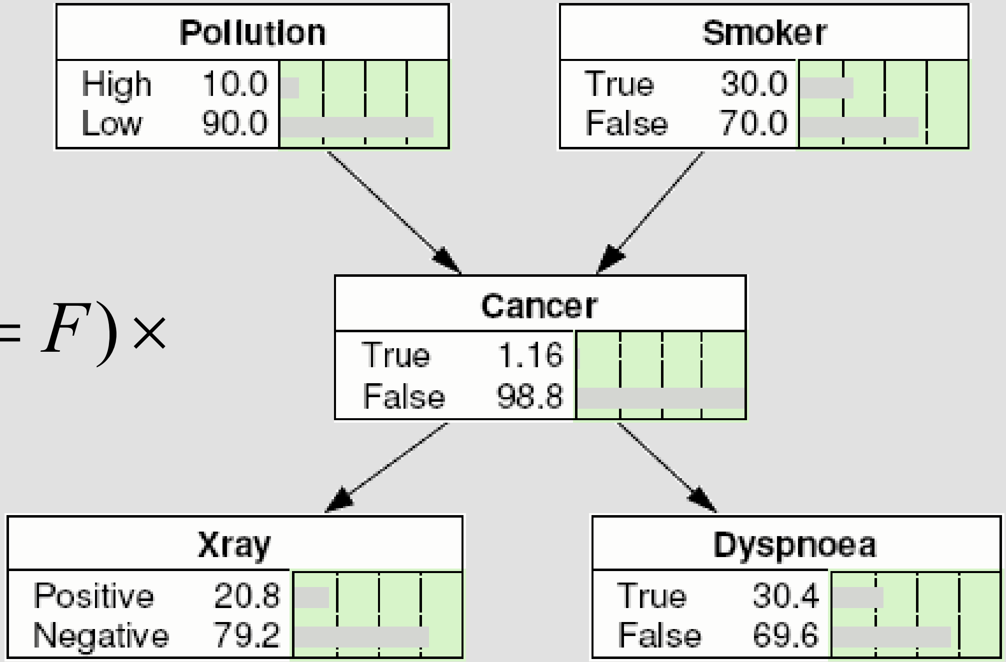
Representing Joint Probability Distribution: Example

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

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

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

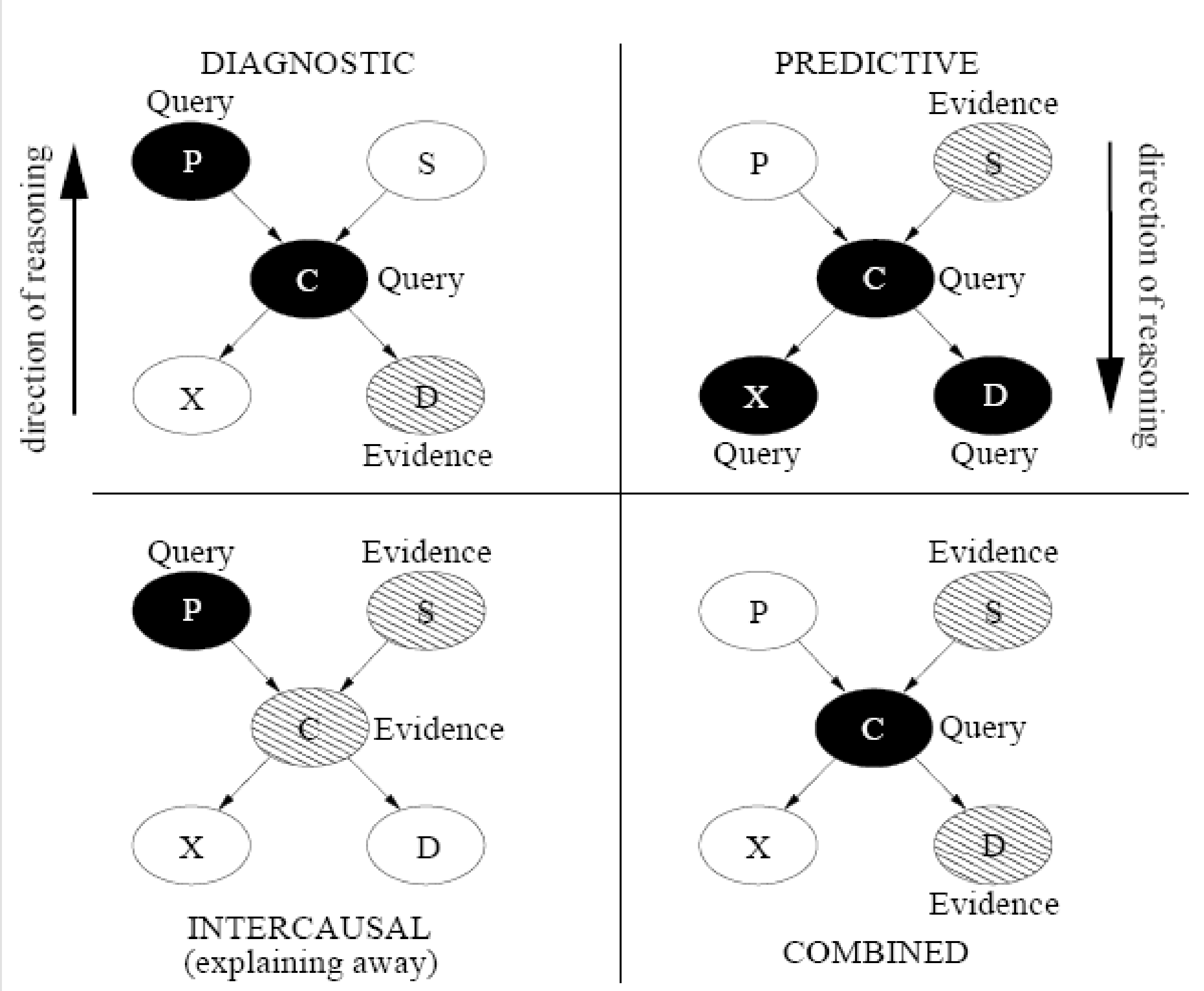
$$\Pr(X = pos \mid C = T) \times$$
$$\Pr(D = T \mid C = T)$$



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



K&N Figure 2.2

Example – Earthquake (Pearl 1988)

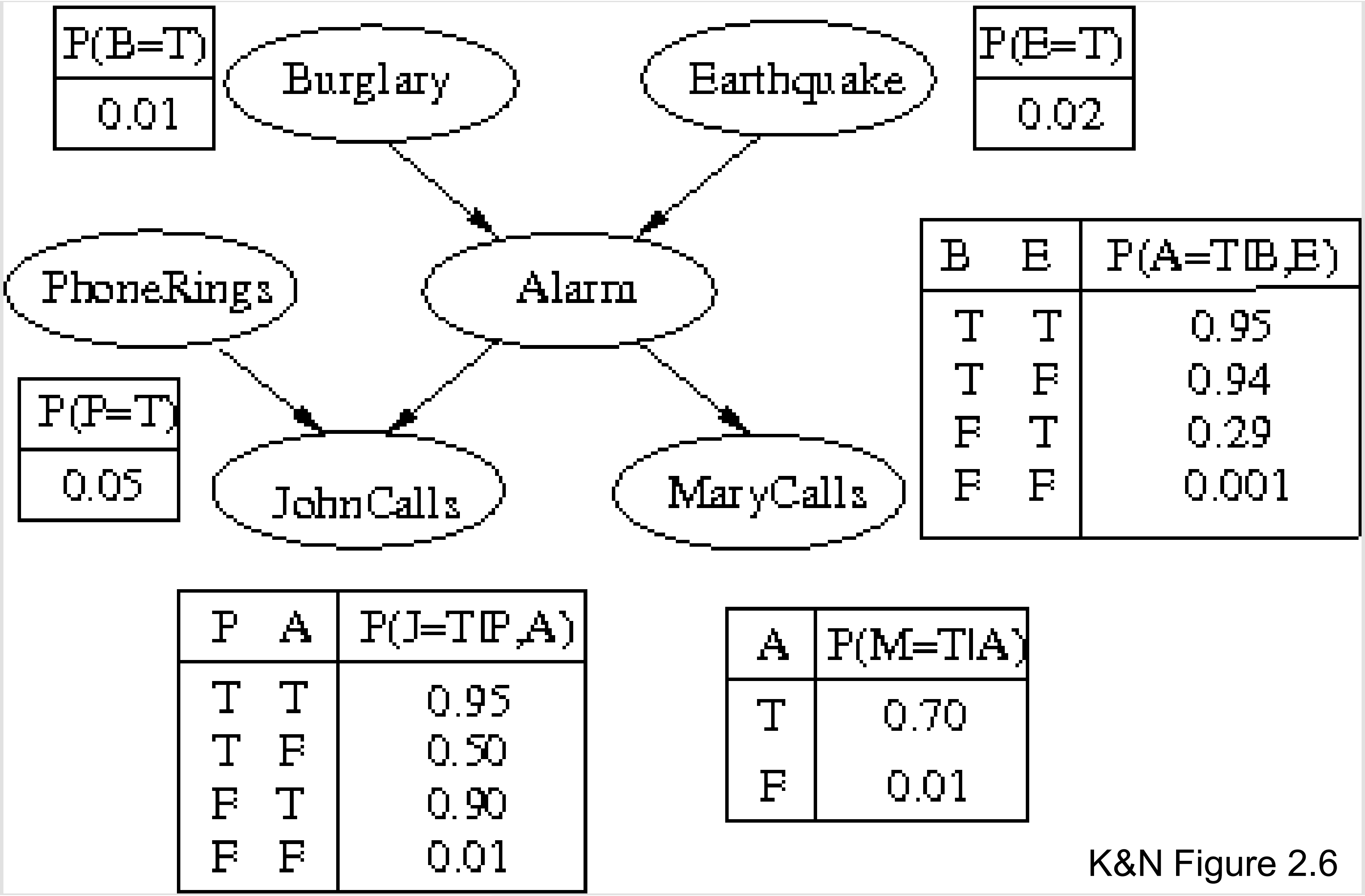
You have a new burglar alarm installed. It reliably detects burglary, but also responds to minor earthquakes. Two neighbours, John and Mary, promise to call the police when they hear the alarm. John always calls when he hears the alarm, but sometimes confuses the alarm with the phone ringing and calls 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

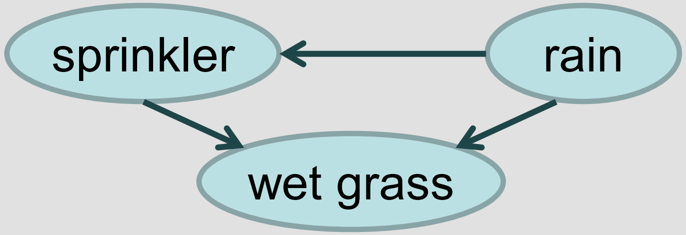
P	A	P(J=T|P,A)
T	T	0.95
T	F	0.50
F	T	0.90
F	F	0.01

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

K&N Figure 2.6

Causality?

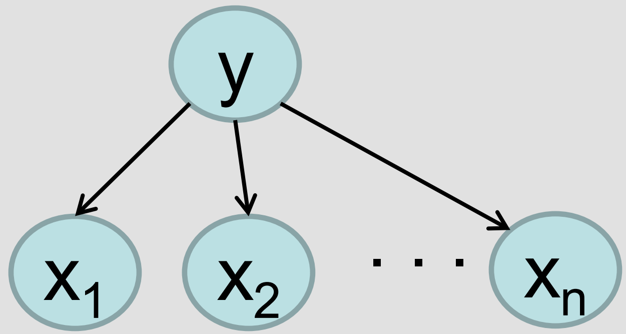
- **When Bayesian networks reflect causal patterns:**
  - Often simpler (nodes have fewer parents)
  - Often easier to think about
  - Often easier to elicit from experts
- **BNs need not actually be causal, but it is good practice**



  - 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**

Low pressure Rain Traffic

$$X \rightarrow Y \rightarrow Z$$

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

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

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

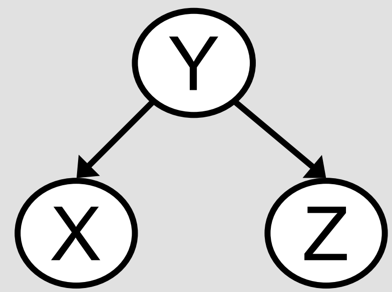
$$= \mathbf{Pr}(z|y)$$

- **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



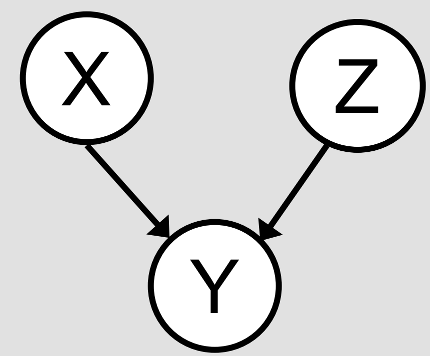
$$\Pr(z|x,y) = \frac{\Pr(x,y,z)}{\Pr(x,y)} = \frac{\cancel{\Pr(y)}\,\cancel{\Pr(x|y)}\,\Pr(z|y)}{\cancel{\Pr(y)}\,\cancel{\Pr(x|y)}} = \Pr(z|y)$$

- **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

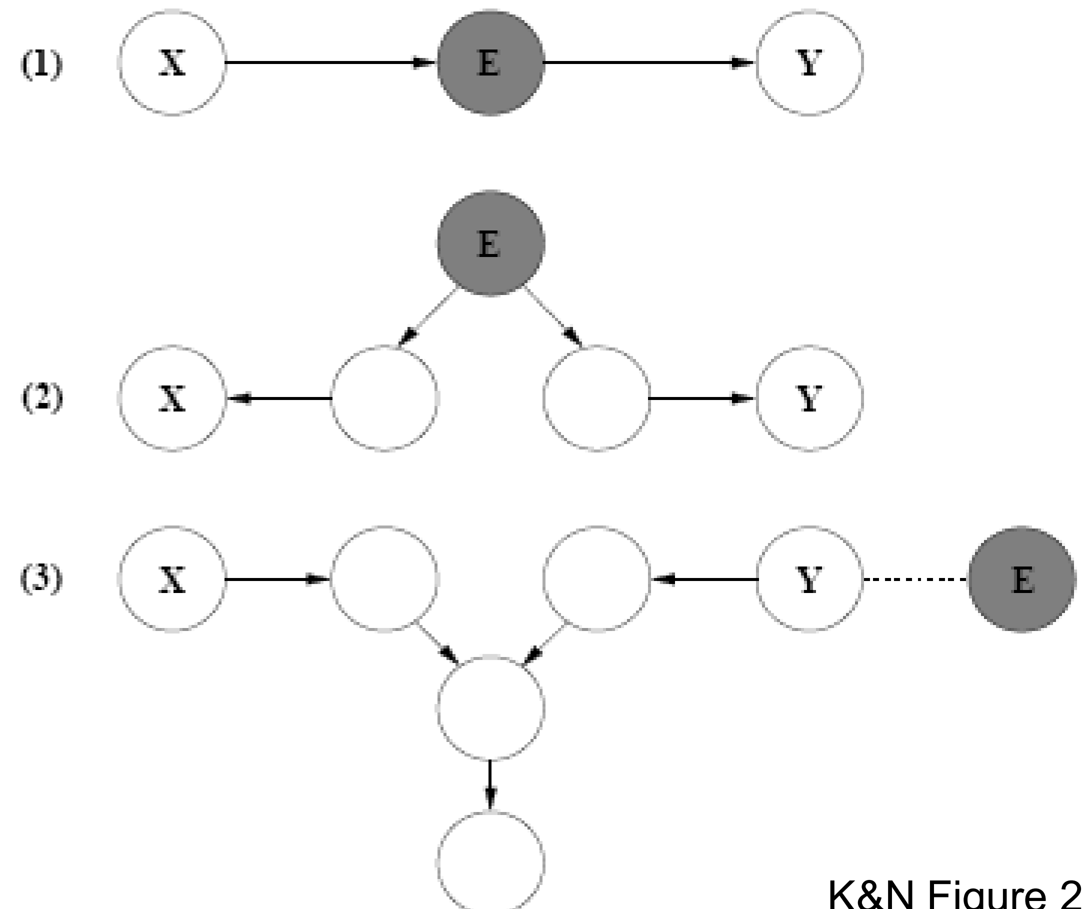
- ***Blocked path*: A path is *blocked* given a set of nodes E, if there is a node Z on the path for which at least one of three conditions holds:**
  1. Z is in **E** and Z has one arrow on the path leading in and one arrow out (**chain**)
  2. Z is in **E** and Z has both path arrows leading out (**common 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



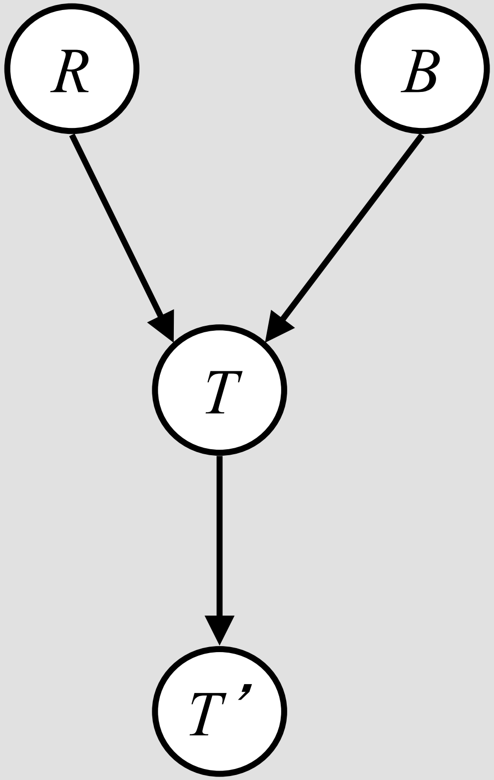
K&N Figure 2.5

D-separation – Example (I)

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

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

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



D-separation – Example (II)

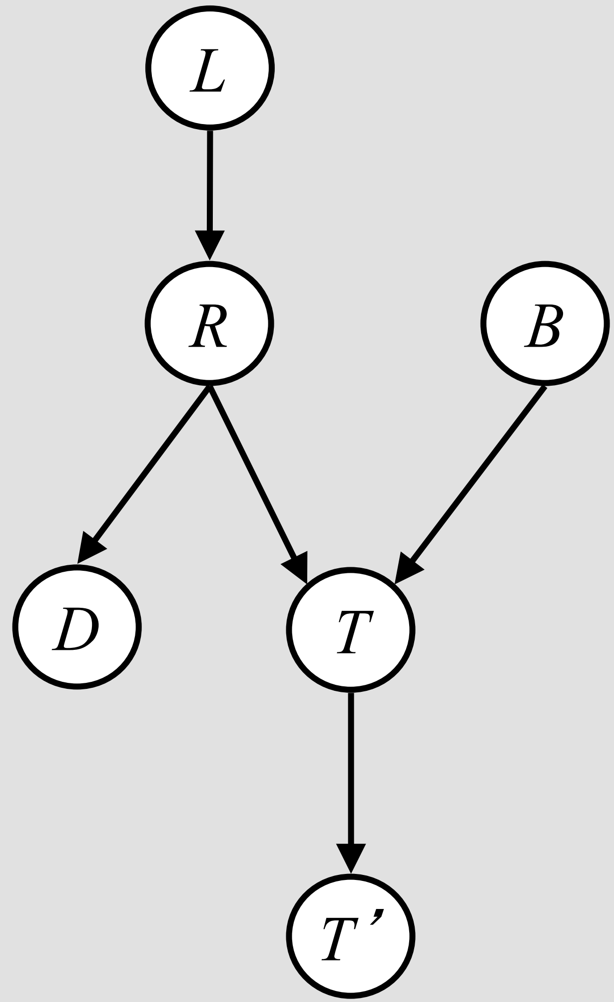
$L \perp\!\!\!\perp T' | T$

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

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

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

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



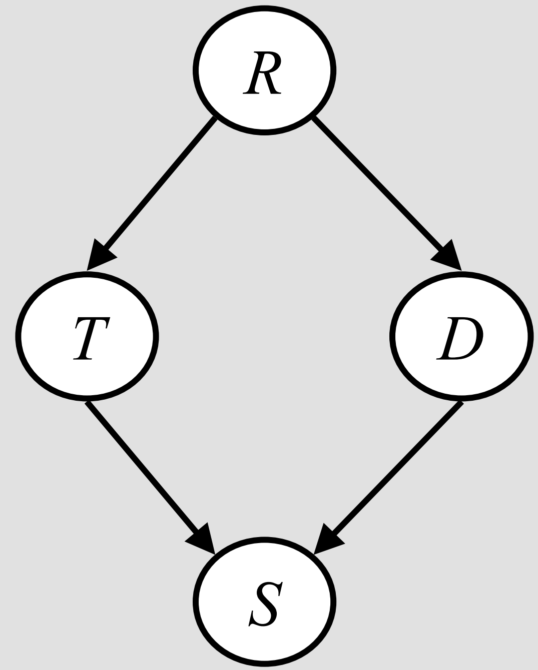
D-separation – Example (III)

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

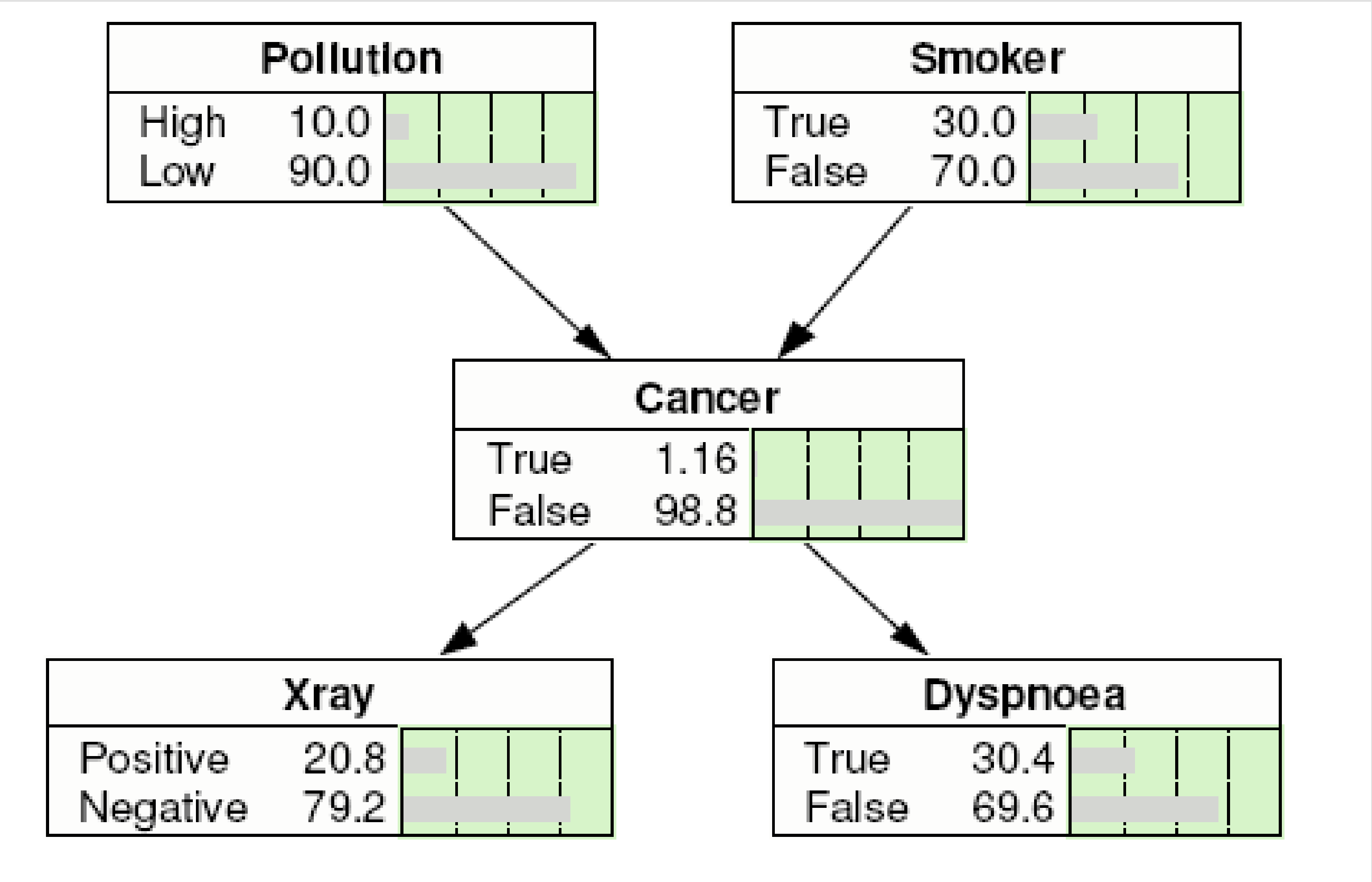
$$T \perp\!\!\!\perp D$$

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

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



Reasoning with Numbers – Using Netica Software



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 reasoning**
  - computes the probability of query variables given evidence variables
  - is flexible – we can enter evidence about any node and update beliefs in other nodes

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
  - Korb, K. and Nicholson, A. (2010), *Bayesian Artificial Intelligence* (2nd ed), Chapman and Hall
    > Chapters 1 & 2
- **Software**
  - Netica – http://www.norsys.com/

Next Lecture Topic

- **Lecture 9 – Bayesian Networks II**
  – BN inference algorithms
  – Bayesian decisions networks
  – Knowledge engineering Bayesian networks
  – Some case studies