Bayesian Network

Md Samiullah

Overview

- → Probability and Conditional probability
- → Bayes Theorem
- → Bayesian Network and variations
- → Issues with ordinary BNs
- → Object Oriented Bayesian Network
- → New challenges
  - ◆ Compilation
  - ◆ Change propagation
  - ◆ Hierarchy and the Yoyo-problem
- → Remarks and directions

Probability and Probabilistic Analysis

- Uncertainty – the quality or state of being not precisely known
  - distinguishes deductive knowledge from inductive belief
- Sources of uncertainty
  - Ignorance
  - Complexity
  - Physical randomness
  - Vagueness
- Uncertainty is the biggest challenge
  - How will be the weather tomorrow? Can we play the match?
  - Which team will win?
  - Should I bet? Should I invest money in this project?
- Probabilistic analysis is a weapon to fight
  - Take a decision based on prior situations
  - Find dependency and conditions
  - Get posterior and take an appropriate action

Random Variables

- A random variable represents some aspect of the world about which we may have uncertainty
  - R = Is it raining?
  - D = How long will it take to drive to work?
  - L = Where am I?
- We denote random variables with capital letters
- Random variables have domains
  - R in {true, false}   (sometimes write as {+r, -r})
  - D in [0, inf)
  - L in possible locations, maybe {(0,0), (0,1), ...}

Probabilistic Inference

- Probabilistic inference: compute a desired probability from other known probabilities
- We generally compute conditional probabilities
  - They represent an agent's beliefs given the evidence
  - E.g., Pr(on time | no reported accidents) = 0.90
- Probabilities change with new evidence:
  - Observing new evidence causes beliefs to be updated
  - E.g., Pr(on time | no accidents, 5 a.m.) = 0.95
  - Pr(on time | no accidents, 5 a.m., raining) = 0.80

Uncertainty and Probabilistic Inference

- General situation:
  - **Evidence**: Agent knows certain things about the state of the world
  - **Hidden variables**: Agent needs to reason about other aspects
  - **Model**: Agent knows something about how the known variables relate to the unknown variables

- Probabilistic reasoning gives us a framework for managing our beliefs and knowledge

Probabilistic Models

- Probabilistic models describe how (a portion of) the world works
- Models are always simplifications
  - May not account for every variable
  - May not account for all interactions between variables
  - "All models are wrong; but some are useful."

    – George E. P. Box
- What do we do with probabilistic models?
  - We (or our agents) need to reason about unknown variables given evidence
    - explanation (diagnostic reasoning)
    - prediction (causal reasoning)
    - value of information

Probability Distribution

- Unobserved random variables have distributions that represent probabilities of value assignments

Pr(Temp)

Temp	Pr
warm	0.5
cold	0.5

Pr(Weather)

Weather	Pr
sunny	0.6
rain	0.1
fog	0.3

- A probability is a single number

$$\Pr(\text{Weather}=\text{rain}) = 0.1 \quad \text{or} \quad \Pr(\text{rain}) = 0.1$$

Probability Calculus

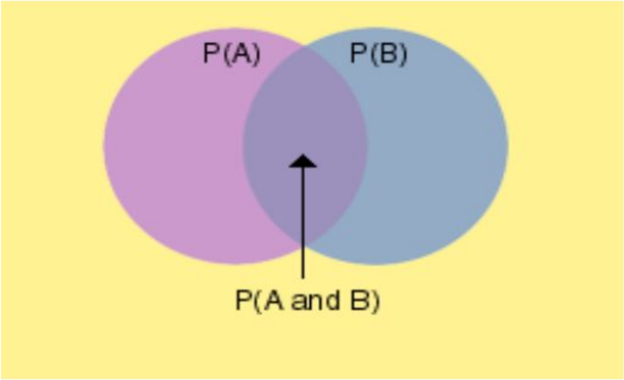
- Classic approach to reasoning under uncertainty (origin: Pascal and Fermat)
- Definitions:
  - **Experiment** – produces one of several possible outcomes
  - **Sample space** – the set of all possible outcomes
  - **Event** – a subset of the sample space: a set of outcomes
  - **Random variable** – a variable whose value is determined by the outcome of an experiment
  - **Probability function** – a function that assigns a probability to every possible outcome of an experiment
    - Given a probability function we can define a probability for each value of a random variable

Probability

- Classical (Laplace,1820)
    - The ratio of the number of favorable cases to the total number of “equally possible” cases
- Frequentism (von Mises,1928)
    - the limiting frequency of favorable cases to total cases in a random sequence
- Bayesianism (Ramsey,1931)
    - the strength of a rational agent’s belief in some proposition
    - rationality: minimally adherence to the probability axioms (subjectivism); possibly adherence to additional principles (e.g., objective Bayesianism, max entropy Bayesianism, etc.)

Probability

- Kolmogorov's Axioms
  - $P(U) = 1$
  - $\forall\ A \subseteq U, P(A) \geq 0$
  - $\forall\ A, B$, if $A \cap B = \varnothing$, then $P(A \vee B) = P(A) + P(B)$



Some probability terms

- Techniques to quantify probabilities of multiple variables
  - Marginal probability:
    - Probability of an event irrespective of the outcome of another variable
    - Probability of event X=*A* given variable *Y*
    - Probability of a specific value of one input variable is the marginal probability across the values of the other input variables
    - $P(X=A) = \sum_{\text{all } y} P(X=A, Y=yi)$
  - Joint probability:
    - The probability of two events occurring simultaneously
    - Probability of events *A* and *B*
    - The probability of a row of data is the joint probability across each input variable
    - P(A and B) = P(A given B) * P(B) = P(B given A) * P(A)
  - Conditional probability:
    - Probability of one event occurring in the presence of a second event
    - Probability of event *A* given event *B*
    - The predictive model itself is an estimate of the conditional probability of an output given an input example

Some probability terms …

- ➔ $P(X)$: singular (also marginal when $P(x) = \sum_y P(x, y)$)
- ➔ $P(X|Y)$: conditional
- ➔ $P(X,Y)$: joint

- ➔ X and Y are binary, i.e., states(X) = {+x, -x} and states(Y) = {+y, -y}
- ➔ For $P(X, Y)$:
  - ◆ $P(x) = \sum_y P(x, y)$
  - ◆ $P(y) = \sum_x P(x, y)$
- ➔ $P(X \mid Y) = P(X, Y)/P(Y) = P(X, Y) / \sum_x P(x, y)$
- ➔ $P(X, Y) = P(X|Y)P(Y) = P(X)P(Y|X)$

$P(X \mid Y) = P(X, Y)/P(Y) = P(Y|X)\, P(X) / P(Y)$

Joint Distributions

- A joint distribution over a set of random variables X1, …, Xn specifies a real number for each value assignment (or outcome):

$$\Pr(X_1{=}x_1, \ldots, X_n{=}x_n) \text{ or } \Pr(x_1, \ldots, x_n)$$

- Size of distribution of n variables with domain sizes d? d^n
- Must obey:

$$\forall x_i \quad \Pr(x_1, \ldots, x_n) \geq 0$$

$$\sum_{x_1,\ldots,x_n} \Pr(x_1, \ldots, x_n) = 1$$

- For all but small distributions, impractical to write out

$\Pr(W,T)$

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

Probabilistic Models

- A probabilistic model is a joint distribution over a set of variables

$$\Pr(X_1, X_2, \dots, X_n)$$

- Given a joint distribution, we can reason about unobserved variables given evidence
- General form of a query:

$$\Pr(X_q \mid e_1, \dots, e_k)$$

Stuff you care about

Stuff you already know

- This kind of posterior distribution is also called the belief function of an agent who uses this model

Events in a Joint Distribution

$$\Pr(E) = \sum_{\{x_1,\ldots,x_n\}\in E} \Pr(x_1, \ldots, x_n)$$

- From a joint distribution, we can calculate the probability of any event
  - Probability that it is hot AND sunny
  - Probability that it is hot
  - Probability that it is hot OR sunny

$\Pr(W, T)$

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

- Typically, the events we care about are partial assignments, like Pr(T=hot)

Marginal Distributions

➔ Marginal distributions are sub-tables that eliminate variables

➔ Marginalization (summing out): Combine collapsed rows by adding

$\Pr(W,T)$

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

$$\Pr(t) = \sum_{w \in \{sun, rain\}} \Pr(t, w)$$

$\Pr(T)$

T	Pr
hot	0.5
cold	0.5

$$\Pr(w) = \sum_{t \in \{hot, cold\}} \Pr(t, w)$$

$\Pr(W)$

W	Pr
sun	0.6
rain	0.4

Conditional Distributions

➔ Conditional distributions are probability distributions over some variables given fixed values of others

Joint Distribution

$\Pr(W, T)$

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

Conditional Distributions

$\Pr(W|T)$

$\Pr(W|T = hot)$

W	Pr
sun	0.8
rain	0.2

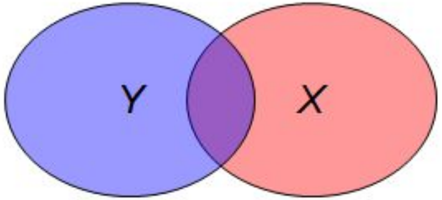
$\Pr(W|T = cold)$

W	Pr
sun	0.4
rain	0.6

Conditional Distributions

$$\Pr(X \mid Y) = \frac{\Pr(X \wedge Y)}{\Pr(Y)}$$

$\Pr(X \cap Y)$



$\Pr(Y)$ $\Pr(X)$

$\mathbf{Pr}(\boldsymbol{W}, \boldsymbol{T})$

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

$$\Pr(W = rain|T = cold) = ?$$

Normalization Trick

- A trick to get a whole conditional distribution at once:
  - Select the joint probabilities matching the evidence
  - Normalize the selection (make it sum to 1)

$\Pr(W,T)$

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

Select

$\Pr(T,rain)$

T	R	Pr
hot	rain	0.1
cold	rain	0.3

Normalize

$\Pr(T\,|\,rain)$

T	Pr
hot	0.25
cold	0.75

- Why does this work?

$$\Pr(x_1|x_2) = \frac{\Pr(x_1,x_2)}{\Pr(x_2)} = \frac{\Pr(x_1,x_2)}{\sum_{x_1}\Pr(x_1,x_2)}$$

Example: Joint to Marginal Probability

Joint: P(X, Y)		
X	**Y**	**P**
+x	+y	0.2
+x	-y	0.3
-x	+y	0.4
-x	-y	0.1

$P(x) = \sum_y P(x, y)$

0.2+0.3

0.4+0.1

0.2+0.4

0.3+0.1

$P(y) = \sum_x P(x, y)$

Marginal: P(X)	
X	**P**
+x	0.5
-x	0.5

Marginal: P(Y)	
Y	**P**
+y	0.6
-y	0.4

Example

Joint: P(X, Y)		
X	**Y**	**P**
+x	+y	0.2
+x	-y	0.3
-x	+y	0.4
-x	-y	0.1

$P(X|Y) = P(X, Y) / \sum_x P(x, y)$

$P(+x|+y) = P(+x, +y) / (P(+x, +y) + P(-x, +y))$
$= 0.2/(0.2+0.4) = 1/3$

$P(+x|-y) = P(+x, -y) / (P(+x, -y) + P(-x, -y))$
$= 0.3/(0.3+0.1) = 3/4$

$P(-x|+y) = P(-x, +y) / (P(+x, +y) + P(-x, +y))$
$= 0.4/(0.2+0.4) = 2/3$

$P(-x|-y) = P(-x, -y) / (P(+x, -y) + P(-x, -y))$
$= 0.1/(0.3+0.1) = 1/4$

Conditional: P(X|Y)		
X	**Y**	**P**
+x	+y	1/3
+x	-y	3/4
-x	+y	2/3
-x	-y	1/4

P(Y|X)		
X	**Y**	**P**
+x	+y	
+x	-y	
-x	+y	
-x	-y	

Example

Conditional: P(X|Y)		
X	**Y**	**P**
+x	+y	1/3
+x	-y	3/4
-x	+y	2/3
-x	-y	1/4

P(Y)	
Y	**P**
+y	0.6
-y	0.4

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

P(+x, +y) = P(+x|+y) P(+y) = ⅓ * 0.6 = 0.2

P(+x, -y) = P(+x|-y) P(-y) = 3/4 * 0.4 = 0.3

P(-x, +y) = P(-x|+y) P(+y) = 2/3 * 0.6 = 0.2

P(-x, -y) = P(-x|-y) P(-y) = 1/4 * 0.4 = 0.2

Joint: P(X, Y)		
X	**Y**	**P**
+x	+y	0.2
+x	-y	0.3
-x	+y	0.4
-x	-y	0.1

Independence and Conditional Indep.

- Two variables are <u>independent</u> if:

$$\Pr(X, Y) = \Pr(X)\Pr(Y)$$

$$\forall x,y \ \Pr(x,y) = \Pr(x)\Pr(y) \ \text{ or } \ \Pr(x|y) = \Pr(x)$$

- Independence : $X \perp\!\!\!\perp Y$ $\qquad X \perp\!\!\!\perp Y$
  - P(X, Y) = P(X|Y)P(Y) = P(X)P(Y) since P(X|Y) = P(X)
  - P(X|Y) = P(X, Y)/P(Y) = P(X)P(Y)/P(Y) = P(X)
- Independence : $X \perp\!\!\!\perp Y \mid Z$
  - P(X, Y | Z) = P(X|Z) P(Y|Z)
  - P(X|Z, Y) = P(X|Z)
- Independence is a simplifying modeling assumption
  - Empirical joint distributions: at best "close" to independent
  - What could we assume for {Weather, Traffic, Cavity, Toothache}?
- More on independence using BNs (later)

Which Variables are Independent?

$\Pr(T)$

T	Pr
warm	0.5
cold	0.5

$\Pr_1(T, W)$

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

$\Pr_2(T, W)$

T	W	Pr
warm	sun	0.3
warm	rain	0.2
cold	sun	0.3
cold	rain	0.2

$\Pr(W)$

W	Pr
sun	0.6
rain	0.4

Conditional Independence

- Unconditional (absolute) independence is very rare
- Conditional independence
  - Our most basic and robust form of knowledge about uncertain environments
  - Notation

$$\left.\begin{array}{ll} \forall x,y,z & \Pr(x,y|z) = \Pr(x|z)\Pr(y|z) \ \text{ or} \\ \forall x,y,z & \Pr(x|y,z) = \Pr(x|z) \\ & \Pr(X,Y|Z) = \Pr(X|Z)\Pr(Y|Z) \\ & \Pr(X|Y,Z) = \Pr(X|Z) \end{array}\right\} X \perp\!\!\!\perp Y \mid Z$$

  - Employs domain knowledge to simplify probabilistic models
  - Example:

    Pr(Traffic|Umbrella,Rain)=Pr(Traffic|Rain) or

    Pr(Traffic,Umbrella|Rain)= Pr(Umbrella|Rain) x Pr(Traffic|Rain)

- Bayesian networks (graphical models) help us express conditional independence assumptions

The Product Rule

- ➔ Sometimes we have conditional distributions but want the joint distribution

$$\Pr(x|y) = \frac{\Pr(x,y)}{\Pr(y)} \quad \Longleftrightarrow \quad \Pr(x,y) = \Pr(x|y)\Pr(y)$$

- ➔ Example:

$\mathbf{Pr}(\boldsymbol{W})$

W	Pr
sun	0.8
rain	0.2

$\mathbf{Pr}(\boldsymbol{T}|\boldsymbol{W})$

T	W	Pr
cold	sun	0.1
hot	sun	0.9
cold	rain	0.7
hot	rain	0.3

$\Longleftrightarrow$

$\mathbf{Pr}(\boldsymbol{T},\boldsymbol{W})$

T	W	Pr
cold	sun	0.08
hot	sun	0.72
cold	rain	0.14
hot	rain	0.06

The Chain Rule

- We can always write a joint distribution as an incremental product of conditional distributions

$$\Pr(x_1, \dots, x_n) = \prod_{i=1}^{n} \Pr(x_i | x_1, \dots, x_{i-1})$$

- Example:

Pr(Traffic,Umbrella,Rain)=

Pr(Umbrella|Rain,Traffic) x Pr(Traffic|Rain) x Pr(Rain)

- Why is this true?

$$\Pr(x_1,\dots,x_n) = \Pr(x_n \mid x_1,\dots,x_{(n-1)}) \Pr(x_1,\dots,x_{(n-1)})$$

$$= \Pr(x_n \mid x_1,\dots,x_{(n-1)}) \Pr(x_{(n-1)} \mid x_1,\dots,x_{(n-2)}) \Pr(x_1,\dots,x_{(n-2)})$$

Bayes Rule

- A joint distribution over a set of random variables X1, …, Xn specifies a real number for each value assignment (or outcome):
- Two ways to factor a joint distribution over two variables:

$$\Pr(x, y) = \Pr(x|y)\Pr(y) = \Pr(y|x)\Pr(x)$$

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

- Why is this helpful?
  - Lets us build one conditional from its rever
  - Often one conditional is tricky but the other one is simple
  - Foundation of many systems (e.g., ASR, MT)
- Forward inference tells us likelihoods
- Finding priors is the main problem in applying Bayes' Rule

Bayes' Inverse Inference Rule:

$$P(h|e) = \frac{P(e|h)P(h)}{P(e)}$$

$$\text{posterior} = (\text{likelihood} \times \text{prior})\alpha$$

Bayes Rule: Conditionalization

- Attributed to Rev. Thomas Bayes

$$\Pr(h \mid e) = \frac{\Pr(e \mid h)\Pr(h)}{\Pr(e)}$$

- Also called Conditionalization: $\Pr'(h) = \Pr(h \mid e)$

- Also read as

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Prob of evidence}}$$

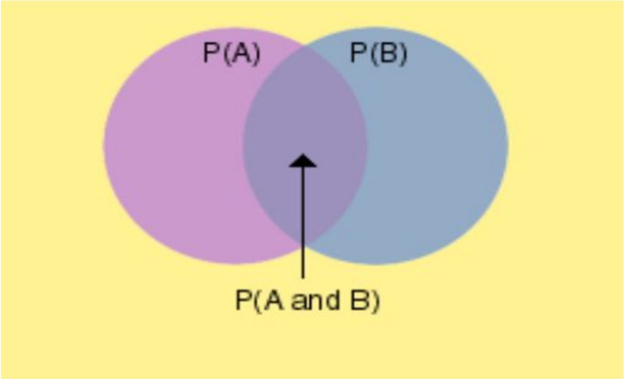
- Assumptions:
  - Joint priors over $\{h_i\}$ and e exist
  - Total evidence: e is observed

Inference with Bayes Rule – Example

- Diagnosis of breast cancer (hypothesis), given xray (evidence)
  - Let Pr(h)=0.01, Pr(e|h)=0.8 and Pr(e|~h)=0.1
  - Bayes theorem yields

$$\begin{aligned}
\Pr(h \mid e) &= \frac{\Pr(e \mid h)\Pr(h)}{\Pr(e)} \\
&= \frac{\Pr(e \mid h)\Pr(h)}{\Pr(e \mid h)\Pr(h) + \Pr(e \mid \sim h)\Pr(\sim h)} \\
&= \frac{0.8 \times 0.01}{0.8 \times 0.01 + 0.1 \times 0.99} \\
&= \frac{0.008}{0.008 + 0.099} = \frac{0.008}{0.107} \approx 0.075
\end{aligned}$$

Conditional Probability



- "Probability of A given that we learn (exactly) B"

  $= P(A \mid B)$

- $P(X|Y)$: 2-D table giving values

  $P(X=xi \mid Y = yj)$

- Conditional probabilities can be defined in terms of unconditional probabilities.

  $P(X \mid Y) = P(X \land Y) / P(Y)$

Bayesian Reasoning

Cancer Problem

- ➔ *You have been referred to a specialty clinic. You are told that one in one hundred appearing at the clinic have cancer X and that the tests are positive given cancer X 80% of the time and they also are positive 10% of the time in people without cancer X.*

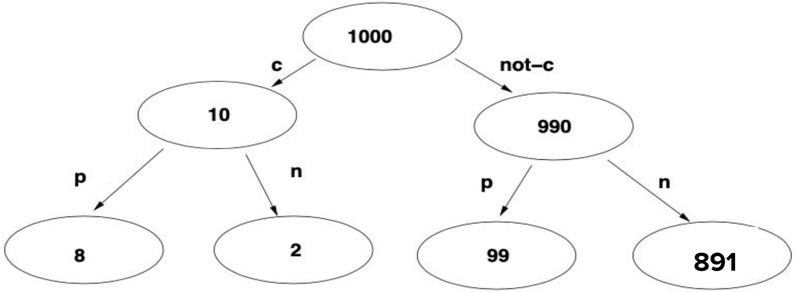
- ➔ What is the probability that you have cancer X?
  - 99%
  - 80%
  - 50-55
  - 30%
  - 7%

Bayesian Reasoning: Cancer

$$
\begin{aligned}
P(c|p) &= \frac{P(p|c)P(c)}{P(p|c)P(c) + P(p|\neg c)P(\neg c)} \\
&= \frac{.8 \times .01}{.8 \times .01 + .1 \times .99} \\
&= \frac{.008}{.008 + .099} \\
&= \frac{.008}{.107} \approx .075
\end{aligned}
$$

Cancer: Frequency Formats

It's easy — multiply!



Classification tree for breast cancer

$$P(c|p) = \frac{P(c,p)}{P(p)} = \frac{8}{8+99}$$

Even easier: use Bayesian Network!

Bayesian Reasoning

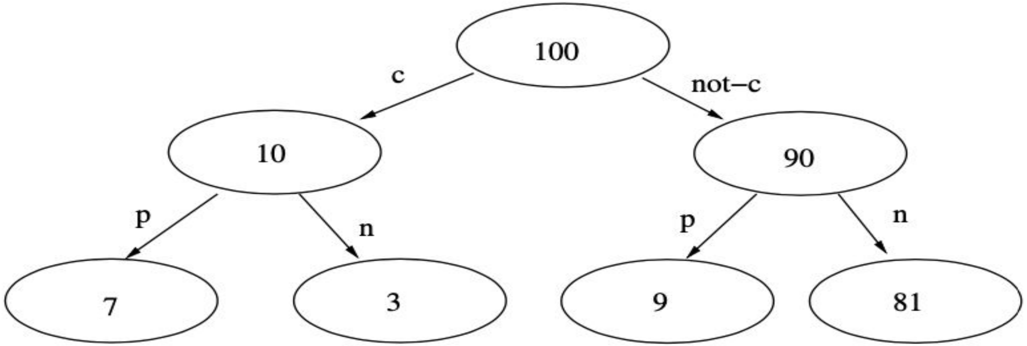
Cancer Problem

- ➔ You may have cancer X. You are in a suspect category where the rate is 10%. A test Y is positive given cancer 70% of the time and it is also positive 10% of the time without cancer.

- ➔ What is the probability that you have cancer X?
  - 90%
  - 70%
  - 50-55
  - 60%
  - 10%

- ➔ OK, what's the probability given a positive test?

Cancer X: Frequency Format



$$P(c|p) = \frac{P(c,p)}{P(p)} = \frac{7}{7+9} \approx 0.44$$

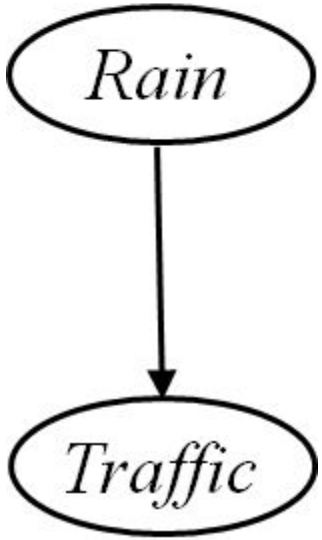
Bayesian Network: The Big Picture

- 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

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 (more later)
- For now, imagine that arrows mean direct causation



Conditional Independence (a 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

BN - Definition

- A data structure that represents the dependence between random variables
- A Bayesian Network is a directed acyclic graph (DAG) in which the following holds:
  - A set of random variables makes up the nodes in the network
  - A set of directed links connects pairs of nodes
  - Each node has a probability distribution that quantifies the effects of its parent nodes

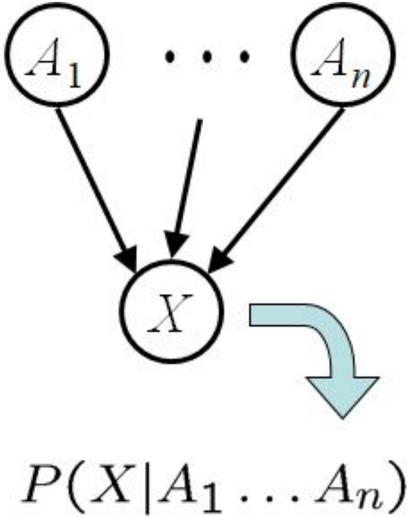
    > Discrete nodes have Conditional Probability Tables (CPTs)
- Gives a concise specification of the joint probability distribution of the variables

BN - Definition ...

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

$$\Pr(X|a_1, \ldots, a_n)$$

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



Bayesian network = Topology (graph) + Local Conditional Probabilities

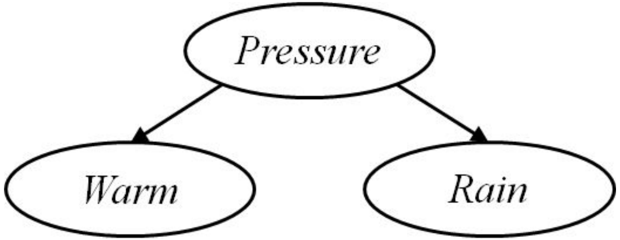
Probabilities in BN

- 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 topology can represent every joint distribution

Building the Joint Distribution - Example



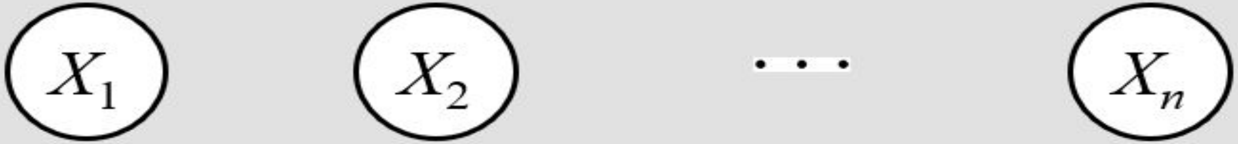
$$\Pr(+high, +warm, \neg rain) = ?$$

$$= \Pr(\neg rain \mid +high, +warm)\ \Pr(+warm \mid +high)\ \Pr(+high)$$

$$= \Pr(\neg rain \mid +high)\ \Pr(+warm \mid +high)\ \Pr(+high)$$

Conditional independence

Example: Coin Flips



$\Pr(X_1)$

h	0.5
t	0.5

$\Pr(X_2)$

h	0.5
t	0.5

$\cdots$

$\Pr(X_n)$

h	0.5
t	0.5

$$\Pr(h, t, t, h) = 0.5 \times 0.5 \times 0.5 \times 0.5$$

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

Designing a BN: Nodes and Values

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

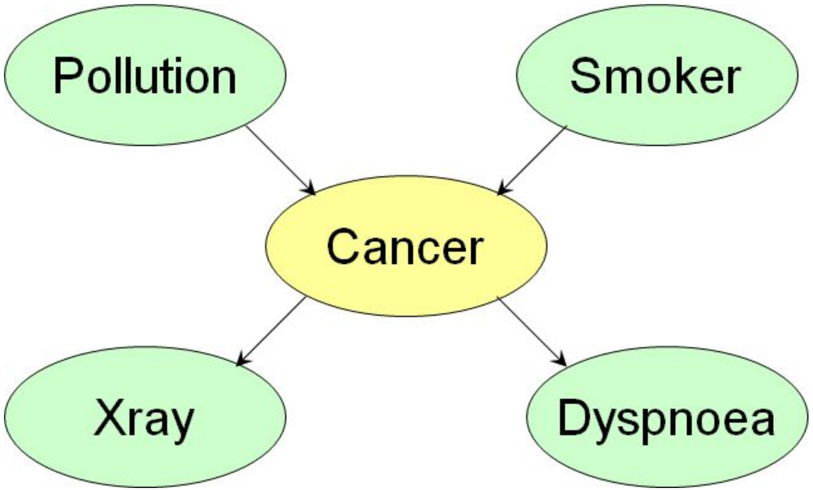
The doctor knows that other relevant information includes whether or not the patient is a smoker (increasing the chances of cancer) and what sort of air pollution he has been exposed to. A positive X-ray would indicate lung cancer.

- 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*}

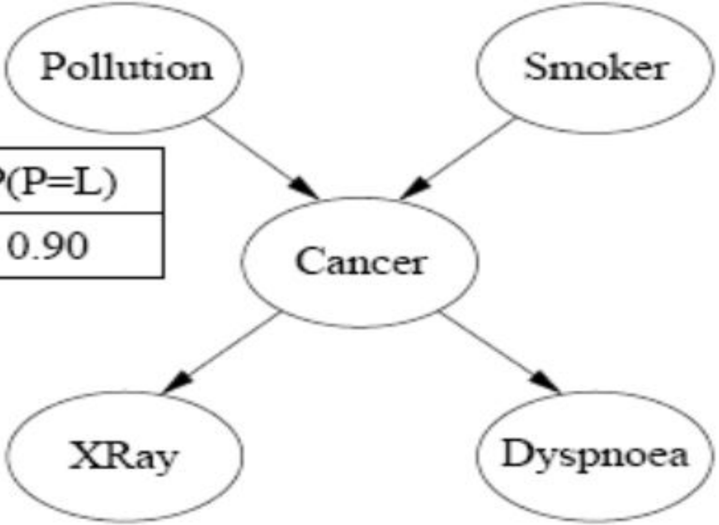


CPTs - Conditional Probability Tables

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(P=L)
0.90

P(S=T)
0.30

The value for =F is the complement

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

Understanding BNs

- 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

BN construction

- Causal structure
  - Expert knowledge and elicitation
  - Automated Learning from data
  - Combination of Elicitation & Automated learning

- Parameter
  - Expert elicitation
  - Learning/estimation from data

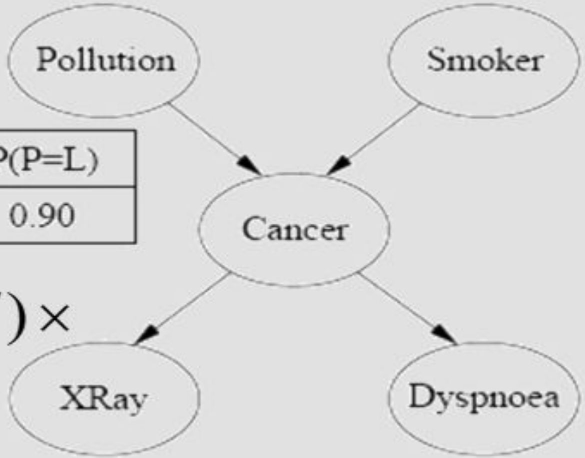
Representing Joint Probability Distribution: Example

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

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

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

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



P(S=T)
0.30

P(P=L)
0.90

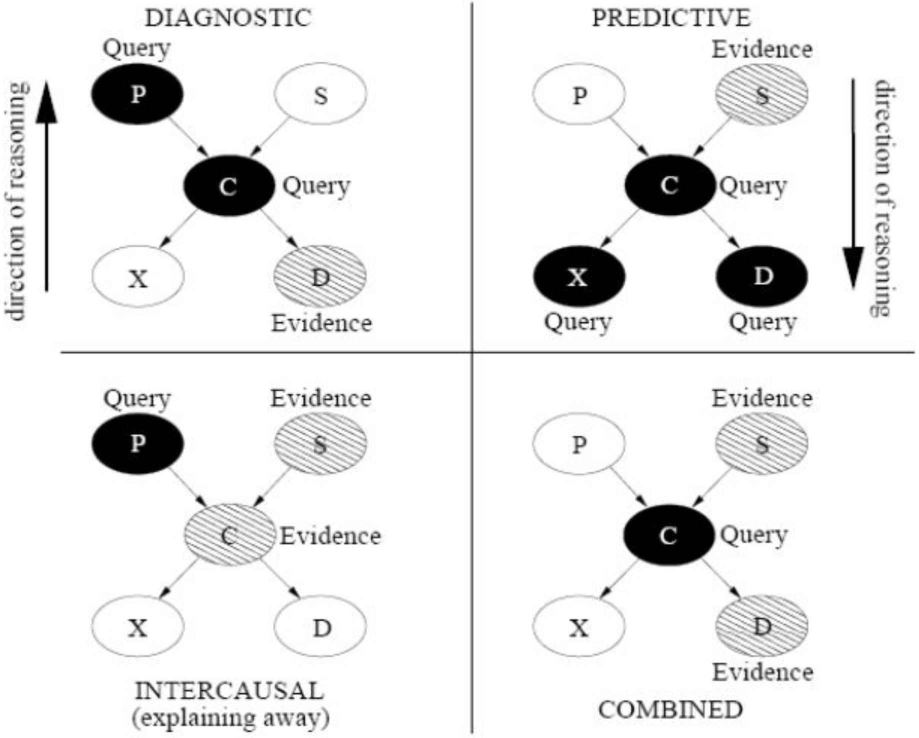
P	S	P(C=T|P,S)
H	T	0.05
H	F	0.02
L	T	0.03
L	F	0.001

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

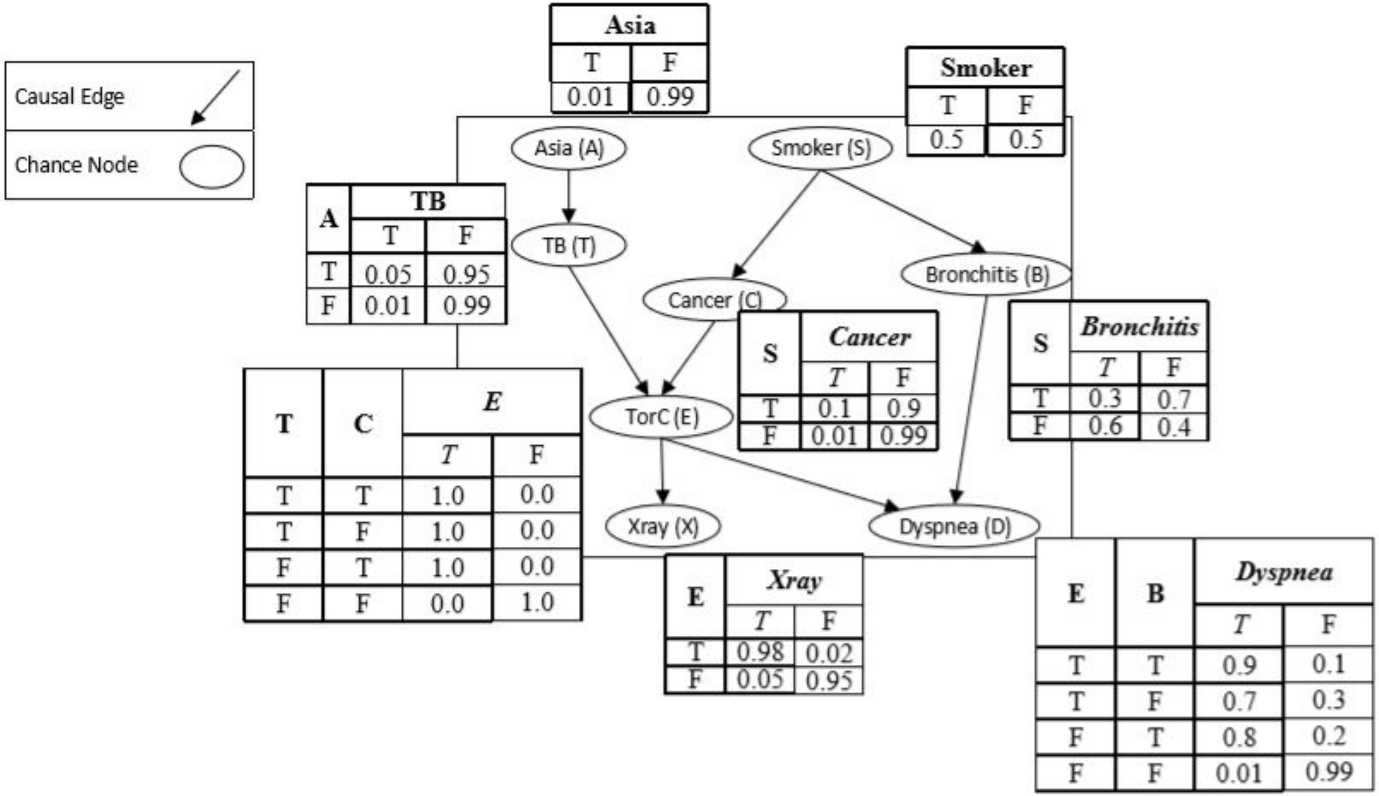
C	P(D=T|C)
T	0.65
F	0.30

Reasoning with BNs and its types

- 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



An example BN: Reasoning



Asia	
T	F
0.01	0.99

Smoker	
T	F
0.5	0.5

A	TB	
	T	F
T	0.05	0.95
F	0.01	0.99

S	*Cancer*	
	T	F
T	0.1	0.9
F	0.01	0.99

S	*Bronchitis*	
	T	F
T	0.3	0.7
F	0.6	0.4

T	C	*E*	
		T	F
T	T	1.0	0.0
T	F	1.0	0.0
F	T	1.0	0.0
F	F	0.0	1.0

E	*Xray*	
	T	F
T	0.98	0.02
F	0.05	0.95

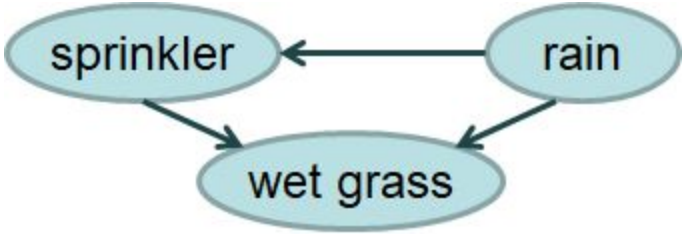
E	B	*Dyspnea*	
		T	F
T	T	0.9	0.1
T	F	0.7	0.3
F	T	0.8	0.2
F	F	0.01	0.99

Reasoning in BNs

This BN and inference in the BN helps in diagnosing a patient, i.e., without any evidence, probability of each disease is P(B=**T**) = 45%, P(T=**T**) = 1.04%, and P(C=**T**) = 5.5%. An evidence on the patient being a smoker (prediction) changes disease probabilities to P(B=**T** | S=**T**) = 30%, P(T=**T** | S=**T**) = 1.04%, and P(C=**T** | S=**T**) = 10% and also the probability of having dyspnea is P(D=**T** | S=**T**) = 30.31% (note that before entering the evidence on "S", the chance of having dyspnea was 39.7%). If patient then presents with shortness of breath, disease probabilities go up, namely P(B=**T** | S=**T**, D=**T**) = 80.26%, P(T=**T** | S=**T**, D=**T**) = 2.61%, and P(C=**T** | S=**T**, D=**T**) = 25.07%, but if Xray result comes back negative, then the probability of T and C go down, B goes up further, namely P(B=**T** | S=**T**, D=**T**, X=**F**) = 96.68%, P(T=**T** | S=**T**, D=**T**, X=**F**) = 0.08%, and P(C=**T** | S=**T**, D=**T**, X=**F**) = 0.72%.

Causality?

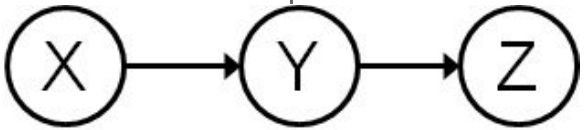
- When Bayesian networks reflect causal patterns:
  - Often simpler (nodes have fewer parents)
  - Often easier to think about
  - Often easier to elicit from experts
- BNs need not actually be causal, but it is good practice



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

Independence in a BN

- Important question about a BN:
  - Are two nodes independent given certain evidence?
  - If yes, can prove using algebra (tedious in general)
  - If no, can prove with a counter example
- Example:



  - Question: are X and Z necessarily independent?
    - Answer: no. Example: low pressure causes rain, which causes traffic.
    - X can influence Z, Z can influence X (via Y)
  - Addendum: they could be independent: how?

Causal Chains

- **This configuration is a "causal chain"**

X: Low pressure

Y: Rain

Z: Traffic

$$P(x,y,z) = P(x)P(y|x)P(z|y)$$

–Is X independent of Z given Y?

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

$$= P(z|y)$$

Yes!

–Evidence along the chain "blocks" the influen

Common Cause

- **Another basic configuration: two effects of the same cause**

–Are X and Z independent?

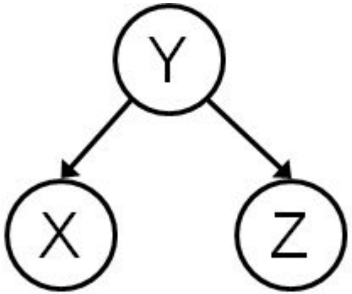
–Are X and Z independent given Y?

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

$$= P(z|y)$$

Yes!

–O en effects.



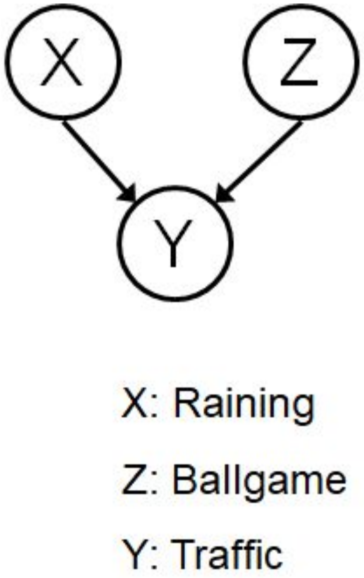
Y: Project due

X: Forum busy

Z: Lab full

Common Effect

- **Last configuration: 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

>Still need to prove they must be (try it!)

–Are X and Z independent given Y?

>No: seeing traffic puts the rain and the ballgame in competition as explanation?

–This is backwards from the other cases

>Observing an effect activates influence between possible causes.

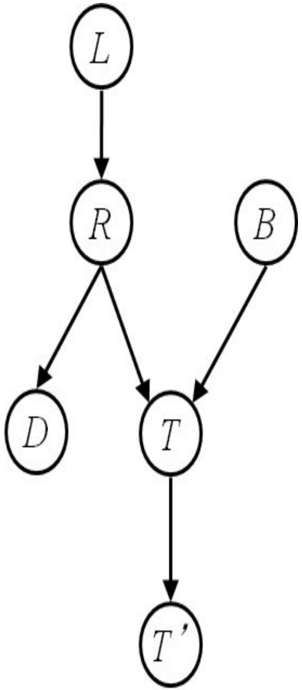
General Case

- 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

$L \perp\!\!\!\perp T' \mid T$	Yes
$L \perp\!\!\!\perp B$	Yes
$L \perp\!\!\!\perp B \mid T$	
$L \perp\!\!\!\perp B \mid T'$	
$L \perp\!\!\!\perp B \mid T, R$	Yes



Independence & D-Separation in BNs …

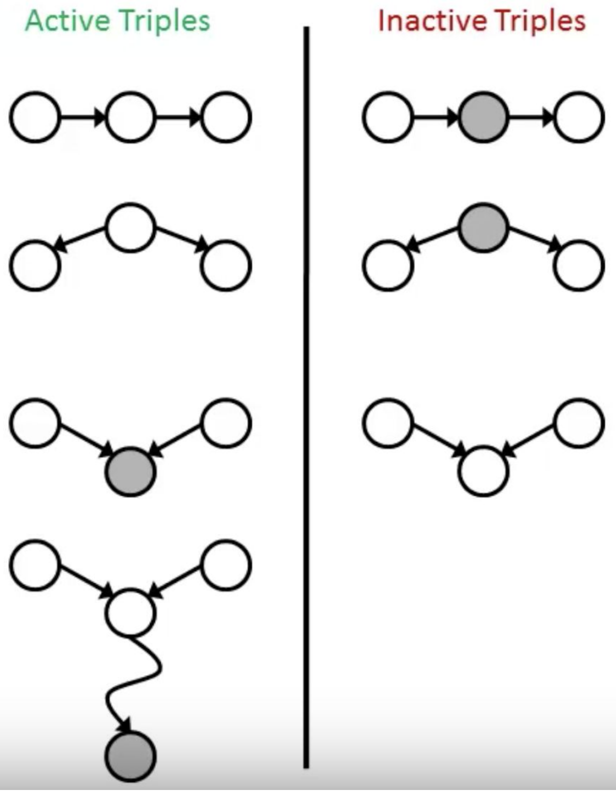
- Backbone of a BN causal structure learning
- P(X|Y, Z) equal to P(X | Y)?
  - Given Y, does learning the value of Z tell us nothing new about X?
  - Is X conditionally independent of Z?
- $X_i \perp\!\!\!\perp X_j \mid (X_{k1}, X_{k2}, \ldots, X_{kn})$?
  - Is X_i independent of X_j i.e. X_i is D-separated from X_j
- Causal Chain: X → Y → Z
- Common cause: Y ← X → Z
- Common effect: X → Z ← Y

Independence & D-Separation in BNs …

- Question: Are X and Y conditionally independent given evidence variables {Z}?
  - Yes, if X and Y "d-separated" by Z
  - Consider all (undirected) paths from X to Y
  - No active paths = independence!

- A path is active if each triple is active:
  - Causal chain $A \rightarrow B \rightarrow C$ where B is unobserved (either direction)
  - Common cause $A \leftarrow B \rightarrow C$ where B is unobserved
  - Common effect (aka v-structure)
    $A \rightarrow B \leftarrow C$ where B *or one of its descendents* is observed

- All it takes to block a path is a single inactive segment



Independence & D-Separation in BNs …

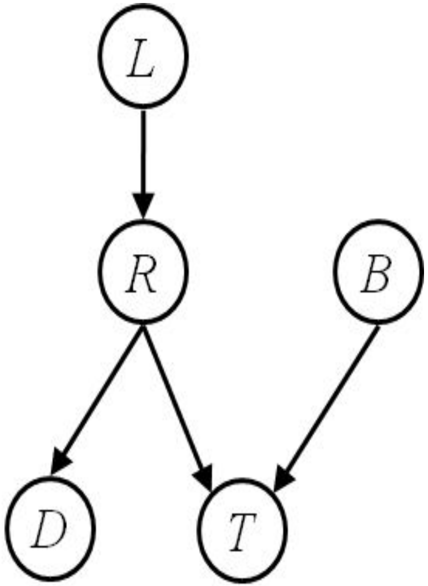
- Find inactive path
  - Inactive corresponds to independence
  - Any inactive segment (Chain/cause/effect) in a path means inactive path, i.e. independence
- In chain and common cause (CCC):
  - Blocked/shaded/evident node in the middle => Inactive
- In “common effect”:
  - NO Blocked/shaded/evident node in the middle => Inactive
- If X and Z have all undirected path inactive: X and Z are independent
  - If any segment is active, independence not guaranteed

Reachability

- Recipe: shade evidence nodes

- Attempt 1: if two nodes are connected by an undirected path not blocked by a shaded node, they are not conditionally independent

- Almost works, but not quite
  - Where does it break?
  - Answer: the v-structure at T doesn't count as a link in a path unless "active"

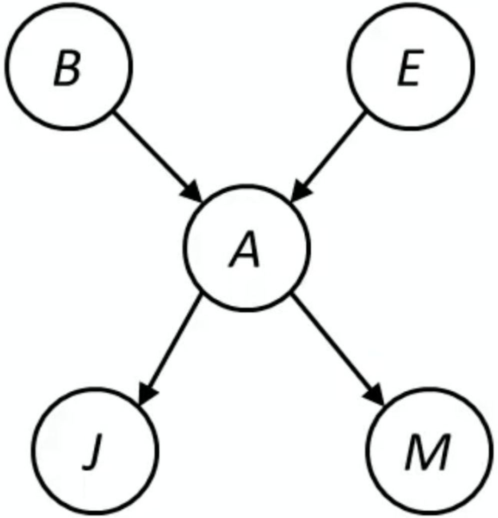


Bayesian Networks (BNs) and variations

- ➔ A technique to describe complex joint distributions using CPTs (simple, local distributions)
- ➔ Graphical Structure (topology) + Probabilities (Local conditional distribution)
    - ◆ DAG + CPTs
- ➔ Consists of:
    - ◆ Nodes (variables with domains)
        - Can be observed or not observed (i.e., assigned or not assigned)
    - ◆ Arcs (interactions)
        - Encode conditional independence by indicating direct influence between variables
- ➔ In order for a Bayesian network to model a probability distribution, the following must be true by definition:
    - ◆ Each variable is conditionally independent of all its non-descendants in the graph given the value of all its parents.
    - ◆ It implies: $P(X_1, X_2,\dots X_n) = \prod_{i=1 \text{ to } n} P(X_i \mid \text{parents}(X_i))$

Alarm BN: Example of joint prob Calc.

B	P(B)
+b	0.001
-b	0.999



E	P(E)
+e	0.002
-e	0.998

A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

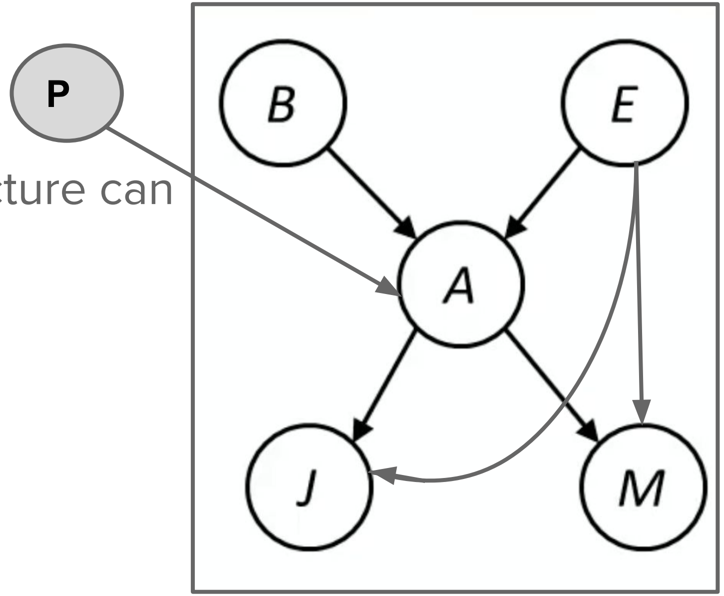
$$P(+b, -e, +a, -j, +m) =$$

$$P(+b)P(-e)P(+a|+b, -e)P(-j|+a)P(+m|+a) =$$

$$0.001 \times 0.998 \times 0.94 \times 0.1 \times 0.7$$

Alarm BN: Extended

- Based on the problem statement, the causal structure can be built
- What about the probability/Parameters?



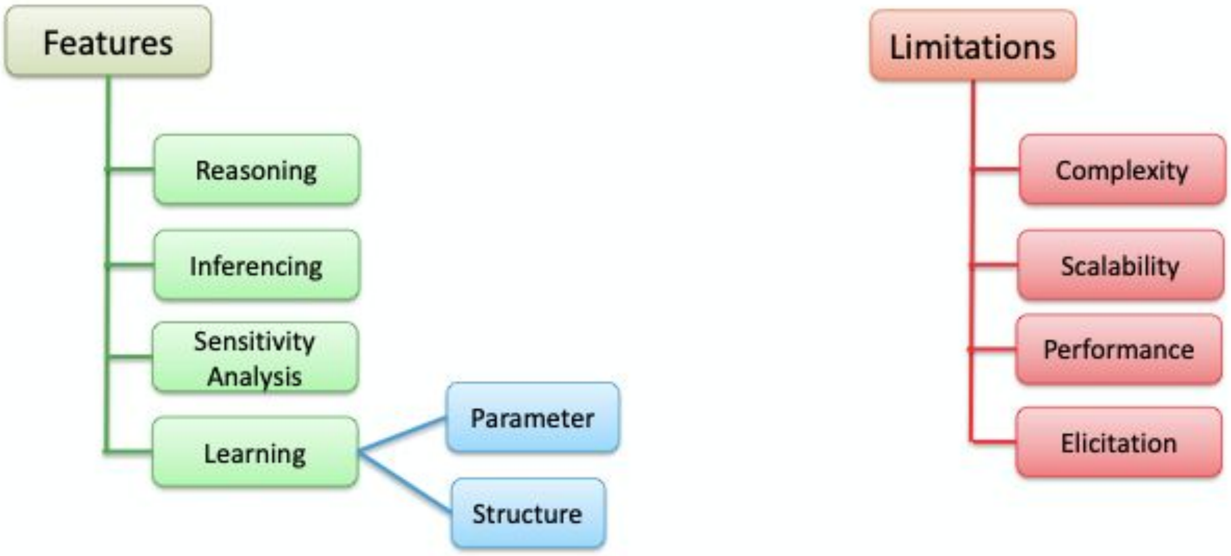
Size of a BN

- 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

Why use BNs?

- Size of a full-joint distribution
  - For N boolean variables: 2^N
- Size of a BN:
  - For N nodes, with nodes having K parents (max)
  - $2.N.2^K = N.2^{K+1}$ (every node has maximum 2^K distributions per state)
- Huge space savings
- Easier to elicit expert knowledge/opinion
- Faster in answering queries
- Cause to effect and effect to cause reasoning: VS If-else-if framework

Why use BNs?



Features
Reasoning
Inferencing
Sensitivity Analysis
Learning
Parameter
Structure
Limitations
Complexity
Scalability
Performance
Elicitation

Causal structure learning

- Automated learning of BN structures from Data
  - PC (Peter and Clark - Prototypical Constraint based approach)
  - IC (Inductive Causation)
  - CaMML (Causal MML)
- Expert's elicitation

Issues with ordinary BNs

- ➔ Scalability
- ➔ Re-usability
- ➔ Inference
- ➔ Automated Learning
- ➔ Knowledge Engineering

BN Summary

- 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

Decision Network

Bayesian Conception of 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

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

Decision Network

- 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)

Bayesian Decision Theory - Example

- Frank Ramsey (1926)
- Decision making under uncertainty – what action to take when the state of the world is unknown

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

Expected utilities:

- E(Take umbrella) $= 60\times0.4 + (-10)\times0.6 = 18$
- E(Leave umbrella) $= -100\times0.4 + 50\times0.6 = -10$

- Bayesian answer –

Find the utility of each possible outcome (action-state pair), and take the action that maximizes the ***expected utility***

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

ID : 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

ID: 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

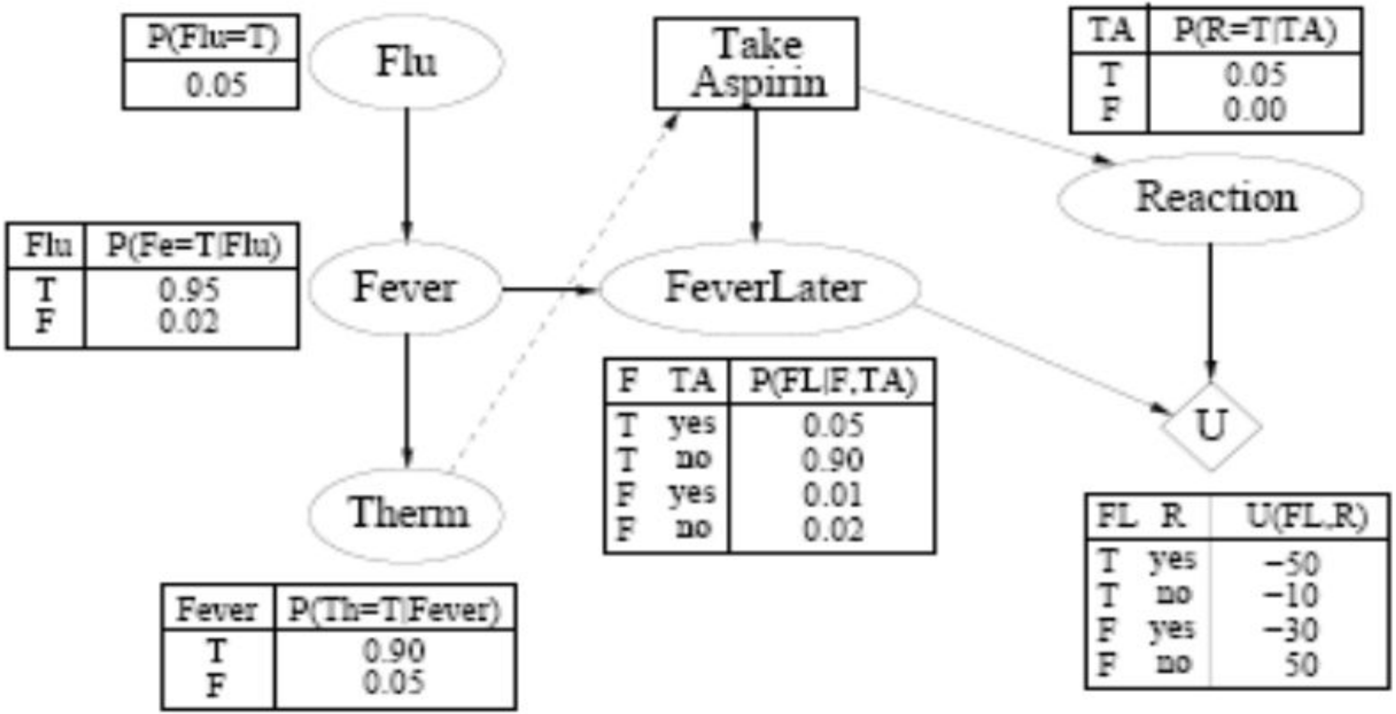
- ***Conditioning links*** **– indicate the variables on which the probability assignment to a chance node will be conditioned**

Example: Fever Problem

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.

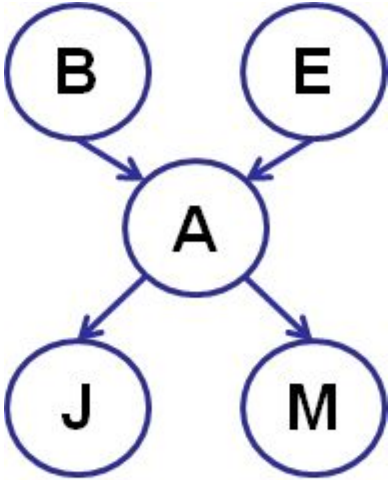
Fever Decision Network



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

Inference



- Calculating some useful quantity from a joint probability distribution
- Posterior probability of a query variable *Q*

  $\Pr(Q \mid E_1=e_1,\ldots,E_k=e_k)$
- Most likely explanation of evidence

$$\operatorname*{argmax}_q \Pr(Q = q \mid E_1 = e_1, \ldots, E_k = e_k)$$

Inference

- Generally aimed at computing marginal probability
- Calculating posterior probability of a set of query variables from the joint probability wrt a set of evidence
  - $P(Q|E_1 = e_1, \ldots, E_k = e_k)$
- Most likely explanation:
  - $\text{argmax}_q\ P(Q|E_1 = e_1, \ldots, E_k = e_k)$

Inference by Enumeration

- ➔ Pr(sun)?
- ➔ Pr(sun | summer)?
- ➔ Pr(sun | winter, hot)?

S	T	W	Pr
summer	hot	sun	0.30
summer	hot	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	hot	sun	0.10
winter	hot	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.20

Inference by Enumeration

- General case:
  - Evidence variables: $E_1, \dots, E_k = e_1, \dots, e_k$
  - Query variable(s): Q
  - Unknown variables: $U_1, \dots, U_r$

$\left.\right\}$ $X_1, \dots, X_n$ *All variables*

- We want: $pr(Q|e_1, \dots, e_k)$
- Procedure
  - Select the entries that are consistent with the evidence
  - Sum out U to get the joint probability of Query and Evidence:

$$\Pr(Q, e_1, \dots, e_k) = \sum_{u_1,\dots,u_r} \Pr\underbrace{(Q, u_1, \dots, u_r, e_1, \dots, e_k)}_{X_1, \dots, X_n}$$

  - Normalize the remaining entries to conditionalize
- Problems:
  - Worst-case time complexity $O(d^n)$
  - Space complexity $O(d^n)$ to store the joint distribution

Inference by Enumeration - Example

- ➔ Pr(sun | summer)
  - ◆ Evidence variables?
  - ◆ Query variables?
  - ◆ Unknown variables?
- ➔ Procedure
  - ◆ Select entries consistent with the evidence
  - ◆ Sum out U to get a joint probability of Q and E
  - ◆ Normalize the remaining entries to conditionalize

summer, sun

S	T	W	Pr
summer	hot	sun	0.30
summer	hot	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	hot	sun	0.10
winter	hot	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.20

summer, rain

Inference by enumeration

- General case:
  - Evidence variables: $E_1 \dots E_k = e_1 \dots e_k$
  - Query* variable: Q
  - Hidden variables: $H_1 \dots H_r$

$\left.\begin{matrix} E_1 \dots E_k = e_1 \dots e_k \\ Q \\ H_1 \dots H_r \end{matrix}\right\} X_1, X_2, \dots X_n$

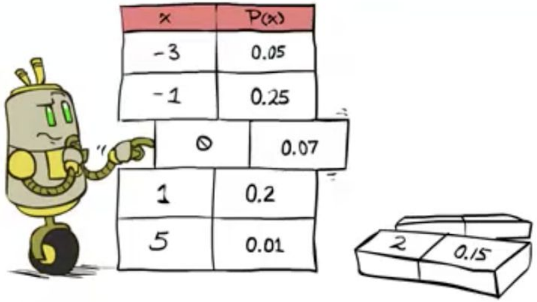
All variables

- We want:

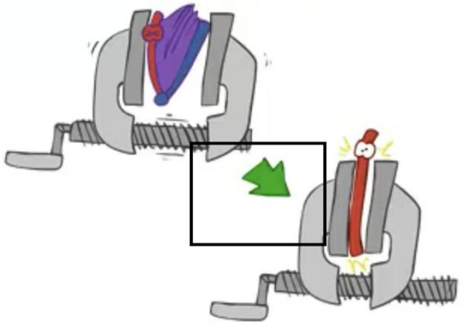
** Works fine with multiple query variables, too*

$$P(Q|e_1 \dots e_k)$$

- Step 1: Select the entries consistent with the evidence



- Step 2: Sum out H to get joint of Query and evidence



- Step 3: Normalize

$$\times \frac{1}{Z}$$

$$Z = \sum_q P(Q, e_1 \cdots e_k)$$

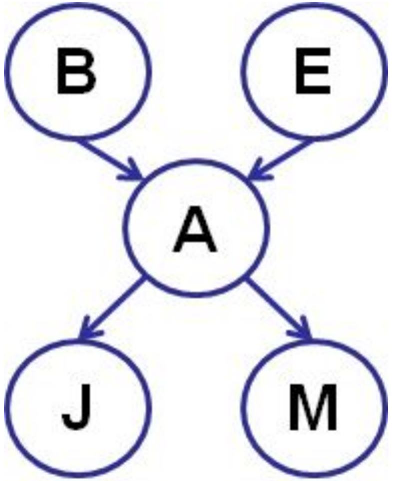
$$P(Q, e_1 \dots e_k) = \sum_{h_1 \dots h_r} \underbrace{P(Q, h_1 \dots h_r, e_1 \dots e_k)}_{X_1, X_2, \dots X_n}$$

$$P(Q|e_1 \cdots e_k) = \frac{1}{Z} P(Q, e_1 \cdots e_k)$$

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:**

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

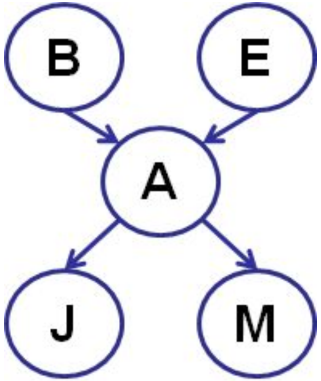


- **There is no need to calculate the denominator**
  - –Compute the numerator for $\Pr(-b|+j,+m)$
  - –Replace the denominator with α in both equations

–$\Pr(+b|+j,+m) + \Pr(-b|+j,+m) = \alpha \Pr(+b,+j,+m) + \alpha \Pr(-b,+j,+m) = 1$

Enumeration Example

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



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

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

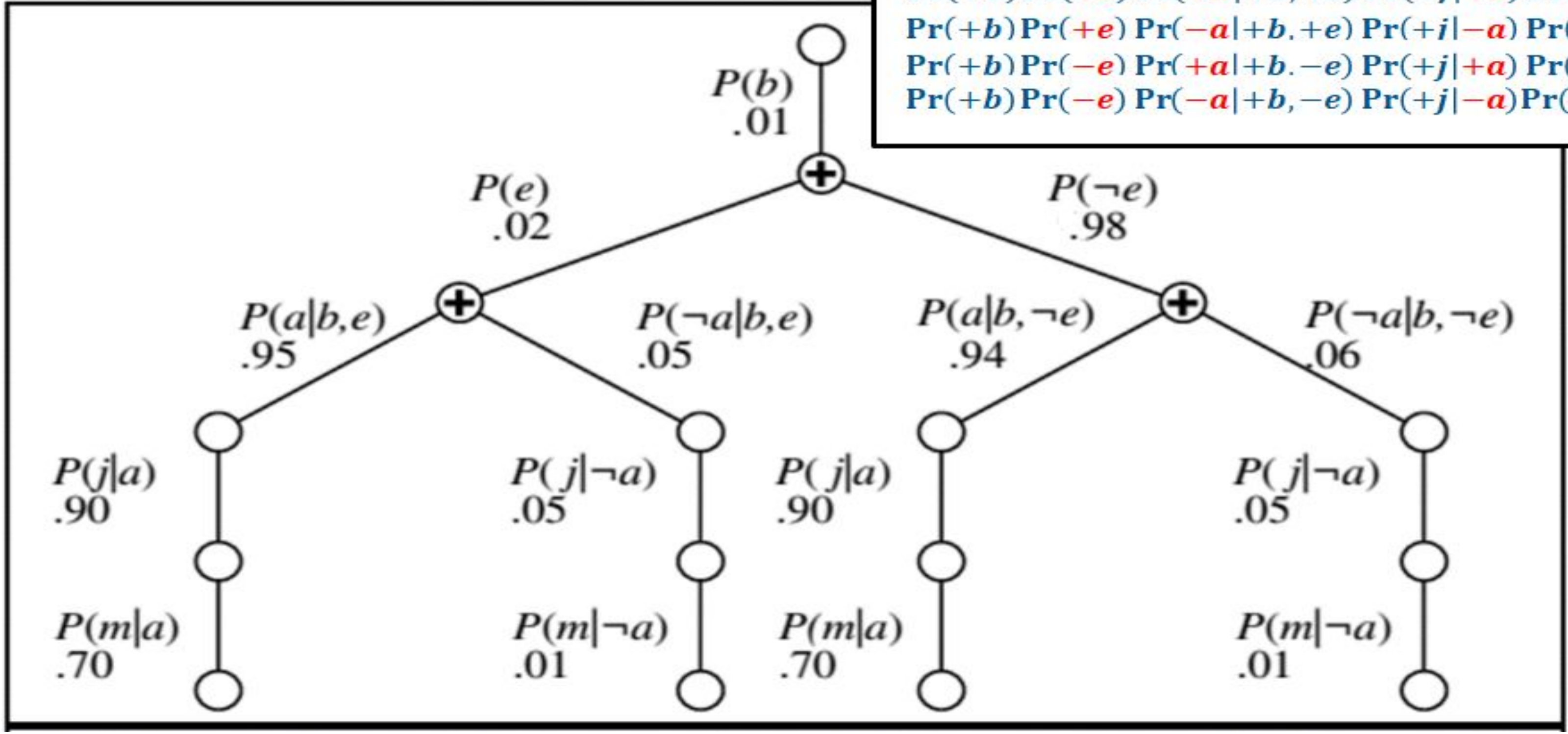
$$\Pr(+b)\Pr(+e)\Pr(-a|+b,+e)\Pr(+j|-a)\Pr(+m|-a) +$$

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

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

Enumeration Example ...

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



Enumeration example for Asia BN

$P(A|X=f, D=t)$

$\propto_B P(A, X=f, D=t)$

$= \sum_{s,\, t,\, c,\, b,\, e} P(A, s, t, c, b, X=f, D=t, e)$

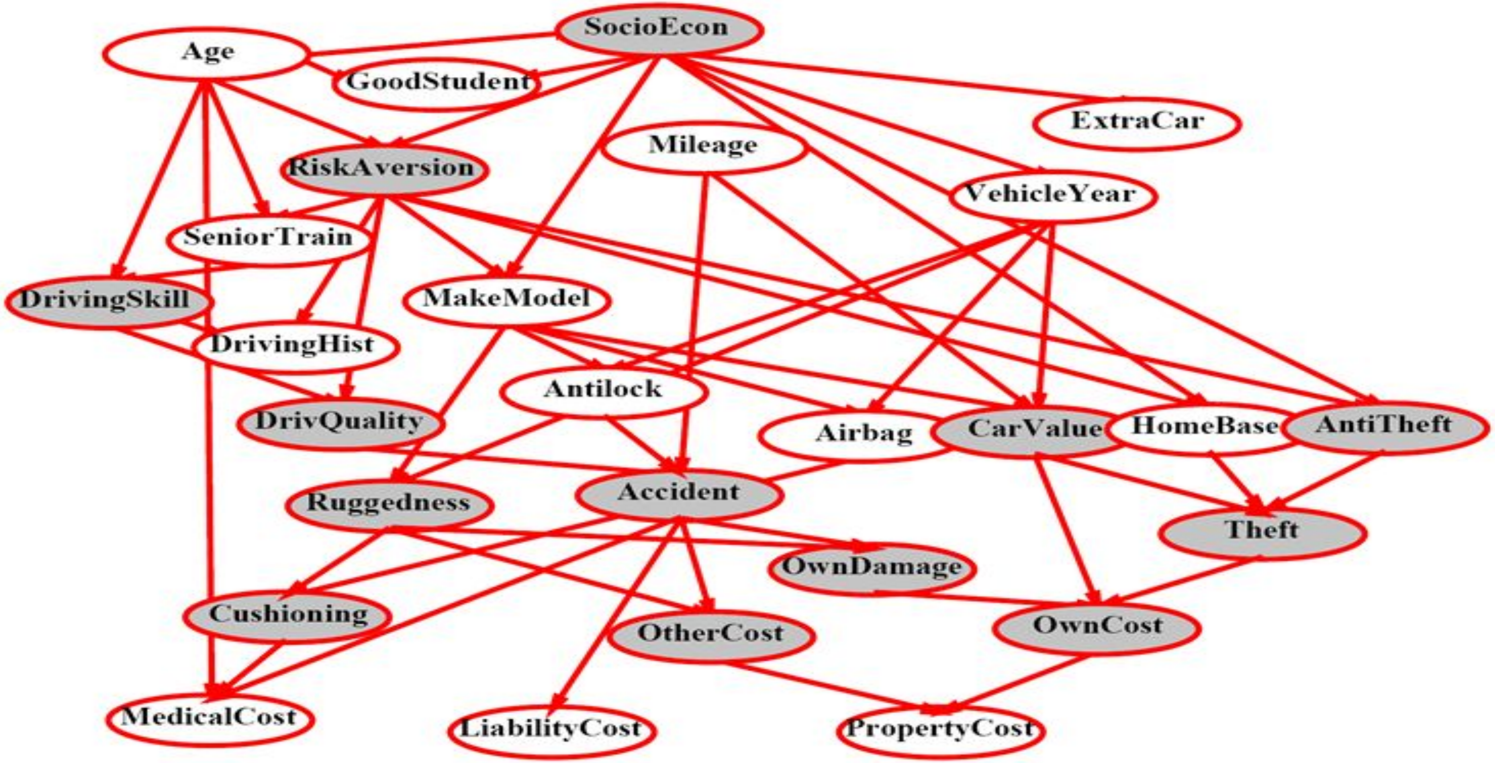
[lowercase to represent instantiated values]

$= \sum_{s,\, t,\, c,\, b,\, e} P(A)\, P(s)\, P(t \mid A)\, P(c \mid s)\, P(b \mid s)\, P(e \mid t, c)\, P(X=f \mid e)\, P(D=t \mid b, e)$

= sum of 32 terms …

We need infinite time for more complicated scenarios!!!

Inference by Enumeration?



Variable Elimination

- **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

Inference by variable elimination

- VE works by eliminating all variables in turn until there is a factor with only query variable
- Factor is a function from some set of variables into a specific value
  - CPTs are factors, e.g., P(A|B,E) function of A, B, E
- Main steps in eliminating a variable:
  - Join all factors containing that variable
  - Sum out the influence of the variable on new factor
  - Exploit product form of joint distributions

VE example

$P(A|X=f, D=t)$

$= \Sigma_{s, t, c, b, e}\ P(A, s, t, c, b, X=f, D=t, e)$

$= \Sigma_{s, t, c, b, e}\ P(A)\ P(s)\ P(t \mid A)\ P(c \mid s)\ P(b \mid s)\ P(e \mid t, c)\ P(X=f \mid e)\ P(D=t \mid b, e)$

Variable orders: A, S, T, C, B, E, X, D => S, T, C, B, E

$= \Sigma_{t, c, b, e}\ P(A)\ \Sigma_s\ P(s)P(c \mid s)P(b \mid s)\ P(t \mid A)\ P(e \mid t, c)\ P(X=f \mid e)\ P(D=t \mid b, e)$

$= \Sigma_{c, b, e}\ P(A)\ f1(c, b)\ \Sigma_t P(t \mid A)P(e \mid t, c)P(D=t \mid b, e)\ P(X=f \mid e)$ *[$\Sigma s\ P(s)P(c \mid s)P(b \mid s) = f1(S, c, b) = f1(c, b)$]*

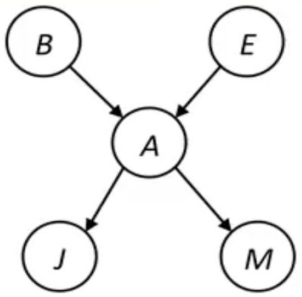
$= P(A)\Sigma_e\Sigma_b\ \Sigma_c f1(c, b)\ f2(e, A, c, b, D=t)\ P(X=f \mid e)$ *[$\Sigma\ tP(t \mid A)P(e \mid t, c)P(D=t \mid b, e) = f2(e, A, c, b, D=t)$]*

$= P(A)\Sigma_e\Sigma_b\ f3(b, e, A, D=t)\ P(X=f \mid e)$ *[$\Sigma\ cf1(c, b)\ f2(e, A, c, b, D=t) = f3(b, e, A, D=t)$]*

$= P(A)\Sigma_e f4(e, A, D=t)\ P(X=f \mid e)$ *[$\Sigma_b\ f3(b, e, A, D=t) = f4(e, A, D=t)$]*

$= P(A)f5(A, D=t, X=f)$ *[$\Sigma_e f4(e, A, D=t)\ P(X=f \mid e) = f5(A, D=t, X=f)$]*

VE another example



$$P(B|j,m) \propto P(B,j,m)$$

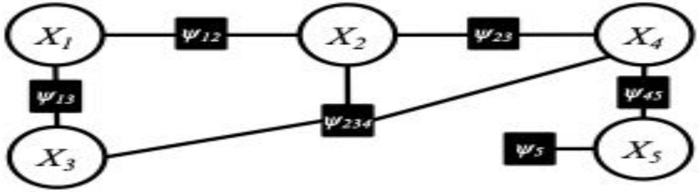
$$\boxed{P(B) \qquad P(E) \qquad P(A|B,E) \qquad P(j|A) \qquad P(m|A)}$$

$$\begin{aligned}
P(B|j,m) &\propto P(B,j,m) \\
&= \sum_{e,a} P(B,j,m,e,a) && \text{marginal can be obtained from joint by summing out} \\
&= \sum_{e,a} P(B)P(e)P(a|B,e)P(j|a)P(m|a) && \text{use Bayes' net joint distribution expression} \\
&= \sum_{e} P(B)P(e)\sum_{a} P(a|B,e)P(j|a)P(m|a) && \text{use x*(y+z) = xy + xz} \\
&= \sum_{e} P(B)P(e)f_1(B,e,j,m) && \text{joining on a, and then summing out gives } f_1 \\
&= P(B)\sum_{e} P(e)f_1(B,e,j,m) && \text{use x*(y+z) = xy + xz} \\
&= P(B)f_2(B,j,m) && \text{joining on e, and then summing out gives } f_2
\end{aligned}$$

Merits of VE

The Variable Elimination Algorithm

Instead, capitalize on the factorization of $p(\mathbf{x})$.



$$
\begin{aligned}
Z &= \sum_{x_1}\sum_{x_2}\psi_{12}(x_1,x_2)\sum_{x_3}\psi_{13}(x_1,x_3)\sum_{x_4}\psi_{24}(x_2,x_4)\psi_{234}(x_2,x_3,x_4)\sum_{x_5}\psi_{45}(x_4,x_5)\psi_5(x_5)\\
&= \sum_{x_1}\sum_{x_2}\psi_{12}(x_1,x_2)\sum_{x_3}\psi_{13}(x_1,x_3)\sum_{x_4}\psi_{24}(x_2,x_4)\psi_{234}(x_2,x_3,x_4)m_5(x_4)\\
&= \sum_{x_1}\sum_{x_2}\psi_{12}(x_1,x_2)\sum_{x_3}\psi_{13}(x_1,x_3)m_4(x_2,x_3)\\
&= \sum_{x_1}\sum_{x_2}\psi_{12}(x_1,x_2)m_3(x_1,x_2)\\
&= \sum_{x_1}m_2(x_1)
\end{aligned}
$$

2^2 additions

2^3 additions

2^3 additions

2^2 additions

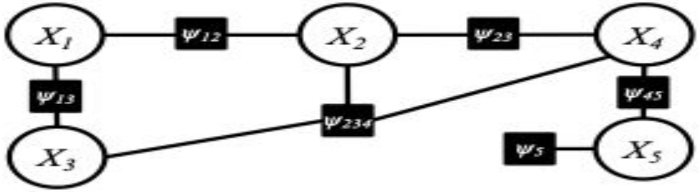
2 additions

Naïve solution requires $2^5 = 32$ additions.
Variable elimination only requires $4+8+8+4+2 = 26$ additions.

Merits of VE ...

The Variable Elimination Algorithm

Instead, capitalize on the factorization of $p(\mathbf{x})$.



$$
\begin{aligned}
Z &= \sum_{x_1}\sum_{x_2}\psi_{12}(x_1,x_2)\sum_{x_3}\psi_{13}(x_1,x_3)\sum_{x_4}\psi_{24}(x_2,x_4)\psi_{234}(x_2,x_3,x_4)\sum_{x_5}\psi_{45}(x_4,x_5)\psi_5(x_5) \\
&= \sum_{x_1}\sum_{x_2}\psi_{12}(x_1,x_2)\sum_{x_3}\psi_{13}(x_1,x_3)\sum_{x_4}\psi_{24}(x_2,x_4)\psi_{234}(x_2,x_3,x_4)m_5(x_4) \\
&= \sum_{x_1}\sum_{x_2}\psi_{12}(x_1,x_2)\sum_{x_3}\psi_{13}(x_1,x_3)m_4(x_2,x_3) \\
&= \sum_{x_1}\sum_{x_2}\psi_{12}(x_1,x_2)m_3(x_1,x_2) \\
&= \sum_{x_1}m_2(x_1)
\end{aligned}
$$

3^2 additions

3^3 additions

3^3 additions

3^2 additions

3 additions

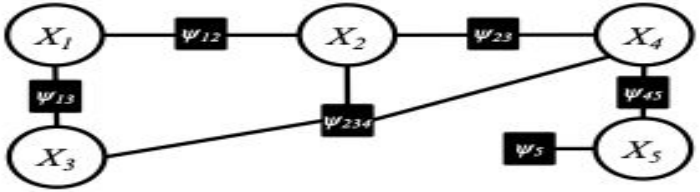
Naïve solution requires $3^5=243$ additions.

Variable elimination only requires $3+3^2+3^3+3^3+3^2 = 75$ additions.

Merits of VE …

Variable Elimination Complexity

Instead, capitalize on the factorization of $p(\mathbf{x})$.



Naïve solution is $O(k^n)$

Variable elimination is $O(nk^r)$

where
n = # of variables
k = max # values a variable can take
r = # variables participating in largest "intermediate" table

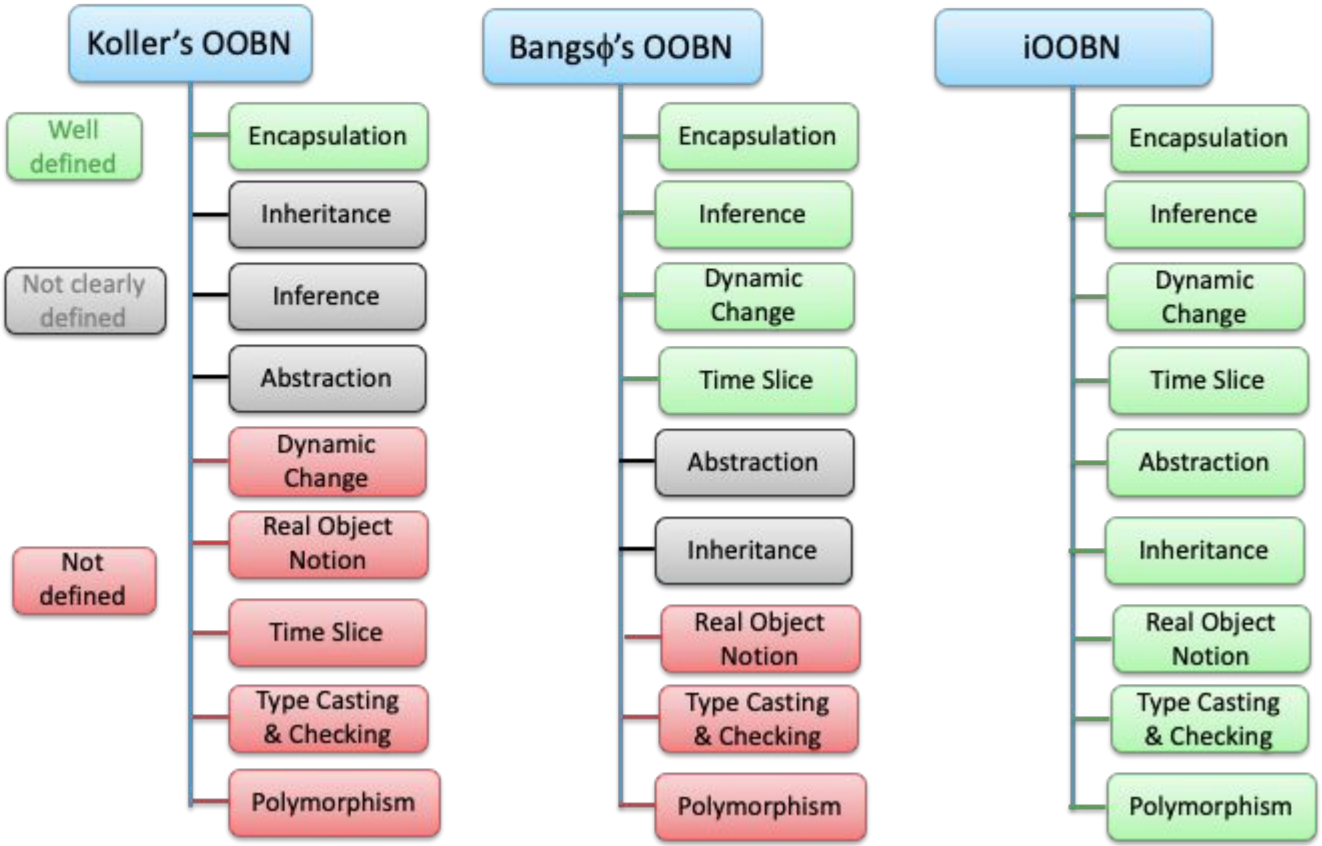
Alternative to BNs

- To eradicate the limitations of BNs
  - PGM
    - OOBNs
    - Templates
    - MSBN
  - PRMSs
    - Plain
    - OOPRMS
    - First Order PMs
    - MEBNs

Object Oriented Bayesian Networks (OOBNs)

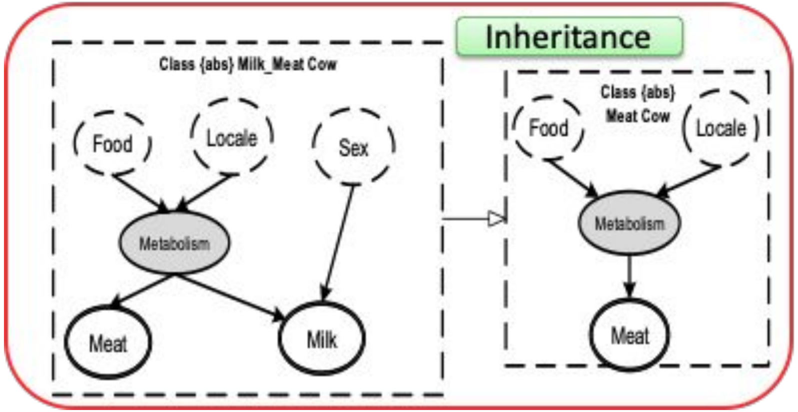
- BN with OO features
  - Encapsulation
  - Abstraction
  - Inheritance
  - Polymorphism
  - Type casting
  - Type checking

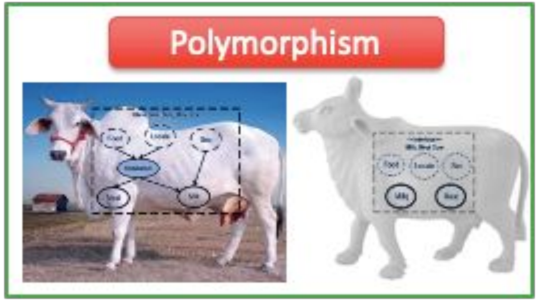
Object Oriented Bayesian Networks (OOBNs)



Object Oriented Bayesian Networks (OOBNs)

- Developing full-fledged OOBN framework: IOOBN
  - Encapsulation & Abstraction
    - Class (Abstract-concrete)
    - Interface
  - Inheritance
  - Polymorphism





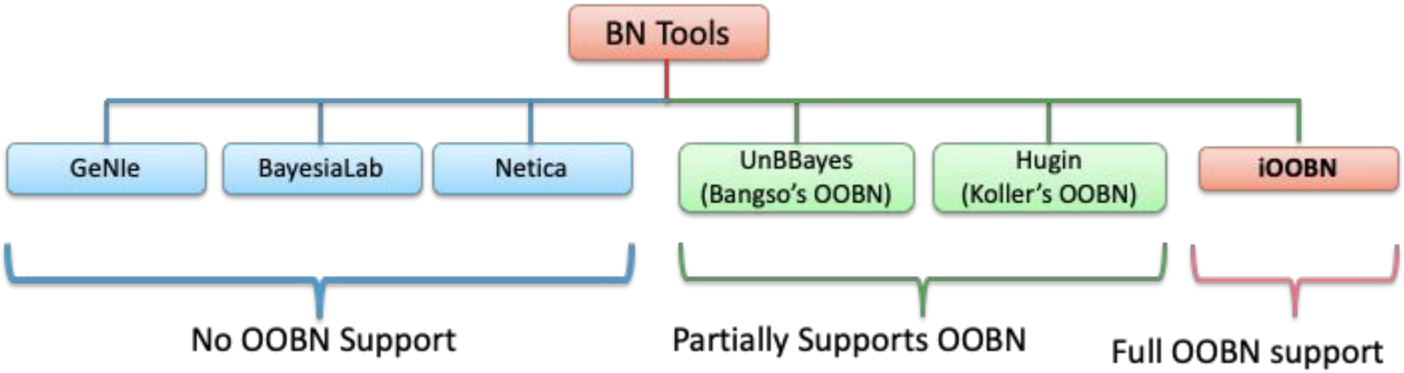
New Challenges

- ➔ Inference
- ➔ Automated learning
- ➔ Knowledge Engineering
- ➔ Library of classes

Tools for BNs

- Commercials
  - Netica
  - GeNIe
  - BayesiaNet
  - Hugin
- Open source
  - UnBBayes
  - BNJ
  - iOOBN

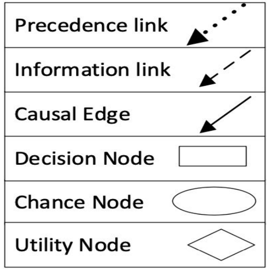
Tools for BNs …



Tools for BNs …

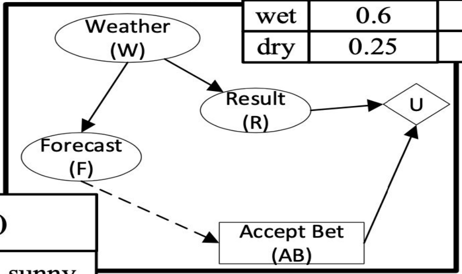
	Netica	GeNIe	HUGIN
DBNs	**	***	*
Submodels	–	***	***
OOBNs	–	–	***
Sensitivity to Findings	**	**	**
Sensitivity to Parameters	–	–	**
Continuous Nodes	*	**	***
Equations	**	*	***
Learning Parameters from Data	***	**	*
Learning Structure from Data	*	**	*
Ease of Use	***	**	*
Cost	**	***	*

An example BN (BDN) in Hugin



Weather (W)	
wet	dry
0.3	0.7

W	Result (R)	
	win	lose
wet	0.6	0.4
dry	0.25	0.75



R	AB	U(R, AB)
win	yes	40
win	no	20
lose	yes	-5
lose	no	-20

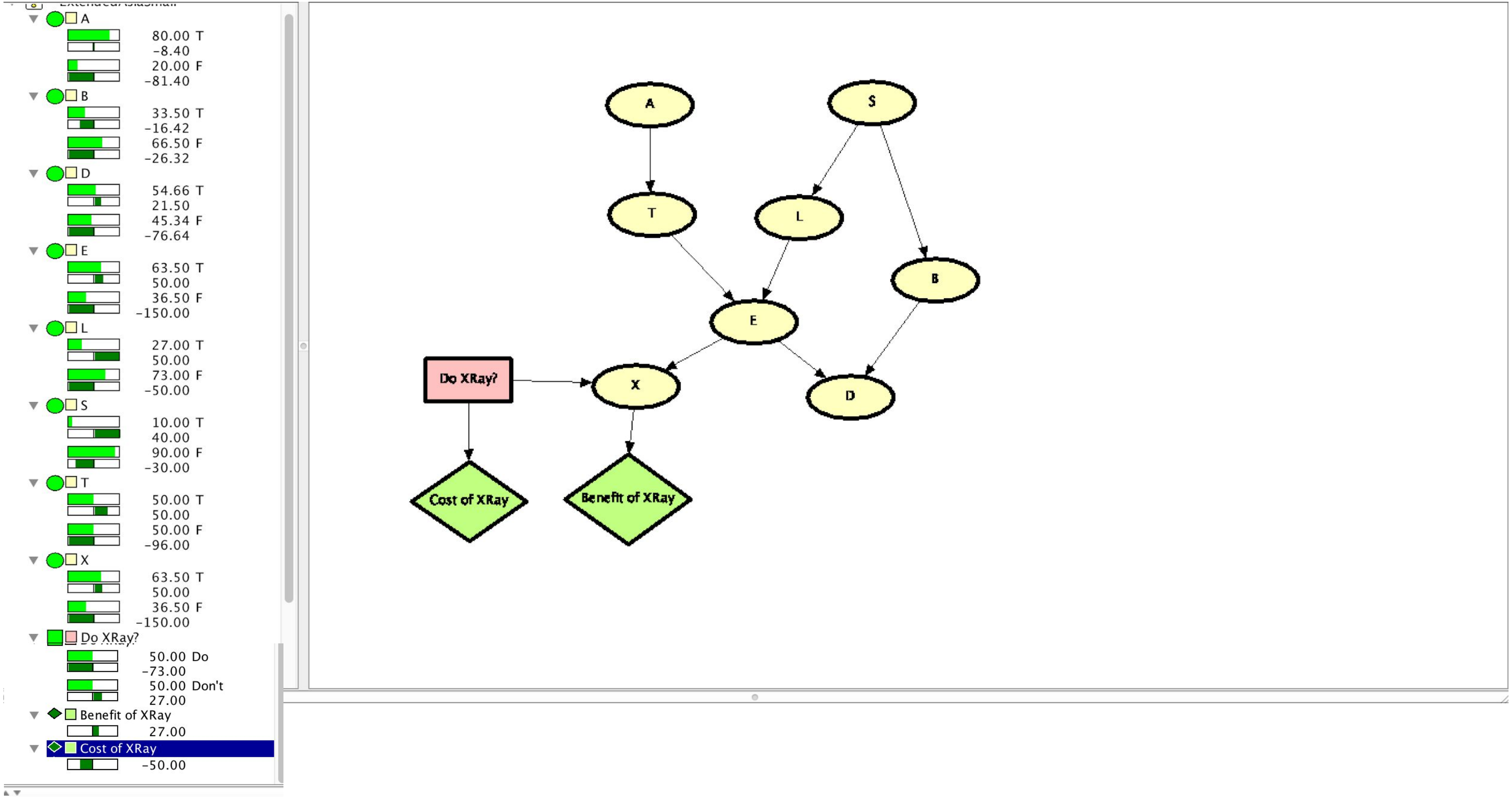
W	Forecast (F)		
	rainy	cloudy	sunny
wet	0.6	0.25	0.15
dry	0.1	0.4	0.5

Accept Bet (AB)	EU
yes	???
no	???

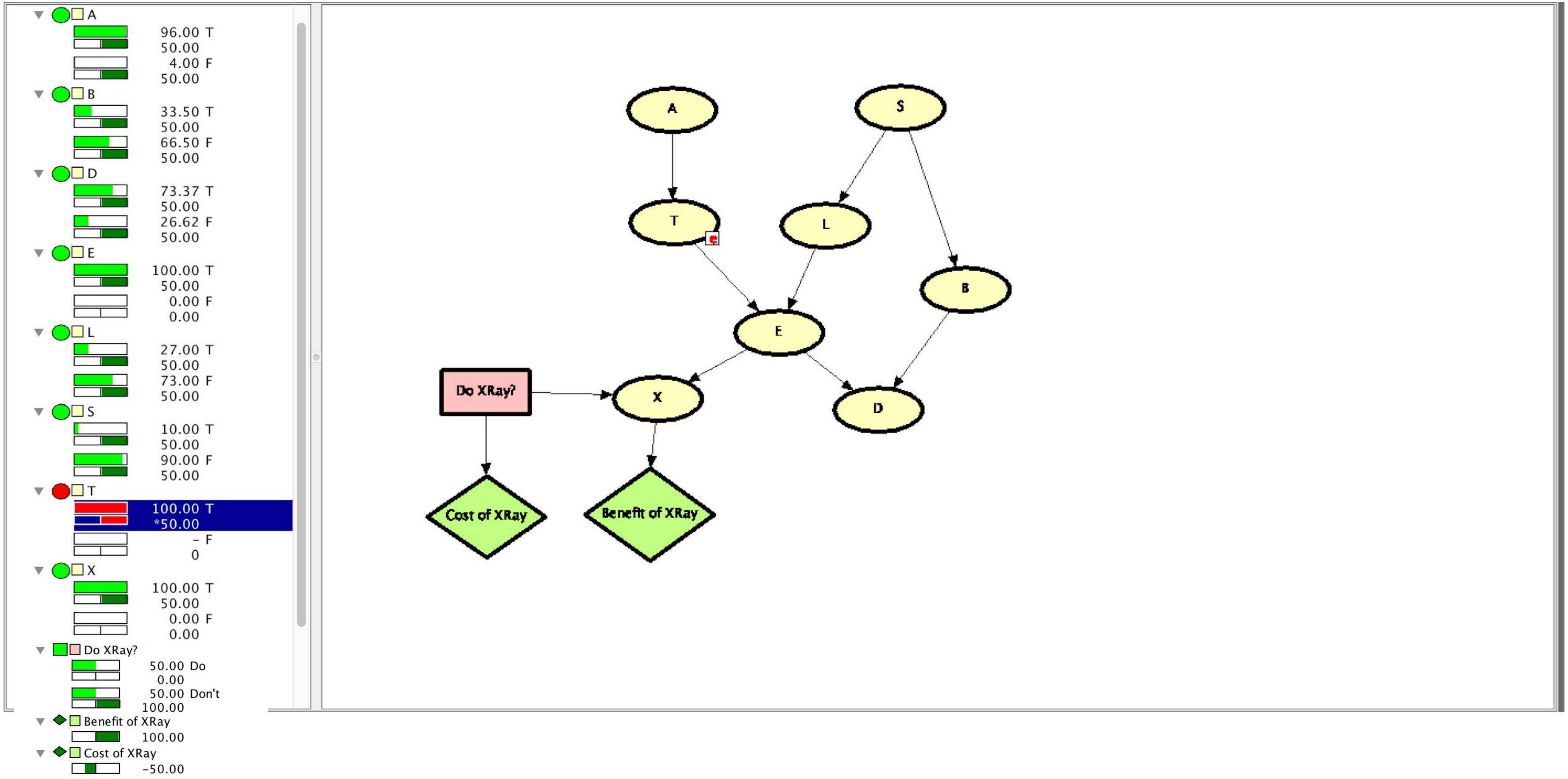
			Accept Bet		
F	**P(W=wet)**	**P(R=win)**	***EU(yes)***	***EU(no)***	**Decision**
rainy	0.720	0.502	**10.12**	7.55	yes
cloudy	0.211	0.324	-0.56	**3.10**	no
sunny	0.114	0.290	-2.61	**2.25**	no

Expected Utility calculation to make a decision on “Accept

Inference in BNs using Hugin



Inference in BNs (Evidence) using Hugin



Scopes of Research

- Compilation (Inference-Reasoning)
  - Continuous
  - Approximate
  - Efficient: Ordinary BN or OOBN
- Automated Learning of Causal Structures
- Knowledge Engineering or Reverse Engineering Applications
  - Crime prediction and criminal detection
  - Disease prediction (Human, Animals and Plants)
  - Profit prediction or Risk Analysis in Investment, Stock Exchange, Defence and etc
- Multi-modal explanation of Causal BN
- Applying BNs in other domains: security, crypto-currency and etc
- Implementation of BN package for python