Probabilistic Reasoning

Why do we need reasoning under uncertainty?

- There are many situations where uncertainty arises:
  - When you travel you reason about the possibility of delays
  - When an insurance company offers a policy it has calculated the risk that you will claim
  - When your brain estimates what an object is it filters random noise and fills in missing details
  - When you play a game you cannot be certain what the other player will do
  - A medical expert system that diagnoses disease has to deal with the results of tests that are sometimes incorrect
- Systems which can reason about the effects of uncertainty should do better than those that don’t.
- But how should uncertainty be represented?

Introduction

- How to build network models to reason under uncertainty according to the laws of probability theory?

- **Bayesian Networks**

  -belief network/probabilistic network/causal network/knowledge map

  -define the syntax & semantics of these networks show how they can be used to capture uncertain knowledge in a natural & efficient way.

What are Bayesian nets?

- Bayesian nets (BN) are a network-based framework for representing and analyzing models involving uncertainty;

- BN are different from other knowledge-based systems tools because uncertainty is handled in mathematically rigorous yet efficient and simple way

- BN are different from other probabilistic analysis tools because of network representation of problems, use of Bayesian statistics, and the synergy between these

Definition of a Bayesian Network

Knowledge structure:

- variables are nodes
- arcs represent probabilistic dependence between variables
- conditional probabilities encode the strength of the dependencies

Computational architecture:

- computes posterior probabilities given evidence about some nodes
- exploits probabilistic independence for efficient computation

Bayesian networks

- A simple, graphical notation for conditional independence assertions & hence for compact specification of full joint distributions
- A Bayesian network is a directed graph in which each node is annotated with quantitative probability information.

Syntax:

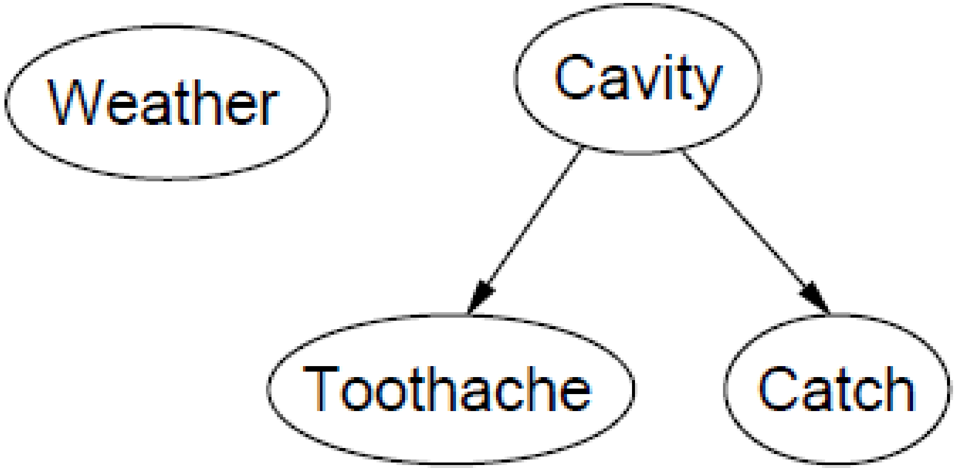
- A set of nodes, one per variable
- A set of directed links/arrows connects pairs of nodes. If there is an arrow from node X to node Y, X is said to be a *parent* of Y
- A directed, acyclic graph (link ≈ "directly influences")
- A conditional distribution for each node given its parents:

$$P(X_i|Parents(X_i))$$

- In the simplest case, conditional distribution represented as a conditional probability table (CPT) giving the distribution over X_i for each combination of parent values

Example

Topology of network encodes conditional independence assertions:

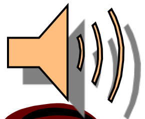


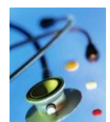
$Weather$ is independent of the other variables

$Toothache$ and $Catch$ are conditionally independent given $Cavity$

Cavity is a direct cause of *Toothache* & *Catch*, whereas no direct causal relationship exists between *Toothache & Catch*

Another Example





- You have a new burglar alarm installed at home. It is fairly reliable at detecting a burglary, but also responds on occasion to minor earthquakes.
- You also have 02 neighbors, John & Mary, who have promised to call you at work when they hear the alarm.
- John always call hears the alarm, but sometimes confuses the telephone ringing with the alarm & calls then, too.
- Mary, on the other hand, likes rather loud music & sometimes misses the alarm altogether.
- Given the evidence who has or has not called,

-estimate the probability of a burglary

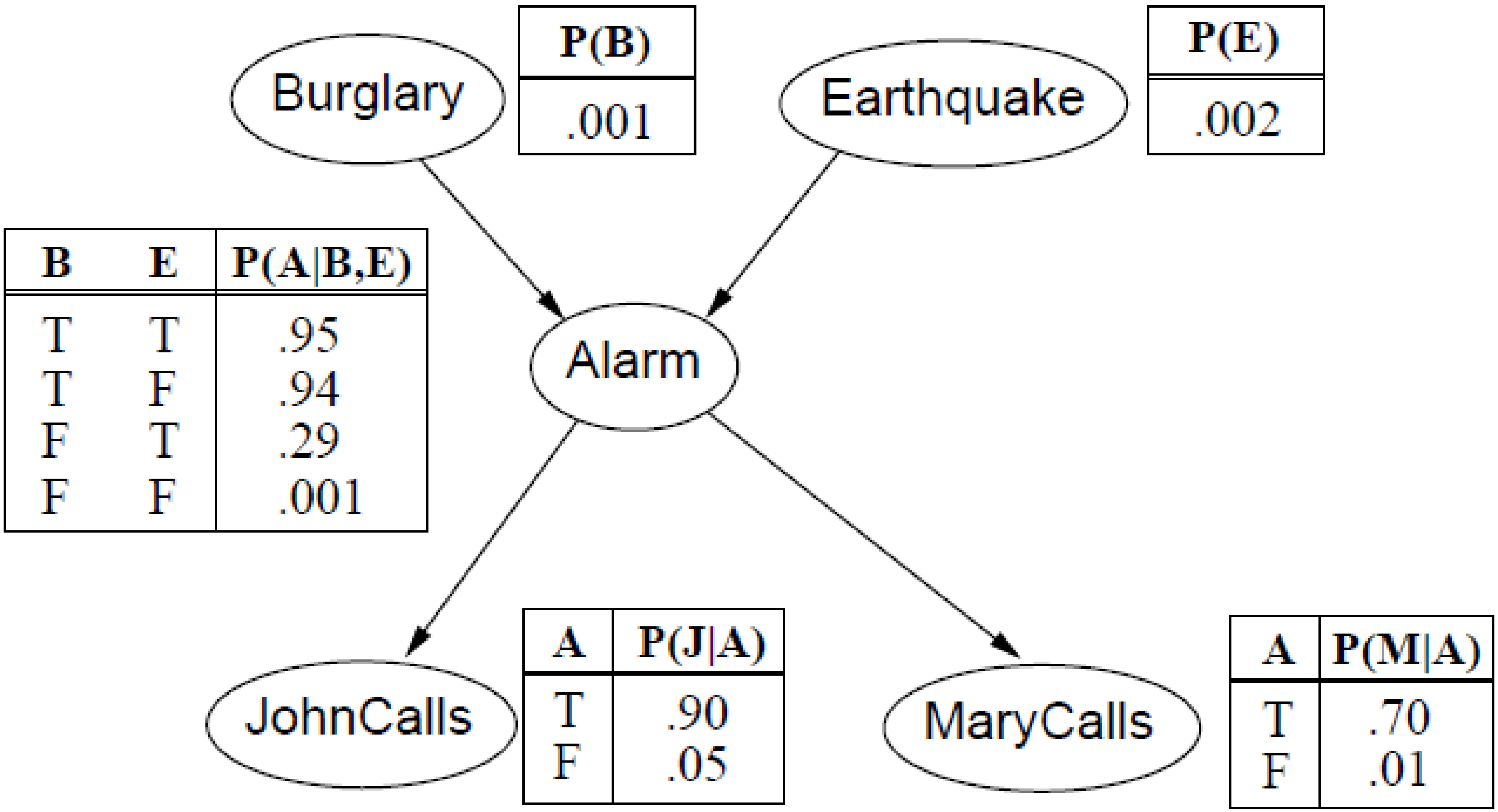
Example

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

Variables: $Burglar$, $Earthquake$, $Alarm$, $JohnCalls$, $MaryCalls$
Network topology reflects "causal" knowledge:

- A burglar can set the alarm off
- An earthquake can set the alarm off
- The alarm can cause Mary to call
- The alarm can cause John to call

Example contd.



P(B)
.001

P(E)
.002

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

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

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

Semantics of Bayesian Networks

Two ways:

1. See the network as a representation of the joint probability distribution

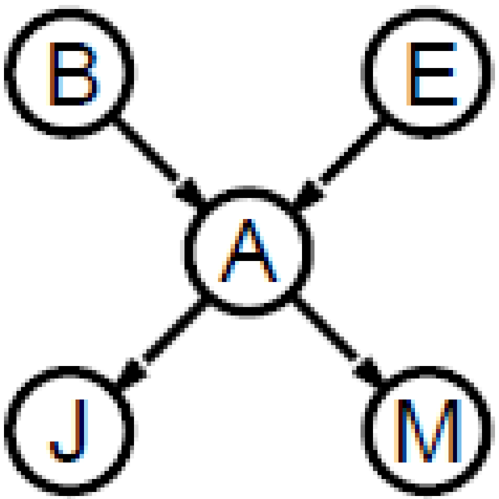
   -helpful in understanding how to construct networks

2. View the network as an encoding of a collection of conditional independence statements.

   -helpful in designing inference procedures

Compactness

A CPT for Boolean X_i with k Boolean parents has 2^k rows for the combinations of parent values



Each row requires one number p for $X_i = true$ (the number for $X_i = false$ is just $1 - p$)

If each variable has no more than k parents, the complete network requires $O(n \cdot 2^k)$ numbers

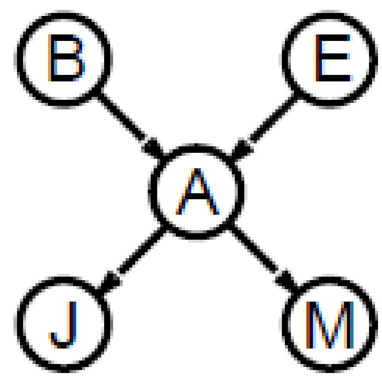
I.e., grows linearly with n, vs. $O(2^n)$ for the full joint distribution

For burglary net, $1 + 1 + 4 + 2 + 2 = 10$ numbers (vs. $2^5 - 1 = 31$)

Global semantics

"Global" semantics defines the full joint distribution as the product of the local conditional distributions:

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | parents(X_i))$$



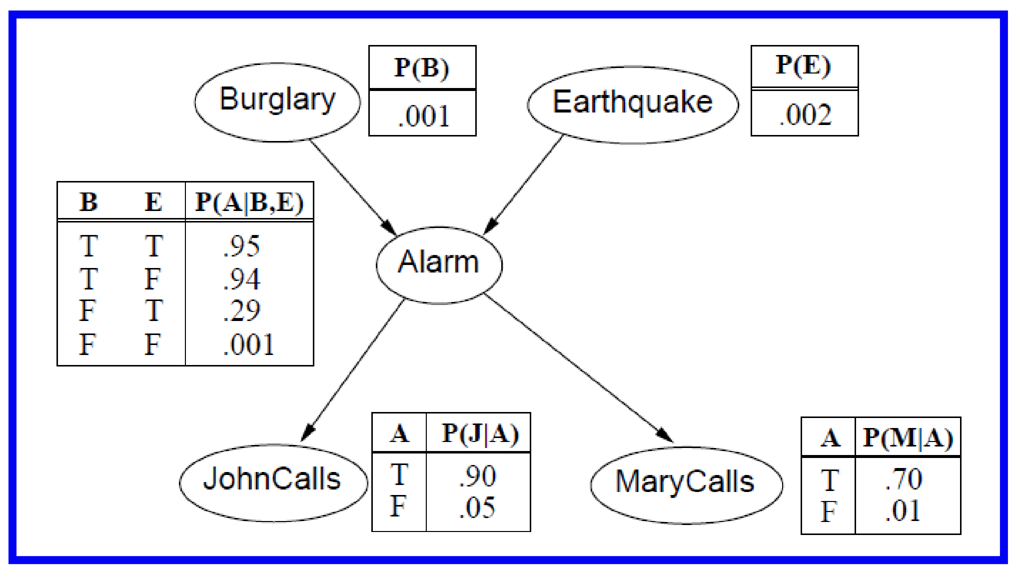
e.g., $P(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$

$$= P(j|a)P(m|a)P(a|\neg b, \neg e)P(\neg b)P(\neg e)$$

$$= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998$$

$$\approx 0.00063$$

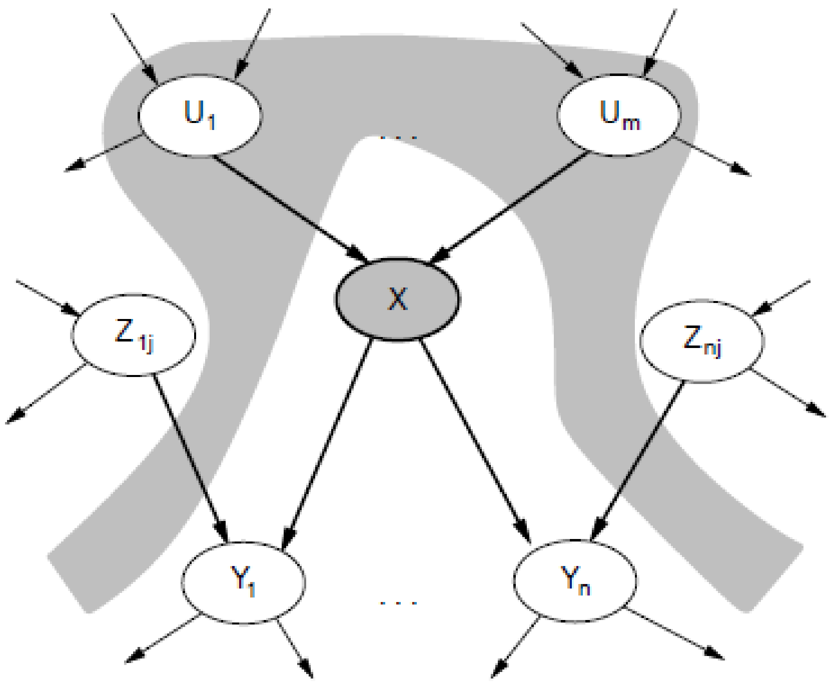
John calls (J) and Mary calls (M). What is the probability that neither burglary nor earthquake occurred if the alarm rang?



Local semantics

- Each node is conditionally independent of its non-descendants given its parents

A node X is conditionally independent of its non-descendants (Z_{ij}S) given its parents (the U_is shown in the gray area)

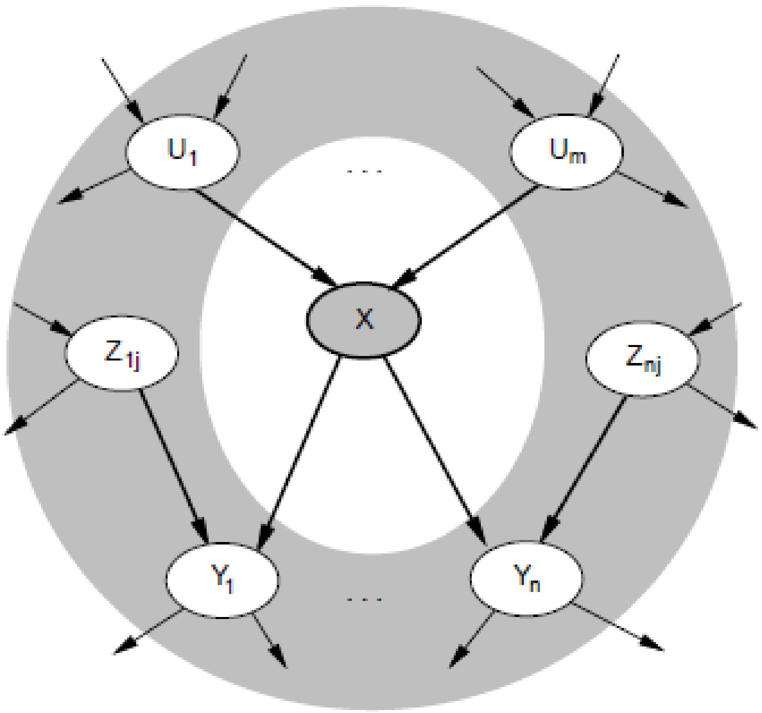


Johncalls is independent of Burglary & Earthquake

Markov Blanket

- Each node is conditionally independent of all others given its

Markov blanket: parents + children + children's parents



Burglary is independent of JohnCalls & MaryCalls, given Alarm & Earthquake

Constructing Bayesian networks

Need a method such that a series of locally testable assertions of conditional independence guarantees the required global semantics

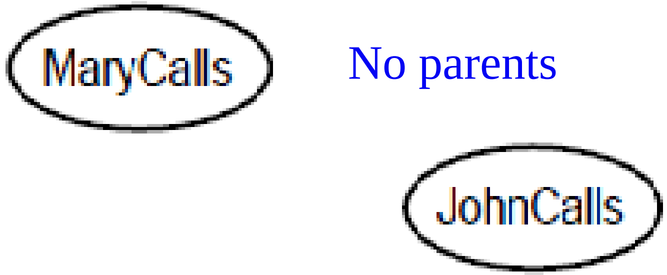
1. Choose an ordering of variables $X_1, \ldots, X_n$
2. For $i = 1$ to n
   add X_i to the network
   select parents from $X_1, \ldots, X_{i-1}$ such that
   $\mathbf{P}(X_i | Parents(X_i)) = \mathbf{P}(X_i | X_1, \ldots, X_{i-1})$

This choice of parents guarantees the global semantics:

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

Example

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

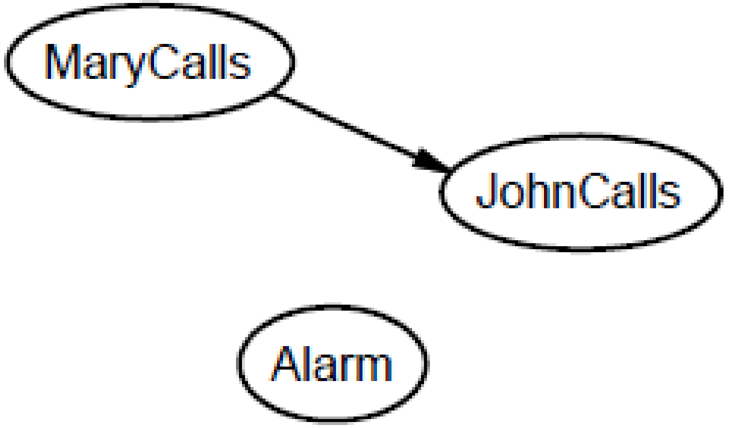


No parents

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

Example

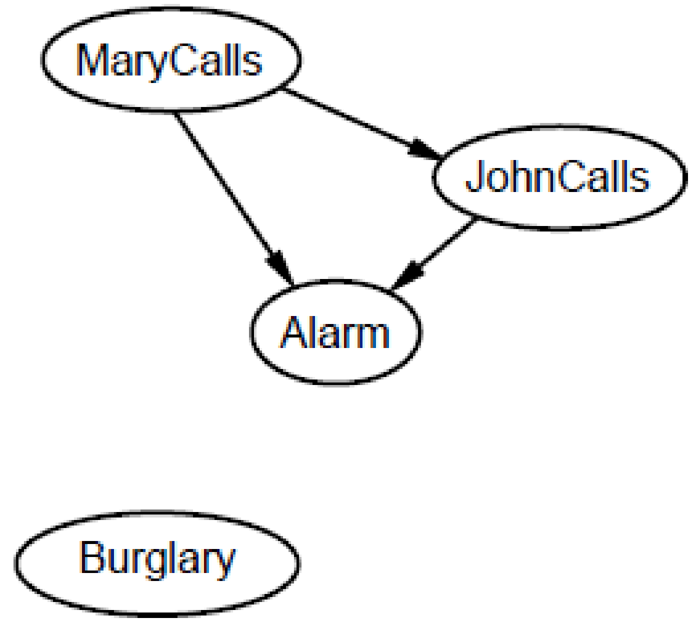
Suppose we choose the ordering M, J, A, B, E



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

Example

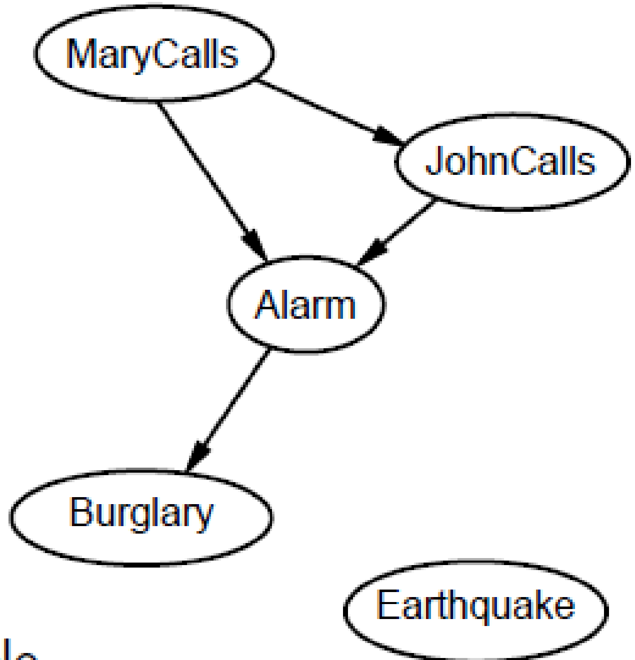
Suppose we choose the ordering M, J, A, B, E



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

Example

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



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

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

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

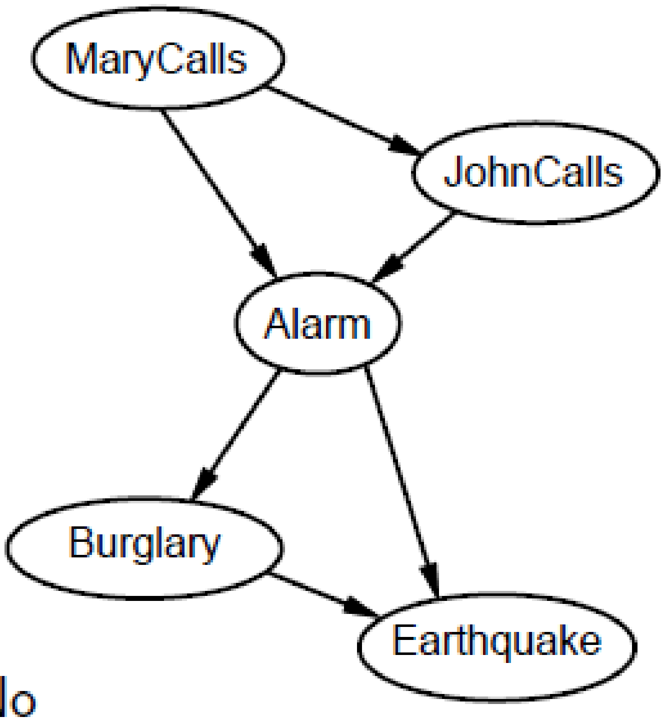
$P(B|A,J,M) = P(B)$? No

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

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

Example

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



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

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

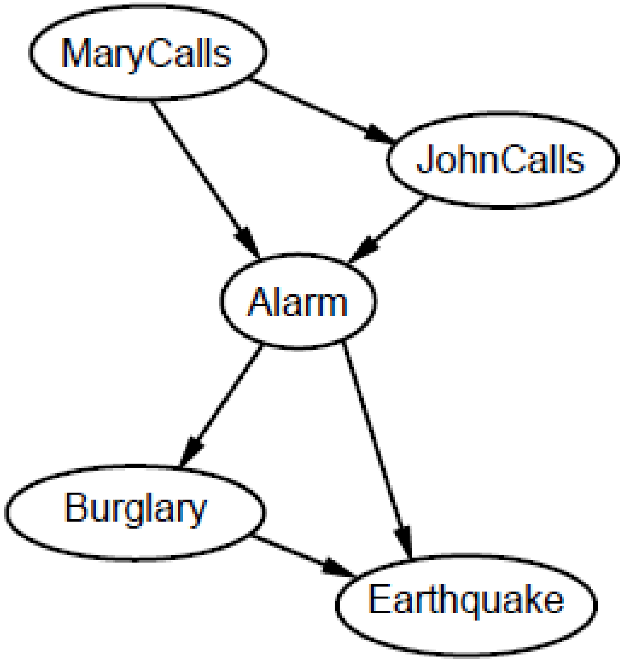
$P(B|A,J,M) = P(B|A)$? Yes

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

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

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

Example



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

Compact Conditional Distributions

CPT grows exponentially with number of parents **O (2^k)**

CPT becomes infinite with continuous-valued parent or child

Solution: canonical distributions that are defined compactly

Deterministic nodes are the simplest case:

$X = f(Parents(X))$ for some function f

E.g., Boolean functions

$NorthAmerican \Leftrightarrow Canadian \vee US \vee Mexican$

E.g., numerical relationships among continuous variables

$$\frac{\partial Level}{\partial t} = \text{ inflow + precipitation - outflow - evaporation}$$

Compact Conditional Distributions

P(¬fever|cold, ¬flu, ¬malaria) = 0.6
P(¬fever| ¬cold, flu, ¬malaria) = 0.2
P(¬fever| ¬cold, ¬flu, malaria) = 0.1

Noisy-OR distributions model multiple noninteracting causes

1) Parents $U_1 \ldots U_k$ include all causes (can add leak node)
2) Independent failure probability q_i for each cause alone

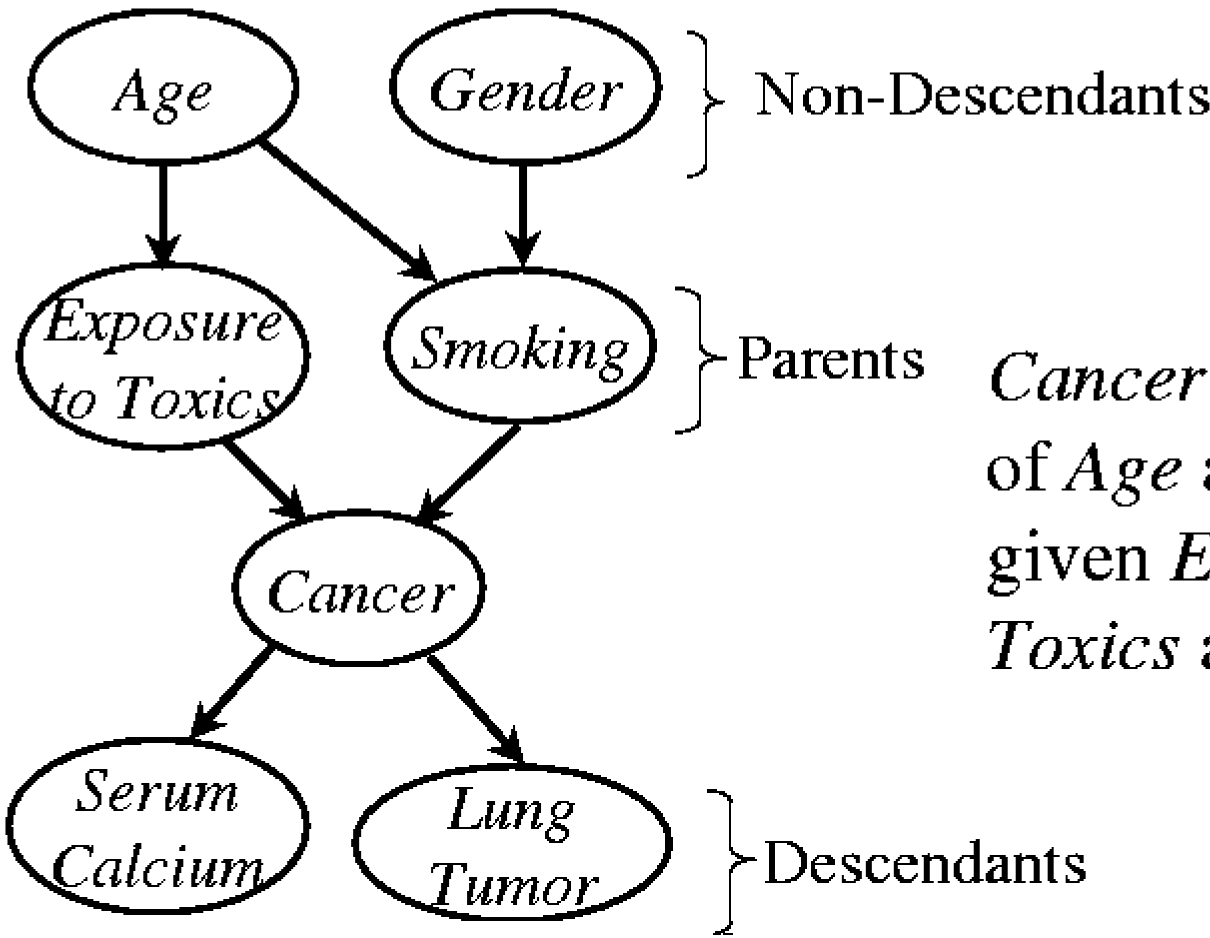
$$\Rightarrow P(X|U_1 \ldots U_j, \neg U_{j+1} \ldots \neg U_k) = 1 - \Pi_{i=1}^{j} q_i$$

$Cold$	Flu	$Malaria$	$P(Fever)$	$P(\neg Fever)$
F	F	F	$\mathbf{0.0}$	1.0
F	F	T	0.9	$\mathbf{0.1}$
F	T	F	0.8	$\mathbf{0.2}$
F	T	T	0.98	$0.02 = 0.2 \times 0.1$
T	F	F	0.4	$\mathbf{0.6}$
T	F	T	0.94	$0.06 = 0.6 \times 0.1$
T	T	F	0.88	$0.12 = 0.6 \times 0.2$
T	T	T	0.988	$0.012 = 0.6 \times 0.2 \times 0.1$

Number of parameters **linear** in number of parents

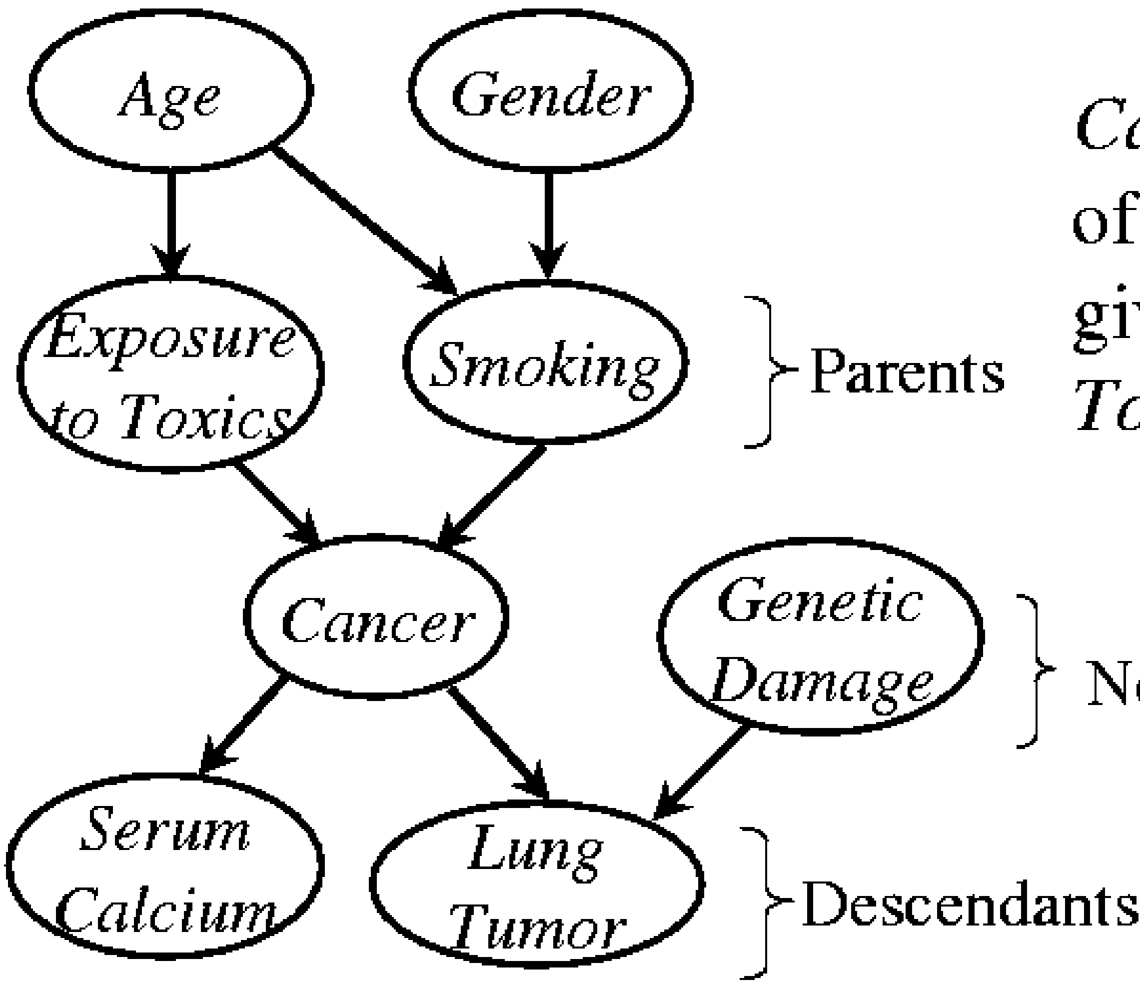
Conditional Independence

A variable (node) is conditionally independent of its non-descendants given its parents.



Cancer is independent of *Age* and *Gender* given *Exposure to Toxics* and *Smoking*.

Another non-descendant



Cancer is independent of *Genetic Damage* given *Exposure to Toxics* and *Smoking*.

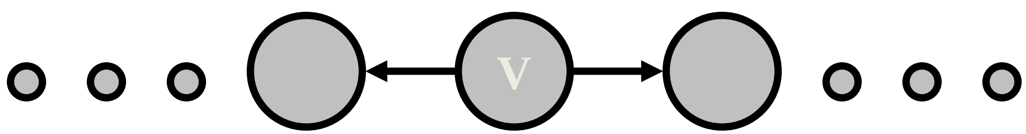
D-Separation of variables

- Fortunately, there is a relatively simple algorithm for determining whether two variables in a Bayesian network are conditionally independent: *d-separation*.
- Definition: X and Z are *d-separated* by a set of evidence variables E iff every undirected path from X to Z is “blocked”.

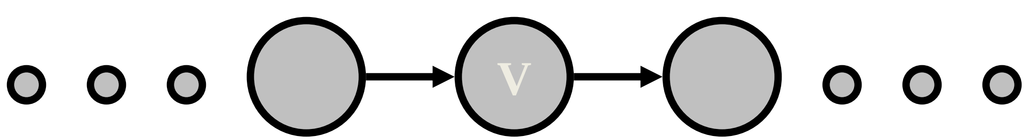
Conditions for D-Seperation

A path is “blocked” iff one or more of the following conditions is true:

1. There exists a variable *V* on the path such that
   – it is in the evidence set *E*
   – the arcs putting *V* in the path are “tail-to-tail”

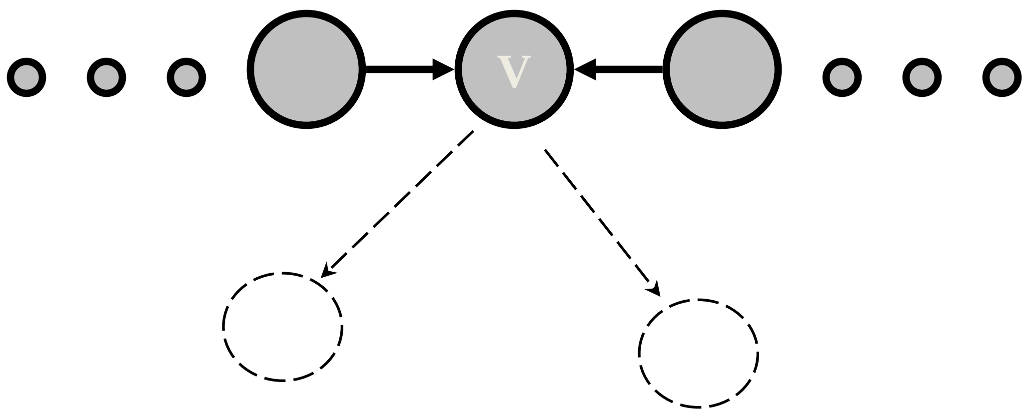


2. There exists a variable *V* on the path such that
   – it is in the evidence set *E*
   – the arcs putting *V* in the path are “tail-to-head”



Conditions for D-Seperation

3. there exists a variable *V* on the path such that
   - it is NOT in the evidence set *E*
   - neither are any of its descendants
   - the arcs putting *V* on the path are "head-to-head"



D-Separation and Independence

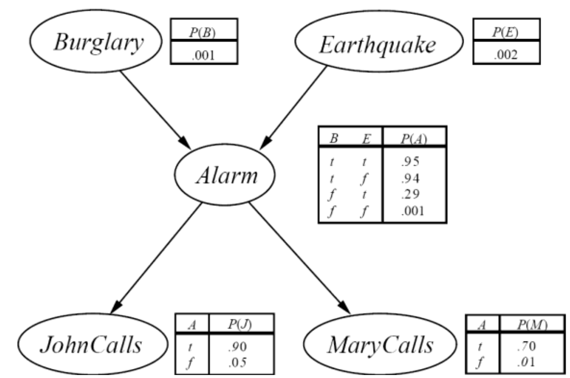
- **Theorem [Verma & Pearl, 1998]:**
  - If a set of evidence variables E d-separates X and Z in a Bayesian network's graph, then X and Z will be independent.

- *d*-separation can be computed in linear time.

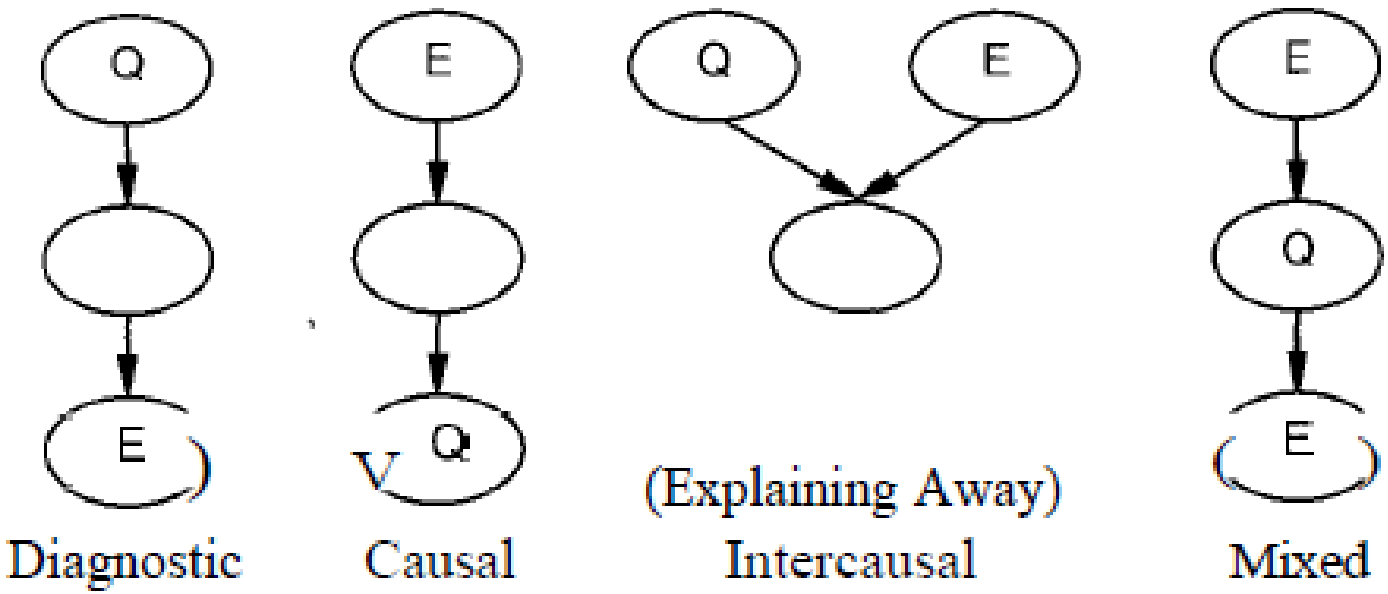
- Thus we now have a fast algorithm for automatically inferring whether learning the value of one variable might give us any additional hints about some other variable, given what we already know.

Inference (Reasoning) in Bayesian Networks

- Consider answering a query in a Bayesian Network
  - Q = set of query variables
  - e = evidence (set of instantiated variable-value pairs)
  - Inference = computation of conditional distribution P(Q | e)
- Examples
  - P(burglary | alarm)
  - P(earthquake | JCalls, MCalls)
  - P(JCalls, MCalls | burglary, earthquake)



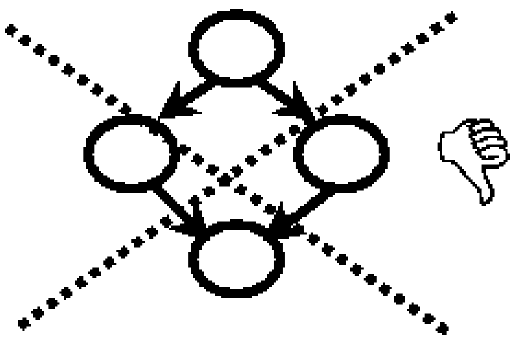
- Can we use the structure of the Bayesian Network to answer such queries efficiently? Answer = yes
  - Generally speaking, complexity is inversely proportional to sparsity of graph
- Can derive an algorithm for BELIEF-NET-ASK (Artificial Intelligence-A Modern Approach by Stuart J. Russell and Peter Norvig, Page-447)

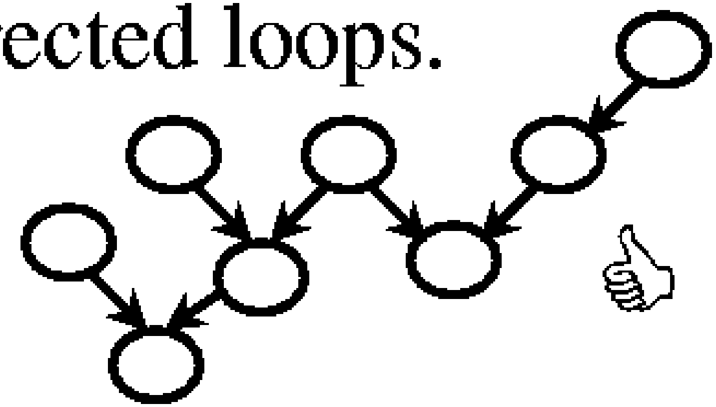


Four patterns of reasoning that can be handled by belief networks.
E represents an evidence variable and Q is a query variable.

Polytrees

- A network is *singly connected* (a *polytree*) if it contains no undirected loops.





Theorem: Inference in a singly connected network can be done in linear time*.

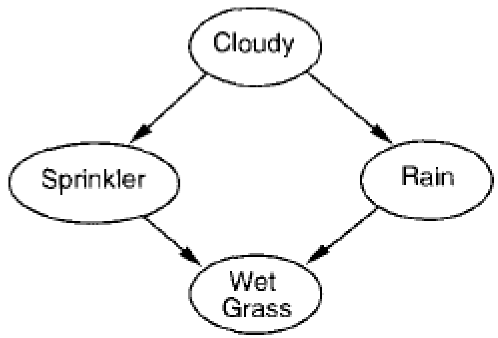
Main idea: in variable elimination, need only maintain distributions over single nodes.

Join tree

- **A multiply connected** graph is one in which two nodes are connected by more than one path.
- Three basic classes of algorithms for evaluating multiply connected networks:
  1. **Clustering**
  2. **Conditioning**
  3. **Stochastic simulation**

Example

P(C) = .5



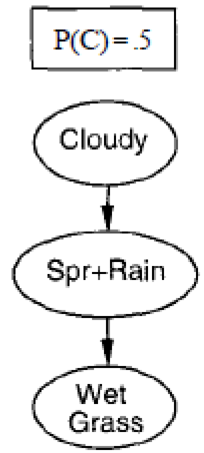
C	P(S)
T	.10
F	.50

C	P(R)
T	.80
F	.20

S	R	P(W)
T	T	.99
T	F	.90
F	T	.90
F	F	.00

A multiply connected network with conditional probability tables.

Example

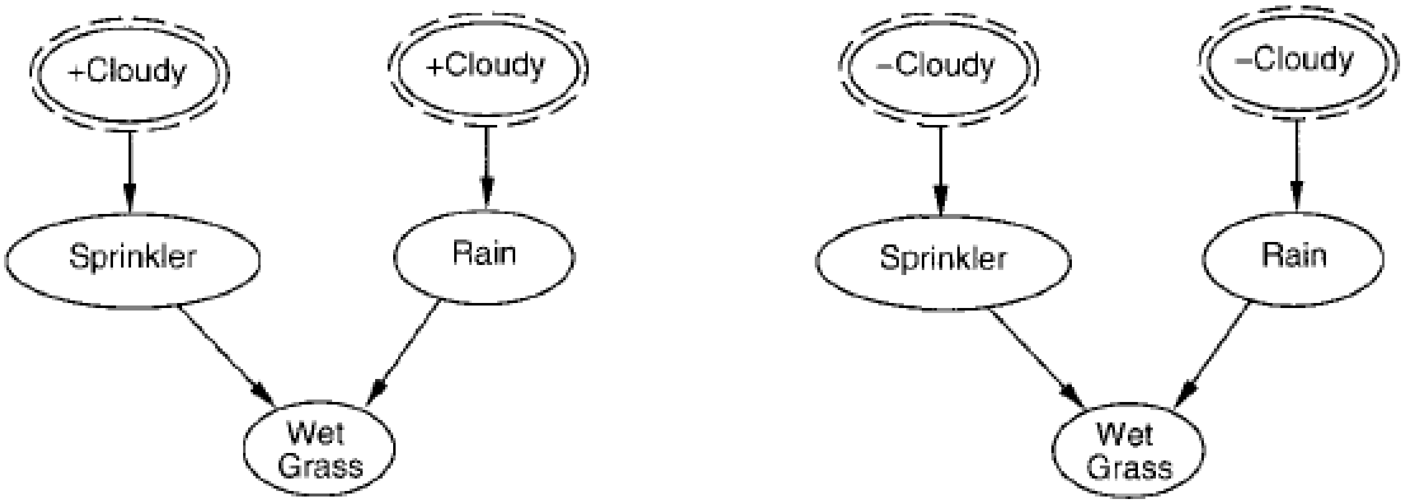


C	P(S+R=x) TT	TF	FT	FF
T	.08	.02	.72	.18
F	.40	.10	.40	.10

S+R	P(W)
T T	.99
T F	.90
F T	.90
F F	.00

A clustered equivalent of the multiply connected network.

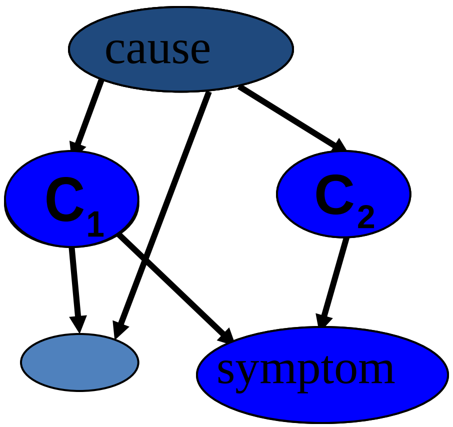
Example

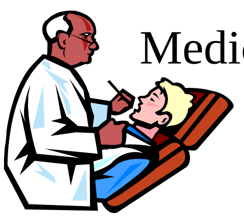


Networks created by instantiation.

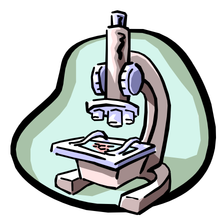
What Bayesian Networks are good for?

- Diagnosis: P(cause|symptom)=?
- Prediction: P(symptom|cause)=?
- Classification: $\max_{class} P(class|data)$
- Decision-making (given a cost function)

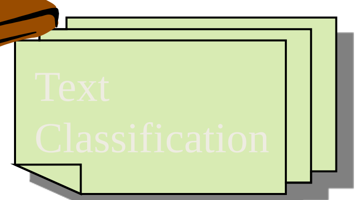


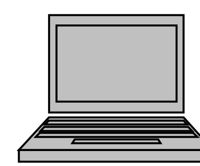

Medicine

Speech recognition



Bio-informatics

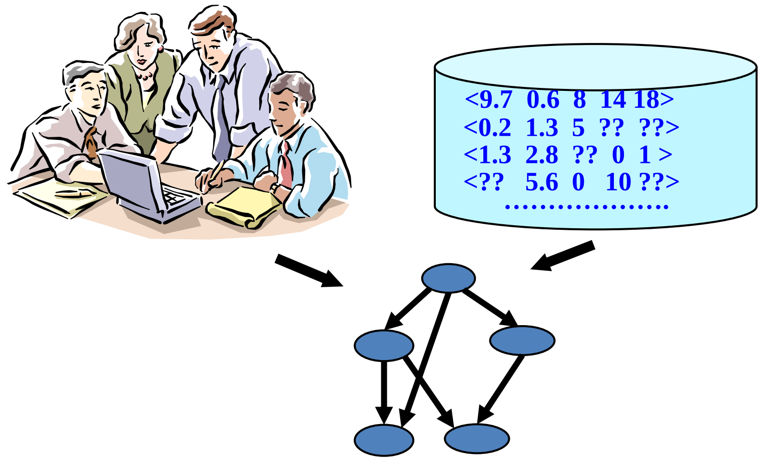

Stock market




Computer troubleshooting

Why learn Bayesian networks?

- Combining domain expert knowledge with data



- Efficient representation and inference

- Incremental learning

- Handling missing data: **<1.3 2.8 ?? 0 1 >**

- Learning causal relationships:

THE END