BAYESIAN NETWORKS

Today

- Reading
  - AIMA Chapter 14.1-14.4
- Goals
  - Bayesian networks
  - (Exact inference in Bayesian networks)

Summary of distributions so far

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

Marginal

Conditional

Joint $p(x,y)$

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

The Product Rule

- Given the conditional and marginal distributions, we can compute the joint distribution using the Product Rule:

$$p(x|y) = \frac{p(x,y)}{p(y)} \quad \Rightarrow \quad p(x,y) = p(x|y) \cdot p(y)$$

- Represents the joint distribution in a causal and more natural way:
  - Intelligence = {high, low}
  - SAT = {high, low}
  - p(Intelligence, SAT) = p(SAT | Intelligence) p(Intelligence)

The Chain Rule

- In general, the joint distribution of a set of random variables can be expressed as a product of conditional and marginal distributions

$$p(x_1, \ldots, x_n) = p(x_1) \cdot p(x_2|x_1) \ldots p(x_n|x_1, \ldots, x_{n-1})$$
$$= \prod_i p(x_i|x_1, \ldots, x_{i-1})$$

- Derived from repeated applications of the Product rule

Independence

- Two variables are independent if knowing the value of one variable **does not** alter the distribution of the other variable
- Mathematical definition:

$$p(X = x, Y = y) = p(X = x|Y = y) \cdot p(Y = y) \qquad \forall x, y$$
$$= p(X = x) \cdot p(Y = y)$$

The value of X is independent of the value of Y

- The joint distribution now factors into the product of simpler distributions
- Example
  - p(CoinToss1, CoinToss2) = p(CoinToss1) p(CoinToss2)
  - p(CarAccident, 49ersWin) = p(CarAccident) p(49ersWin)

Conditional independence

- Two variables are conditionally independent if

$$p(X = x, Y = y|Z = z) = p(X = x|Z = z) \cdot p(Y = y|Z = z)$$

- In other words, given Z the variables X and Y are independent
- Examples
  - p(Fever, Headache) = p(Fever | Headache) p(Headache)
  - p(Fever, Headache | Flu) = p(Fever | Flu) p(Headache | Flu)

Moving away from numerical quantities

"The traditional definition of independence uses equality of numerical quantities, as in

p(x, y) = p(x)p(y)

suggesting that one must test whether the joint distribution of X and Y is equal to the product of their marginals in order to determine whether X and Y are independent. By contrast people can easily and confidently detect dependencies, even though they may not be able to provide precise numerical estimates of probabilities. A person who is reluctant to estimate the probability of being burglarized the next day or of having a nuclear war within five years can nevertheless state with ease whether the two events are dependent, namely, whether knowing the truth of one proposition will alter the belief of the other."

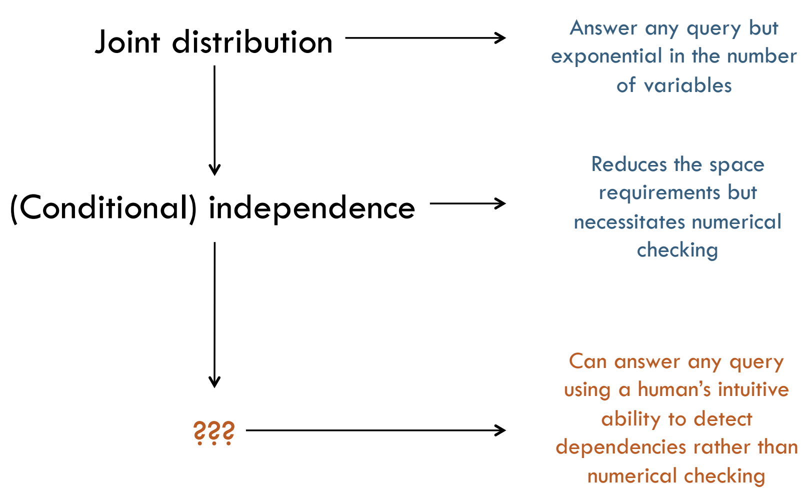
\- Judea Pearl

Moving away from numerical quantities

"It is usually easy for a domain expert to decide what direct influences exist in the domain – much easier, in fact, than actually specifying the probabilities themselves"

- Humans can "easily and confidently" detect dependencies
- Move away from numerical representation of the joint distribution (or the conditional distributions) to a representation that encodes dependencies

Probabilistic Inference



Bayesian Network

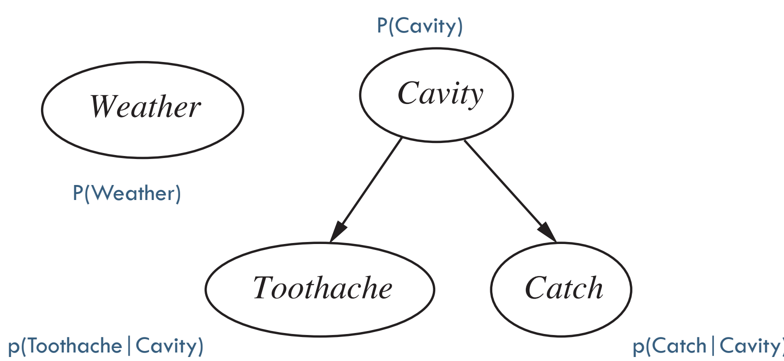
- Bayesian networks represent dependencies among variables and concisely encode the full joint dist.
- A Bayesian network is a directed acyclic graph where:
  - Nodes correspond to random variables
  - Directed edge btw pairs of nodes represent direct influence
  - Each node has a conditional probability distribution

$$p(X_i \mid Parents(X_i))$$

Bayesian Network = Topology + CPT

Bayesian Network Examples

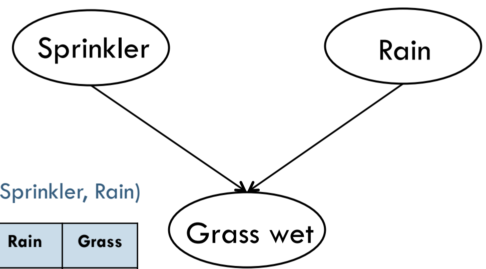
- Weather = {rainy, sunny, cloudy, snowy}
- Cavity = { yes, no}
- Toothache = {yes, no}
- Catch = {yes, no}



Bayesian Network Examples

p(Sprinkler) = 0.5

p(Rain) = 0.2



p(Grass | Sprinkler, Rain)

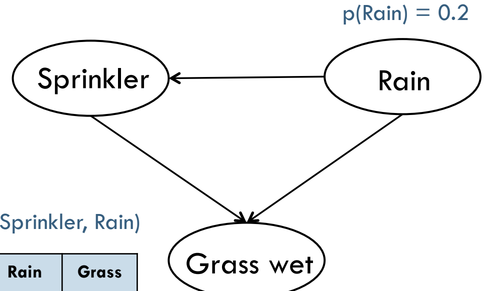
Sprink	Rain	Grass
F	F	0.0
F	T	0.8
T	F	0.9
T	T	0.99

Bayesian Network Examples

p(Sprinkler | Rain)

Rain	Sprink
F	0.5
T	0.01

p(Rain) = 0.2



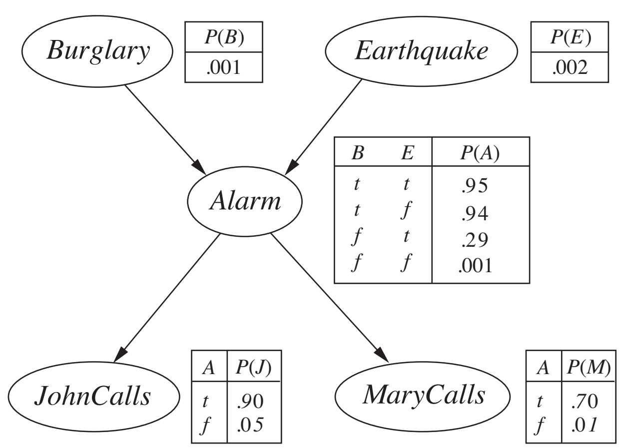
p(Grass | Sprinkler, Rain)

Sprink	Rain	Grass
F	F	0.0
F	T	0.8
T	F	0.9
T	T	0.99

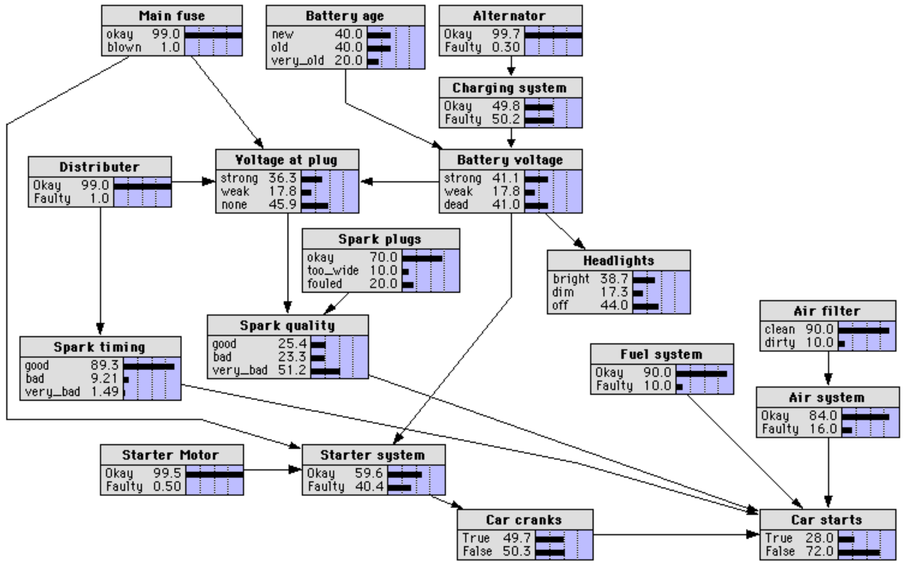
Bayesian Network Examples

- Burglary = {yes, no}
- Earthquake = { yes, no}
- Alarm = {yes, no}
- MaryCalls = {yes, no}
- JohnCalls = {yes no}

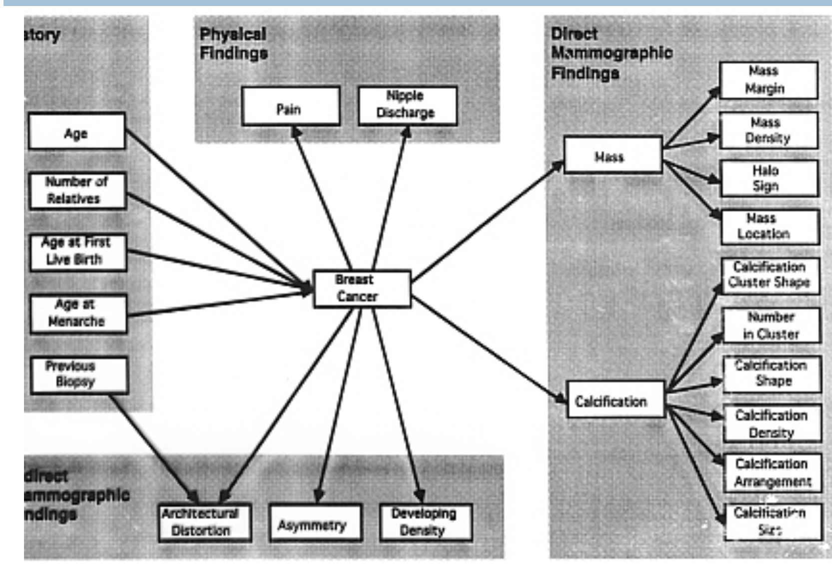
Bayesian Network Examples



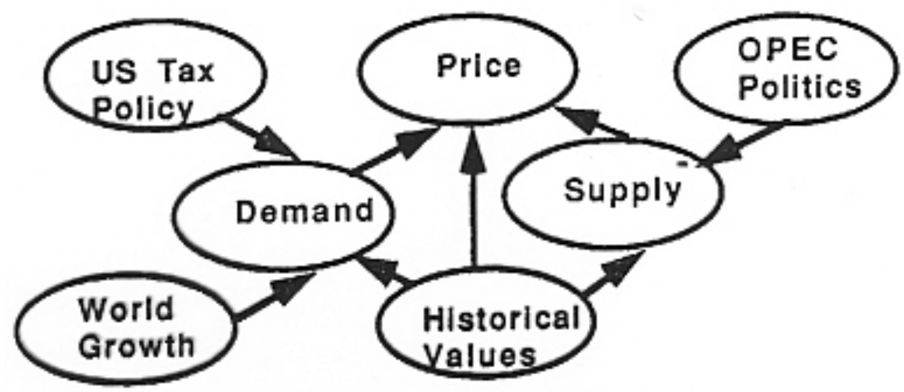
Example: Car diagnosis



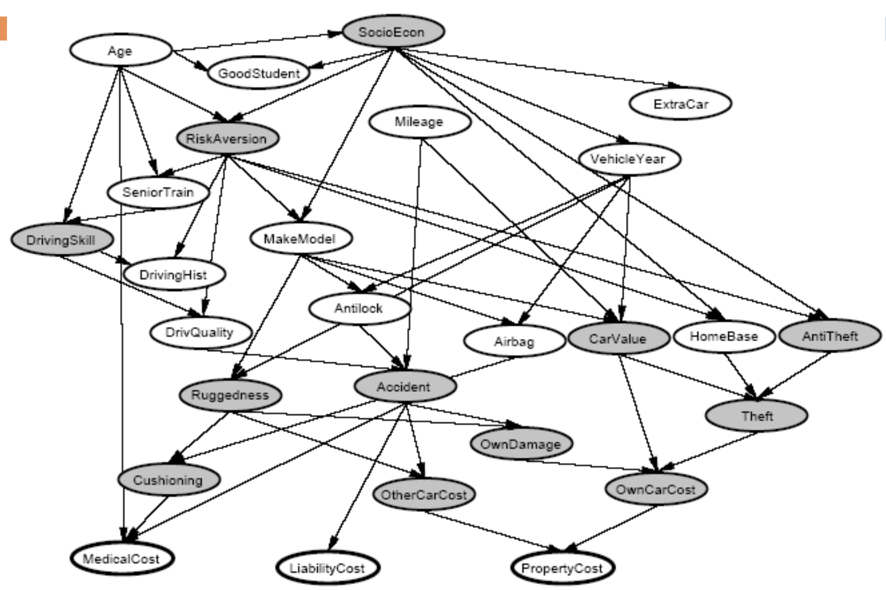
Example: MammoNet



Example: ARCO1 (Forecasting Oil Prices)



Example: Insurance



Representing the joint distribution

- The joint distribution is given by a product of the conditional distributions

$$\begin{aligned} p(j, m, a, \neg b, \neg e) &= p(j|m, a, \neg b, \neg e)p(m|a, \neg b, \neg e)p(a|\neg b, \neg e)p(b|\neg e)p(e) \\ &= p(j|a)p(m|a)p(a|\neg b, \neg e)p(b)p(e) \\ &= 0.9 \cdot 0.7 \cdot 0.001 \cdot 0.999 \cdot 0.998 \end{aligned}$$

- If each variable has k parents, how many probabilities are required?

- N=30 binary variables and k = 5 parents each
  - Bayesian Network requires 960 probabilities
  - The full joint requires over a billion

Constructing a Bayesian Network

Step One: Determine an ordering of the random variables

$\{X_1, X_2, \ldots, X_n\}$

Step Two: For each variable X_i, choose minimal set of nodes from $\{X_1,\ldots,X_{i-1}\}$ required to specify the conditional distribution

$p(X_i \mid Parents(X_i))$

Step Three: Specify the conditional probability tables (CPTs):

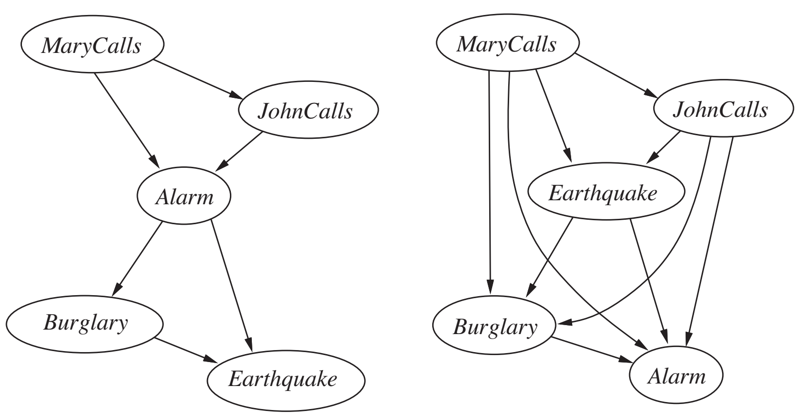
- Interview experts
- Learn from data
  - Learn discrete probabilities
  - Specify a parametric formula, e.g. Gaussian distribution, and learn parameters, e.g. mean and variance, from data

Constructing a Bayesian Network

(MaryCalls, JohnCalls, Alarm, Burglary, Earthquake)

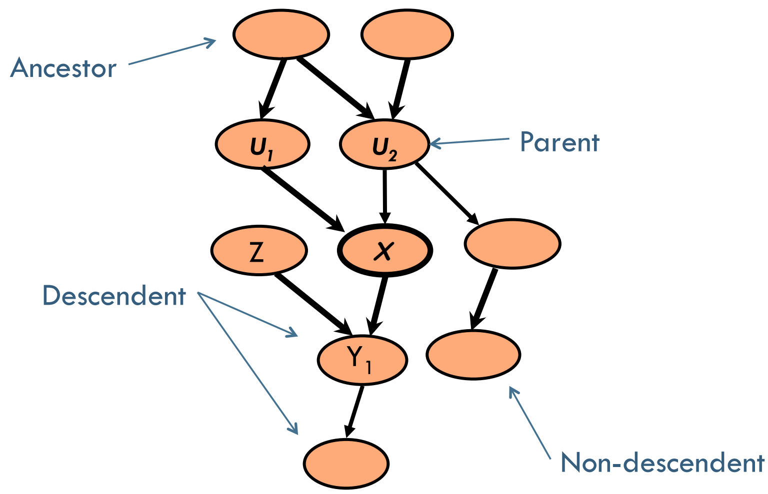
Constructing a Bayesian Network

(MaryCalls, JohnCalls, Alarm, Burglary, Earthquake)

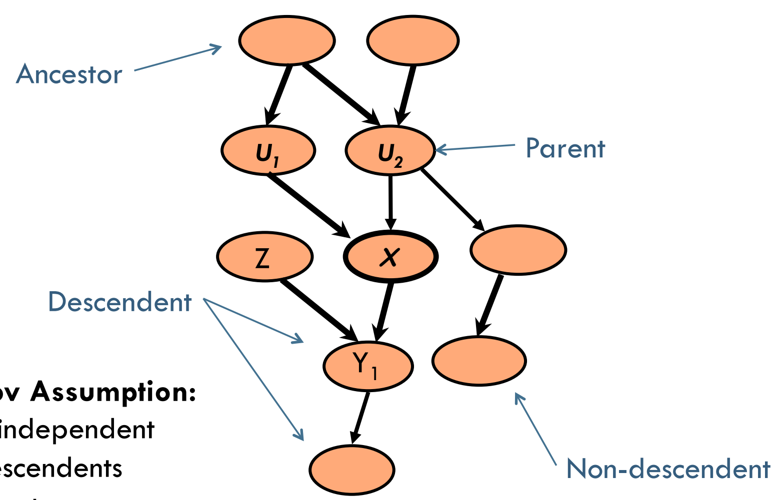


(a) (b)

Bayesian Networks terminology

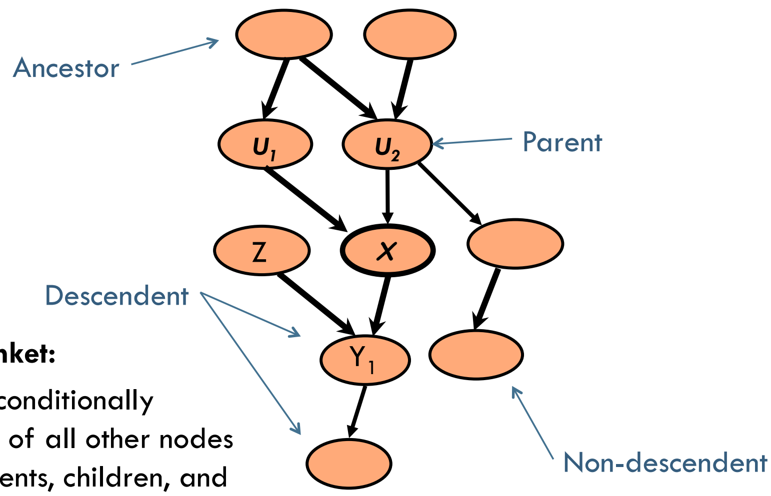


Independence assumptions encoded in the Bayesian Network



Local Markov Assumption:
A node X is independent of its non-descendents given its parents

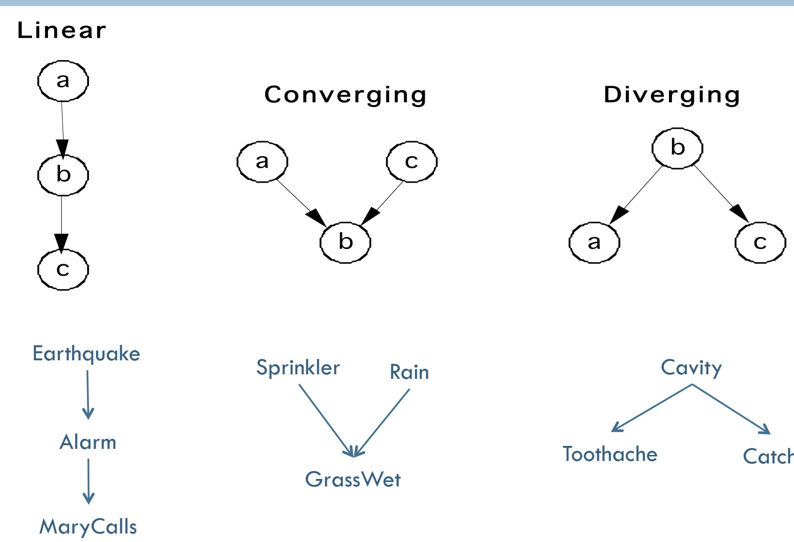
Independence assumptions encoded in the Bayesian Network



Markov Blanket:

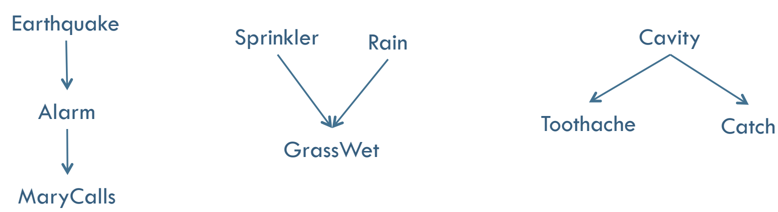
A node X is conditionally independent of all other nodes given its parents, children, and children's parents

Three Types of Connections



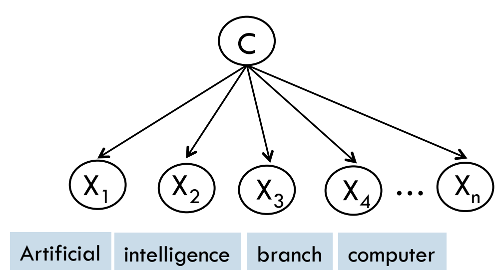
Connection patterns and independence

- **Linear connection:** The two end variables are dependent on each other. The middle variable renders them independent.
- **Converging connection:** The two end variables are independent of each other. The middle variable renders them dependent.
- **Divergent connection:** The two end variables are dependent on each other. The middle variable renders them independent.



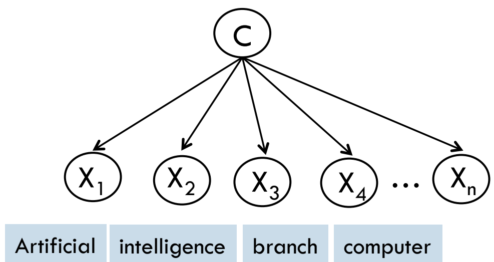
Commonly used Bayesian Networks

- Naïve Bayes Classifier
  - Commonly used for text classification (and medical diagnosis)
  - C is the class (topic or label) of the document
  - The X variables represent the words in the document



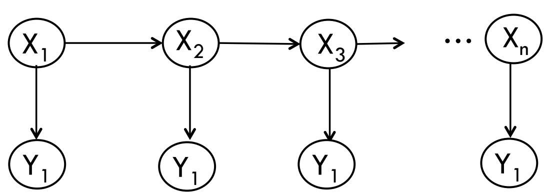
Commonly used Bayesian Networks

- Naïve Bayes Classifier
  - What are the independence assumptions encoded in this BN?
  - Given these independence assumptions, how does the joint distribution factor?
  - What are the distributions that must be specified?



Commonly used Bayesian Networks

- Hidden markov model
  - Used for time series, e.g. speech recognition
  - What are the independence assumptions?



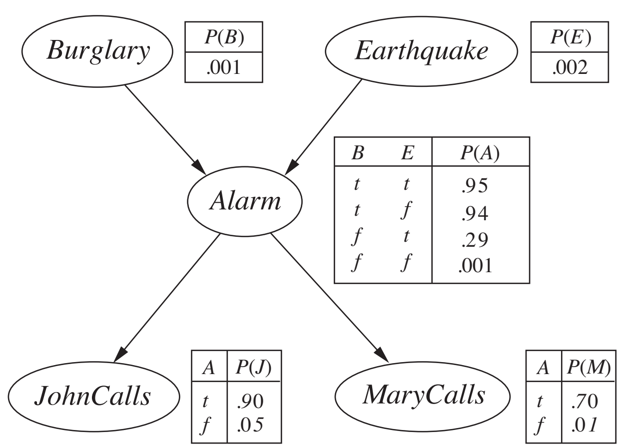
Inference in Bayesian Networks

- Probabilistic inference refers to the task of computing some desired probability given other known probabilities (evidence)

- Exact Inference
  - Enumeration
  - Variable elimination

- Approximate Inference
  - Direct sampling
  - Rejection sampling
  - Likelihood weighting
  - MCMC

Recall: Burglary network



Inference by Enumeration

~~Step One: select the entries in the table consistent with the evidence (this becomes our world)~~

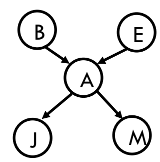
Step Two: sum over the H variables to get the joint distribution of the query and evidence variables

Step Three: Normalize

$$p(b|j,m) \propto \sum_e \sum_a p(b,j,m,e,a)$$ ← Conditional and joint differ only by the normalizing constant

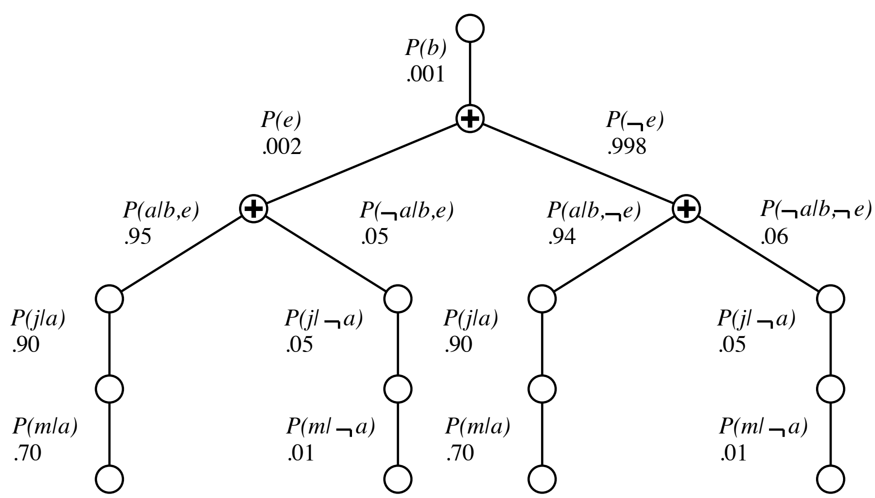
$$= \sum_e \sum_a p(b) \cdot p(e) \cdot p(j|a) \cdot p(m|a) \cdot p(a|b,e)$$ ← Independencies read from BN

$$= p(b) \sum_e p(e) \sum_a p(j|a) \cdot p(m|a) \cdot p(a|b,e)$$ ← Algebraic simplifications



- Compute p(b|j,m) and p(-b|j,m) and then normalize
- May compute the same expression more than once

Inference by Enumeration



Inference by Variable Elimination

- Carry out sums from right to left storing intermediate results to avoid recomputation

$$
\begin{aligned}
p(B|j,m) &= \alpha\; p(B) \sum_e p(e) \sum_a p(a|B,e)\; p(j|a)\; p(m|a) \\
&= \alpha\; f_1(B) \sum_e f_2(e) \sum_a f_3(A,B,E)\; f_4(A)\; f_5(A) \\
&= \alpha\; f_1(B) \sum_e f_2(e)\; f_6(B,E) \\
&= \alpha\; f_1(B)\; f_7(B)
\end{aligned}
$$

- Results are stored in factors (matrices)
- Two operations: pointwise multiplication and summation

Inference by Variable Elimination

- Point-wise multiplication of two factors

A	B	f_1(A,B)	B	C	f_2(B,C)	A	B	C	f_3(A,B,C)
T	T	.3	T	T	.2	T	T	T	
T	F	.7	T	F	.8	T	T	F	
F	T	.9	F	T	.6	T	F	T	
F	F	.1	F	F	.4	T	F	F	
						F	T	T	
						F	T	F	
						F	F	T	
						F	F	F	

- Summing out a variable corresponds to adding submatrices

Inference by Variable Elimination

- Every variable that is not an ancestor of a query variable or evidence variable is irrelevant
- Ordering of variables for summing out affects the time and space of VE
  - For polytrees (at most one path between any two nodes), VE is linear in the size of the network
  - In general, time and space are exponential