Introduction to Bayesian Networks

Syntax, Semantics, and Examples

CS 6316 – Machine Learning
Fall 2017

OUTLINE

- Preface
- Bayesian Networks: Introduction
- Representation
- Conditional Independence
- Alarm Example
- Constructing Bayesian Networks
- Exercise

Preface

- We can divide the large variety of **classification approaches** into roughly three major types:
  1. Discriminative
     - Directly estimate a decision rule/boundary
     - E.g. decision tree (*done*), SVM
  2. Generative
     - Build a generative statistical model
     - E.g. Bayesian networks <----- ***this lecture!***
  3. Instance based classifiers
     - Use observation directly (no models)
     - E.g. K nearest neighbors (*done*)

Bayesian Networks

Material adapted from- I.Rish IBM T.J.Watson Research Center,

A.Moore@cmu-*Bayes Nets for Representing and Reasoning about Uncertainty,*

Norvig et al. *Bayesian Networks,* K.Murphy-*Graphical Models and Bayesian Networks*

Bayesian Networks

- Structured, graphical representations of probabilistic relationships between several random variables
- Explicit representation of conditional independencies
- Missing arcs (edges) encode conditional independence
- Efficient representation of joint PDF P(X)

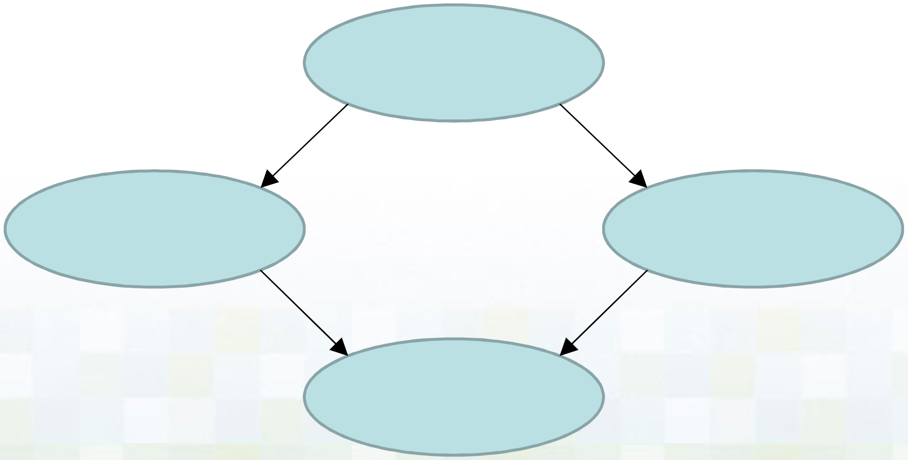
- Generative model (not just discriminative):
  - Allows for arbitrary queries to be answered, e.g.:
  - P(lung cancer=yes | smoking=no, pos X-ray=yes)=?

Bayesian Networks

- A clean, clear, manageable language and methodology for expressing what you're certain and uncertain about
- Already many practical applications in
  - Medicine, Factories, Helpdesks
  - P(this problem | these symptoms) ~ Inference
  - Anomaly detection
  - Choosing next diagnostic test | these observations ~ Active data collection
  - Etc (some more examples later)

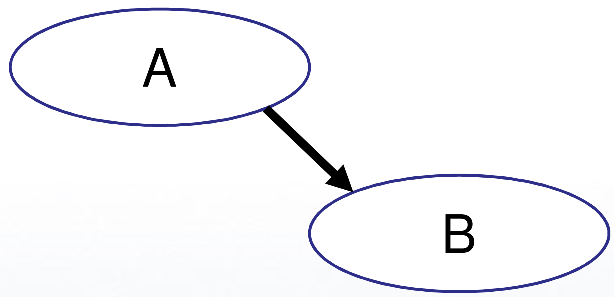
Representation

- Probabilistic graphical models are graphics in which **nodes** represent random variables
- A directed, acyclic graph. The (lack of) **arcs** represent conditional independence assumptions
- Provides a *compact representation* of joint probability distributions



Representation

- *Directed* graphical models (Bayesian Networks or Belief Networks) have represent the notion of independence by taking into account the directionality of the arcs
- An arc from node A to node B indicates that "A causes B"
  - This can be used as a guide to construct the graph structure

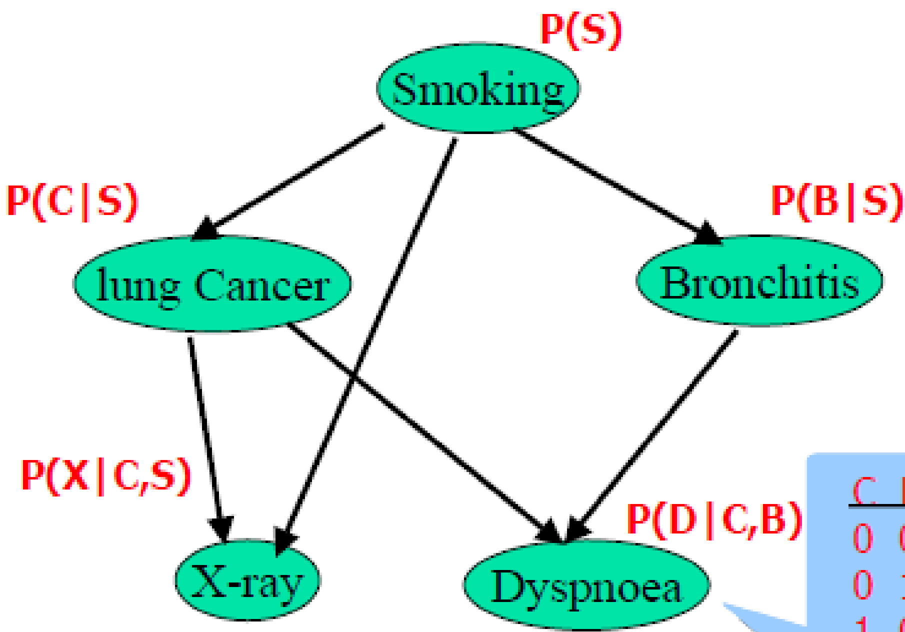


Representation

- For a directed model, we must specify the **Conditional Probability Distribution (CPD)** at each node
- If the variables are **discrete**, this can be represented as a **Conditional Probability Table (CPT)**, which lists the probability that the child node takes on each of its different values for each combination of values of its parents.

Bayesian Network

$BN = (G, \Theta)$



G - directed acyclic graph (DAG)
nodes – random variables
edges – direct dependencies

Θ - set of parameters in all conditional probability distributions (CPDs)

CPD:

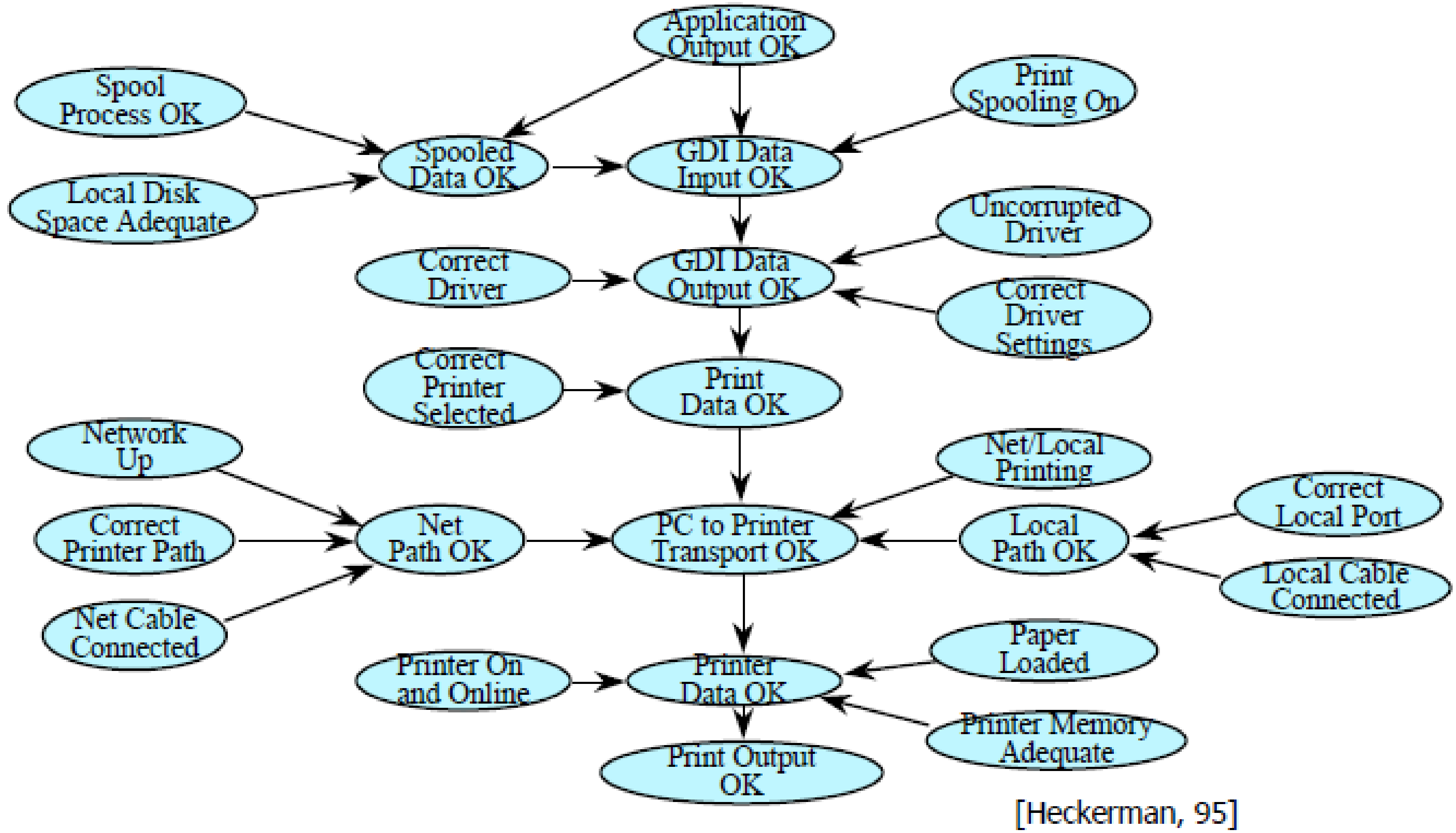
C	B	D=0	D=1
0	0	0.1	0.9
0	1	0.7	0.3
1	0	0.8	0.2
1	1	0.9	0.1

CPD of node X: P(X|parents(X))

- Compact representation of joint distribution is a product form (chain rule) $P(S, C, B, X, D) = P(S)\, P(C|S)\, P(B|S)\, P(X|C,S)\, P(D|C,B)$

$1+2+2+4+4 = 13$ parameters, instead of $2^5 = 32$

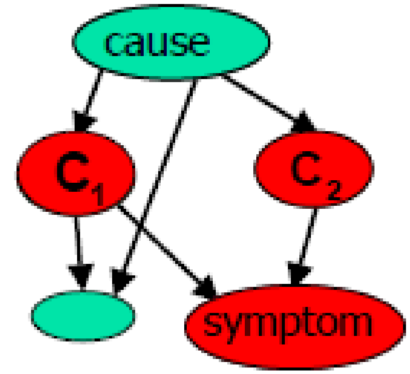
Example: Printer Troubleshooting

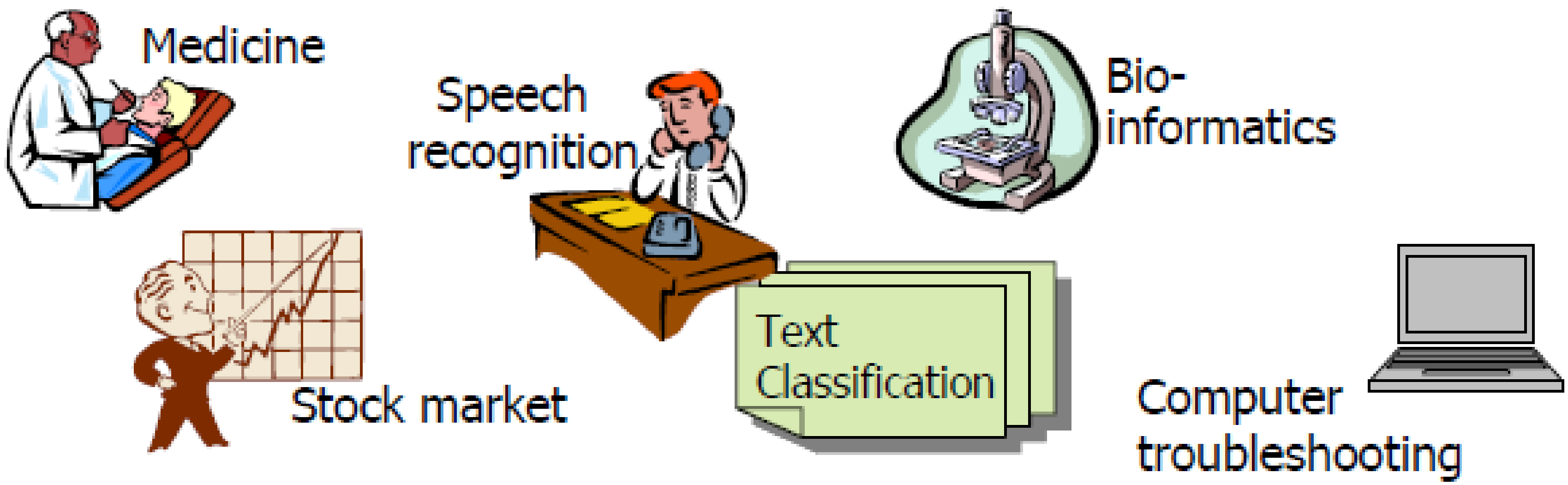


- 26 variables. Instead of 2^{26} parameters (>67 mill) we get **99** $= 17\text{x}1 + 1\text{x}2^1 + 2\text{x}2^2 + 3\text{x}2^3 + 3\text{x}2^4$

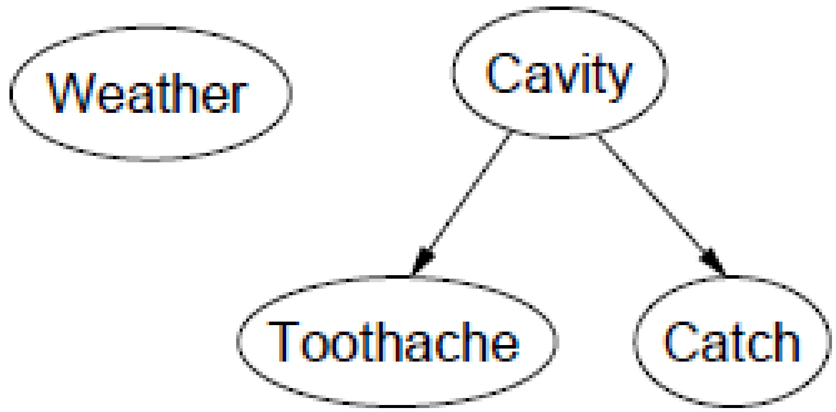
What are Bayesian Networks Useful for?

- Diagnosis: P(cause | symptom) = ?
- Prediction: P(symptom | cause) = ?
- Classification: Max class P(class | data)
- Decision-making (given a cost function)





Conditional Independence Assertions



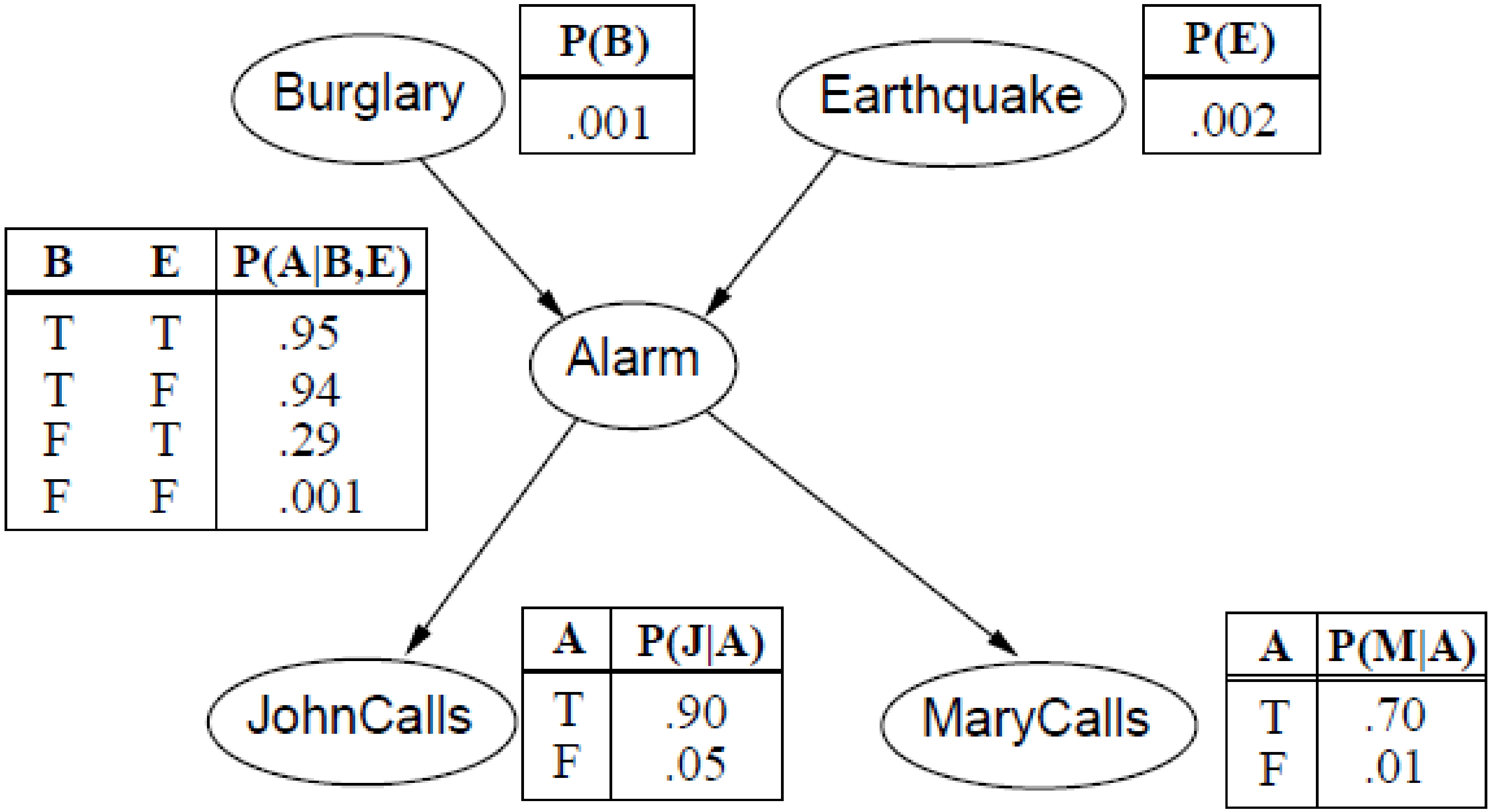
- *Weather* is independent of the other variables
- *Toothache* and *Catch* are conditionally independent given *Cavity*

Alarm 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 "casual" 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

Alarm Example



P(B)
.001

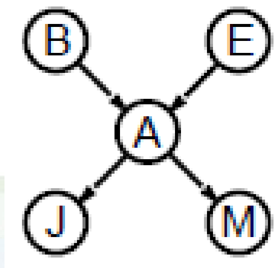
P(E)
.002

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

Alarm Example: 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 $O(n\ 2^k)$ 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. $2^5 - 1 = 31$)

Alarm Example: Global Semantics

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

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

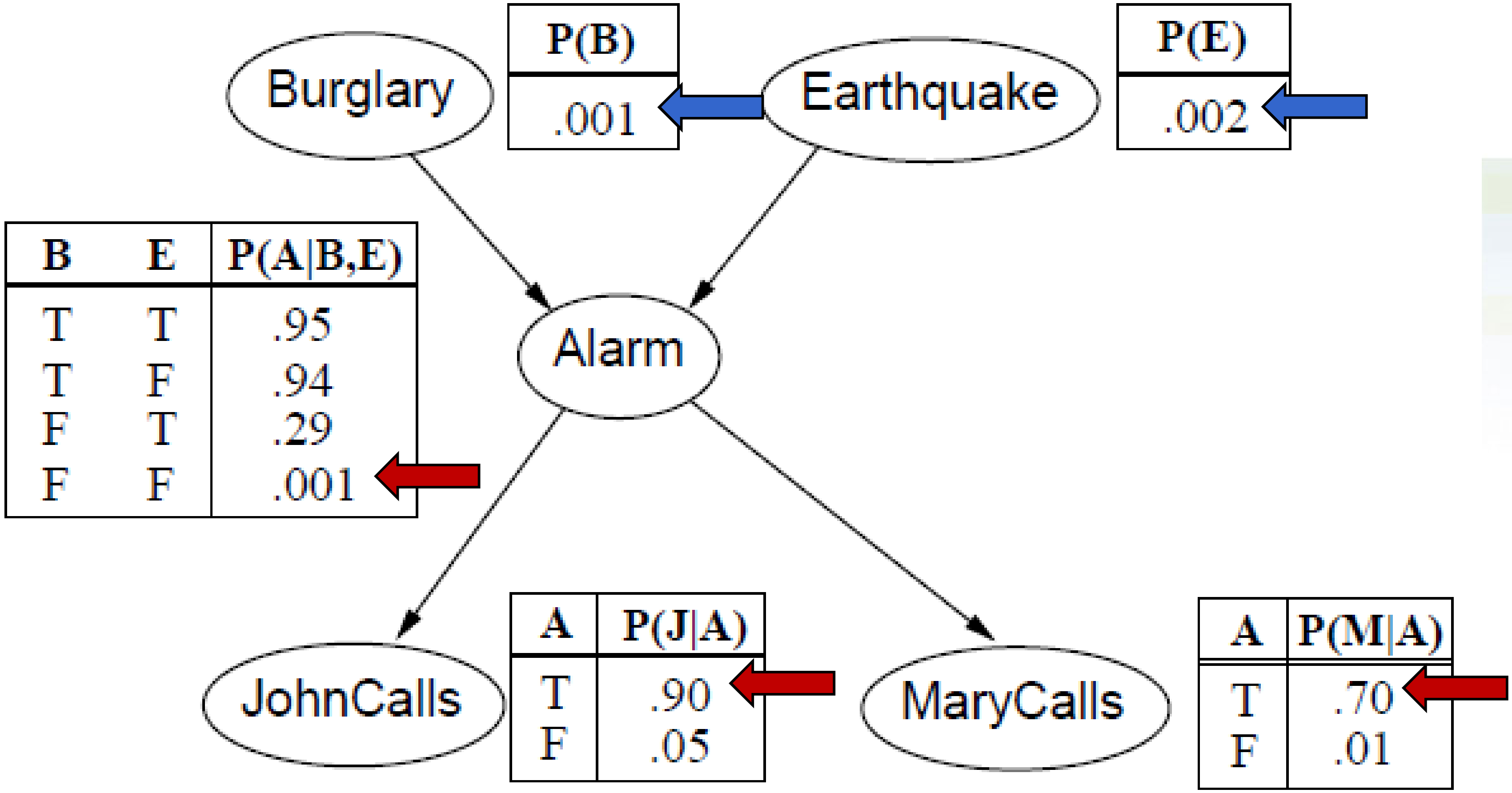
e.g., $P(j \land m \land a \land \neg b \land \neg e)$

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

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

≈ 0.00063

The product of probabilities



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

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

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

$$\approx 0.00063$$

Constructing Bayesian Networks

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

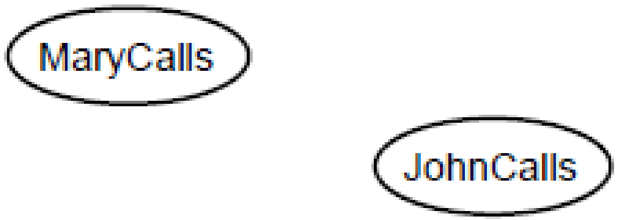
1. Choose an ordering of variables $X_1, \ldots, X_n$
2. For $i = 1$ to n
   add X_i to the network
   select parents from $X_1, \ldots, X_{i-1}$ such that
   $$\mathbf{P}(X_i|Parents(X_i)) = \mathbf{P}(X_i|X_1, \ldots, X_{i-1})$$

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

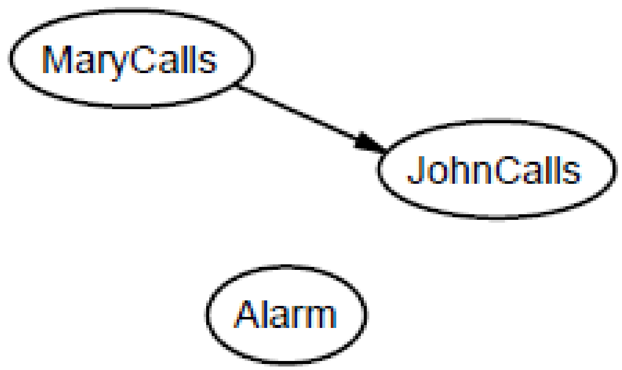
Suppose we choose the ordering M, J, A, B, E



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

Example

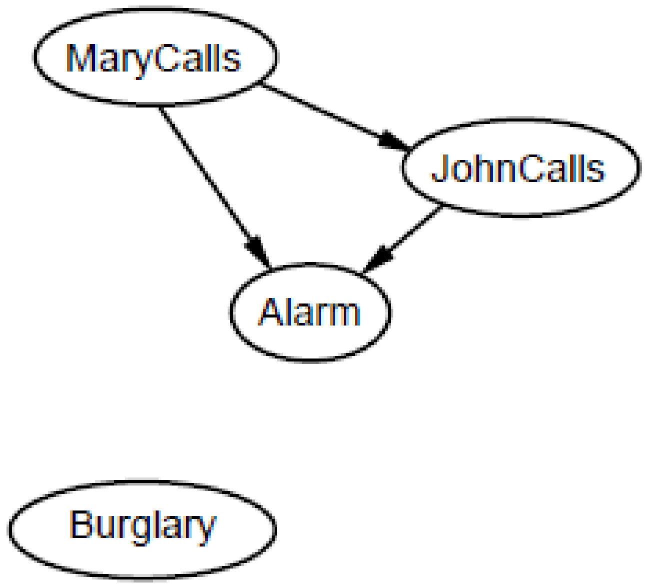
Suppose we choose the ordering M, J, A, B, E



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

Example

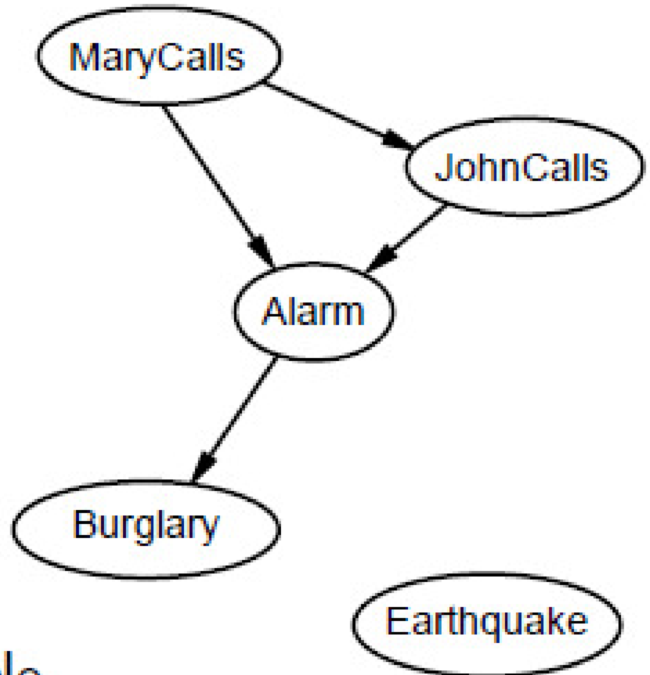
Suppose we choose the ordering M, J, A, B, E



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

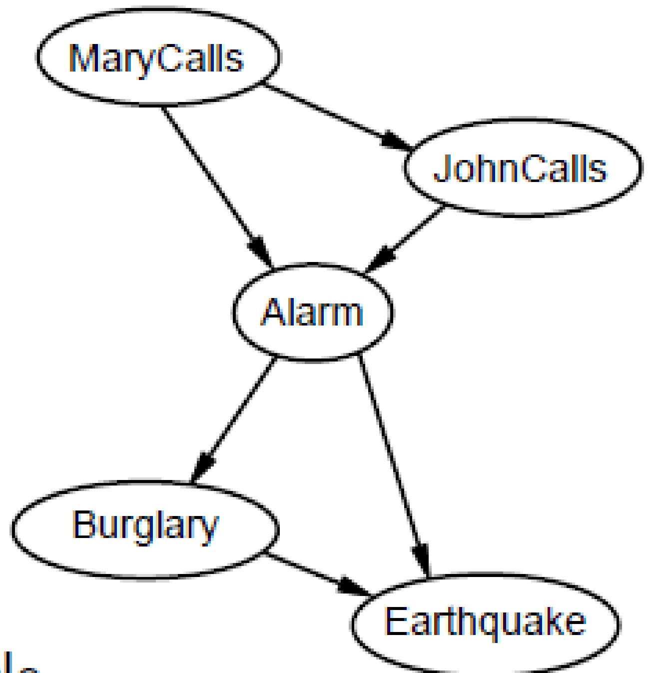
$P(B|A, J, M) = P(B)$? No

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

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

Example

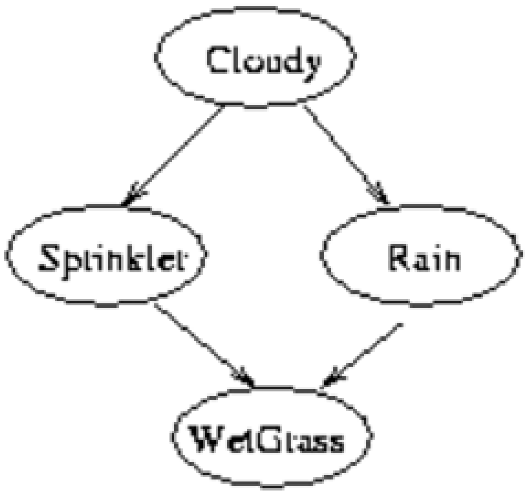
Suppose we choose the ordering M, J, A, B, E



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

Exercise

- Consider the following example, in which all nodes are binary (i.e. have two possible values) denoted T (true) and F (false)



- We see that the event "grass is wet" (W=true) has two possible causes: either the water sprinkler is on (S=true) or it is raining (R=true)

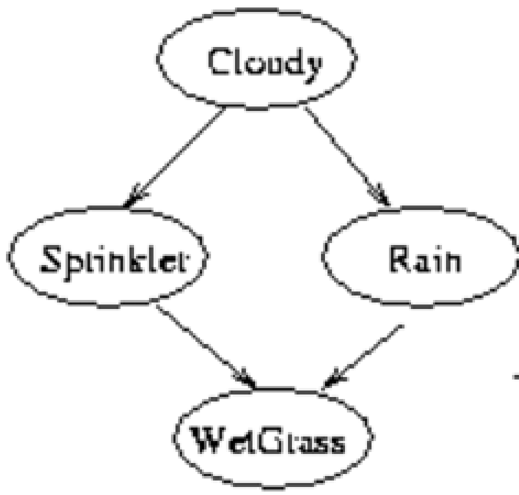
Exercise

Cloudy

P(C=F)	P(C=T)
0.5	0.5

Sprinkler

C	P(S=F)	P(S=T)
F	0.5	0.5
T	0.9	0.1



Rain

C	P(R=F)	P(R=T)
F	0.8	0.2
T	0.2	0.8

WetGrass

S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

Choose values for C, S, R, and W.

Calculate the probability: P(C, S, R, W)

(see next slide for assistance)

To think about:

P(S=T | W=T) vs. P(R=T | W=T) ??

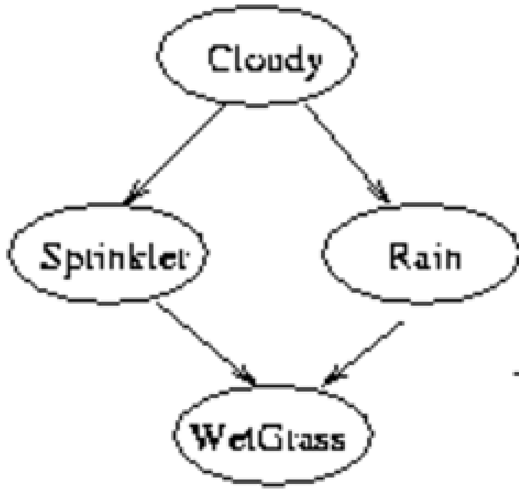
Exercise

Cloudy

P(C=F)	P(C=T)
0.5	0.5

Sprinkler

C	P(S=F)	P(S=T)
F	0.5	0.5
T	0.9	0.1



Rain

C	P(R=F)	P(R=T)
F	0.8	0.2
T	0.2	0.8

WetGrass

S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

- By Chain rule of probability, the **joint probability** of all the nodes in the graph above is:
  P(C, S, R, W) = P(C) * P(S|C) * P(R|C,S) * P(W|C,S,R)
- By using **conditional independence relationships**:
  P(C, S, R, W) = P(C) * P(S|C) * P(R|C) * P(W|S,R)

Additional Material

Some background and additional information

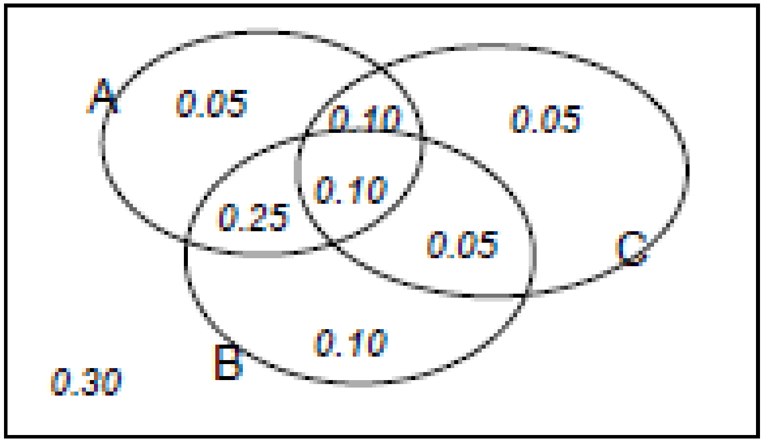
The Joint Distribution

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have 2^M rows).
2. For each combination of values, say how probable it is.
3. If you subscribe to the axioms of probability, those numbers must sum to 1.

Example: Boolean variables A, B, C

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10



Using the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Once you have the JD you can ask for the probability of any logical expression involving your attribute

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

Inference with the Joint

gender	hours_worked	wealth		
Female	v0:40.5-	poor	0.253122	
		rich	0.0246895	
	v1:40.5+	poor	0.0421768	
		rich	0.0116293	
Male	v0:40.5-	poor	0.331313	
		rich	0.0971295	
	v1:40.5+	poor	0.134106	
		rich	0.105933	

$$P(E_1 \mid E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{rows matching } E_1 \text{ and } E_2} P(\text{row})}{\sum_{\text{rows matching } E_2} P(\text{row})}$$

P(Male | Poor) = 0.4654 / 0.7604 = 0.612

Are "Bayesian networks" Bayesian?

- Despite the name, Bayesian networks do not necessarily imply a commitment to Bayesian statistics
- Rather, they are so called because they use Bayes' rule for probabilistic inference (explained next)
- Nevetherless, Bayes nets are a useful representation for hierarchical Bayesian models, which form the foundation of applied Bayesian statistics
- In such a model, the parameters are treated like any other random variable, and becomes nodes in the graph

Inference in Bayesian Networks

The most common task we wish to solve using Bayesian networks is probabilistic inference. For example, consider the water sprinkler network, and suppose we observe the fact that the grass is wet. There are two possible causes for this: either it is raining, or the sprinkler is on. Which is more likely?

We can use Bayes' rule to compute the posterior probability of each explanation (where 0==false and 1==true).

Inference in Bayesian Networks

$$\Pr(S=1|W=1) = \frac{\Pr(S=1,W=1)}{\Pr(W=1)} = \frac{\sum_{c,r}\Pr(C=c,S=1,R=r,W=1)}{\Pr(W=1)} = 0.2781/0.6471 = 0.430$$

$$\Pr(R=1|W=1) = \frac{\Pr(R=1,W=1)}{\Pr(W=1)} = \frac{\sum_{c,s}\Pr(C=c,S=s,R=1,W=1)}{\Pr(W=1)} = 0.4581/0.6471 = 0.708$$

$$\Pr(W=1) = \sum_{c,r,s}\Pr(C=c,S=s,R=r,W=1) = 0.6471$$

- Where Pr(W=1) is a normalizing constant, equal to the probability (*likelihood*) of the data. So we see that it is more likely that the grass is wet because it is raining: the likelihood ratio is 0.7079/0.4298 = 1.647