Bayesian Classifiers

CS510

Lecture #18

4/22/13

Colorado State University

Where are we?

- Learning the basics of classifiers
  - Goal: become an intelligent user
  - SVMs : Linear classifiers
    - Learn more in CS548
  - Feedforward Networks : Non-linear classifiers
    - Learn more in CS545
  - Today: Bayesian classifiers
    - Learn more in CS440

Review: Probability Basics

- Let X be variable whose value takes on a discrete set of labels, $X = \{x_1, \ldots, x_n\}$
- The $P(X=x_i)$ is a probability function if:

$$\forall x_i : 0 \leq p\left(X = x_i\right) \leq 1$$

$$\sum_i p\left(X = x_i\right) = 1$$

- We abbreviate $P(X=x_i)$ as $P(x_i)$

Review: Probability Basics (II)

- Probabilities reflect the likelihood that a statement $X=x_i$ is true
  - If the statement is true, $P(x_i) = 1$
  - If it is false, $P(x_i) = 0$
  - Otherwise, higher values indicate more likely
- *Frequentist* probabilities represent samplings of random draws
- Subjective probabilities may not

Bayes Rule

- Given $X = \{x_1, \ldots, x_n\}, Y = \{y_1, \ldots, y_m\}$

$$P(x_i \wedge y_j) = P(x_i \mid y_j)P(y_j) = P(y_j \mid x_i)P(x_i)$$

- Or, put another way,

$$P(x_i \mid y_j) = \frac{P(y_j \mid x_i)P(x_i)}{P(y_j)}$$

Bayesian Classification Example

- Let's say you extract a circle from an image, and want to know if it's a tire?

$$\text{Object(O)} = \{Wheel, handlebar, road, dirt, \ldots\}$$

$$\text{Feature(F)} = \{Circle, \neg Circle\}$$

$$P(Wheel \mid Circle\} = \frac{P(Circle \mid Wheel)P(Wheel)}{P(Circle)}$$

- Note that P(Circle|Wheel) is easier to estimate than P(Wheel|Circle)

Naïve Bayes Classifiers

- Assume that features are independent, so

$$P\left(x \mid f_1,\ldots,f_m\right) = P\left(x \mid f_1\right)P\left(x \mid f_2\right)\ldots P\left(x \mid f_m\right)$$

- And of course

$$P\left(x \mid f_i\right) = \frac{P\left(f_i \mid x\right)P\left(x\right)}{P\left(f_i\right)}$$

- So

$$P\left(x \mid f_1,\ldots,f_m\right) = P\left(f_1 \mid x\right)\ldots P\left(f_m \mid x\right)\frac{P\left(x\right)^m}{\Pi P\left(f_i\right)}$$

Naïve Bayes Classifiers (II)

- The fractional term in the last equation is constant for all x_i.
- So the most likely x is the one that maximizes

$$P\left(f_1 \mid x\right) \ldots P\left(f_m \mid x\right)$$

- You can recover the true probabilities by normalizing for all x_i
- Works well with PCA (where features are approximately independent)

Conditional Independence

- Unfortunately, most random variables of interest are not independent
- A more useful notion is *conditional independence*
- Two variables X and Y are conditionally independent given Z if
  - $P(X = x|Y = y, Z=z) = P(X = x|Z=z)$ for all values x,y,z
  - That is, learning the values of Y does not change prediction of X once we know the value of Z
  - Notation: $I(X ; Y | Z)$

Conditionally Chained Inference

- Independence is a strong assumption
- Often, x depends on multiple features that are not independent of each other
- If the features can be chained so that each depends only on previous features other…

$$P(x \mid f_1, f_2, f_3) = P(x \mid f_1) P(f_1 \mid f_2, f_3) P(f_2, f_3)$$
$$= P(x \mid f_1) P(f_1 \mid f_2) P(f_2 \mid f_3) P(f_3)$$

Purpose of Bayesian Networks

- Facilitate the description of a collection of beliefs by
  - making causality relations explicit
  - exploiting conditional independence
- Provide efficient methods for:
  - Representing a joint probability distribution
  - Updating belief strengths when new evidence is observed

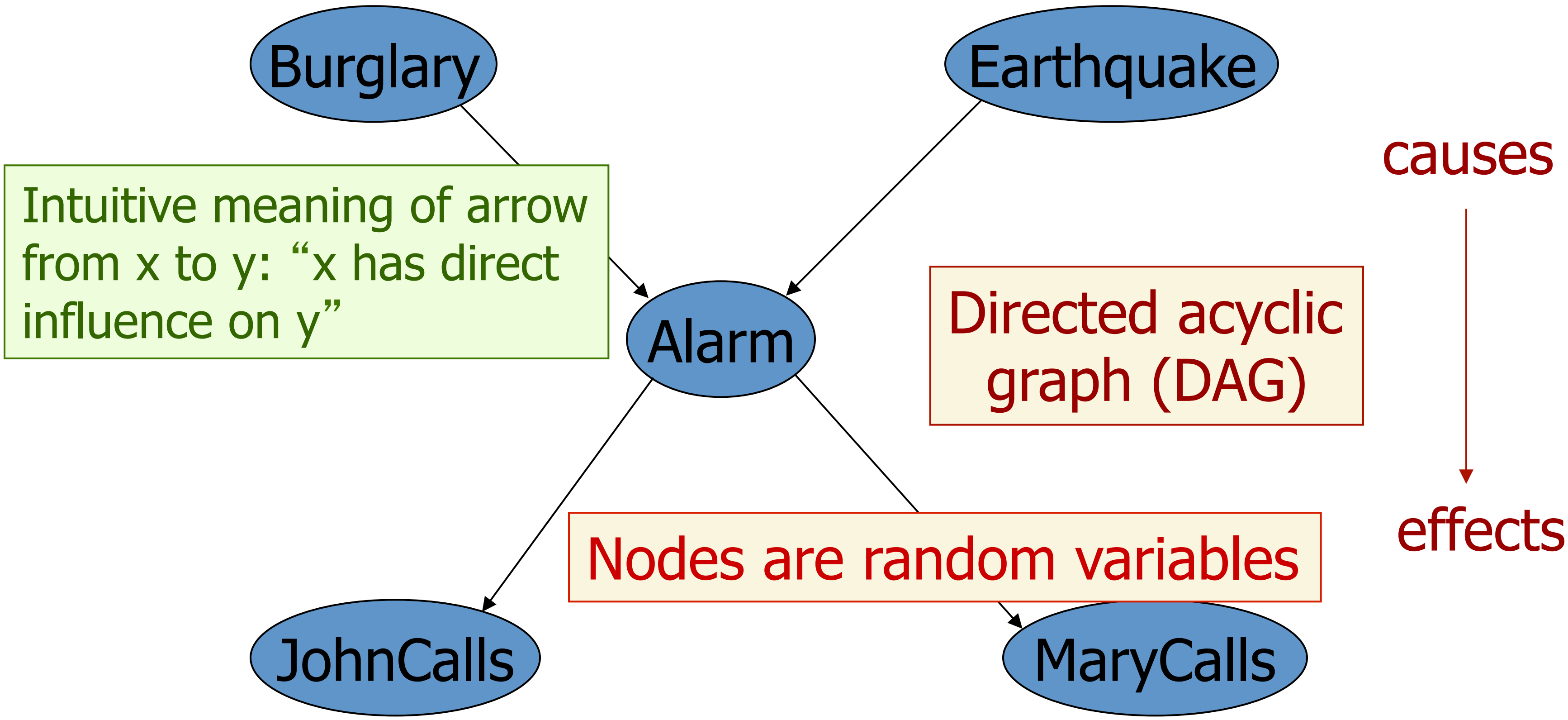
Example

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

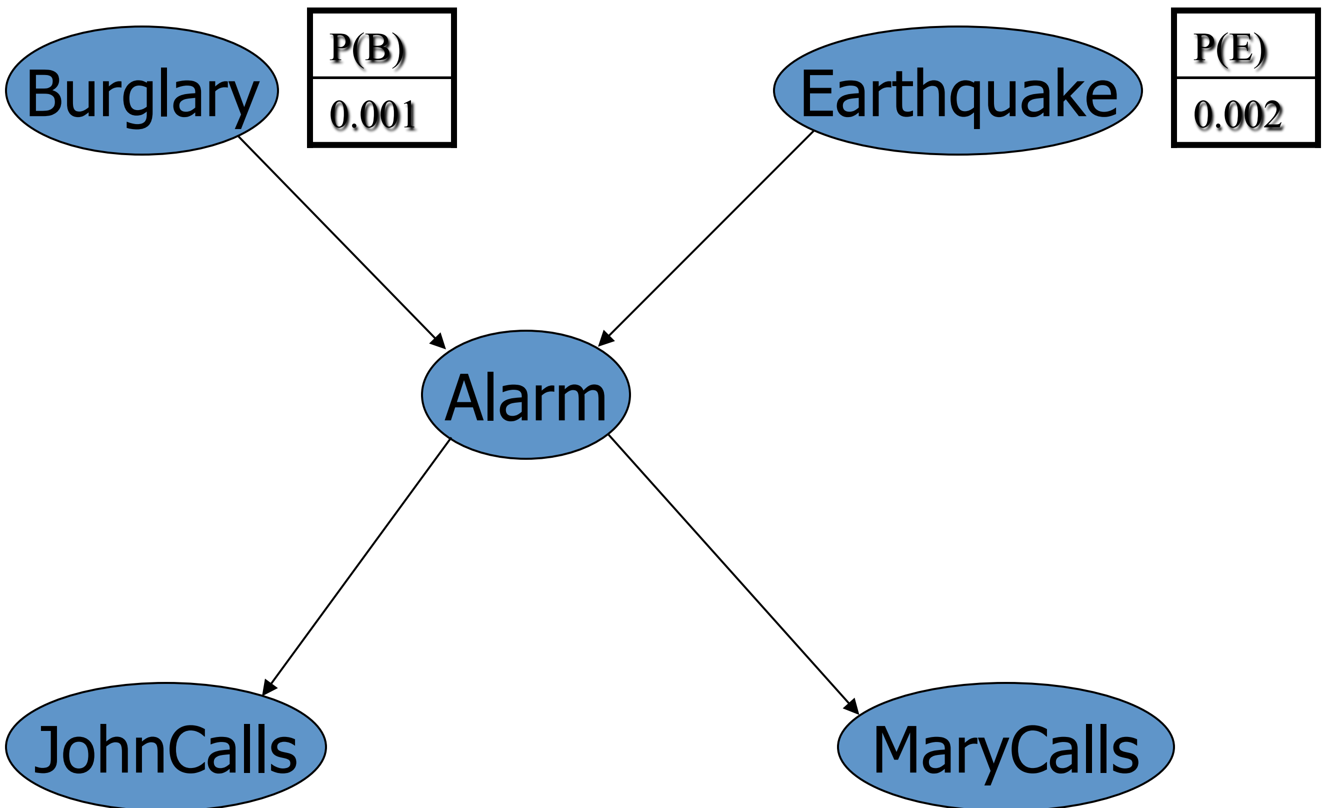
Variables: Burglary, Earthquake, Alarm, JohnCalls, MaryCalls

Network topology reflects "causal" knowledge:
- A burglary 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

A Simple Network

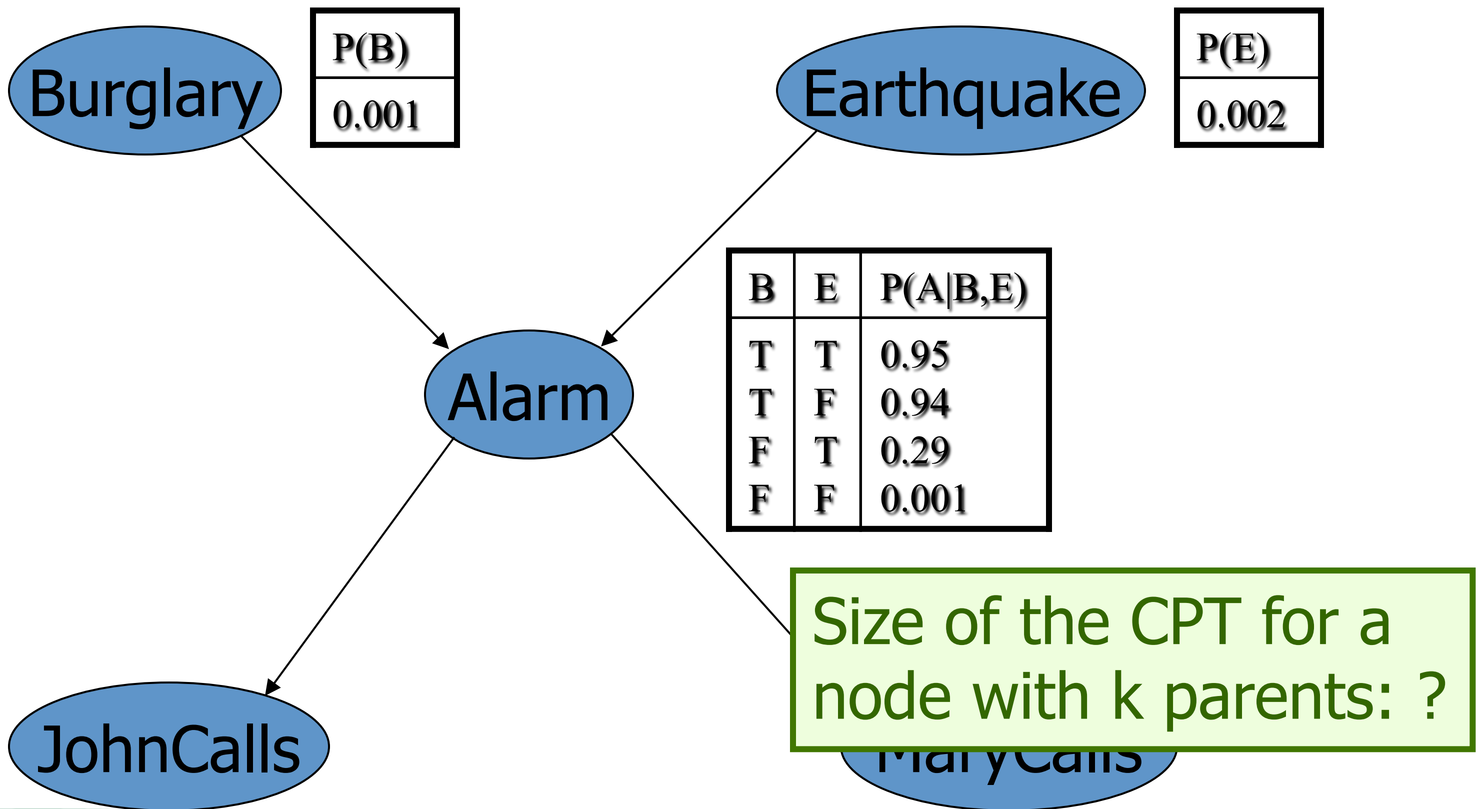


Assigning Probabilities to Roots



Colorado State University

Conditional Probability Tables



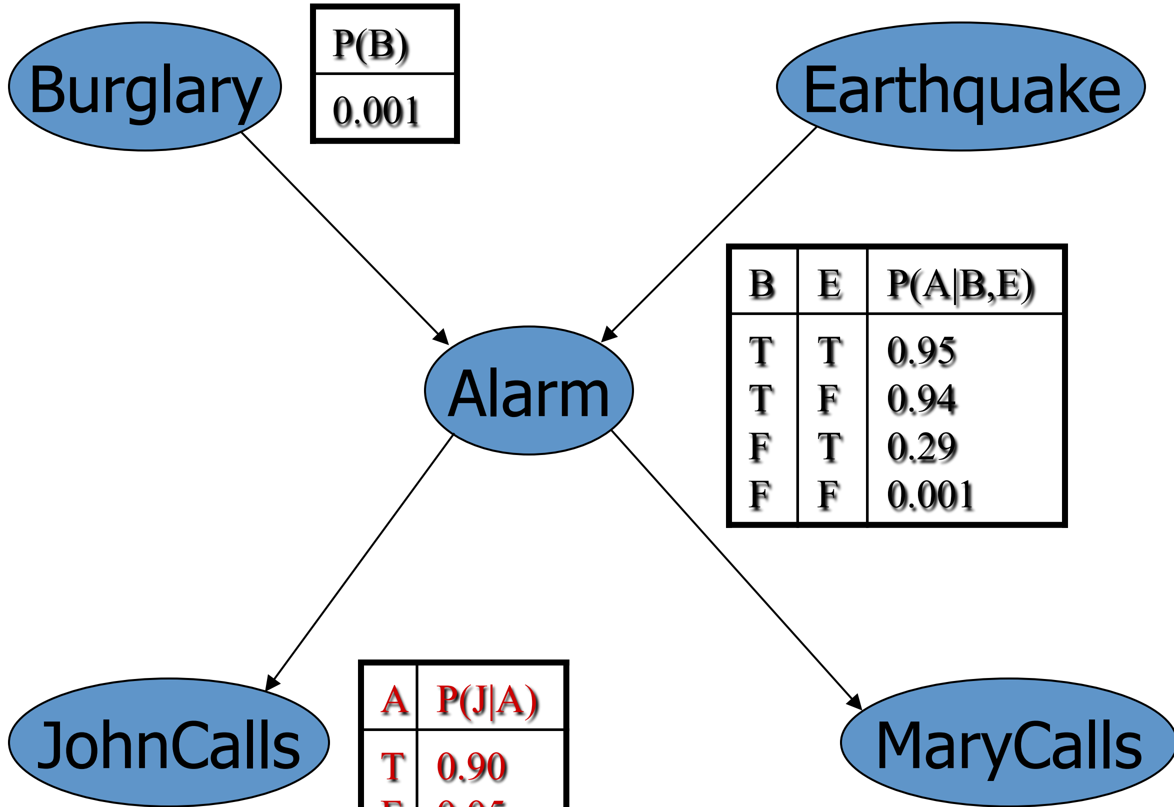
P(B)
0.001

P(E)
0.002

B	E	P(A|B,E)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

Size of the CPT for a node with k parents: ?

Conditional Probability Tables



P(B)
0.001

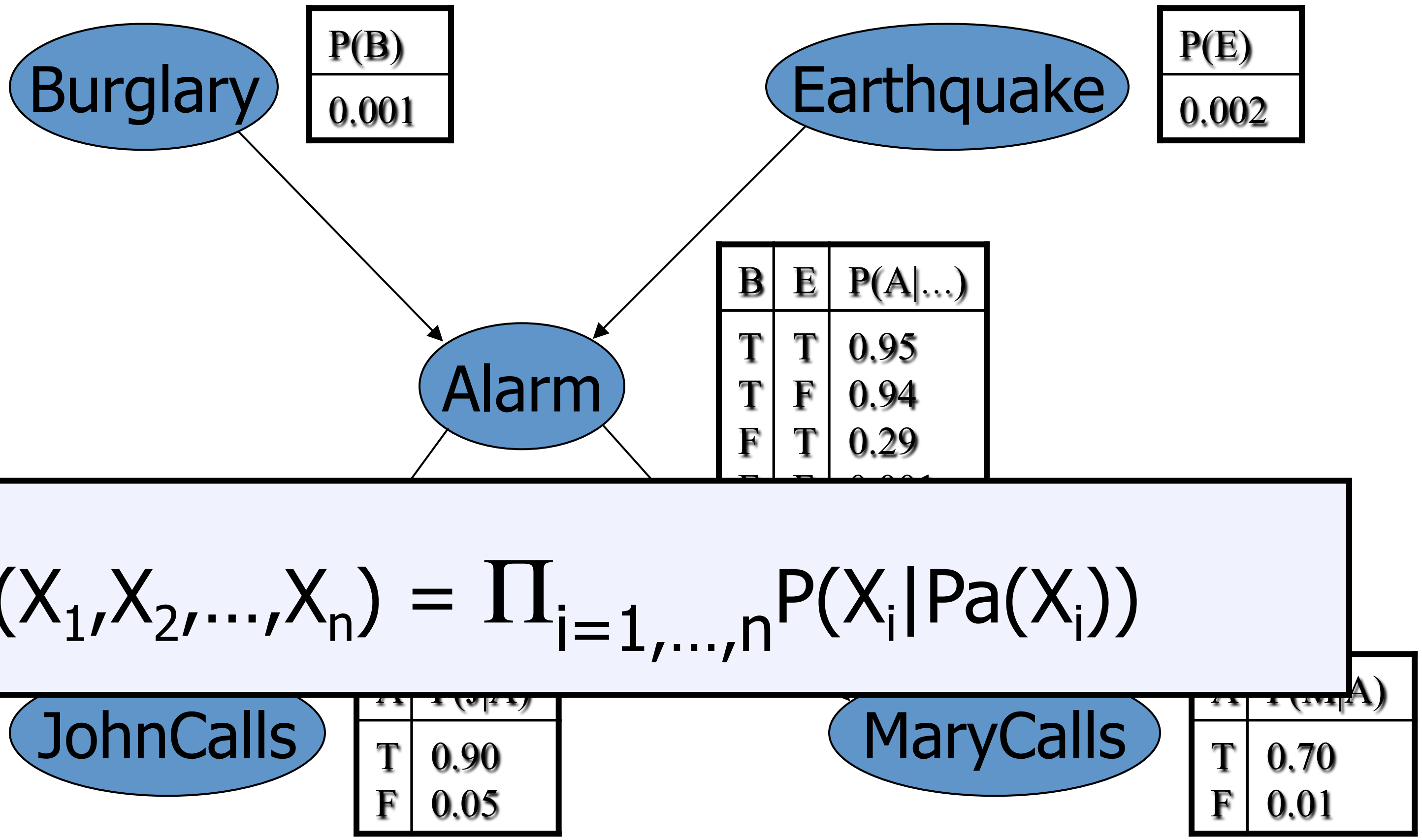
P(E)
0.002

B	E	P(A|B,E)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

A	P(J|A)
T	0.90
F	0.05

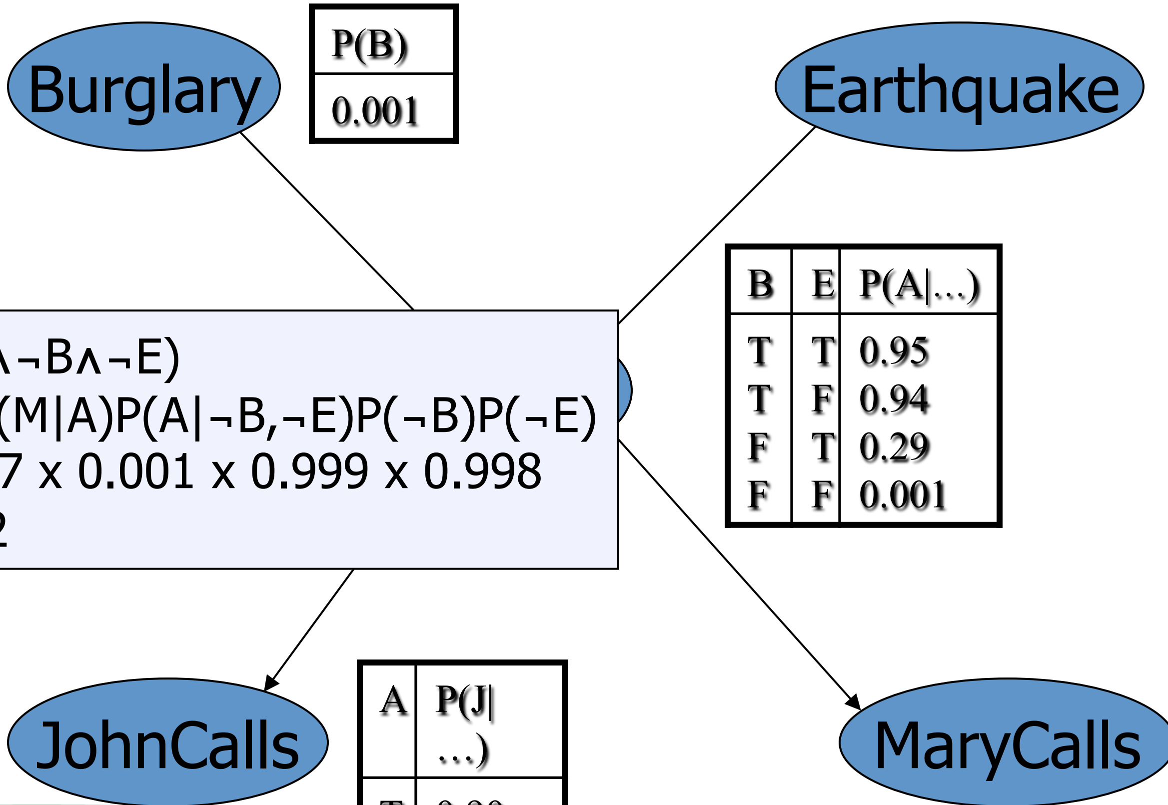
A	P(M|A)
T	0.70
F	0.01

What the BN Means



$$P(X_1,X_2,\ldots,X_n) = \Pi_{i=1,\ldots,n} P(X_i|Pa(X_i))$$

Calculation of Joint Probability



P(B)
0.001

P(E)
0.002

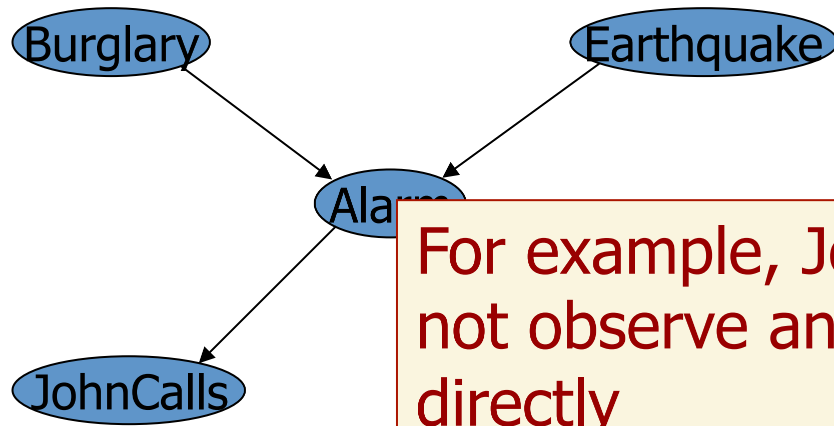
B	E	P(A|…)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

A	P(J|…)
T	0.90
F	0.05

A	P(M|…)
T	0.70
F	0.01

$P(J \wedge M \wedge A \wedge \neg B \wedge \neg E)$
$= P(J|A)P(M|A)P(A|\neg B,\neg E)P(\neg B)P(\neg E)$
$= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998$
$= 0.00062$

What the BN Encodes



For example, John does not observe any burglaries directly

- Each of the beliefs JohnCalls and MaryCalls is independent of Burglary and Earthquake given Alarm or ¬Alarm

- The beliefs JohnCalls and MaryCalls are independent given Alarm or ¬Alarm

Independence

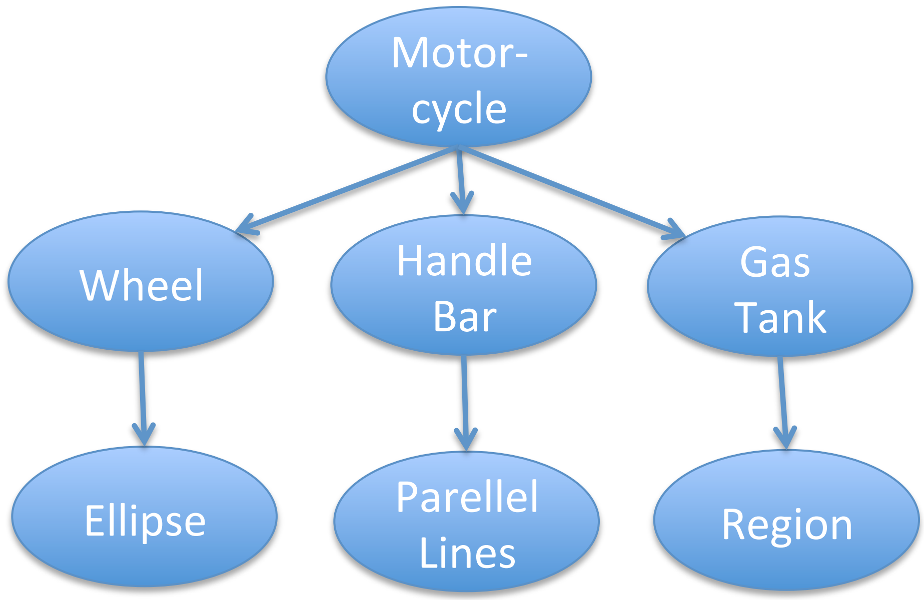
- The expression:

  $$P(X_1,X_2,\ldots,X_n) = \Pi_{i=1,\ldots,n} P(X_i|Pa(X_i))$$

  means that each belief is independent of its predecessors in the BN given its parents
- Said otherwise, the parents of a belief X_i are all the beliefs that "directly influence" X_i
- Usually (but not always) the parents of X_i are its causes and X_i is the effect of these causes

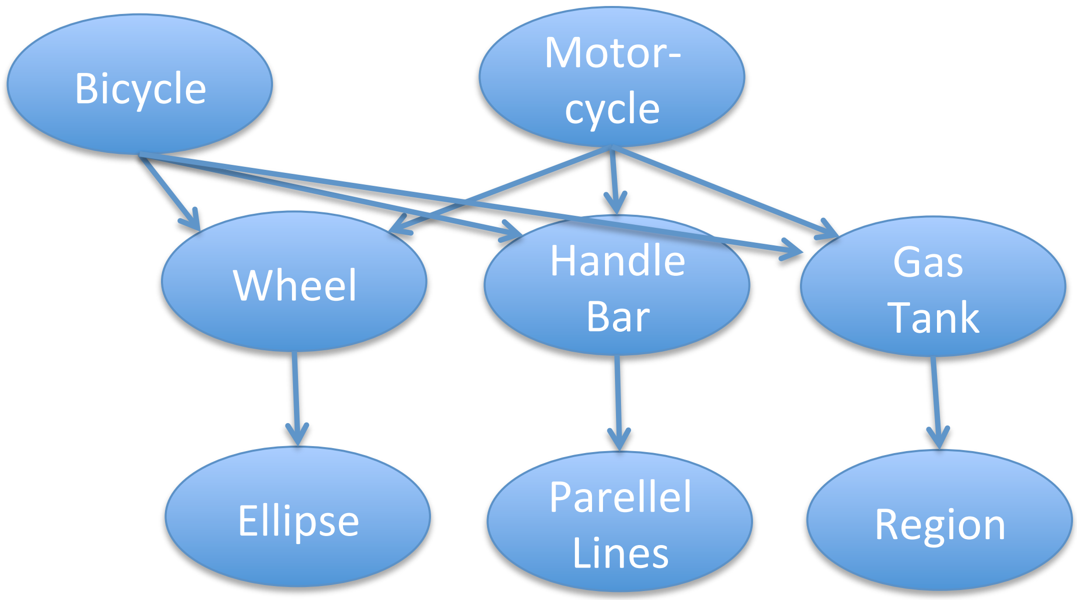
E.g., JohnCalls is influenced by Burglary, but not directly. JohnCalls is directly influenced by Alarm

Computer Vision Example



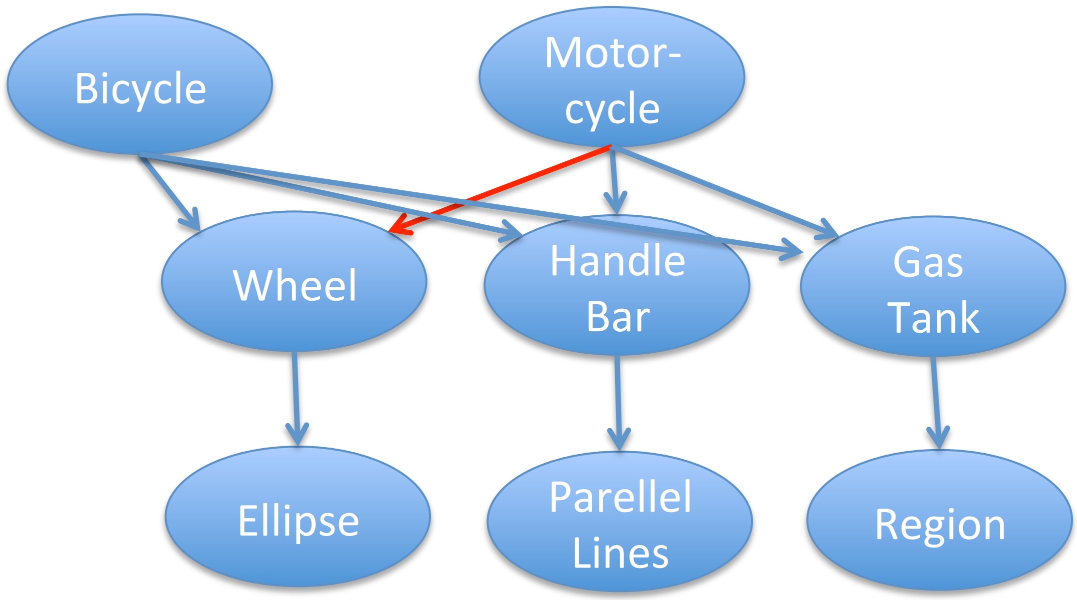
One purpose for Bayesian nets is to separate object variation information from feature extraction probabilities

Computer Vision Example (II)



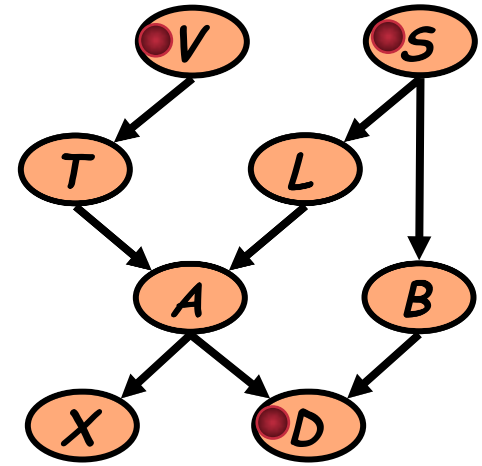
Another is to exploit shared subpart. Bicycles have wheels and handle bars too, but not gas tanks, so P(Gas Tank | Bicycle) ≈ 0.

Computer Vision Example (III)



Finally, in the constellation model the probability functions can depend on location as well as existence

Solving for Probabilities



- How do we deal with evidence?

- Suppose get evidence $V = t, S = f, D = t$ and want to compute $P(L|V = t, S = f, D = t)$

$$P(L \mid V = t, S = f, D = t) = \frac{P(L, V = t, S = f, D = t)}{P(V = t, S = f, D = t)}$$

Variable Elimination

- There is an algorithm for this
  - It's called variable elimination
  - Efficient for trees; inefficient for DAGs
  - Its taught in CS440 (see Russell & Norvig)
- As a user, you need to know:
  - VE can compute the likelihood of any node in a Bayesian net, given any set of evidence
  - But, computing P(X=x) is NP-hard (in general)

Approaches to inference

- Exact inference
  - Variable elimination
  - Join tree algorithm (also NP)
- Approximate inference
  - Simplify the structure of the network to make exact inference efficient (variational methods, loopy belief propagation)
- Probabilistic methods
  - Stochastic simulation / sampling methods
  - Markov chain Monte Carlo methods