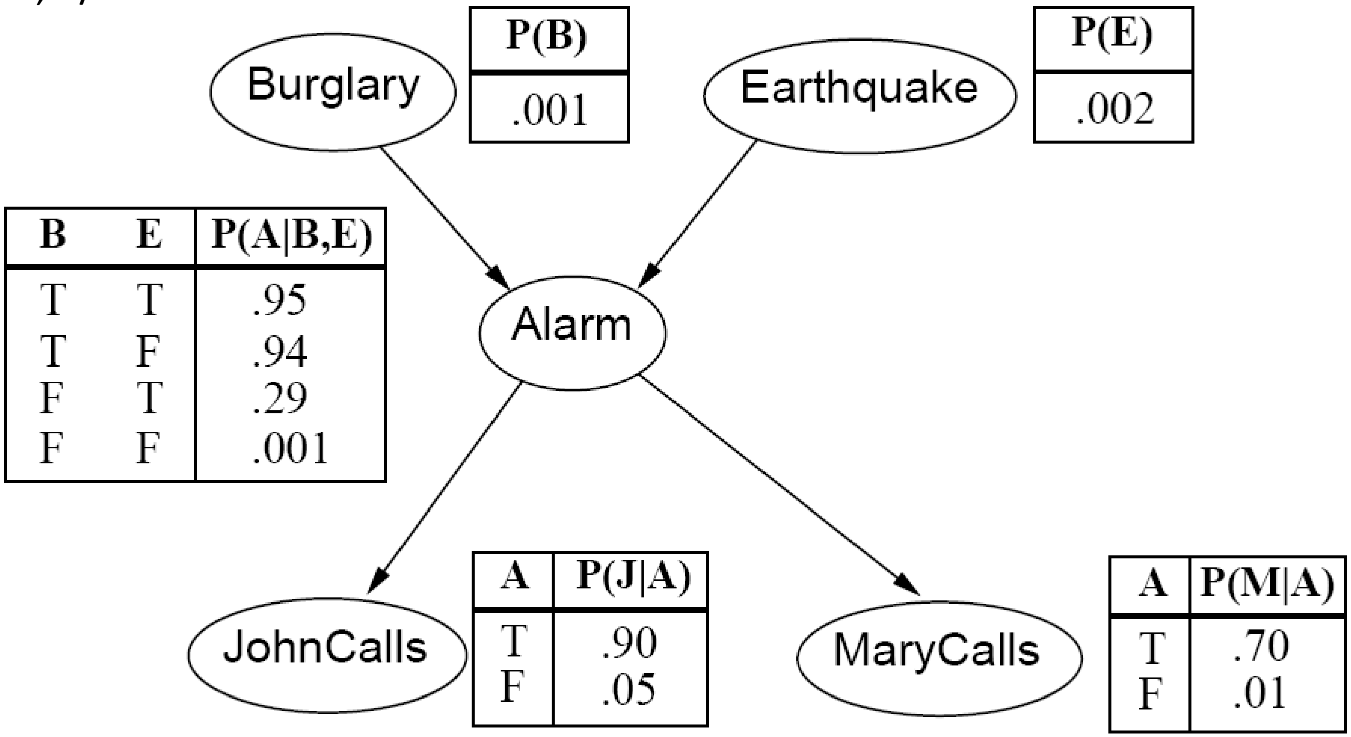
CS440/ECE448 Lecture 17: Bayesian Networks

Slides by Svetlana Lazebnik, 10/2016

Modified by Mark Hasegawa-Johnson, 3/2018



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

Review: Bayesian inference

- A general scenario:
  - Query *variables:* **X**
  - *Evidence* (*observed*) variables and their values: **E** = **e**
- **Inference problem**: answer questions about the query variables given the evidence variables
- This can be done using the posterior distribution P(**X** | **E** = **e**)
- Example of a useful question: **Which X is true?**
  - More formally: what value of **X** has the least probability of being wrong?
  - Answer: **MPE = MAP** (argmin P(error) = argmax P(X=x|E=e))

Today: What if P(X,E) is complicated?

- Very, very common problem: P(X,E) is complicated because both X and E depend on some hidden variable Y
- SOLUTION:
  - Draw a bunch of circles and arrows that represent the dependence
  - When your algorithm performs inference, make sure it does so in the order of the graph
- FORMALISM: Bayesian Network

Hidden Variables

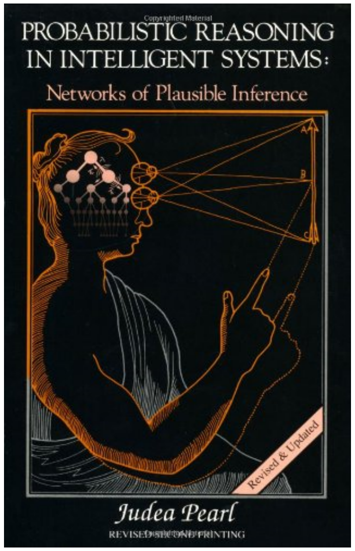
- A general scenario:
  - Query *variables:* **X**
  - *Evidence* (*observed*) variables and their values: **E** = **e**
  - *Unobserved* variables: **Y**
- **Inference problem**: answer questions about the query variables given the evidence variables
  - This can be done using the posterior distribution P(**X** | **E** = **e**)
  - In turn, the posterior needs to be derived from the full joint P(**X**, **E**, **Y**)

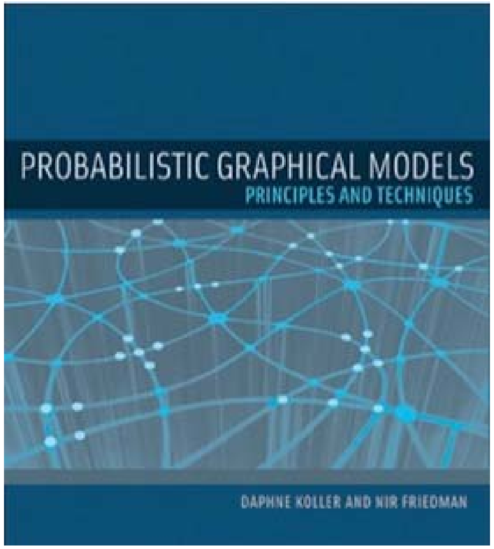
$$P(\boldsymbol{X} \mid \boldsymbol{E} = \boldsymbol{e}) = \frac{P(\boldsymbol{X}, \boldsymbol{e})}{P(\boldsymbol{e})} \propto \sum\nolimits_{\boldsymbol{y}} P(\boldsymbol{X}, \boldsymbol{e}, \boldsymbol{y})$$

- Bayesian networks are a tool for representing joint probability distributions efficiently

Bayesian networks

- More commonly called *graphical models*
- A way to depict conditional independence relationships between random variables
- A compact specification of full joint distributions

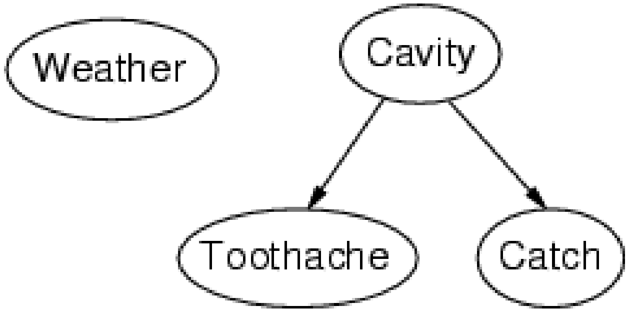



PROBABILISTIC GRAPHICAL MODELS
PRINCIPLES AND TECHNIQUES
DAPHNE KOLLER AND NIR FRIEDMAN

Outline

- Review: Bayesian inference
- Bayesian network: graph semantics
- The Los Angeles burglar alarm example
- Conditional independence ≠ Independence
- Constructing a Bayesian network: Structure learning
- Constructing a Bayesian network: Hire an expert

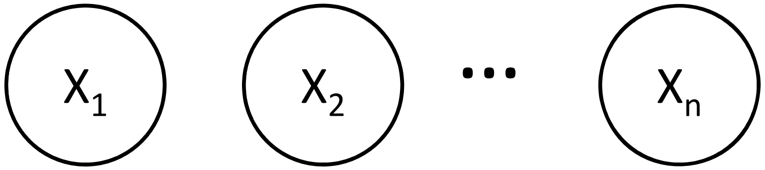
Bayesian networks: Structure



- **Nodes:** random variables
- **Arcs:** interactions
  - An arrow from one variable to another indicates direct influence
  - Must form a directed, *acyclic* graph

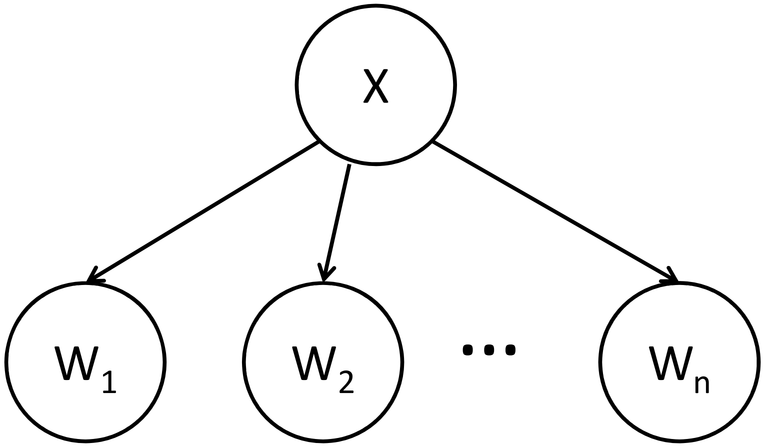
Example: N independent coin flips

- Complete independence: no interactions



Example: Naïve Bayes document model

- Random variables:
  - X: document class
  - $W_1, \ldots, W_n$: words in the document



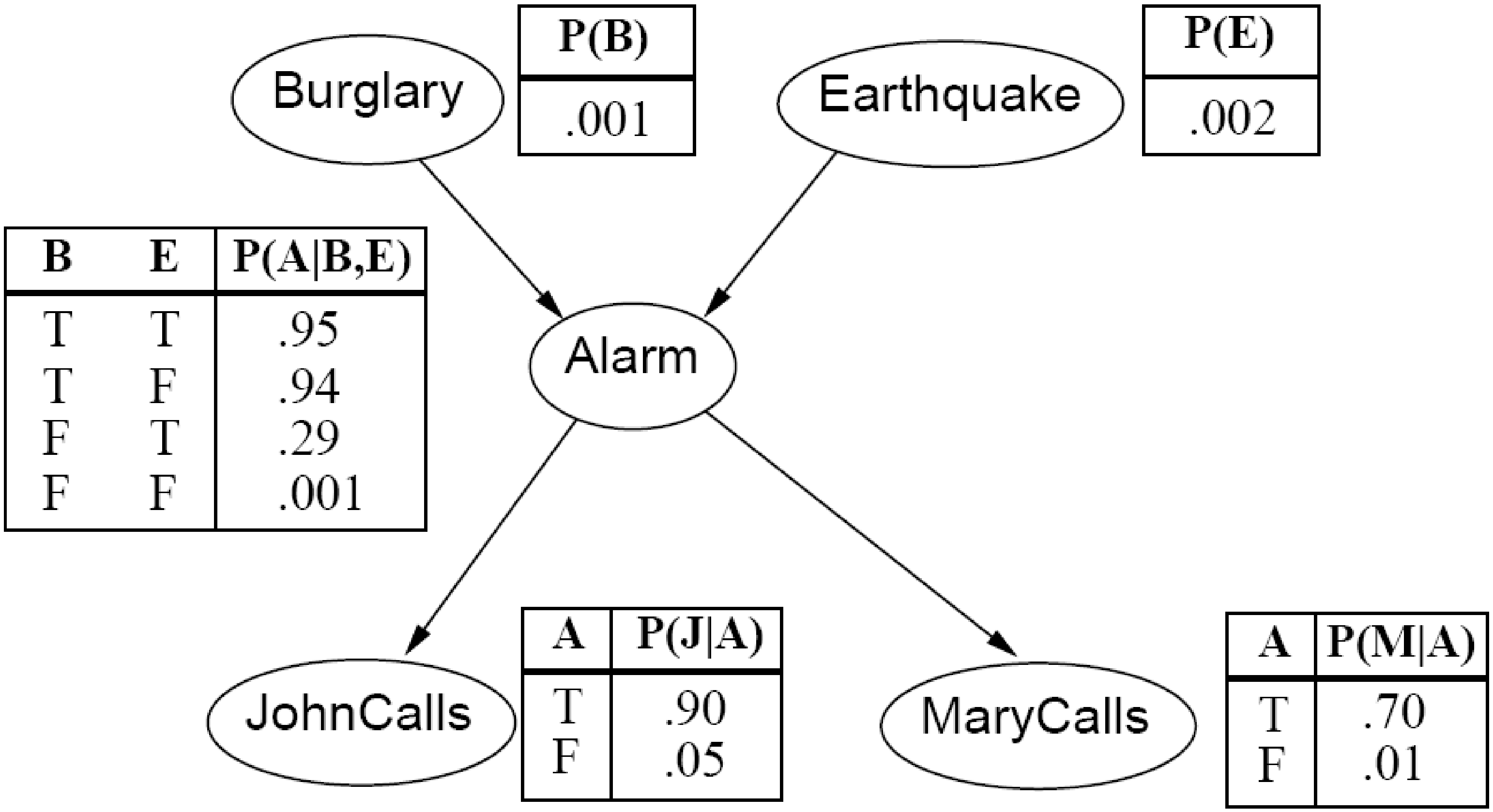
Outline

- Review: Bayesian inference
- Bayesian network: graph semantics
- The Los Angeles burglar alarm example
- Conditional independence ≠ Independence
- Constructing a Bayesian network: Structure learning
- Constructing a Bayesian network: Hire an expert

Example: Los Angeles Burglar Alarm

- I have a burglar alarm that is sometimes set off by minor earthquakes. My two neighbors, John and Mary, promised to call me at work if they hear the alarm
  - Example inference task: suppose Mary calls and John doesn’t call. What is the probability of a burglary?
- What are the random variables?
  - *Burglary, Earthquake, Alarm, JohnCalls, MaryCalls*
- What are the direct influence relationships?
  - 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: Burglar Alarm



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

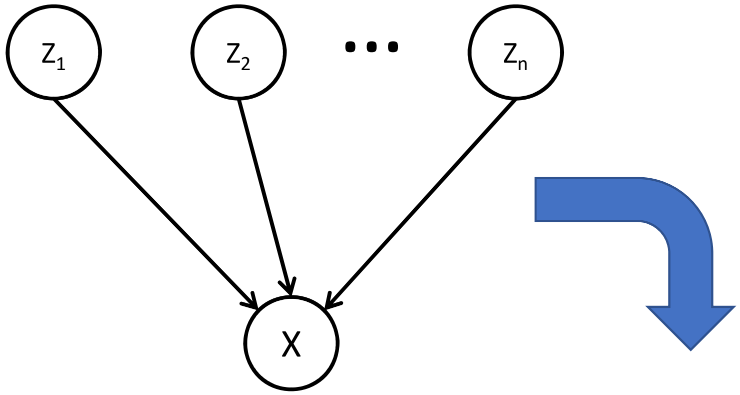
Conditional independence and the joint distribution

- Key property: each node is conditionally independent of its *non-descendants* given its *parents*
- Suppose the nodes $X_1, \ldots, X_n$ are sorted in topological order
- To get the joint distribution $P(X_1, \ldots, X_n)$, use chain rule:

$$
\begin{aligned}
P(X_1, \ldots, X_n) &= \prod_{i=1}^{n} P\left(X_i \mid X_1, \ldots, X_{i-1}\right) \\
&= \prod_{i=1}^{n} P\left(X_i \mid Parents(X_i)\right)
\end{aligned}
$$

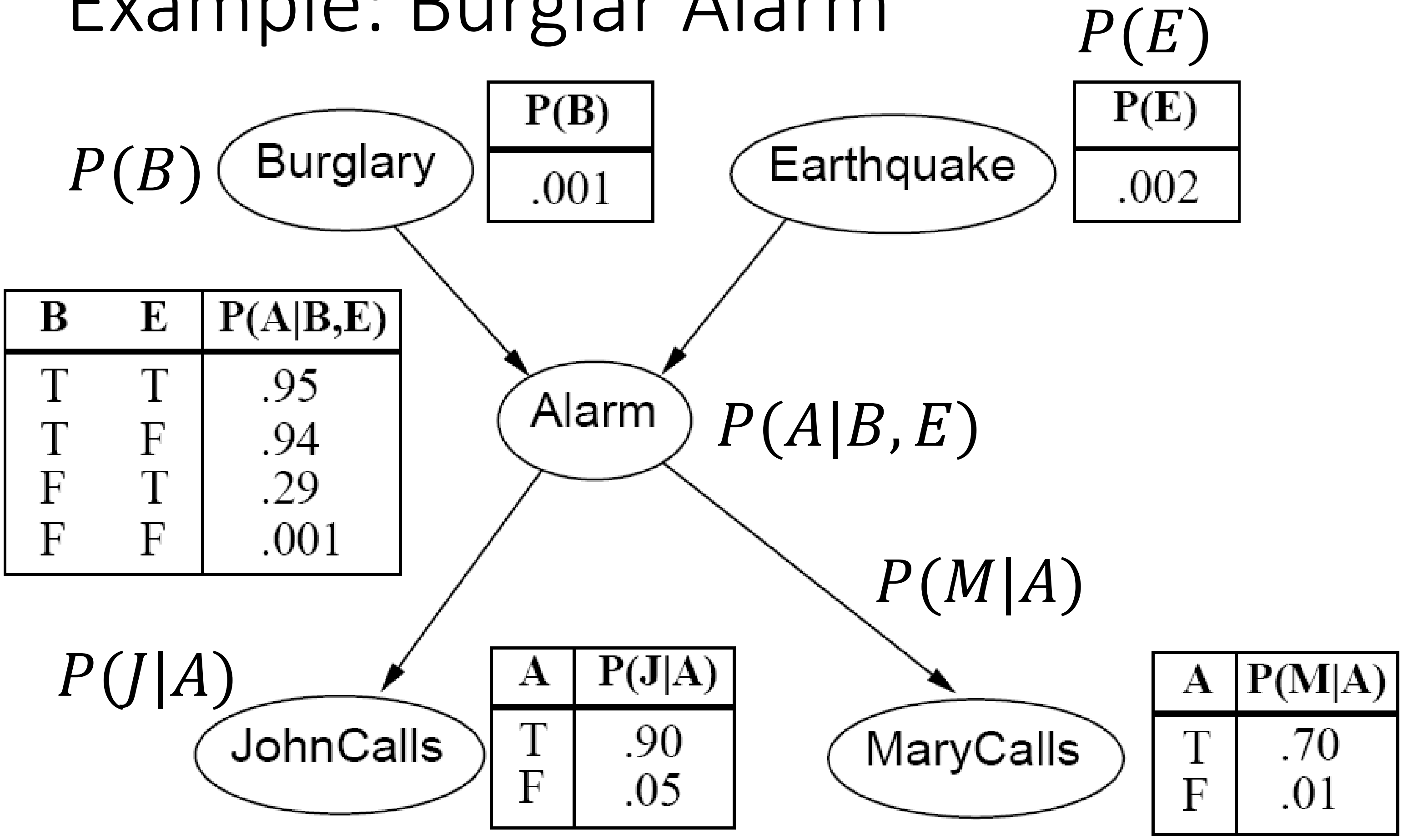
Conditional probability distributions

- To specify the full joint distribution, we need to specify a *conditional* distribution for each node given its parents: P (X | Parents(X))



P (X | Z_1, …, Z_n)

Example: Burglar Alarm



$P(B)$

P(B)
.001

$P(E)$

P(E)
.002

$P(A|B,E)$

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

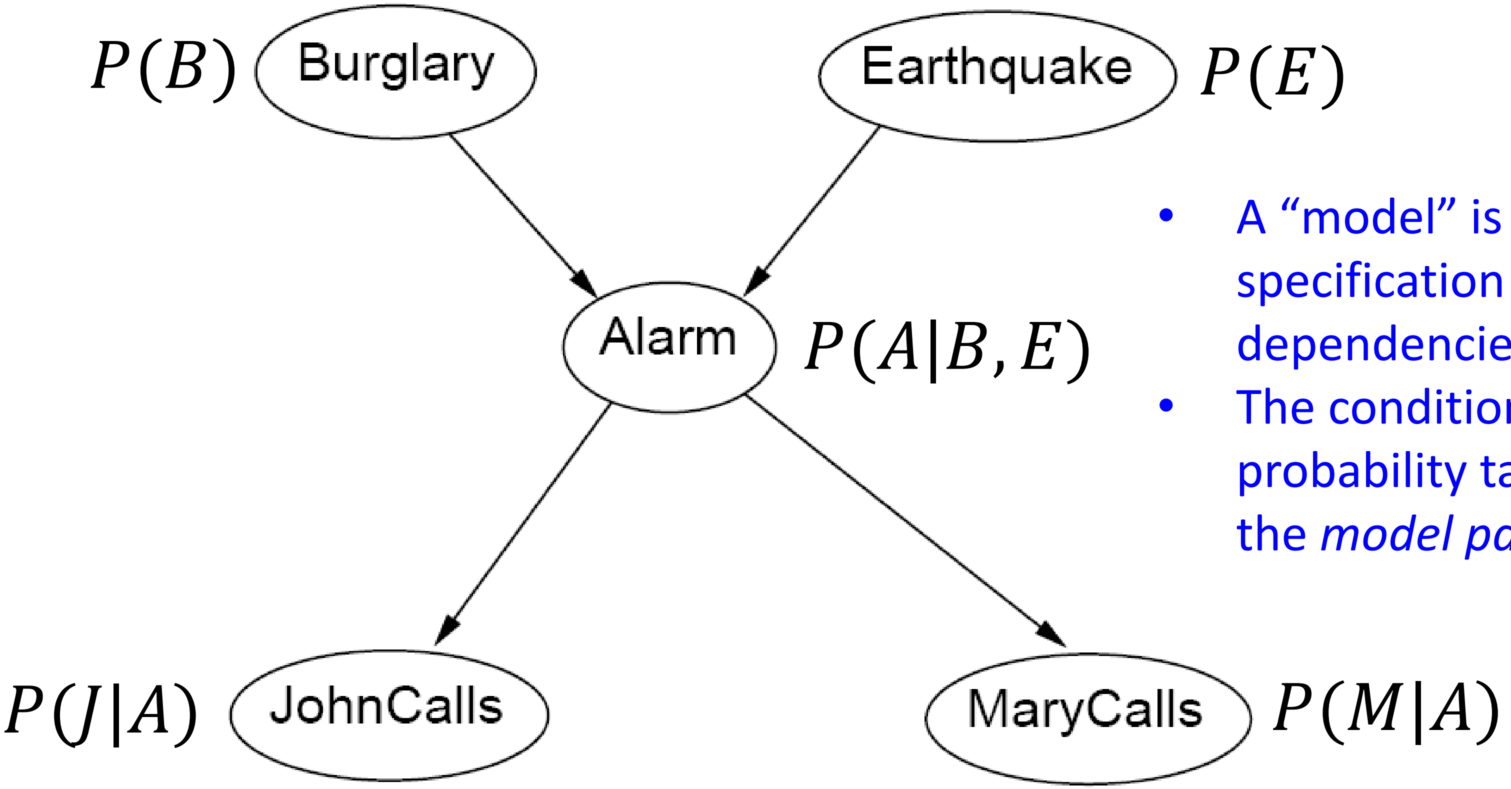
$P(J|A)$

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

$P(M|A)$

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

Example: Burglar Alarm



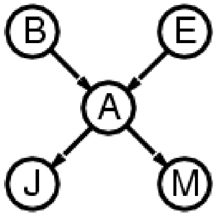
- A “model” is a complete specification of the dependencies.
- The conditional probability tables are the *model parameters.*

Outline

- Review: Bayesian inference
- Bayesian network: graph semantics
- The Los Angeles burglar alarm example
- Conditional independence ≠ Independence
- Constructing a Bayesian network: Structure learning
- Constructing a Bayesian network: Hire an expert

The joint probability distribution

$$\boxed{P(X_1,\mathrm{K}\ ,X_n) = \prod_{i=1}^{n} P\big(X_i \mid Parents(X_i)\big)}$$



For example,

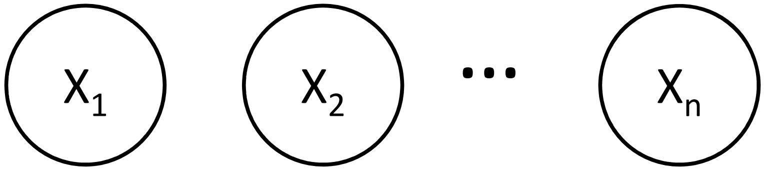
$$P(j, m, a, \neg b, \neg e) = P(\neg b)\, P(\neg e)\, P(a|\neg b, \neg e)\, P(j|a)\, P(m|a)$$

Independence

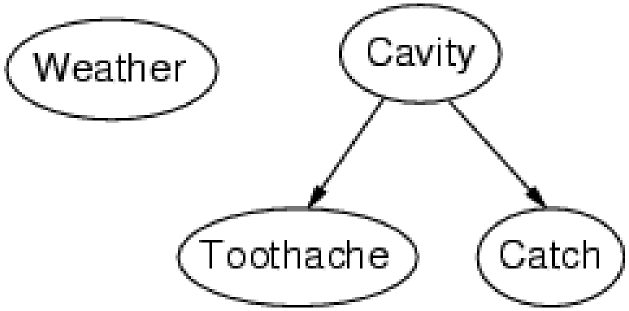
- By saying that X_i and X_j are independent, we mean that

$$\mathrm{P}(X_j, X_i) = \mathrm{P}(X_i)\mathrm{P}(X_j)$$

- X_i and X_j are independent if and only if they have no common ancestors
- Example: *independent coin flips*

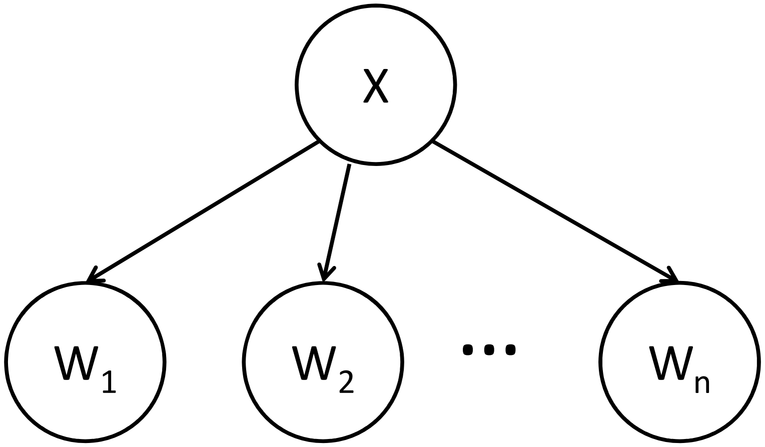


- Another example: Weather is independent of all other variables in this model.



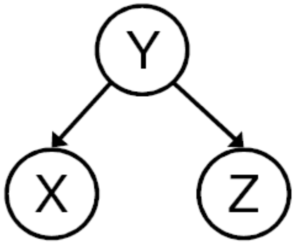
Conditional independence

- By saying that W_i and W_j are conditionally independent given X, we mean that
$$\mathrm{P}\left(W_i, W_j \middle| X\right) = \mathrm{P}(W_i|X)\mathrm{P}(W_j|X)$$
- W_i and W_j are conditionally independent given X if and only if they have no common ancestors other than the ancestors of X.
- Example: *naïve Bayes model:*



Conditional independence ≠ Independence

Common cause: Conditionally Independent



Y: Project due

X: Newsgroup busy

Z: Lab full

Are X and Z independent? **No**

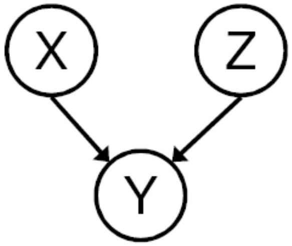
$$P(Z,X) = \sum_Y P(Z|Y)P(X|Y)P(Y)$$

$$P(Z)P(X) = \left(\sum_Y P(Z|Y)P(Y)\right)\left(\sum_Y P(X|Y)P(Y)\right)$$

Are they conditionally independent given Y? **Yes**

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

Common effect: Independent



X: Raining

Z: Ballgame

Y: Traffic

Are X and Z independent? **Yes**

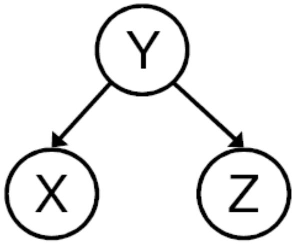
$$P(X,Z) = P(X)P(Z)$$

Are they conditionally independent given Y? **No**

$$P(Z,X|Y) = \frac{P(Y|X,Z)P(X)P(Z)}{P(Y)} \neq P(Z|Y)P(X|Y)$$

Conditional independence ≠ Independence

Common cause: Conditionally Independent



Y: Project due

X: Newsgroup busy

Z: Lab full

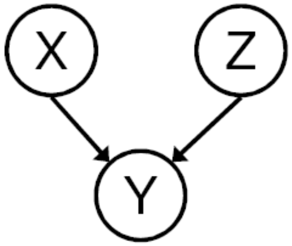
Are X and Z independent? **No**

Knowing X tells you about Y, which tells you about Z.

Are they conditionally independent given Y? **Yes**

If you already know Y, then X gives you no useful information about Z.

Common effect: Independent



X: Raining

Z: Ballgame

Y: Traffic

Are X and Z independent? **Yes**

Knowing X tells you nothing about Z.

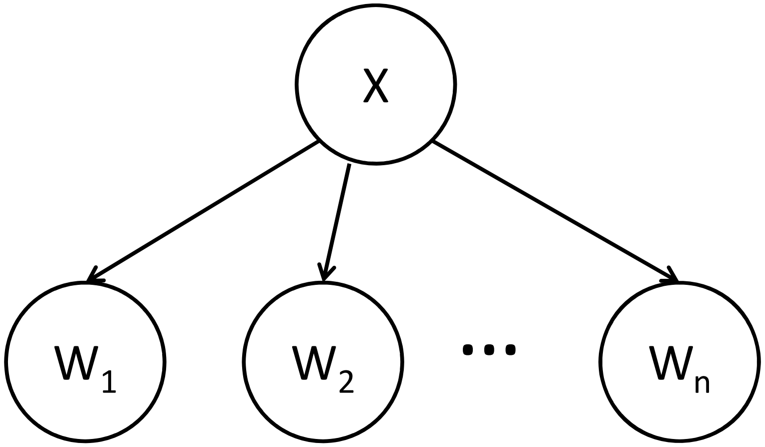
Are they conditionally independent given Y? **No**

If Y is true, then either X or Z must be true.

Knowing that X is false means Z must be true.

We say that X “explains away” Z.

Conditional independence ≠ Independence

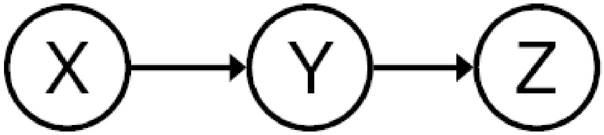


Being conditionally independent given X does NOT mean that W_i and W_j are independent. Quite the opposite.

Suppose $\mathrm{P}(X) = 1/2$, and $w_i = x$, and $w_j = x.$ Then $P(W_i) = \frac{1}{2}$, but $P\left(W_i|W_j\right) = 1$.

Conditional independence

Another example: *causal chain*



X: Low pressure

Y: Rain

Z: Traffic

- X and Z are conditionally independent given Y, because they have no common ancestors other than the ancestors of Y.

- Being conditionally independent given Y does NOT mean that X and Z are independent. Quite the opposite. For example, suppose $\mathrm{P}(X) = 0.5$, $\mathrm{P}(Y|X) = 0.8$, $\mathrm{P}(Y|\neg X) = 0.1$, $\mathrm{P}(Z|Y) = 0.7$, and $\mathrm{P}(Z|\neg Y) = 0.4$. Then we can calculate that $\mathrm{P}(Z|X) = 0.64$, but $\mathrm{P}(Z) = 0.535$

Outline

- Review: Bayesian inference
- Bayesian network: graph semantics
- The Los Angeles burglar alarm example
- Conditional independence ≠ Independence
- Constructing a Bayesian network: Structure learning
- Constructing a Bayesian network: Hire an expert

Constructing a Bayes Network: Two Methods

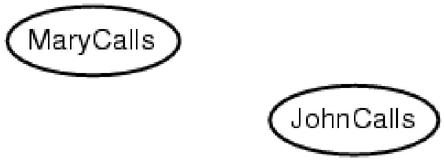
1. “Structure Learning” a.k.a. “Analysis of Causality:”
    1. Suppose you know the variables, but you don’t know which variables depend on which others. You can learn this from data.
    2. This is an exciting new area of research in statistics, where it goes by the name of “analysis of causality.”
    3. … but it’s almost harder than method #2, and I don’t recommend using it unless you really have to.
2. “Hire an Expert:”
    1. Find somebody who knows how to solve the problem.
    2. Get her to tell you what are the important variables, and which variables depend on which others.
    3. THIS IS ALMOST ALWAYS THE BEST WAY.

Constructing Bayesian networks

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
     $P(X_i \mid Parents(X_i)) = P(X_i \mid X_1, \ldots X_{i-1})$

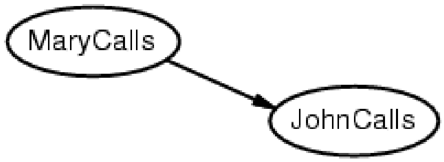
Example

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



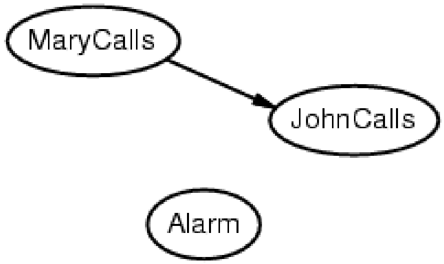
Example

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



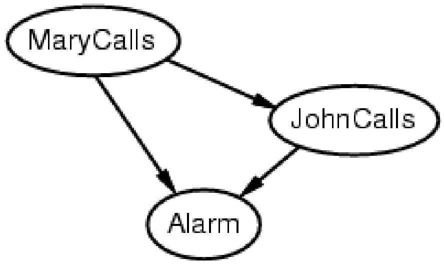
Example

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



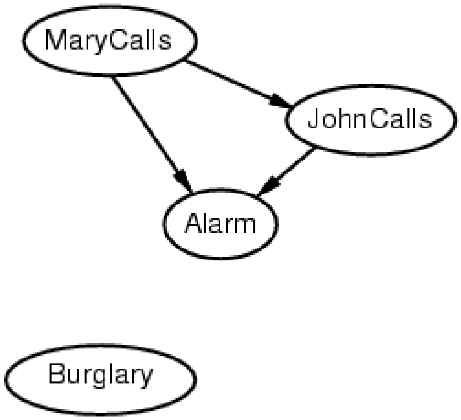
Example

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



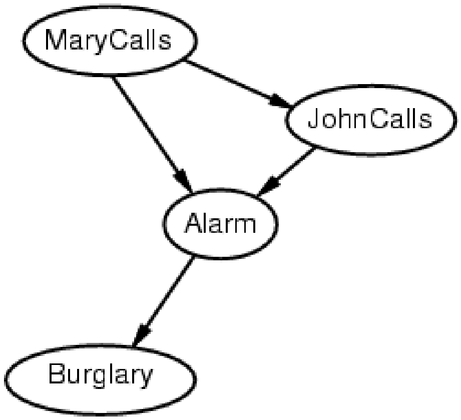
Example

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



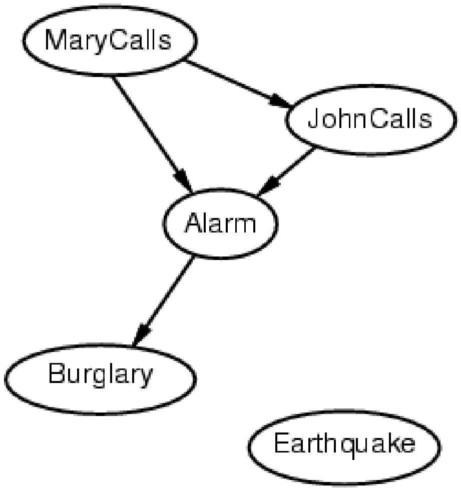
Example

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



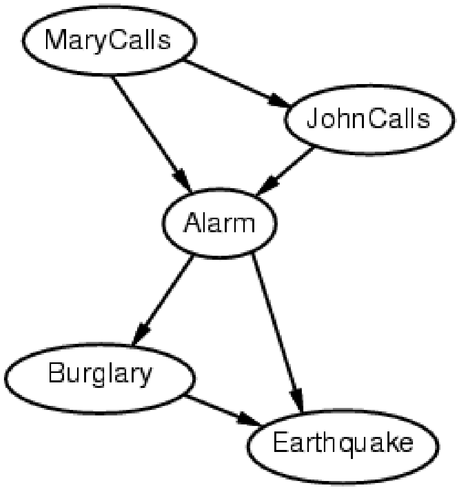
Example

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

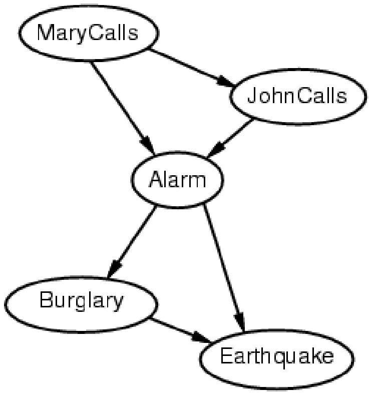


Example

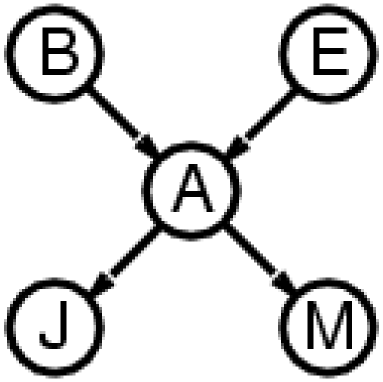
- Suppose we choose the ordering M, J, A, B, E



Example contd.



versus

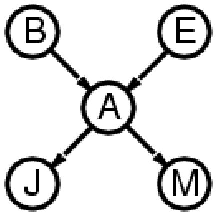


- Deciding conditional independence is hard in noncausal directions
    - The causal direction seems much more natural
- Network is less compact: 1 + 2 + 4 + 2 + 4 = 13 numbers needed (vs. 1+1+4+2+2=10 for the causal ordering)

Why store it in causal order? A: Saves memory

- Suppose we have a Boolean variable X_i with k Boolean parents. How many rows does its conditional probability table have?
  - 2^k rows for all the combinations of parent values
  - Each row requires one number for $P(X_i = \text{true} \mid \text{parent values})$
- If each variable has no more than k parents, how many numbers does the complete network require?
  - $O(n \cdot 2^k)$ numbers – vs. $O(2^n)$ for the full joint distribution
- How many nodes for the burglary network?

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

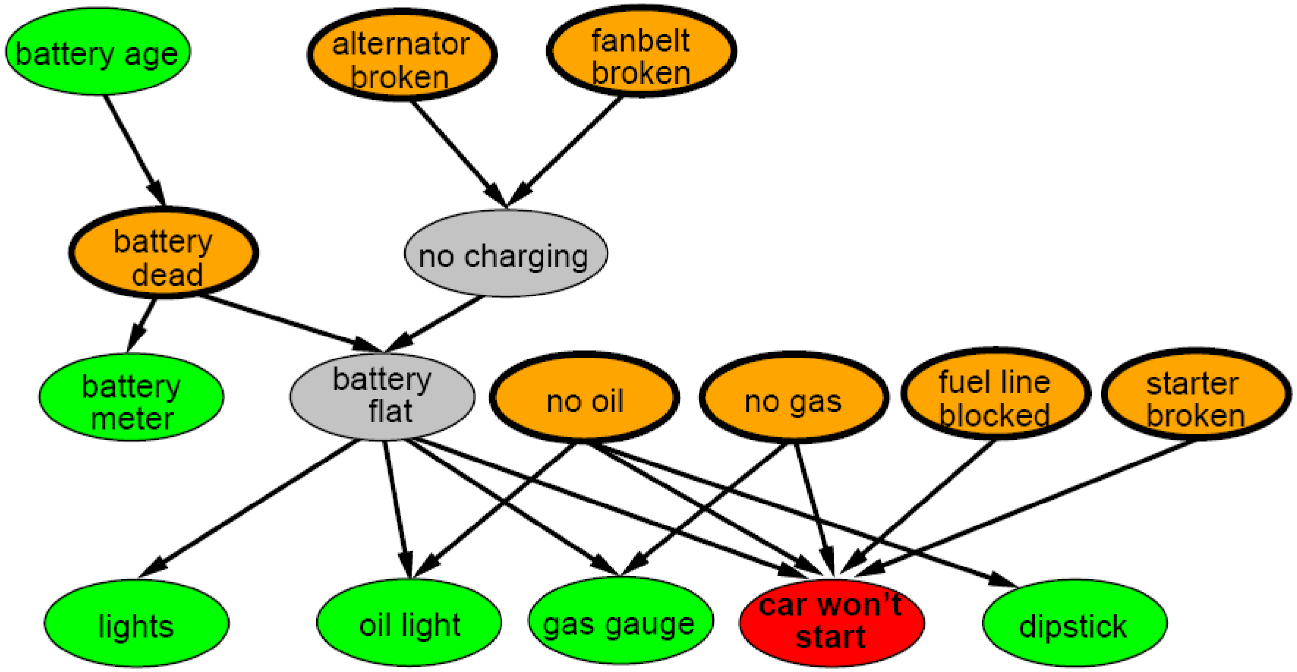


Outline

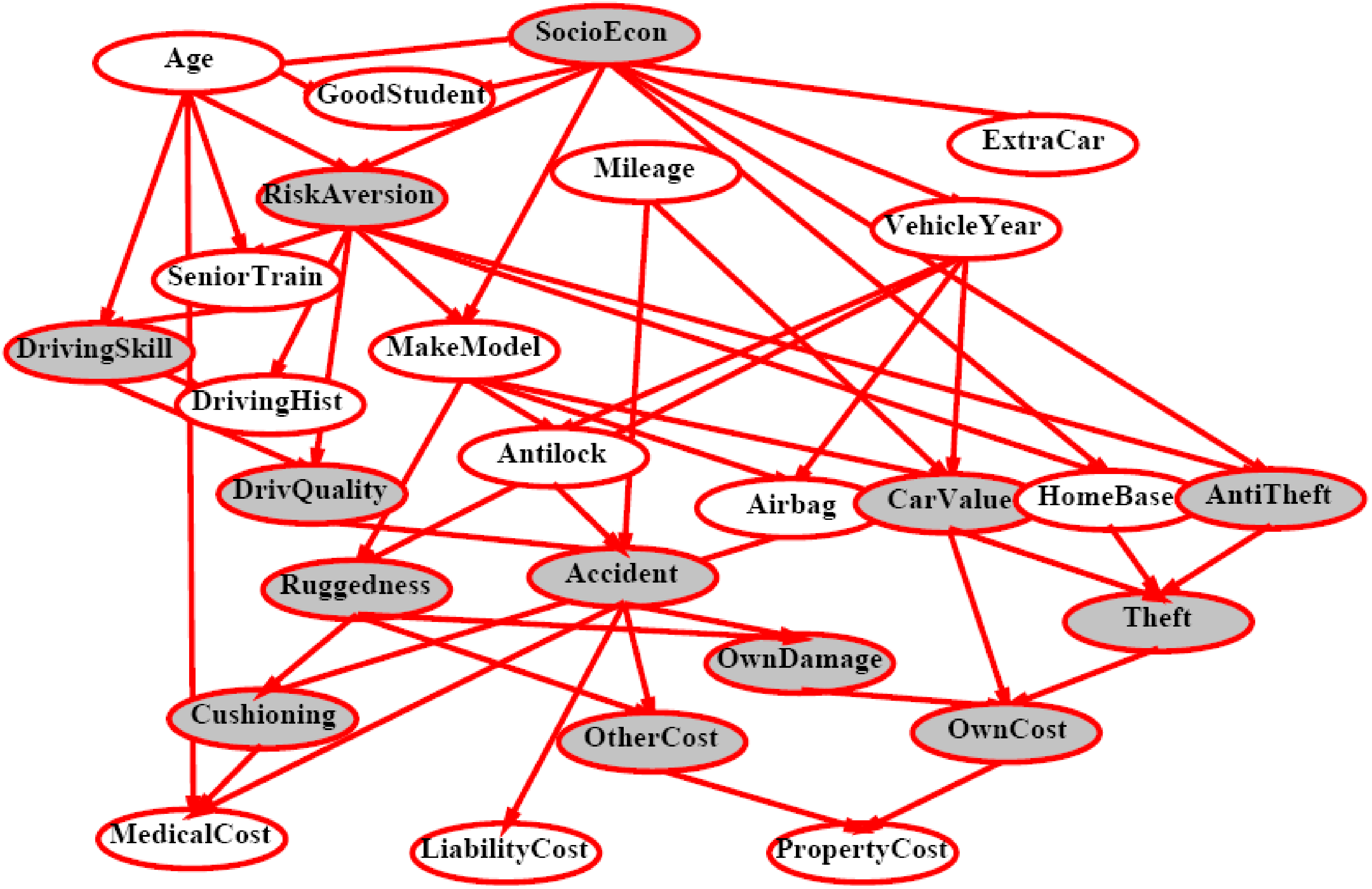
- Review: Bayesian inference
- Bayesian network: graph semantics
- The Los Angeles burglar alarm example
- Conditional independence ≠ Independence
- Constructing a Bayesian network: Structure learning
- Constructing a Bayesian network: Hire an expert

A more realistic Bayes Network: Car diagnosis

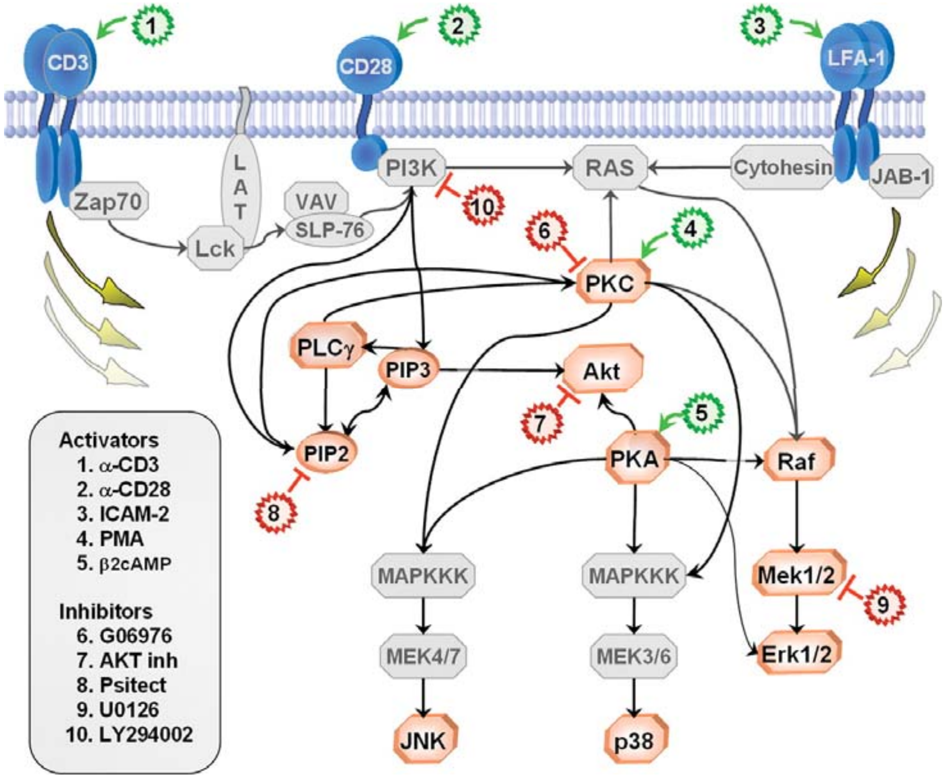
- Initial observation: car won’t start
- Orange: “broken, so fix it” nodes
- Green: testable evidence
- Gray: “hidden variables” to ensure sparse structure, reduce parameters



Car insurance



In research literature…



Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A. Lauffenburger, and Garry P. Nolan

(22 April 2005) ***Science*** **308** (5721), 523.

In research literature…

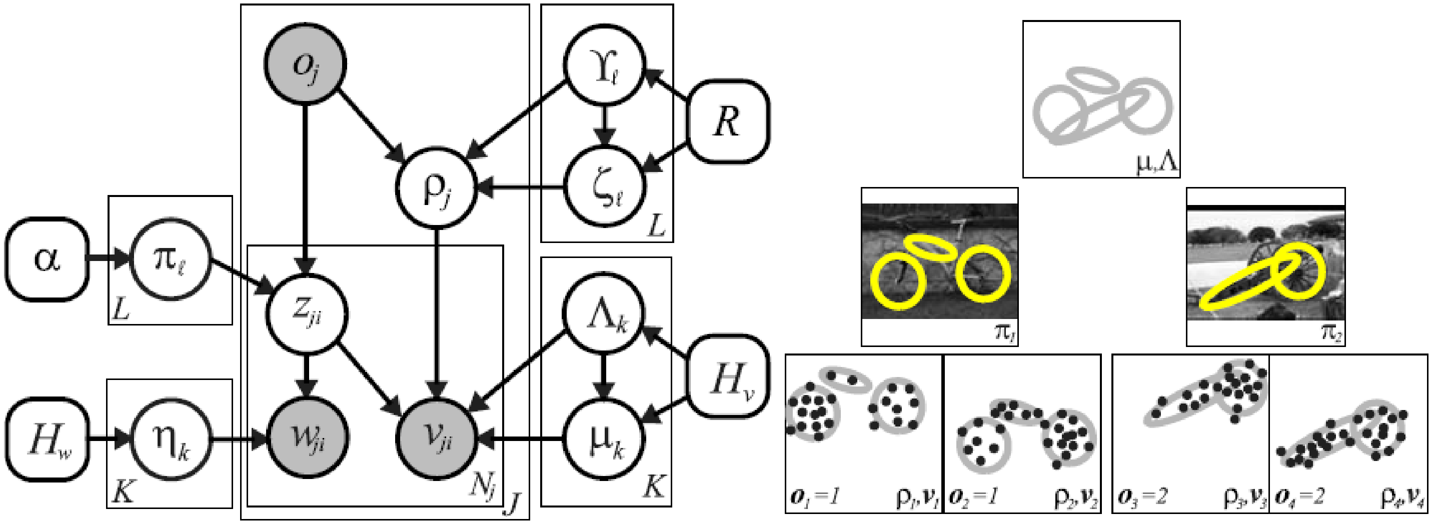


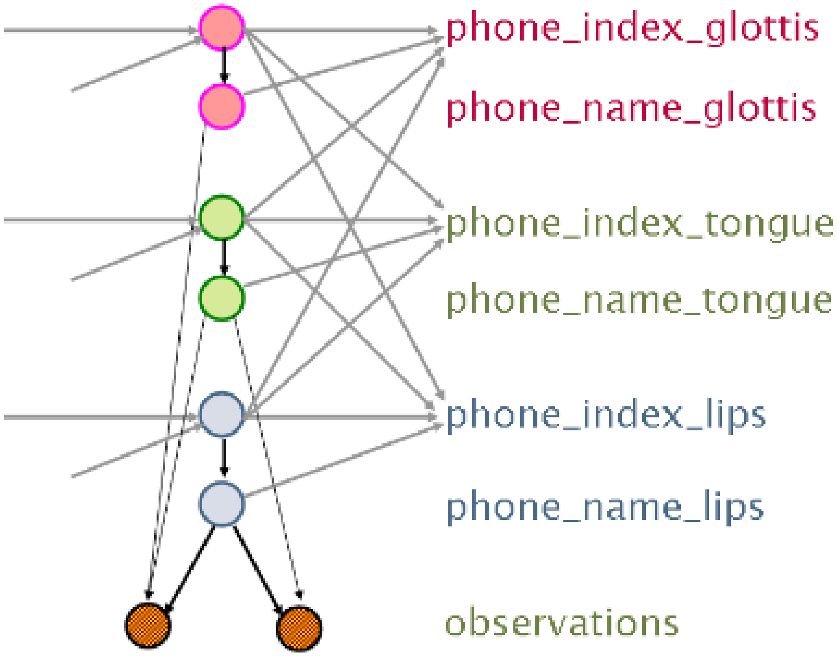
Fig. 3 A parametric, fixed-order model which describes the visual appearance of L object categories via a common set of K shared parts. The j^{th} image depicts an instance of object category o_j, whose position is determined by the reference transformation ρ_j. The appearance w_{ji} and position v_{ji}, relative to ρ_j, of visual features are determined by assignments $z_{ji} \sim \pi_{o_j}$ to latent parts. The cartoon example illustrates how a wheel part might be shared among two categories, *bicycle* and *cannon*. We show feature positions (but not appearance) for two hypothetical samples from each category

Describing Visual Scenes Using Transformed Objects and Parts

E. Sudderth, A. Torralba, W. T. Freeman, and A. Willsky.

International Journal of Computer Vision, No. 1-3, May 2008, pp. 291-330.

In research literature…

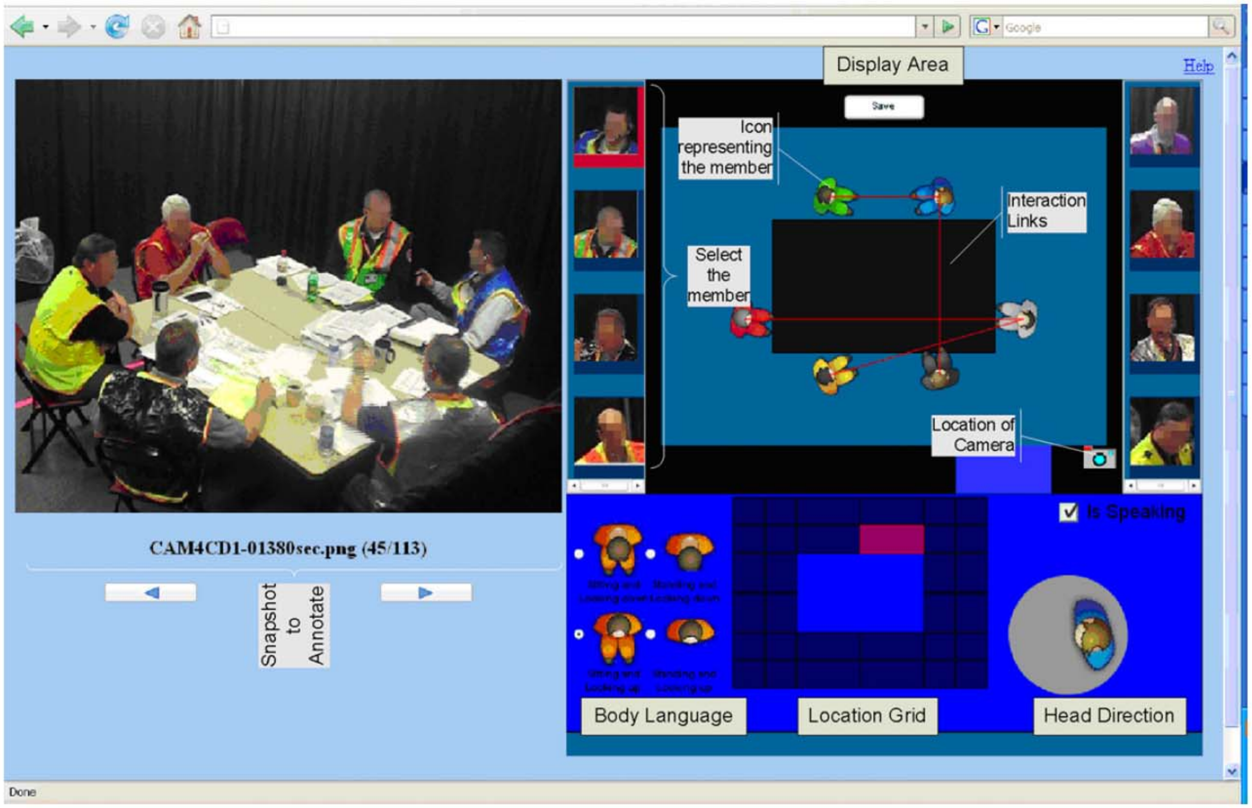


[Audiovisual Speech Recognition with Articulator Positions as Hidden Variables]

Mark Hasegawa-Johnson, Karen Livescu, Partha Lal and Kate Saenko

International Congress on Phonetic Sciences 1719:299-302, 2007

In research literature…

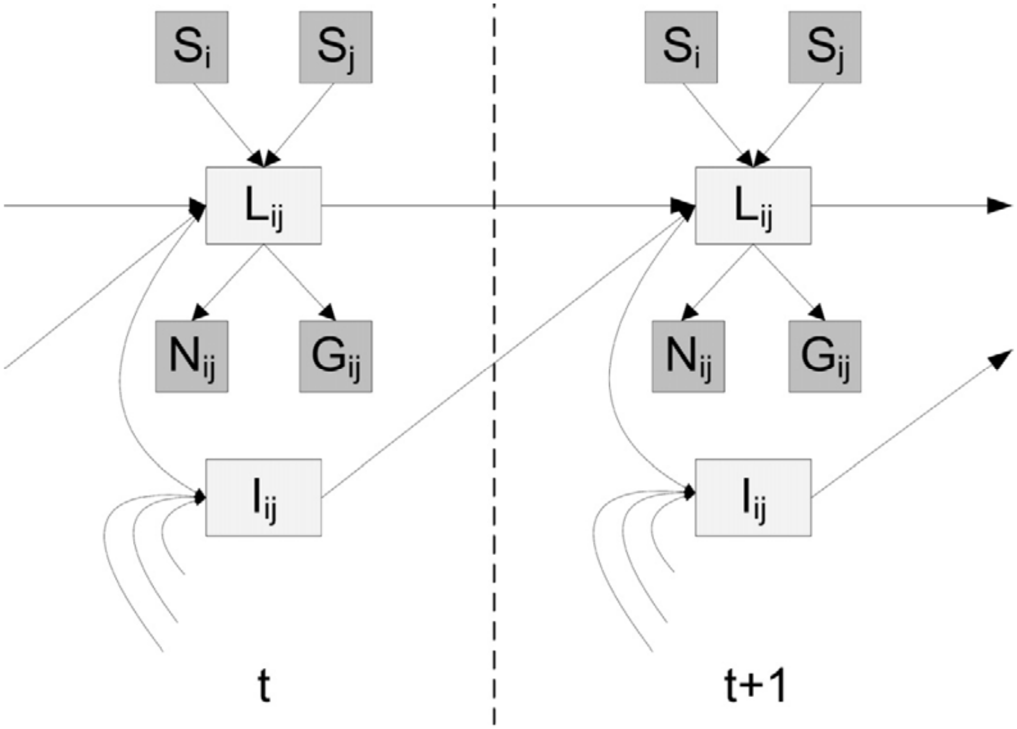


[Detecting interaction links in a collaborating group using manually annotated data]

S. Mathur, M.S. Poole, F. Pena-Mora, M. Hasegawa-Johnson, N. Contractor

Social Networks 10.1016/j.socnet.2012.04.002

In research literature…



- **Link:** $L_{ij} = 1$ if #i is listening to #j.
- **Indirect:** $I_{ij} = 1$ if #i and #j are both listening to the same person.
- **Speaking**: $S_i = 1$ if the i'th person is speaking.
- **Gaze:** $G_{ij} = 1$ if #i is looking at #j.
- **Neighborhood:** $N_{ij} = 1$ if they're near one another

Detecting interaction links in a collaborating group using manually annotated data

S. Mathur, M.S. Poole, F. Pena-Mora, M. Hasegawa-Johnson, N. Contractor

Social Networks 10.1016/j.socnet.2012.04.002

Summary

- Bayesian networks provide a natural representation for (causally induced) conditional independence
- Topology + conditional probability tables
- Generally easy for domain experts to construct