Bayesian Networks

CS171, Fall 2016

Introduction to Artificial Intelligence

Prof. Alexander Ihler

Reading: R&N Ch 14

BREN:ICS INFORMATION AND COMPUTER SCIENCES

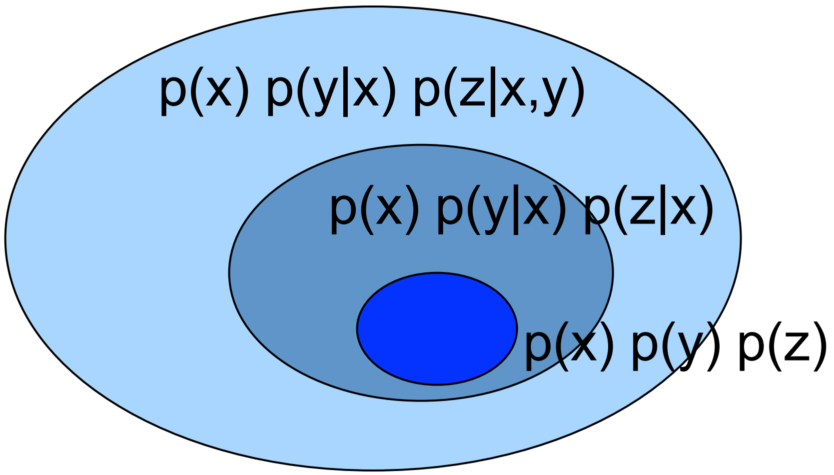
UNIVERSITY of CALIFORNIA IRVINE

Why Bayesian Networks?

- Knowledge Representation & Reasoning (Inference)
  - Propositional Logic
    - Knowledge Base : Propositional logic sentences
    - Reasoning : KB |= Theory
      - Find a model or Count models
  - Probabilistic Reasoning
    - Knowledge Base : Full joint probability over all random variables
    - Reasoning: Compute Pr (KB |= Theory)
      - Find the most probable assignments
      - Compute marginal / conditional probability
- Why Bayesian Net?
  - Manipulating full joint probability distribution is very hard!
  - Exploit conditional independence properties of our distribution
  - Bayesian Network captures conditional independence
    - Graphical Representation (Probabilistic Graphical Models)
    - Tool for Reasoning, Computation (Probabilistic Reasoning bases on the Graph)

Conditional independence

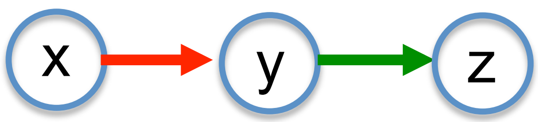
- Recall: chain rule of probability
  - $p(x,y,z) = p(x)\, p(y|x)\, p(z|x,y)$
- *Some* of these models will be conditionally independent
  - e.g., $p(x,y,z) = p(x)\, p(y|x)\, p(z|x)$
- *Some* models may have even *more* independence
  - E.g., $p(x,y,z) = p(x)\, p(y)\, p(z)$



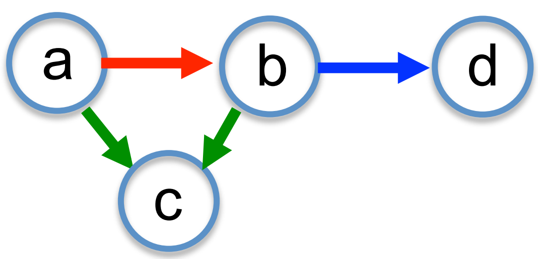
Bayesian networks

- Directed graphical model
- Nodes associated with variables
- "Draw" independence in conditional probability expansion
  - Parents in graph are the RHS of conditional

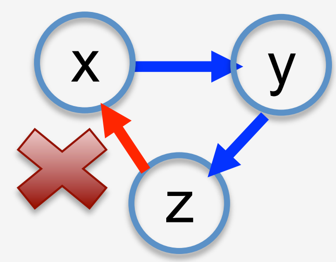
- Ex: $p(x, y, z) = p(x)\ p(y \mid x)\ p(z \mid y)$



- Ex: $p(a, b, c, d) = p(a)\ p(b \mid a)\ p(c \mid a, b)\ p(d \mid b)$



*Graph must be **acyclic***



Corresponds to an order over the variables (chain rule)

Example from Russell & Norvig

Example

- Consider the following 5 binary variables:
  - B = a burglary occurs at your house
  - E = an earthquake occurs at your house
  - A = the alarm goes off
  - J = John calls to report the alarm
  - M = Mary calls to report the alarm

  - What is P(B | M, J) ? (for example)

  - We can use the full joint distribution to answer this question
    - Requires $2^5 = 32$ probabilities

    - Can we use prior domain knowledge to come up with a Bayesian network that requires fewer probabilities?

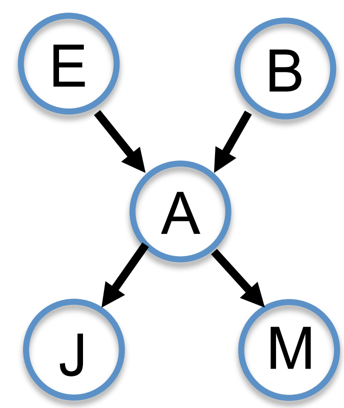
Constructing a Bayesian network

- Order the variables in terms of causality (may be a partial order)
  - e.g., $\{\ E,\ B\ \} \longrightarrow \{\ A\ \} \longrightarrow \{\ J,\ M\ \}$

- Now, apply the chain rule, and simplify based on assumptions

$$\begin{aligned} p(J, M, A, E, B) &= p(E, B)\ p(A \mid E, B)\ p(J, M \mid A, E, B) \\ &= p(E)\ p(B)\ p(A \mid E, B)\ p(J, M \mid A) \\ &= p(E)\ p(B)\ p(A \mid E, B)\ p(J \mid A)\ p(M \mid A) \end{aligned}$$

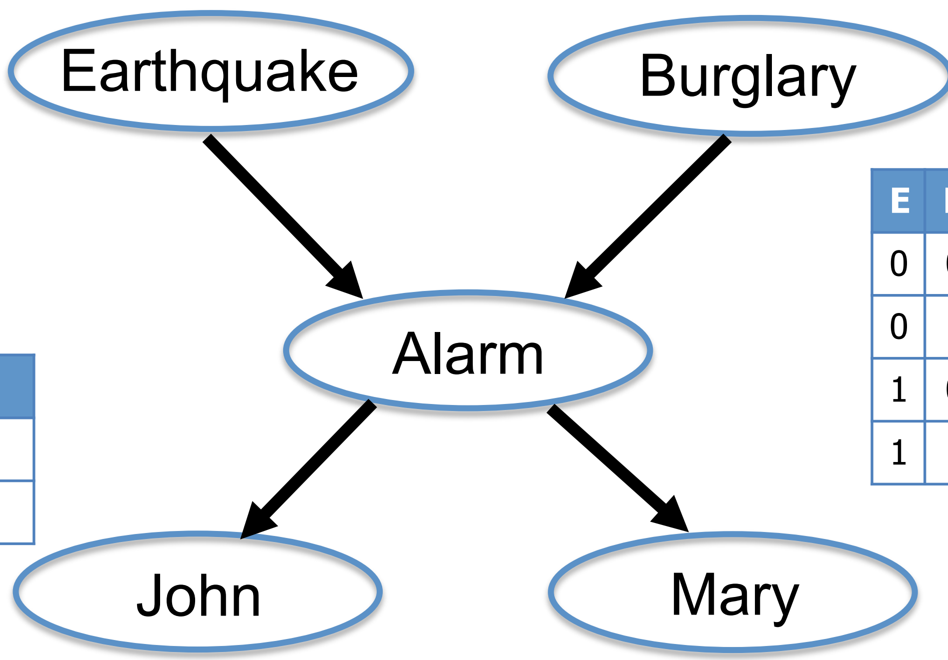
  - These assumptions are reflected in the graph structure of the Bayesian network



Constructing a Bayesian network

- Given $p(J, M, A, E, B) = p(E)\ p(B)\ p(A \mid E, B)\ p(J \mid A)\ p(M \mid A)$
- Define probabilities: 1 + 1 + 4 + 2 + 2
- Where do these come from?
  - Expert knowledge; estimate from data; some combination

P(E)
0.002



P(B)
0.001

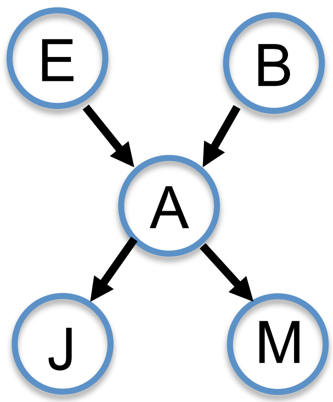
E	B	P(A|E,B)
0	0	0.001
0	1	0.29
1	0	0.94
1	1	0.95

A	P(J|A)
0	0.05
1	0.90

A	P(M|A)
0	0.01
1	0.70

Constructing a Bayesian network

- Joint distribution



Full joint distribution:
$2^5 = 32$ probabilities

Structured distribution:
specify 10 parameters

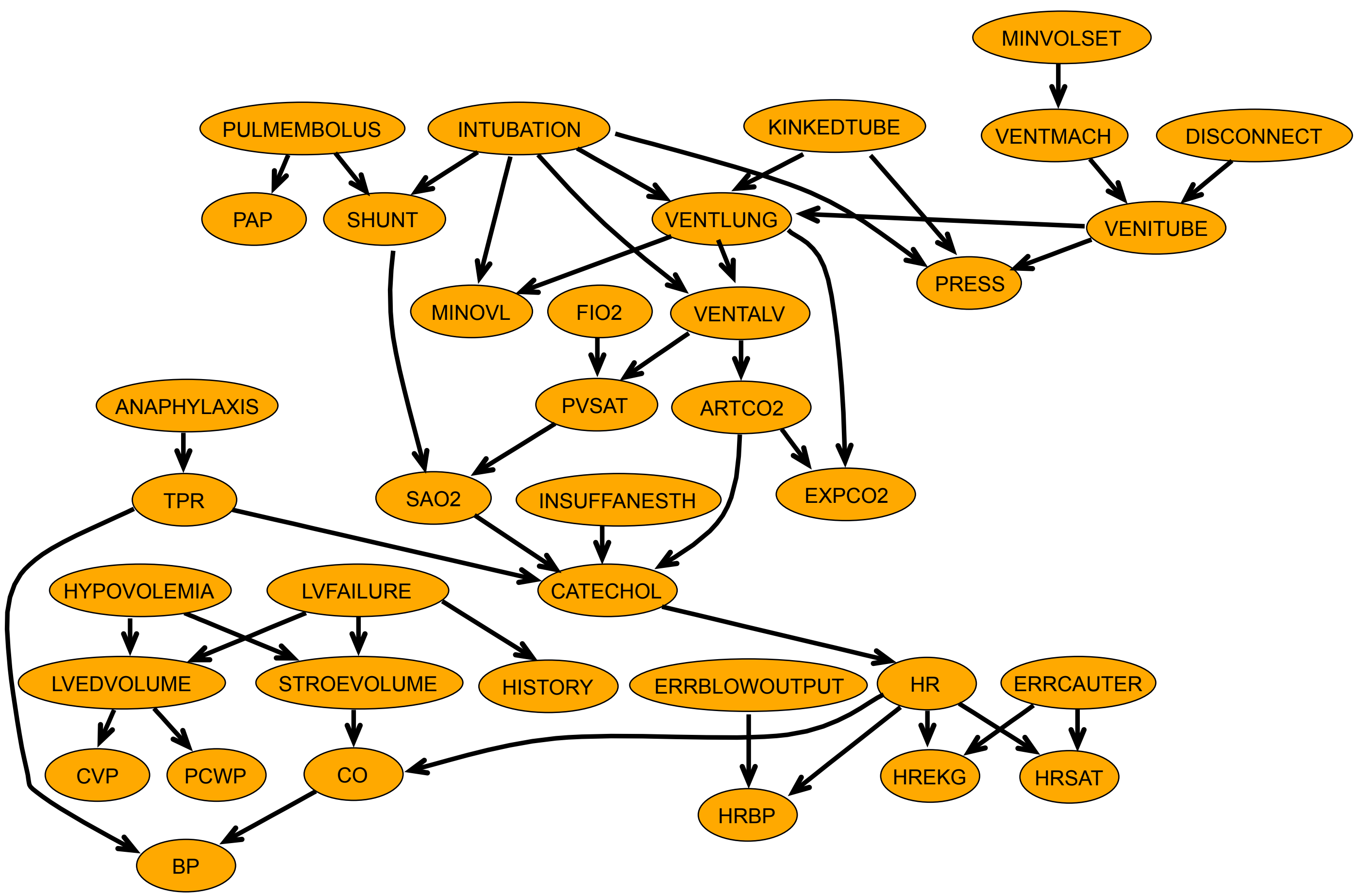
E	B	A	J	M	P(…)
0	0	0	0	0	.93674
0	0	0	0	1	.00133
0	0	0	1	0	.00005
0	0	0	1	1	.00000
0	0	1	0	0	.00003
0	0	1	0	1	.00002
0	0	1	1	0	.00003
0	0	1	1	1	.00000
0	1	0	0	0	.04930
0	1	0	0	1	.00007
0	1	0	1	0	.00000
0	1	0	1	1	.00000
0	1	1	0	0	.00027
0	1	1	0	1	.00016
0	1	1	1	0	.00025
0	1	1	1	1	.00000

E	B	A	J	M	P(…)
1	0	0	0	0	.00946
1	0	0	0	1	.00001
1	0	0	1	0	.00000
1	0	0	1	1	.00000
1	0	1	0	0	.00007
1	0	1	0	1	.00004
1	0	1	1	0	.00007
1	0	1	1	1	.00000
1	1	0	0	0	.00050
1	1	0	0	1	.00000
1	1	0	1	0	.00000
1	1	0	1	1	.00000
1	1	1	0	0	.00063
1	1	1	0	1	.00037
1	1	1	1	0	.00059
1	1	1	1	1	.00000

Alarm network

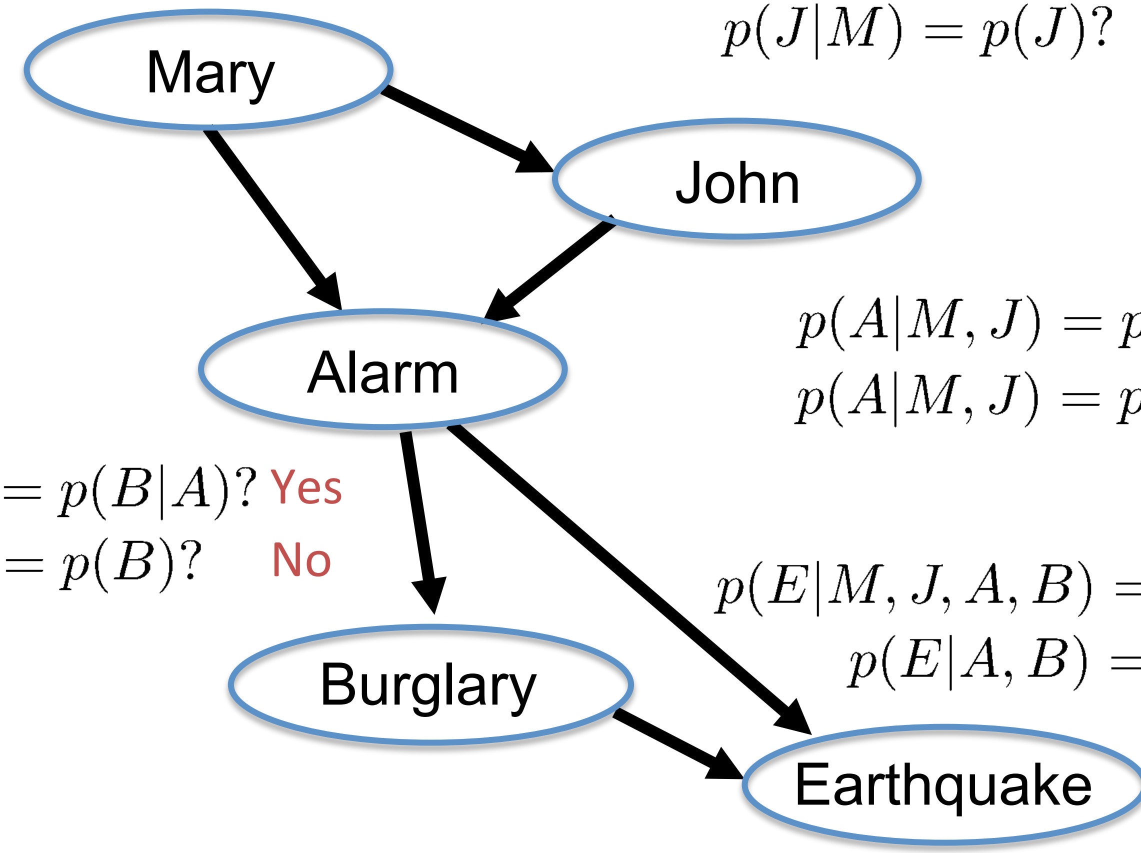
[Beinlich et al., 1989]

The “alarm” network: 37 variables, 509 parameters (rather than $2^{37} = 10^{11}$!)



Network structure and ordering

- The network structure depends on the conditioning order
  - Suppose we choose ordering M, J, A, B, E



$p(M)$

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

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

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

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

$p(B|A) = p(B)?$ No

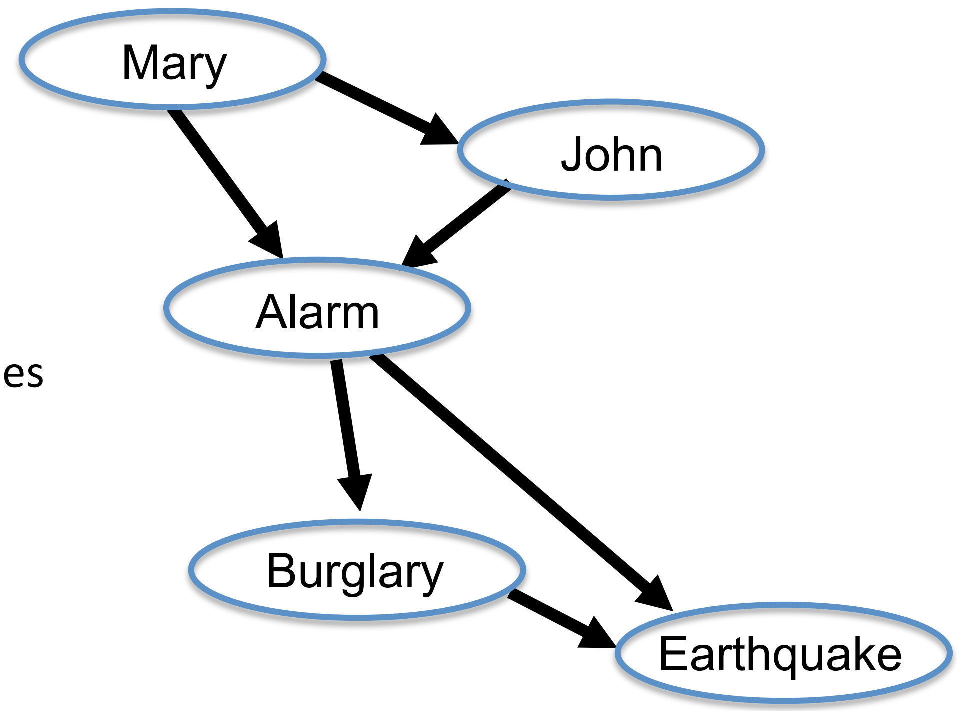
$p(E|M, J, A, B) = p(E|A, B)?$ Yes

$p(E|A, B) = p(E|A)?$ No

Network structure and ordering

- The network structure depends on the conditioning order
  - Suppose we choose ordering M, J, A, B, E
- "Non-causal" ordering
  - Deciding independence is harder
  - Selecting probabilities is harder
  - Representation is less efficient

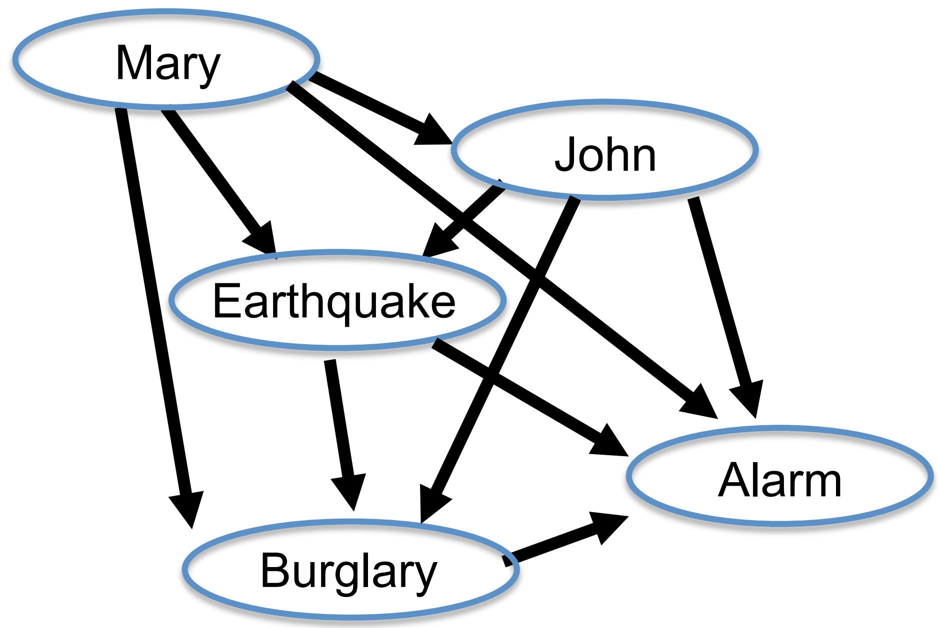
1 + 2 + 4 + 2 + 4 = 13 probabilities



Network structure and ordering

- The network structure depends on the conditioning order
  - Suppose we choose ordering M, J, A, B, E

- "Non-causal" ordering
  - Deciding independence is harder
  - Selecting probabilities is harder
  - Representation is less efficient

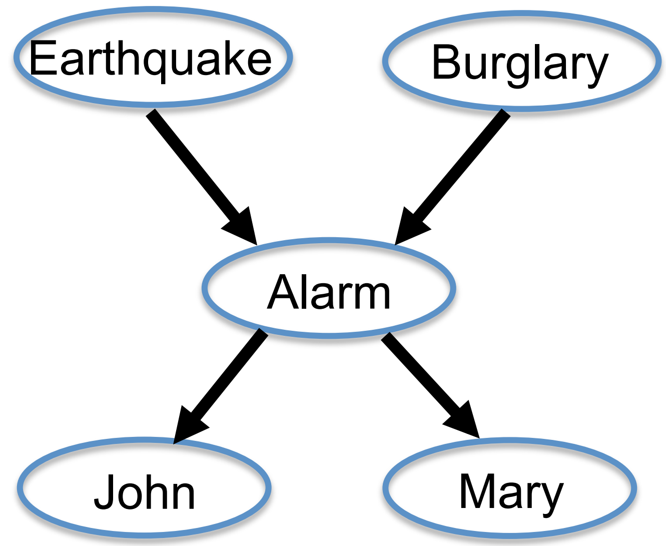


  - Some orders may not reveal any independence!

$$p(J, M, A, E, B) = p(M)\, p(J|M)\, p(E|M, J)\, p(B|M, J, E)\, p(A|M, J, E, B)$$

Reasoning in Bayesian networks

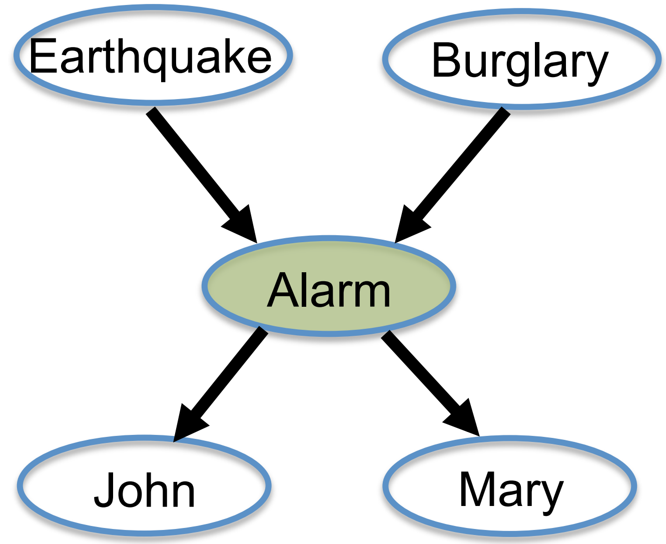
- Suppose we observe J
  - Observing J makes A more likely
  - A being more likely makes B more likely
- Suppose we observe A
  - Makes M more likely
- Observe A and J?
  - J doesn't add anything to M
  - Observing A makes J, M independent
- How can we read independence directly from the graph?



Reasoning in Bayesian networks

- How are J,M related given A?
  - P(M)   = 0.0117
  - P(M|A)   = 0.7
  - P(M|A,J) = 0.7
  - Conditionally independent

(we actually know this by construction!)

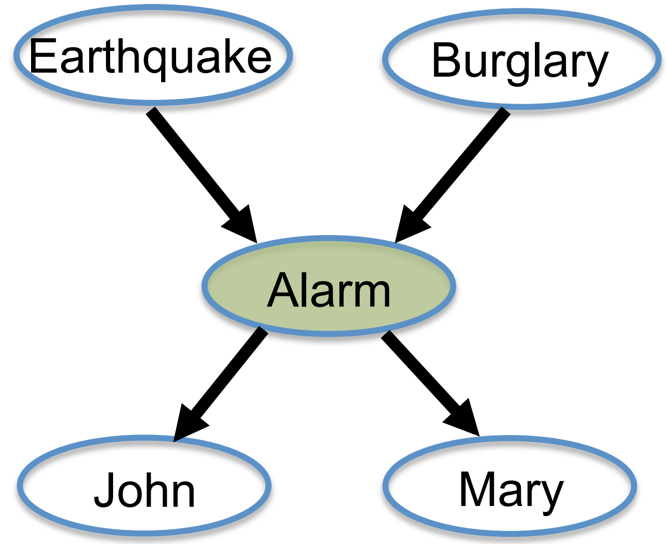


- Proof:

$$
\begin{aligned}
p(J, M|a) &\propto \sum_{e,b} p(e)\ p(b)\ p(a|e,b)\ p(J|a)\ p(M|a) \\
&= \Big(\sum_{e,b} p(e,b,a) \Big)\ p(J|a)\ p(M|a) \\
&= p(a)\ p(J|a)\ p(M|a) \\
&= \ c_a \ \ f_a(J)\ g_a(M)
\end{aligned}
$$

Reasoning in Bayesian networks

- How are J,B related given A?
  - P(B) = 0.001
  - P(B|A) = 0.3735
  - P(B|A,J) = 0.3735
  - Conditionally independent



- Proof:

$$p(J, B|a) \propto \sum_{e,m} p(e)\ p(B)\ p(a|e, B)\ p(J|a)\ p(m|a)$$

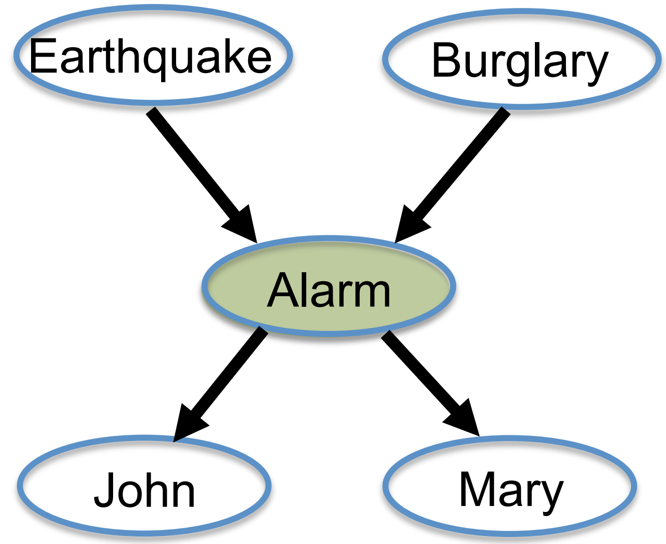
$$= \left(\sum_{e} p(e, B, a)\right)\ p(J|a)\ \left(\sum_{m} p(m|a)\right)$$

$$= p(B, a)\ p(J|a)$$

$$= \ f_a(B)\ \ g_a(J)$$

Reasoning in Bayesian networks

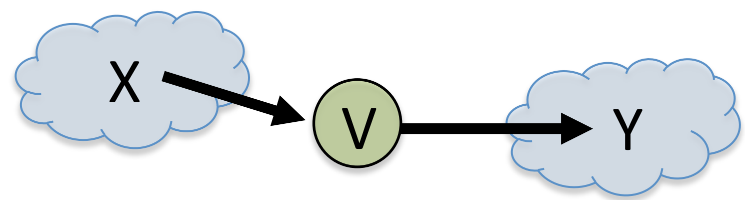
- How are E,B related?
  - P(B) = 0.001
  - P(B|E) = 0.001
  - (Marginally) independent
- What about given A?
  - P(B|A) = 0.3735
  - P(B|A,E) = 0.0032
  - Not conditionally independent!
  - The "causes" of A become coupled by observing its value
  - Sometimes called "explaining away"



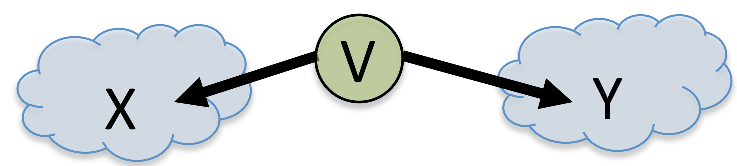
D-Separation

- Prove sets X,Y independent given Z?
- Check all *undirected* paths from X to Y
- A path is "inactive" if it passes through:

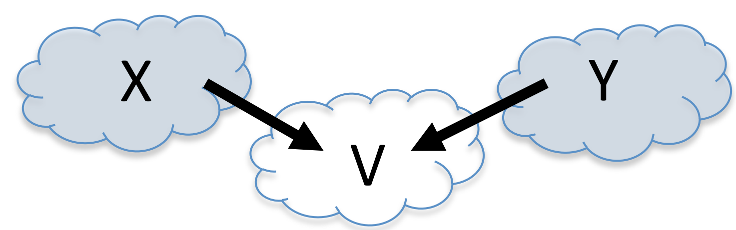
(1) A "chain" with an observed variable



(2) A "split" with an observed variable



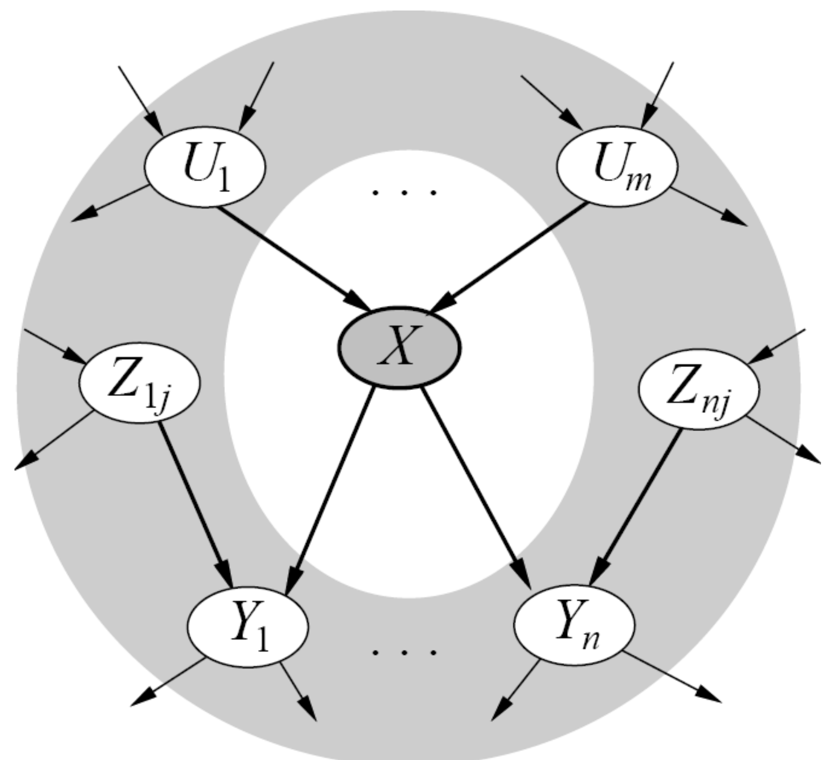
(3) A "vee" with **only unobserved** variables below it



- If all paths are inactive, conditionally independent!

Markov blanket

A node is conditionally independent of all other nodes in the network given its Markov blanket (in gray)

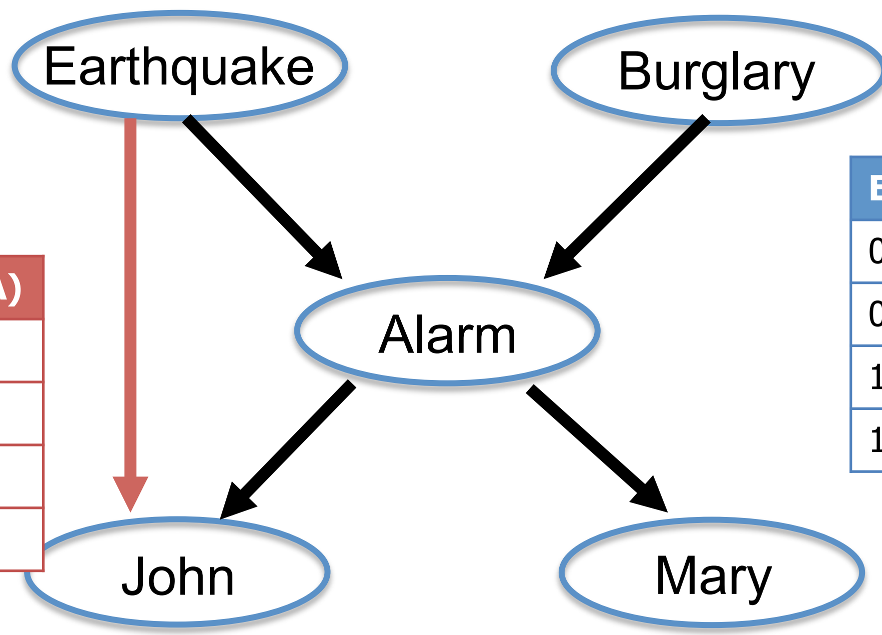


Graphs and Independence

- Graph structure allows us to infer independence in p(.)
  - X,Y d-separated given Z?
- Adding edges
  - Fewer independencies inferred, but still valid to represent p(.)
  - Complete graph: can represent any distribution p(.)

P(E)
0.002

E	A	P(J|E,A)
0	0	0.05
0	1	0.90
1	0	0.05
1	1	0.90



P(B)
0.001

E	B	P(A|E,B)
0	0	0.001
0	1	0.29
1	0	0.94
1	1	0.95

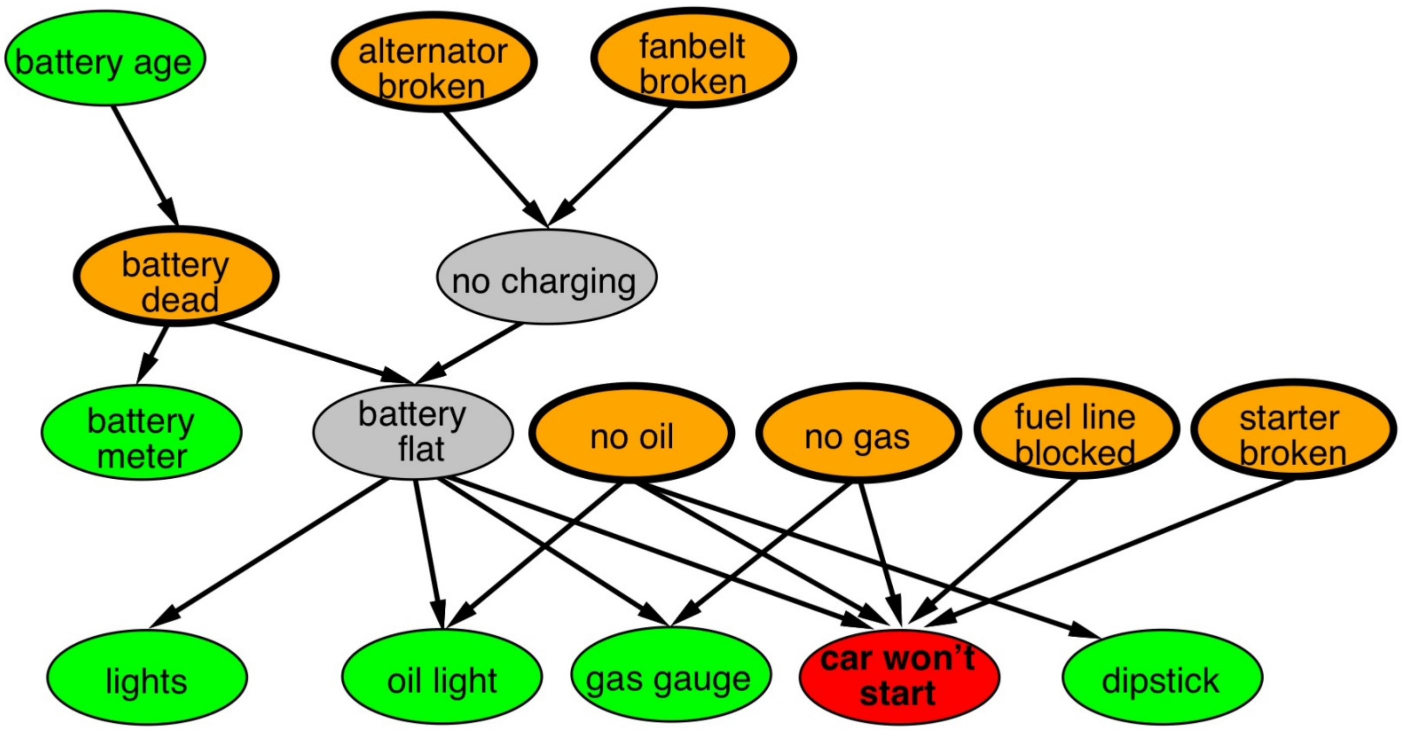
A	P(M|A)
0	0.01
1	0.70

Example: Car diagnosis

Initial evidence: car won't start

Testable variables (green), "broken, so fix it" variables (orange)

Hidden variables (gray) ensure sparse structure, reduce parameters



Compact conditional distributions contd.

Noisy-OR distributions model multiple noninteracting causes

1) Parents $U_1 \ldots U_k$ include all causes (can add leak node)

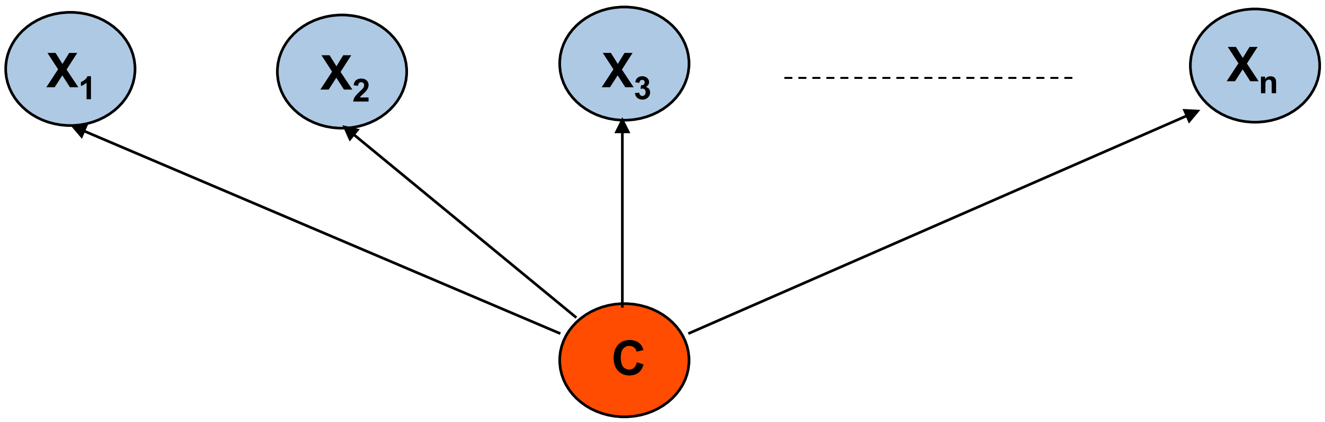
2) Independent failure probability q_i for each cause alone

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

$Cold$	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

Naïve Bayes Model



$$P(C \mid X_1,\ldots,X_n) = \alpha \prod P(X_i \mid C) \; P(C)$$

Features X are conditionally independent given the class variable C

Widely used in machine learning
    e.g., spam email classification: X's = counts of words in emails

Probabilities P(C) and P(Xi | C) can easily be estimated from labeled data

Naïve Bayes Model (2)

$$P(C \mid X_1, \ldots X_n) = \alpha \prod P(X_i \mid C) \; P(C)$$

<Learning Naïve Bayes Model>

Probabilities P(C) and P(Xi | C) can easily be estimated from labeled data

P(C = cj) ≈ #(Examples with class label cj) / #(Examples)
P(Xi = xik | C = cj)
≈ #(Examples with Xi value xik and class label cj)
/ #(Examples with class label cj)

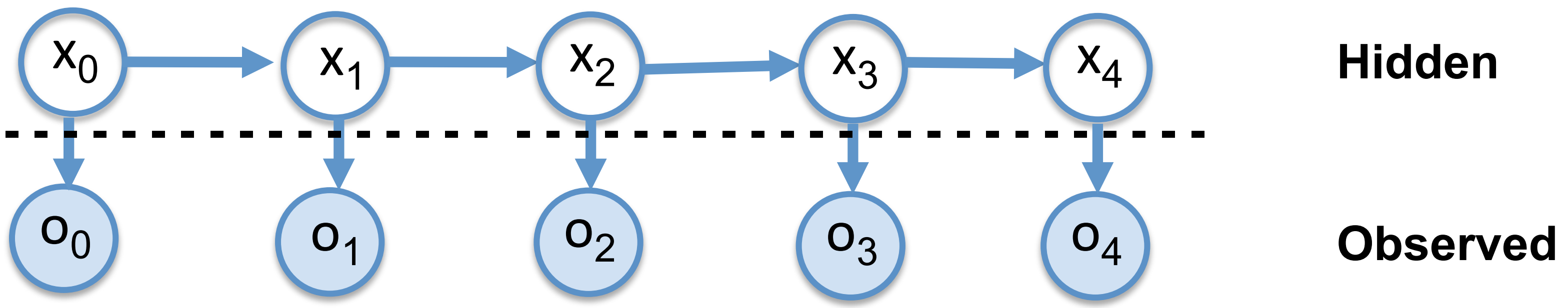
Usually easiest to work with logs

$$\log [\, P(C \mid X_1, \ldots X_n) \,]$$
$$= \log \alpha + \sum [\, \log P(X_i \mid C) + \log P(C) \,]$$

DANGER: Suppose ZERO examples with Xi value xik and class label cj ?
An unseen example with Xi value xik will NEVER predict class label cj !

Practical solutions: Pseudocounts, e.g., add 1 to every #() , etc.
Theoretical solutions: Bayesian inference, beta distribution, etc.

Hidden Markov Models



- Two key assumptions
  - Hidden state sequence is Markov
  - Observations o_t is conditionally independent given state x_t
- Widely used in:
  - speech recognition, protein sequence models, ...
- Bayesian network is a tree, so inference is linear in n
  - Exploit graph structure for efficient computation (as in CSPs)

You should know…

- Basic concepts and vocabulary of Bayesian networks.
  - Nodes represent random variables.
  - Directed arcs represent (informally) direct influences.
  - Conditional probability tables, P(Xi | Parents(Xi)).

- Given a Bayesian network:
  - Write down the full joint distribution it represents.

- Given a full joint distribution in factored form:
  - Draw the Bayesian network that represents it.

- Given a variable ordering and some background assertions of conditional independence among the variables:
  - Write down the factored form of the full joint distribution, as simplified by the conditional independence assertions.

Summary

- Bayesian networks represent a joint distribution using a graph
- The graph encodes a set of conditional independence assumptions
- Answering queries (or inference or reasoning) in a Bayesian network amounts to efficient computation of appropriate conditional probabilities
- Probabilistic inference is intractable in the general case
  - But can be carried out in linear time for certain classes of Bayesian networks