Representing Knowledge with Bayesian Networks

CMPSCI 383
October 27, 2011

Today's topics

- Quick review of probability
- Representing joint probability distributions
- Bayesian networks
  - Syntax and semantics
  - Independence relations they encode

Probability terminology

- This type of distribution gives the probability of conjunctions of propositions.
- This method extracts the probability of a subset of variables from a joint distribution.
- This method extracts the joint probability of a subset of variables given a set of others.

- *What is a* **Joint distribution**
- *What is* **Marginalization**
- *What is* **Conditioning**

Joint probability distributions

- Joint probability distributions describe the probabilities of conjunctions of propositions.
- Example
  - Probability of passing all your courses

$$p(P_{383}, P_{377}, P_{History}, P_{English}, \ldots)$$

Example: Visiting the dentist

	T		$\neg T$	
	C	$\neg C$	C	$\neg C$
V	0.108	0.012	0.072	0.008
$\neg V$	0.016	0.064	0.144	0.576

V = Cavity; T = Toothache; C = Catch

What is the sum of probabilities in this table?

Marginalization

	T		$\neg T$	
	C	$\neg C$	C	$\neg C$
V	0.108	0.012	0.072	0.008
$\neg V$	0.016	0.064	0.144	0.576

What is the p(C)?

p(C) = 0.108 + 0.016 + 0.072 + 0.144 = 0.20

Conditioning

	T		$\neg T$	
	C	$\neg C$	C	$\neg C$
V	0.108	0.012	0.072	0.008
$\neg V$	0.016	0.064	0.144	0.576

What is the p(CIT)?

(0.108+0.016)/(0.108+0.012+0.016+0.064)

0.62

Conditioning

	T		$\neg T$	
	C	$\neg C$	C	$\neg C$
V	0.108	0.012	0.072	0.008
$\neg V$	0.016	0.064	0.144	0.576

This is just an application of the definition:

$$p(C|T) = p(C,T) / p(T)$$

Joint distributions are powerful

- From a joint distribution of a set of variables, you can calculate
  - The joint probability distribution of any subset of those variables
  - The conditional probability distribution of any subset given any other subset
- The joint distributions is “everything you need to know” about a set of variables

…but there is a problem: dimensionality

	T		$\neg T$	
	C	$\neg C$	C	$\neg C$
V	0.108	0.012	0.072	0.008
$\neg V$	0.016	0.064	0.144	0.576

- Simple conditional probability tables grow exponentially with the number of variables
- Essentially impossible once you have more than about 10 variables.

Solution: Exploit independence

	T		$\neg T$	
	C	$\neg C$	C	$\neg C$
V	0.108	0.012	0.072	0.008
$\neg V$	0.016	0.064	0.144	0.576

- Do we need to include the weather?
- Do we need to include it in every part of the table?

Bayesian networks

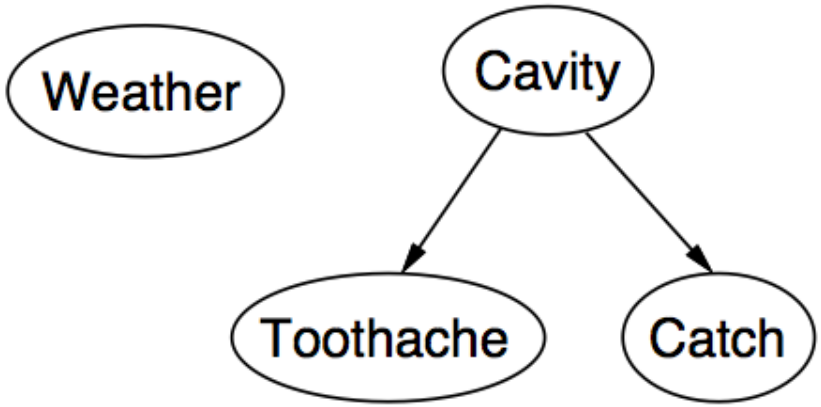
- A simple, graphical notation for conditional independence assertions and hence for compact specification of joint distributions
- Syntax:
  - a set of nodes, one per variable
  - a directed, acyclic graph (link ≈ "directly influences")
  - a conditional distribution for each node given its parents:

$$\mathbf{P}\,(X_i \mid \text{Parents}\,(X_i))$$

- In the simplest case, conditional distribution represented as a conditional probability table (CPT) giving the distribution over X_i for each combination of parent values

Example

- Topology of network encodes conditional independence assertions:

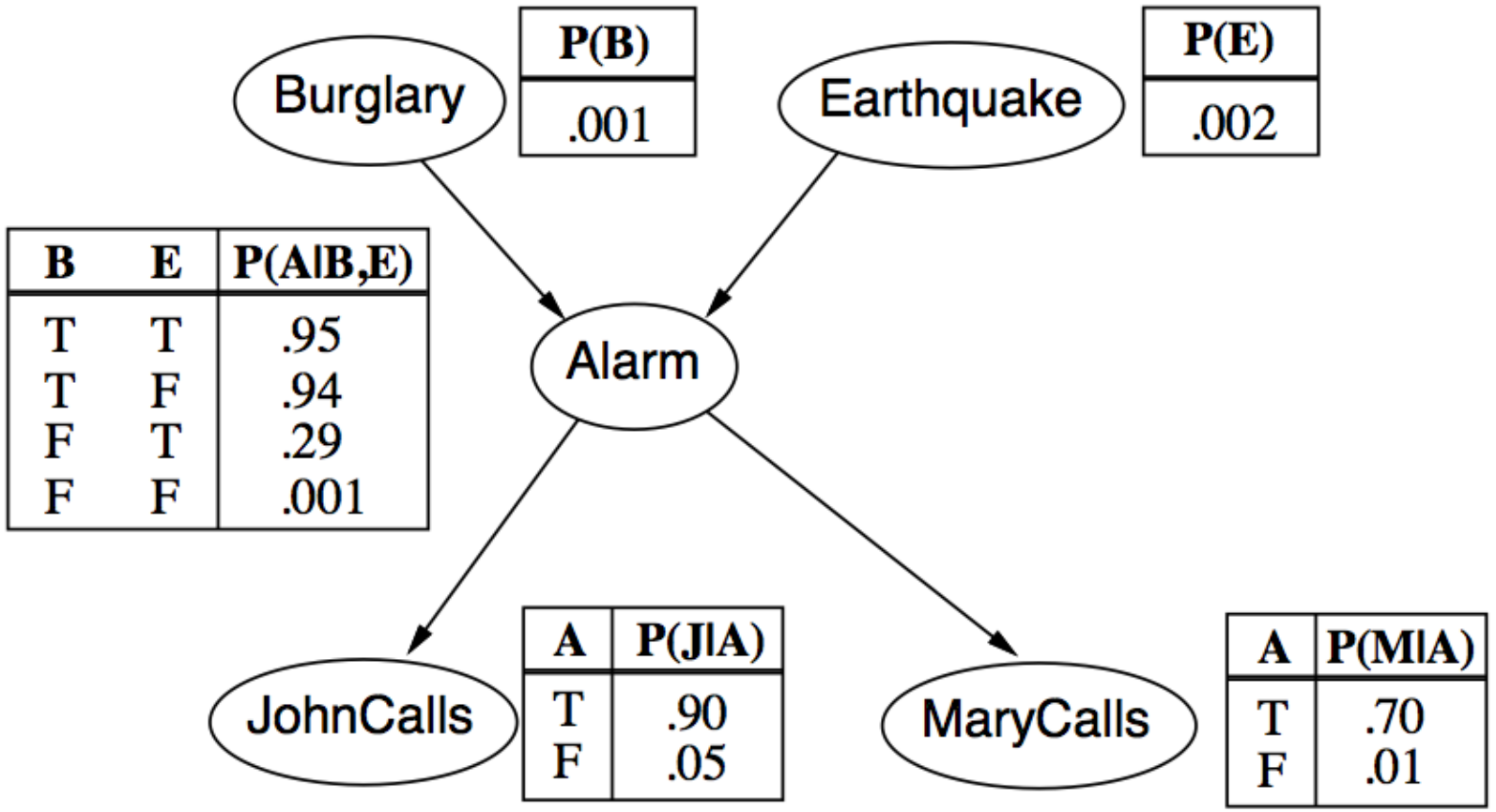


- *Weather* is independent of the other variables
- *Toothache* and *Catch* are conditionally independent given *Cavity*

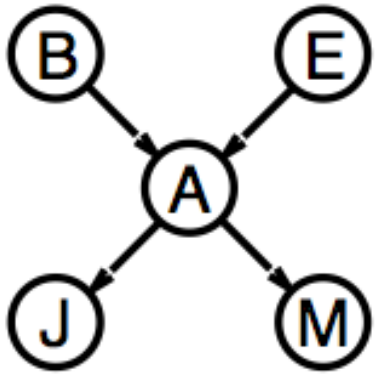
Example: Home security

- I'm at work, and my neighbor John calls to say my alarm is ringing, but my neighbor Mary doesn't call. We live in California, and sometimes the alarm is set off by minor earthquakes.
- Is there a burglar?
- Variables:
  Burglary, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*
- We have some probabilistic “causal” 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

Example: Home security



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

Semantics

- The full joint distribution is defined as the product of the local conditional distributions:

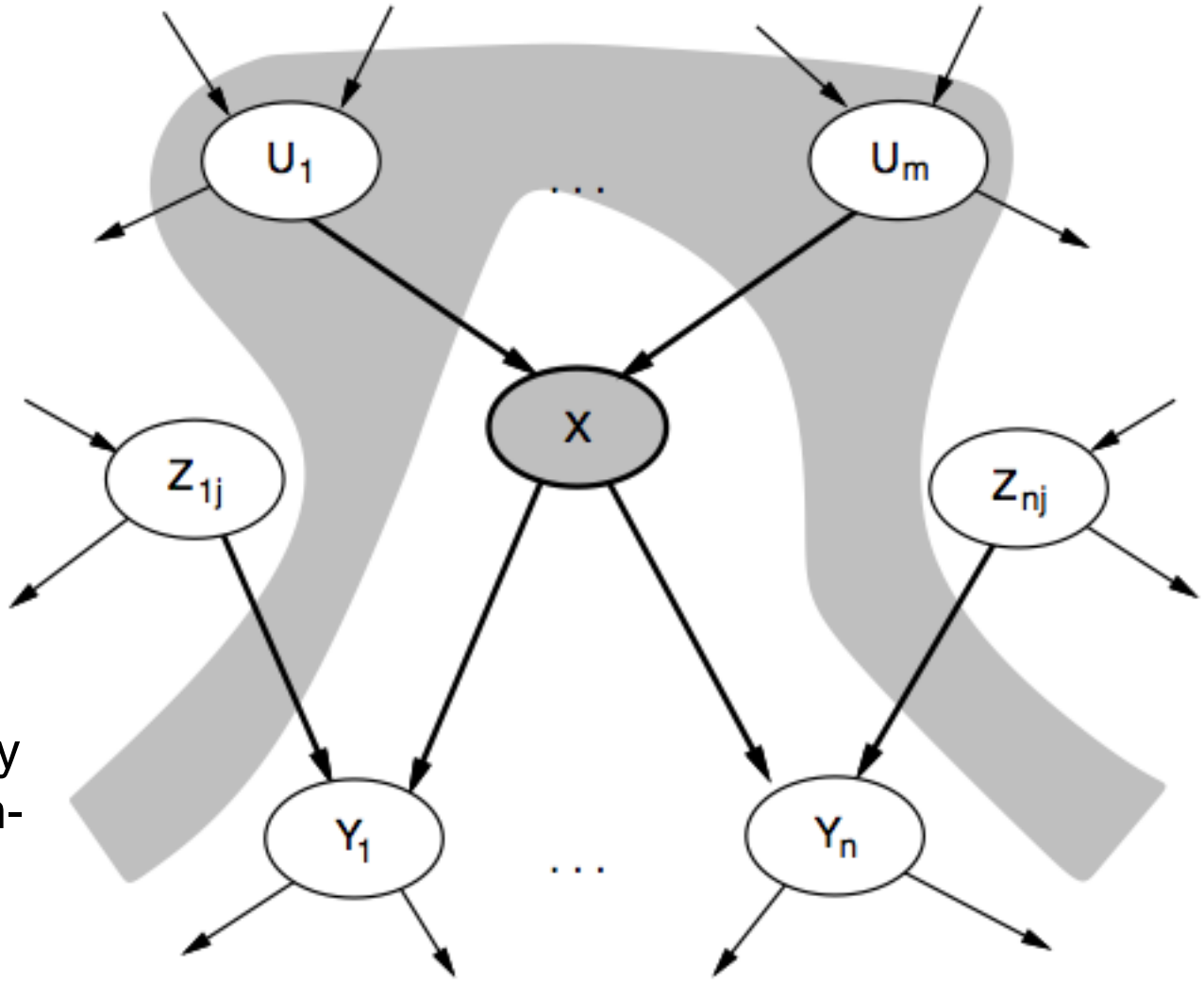
$$\mathbf{P}(X_1, \ldots, X_n) = \pi_{i=1}^{n} \mathbf{P}(X_i \mid Parents(X_i))$$

- Example

  $\mathbf{P}(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$

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

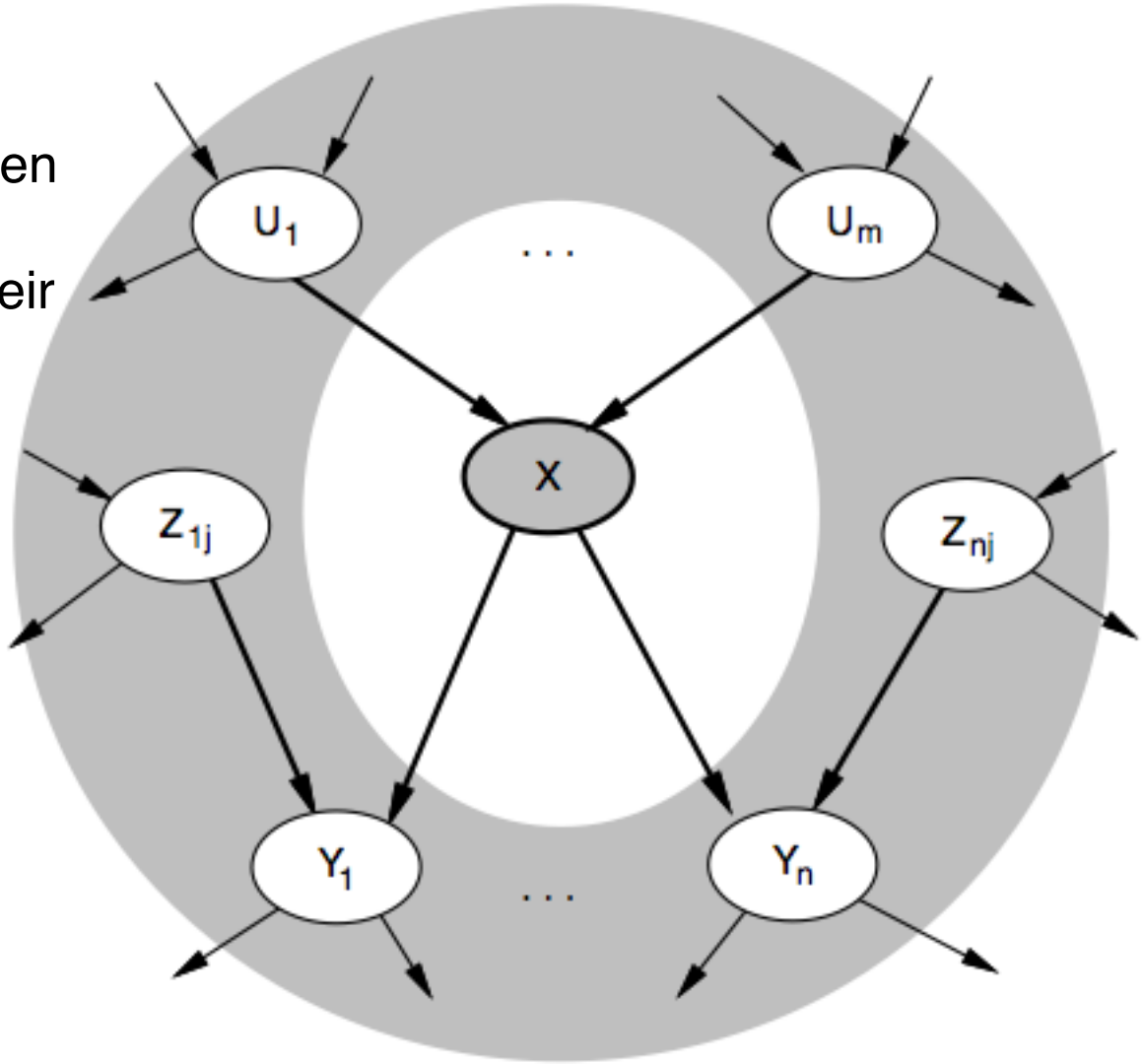
Independence



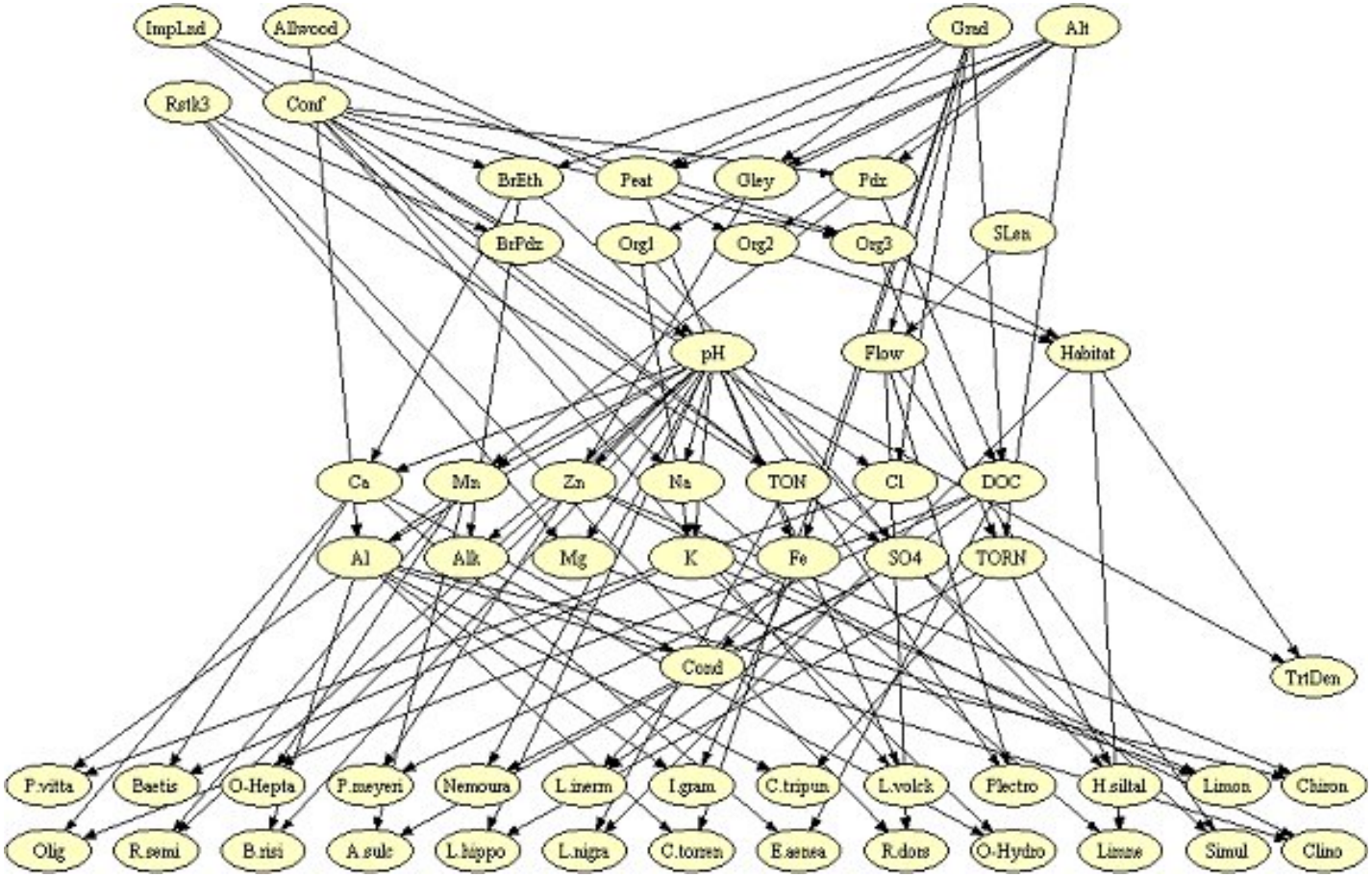
Node X is conditionally independent of its non-descendants given its parents.

Independence

Node X is conditionally independent of all other nodes in the network given its “Markov blanket” (its parents, children, and their parents).

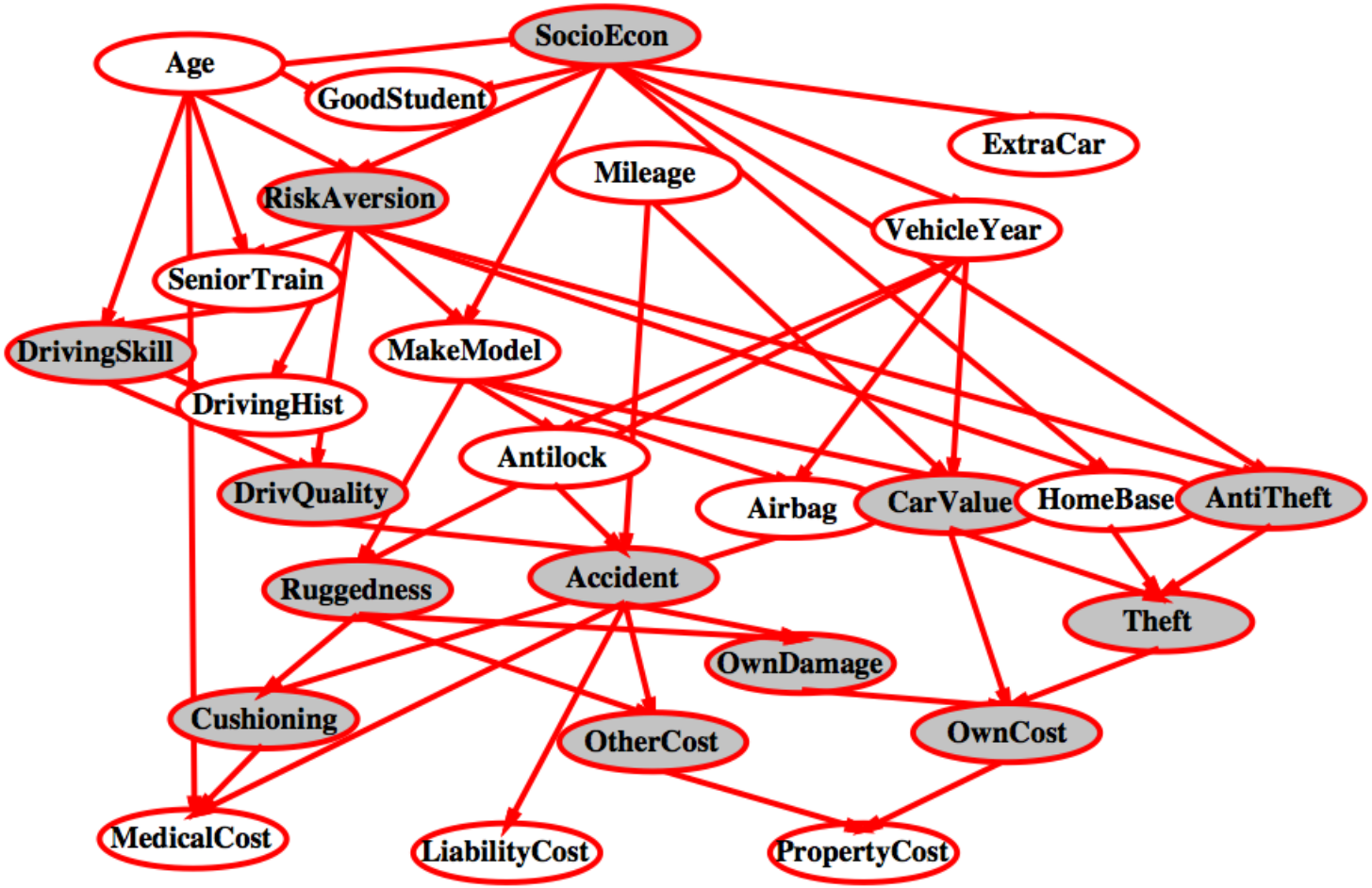


Example: River Pollution Diagnosis

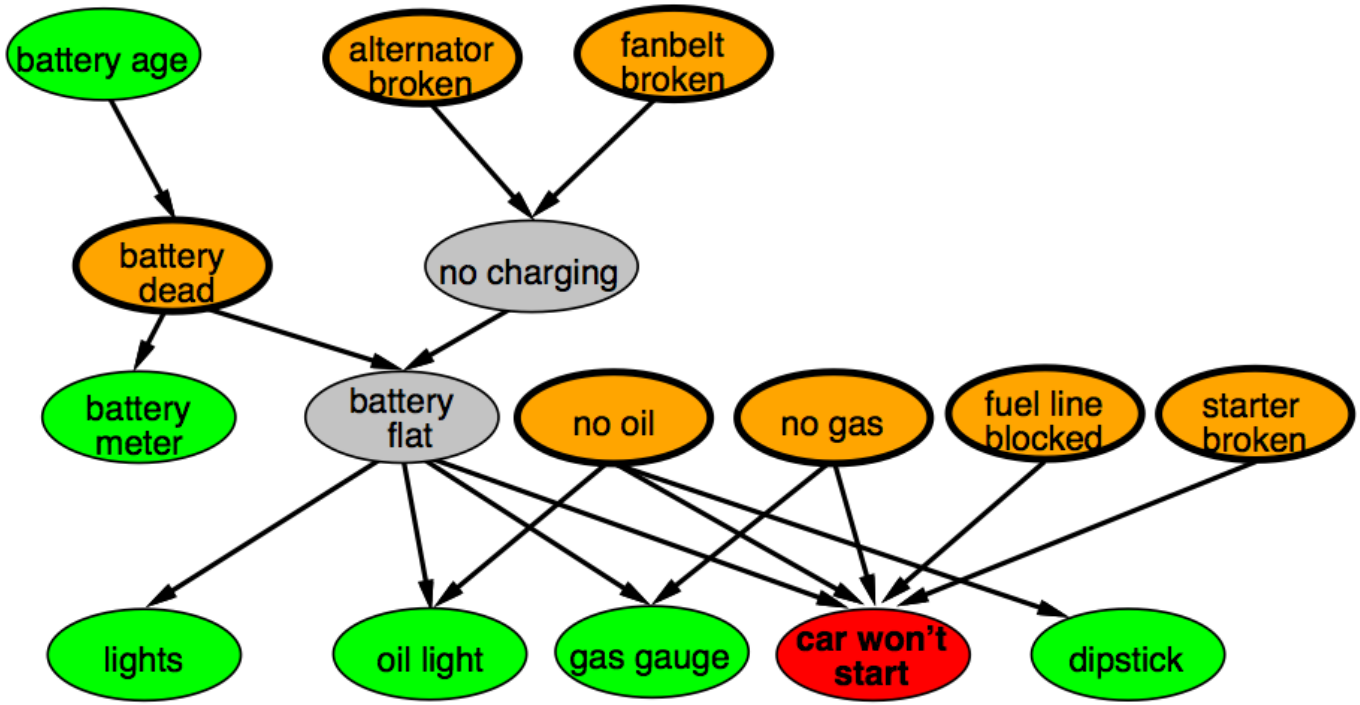


Source: http://www.soc.staffs.ac.uk/research/groups/cies2/project.htm

Example: Estimating auto insurance risk



Example: Car diagnosis



Initial evidence, Testable variables, Hidden variables

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

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

This choice of parents guarantees:

$$\mathbf{P}(X_1, \ldots, X_n) = \pi_{i=1}^{n} \mathbf{P}(X_i \mid X_1, \ldots, X_{i-1}) \quad \text{(chain rule)}$$

$$= \pi_{i=1}^{n} \mathbf{P}(X_i \mid \text{Parents}(X_i)) \quad \text{(by constr.)}$$

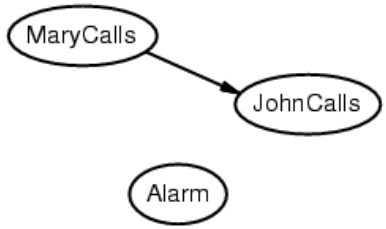
Example

- 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*



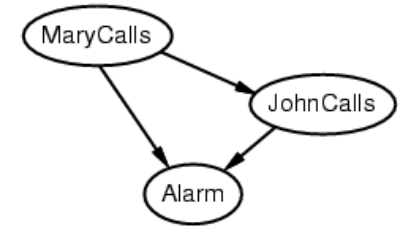
P(*J* | *M*) = ***P***(*J*)?

No

P(*A* | *J*, *M*) = ***P***(*A*)? ***P***(*A* | *J*, *M*) = ***P***(*A* | *J*)? ***P***(*A* | *J*, *M*) = ***P***(*A* | *M*)?

Example

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



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

No

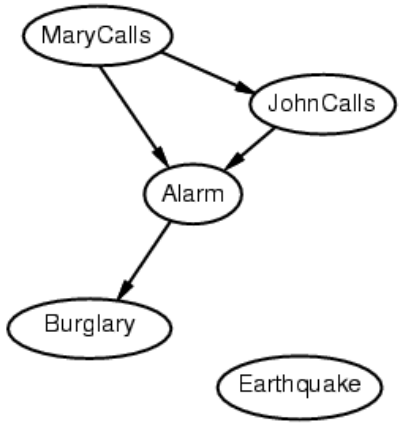
P(*A* | *J*, *M*) = ***P***(*A*)? ***P***(*A* | *J*, *M*) = ***P***(*A* | *J*)? ***P***(*A* | *J*, *M*) = ***P***(*A* | *M*)? **No**

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

P(*B* | *A*, *J*, *M*) = ***P***(*B* | *A*)? **Yes**

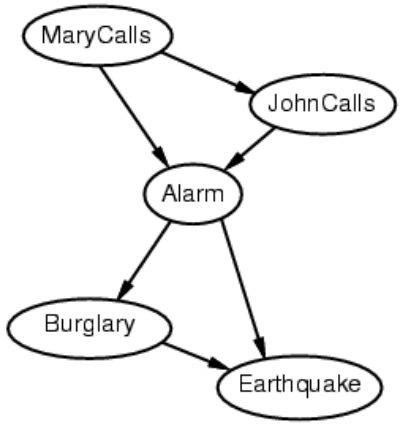
P(*B* | *A*, *J*, *M*) = ***P***(*B*)? **No**

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

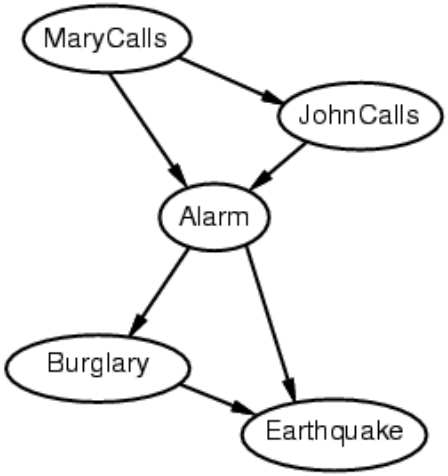
P(*B* | *A*, *J*, *M*) = ***P***(*B* | *A*)? **Yes**

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

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

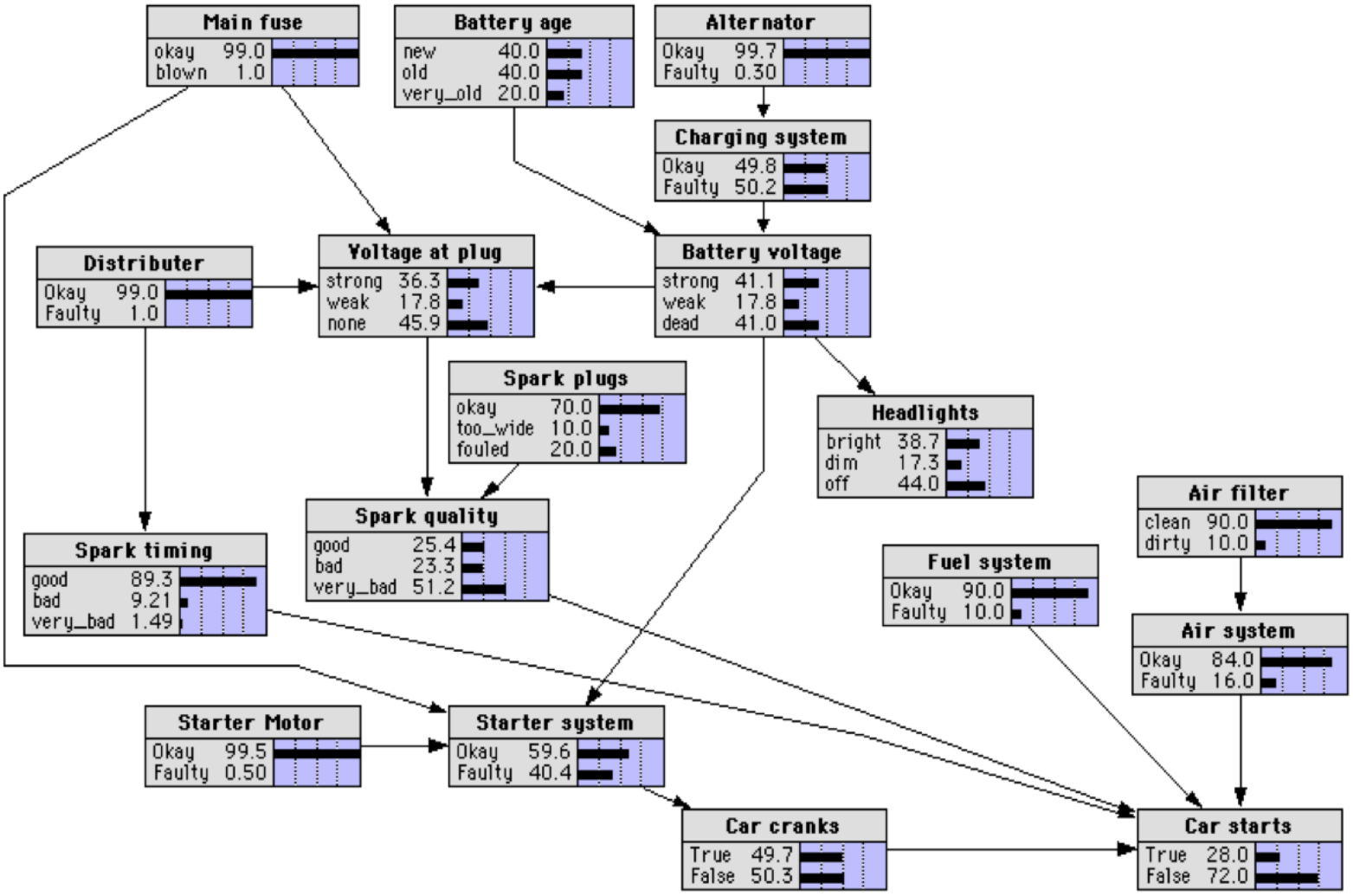
P(*E* | *B*, *A*, *J*, *M*) = ***P***(*E* | *A*, *B*)? **Yes**

Example contd.



- Deciding conditional independence is difficult in noncausal directions
- (Causal models and conditional independence seem hardwired for humans!)
- Network is less compact: 1 + 2 + 4 + 2 + 4 = 13 numbers needed

Example: Car diagnosis



Summary

- Bayesian network:
  - Directed acyclic graph whose nodes correspond to r.v.s; each note has a conditional distribution for its values given its parents.
  - Provides a concise way to represent conditional independence relations
  - Specifies the full joint distribution
  - Often exponentially smaller than explicit representation of the joint distribution

Next Class

- Inference in Bayesian Networks
- Secs. 14.4, 14.5