Chapter 14
Probabilistic Reasoning
Sections 14.1 – 14.3
Bayesian Belief Networks (BBNs)
Representation

CS5811 - Artificial Intelligence

Nilufer Onder
Department of Computer Science
Michigan Technological University

Outline

Motivation

Consider data that classifies N=800 boys with respect to boy scout status (B: true, false), juvenile delinquency (D: true, false), and socioeconomic status (S: low, medium, high).

We would like to use a scheme that allows efficient representation and reasoning of probabilistic information.

Variable B	Variable D	Variable S	Number
y	y	l	11
y	y	m	14
y	y	h	8
y	n	l	43
y	n	m	104
y	n	h	196
n	y	l	42
n	y	m	20
n	y	h	2
n	n	l	169
n	n	m	132
n	n	h	59
Total			800

Bayesian belief networks (BBNs)

A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions.

Syntax:

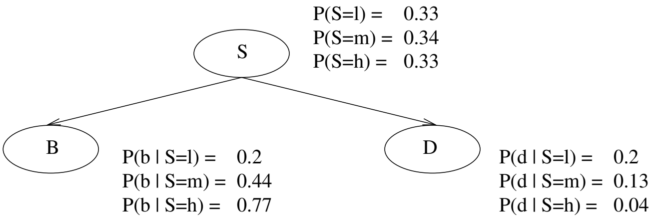
- a set of nodes
  each node represents a variable
- a directed, acyclic graph
  the existence of a link usually means "directly influences"
- a conditional distribution for each node given its parents

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

$\mathbb{P}(X_i \mid \text{Parents } (X_i))$

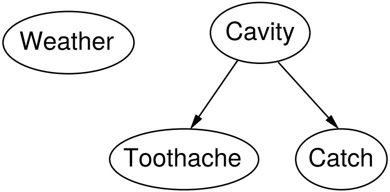
A BBN network with three variables

Suppose that after analysis, we find that juvenile delinquency (D) and boy scout status (B) are conditionally independent given socioeconomic status (S). This coincides with the intuition that socioeconomic status is the common cause for both.
We can represent this as a BBN.



Network topology

The topology of the network encodes conditional independence assertions.



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

Burglary example

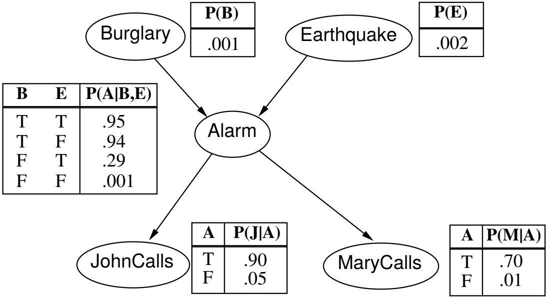
Example from Judea Pearl at UCLA:
I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes it's set off by minor earthquakes. Is there a burglar?

Variables: Burglar, Earthquake, Alarm, JohnCalls, MaryCalls

Network topology reflects "causal" knowledge:

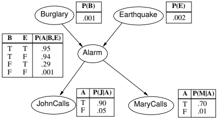
- A burglar can set the alarm off
- An earthquake can set the alarm off
- The alarm can cause Mary to call
- The alarm can cause John to call

BBN for the burglary example



Compactness

A CPT for Boolean node X_i with k Boolean parents needs 2^k rows,
one for each combination of the parent values.



Each row requires one number p for $X_i = \text{true}$.
The number for $X_i = \text{false}$ is just $1 - p$.
If each variable has no more than k parents,
the complete network requires $O(n \times 2^k)$ numbers.

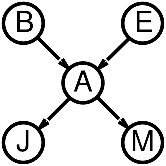
The size of the network grows linearly with n, the number of variables.

In comparison, a full joint probability distribution (JPD) table requires $O(2^n)$ rows, i.e., grows exponentially with n.
For the burglary network,
the BBN requires $1 + 1 + 4 + 2 + 2 = 10$ numbers,
the full JPD table requires $2^5 - 1 = 31$ numbers.

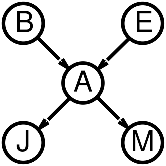
How many numbers are needed for the boy scouts BBN and table?

Semantics of Bayesian nets

In general, *semantics* = "what things mean."
Here we are interested in what a Bayesian net means.
We'll look at *global* and *local* semantics



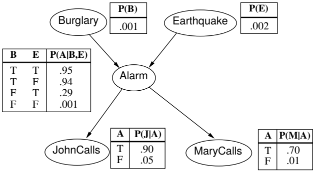
Global semantics



The *global semantics* defines the full joint distribution as the product of the local conditional distributions.
If $X_1, \ldots, X_n$ are all of the random variables, then by combining the chain rule and conditional independence, we get
$\mathbb{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \mid \text{ Parents } (X_i))$

$$\begin{aligned} \text{E.g.,} \quad & P(j \wedge m \wedge a \wedge \neg b \wedge \neg e) \\ = \; & P(j \mid m, a, \neg b, \neg e) P(m \mid a, \neg b, \neg e) P(a \mid \neg b, \neg e) P(\neg b \mid \neg e) P(\neg e) \\ = \; & P(j \mid a) P(m \mid a) P(a \mid \neg b, \neg e) P(\neg b) P(\neg e) \end{aligned}$$

Plug in the values



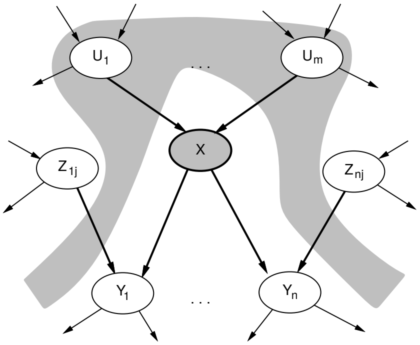
The global semantics defines the full joint distribution as the product of the local conditional distributions
$\mathbb{P}(X-1,\ldots,X_n)=\prod_{i=1}^{n}\mathbb{P}(X_i \mid \text{ Parents } (X_i))$

E.g.,

$$\begin{aligned} & P(j \wedge m \wedge a \wedge \neg b \wedge \neg e) \\ = \; & P(j \mid a)P(m \mid a)P(a \mid \neg b, \neg e)P(\neg b)P(\neg e) \\ = \; & 0.9 \times 0.7 \times 0.01 \times (1-0.001) \times (1-0.002) \\ = \; & 0.06224526 \end{aligned}$$

Local semantics

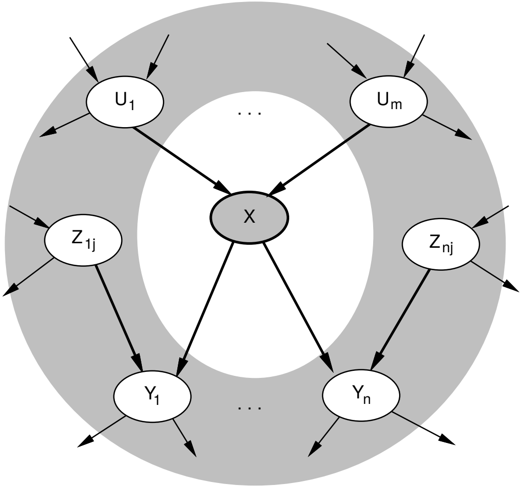
Local semantics: Each node is conditionally independent of its nondescendants given its parents



Theorem: local semantics ⇔ global semantics

Markov blanket

Each node is conditionally independent of all others given its *Markov blanket*: parents + children + children's parents



Constructing Bayesian networks

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

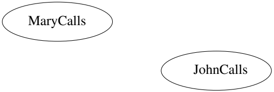
1. Choose an ordering of variables $X_1, \ldots, X_n$
   In principle *any* ordering will work
2. For $i = 1$ to n
   Add X_i to the network
   Select parents from $X_1, \ldots, X_{i-1}$ such that
   $\mathbb{P}(X_i \mid \text{Parents } (X_i)) = \mathbb{P}(X_i \mid X_1, \ldots, X_{i-1})$

This choice of parents guarantees the global semantics

$$\begin{aligned}\mathbb{P}(X_1, \ldots, X_{i-1}) &= \textstyle\prod_{i=1}^{n} \mathbb{P}(X_i | X_1, \ldots, X_{i-1}) \text{ (chain rule)} \\ &= \textstyle\prod_{i=1}^{n} \mathbb{P}(X_i | \text{ Parents } (X_i)) \text{ (by construction)}\end{aligned}$$

Example

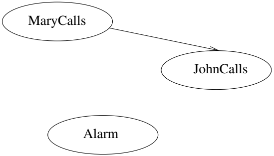
Suppose we choose the ordering M, J, A, B, E



$P(J \mid M) = P(J)$?

Example

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

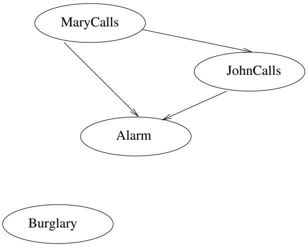


$P(J \mid M) = P(J)$? **No**

$P(A \mid J, M) = P(A \mid J)$? $\quad P(A \mid J, M) = P(A)$?

Example

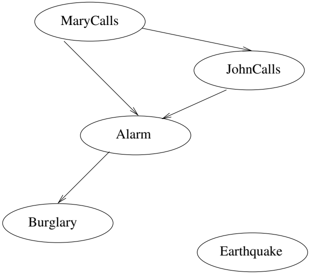
Suppose we choose the ordering M, J, A, B, E



$P(J \mid M) = P(J)$? **No**
$P(A \mid J, M) = P(A \mid J)$? $P(A \mid J, M) = P(A)$? **No**
$P(B \mid A, J, M) = P(B \mid A)$?
$P(B \mid A, J, M) = P(B)$?

Example

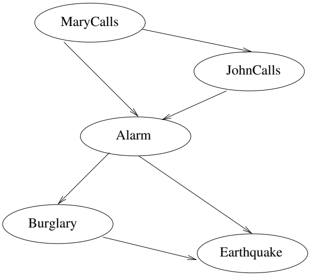
Suppose we choose the ordering M, J, A, B, E



$P(J \mid M) = P(J)$? **No**
$P(A \mid J, M) = P(A \mid J)$? $P(A \mid J, M) = P(A)$? **No**
$P(B \mid A, J, M) = P(B \mid A)$? **Yes**
$P(B \mid A, J, M) = P(B)$? **No**
$P(E \mid B, A, J, M) = P(E \mid A)$?
$P(E \mid B, A, J, M) = P(E \mid A, B)$?

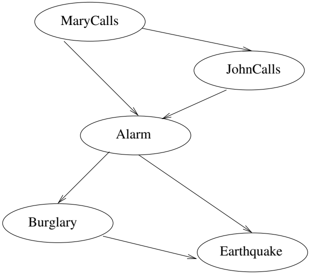
Example

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



$P(J \mid M) = P(J)$? **No**
$P(A \mid J, M) = P(A \mid J)$? $\quad P(A \mid J, M) = P(A)$? **No**
$P(B \mid A, J, M) = P(B \mid A)$? **Yes**
$P(B \mid A, J, M) = P(B)$? **No**
$P(E \mid B, A, J, M) = P(E \mid A)$? **No**
$P(E \mid B, A, J, M) = P(E \mid A, B)$? **Yes**

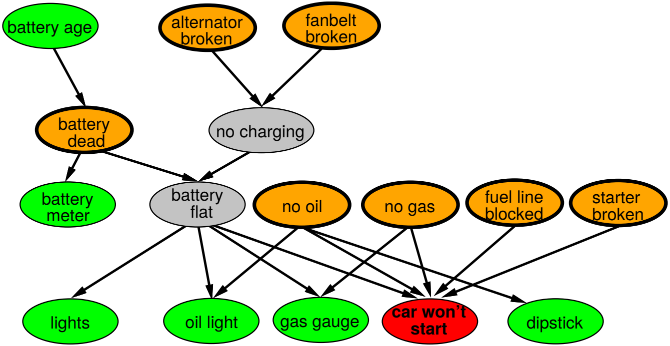
Example



Deciding conditional independence is hard in noncausal directions.
(Causal models and conditional independence seem hardwired for humans!)
Assessing conditional probabilities is hard in noncausal directions.
Network is less compact: $1 + 2 + 4 + 2 + 4 = 13$ numbers needed, rather than 10.

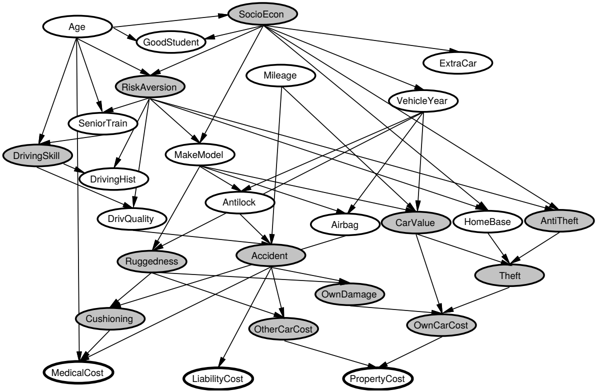
Car diagnosis example

Initial evidence: car won't start
Green variables are "testable variables"
Orange variables are "broken, so fix it variables"
Gray variables are "hidden variables" to ensure sparse structure and reduce parameters



Car insurance example

Estimating the expected claim costs for a policy holder:
MedicalCost, LiabilityCost, PropertyCost
Unshaded variables are the data on the application form
Gray variables are "hidden variables"



Sources for the slides

- AIMA textbook (3^{rd} edition)
- Dana Nau's CMSC421 slides. 2010. http://www.cs.umd.edu/~nau/cmsc421/chapter14a.pdf
- Penn State online Stat 504 – Analysis of Discrete Data course. https://onlinecourses.science.psu.edu/stat504/print/book/export/html/112