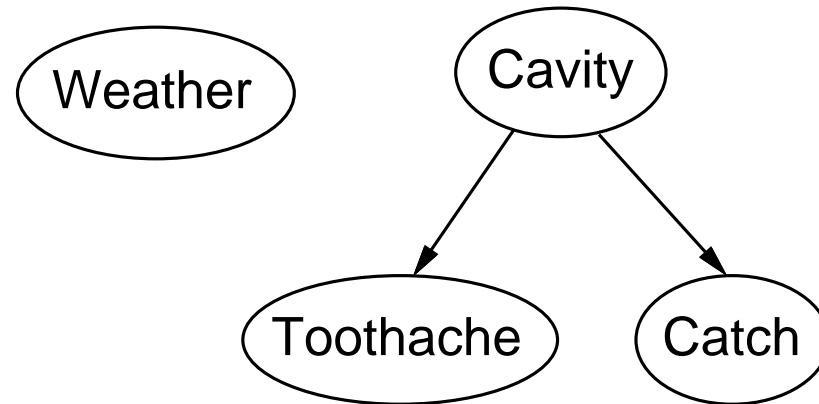


REASONING WITH BAYESIAN NETWORKS

Example of a Bayesian network



Weather is independent of the other variables

Toothache and *Catch* are conditionally independent given *Cavity*

Topology of network encodes conditional independence assertions.

Bayesian networks

Syntax:

a set of nodes, one per variable

a directed, acyclic graph (link \approx “directly influences”)

a conditional distribution for each node given its parents:

$$\mathbf{P}(X_i | Parents(X_i))$$

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

Bayesian nets provide a simple, graphical notation for conditional independence assertions.

Therefore, Bayesian networks provide a compact specification of full joint distributions.

Example

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?

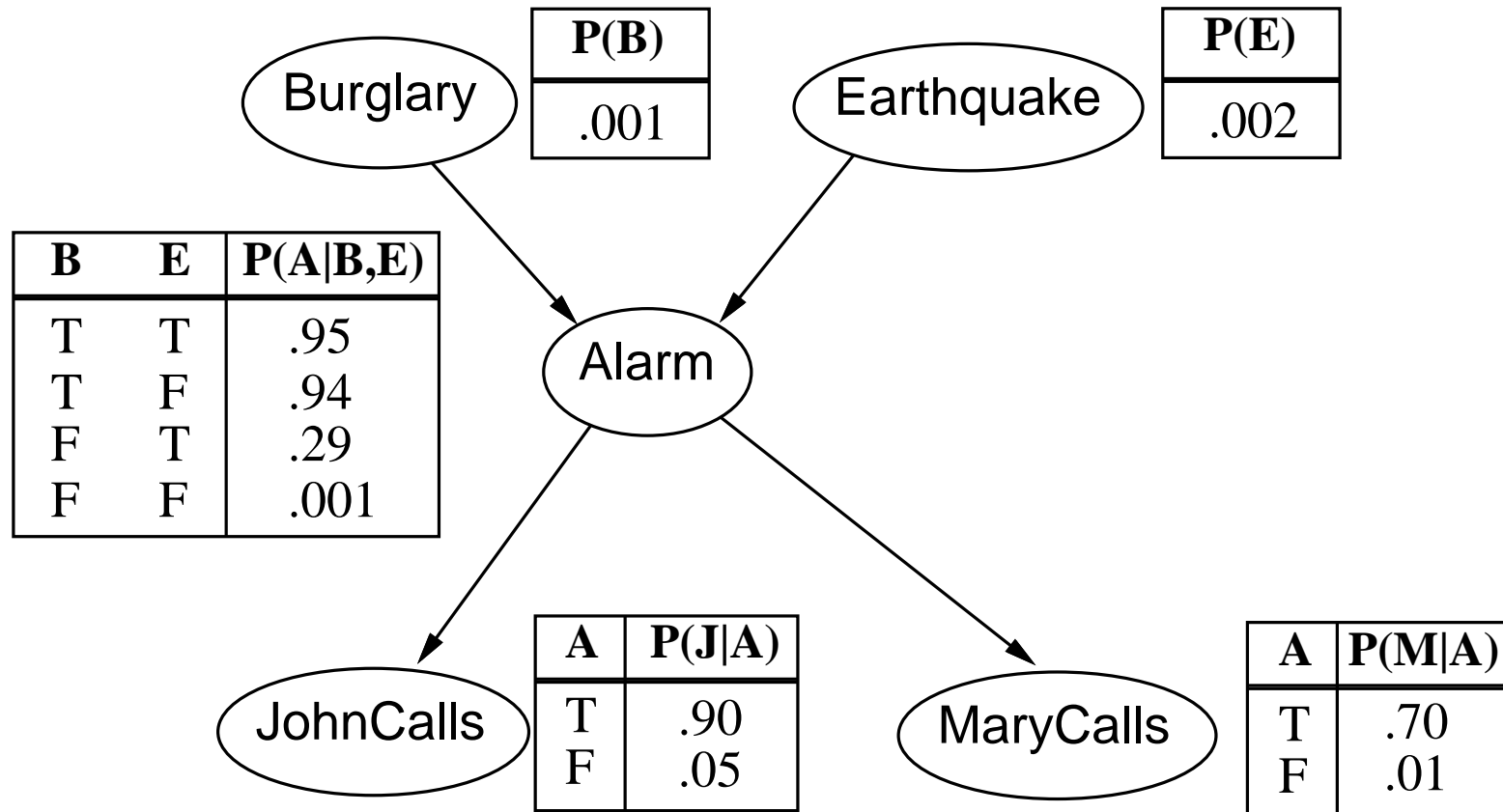
Variables: *Burglar*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*

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

This causal knowledge is incorporated into network topology.

Example contd.



Compactness

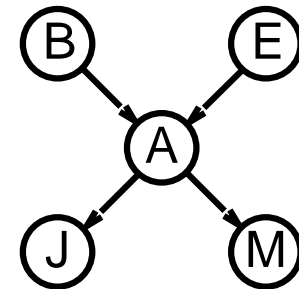
A CPT for Boolean X_i with k Boolean parents has 2^k rows for the combinations of parent values

Each row requires one number p for $X_i = \text{true}$ (the number for $X_i = \text{false}$ is just $1 - p$)

If each variable has no more than k parents, the complete n -node 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 of Bayesian nets

We have described what a network is, not what it *means*.

Two ways of understanding semantics:

Global semantics: network represents the full joint distribution

Local semantics: network encodes assumptions about conditional independence

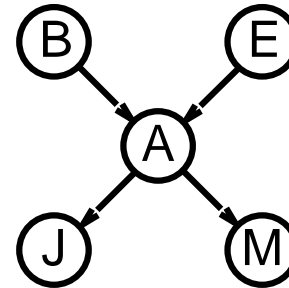
Global semantics

Full joint distribution defined as the product of the local conditional distributions:

$$\mathbf{P}(X_1, \dots, X_n) = \prod_{i=1}^n \mathbf{P}(X_i | \text{Parents}(X_i))$$

e.g., $P(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$

=



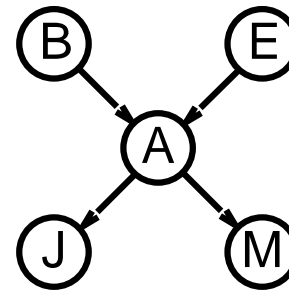
Global semantics

Full joint distribution defined as the product of the local conditional distributions:

$$\mathbf{P}(X_1, \dots, X_n) = \prod_{i=1}^n \mathbf{P}(X_i | \text{Parents}(X_i))$$

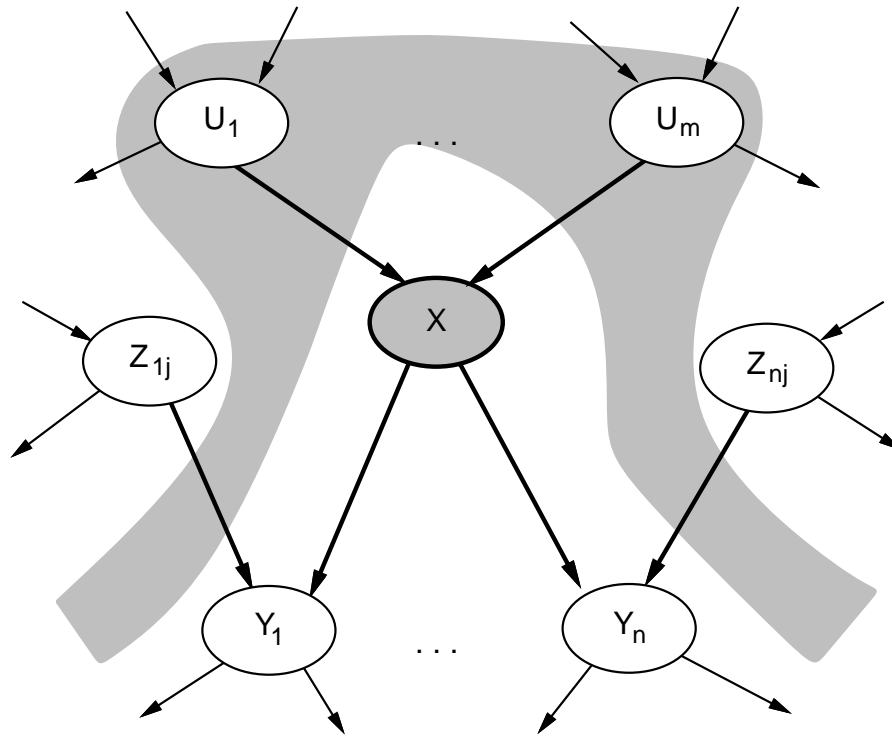
e.g., $P(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$

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



Local semantics

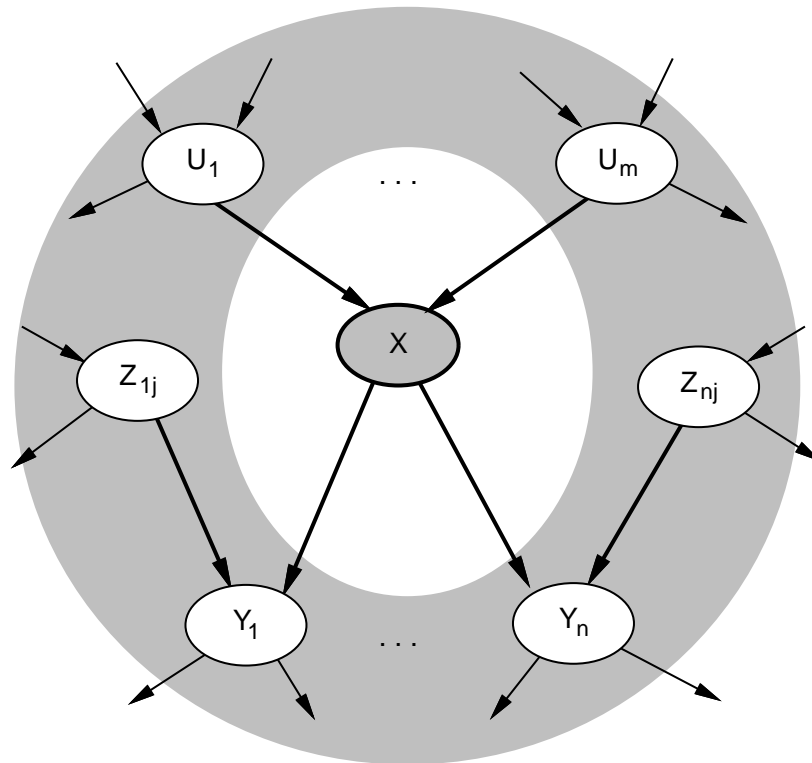
Each node is conditionally independent of its nondescendants given its parents.



Theorem: Local semantics \Leftrightarrow global semantics

Markov blanket

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



Constructing Bayesian networks

Here is a method of constructing a network that guarantees global semantics, yet requires only locally testable assertions of conditional independence.

1. Choose an ordering of variables X_1, \dots, X_n
2. For $i = 1$ to n
 - add X_i to the network
 - select parents from X_1, \dots, X_{i-1} such that
$$\mathbf{P}(X_i | \text{Parents}(X_i)) = \mathbf{P}(X_i | X_1, \dots, X_{i-1})$$

This choice of parents guarantees the global semantics:

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

Ordering of variables is very important.

Example

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

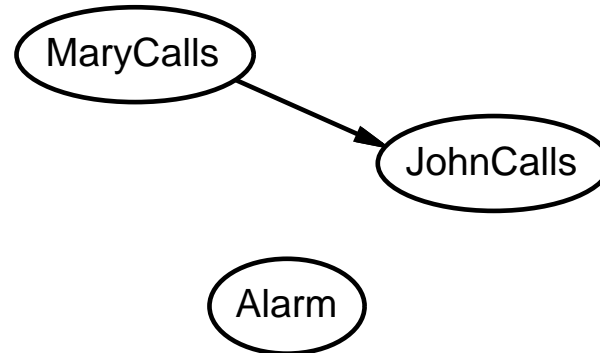
MaryCalls

JohnCalls

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

Example

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

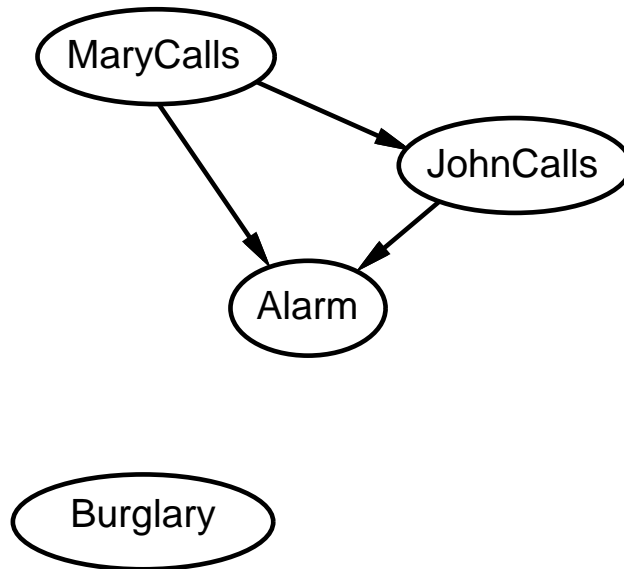


$P(J|M) = P(J)$? No

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

Example

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



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

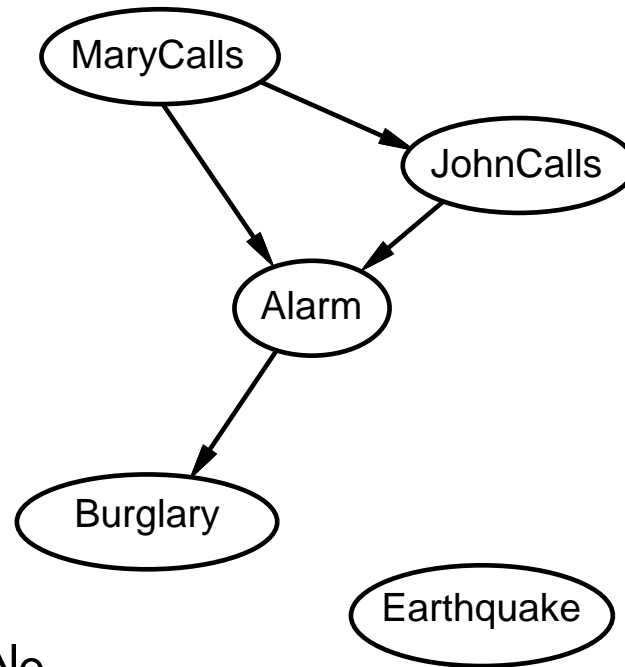
$P(A|J, M) = P(A|J)$? $P(A|J, M) = P(A)$? No

$P(B|A, J, M) = P(B|A)$?

$P(B|A, J, M) = P(B)$?

Example

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



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

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

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

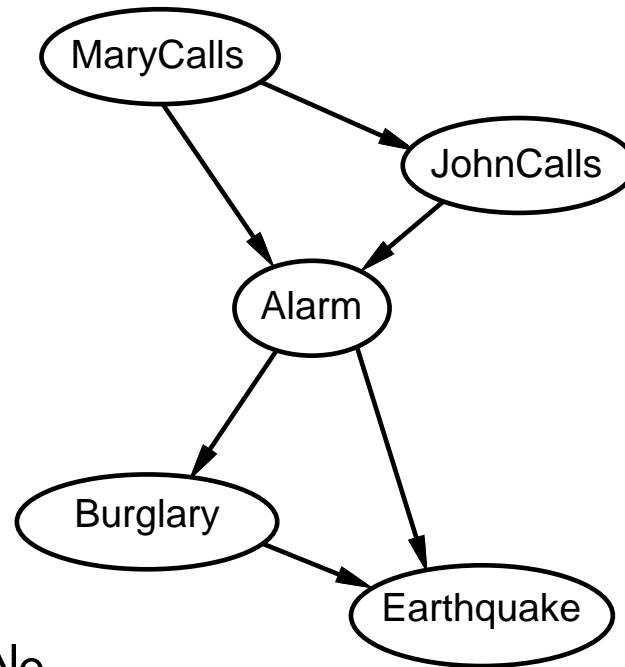
$P(B|A, J, M) = P(B)$? No

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

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

Example

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



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

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

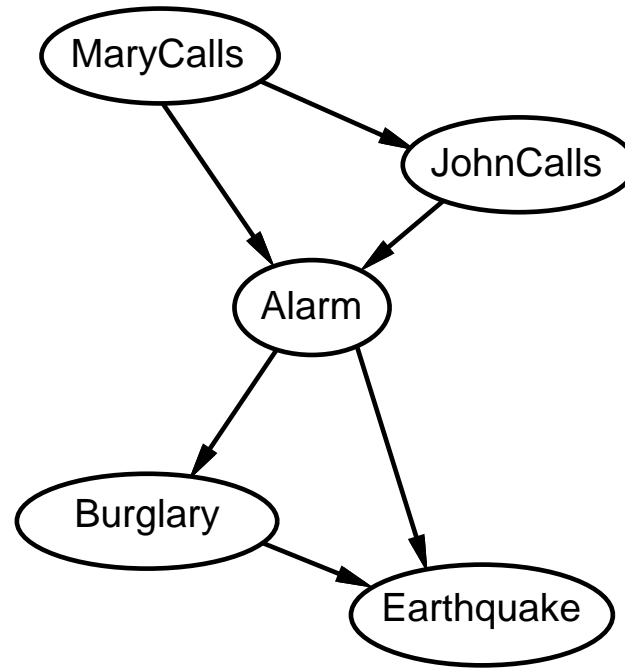
$P(B|A, J, M) = P(B|A)$? Yes

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

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

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

Problems with this network



Network is less compact: $1 + 2 + 4 + 2 + 4 = 13$ numbers needed

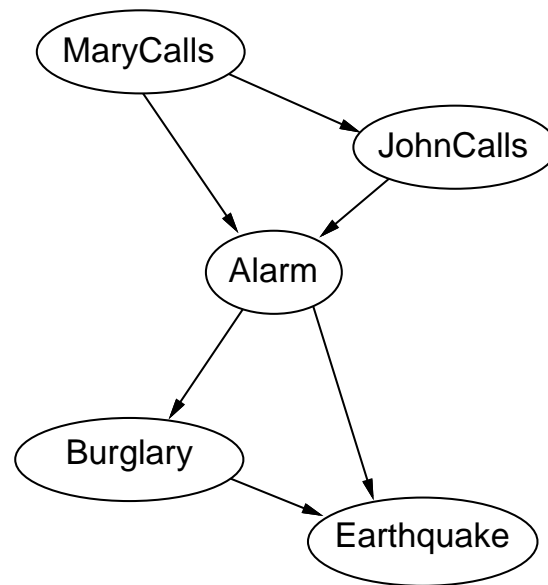
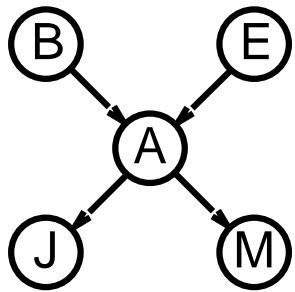
Deciding conditional independence is hard in noncausal directions

(Causal models and conditional independence seem hardwired for humans!)

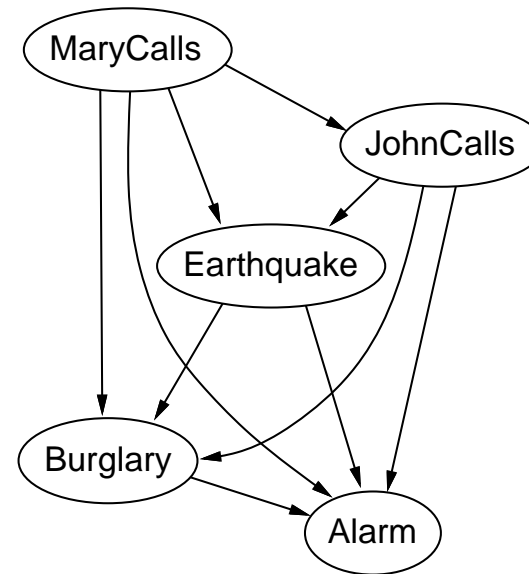
Assessing conditional probabilities is hard in noncausal directions

Ordering of variables

Because it is easier for us to apply knowledge in the causal direction, the correct order to add nodes is to start with "root causes", then add the variables that they influence, and so on, until we reach the variables that have no direct causal influence on other variables.



(a)



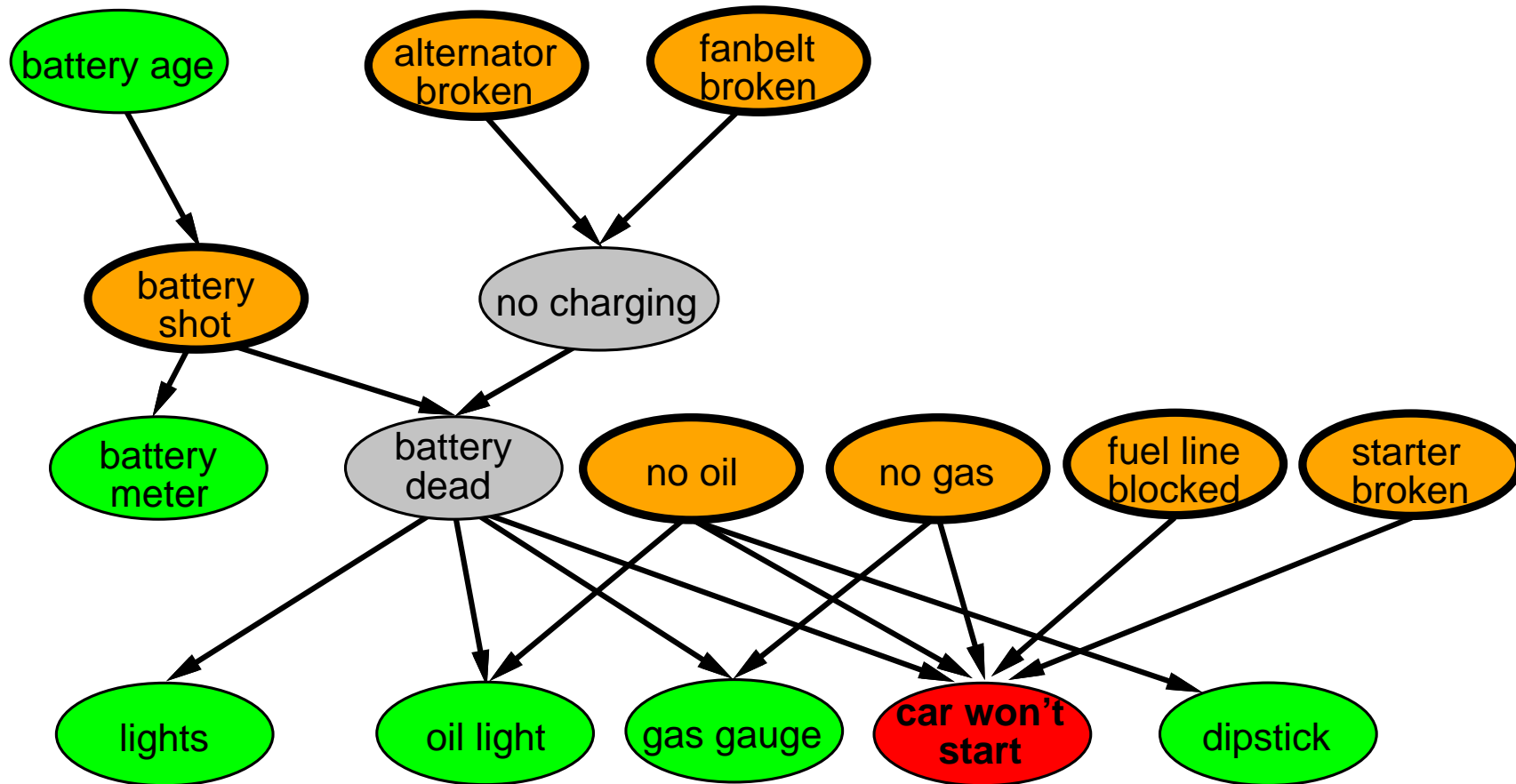
(b)

Example: Car diagnosis

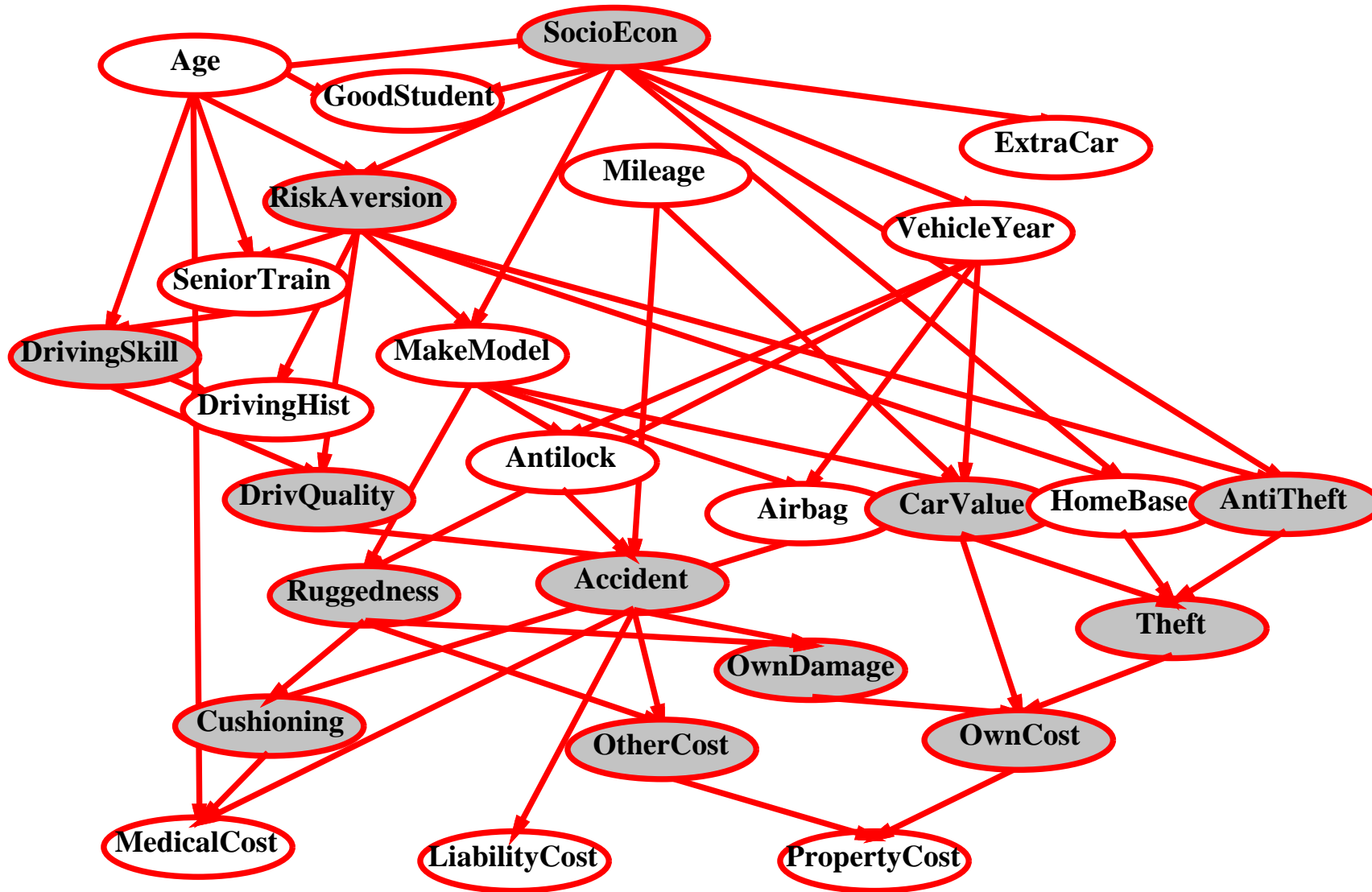
Initial evidence: car won't start

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

Hidden variables (gray) ensure sparse structure, reduce parameters



Example: Car insurance



Compact conditional distributions

CPT grows exponentially with number of parents

CPT becomes infinite with continuous-valued parent or child

Solution: **canonical** distributions that are defined compactly

Deterministic conditional distribution

Deterministic nodes are the simplest case:

$$X = f(\text{Parents}(X)) \text{ for some function } f$$

E.g., Boolean functions

$$\text{NorthAmerican} \Leftrightarrow \text{Canadian} \vee \text{US} \vee \text{Mexican}$$

E.g., numerical relationships among continuous variables

$$\frac{\partial \text{Level}}{\partial t} = \text{inflow} + \text{precipitation} - \text{outflow} - \text{evaporation}$$

Noisy-OR conditional distribution

Noisy-OR distributions model multiple, noninteracting, unreliable causes.

Cold alone causes a fever with probability .4.

Flu alone causes a fever with probability .8.

Malaria alone causes a fever with probability .9.

$$P(\text{fever}|\text{cold}, \neg\text{flu}, \neg\text{malaria}) = .4$$

$$P(\text{fever}|\neg\text{cold}, \text{flu}, \neg\text{malaria}) = .8$$

$$P(\text{fever}|\neg\text{cold}, \neg\text{flu}, \text{malaria}) = .9$$

Cold and flu together *fail* to cause a fever only if cold alone fails and flu alone fails.

$$P(\neg\text{fever}|\text{cold}, \text{flu}, \neg\text{malaria}) =$$

$$P(\neg\text{fever}|\neg\text{cold}, \neg\text{flu}, \text{malaria})P(\neg\text{fever}|\text{cold}, \neg\text{flu}, \neg\text{malaria})$$

$$P(\text{fever}|\text{cold}, \text{flu}, \neg\text{malaria}) = 1 -$$

$$[1 - P(\text{fever}|\neg\text{cold}, \neg\text{flu}, \text{malaria})][1 - P(\text{fever}|\text{cold}, \neg\text{flu}, \neg\text{malaria})]$$

Noisy OR

Formally, consider all *potential* causes $U_1 \dots U_k$ of some effect X (can add *leak node*).

Some subset of causes (e.g., $1 \dots j$) are *present*, i.e., $u_1 \dots u_j, \neg u_{j+1} \dots \neg u_k$.

r_i : probability that (present) cause i alone produces effect

Effect fails to occur if all present causes *independently* fail to produce effect.

$$P(\neg x | u_1 \dots u_j, \neg u_{j+1} \dots \neg u_k) = \prod_{i=1}^j (1 - r_i)$$

$$P(x | u_1 \dots u_j, \neg u_{j+1} \dots \neg u_k) = 1 - \prod_{i=1}^j (1 - r_i)$$

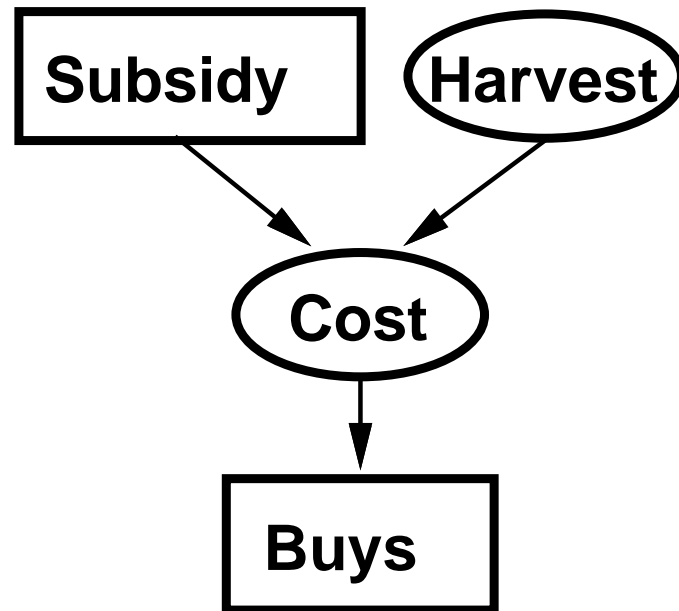
Noisy OR

<i>Cold</i>	<i>Flu</i>	<i>Malaria</i>	$P(\text{fever} \text{Cold}, \text{Flu}, \text{Malaria})$	$P(\neg \text{fever} \text{Cold}, \text{Flu}, \text{Malaria})$
F	F	F	0.0	1.0
F	F	T	0.9	0.1
F	T	F	0.8	0.2
F	T	T	0.98	$0.02 = 0.2 \times 0.1$
T	F	F	0.4	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

Hybrid (discrete+continuous) networks

Discrete (*Subsidy* and *Buys*); continuous (*Harvest* and *Cost*)



Option 1: discretization—possibly large errors, large CPTs

Option 2: finitely parameterized canonical families

- 1) Continuous variable, discrete+continuous parents (e.g., *Cost*)
- 2) Discrete variable, continuous parents (e.g., *Buys*)

Continuous child variables

Need one **conditional density** function for child variable given continuous parents, for each possible assignment to discrete parents

Most common is the **linear Gaussian** model, e.g.,

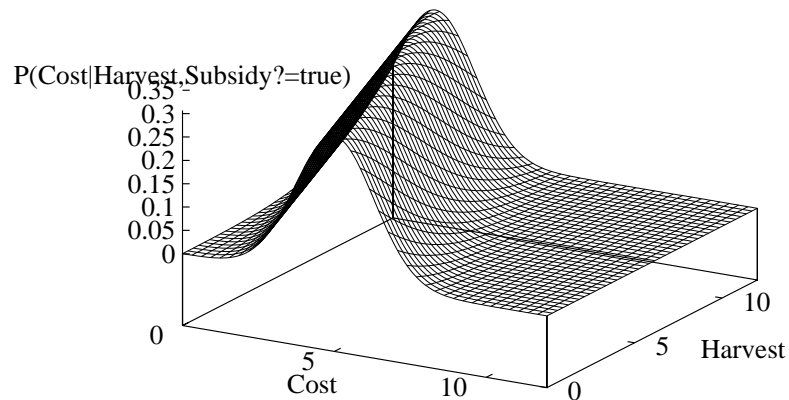
$$\begin{aligned} P(\text{Cost} = c | \text{Harvest} = h, \text{Subsidy} = \text{true}) \\ &= N(a_t h + b_t, \sigma_t)(c) \\ &= \frac{1}{\sigma_t \sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{c - (a_t h + b_t)}{\sigma_t}\right)^2\right) \end{aligned}$$

Mean *Cost* varies linearly with *Harvest*, variance is fixed

Linear variation is unreasonable over the full range

but works OK if the **likely** range of *Harvest* is narrow

Continuous child variables



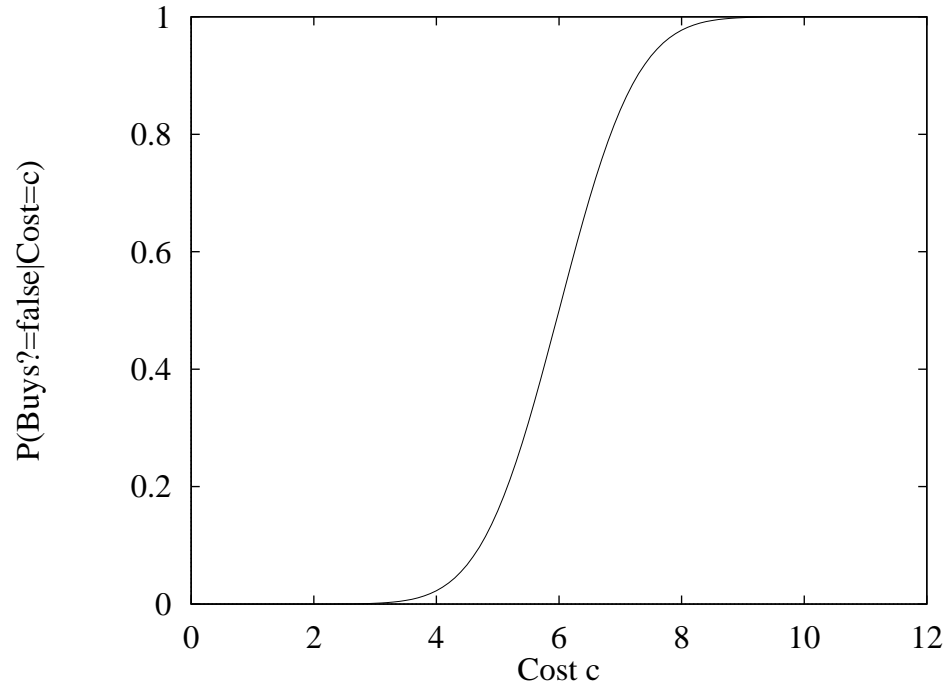
All-continuous network with LG distributions

⇒ full joint distribution is a multivariate Gaussian

Discrete+continuous LG network is a **conditional Gaussian** network i.e., a multivariate Gaussian over all continuous variables for each combination of discrete variable values

Discrete variable w/ continuous parents

Probability of *Buys* given *Cost* should be a “soft” threshold:



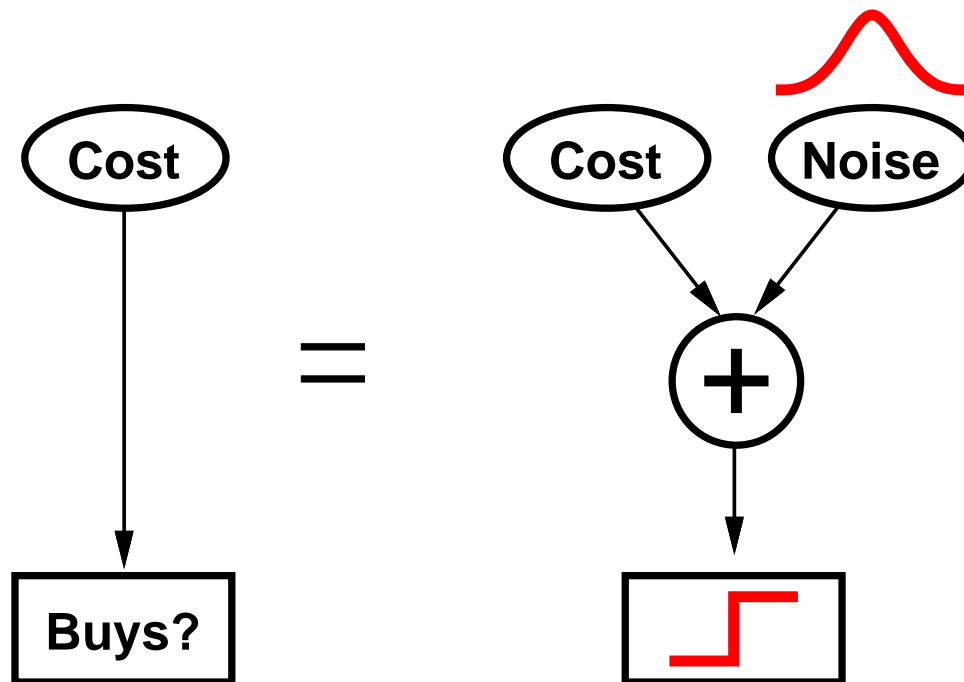
Probit distribution uses integral of Gaussian:

$$\Phi(x) = \int_{-\infty}^x N(0,1)(x)dx$$

$$P(\text{Buys} = \text{true} \mid \text{Cost} = c) = \Phi((-c + \mu)/\sigma)$$

Why the probit?

1. It's sort of the right shape
2. Can view as hard threshold whose location is subject to noise

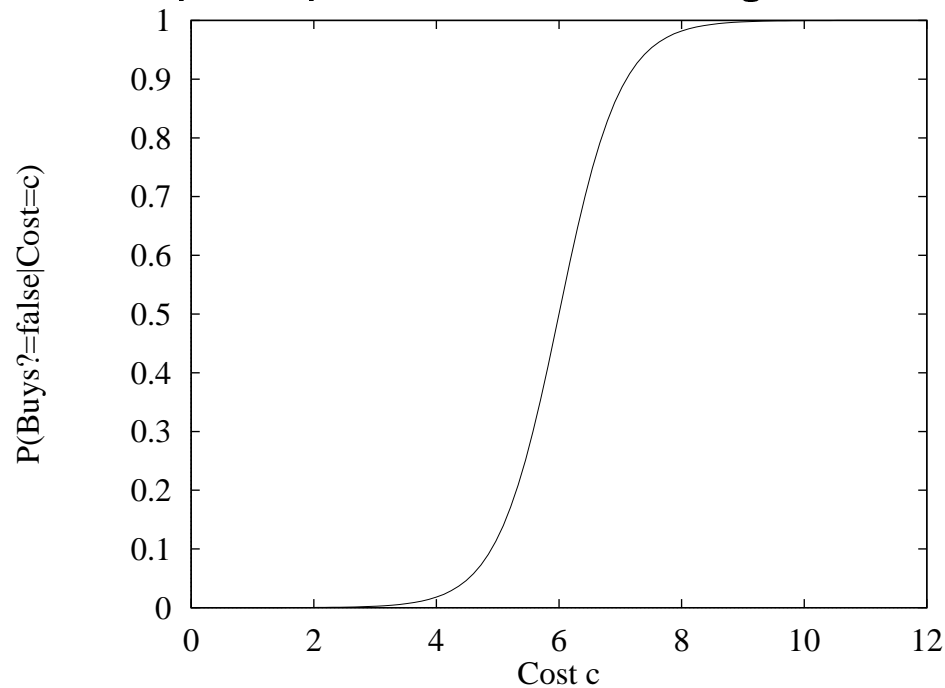


Discrete variable contd.

Sigmoid (or logit) distribution also used in neural networks:

$$P(\text{Buys} = \text{true} \mid \text{Cost} = c) = \frac{1}{1 + \exp\left(-2\frac{-c+\mu}{\sigma}\right)}$$

Sigmoid has similar shape to probit but much longer tails:



Summary

Given causal knowledge, Bayes nets provide a natural representation for encoding conditional independence.

Topology + CPTs = compact representation of joint distribution

Generally easy for (non)experts to construct

Canonical distributions (e.g., noisy-OR) = compact representation of CPTs

Continuous variables \Rightarrow parameterized distributions (e.g., linear Gaussian)