EECS 391
Intro to AI

Bayesian Networks

(aka: Belief Nets, Bayesian Belief Nets,
or more generally directed graphical models)

L13+14 Thu Oct 19 & Thu Oct 26

Recap of inference with small number of variables

- Probability: *precise representation of uncertainty*
- Probability theory: *optimal updating of knowledge based on new information*

- Bayesian Inference with Boolean variables

likelihood *prior*

posterior $$P(D|T) = \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|\bar{D})P(\bar{D})}$$

normalizing constant

- Inferences combines sources of knowledge

$$P(D|T) = \frac{0.9 \times 0.001}{0.9 \times 0.001 + 0.1 \times 0.999} = 0.0089$$

- Inference is sequential

$$P(D|T_1, T_2) = \frac{P(T_2|D)P(T_1|D)P(D)}{P(T_2)P(T_1)}$$

How do you model a world?

How to you reason about it?

Today: Inference with more complex dependencies

- How do we represent (model) more complex probabilistic relationships?
- How do we use these models to draw inferences?

The wet grass example

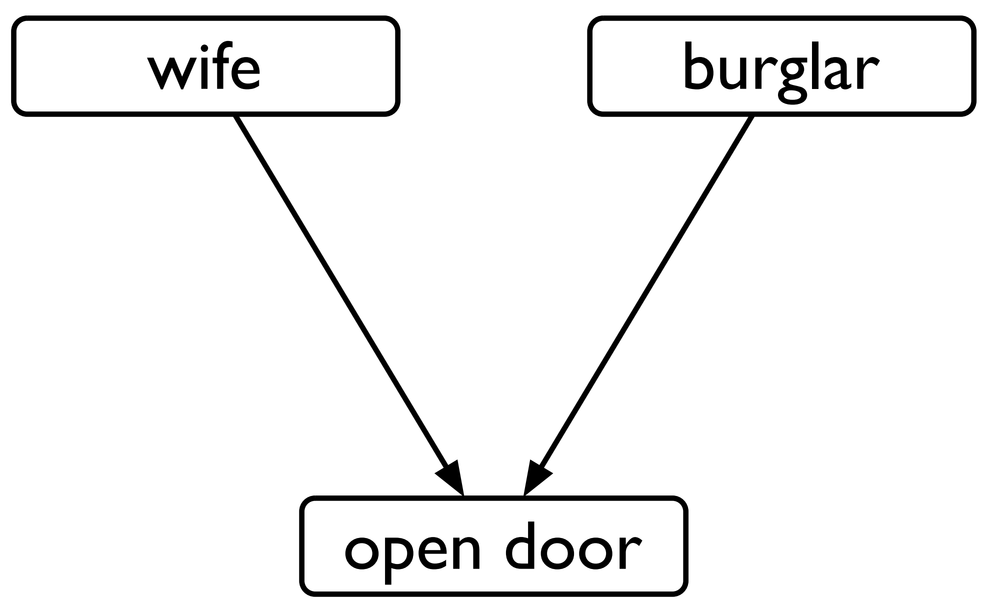
- A women leaves her house and notices the grass is wet.
- Did she forget to turn off the sprinkler?

Probabilistic reasoning

- I go home and notice that the front door is open.
  - Is it a burglar? Should I go in or call the police?
- *How should we represent these possibilities?*

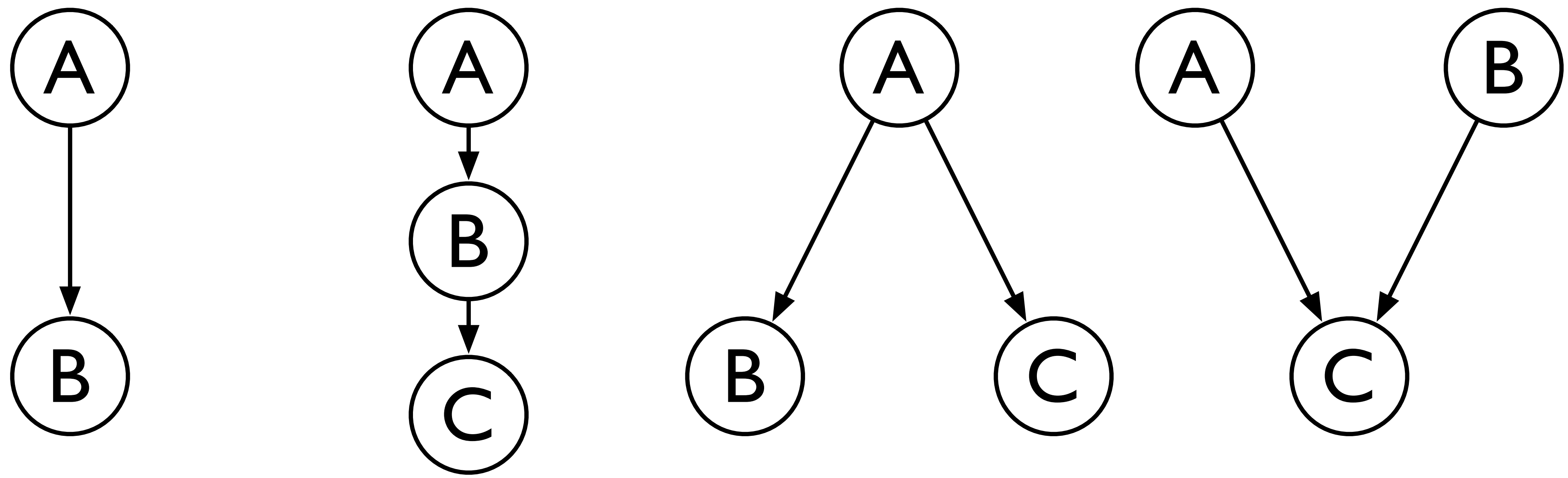
Belief networks

- In Belief networks, *causal relationships* are represented in directed acyclic graphs.
- Arrows indicate causal relationships between the nodes.



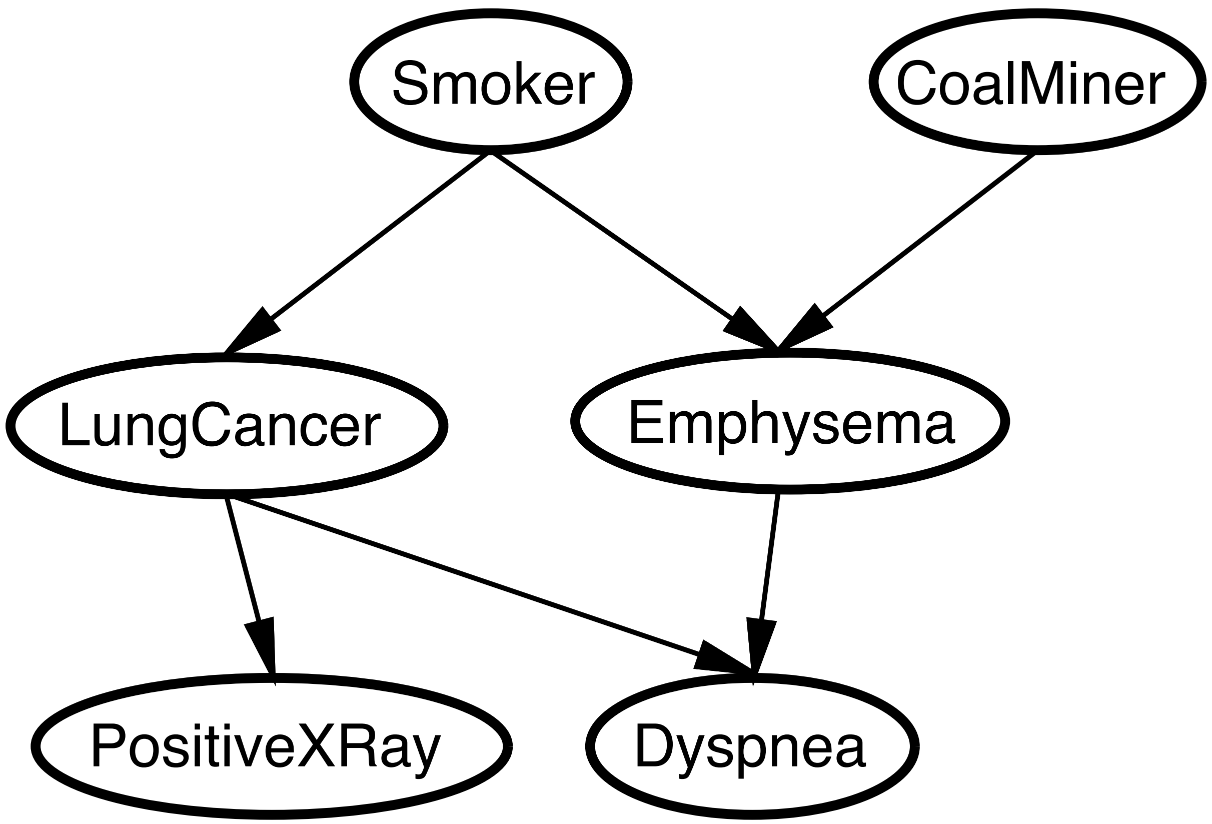
Types of probabilistic relationships

- How do we represent these relationships?

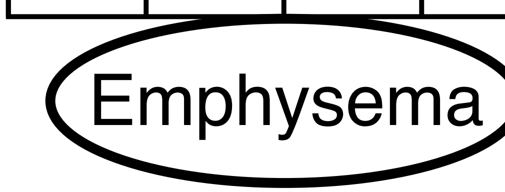


Direct cause	Indirect cause	Common cause	Common effect
P(B|A)	P(B|A) P(C|B)	P(B|A) P(C|A)	P(C|A,B)
	C is *independent* of A given B	Are B and C independent?	Are A and B independent?

Another example of a Bayesian network

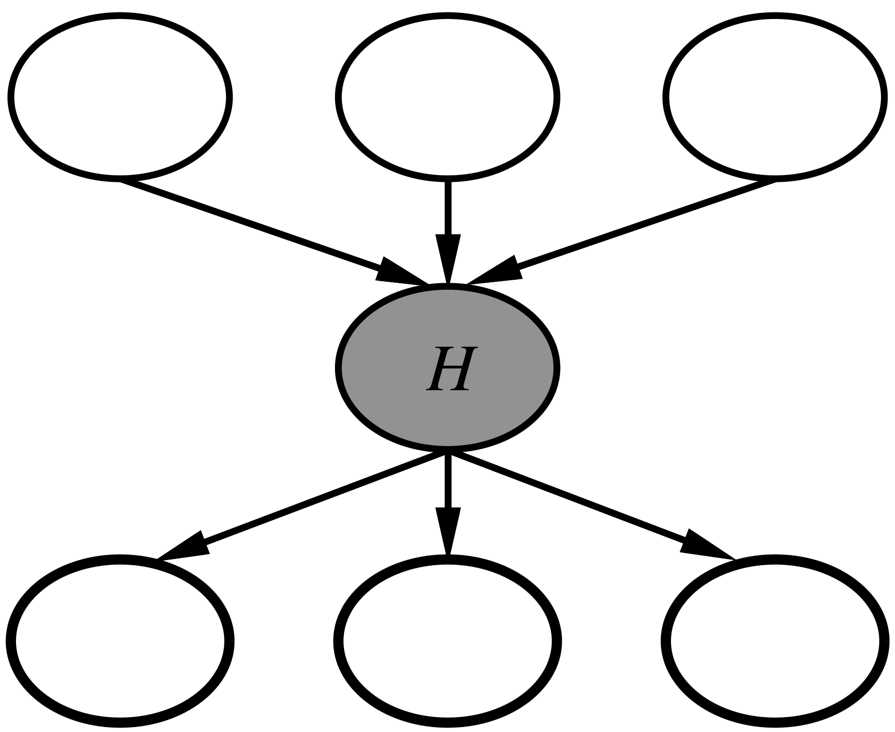


	S C	S ~C	~S C	~S~C
E	0.9	0.3	0.5	0.1
~E	0.1	0.7	0.5	0.9

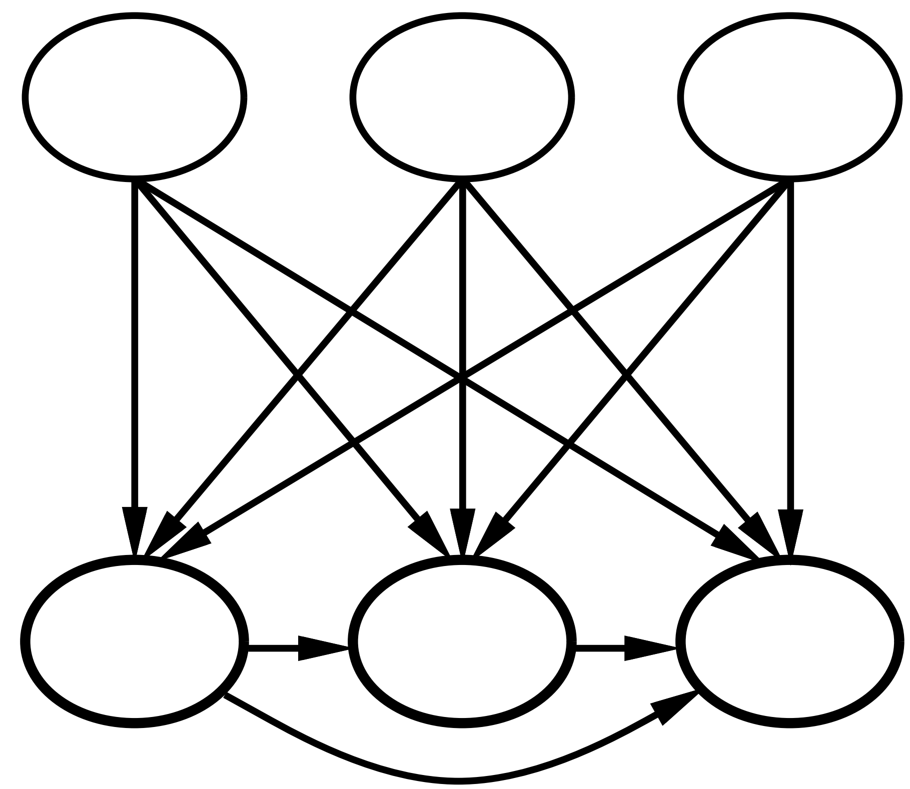


The joint probability of the network is specified in terms of the **conditional probabilities**

Explanations can be simpler using hidden causes



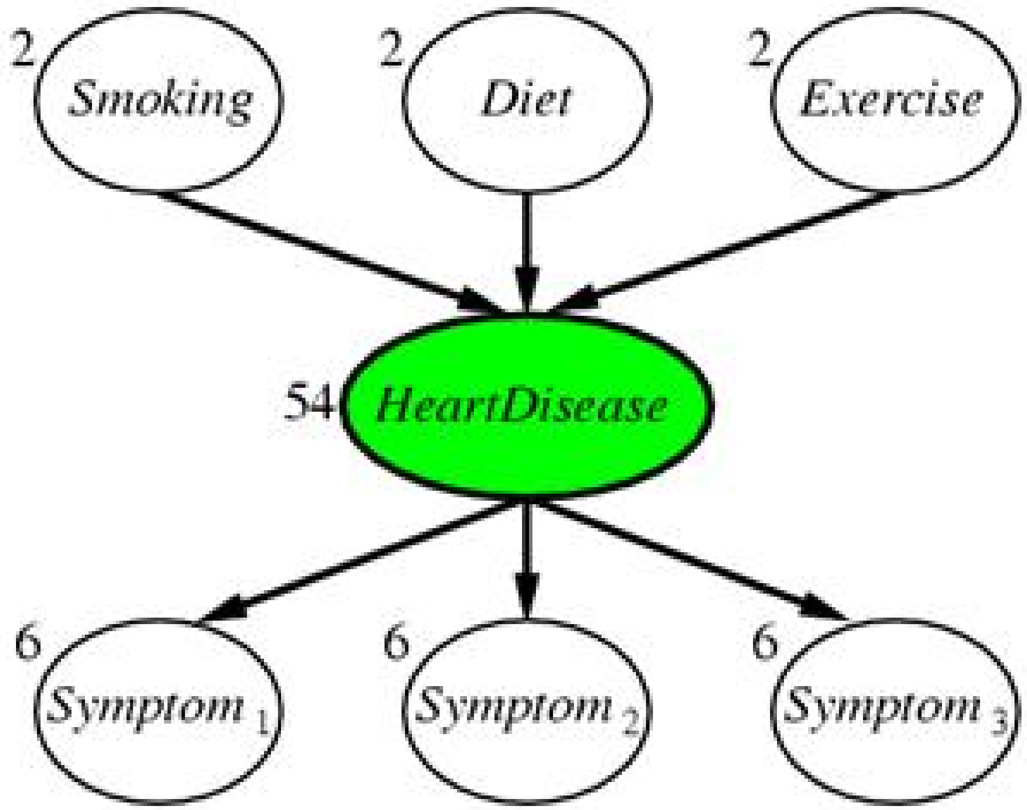
one-hidden cause:
45 indep. params

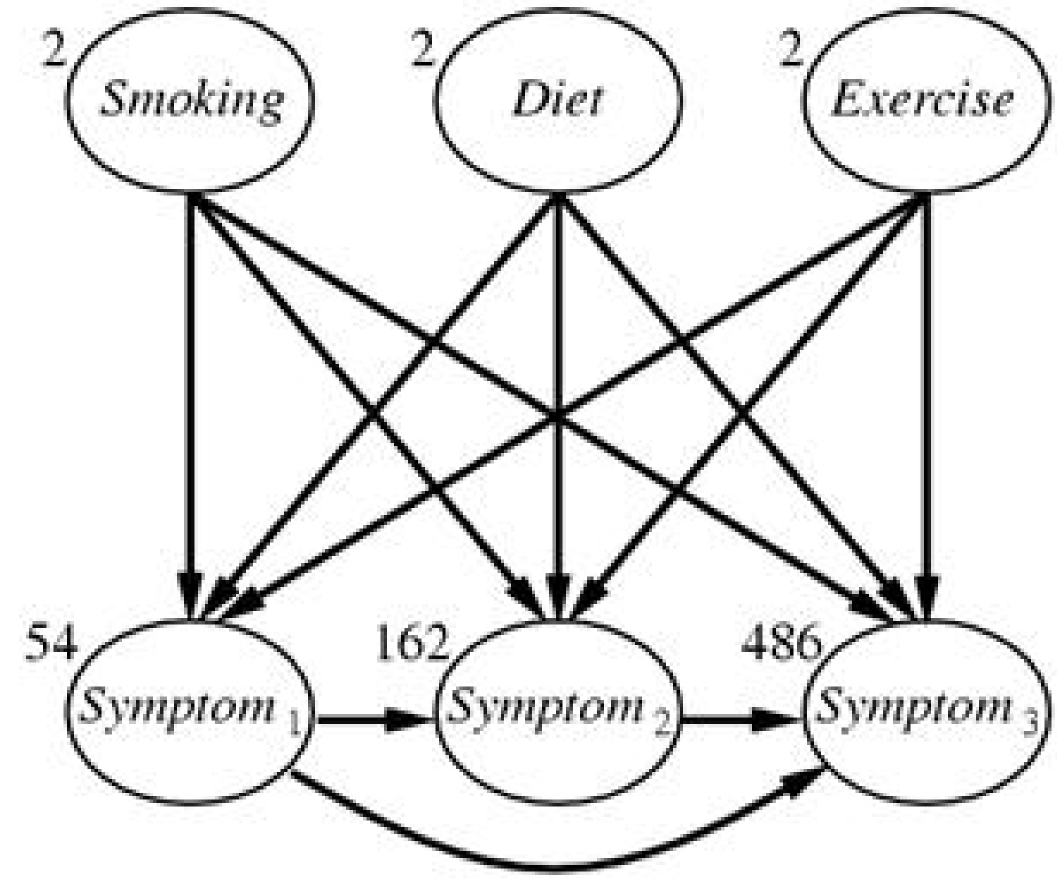


a fully observable network:
708 indep. params

The network structure should reflect the real-world structure.
Models with hidden causes are also called **latent variable models**.

A concrete example (from Kevin Murphy)





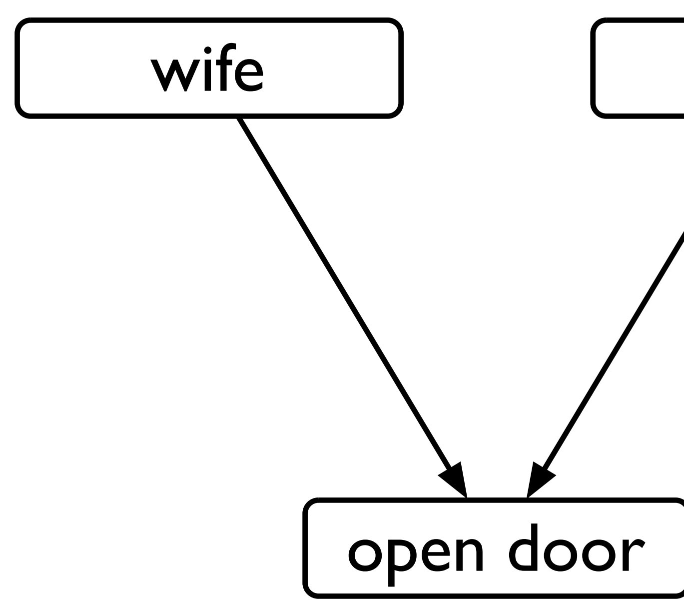
Example: Pathfinder System

- Heckerman (Probabilistic Similarity Networks, MIT Press)
- Diagnostic system for lymph node disease
- 60 diseases and 100 symptoms and test results.
- 14,000 probabilities
- Expert consulted to make network
  - 8 hours to determine variables
  - 35 hours for net topology
  - 40 hours for probability tables
- Experts found it easy to specify causal links and probabilities
- Outperforms world experts in diagnosis
- Extended to many other medical domains

Belief networks

- In Belief networks, *causal relationships* are represented in directed acyclic graphs.
- Arrows indicate causal relationships between the nodes.

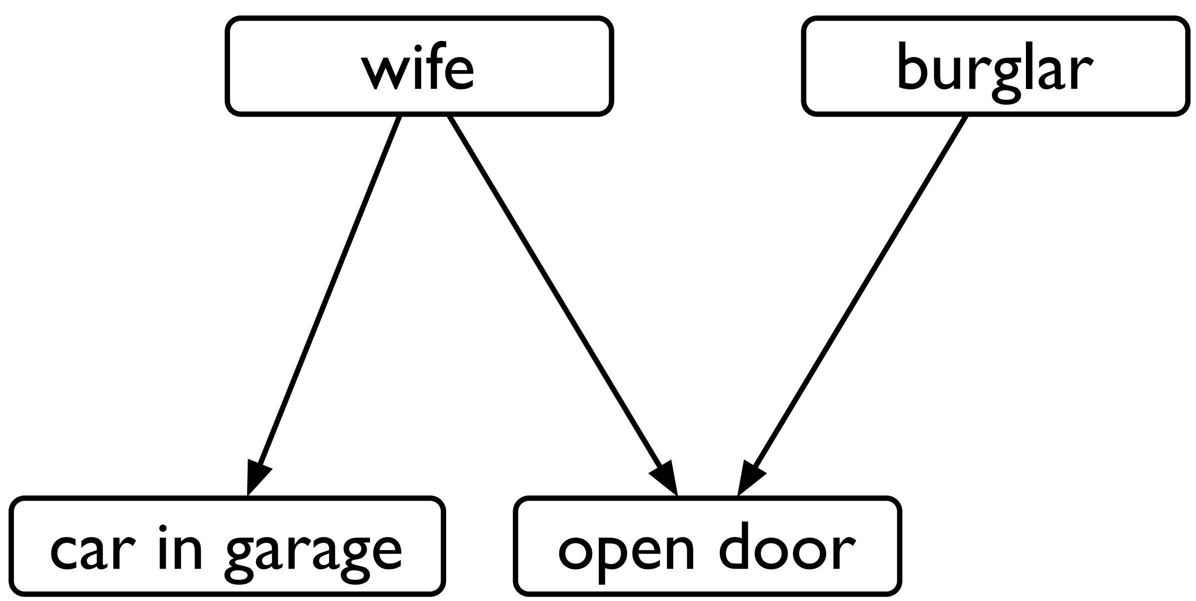
How can we determine what is happening *before* we go in?



We need more information. What else can we observe?

Explaining away

- Suppose we notice that the car is in the garage.
- Now we infer that it's probably my wife, and not a burglar.
- This fact "explains away" the hypothesis of a burglar.



Note that there is no direct causal link between "burglar" and "car in garage".

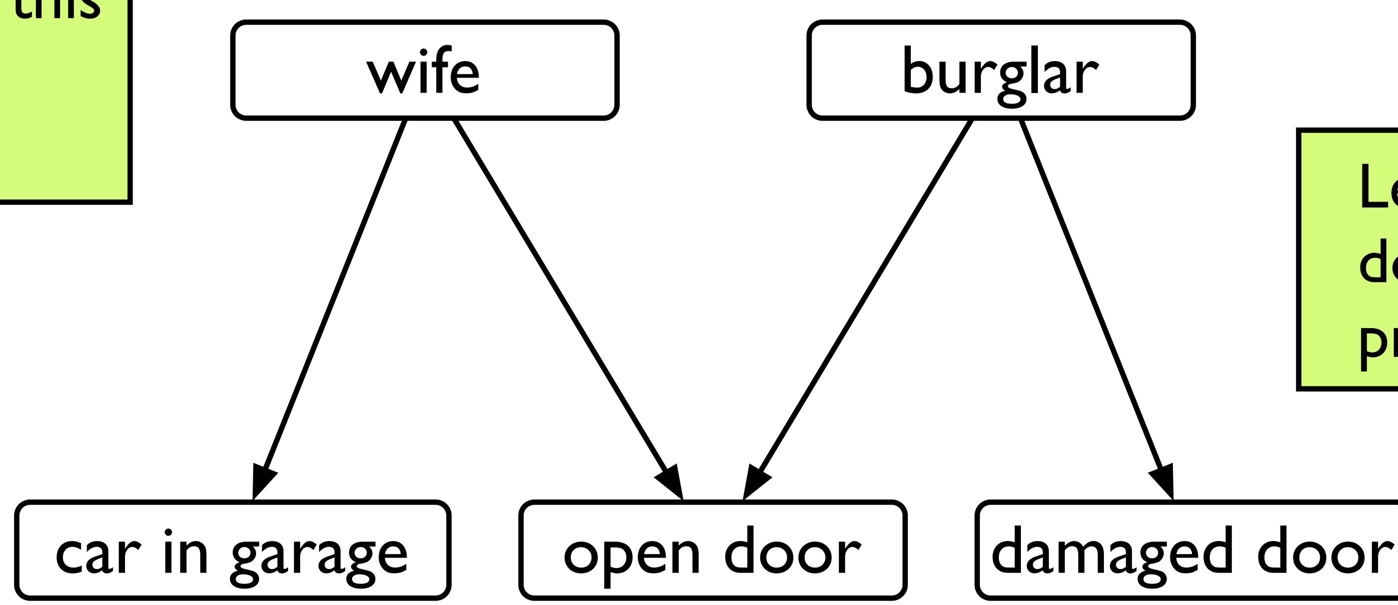
Yet, seeing the car changes our beliefs about the burglar.

Explaining away

- Suppose we notice that the car is in the garage.
- Now we infer that it's probably my wife, and not a burglar.
- This fact "explains away" the hypothesis of a burglar.

- We could also notice the door was damaged, in which case we reach the opposite conclusion.

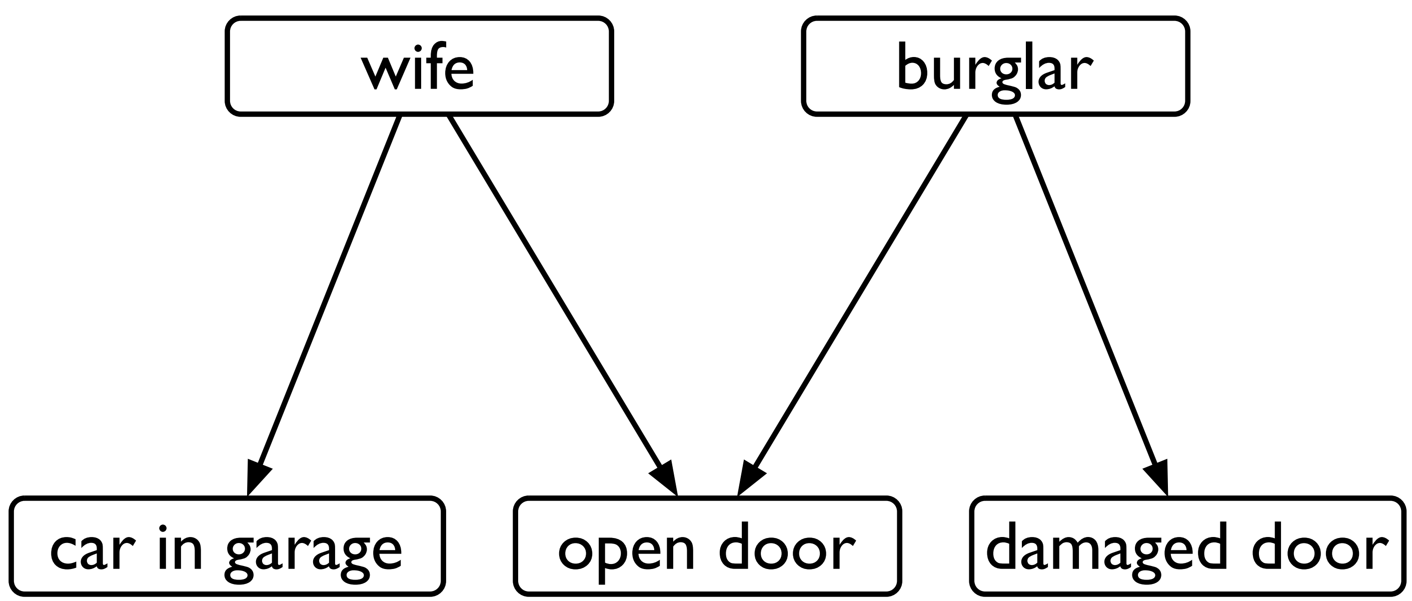
How do we make this inference process more precise?



Let's start by writing down the conditional probabilities.

Defining the belief network

- Each link in the graph represents a conditional relationship between nodes.
- To compute the inference, we must specify the conditional probabilities.



Defining the belief network

- Each link in the graph represents a conditional relationship between nodes.
- To compute the inference, we must specify the conditional probabilities.
- Let's start with the open door. What do we specify?

What else do we need to specify?

The priors probabilities.

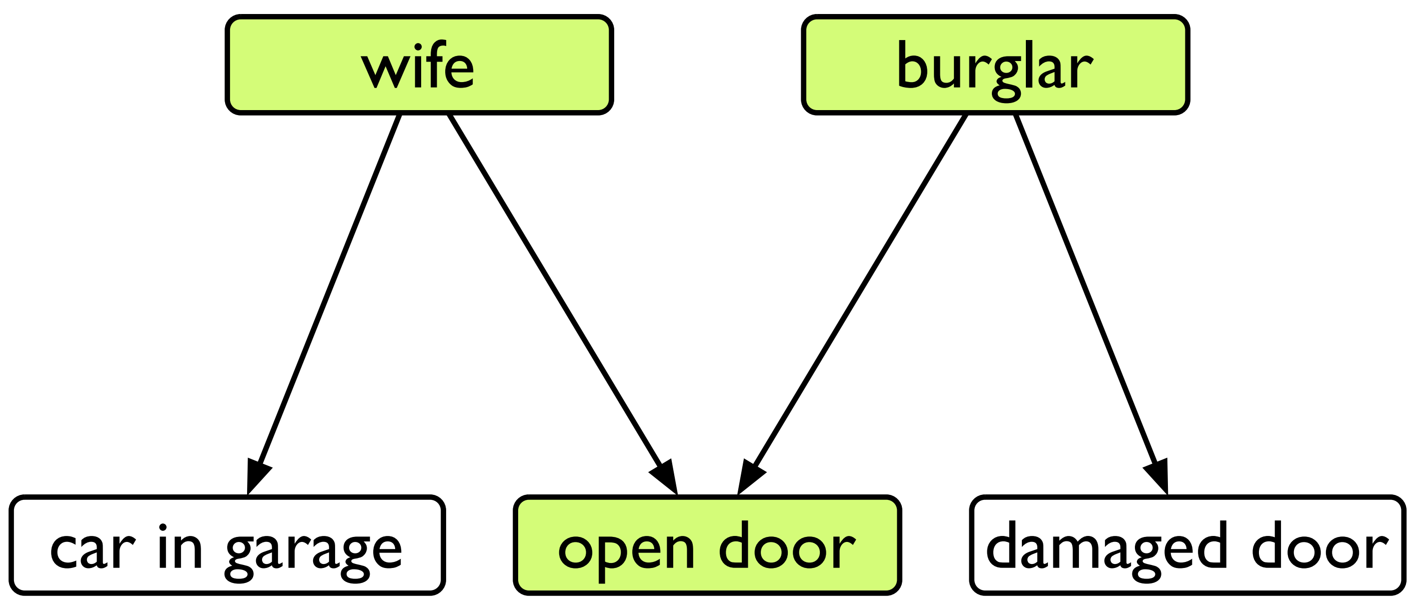
W	B	P(O|W,B)
F	F	0.01
F	T	0.25
T	F	0.05
T	T	

Check: Does this column have to sum to one?

No! Only the full joint distribution does. This is a conditional distribution.

But note that:

$P(\neg O|W,B) = 1 - P(O|W,B)$



Defining the belief network

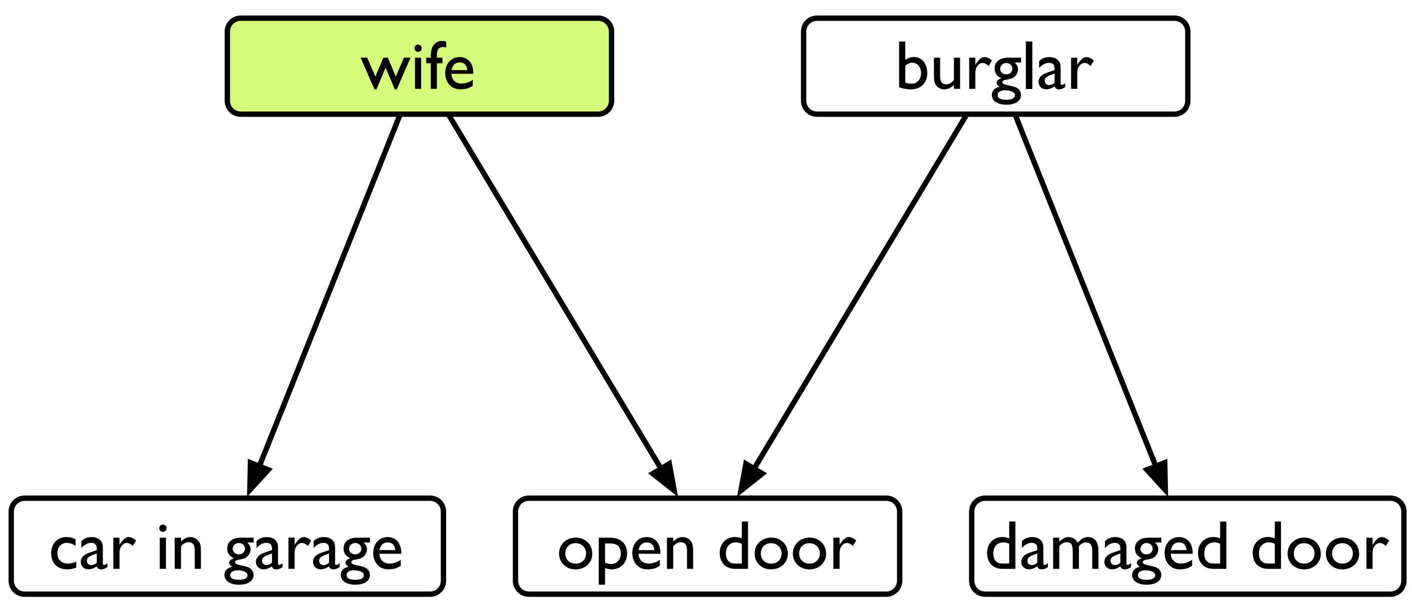
- Each link in the graph represents a conditional relationship between nodes.
- To compute the inference, we must specify the conditional probabilities.
- Let's start with the open door. What do we specify?

What else do we need to specify?

The priors probabilities.

P(W)
0.05

W	B	P(O|W,B)
F	F	0.01
F	T	0.25
T	F	0.05
T	T	0.75



Defining the belief network

- Each link in the graph represents a conditional relationship between nodes.
- To compute the inference, we must specify the conditional probabilities.
- Let's start with the open door. What do we specify?

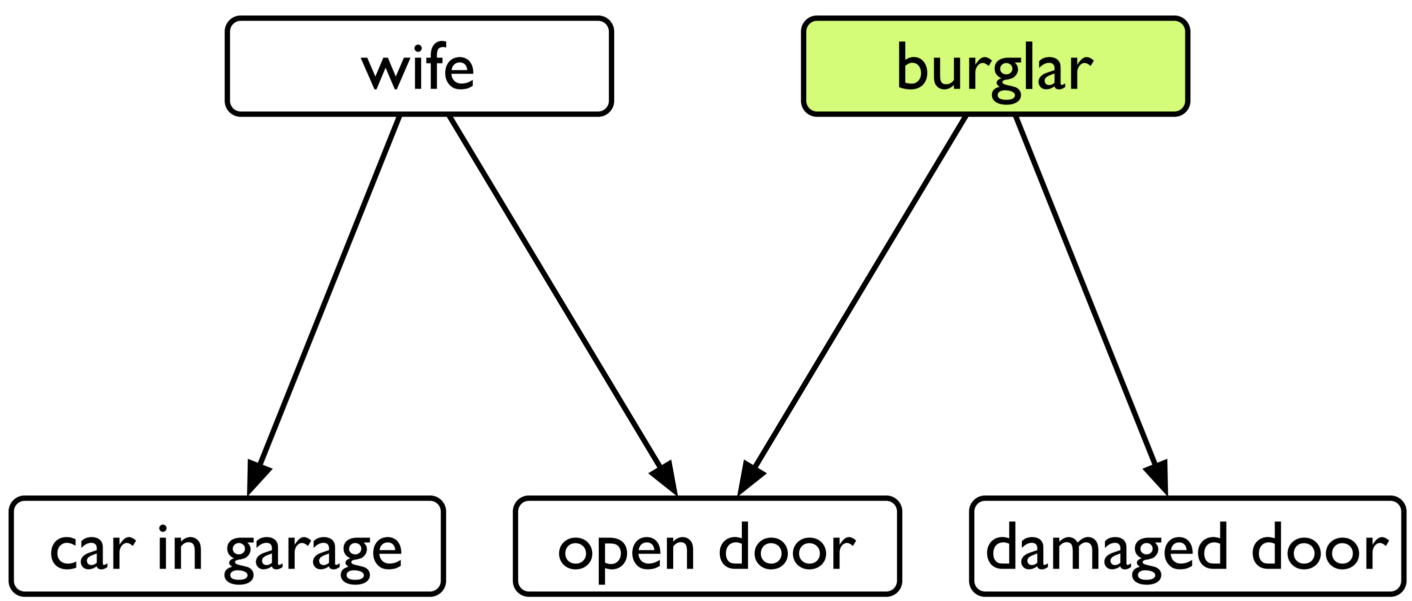
What else do we need to specify?

The priors probabilities.

P(W)
0.05

W	B	P(O|W,B)
F	F	0.01
F	T	0.25
T	F	0.05
T	T	0.75

P(B)
0.001



Defining the belief network

- Each link in the graph represents a conditional relationship between nodes.
- To compute the inference, we must specify the conditional probabilities.
- Let's start with the open door. What do we specify?

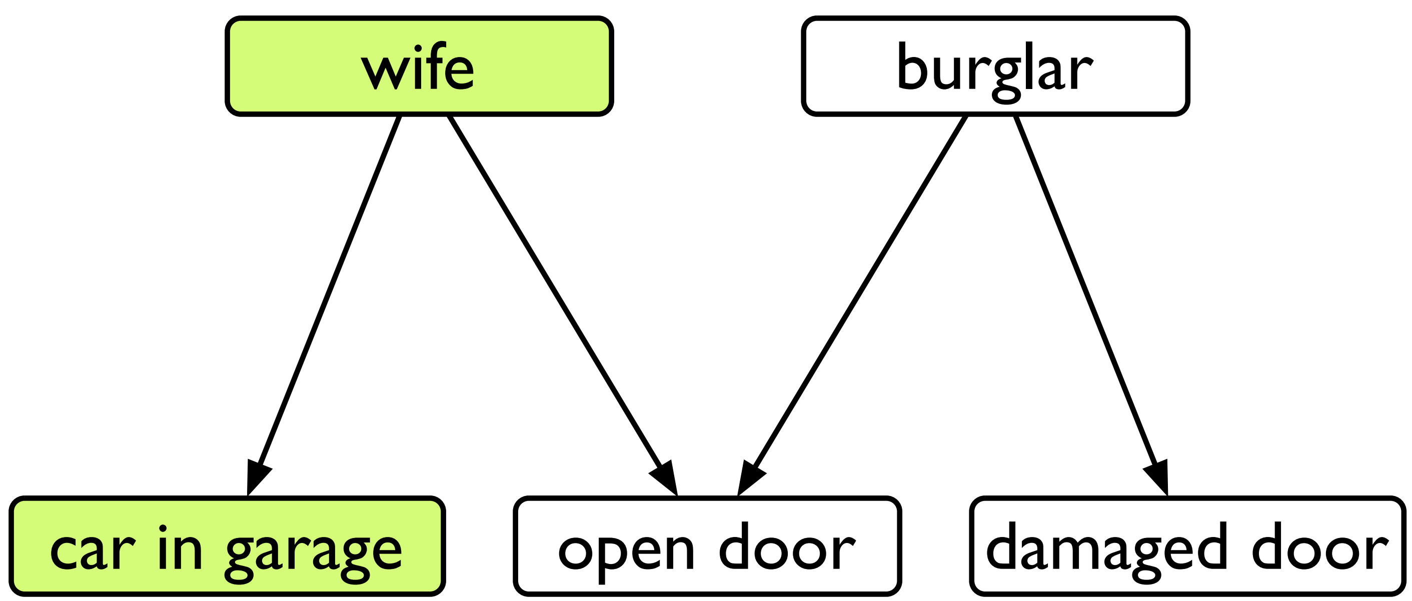
W	B	P(O|W,B)
F	F	0.01
F	T	0.25
T	F	0.05
T	T	0.75

P(W)
0.05

P(B)
0.001

Finally, we specify the remaining conditionals

W	P(C|W)
F	0.01
T	0.95



Defining the belief network

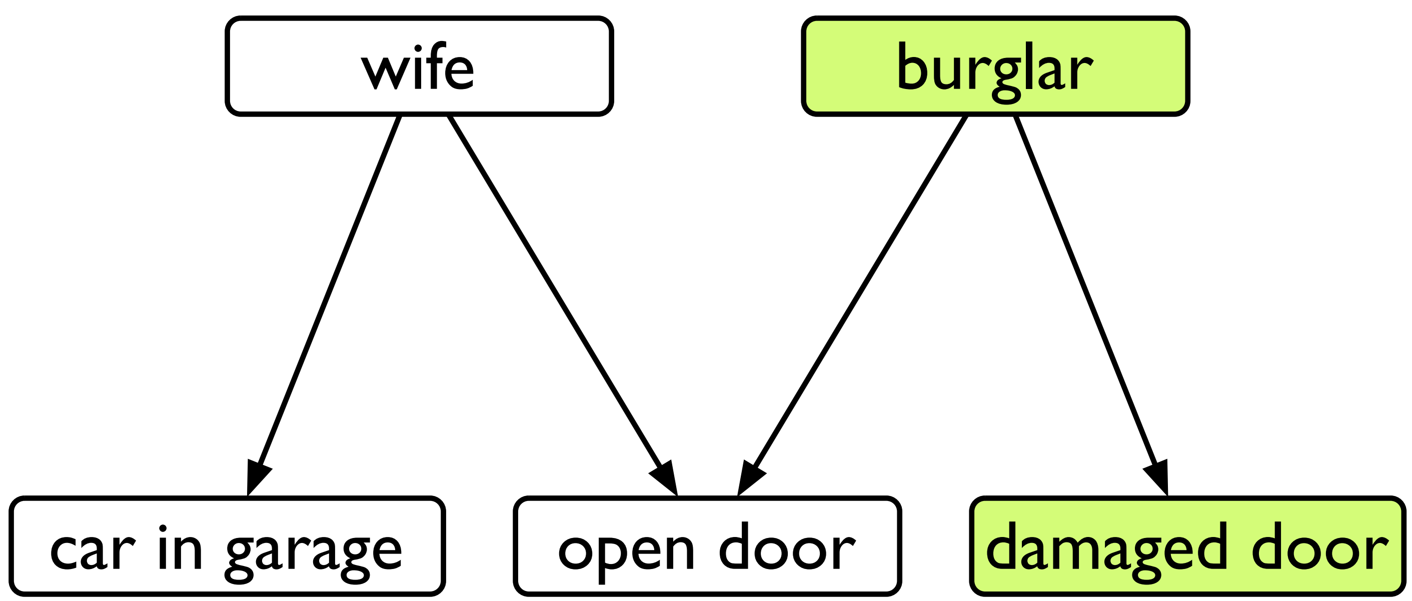
- Each link in the graph represents a conditional relationship between nodes.
- To compute the inference, we must specify the conditional probabilities.
- Let's start with the open door. What do we specify?

W	B	P(O|W,B)
F	F	0.01
F	T	0.25
T	F	0.05
T	T	0.75

P(W)
0.05

P(B)
0.001

Finally, we specify the remaining conditionals



W	P(C|W)
F	0.01
T	0.95

B	P(D|B)
F	0.001
T	0.5

Now what?

Calculating probabilities using the joint distribution

- What the probability that the door is open, it is my wife and not a burglar, we see the car in the garage, and the door is not damaged?
- Mathematically, we want to compute the expression: $P(o,w,\neg b,c,\neg d) = ?$
- We can just repeatedly apply the rule relating joint and conditional probabilities.
  - $P(x,y) = P(x|y)\,P(y)$

Summary of inference with the joint probability distribution

- The complete (probabilistic) relationship between variables is specified by the joint probability:

$$P(X_1, X_2, \ldots, X_n)$$
$$= P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

- All conditional and marginal distributions can be derived from this using the basic rules of probability, the sum rule and the product rule

$$P(X) = \sum_Y P(X, Y)$$ sum rule, "marginalization"

$$P(X, Y) = P(Y|X)P(X) = P(X|Y)P(Y)$$ product rule

$$P(Y|X) = \frac{P(X, Y)}{P(Y)}$$ corollary, conditional probability

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$ corollary, Bayes rule

Calculating probabilities using the joint distribution

- The probability that the door is open, it is my wife and not a burglar, we see the car in the garage, and the door is not damaged.
- $P(o,w,\neg b,c,\neg d) = P(o|w,\neg b,c,\neg d)P(w,\neg b,c,\neg d)$

$$= P(o|w,\neg b)P(w,\neg b,c,\neg d)$$

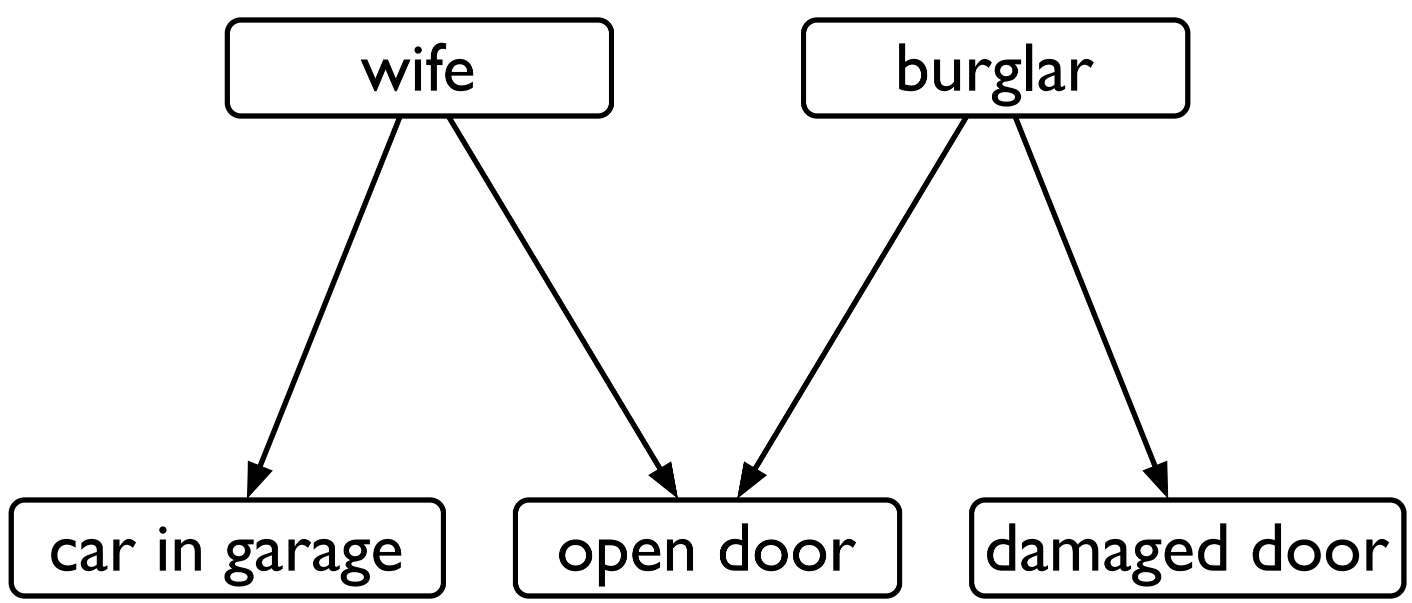
$$= P(o|w,\neg b)P(c|w,\neg b,\neg d)P(w,\neg b,\neg d)$$

$$= P(o|w,\neg b)P(c|w)P(w,\neg b,\neg d)$$

$$= P(o|w,\neg b)P(c|w)P(\neg d|w,\neg b)P(w,\neg b)$$

$$= P(o|w,\neg b)P(c|w)P(\neg d|\neg b)P(w,\neg b)$$

$$= P(o|w,\neg b)P(c|w)P(\neg d|\neg b)P(w)P(\neg b)$$

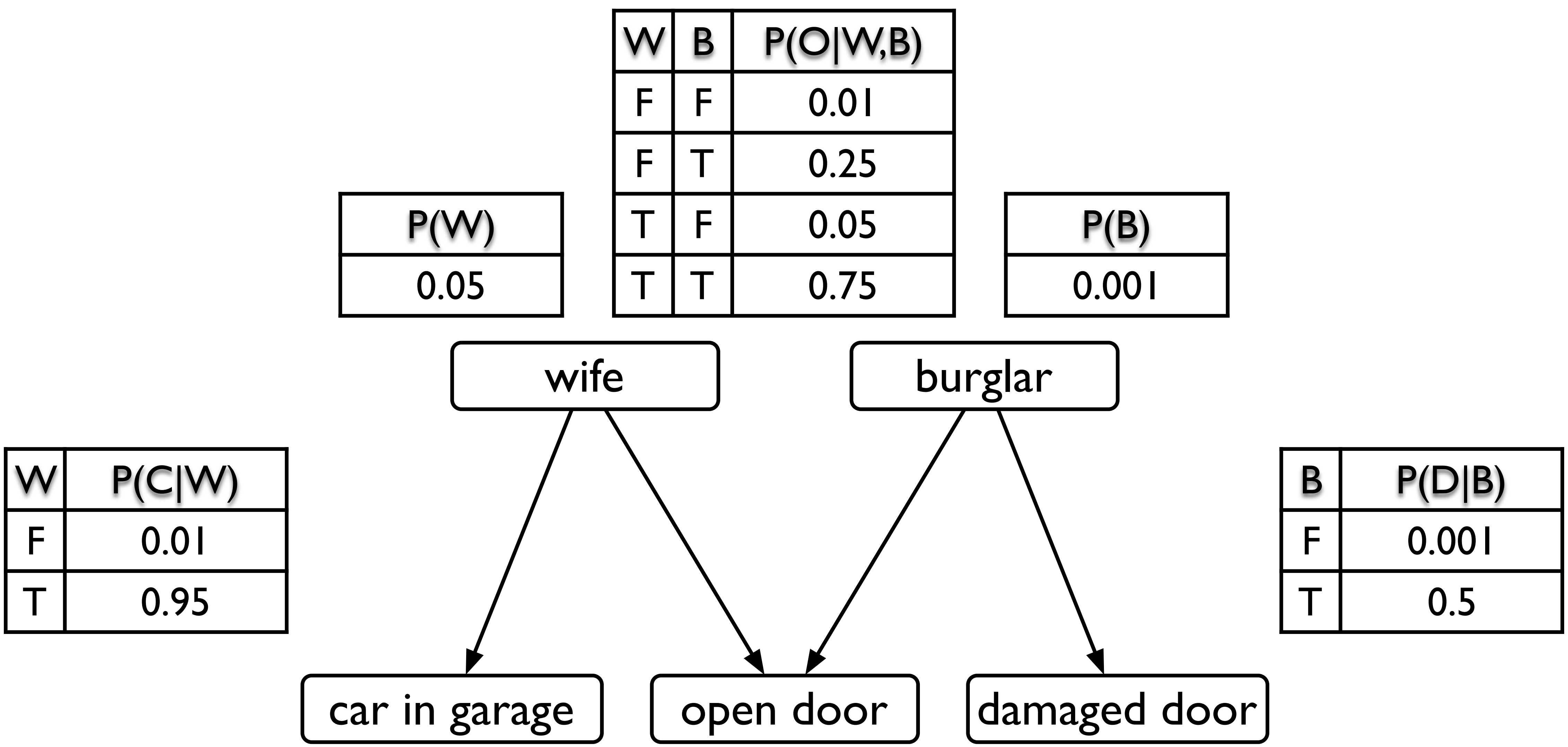


Calculating probabilities using the joint distribution

- $P(o,w,\neg b,c,\neg d) = P(o|w,\neg b)P(c|w)P(\neg d|\neg b)P(w)P(\neg b)$

  $= \quad 0.05 \times 0.95 \times 0.999 \times 0.05 \times 0.999 = 0.0024$

- This is essentially the probability that my wife is home and leaves the door open.



W	B	P(O|W,B)
F	F	0.01
F	T	0.25
T	F	0.05
T	T	0.75

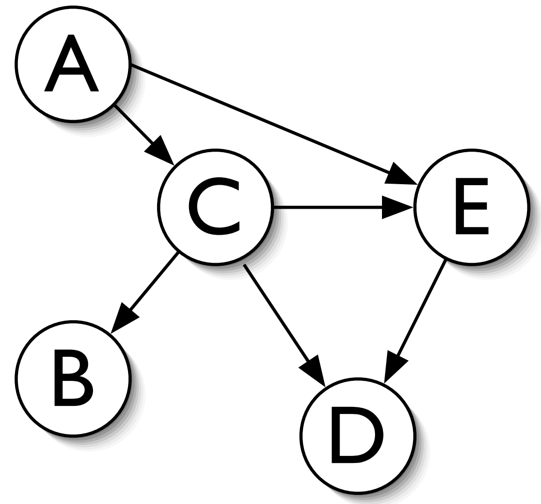
P(W)
0.05

P(B)
0.001

W	P(C|W)
F	0.01
T	0.95

B	P(D|B)
F	0.001
T	0.5

Calculating probabilities in a general Bayesian belief network



- Note that by specifying all the conditional probabilities, we have also specified the joint probability. For the directed graph above:

  P(A,B,C,D,E) = P(A) P(B|C) P(C|A) P(D|C,E) P(E|A,C)

- The general expression is:

$$
\begin{aligned}
P(x_1, \ldots, x_n) &\equiv P(X_1 = x_1 \wedge \ldots \wedge X_n = x_n) \\
&= \prod_{i=1}^{n} P(x_i | \mathrm{parents}(X_i))
\end{aligned}
$$

- With this we can calculate (in principle) the probability of any joint probability.
- This implies that we can also calculate any conditional probability.

Calculating conditional probabilities

- Using the joint we can compute any conditional probability too
- The conditional probability of any one subset of variables given another disjoint subset is

$$P(S_1|S_2) = \frac{P(S_1 \wedge S_2)}{P(S_2)} = \frac{\sum p \in S_1 \wedge S_2}{\sum p \in S_2}$$

where p ∈ S is shorthand for all the entries of the joint matching subset S.

- How many terms are in this sum? 2^N

The number of terms in the sums is *exponential* in the number of variables.

In fact, general querying Bayes nets is NP complete.

So what do we do?

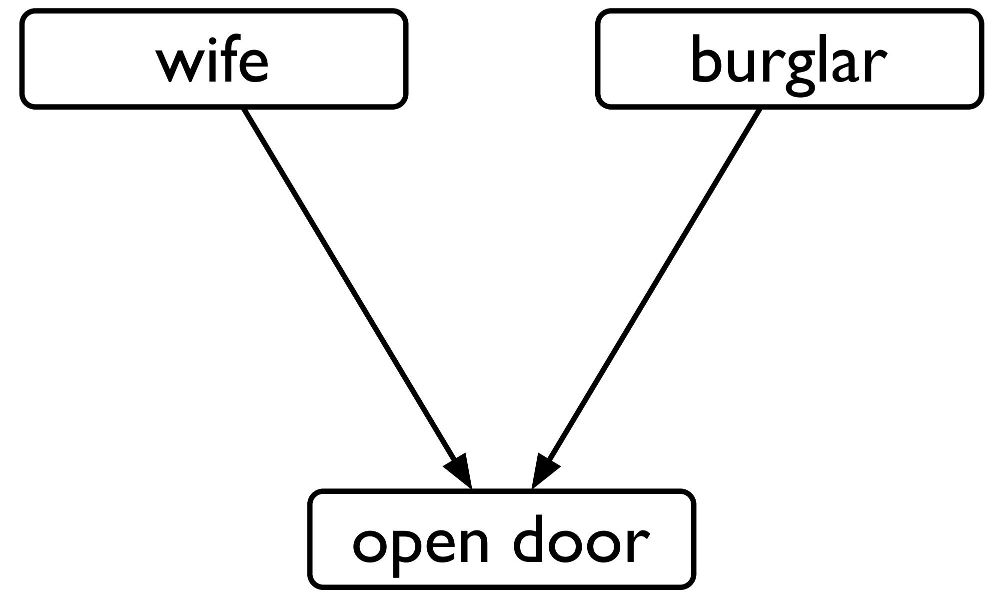
- There are also many approximations:
  - stochastic (MCMC) approximations
  - approximations
- The are special cases of Bayes nets for which there are fast, exact algorithms:
  - variable elimination
  - belief propagation

Belief networks with multiple causes

- In the models above, we specified the joint conditional density by hand.
- This specified the probability of a variable given each possible value of the causal nodes.
- Can this be specified in a more generic way?
- Can we avoid having to specify every entry in the joint conditional pdf?
- For this we need to specify:

  P(X | parents(X))

- One classic example of this function is the "Noisy-OR" model.

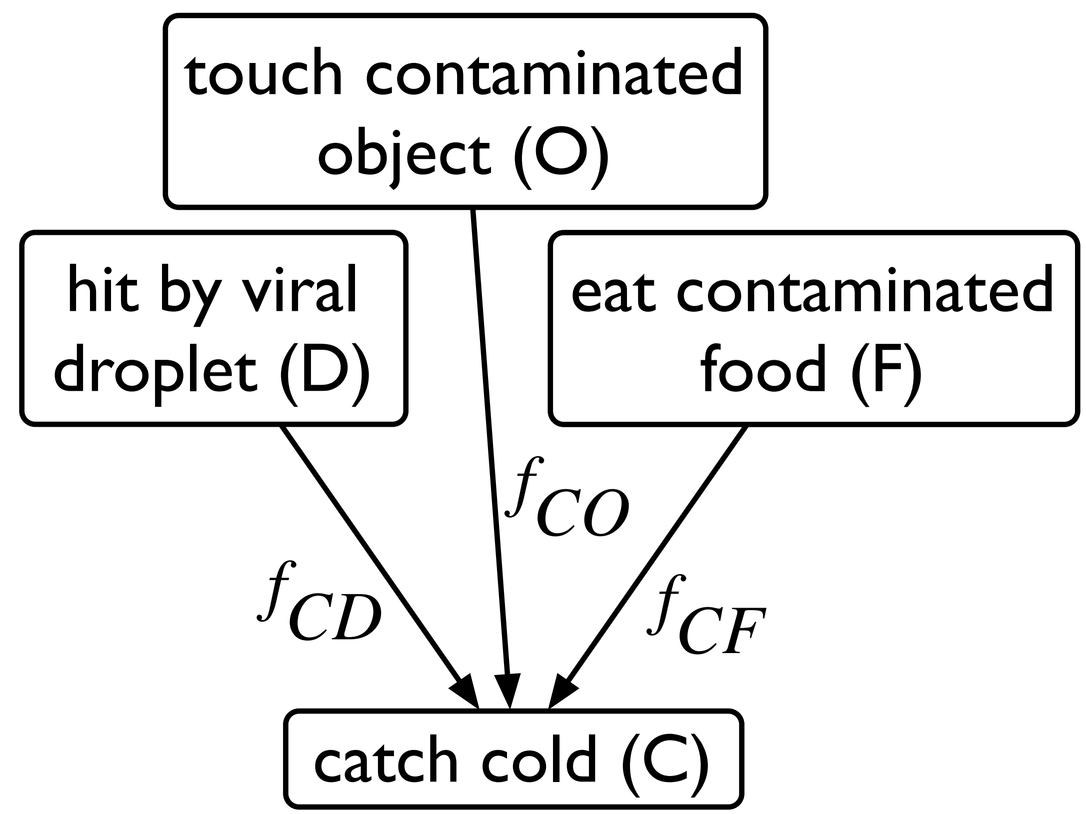


W	B	P(O|W,B)
F	F	0.01
F	T	0.25
T	F	0.05
T	T	0.75

Beyond tables: modeling causal relationships using Noisy-OR

- We assume each cause C_j can produce effect E_i with probability f_{ij}.
- The noisy-OR model assumes the parent causes of effect E_i contribute independently.
- The probability that none of them caused effect E_i is simply the product of the probabilities that each one *did not* cause E_i.
- The probability that any of them caused E_i is just one minus the above, i.e.

$$
\begin{aligned}
P(E_i|\mathrm{par}(E_i)) &= P(E_i|C_1, \ldots, C_n) \\
&= 1 - \prod_i (1 - P(E_i|C_j)) \\
&= 1 - \prod_i (1 - f_{ij})
\end{aligned}
$$



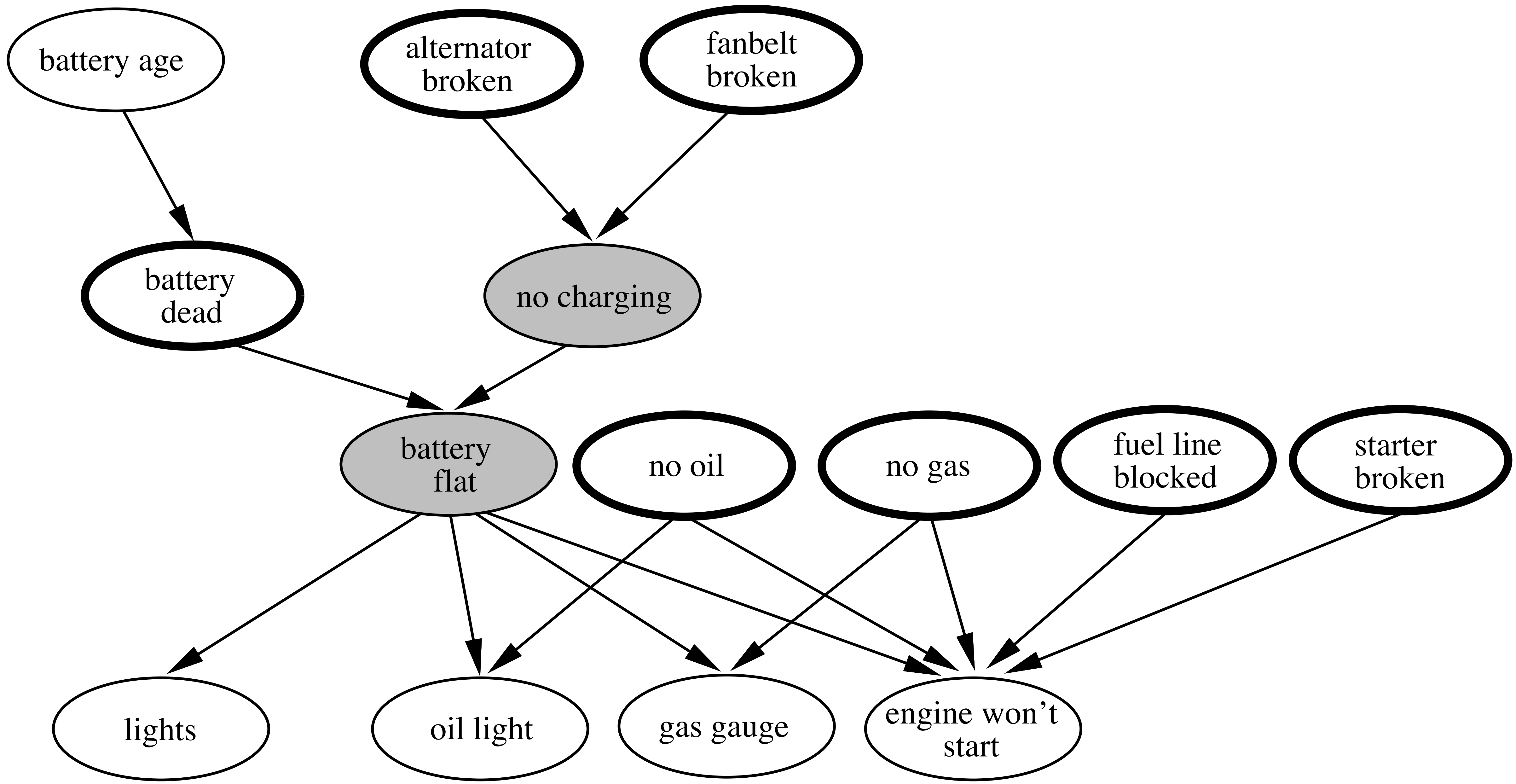
$$
P(C|D,O,F) = 1 - (1 - f_{CD})(1 - f_{CO})(1 - f_{CF})
$$

Another noisy-OR example

Table 2. Conditional probability table for $P(\mathit{Fever} \mid \mathit{Cold}, \mathit{Flu}, \mathit{Malaria})$, as calculated from the noisy-OR model.

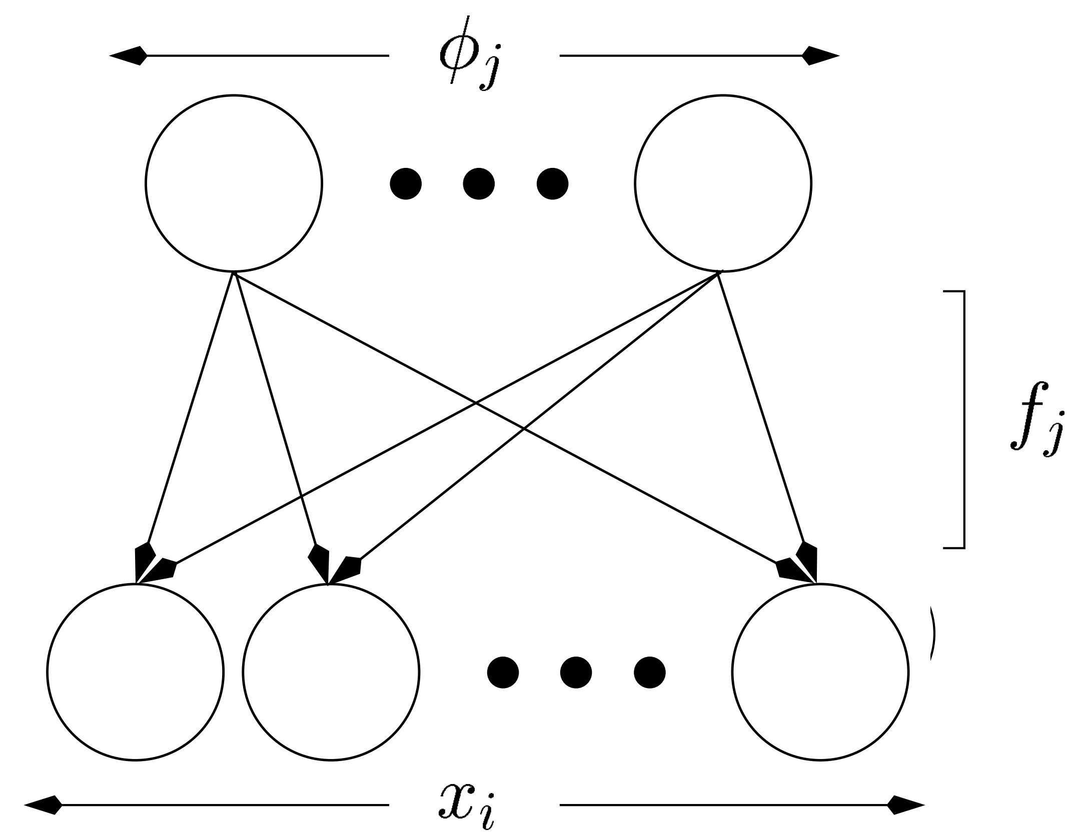
Cold	*Flu*	*Malaria*	$P(\mathit{Fever})$	$P(\neg \mathit{Fever})$
F	F	F	0.0	1.0
F	F	T	0.9	0.1
F	T	F	0.8	0.2
F	T	T	0.98	$0.02 = 0.2 \times 0.1$
T	F	F	0.4	0.6
T	F	T	0.94	$0.06 = 0.6 \times 0.1$
T	T	F	0.88	$0.12 = 0.6 \times 0.2$
T	T	T	0.988	$0.012 = 0.6 \times 0.2 \times 0.1$

A more complex model with noisy-OR nodes

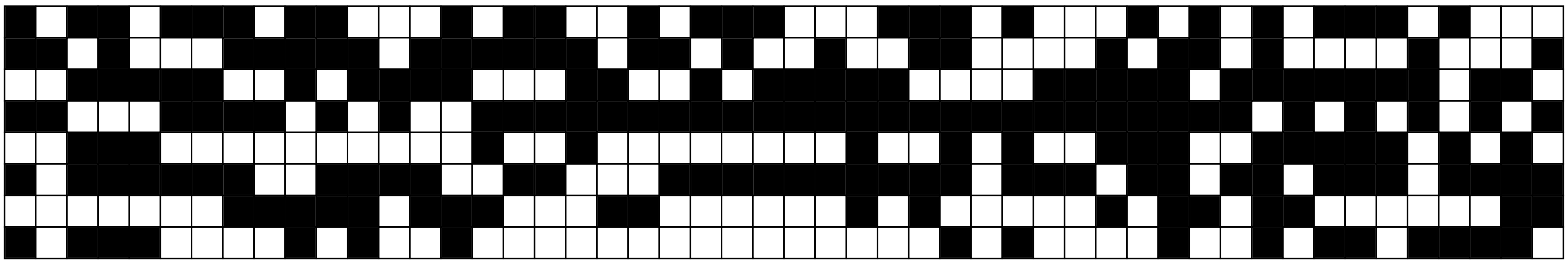


A general one-layer causal network

- Could either model causes and effects
- Or equivalently stochastic binary features.
- Each input x_i encodes the probability that the ith binary input feature is present.
- The set of features represented by ϕj is defined by weights f_{ij} which encode the probability that feature i is an instance of ϕ_j.



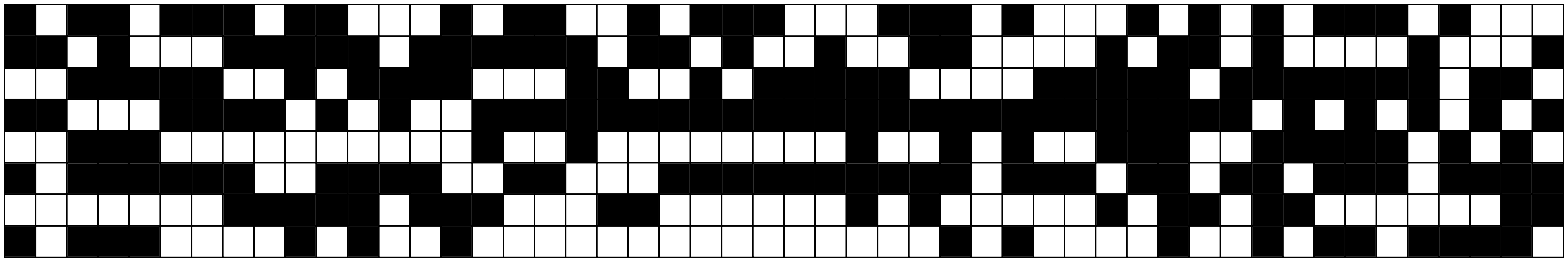
The data: a set of stochastic binary patterns



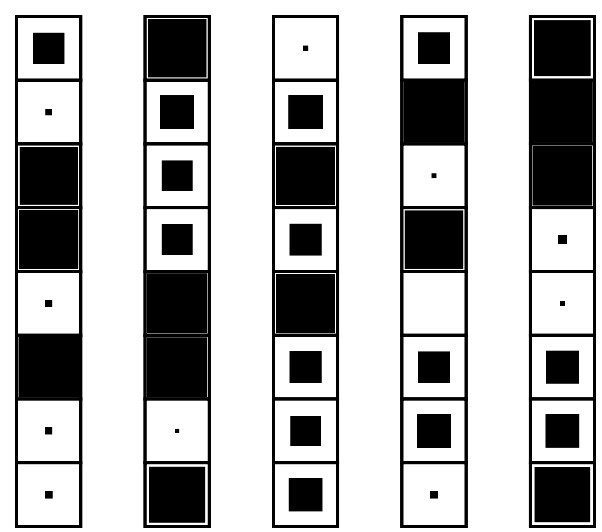
Each column is a distinct eight-dimensional binary feature.

There are five underlying causal feature patterns.
What are they?

The data: a set of stochastic binary patterns



Each column is a distinct eight-dimensional binary feature.

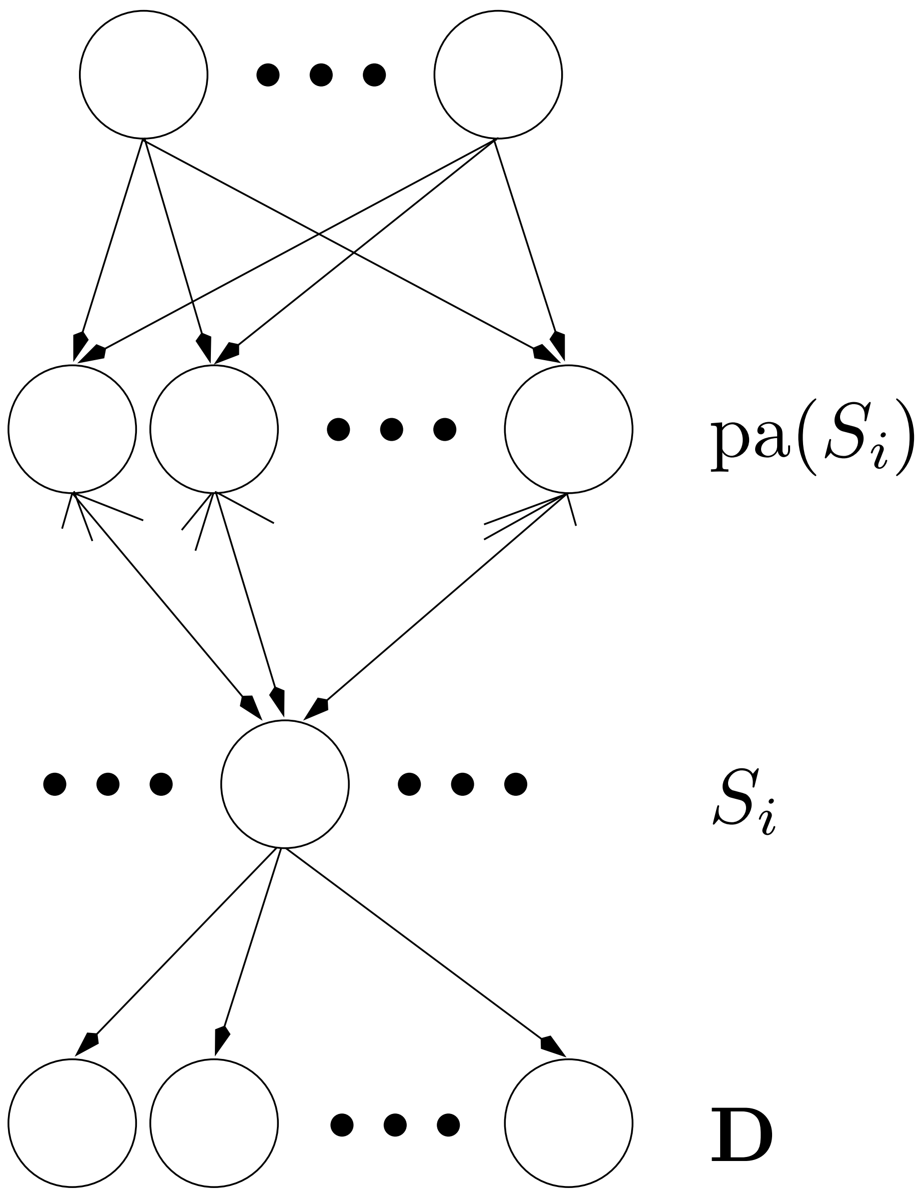


true hidden causes of the data

This is a *learning* problem, which
we'll cover in later lecture.

Hierarchical Statistical Models

A Bayesian belief network:



The joint probability of binary states is

$$P(\mathbf{S}|\mathbf{W}) = \prod_i P(S_i|\mathrm{pa}(S_i), \mathbf{W})$$

The probability S_i depends only on its parents:

$$P(S_i|\mathrm{pa}(S_i), \mathbf{W}) = \begin{cases} h(\sum_j S_j w_{ji}) & \text{if } S_i = 1 \\ 1 - h(\sum_j S_j w_{ji}) & \text{if } S_i = 0 \end{cases}$$

The function h specifies how causes are combined, $h(u) = 1 - \exp(-u)$, $u > 0$.

Main points:

- hierarchical structure allows model to form high order representations
- upper states are priors for lower states
- weights encode higher order features