Probability

Motivation

- Agents don't have complete knowledge about the world.
- Agents need to make decisions in an uncertain world.
- It isn't enough to assume what the world is like. Example: wearing a seat belt.
- An agent needs to reason about its uncertainty.
- When an agent acts under uncertainty, it is gambling.
- Agents who don't use probabilities will do worse than those who do.

Probability

- Belief in a proposition a can be measured by a number between 0 and 1 — this is the *probability of* $a = P(a)$.
  - $P(a) = 0$ means that a is believed false.
  - $P(a) = 1$ means that a is believed true.
- $0 < P(a) < 1$ means the agent is unsure.
- Probability is a measure of ignorance.
- Probability is *not* a measure of degree of truth.

Random Variables

- The variables in probability are called *random variables*.
- Each variable X has a set of possible values.
- A tuple of random variables is written as $X_1, \ldots, X_n$.
- $X = x$ means variable X has value x.
- A *proposition* is a Boolean expression made from assignments of values to variables.
- Example: X_1, X_2 are the values of two dice. $X_1 + X_2 = 7$ is equivalent to $(X_1 = 1 \wedge X_2 = 6) \vee (X_1 = 2 \wedge X_2 = 5) \vee \ldots$

Semantics

- A *possible world* ω is a variable assignment (all the variables).
- Let Ω be the set of all possible worlds.
- Assuming each variable has a finite set of possible values.
  - Define $P(\omega)$ for each world ω so that $0 \leq P(\omega)$ and they sum to 1. This is the *joint probability distribution*.
  - The probability of proposition a is defined by:

$$P(a) = \sum_{\omega \models a} P(\omega)$$

Dice Example

$\mathbf{P}(X_1, X_2)$		
X_1	X_2	P
1	1	1/36
1	2	1/36
1	3	1/36
1	4	1/36
1	5	1/36
1	6	1/36
2	1	1/36
2	2	1/36
2	3	1/36
2	4	1/36
2	5	1/36
2	6	1/36

$\mathbf{P}(X_1, X_2)$		
X_1	X_2	P
3	1	1/36
3	2	1/36
3	3	1/36
3	4	1/36
3	5	1/36
3	6	1/36
4	1	1/36
4	2	1/36
4	3	1/36
4	4	1/36
4	5	1/36
4	6	1/36

$\mathbf{P}(X_1, X_2)$		
X_1	X_2	P
5	1	1/36
5	2	1/36
5	3	1/36
5	4	1/36
5	5	1/36
5	6	1/36
6	1	1/36
6	2	1/36
6	3	1/36
6	4	1/36
6	5	1/36
6	6	1/36

Joint Distribution Example

$\mathbf{P}(A, B, C, D)$

A	B	C	D	P
T	T	T	T	0.04
T	T	T	F	0.04
T	T	F	T	0.32
T	T	F	F	0.00
T	F	T	T	0.00
T	F	T	F	0.08
T	F	F	T	0.16
T	F	F	F	0.16

$\mathbf{P}(A, B, C, D)$

A	B	C	D	P
F	T	T	T	0.01
F	T	T	F	0.01
F	T	F	T	0.02
F	T	F	F	0.00
F	F	T	T	0.00
F	F	T	F	0.08
F	F	F	T	0.04
F	F	F	F	0.04

Axioms of Probability

Three axioms for probabilities (finite case)

1. $0 \leq P(a)$ for any proposition p.
2. $P(a) = 1$ if a is a tautology.
3. $P(a \vee b) = P(a) + P(b)$ if a and b contradict.

For all propositions a and b, these axioms imply:

- $P(\neg a) = 1 - P(a)$
- $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$
- $P(a) = P(a \wedge b) + P(a \wedge \neg b)$
- If variable V has possible values D, then
  $$P(a) = \sum_{d \in D} P(a \wedge V = d)$$

Conditional Probabilities

- Conditional probabilities specify how to revise beliefs based on *evidence*, known values for one or more variables.
- If e is the evidence, the *conditional probability* of h given e is $P(h \mid e) = P(h \wedge e)/P(e)$.
- Chain Rule: $P(a \wedge b \wedge c) =$
  $P(a \mid b \wedge c) P(b \wedge c) = P(a \mid b \wedge c) P(b \mid c) P(c)$
- Bayes' Theorem:
  $P(h \mid e) = P(e \mid h) P(h)/P(e)$
- Update with additional evidence e'.
  $P(h \mid e \wedge e') = P(e' \mid h \wedge e) P(h \mid e)/P(e' \mid e)$

Using Bayes' Theorem

- Often you have causal knowledge:
  $P(symptom \mid disease)$
  $P(\textit{light is off} \mid \textit{status of switches})$
  $P(alarm \mid fire)$
  $P(\textit{looks, swims, quacks like a duck} \mid \textit{a duck})$
- and want to do evidential reasoning:
  $P(disease \mid symptom)$
  $P(\textit{status of switches} \mid \textit{light is off})$
  $P(fire \mid alarm)$.
  $P(\textit{a duck} \mid \textit{looks, swims, quacks like a duck})$
- Bayes' theorem tells you how.

Problem: Lots of Small Numbers

- If there are n binary variables, there are 2^n numbers to be assigned for a joint probability distribution.
- They need to add up to one, so if $2^n - 1$ numbers are assigned, the last one can be solved for. [Doesn't help much.]
- In addition, each number is likely to be very, very small, so it is unrealistic to use frequencies (even with Big Data).
- Reduce work by using knowledge of when one variable is independent of another variable.

Conditional Independence

- X and Y are *independent* if, for any x, y, $P(X=x \wedge Y=y) = P(X=x)P(Y=y)$.
  - One die is independent of the other die.
  - Rain is independent of the day of the week.
- X is *conditionally independent* of Y given Z if $P(X=x \mid Y=y \wedge Z=z) = P(X=x \mid Z=z)$ for any x, y, z. Knowing Z, ignore Y to infer X.
  - A nice day (Y) makes me more likely to exercise (Z), and so more likely to be tired (X).
  - Suppose X, Y, Z randomly chosen, but $X \neq Z$ and $Z \neq Y$.

Naive Bayes

- Suppose we want to determine the probability of a hypothesis given the evidence $P(H \mid E_1, \ldots, E_n)$
- A naive, but often effective, assumption is that the evidence is conditionally independent of the hypothesis.

$$P(H \mid E_1, \ldots, E_n)$$
$$= P(H)\, P(E_1, \ldots, E_n \mid H) \,/\, P(E_1, \ldots, E_n)$$
$$\approx P(H)\, \prod_{i=1}^{n} P(E_i|H) \,/\, P(E_1, \ldots, E_n)$$

- Different values for H have same denominator, so only need to compare numerators.

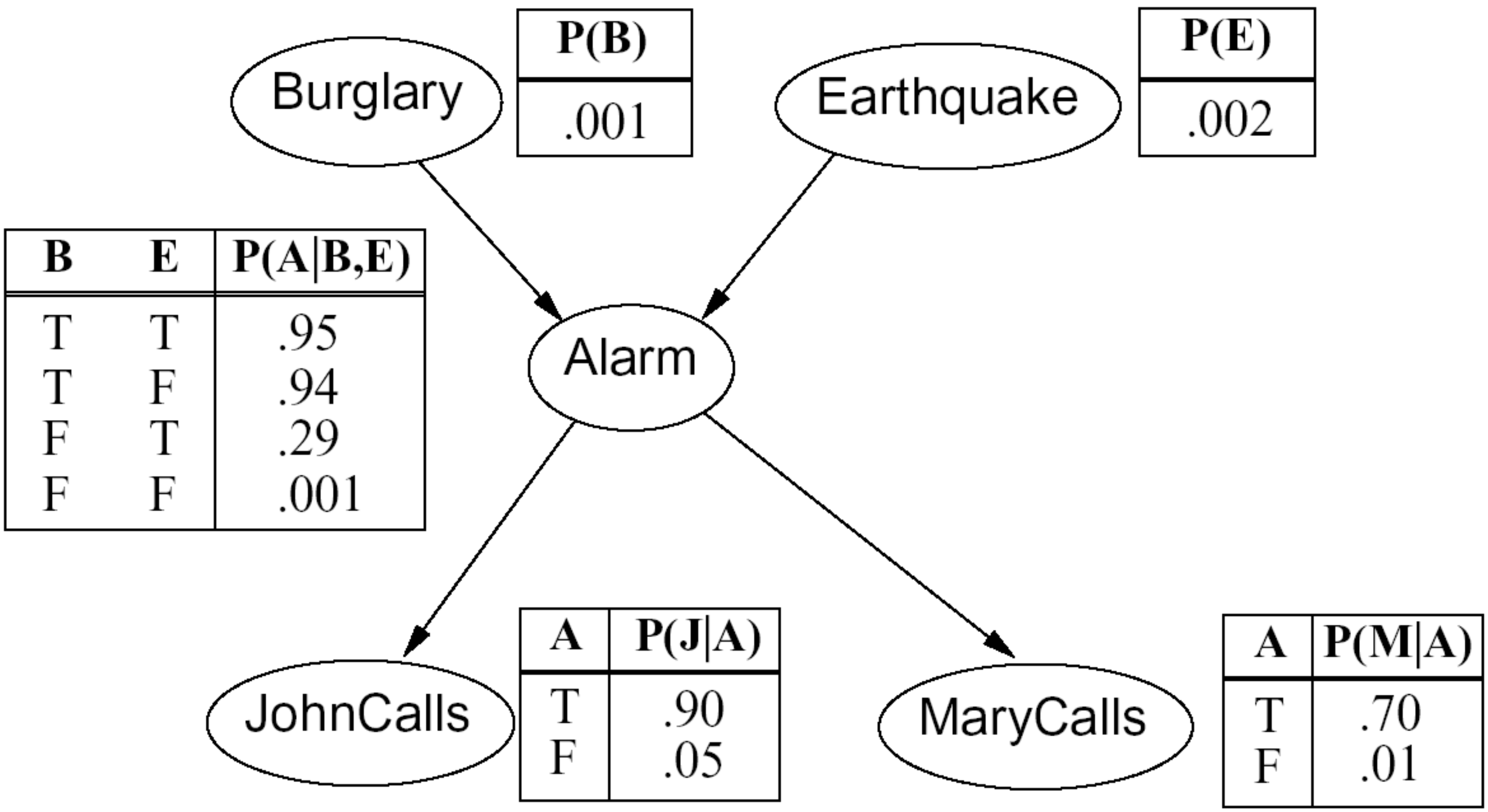
Bayesian Networks

A *Bayesian network** consists of:

- a directed acyclic graph, where nodes correspond to variables,
- a set of possible values for each variable,
- a prior probability table for each variable with no parents, and
- a conditional probability table for each variable with parents, specifying the probability of the variable given its parents.

*I prefer "Bayesian network" over "belief network".

Example 1

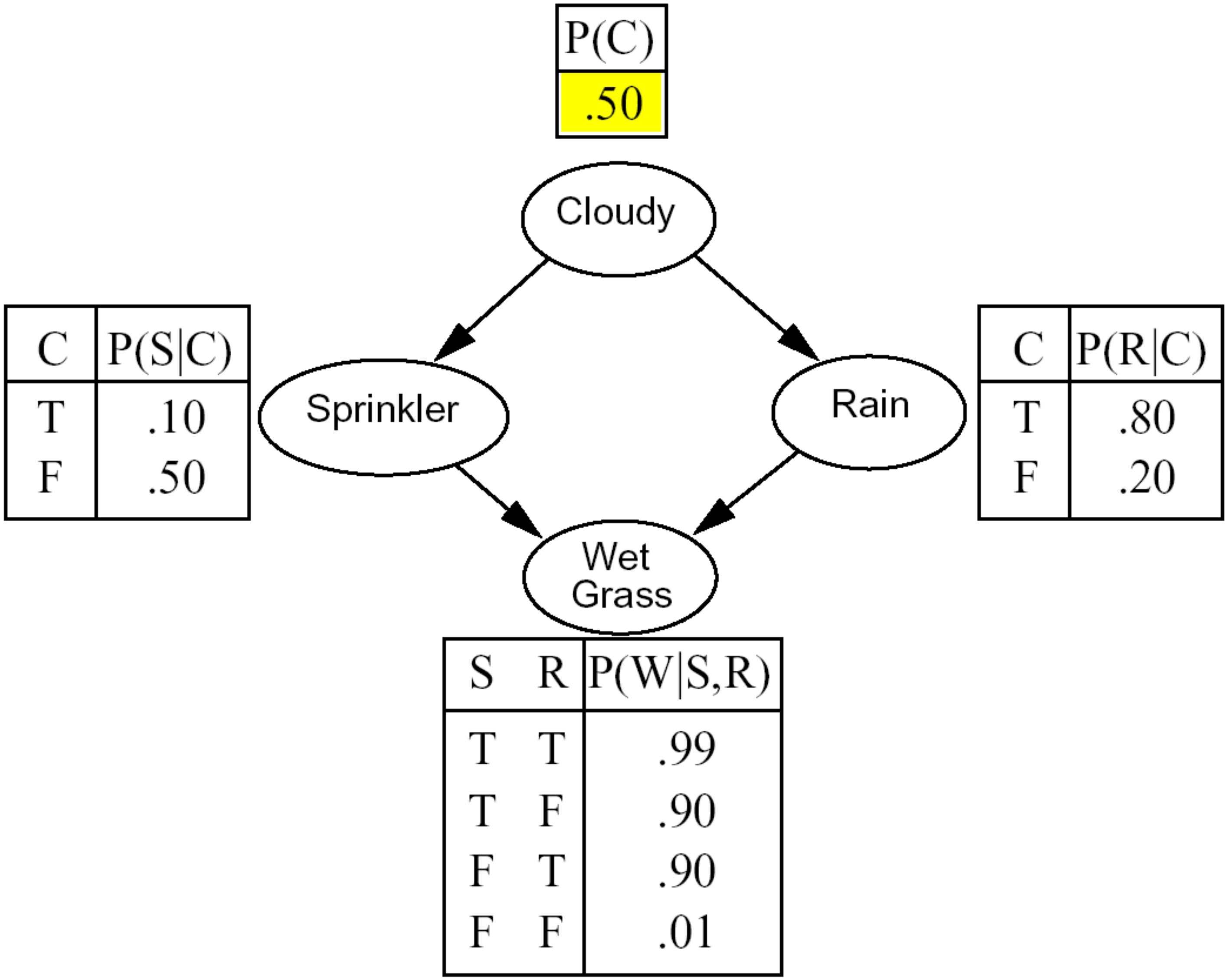


B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J|A)
T	.90
F	.05

A	P(M|A)
T	.70
F	.01

Example 2



P(C)
.50

C	P(S|C)
T	.10
F	.50

C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

Joint Probability Distribution

- Let $X_1, X_2, \ldots, X_n$ be the variables in the Bayesian network.
- Let ω be a assignment of values to variables. Let $\omega(X)$ be the value assigned to X.
- Let $parents(X)$ be the parents of X in the Bayesian network. $parents(X) = \emptyset$ if X has no parents.
- The joint probability distribution of a Bayesian network is specified by:

$$P(\omega) = \prod_{i=1}^{n} P(X_i = \omega(X_i) \mid parents(X_i) = \omega(parents(X_i)))$$

Examples

- Suppose $\neg B, E, A, J, \neg M$ (no burglary, an earthquake, an alarm, John calls, Mary doesn't call)

$$P(\neg B, E, A, J, \neg M) =$$
$$P(\neg B)P(E)P(A\,|\,\neg B, E)P(J\,|\,A)P(\neg M\,|\,A)$$
$$= (0.999)(0.002)(0.29)(0.9)(0.3)$$

- Suppose $\neg C, S, \neg R, W$ (not cloudy, sprinkler was on, no rain, wet grass)

$$P(\neg C, S, \neg R, W)$$
$$= P(\neg C)P(S \mid \neg C)P(\neg R \mid C)P(W \mid S, \neg R)$$
$$= (0.5)(0.5)(0.8)(0.9)$$

Model Size

- Suppose n variables, each with k possible values.
- Defining a joint probability table requires k^n probabilities ($k^n - 1$ to be exact).
- In a Bayesian network, a variable with j parents requires a table of k^{j+1} probabilities ($(k-1) * k^j$ to be exact).
- If no variable has has more than j parents, then less than $n * k^{j+1}$ probabilities are required.
- Number of probabilities reduced from exponential in n to linear in n and exponential in j.

Construction

To represent a problem in a Bayesian network:

- What are the relevant variables?
  - What will you observe?
  - What would you like to infer?
  - Are there hidden variables that would make the model simpler?
- What are the possible values of the variables?
- What is the relationship between them? A cause should be a parent of what it directly affects.
- How does the value of each variable depend on its parents, if any? This is expressed by prior and conditional probabilities.

Brute Force

Brute force calculation of $P(h \mid e)$ is done by:

1. Apply the conditional probability rule.

   $$P(h \mid e) = P(h \wedge e) \,/\, P(e)$$

2. Determine which values in the joint probability table are needed.

   $$P(e) = \sum_{\omega \models e} P(\omega)$$

3. Apply the joint probability distribution for Bayesian networks.

   $$P(\omega) = \prod_{i=1}^{n} P(X_i = \omega(X_i) \mid parents(X_i) = \omega(parents(X_i)))$$

Example Calculation, Part 1

Calculate $P(W \mid C, \neg R)$ in the cloudy example.

1. Apply the conditional probability rule.

$$P(W \mid C, \neg R) = \frac{P(W, C, \neg R)}{P(C, \neg R)}$$

2. Determine which values in the joint probability table are needed.

$$P(C, S, \neg R, W)$$
$$P(C, \neg S, \neg R, W)$$
$$P(C, S, \neg R, \neg W)$$
$$P(C, \neg S, \neg R, \neg W)$$

Example Calculation, Part 2

3. Apply the joint probability distribution for Bayesian networks.

$$P(C, S, \neg R, W) = (0.5)(0.1)(0.2)(0.9) = 0.009$$
$$P(C, \neg S, \neg R, W) = (0.5)(0.9)(0.2)(0.01) = 0.0009$$
$$P(C, S, \neg R, \neg W) = (0.5)(0.1)(0.2)(0.1) = 0.001$$
$$P(C, \neg S, \neg R, \neg W) = (0.5)(0.9)(0.2)(0.99) = 0.0891$$

$$P(W \mid C, \neg R) = \frac{P(W, C, \neg R)}{P(C, \neg R)}$$

$$= \frac{0.009 + 0.0009}{0.009 + 0.0009 + 0.001 + 0.0891} = 0.099$$

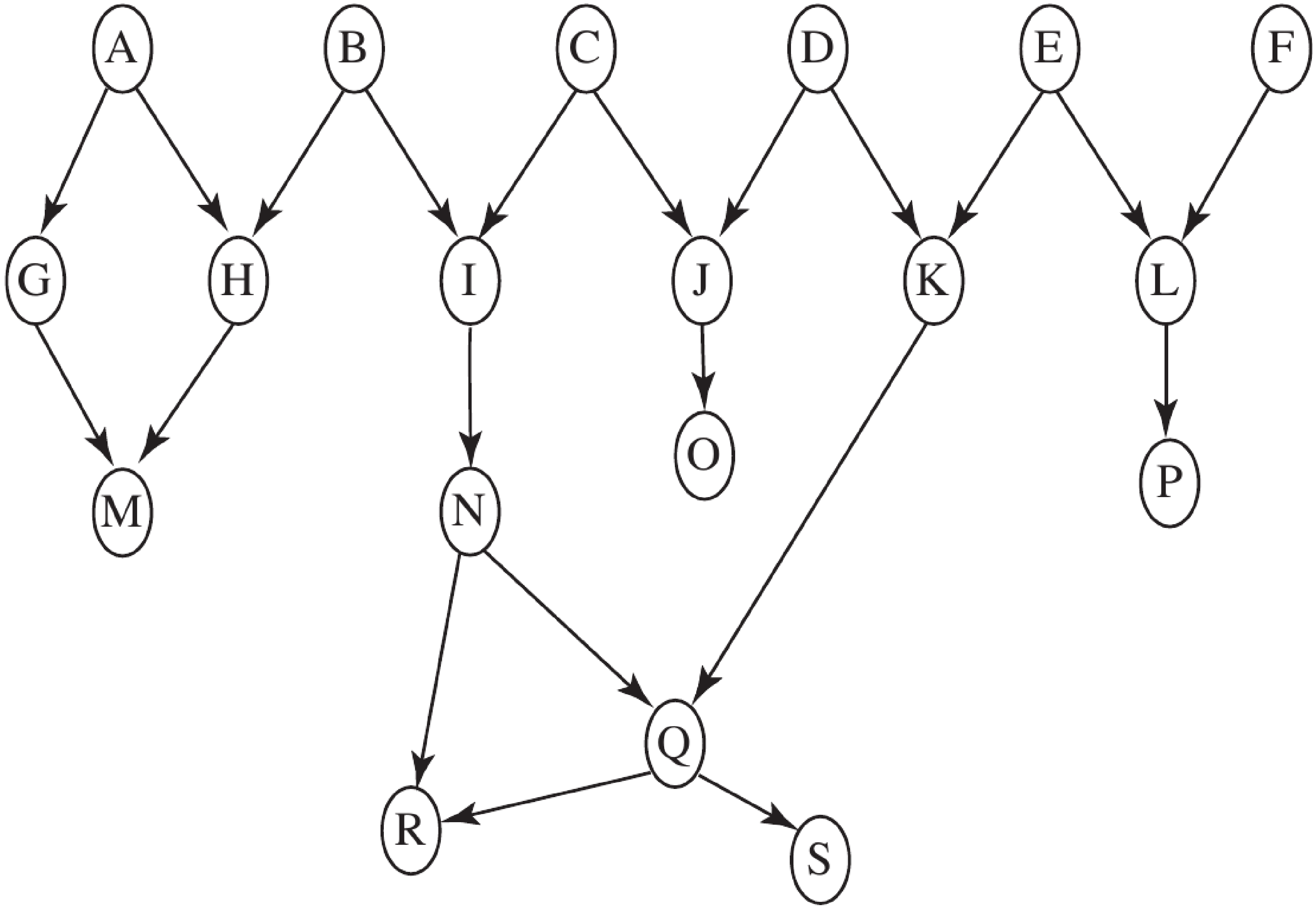
Pruning Irrelevant Variables

Some variables might not be relevant to $P(h \mid e)$.

- Prune any variables that have no observed or queried descendents, that is, not part of e or h.
- Connect the parents of any observed variable.
- Remove arc directions and observed variables.
- Prune any variables not connected to h in the (undirected) graph.
- Calculate $P(h \mid e)$ in original network minus pruned variables.

In example on the next slide, compare $P(H \mid Q, P)$ and $P(H \mid K, Q, P)$.

Example



Variable Elimination Algorithm

Variable elimination is an exact algorithm for computing $P(h \mid e)$. Assume h is one variable.

- Convert each probability table into a *factor*. Let F be all the factors.
- Eliminate each non-query variable X in turn:
  - Identify the factors F' in which X appears.
  - If X is observed, *set* X to the observed value in each factor in F'.
  - Otherwise, *multiply* the factors in F' together, *sum out* X, add the result to F, and remove F' from F.
- *Multiply* the remaining factors and *normalize*.

Convert to Factors

These are the factors of the cloudy network.

C	val
T	0.5
F	0.5

C	S	val
T	T	0.1
T	F	0.9
F	T	0.5
F	F	0.5

C	R	val
T	T	0.8
T	F	0.2
F	T	0.2
F	F	0.8

S	R	W	val
T	T	T	0.99
T	T	F	0.01
T	F	T	0.90
T	F	F	0.10
F	T	T	0.90
F	T	F	0.10
F	F	T	0.01
F	F	F	0.99

Want $P(W \mid C, \neg R)$.

Set Operation

Setting C to T and R to F.

C	val
T	0.5
~~F~~	~~0.5~~

C	S	val
T	T	0.1
T	F	0.9
~~F~~	~~T~~	~~0.5~~
~~F~~	~~F~~	~~0.5~~

C	R	val
~~T~~	~~T~~	~~0.8~~
T	F	0.2
~~F~~	~~T~~	~~0.2~~
~~F~~	~~F~~	~~0.8~~

S	R	W	val
~~T~~	~~T~~	~~T~~	~~0.99~~
~~T~~	~~T~~	~~F~~	~~0.01~~
T	F	T	0.90
T	F	F	0.10
~~F~~	~~T~~	~~T~~	~~0.90~~
~~F~~	~~T~~	~~F~~	~~0.10~~
F	F	T	0.01
F	F	F	0.99

Set Operation

Finish set operation by removing C and R columns.

val
0.5

S	val
T	0.1
F	0.9

val
0.2

S	W	val
T	T	0.90
T	F	0.10
F	T	0.01
F	F	0.99

Multiply and Sum Out

Multiply tables containing S. Result will have all the variables in the old tables.

S	val
T	0.1
F	0.9

$\times$

S	W	val
T	T	0.90
T	F	0.10
F	T	0.01
F	F	0.99

$=$

S	W	val
T	T	$.1(.90) = 0.090$
T	F	$.1(.10) = 0.010$
F	T	$.9(.01) = 0.009$
F	F	$.9(.99) = 0.891$

Sum out S =

W	val
T	$.090 + .009 = 0.099$
F	$.010 + .891 = 0.901$

Multiply and Normalize

Multiply remaining tables (only h is left).

val
0.5

$\times$

val
0.2

$\times$

W	val
T	0.099
F	0.901

$=$

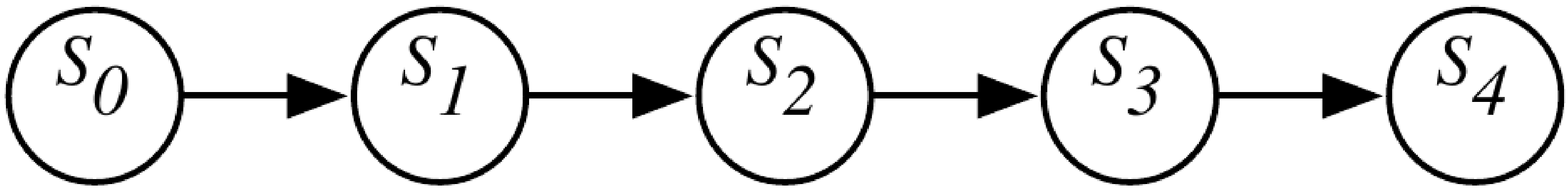
W	val
T	0.0099
F	0.0901

Normalize $=$

W	val
T	0.099
F	0.901

Markov Chains

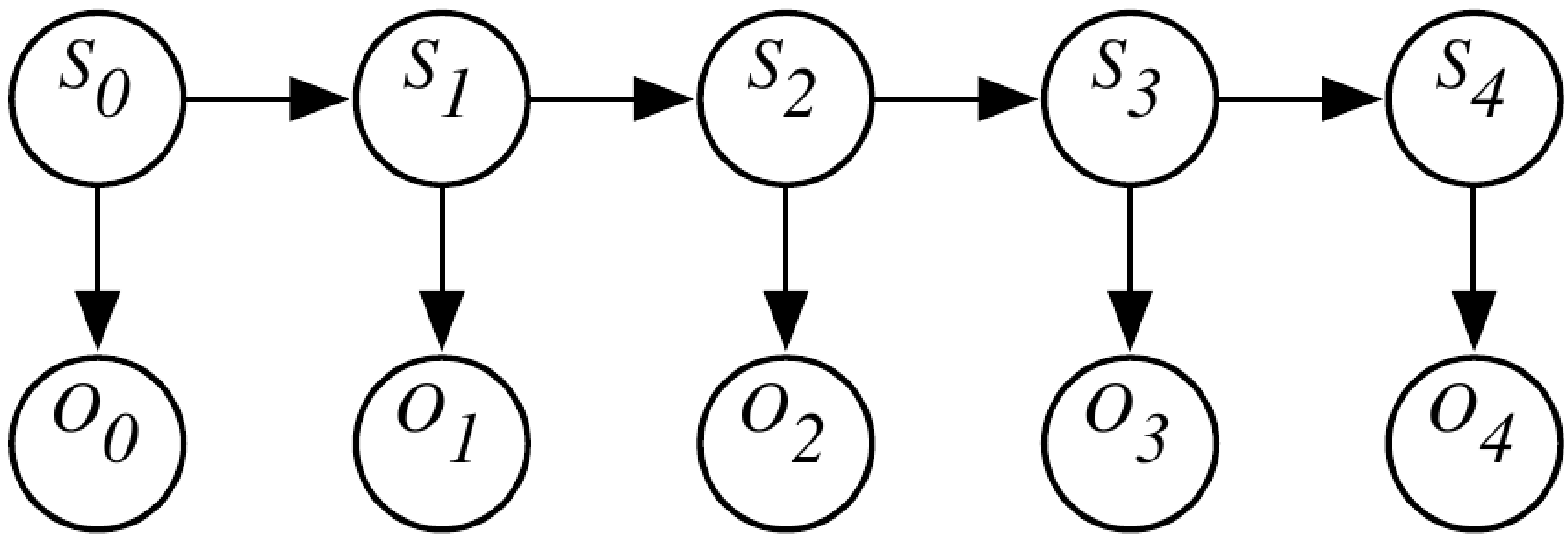
A *Markov chain* is a Bayesian network for representing a sequence of values.



- This represents the *Markov assumption*: $P(S_{t+1} \mid S_0, \ldots, S_t) = P(S_{t+1} \mid S_t)$
- A *stationary* Markov chain has $P(S_{t+1} \mid S_t) = P(S_1 \mid S_0)$
- Sequence of states over time, e.g., queueing theory.
- Probability of next item (word, letter) given previous item.

Hidden Markov Models

A *hidden Markov model* adds observations to a Markov chain.



- States are not directly observable.
- Observations provide evidence for states.
- $P(S_0)$ specifies initial conditions.
- $P(S_{t+1} \mid S_t)$ specifies the dynamics.
- $P(O_t \mid S_t)$ specifies the sensor model.

HMM Probabilities

- To compute $P(S_i \mid O_0, \ldots, O_i, \ldots, O_k)$
  - Compute variable elimination forward. This computes $P(S_i \mid O_0, \ldots, O_{i-1})$.
  - Compute variable elimination backward. This computes $P(O_i, \ldots, O_k \mid S_i)$.
  - $P(S_i \mid O_0, \ldots, O_k) \propto P(S_i \mid O_0, \ldots, O_{i-1}) * P(O_i, \ldots, O_k \mid S_i)$
- To compute most probable sequence of states:
  - Viterbi algorithm.
  - Idea: Find most probable sequence to S_i.
  - Use that information to extend sequence by one state, and repeat.