Probabilistic Inference and Belief Networks

Uncertain Knowledge and Reasoning

[related reading: chapter 8 of the AI textbook]

Probabilities everywhere

- Not just for games of chance!
  - I’m snuffling: am I sick?
  - Email contains “FREE!”: is it spam?
  - Tooth hurts: have cavity?
  - Safe to cross street?
  - 60 min enough to get to the airport?
  - Robot rotated wheel three times, how far did it advance?

- Reasons for uncertainty and randomness:
  - Theoretical and modeling limitations
    - Coin toss example: we may not have a complete model for physics of the environment (e.g. molecular structure of the coin, micro air movements, ...)
  - Sensory and measurement limitations
    - Coin toss example: we have a physical model that requires as input very precise measurements that are not available (e.g. how much energy exactly the coin received, the current state of the environment).
  - Computational limitations
    - Coin toss example: assuming that the model and input data are available, calculations might be too time consuming.

Random Variables

- A random variable is some aspect of the world about which we have uncertainty
  - R = Is it raining?
  - D = How long will it take to drive to work?
  - L = Where am I?
- We denote random variables with capital letters. For their values we use lower case.
- Each random variable has a domain
  - R in {true, false}
  - D in [0, infinity)
  - L in possible locations

Probability distributions

- Unobserved random variables have distributions

$P(T)$

T	P
warm	0.5
cold	0.5

$P(W)$

W	P
sun	0.6
rain	0.1
fog	0.3

- A distribution is a TABLE of probabilities of values
- A probability (lower case value) is a single number

$$P(W = rain) = 0.1 \qquad P(rain) = 0.1$$

- Must have: $\forall x\, P(x) \geq 0$ $\qquad \sum_x P(x) = 1$

Joint distributions

- A *joint distribution* over a set of random variables: $X_1, X_2, \ldots X_n$ is a map from assignments (or *outcomes*, or *atomic events*) to reals:

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

$$P(x_1, x_2, \ldots x_n)$$

  - Size of distribution if n variables with domain sizes d?

- Must obey:

$$0 \leq P(x_1, x_2, \ldots x_n) \leq 1$$

$$\sum_{(x_1, x_2, \ldots x_n)} P(x_1, x_2, \ldots x_n) = 1$$

T	S	P
warm	sun	0.4
warm	rain	0.1
cold	sun	0.2
cold	rain	0.3

Events

T	S	P
warm	sun	0.4
warm	rain	0.1
cold	sun	0.2
cold	rain	0.3

- An *event* is a set E of assignments (or outcomes)

$$P(E) = \sum_{(x_1 \ldots x_n) \in E} P(x_1 \ldots x_n)$$

- From a joint distribution, we can calculate the probability of any event

- Probability that it's warm AND sunny?

- Probability that it's warm?

- Probability that it's warm OR sunny?

Marginalization

- Marginalization (or summing out) is *projecting* a joint distribution to a sub-distribution over subset of variables

$$P(X_1 = x_1) = \sum_{x_2} P(X_1 = x_1, X_2 = x_2)$$

$P(T, S)$

T	S	P
warm	sun	0.4
warm	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P(t) = \sum_{s} P(t, s)$$

$P(T)$

T	P
warm	0.5
cold	0.5

$$P(s) = \sum_{t} P(t, s)$$

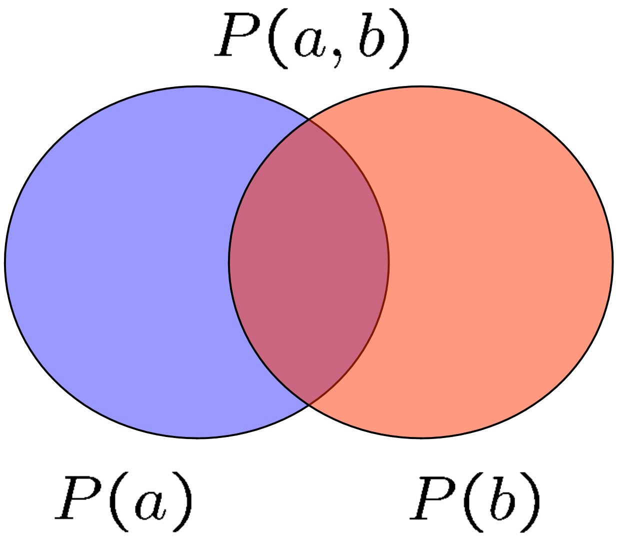
$P(S)$

S	P
sun	0.6
rain	0.4

Conditional Probabilities

- A conditional probability is the probability of an event given another event (usually called evidence)

$$P(a|b) = \frac{P(a,b)}{P(b)}$$



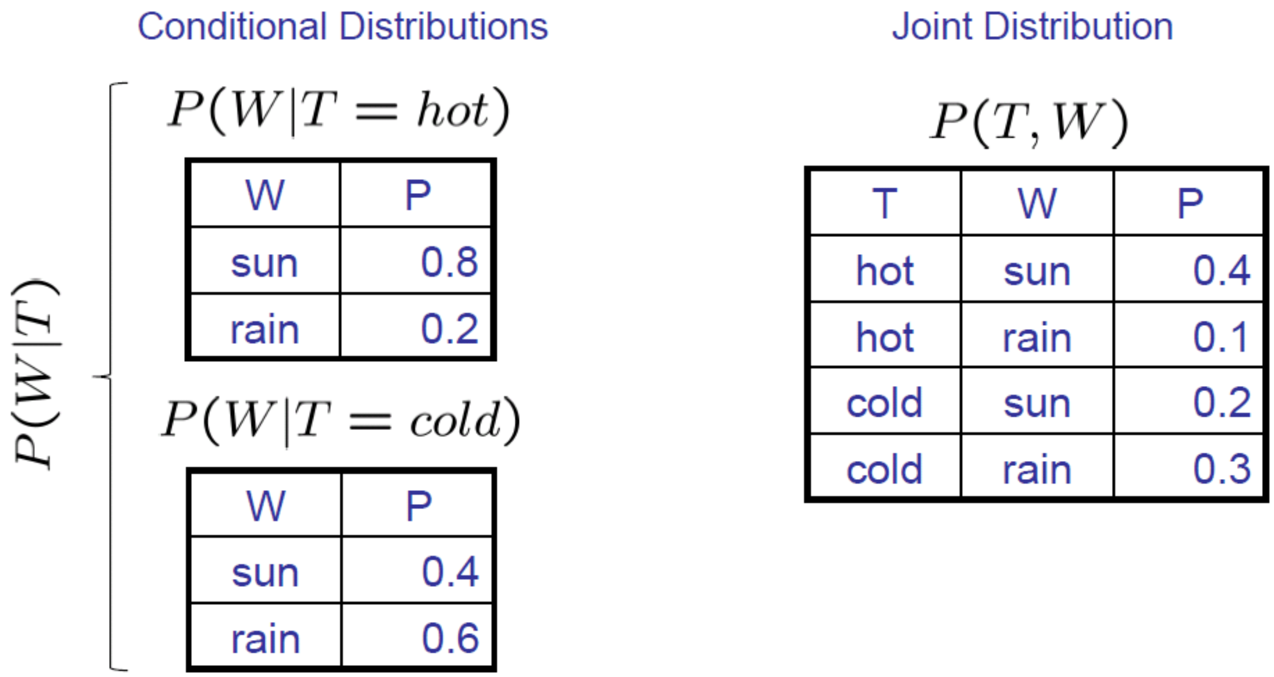
$P(T,S)$

T	S	P
warm	sun	0.4
warm	rain	0.1
cold	sun	0.2
cold	rain	0.3

P(rain | cold) =

Conditional Distributions

- Conditional distributions are probability distributions over some variables given fixed values of others
- Example and notation:
  - P(*hot* | sun) = a single number
  - P(W, T) = a table with 2 x 2 rows summing to 1
  - P(W | T) = Two 2-row tables, each summing to 1



Conditional Distributions

$P(W|T)$

$P(W|T = hot)$

W	P
sun	0.8
rain	0.2

$P(W|T = cold)$

W	P
sun	0.4
rain	0.6

Joint Distribution

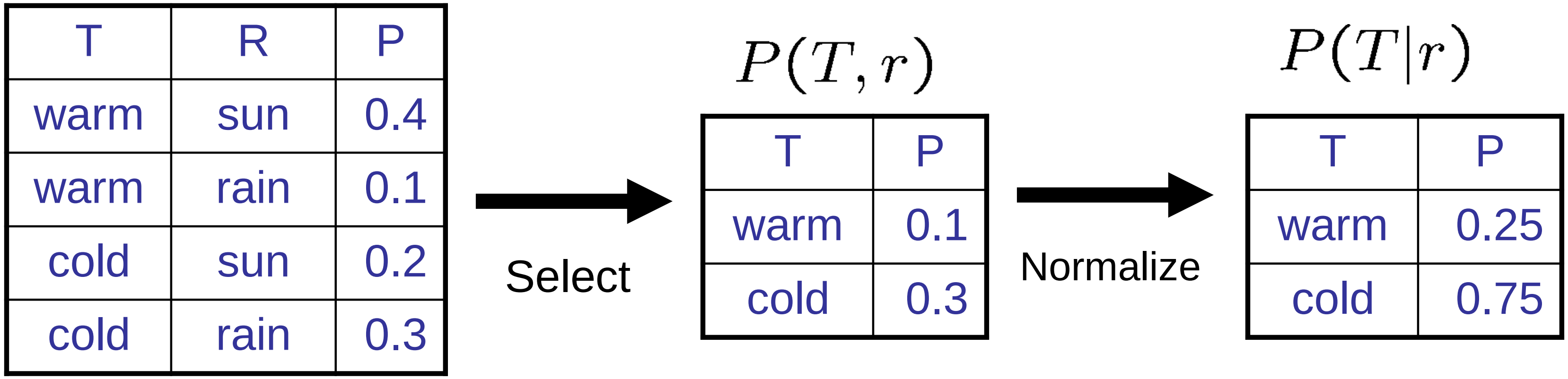
$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

Normalization Trick

- A trick to get the whole conditional distribution at once:
  - Select the joint probabilities matching the evidence
  - Normalize the selection (divide by total sum so they sum to 1)

- Example: find P(T|rain)



T	R	P
warm	sun	0.4
warm	rain	0.1
cold	sun	0.2
cold	rain	0.3

Select

$P(T, r)$

T	P
warm	0.1
cold	0.3

Normalize

$P(T|r)$

T	P
warm	0.25
cold	0.75

$$P(x_1|x_2) = \frac{P(x_1, x_2)}{P(x_2)} = \frac{P(x_1, x_2)}{\sum_{x_1} P(x_1, x_2)}$$

The Product Rule

- Sometimes joint P(X,Y) is easy to get
- Sometimes easier to get conditional P(X|Y)

$$P(x|y) = \frac{P(x,y)}{P(y)} \qquad \Longleftrightarrow \qquad P(x,y) = P(x|y)P(y)$$

The Chain Rule

- More generally, can always write any joint distribution as an incremental product of conditional distributions

$$P(x_1, x_2, x_3) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2)$$

$$P(x_1, x_2, \ldots x_n) = \prod_i P(x_i|x_1 \ldots x_{i-1})$$

Probabilistic Inference

- Probabilistic inference: compute a desired probability from other known probabilities (e.g. conditional from joint)

- We generally compute conditional probabilities
  - P(on time | no reported accidents) = 0.90
  - These represent the agent’ s *beliefs* given the evidence

- Probabilities change with new evidence:
  - P(on time | no accidents, 5 a.m.) = 0.95
  - P(on time | no accidents, 5 a.m., raining) = 0.80
  - Observing new evidence causes *beliefs to be updated*

Inference by Enumeration

- We want: $P(Y_1 \dots Y_m | e_1 \dots e_k)$

  - Evidence variables: $(E_1 \dots E_k) = (e_1 \dots e_k)$
  - Query variables: $Y_1 \dots Y_m$
  - Hidden variables: $H_1 \dots H_r$

  $\left.\begin{array}{l} \\ \\ \\ \end{array}\right\} \quad X_1, X_2, \dots X_n$

  All variables

- First, select the entries consistent with the evidence
- Second, sum out H:

$$P(Y_1 \dots Y_m, e_1 \dots e_k) = \sum_{h_1 \dots h_r} P(\underbrace{Y_1 \dots Y_m, h_1 \dots h_r, e_1 \dots e_k}_{X_1, X_2, \dots X_n})$$

- Finally, normalize the remaining entries.

- Obvious problems:
  - Worst-case time complexity $O(d^n)$
  - Space complexity $O(d^n)$ to store the joint distribution

Inference by Enumeration

- P(R)?

- P(sun | winter)?

  or equivalently P(R=sun | winter)?

- P(R | winter, warm)?

S	T	R	P
summer	warm	sun	0.30
summer	warm	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	warm	sun	0.10
winter	warm	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.20

Probabilistic Models

- Models describe how (a portion of) the world works

- Models are always simplifications
  - May not account for every variable
  - May not account for all interactions between variables
  - “All models are wrong; but some are useful.”
    - – George E. P. Box

- A joint distribution is a probabilistic model.

- What do we do with probabilistic models?
  - We (or our agents) need to reason about unknown variables, given evidence
  - Example: explanation (diagnostic reasoning)
  - Example: prediction (causal reasoning)

Complexity of Models

- Engineers and designers are interested in *simple* and *compact* models (as long as the model is sufficiently good)
  - Simple models are easier to build
  - Simple models are easier to explain (e.g. why/how they work)
  - Compact models take less space
  - Simplicity/Compactness usually implies more efficient computation (lower time complexity)
- One way of measuring the complexity of a probabilistic model is to count the number of (free) parameters (values) that must be specified.

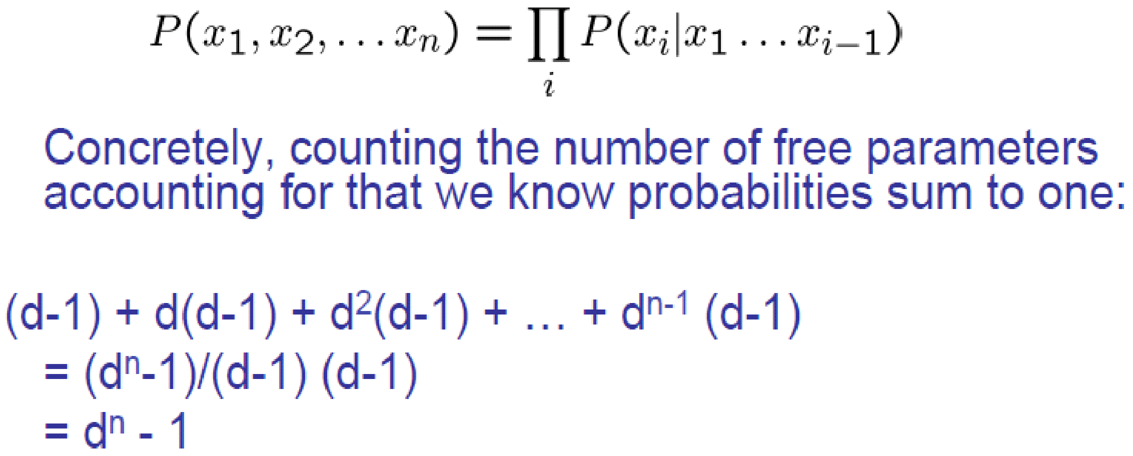
A joint distribution over *n* variables whose domain sizes are *d* (each can take d distinct values) requires d^n entries in the table.

The number of *(free) parameters* is $d^n - 1$. [Because once you specify d^n-*1* of the entries, the last one is (1 – sum of those specified) because the entries should add up to 1.]

Complexity of Models (cont’d)

If a probabilistic model has multiple tables (distributions), the number of its free parameters is the sum of the number of free parameters of the tables.

Question: How many free parameters would we need if instead of a full joint distribution, we used multiple smaller distributions using the chain rule?



$$P(x_1, x_2, \ldots x_n) = \prod_i P(x_i | x_1 \ldots x_{i-1})$$

Concretely, counting the number of free parameters accounting for that we know probabilities sum to one:

$(d-1) + d(d-1) + d^2(d-1) + \ldots + d^{n-1}\,(d-1)$
$= (d^n-1)/(d-1)\,(d-1)$
$= d^n - 1$

It doesn’t make a difference. (i.e. using the chain rule alone doesn’t reduce complexity.)

Independence

- Two variables are *independent* if:

$$P(X, Y) = P(X)P(Y)$$

$$\forall x, y\, P(x, y) = P(x)P(y)$$

$$X \perp\!\!\!\perp Y$$

  - This says that their joint distribution *factors* into a product of two simpler distributions
- We can use independence as a *modeling assumption*
  - Independence can be a simplifying assumption

- How many parameters in the joint model using one table?
- How many parameters when assuming independence?

Example: Independence

- N fair, independent coin flips:

$P(X_1)$

H	0.5
T	0.5

$P(X_2)$

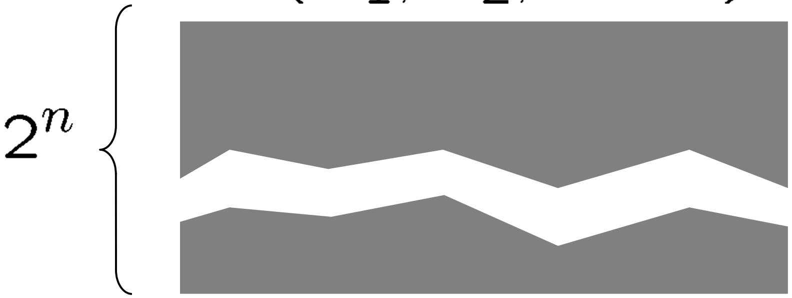
H	0.5
T	0.5

$\ldots$

$P(X_n)$

H	0.5
T	0.5

$P(X_1, X_2, \ldots X_n)$



Conditional Independence

- Unconditional (absolute) independence is very rare.

- Conditional independence:

$$\forall x, y, z : P(x, y|z) = P(x|z)P(y|z)$$
$$\forall x, y, z : P(x|z, y) = P(x|z)$$

$$X \perp\!\!\!\perp Y|Z$$

Conditional Independence: Example

- If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:
  - P(catch | toothache, cavity) = P(catch | cavity)
- The same independence holds if I don’t have a cavity:
  - P(catch | toothache, –cavity) = P(catch| –cavity)
- Catch is *conditionally independent* of Toothache given Cavity:
  - P(Catch | Toothache, Cavity) = P(Catch | Cavity)
- Equivalent statements:
  - P(Toothache | Catch , Cavity) = P(Toothache | Cavity)
  - P(Toothache, Catch | Cavity) = P(Toothache | Cavity) P(Catch | Cavity)
  - One can be derived from the other easily

Chain Rule and Conditional Independence

- How many entries/parameters do we need for the joint distribution P(Toothache, Cavity, Catch)?

  7 independent entries (2^3 - 1)

- Write out full joint distribution using chain rule:
  - P(Toothache, Catch, Cavity)

    = P(Toothache | Catch, Cavity) P(Catch, Cavity)

    = P(Toothache | Catch, Cavity) P(Catch | Cavity) P(Cavity)

    = P(Toothache | Cavity) P(Catch | Cavity) P(Cavity)

- How many (free) parameters does it need now?

  2 + 2 +1 = 5

  (i.e. by assuming conditional independence, the complexity is reduced.)

Belief Networks: Big Picture

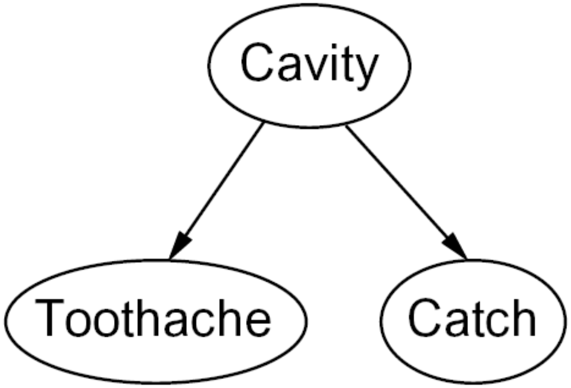
- Two problems with using full joint distribution tables for probabilistic models:
  - Unless there are only a few variables, the joint is WAY too big to represent explicitly
  - Hard to learn (estimate) anything empirically about more than a few variables at a time

- Belief nets: a technique for describing complex joint distributions (models) using simple, local distributions (conditional probabilities)
  - More properly called graphical models
  - We describe how variables locally interact
  - Local interactions chain together to give global, indirect interactions
  - Also known as Bayes' nets or Bayesian networks

Graphical Model Notation

- Nodes: variables (with domains)
  - Can be assigned (observed) or unassigned (unobserved)

- Arcs: influences
  - Allow dependence between variables
  - For now: imagine that arrows mean causation (in general, they don’t have to)



Example: Coin Flips

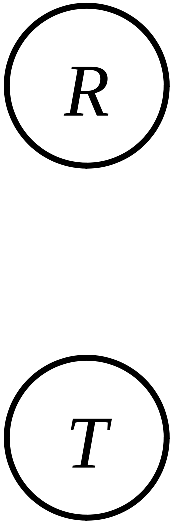
- N independent coin flips

X_1
X_2
$\cdots$
X_n

- No interactions between variables: absolute independence

Example: Traffic

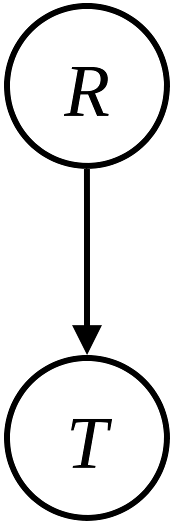
- Variables:
  - R: It rains
  - T: There is traffic



- Model 1: independent

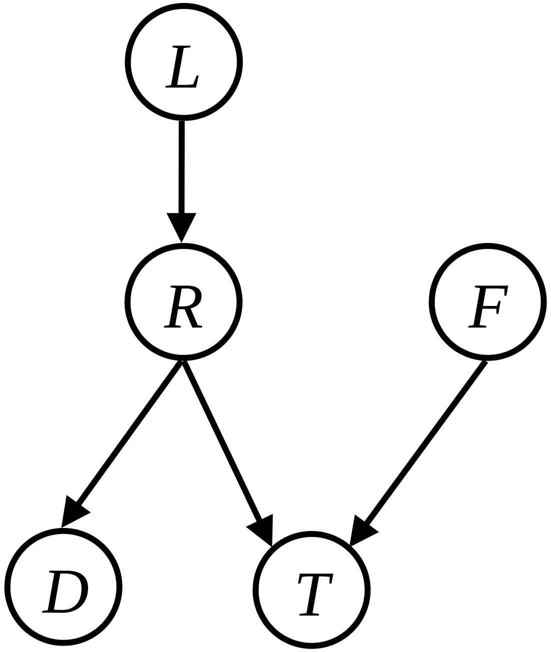
Example: Traffic

- Variables:
  - R: It rains
  - T: There is traffic
- Model 2: rain causes traffic
- An agent using model 2 may perform better.



Example: Traffic II

- Variables
  - T: Traffic
  - R: It rains
  - L: Low pressure
  - D: Roof drips
  - F: Festival



Example: Alarm network

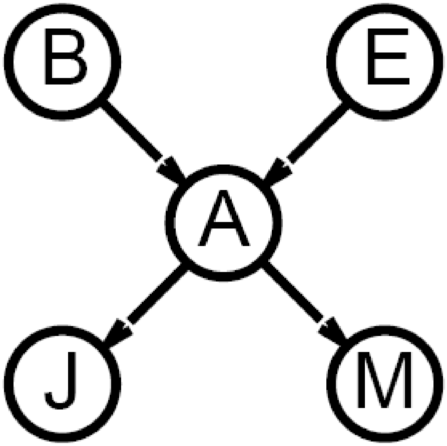
You have a new burglar alarm installed at home. It is fairly reliable at detecting a burglary, but also responds on occasion to minor earthquakes.

You also have two neighbors, John and Mary, who have promised to call you at work when they hear the alarm. John nearly always calls when he hears the alarm, but sometimes confuses the telephone ringing with the alarm and calls then, too. Mary, on the other hand, likes rather loud music and often misses the alarm altogether.

- Example inference task in this domain:
  - Given the evidence of who has or has not called, we would like to estimate the probability of a burglary.
  - We would like to know the probability of the alarm going off.

Example: Alarm network

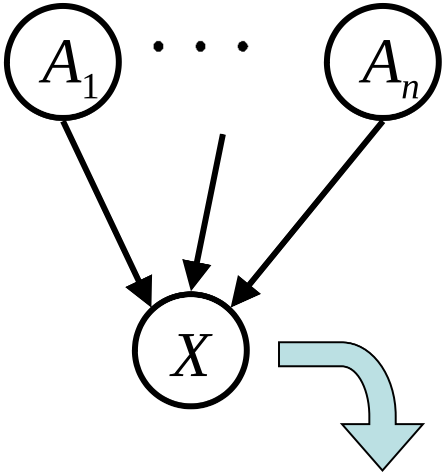
- Variables
  - B: Burglary
  - A: Alarm goes off
  - M: Mary calls
  - J: John calls
  - E: Earthquake



Belief Net Semantics

A belief network is:

- A set of nodes, one per random variable
- A directed, acyclic graph
- A collection of distributions (CPTs) over each node, one for each combination of parents' values



$$P(X|a_1 \ldots a_n)$$

$$P(X|A_1 \ldots A_n)$$

- CPT: conditional probability table
- Description of a noisy "causal" process

A belief net = Topology (graph) + Local Conditional Probabilities

Topological semantics

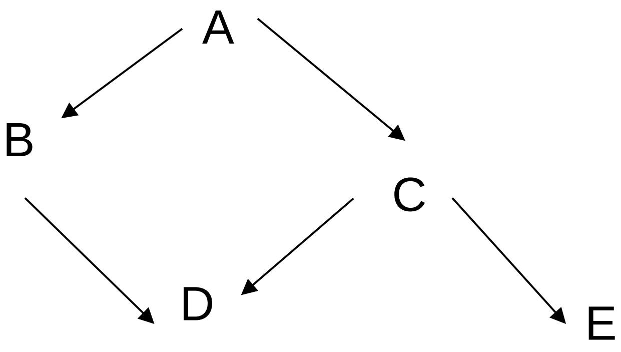
- A node is **conditionally independent** of its **non-descendants** given its **parents**

Also the following but we
won’t cover them in COSC367:

- A node is **conditionally independent** of all other nodes in the network given its parents, children, and children’s parents (also known as its **Markov blanket**)

- The method called **d-separation** can be applied to decide whether a set of nodes X is independent of another set Y, given a third set Z

Joint Probabilities in BNs: Starting with an example



Computing the joint probability for all variables is easy:

P(a, b, c, d, e)

= P(e | a, b, c, d) P(a, b, c, d)   by the product rule

= P(e | c) P(a, b, c, d)   by cond. indep. assumption

= P(e | c) P(d | a, *b, c*) P(a, b, c)

= P(e | c) P(d | b, c) P(c | *a,* b) P(a, b)

= P(e | c) P(d | b, c) P(c | a) P(b | a) P(a)

BNs implicitly encode joint distributions

- Belief nets implicitly encode joint distributions
  - As a product of local conditional distributions
  - To see what probability a BN gives to a full assignment, multiply all the relevant conditionals together:

$$P(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} P(x_i | parents(X_i))$$

- This lets us reconstruct any entry of the full joint
- Not every BN can represent every full joint
  - The topology enforces certain conditional independencies
- By having the joint distribution, we can answer any query (e.g. by using inference by enumeration)
  - Note: more advanced inference algorithms do not require constructing the entire joint.

Example: Coin Flips

$P(X_1)$

h	0.5
t	0.5

$P(X_2)$

h	0.5
t	0.5

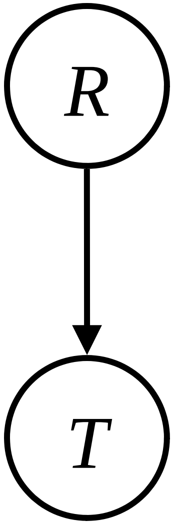
$\dots$

$P(X_n)$

h	0.5
t	0.5

P(h, t, h, h, …, t) =

Example: Traffic



$$P(R)$$

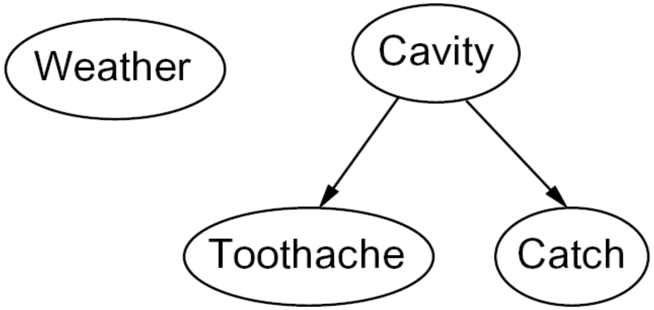
r	1/4
¬ r	3/4

$$P(T|R)$$

r	t	3/4
	¬ t	1/4
¬ r	t	1/2
	¬t	1/2

$$P(r, \neg t) =$$

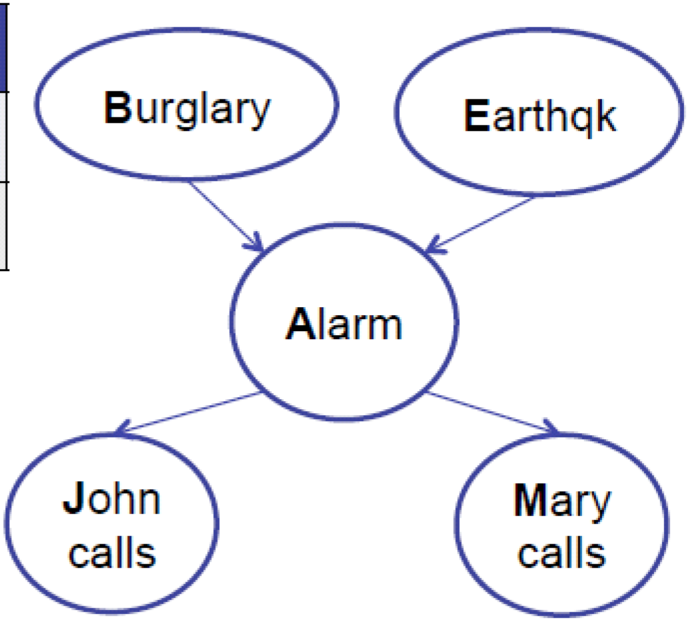
Example: Dental Health



P(¬cavity, catch, ¬toothache, weather) =
P(weather) P(¬cavity) P(¬toothache|¬cavity) P(catch|¬cavity)

Example: Alarm Network

B	P(B)
+b	0.001
¬b	0.999



E	P(E)
+e	0.002
¬e	0.998

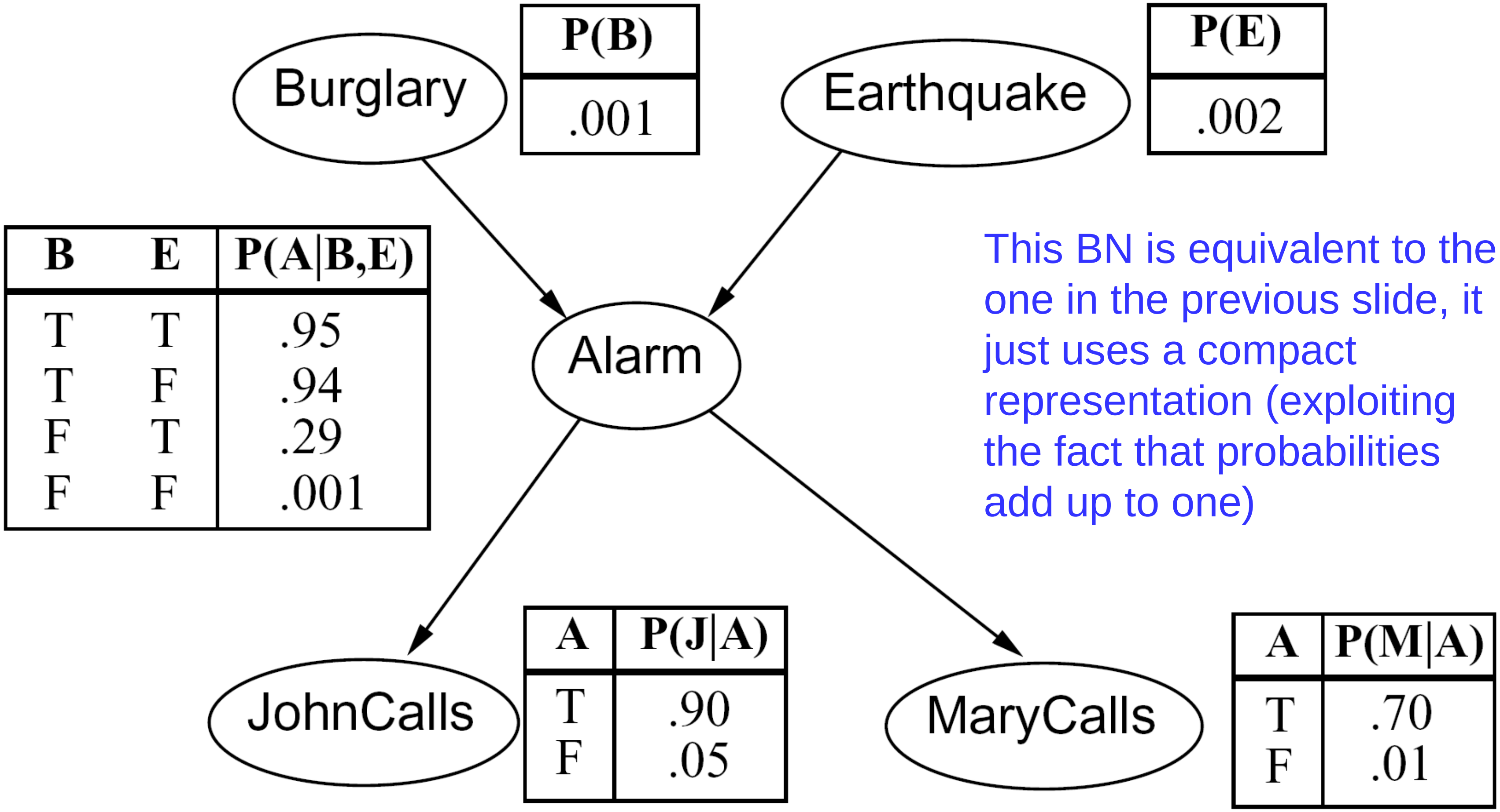
B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	¬a	0.05
+b	¬e	+a	0.94
+b	¬e	¬a	0.06
¬b	+e	+a	0.29
¬b	+e	¬a	0.71
¬b	¬e	+a	0.001
¬b	¬e	¬a	0.999

A	J	P(J|A)
+a	+j	0.9
+a	¬j	0.1
¬a	+j	0.05
¬a	¬j	0.95

A	M	P(M|A)
+a	+m	0.7
+a	¬m	0.3
¬a	+m	0.01
¬a	¬m	0.99

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

Example: Alarm Network



P(B)
.001

P(E)
.002

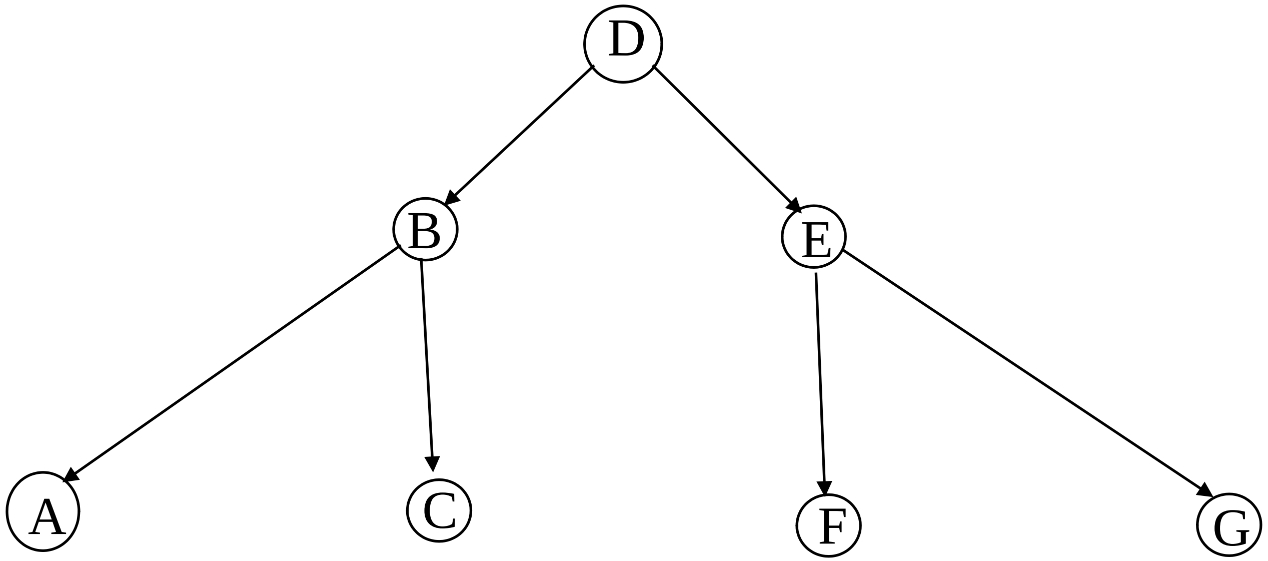
B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

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

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

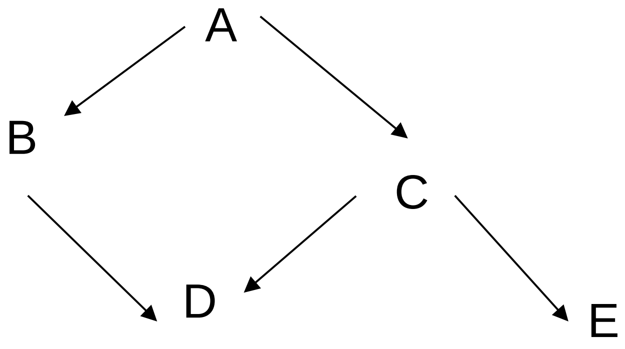
This BN is equivalent to the one in the previous slide, it just uses a compact representation (exploiting the fact that probabilities add up to one)

Example: Tree-Structured Bayesian Network



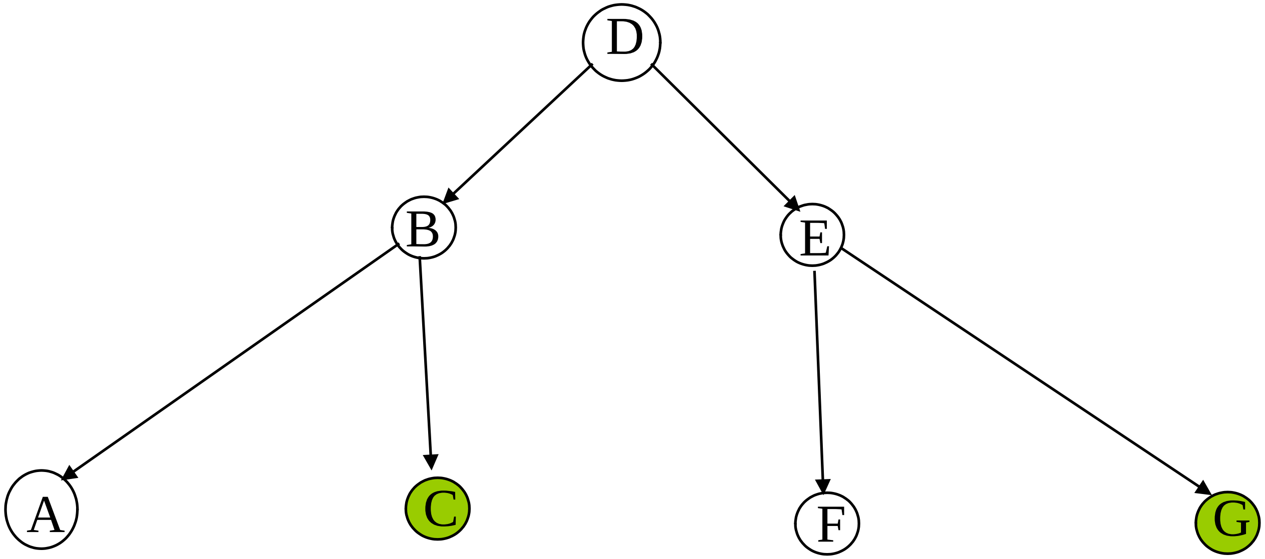
P(A, B, C, D, E, F, G) is modeled as P(A|B) P(C|B) P(F|E) P(G|E) P(B|D) P(E|D) P(D)

Computing probabilities using BN



- Any complete joint can be computed:
  - P(a,b,-c,d,-e) = P(a)P(b|a)P(-c|a)P(d|b,-c)P(-e|-c)
- Some probabilities are directly available in the CPTs. No calculation is needed:
  - P(a)
  - P(b|-a)
  - P(d|b,-c)
  - P(-e|c)
- For other cases, use inference by enumeration.

Example: Using BN and enumeration



$$P(a|c,g) = \alpha\Sigma_{BDEF} P(a,B,D,E,F,c,g)$$

$$\alpha = 1/(\Sigma_{BDEF} P(a,B,D,E,F,c,g) + \Sigma_{BDEF} P(-a,B,D,E,F,c,g))$$

Observe that inside the brackets equals to p(c,g)

Inference cases

- How to answer P(Y)?
  - There is no evidence therefore all variables except the query variable are hidden (must be summed over)

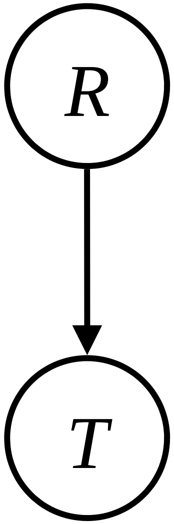
- How to answer P(y|**e**)?
  - First answer P(Y|**e**), then pick the result for Y=y

- How to answer $P(Y_1=y_1,Y_2=y_2|e)$?
  - It is $P(y_1|y_2,\mathbf{e})P(y_2|\mathbf{e})$

Further Discussion

Reverse causality?

- Consider the basic traffic net
- Let’s multiply out the joint



$P(R)$

r	1/4
- r	3/4

r	t	3/4
	-t	1/4

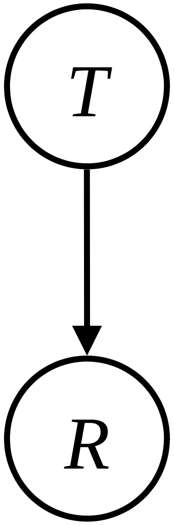
- r	t	1/2
	- t	1/2

$P(T, R)$

r	t	3/16
r	- t	1/16
- r	t	6/16
- r	- t	6/16

Reverse causality? (cont'd)

- Can we express the same joint distribution by a network where arrows no longer mean causality?
  - Yes, but the network is now harder for humans to construct and understand.



t	9/16
- t	7/16

$P(R|T)$

t	r	1/3
	- r	2/3
- t	r	1/7
	- r	6/7

r	t	3/16
r	- t	1/16
- r	t	6/16
- r	- t	6/16

Non-causal arrows?

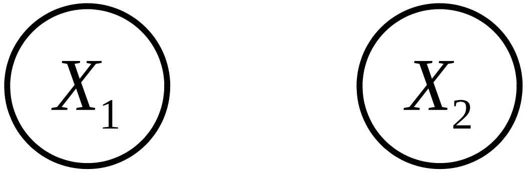
- When Belief nets reflect the true causal patterns:
  - Often simpler (nodes have fewer parents)
  - Often easier to think about
  - Often easier to elicit from experts

- BNs need not actually be causal
  - Sometimes no causal net exists over the domain (especially if variables are missing)
  - E.g. consider the variables *Traffic* and *Drips*
  - End up with arrows that reflect correlation, not causation

- What do the arrows really mean?
  - They allow dependence

Example

- For this graph, you can fiddle with the CPTs all you want, but you won’t be able to represent any distribution in which the flips are dependent!

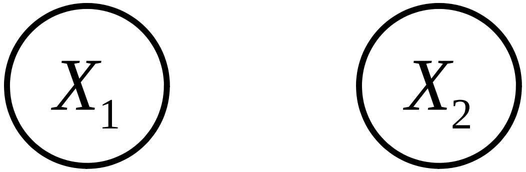


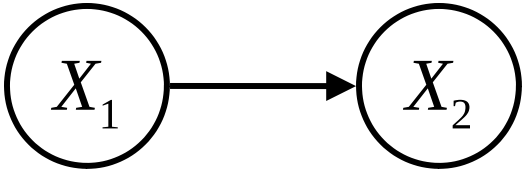
h	0.5
t	0.5

h	0.5
t	0.5

Example

- Arcs don’t prevent independence, they just allow dependence.





h	0.5
t	0.5

h	0.5
t	0.5

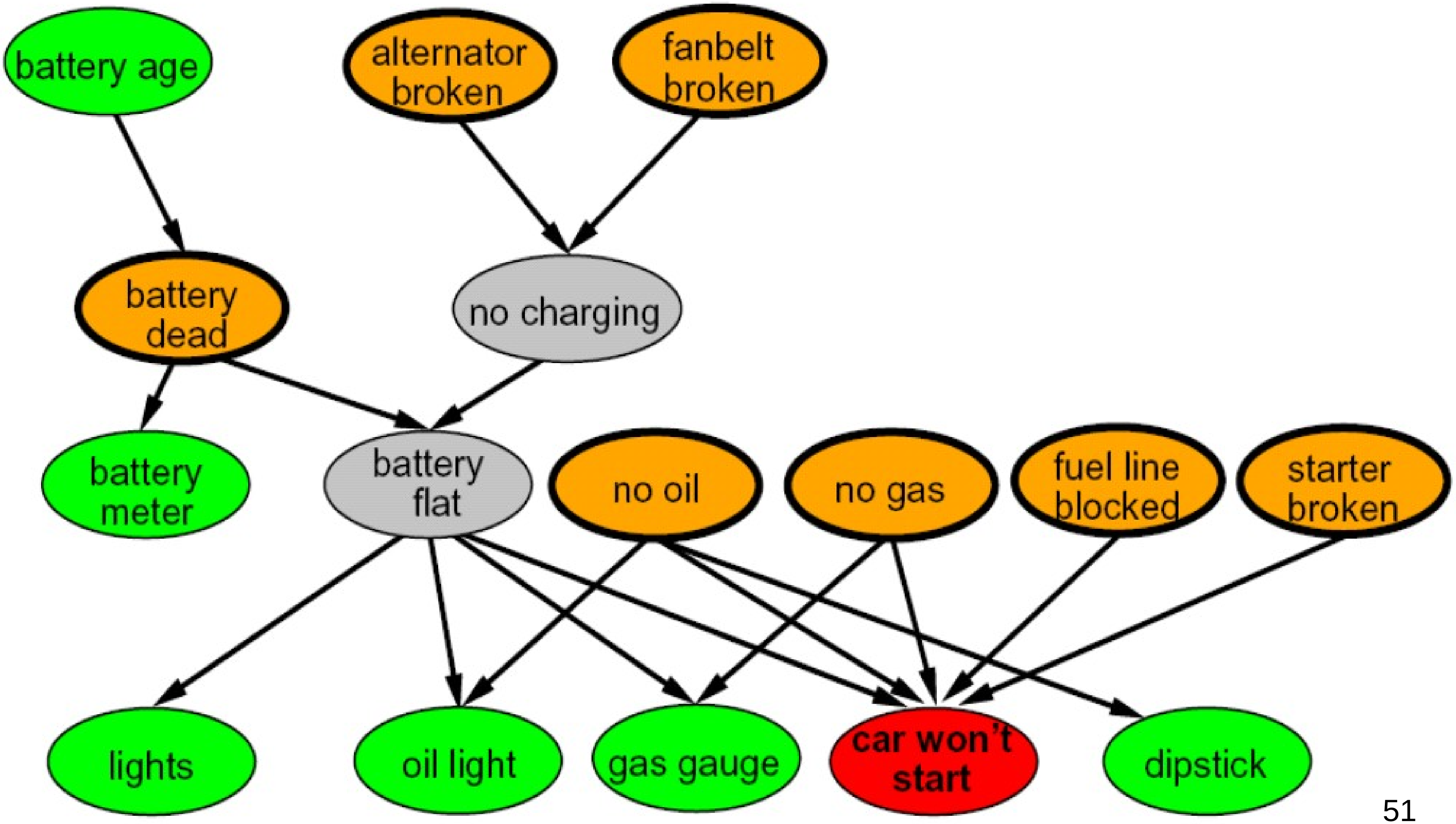
h	0.5
t	0.5

$P(X_2|X_1)$

h | h	0.5
t | h	0.5

h | t	0.5
t | t	0.5

A larger network: Car Breakdown


battery age
alternator broken
fanbelt broken
battery dead
no charging
battery meter
battery flat
no oil
no gas
fuel line blocked
starter broken
lights
oil light
gas gauge
car won't start
dipstick