CS 188: Artificial Intelligence
Spring 2008

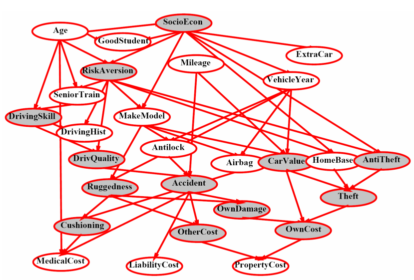
Bayes Nets
2/5/08, 2/7/08

Dan Klein – UC Berkeley

Bayes' Nets

- A Bayes' net is an efficient encoding of a probabilistic model of a domain
- Questions we can ask:
  - Inference: given a fixed BN, what is P(X | e)?
  - Representation: given a fixed BN, what kinds of distributions can it encode?
  - Modeling: what BN is most appropriate for a given domain?

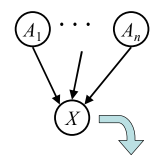
Example Bayes' Net



Bayes' Net Semantics

- A Bayes' net:
  - A set of nodes, one per variable X
  - A directed, acyclic graph
  - A conditional distribution of each variable conditioned on its parents (the *parameters* θ)

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



$P(X|A_1 \dots A_n)$

- Semantics:
  - A BN defines a joint probability distribution over its variables:

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

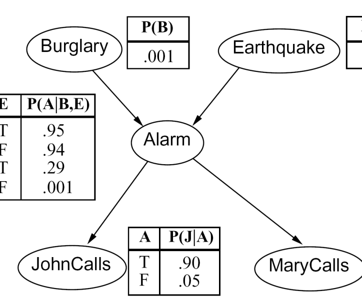
Building the (Entire) Joint

- We can take a Bayes' net and build any entry from the full joint distribution it encodes

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

  - Typically, there's no reason to build ALL of it
  - We build what we need on the fly
- To emphasize: every BN over a domain implicitly represents some joint distribution over that domain, but is specified by local probabilities

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

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

Size of a Bayes' Net

- How big is a joint distribution over N Boolean variables?
- How big is an N-node net if nodes have k parents?
- Both give you the power to calculate $P(X_1, X_2, \ldots X_n)$
- BNs: Huge space savings!
- Also easier to elicit local CPTs
- Also turns out to be faster to answer queries (next class)

Bayes' Nets

- So far: how a Bayes' net encodes a joint distribution
- Next: how to answer queries about that distribution
  - Key idea: conditional independence
  - Last class: assembled BNs using an intuitive notion of conditional independence as causality
  - Today: formalize these ideas
  - Main goal: answer queries about conditional independence and influence
- After that: how to answer numerical queries (inference)

Conditional Independence

- Reminder: independence
  - X and Y are independent if

$$\forall x, y \; P(x, y) = P(x)P(y) \dashrightarrow X \perp\!\!\!\perp Y$$

  - X and Y are conditionally independent given Z

$$\forall x, y, z \; P(x, y|z) = P(x|z)P(y|z) \dashrightarrow X \perp\!\!\!\perp Y | Z$$

  - (Conditional) independence is a property of a distribution

Example: Independence

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

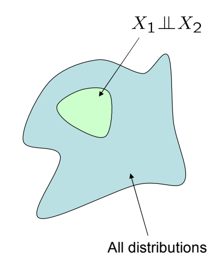
X_1 X_2

$P(X_1)$

h	0.5
t	0.5

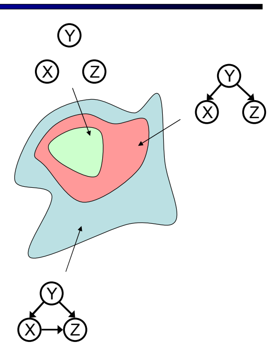
$P(X_2)$

h	0.5
t	0.5



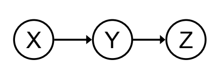
Topology Limits Distributions

- Given some graph topology G, only certain joint distributions can be encoded
- The graph structure guarantees certain (conditional) independences
- (There might be more independence)
- Adding arcs increases the set of distributions, but has several costs



Independence in a BN

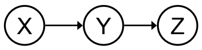
- Important question about a BN:
  - Are two nodes independent given certain evidence?
  - If yes, can calculate using algebra (really tedious)
  - If no, can prove with a counter example
  - Example:



  - Question: are X and Z independent?
    - Answer: not *necessarily*, we've seen examples otherwise: low pressure causes rain which causes traffic.
    - X can influence Z, Z can influence X (via Y)
    - Addendum: they *could* be independent: how?

Causal Chains

- This configuration is a "causal chain"



X: Low pressure
Y: Rain
Z: Traffic

$$P(x, y, z) = P(x)P(y|x)P(z|y)$$

- Is X independent of Z given Y?

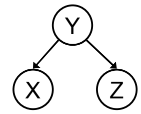
$$P(z|x, y) = \frac{P(x, y, z)}{P(x, y)} = \frac{P(x)P(y|x)P(z|y)}{P(x)P(y|x)}$$
$$= P(z|y)$$

Yes!

- Evidence along the chain "blocks" the influence

Common Cause

- Another basic configuration: two effects of the same cause
  - Are X and Z independent?
  - Are X and Z independent given Y?



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

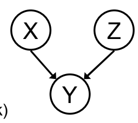
Yes!

Y: Project due
X: Newsgroup busy
Z: Lab full

  - Observing the cause blocks influence between effects.

Common Effect

- Last configuration: two causes of one effect (v-structures)
  - Are X and Z independent?
    - Yes: remember the ballgame and the rain causing traffic, no correlation?
    - Still need to prove they must be (homework)
  - Are X and Z independent given Y?
    - No: remember that seeing traffic put the rain and the ballgame in competition?
  - This is backwards from the other cases
    - Observing the effect enables influence between effects.



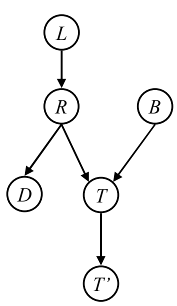
X: Raining
Z: Ballgame
Y: Traffic

The General Case

- Any complex example can be analyzed using these three canonical cases
- General question: in a given BN, are two variables independent (given evidence)?
- Solution: graph search!

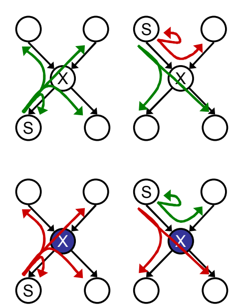
Reachability

- Recipe: shade evidence nodes
- Attempt 1: if two nodes are connected by an undirected path not blocked by a shaded node, they are conditionally independent
- Almost works, but not quite
  - Where does it break?
  - Answer: the v-structure at T doesn't count as a link in a path unless shaded



Reachability (the Bayes' Ball)

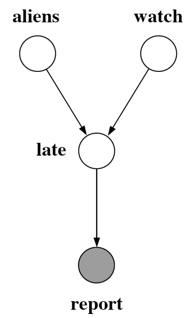
- Correct algorithm:
  - Shade in evidence
  - Start at source node
  - Try to reach target by search
  - States: pair of (node X, previous state S)
  - Successor function:
    - X unobserved:
      - To any child
      - To any parent if coming from a child
    - X observed:
      - From parent to parent
  - If you can't reach a node, it's conditionally independent of the start node given evidence



Example

$A \perp\!\!\!\perp W$ Yes

$A \perp\!\!\!\perp W | R$



Example

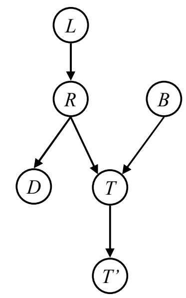
$L \perp\!\!\!\perp T' | T$ Yes

$L \perp\!\!\!\perp B$ Yes

$L \perp\!\!\!\perp B | T$

$L \perp\!\!\!\perp B | T'$

$L \perp\!\!\!\perp B | T, R$ Yes



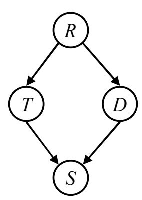
Example

- Variables:
  - R: Raining
  - T: Traffic
  - D: Roof drips
  - S: I'm sad
- Questions:

  $T \perp\!\!\!\perp D$

  $T \perp\!\!\!\perp D | R$ Yes

  $T \perp\!\!\!\perp D | R, S$

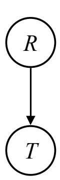


Causality?

- When Bayes' 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
  - E.g. consider the variables *Traffic* and *Drips*
  - End up with arrows that reflect correlation, not causation
- What do the arrows really mean?
  - Topology may happen to encode causal structure
  - Topology only guaranteed to encode conditional independencies

Example: Traffic

- Basic traffic net
- Let's multiply out the joint



$P(R)$

r	1/4
¬r	3/4

$P(T|R)$

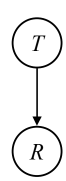
r	t	3/4
	¬t	1/4
¬r	t	1/2
	¬t	1/2

$P(T, R)$

r	t	3/16
r	¬t	1/16
¬r	t	6/16
¬r	¬t	6/16

Example: Reverse Traffic

- Reverse causality?



$P(T)$

t	9/16
¬t	7/16

$P(R|T)$

t	r	1/3
	¬r	2/3
¬t	r	1/7
	¬r	6/7

$P(T, R)$

r	t	3/16
r	¬t	1/16
¬r	t	6/16
¬r	¬t	6/16

Example: Coins

- Extra arcs don't prevent representing independence, just allow non-independence

$P(X_1)$

h	0.5
t	0.5

$P(X_2)$

h	0.5
t	0.5

$P(X_1)$

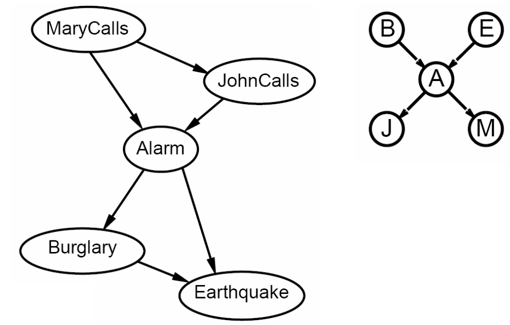
h	0.5
t	0.5

$P(X_2|X_1)$

h | h	0.5
t | h	0.5

h | t	0.5
t | t	0.5

Alternate BNs



Summary

- Bayes nets compactly encode joint distributions
- Guaranteed independencies of distributions can be deduced from BN graph structure
- A Bayes' net may have other independencies that are not detectable until you inspect its specific distribution
- The Bayes' ball algorithm (aka d-separation) tells us when an observation of one variable can change belief about another variable

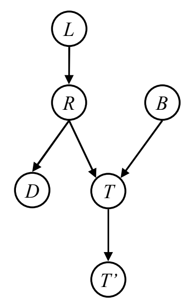
Inference

- Inference: calculating some statistic from a joint probability distribution
- Examples:
  - Posterior probability:

$$P(Q|E_1 = e_1, \ldots E_k = e_k)$$

  - Most likely explanation:

$$\text{argmax}_q \; P(Q = q|E_1 = e_1 \ldots)$$

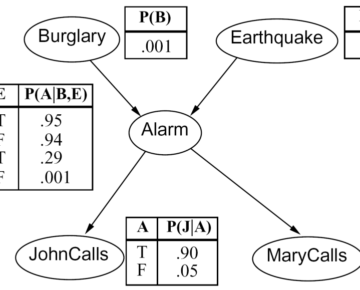


Inference by Enumeration

- P(sun)?
- P(sun | winter)?
- P(sun | 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

Reminder: 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

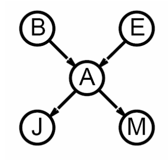
A	P(J|A)
T	.90
F	.05

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

Inference by Enumeration

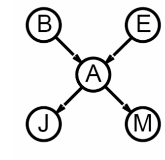
- Given unlimited time, inference in BNs is easy
- Recipe:
  - State the marginal probabilities you need
  - Figure out ALL the atomic probabilities you need
  - Calculate and combine them
- Example:

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



Example

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



$$\begin{aligned} P(b,j,m) = \; & P(b,e,a,j,m) + \\ & P(b,\bar{e},a,j,m) + \\ & P(b,e,\bar{a},j,m) + \\ & P(b,\bar{e},\bar{a},j,m) \\ = \; & \sum_{e,a} P(b,e,a,j,m) \end{aligned}$$

Where did we use the BN structure?

We didn't!

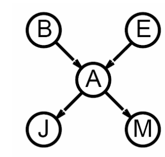
Example

- In this simple method, we only need the BN to synthesize the joint entries

$$\begin{aligned} P(b,j,m) = \\ & P(b)P(e)P(a|b,e)P(j|a)P(m|a) + \\ & P(b)P(e)P(\bar{a}|b,e)P(j|\bar{a})P(m|\bar{a}) + \\ & P(b)P(\bar{e})P(a|b,\bar{e})P(j|a)P(m|a) + \\ & P(b)P(\bar{e})P(\bar{a}|b,\bar{e})P(j|\bar{a})P(m|\bar{a}) \end{aligned}$$

Normalization Trick

$$P(B|j,m) = \frac{P(B,j,m)}{P(j,m)}$$



$$P(b,j,m) = \sum_{e,a} P(b,e,a,j,m)$$

$$P(\bar{b},j,m) = \sum_{e,a} P(\bar{b},e,a,j,m)$$

$$\begin{pmatrix} P(b,j,m) \\ P(\bar{b},j,m) \end{pmatrix} \xrightarrow{\text{Normalize}} \begin{pmatrix} P(b|j,m) \\ P(\bar{b}|j,m) \end{pmatrix}$$

Inference by Enumeration

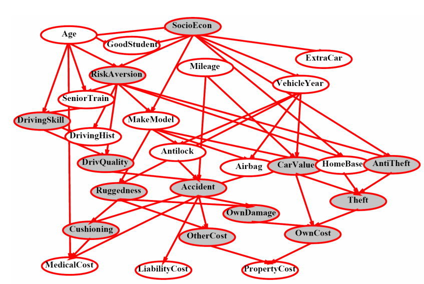
- General case:
  - Evidence variables: $(E_1 \dots E_k) = (e_1 \dots e_k)$
  - Query variables: $Y_1 \dots Y_m$
  - Hidden variables: $H_1 \dots H_r$

  $\left.\right\} X_1, X_2, \dots X_n$ *All variables*
- We want: $P(Y_1 \dots Y_m | e_1 \dots e_k)$
- First, select the entries consistent with the evidence
- Second, sum out H:

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

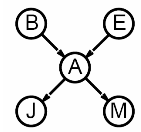
- Finally, normalize the remaining entries to conditionalize
- Obvious problems:
  - Worst-case time complexity O(d^n)
  - Space complexity O(d^n) to store the joint distribution

Inference by Enumeration?



Nesting Sums

- Atomic inference is extremely slow!
- Slightly clever way to save work:
  - Move the sums as far right as possible
  - Example:



$$P(b,j,m) = \sum_{e,a} P(b,e,a,j,m)$$

$$= \sum_{e,a} P(b)P(e)P(a|b,e)P(j|a)P(m|a)$$

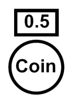
$$= P(b)\sum_{e} P(e)\sum_{a} P(a|b,e)P(j|a)P(m|a)$$

Variable Elimination: Idea

- Lots of redundant work in the computation tree
- We can save time if we cache all partial results
- This is the basic idea behind variable elimination

Sampling

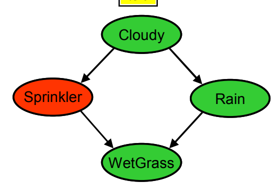
- Basic idea:
  - Draw N samples from a sampling distribution S
  - Compute an approximate posterior probability
  - Show this converges to the true probability P



- Outline:
  - Sampling from an empty network
  - Rejection sampling: reject samples disagreeing with evidence
  - Likelihood weighting: use evidence to weight samples

Prior Sampling

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

Prior Sampling

- This process generates samples with probability

$$S_{PS}(x_1 \ldots x_n) = \prod_{i=1}^{n} P(x_i|\text{Parents}(X_i)) = P(x_1 \ldots x_n)$$

…i.e. the BN's joint probability

- Let the number of samples of an event be $N_{PS}(x_1 \ldots x_n)$
- Then $\lim_{N\to\infty} \hat{P}(x_1,\ldots,x_n) = \lim_{N\to\infty} N_{PS}(x_1,\ldots,x_n)/N$

$$= S_{PS}(x_1,\ldots,x_n)$$

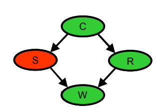
$$= P(x_1 \ldots x_n)$$

- I.e., the sampling procedure is consistent

Example

- We'll get a bunch of samples from the BN:

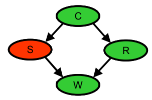
c, ¬s, r, w
c, s, r, w
¬c, s, r, ¬w
c, ¬s, r, w
¬c, s, ¬r, w



- If we want to know P(W)
  - We have counts <w:4, ¬w:1>
  - Normalize to get P(W) = <w:0.8, ¬w:0.2>
  - This will get closer to the true distribution with more samples
  - Can estimate anything else, too
  - What about P(C| ¬r)?   P(C| ¬r, ¬w)?

Rejection Sampling

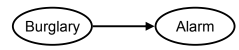
- Let's say we want P(C)
  - No point keeping all samples around
  - Just tally counts of C outcomes
- Let's say we want P(C| s)
  - Same thing: tally C outcomes, but ignore (reject) samples which don't have S=s
  - This is rejection sampling
  - It is also consistent (correct in the limit)



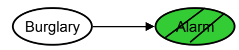
c, ¬s, r, w
c, s, r, w
¬c, s, r, ¬w
c, ¬s, r, w
¬c, s, ¬r, w

Likelihood Weighting

- Problem with rejection sampling:
  - If evidence is unlikely, you reject a lot of samples
  - You don't exploit your evidence as you sample
  - Consider P(B|a)



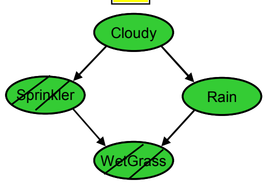
- Idea: fix evidence variables and sample the rest



- Problem: sample distribution not consistent!
- Solution: weight by probability of evidence given parents

Likelihood Sampling

P(C)
.50



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

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

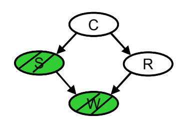
$w = 1.0 \times 0.1 \times 0.99$

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

Likelihood Weighting

- Sampling distribution if z sampled and e fixed evidence

$$S_{WS}(\mathbf{z}, \mathbf{e}) = \prod_{i=1}^{l} P(z_i|\text{Parents}(Z_i))$$



- Now, samples have weights

$$w(\mathbf{z}, \mathbf{e}) = \prod_{i=1}^{m} P(e_i|\text{Parents}(E_i))$$

- Together, weighted sampling distribution is consistent

$$S_{WS}(\mathbf{z}, \mathbf{e}) w(\mathbf{z}, \mathbf{e}) = \prod_{i=1}^{m} P(e_i|\text{Parents}(E_i)) \prod_{i=1}^{m} P(e_i|\text{Parents}(E_i))$$

$$= P(\mathbf{z}, \mathbf{e})$$

Likelihood Weighting

- Note that likelihood weighting doesn't solve all our problems
- Rare evidence is taken into account for downstream variables, but not upstream ones
- A better solution is Markov-chain Monte Carlo (MCMC), more advanced
- We'll return to sampling for robot localization and tracking in dynamic BNs