CSE 473: Artificial Intelligence
Spring 2012

Bayesian Networks - Inference

Dan Weld

Slides adapted from Jack Breese, Dan Klein, Daphne Koller, Stuart Russell, Andrew Moore & Luke Zettlemoyer

Probabilistic Models - Outline

- Bayesian Networks (BNs)
- Independence
- Efficient Inference in BNs
  - Variable Elimination
  - Direct Sampling
  - Markov Chain Monte Carlo (MCMC)
- Learning

Bayes' Nets: Big Picture

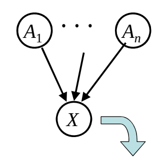
- Problems with using full joint distribution :
  - Unless very few variables, the joint is WAY too big
  - Unless very few variables, hard to learn (estimate empirically)
- Bayesian networks: a technique for describing complex joint distributions (models) using simple, local distributions (conditional probabilities)
  - A kind of "graphical model"
  - We describe how random variables interact, locally
  - Local interactions chain together to give global distribution

Bayes' Net Semantics

Formally:

- A set of nodes, one per variable X
- A directed, acyclic graph
- A CPT for each node
  - CPT = "Conditional Probability Table"
  - Collection of distributions over X, one for each combination of parents' values

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



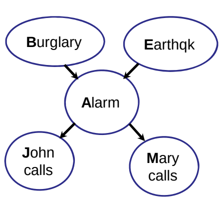
$$P(X|A_1 \ldots A_n)$$

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

Example: Alarm Network

10 params *vs* 31

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

Probabilities in BNs

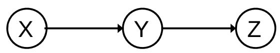
- Bayes' nets implicitly encode joint distributions
  - As a product of local conditional distributions
  - To see what probability a BN gives to a full assignment, multiply all the relevant conditionals together:

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

- This lets us reconstruct any entry of the full joint
- Not every BN can represent every joint distribution
  - The topology enforces certain *independence* assumptions
  - Compare to the exact decomposition according to the chain rule!

Independence in a BN

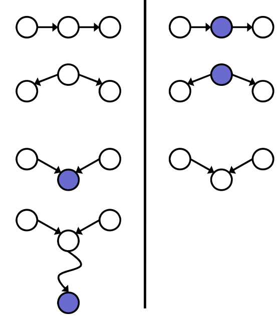
- Important question about a BN:
  - Are two nodes independent given certain evidence?
    - If yes, can prove using algebra (tedious in general)
    - If no, can prove with a counter example
  - Example:



  - Question: are X and Z independent?
    - Answer: no.
      - Example: low pressure causes rain, which causes traffic.
    - Knowledge about X may change belief in Z,
    - Knowledge about Z may change belief in X (via Y)
    - Addendum: they *could* be independent: how?

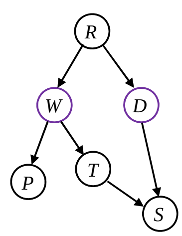
Reachability (D-Separation)

- Question: Are X and Y conditionally independent given evidence vars {Z}?
  - Yes, if X and Y "separated" by Z
  - Look for active paths from X to Y
  - No active paths = independence!
- A path is active if each triple is active:
  - Causal chain A → B → C where B is **unobserved** (either direction)
  - Common cause A ← B → C where B is **unobserved**
  - Common effect (aka v-structure) A → B ← C where B *or one of its descendents* is **observed**
- All it takes to block a path is a single inactive segment

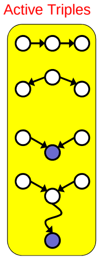


Example

- Variables:
  - R: Raining
  - W: Wet
  - P: Plants growing
  - T: Traffic bad
  - D: Roof drips
  - S: I'm sad
- Questions:
  - $W \perp D$

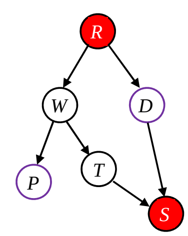


Active Triples

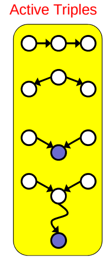


Example

- Variables:
  - R: Raining
  - W: Wet
  - P: Plants growing
  - T: Traffic bad
  - D: Roof drips
  - S: I'm sad
- Questions:
  - $W \perp D$ *No*
  - $P \perp D$ | R, S

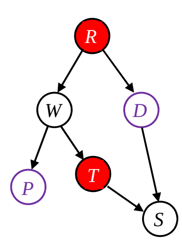


Active Triples

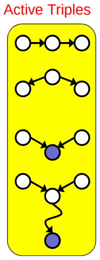


Example

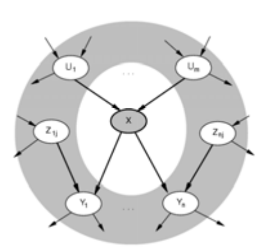
- Variables:
  - R: Raining
  - W: Wet
  - P: Plants growing
  - T: Traffic bad
  - D: Roof drips
  - S: I'm sad
- Questions:
  - $W \perp D$ *No*
  - $P \perp D$ | R, S *No*
  - $P \perp D$ | R, T *Yes*



Active Triples



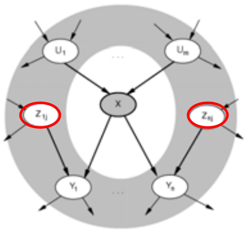
Given Markov Blanket, X is Independent of All Other Nodes



MB(X) = Par(X) ∪ Childs(X) ∪ Par(Childs(X))

12

Given Markov Blanket, X is Independent of All Other Nodes

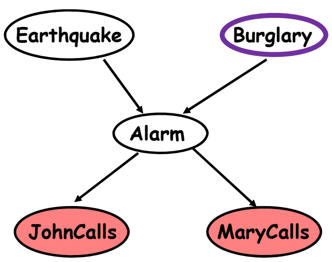


MB(X) = Par(X) ∪ Childs(X) ∪ Par(Childs(X))

Inference in BNs

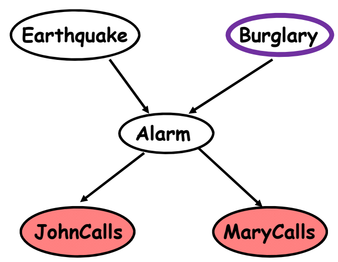
- The graphical independence representation
  - yields efficient inference schemes
- We generally want to compute
  - Marginal probability: *Pr(Z)*,
  - *Pr(Z|**E**)* where ***E*** is (conjunctive) evidence
    - Z: query variable(s),
    - E: evidence variable(s)
    - everything else: hidden variable
- Computations organized by network topology

P(B | J=true, M=true)



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

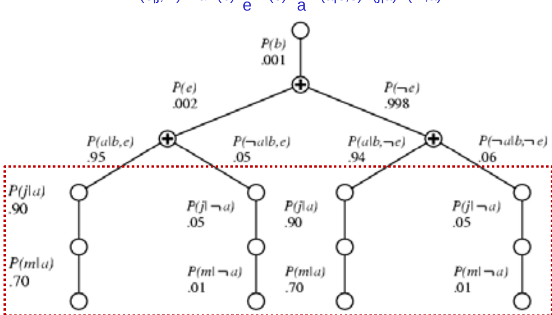
P(B | J=true, M=true)



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

Variable Elimination

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



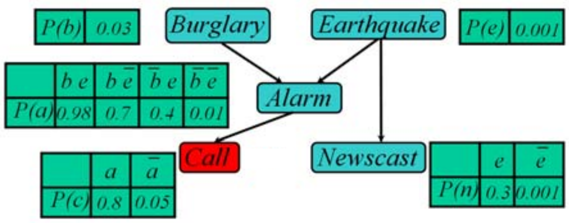
Repeated computations ➔ Dynamic Programming

Approximate Inference in Bayes Nets Sampling based methods

(Based on slides by Jack Breese and Daphne Koller)

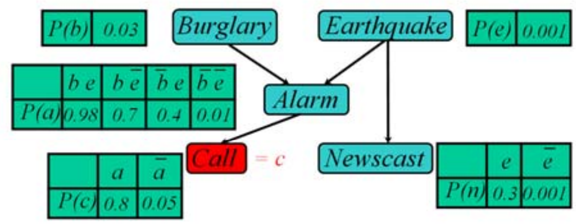
Bayes Net is a generative model

- We can easily generate samples from the distribution represented by the Bayes net
  - Generate one variable at a time in topological order



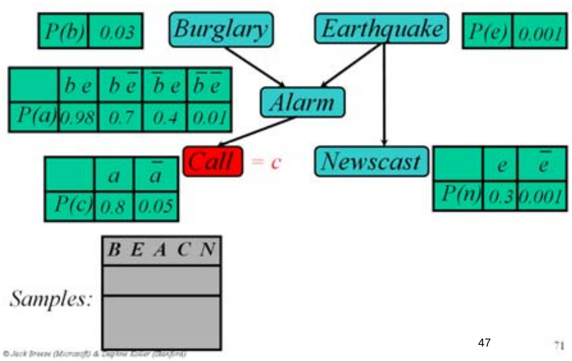
Use the samples to compute marginal probabilities, say P(c)

Stochastic simulation P(B|C)



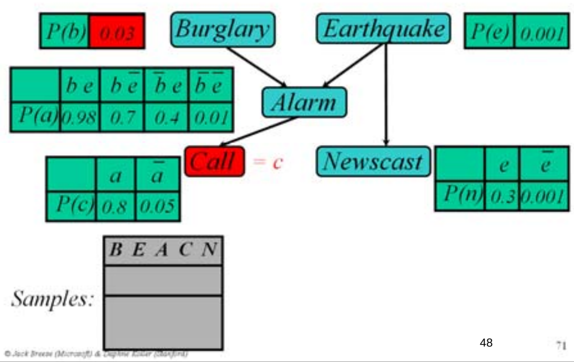
 46 71

Stochastic simulation P(B|C)



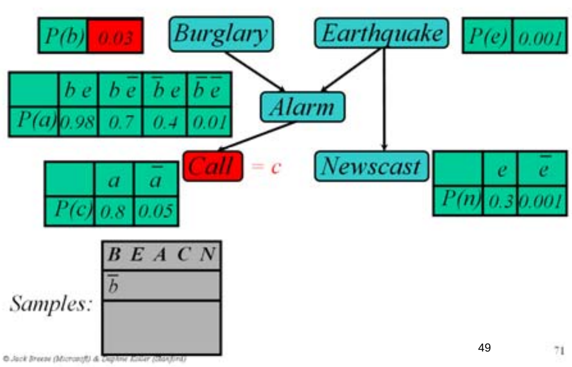
 47 71

Stochastic simulation P(B|C)



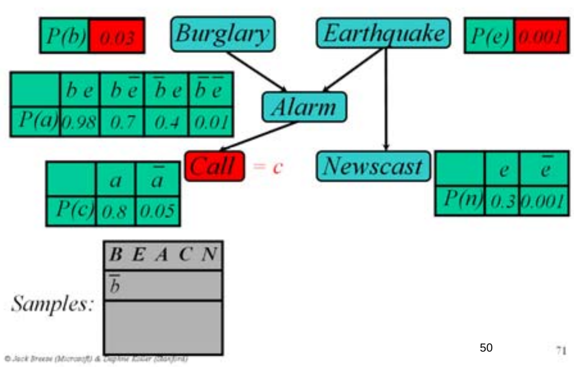
 48 71

Stochastic simulation P(B|C)

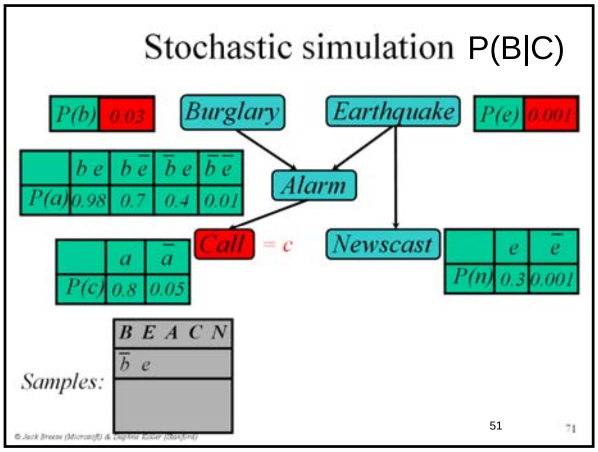


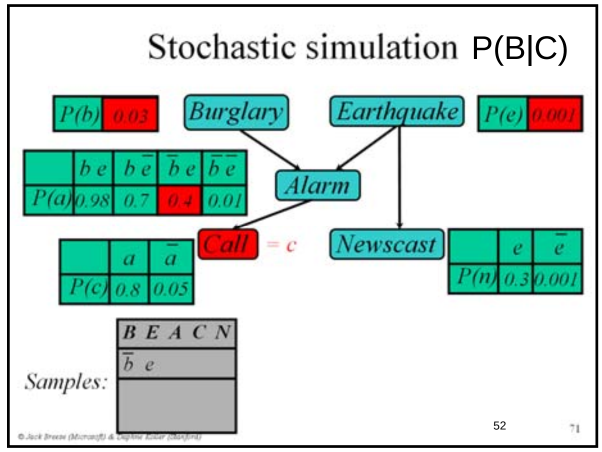
 49 71

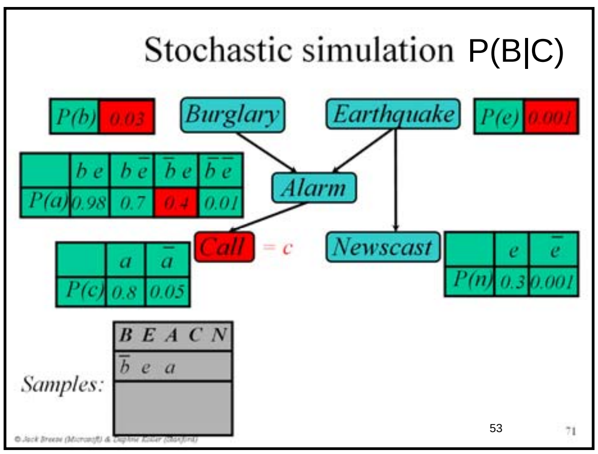
Stochastic simulation P(B|C)

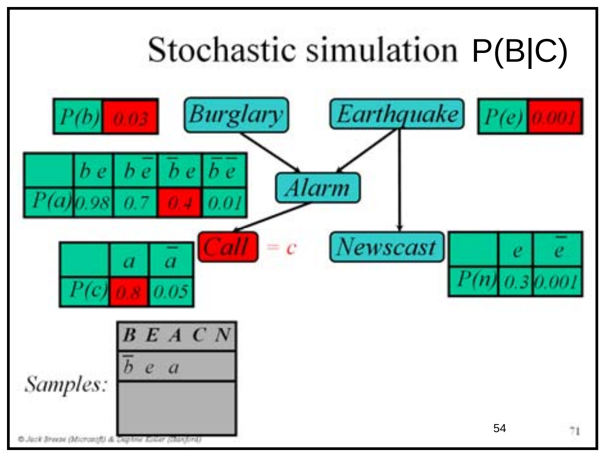


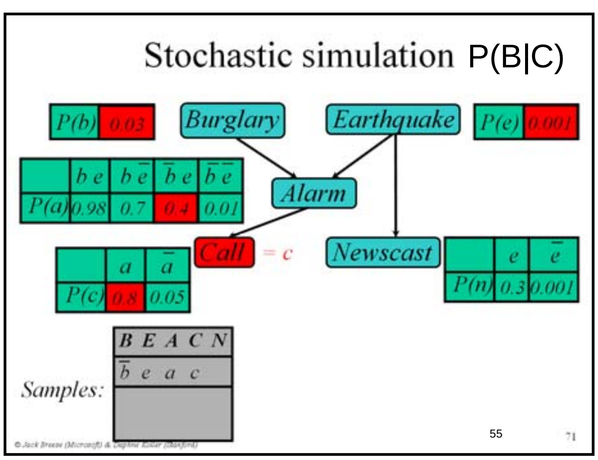
 50 71

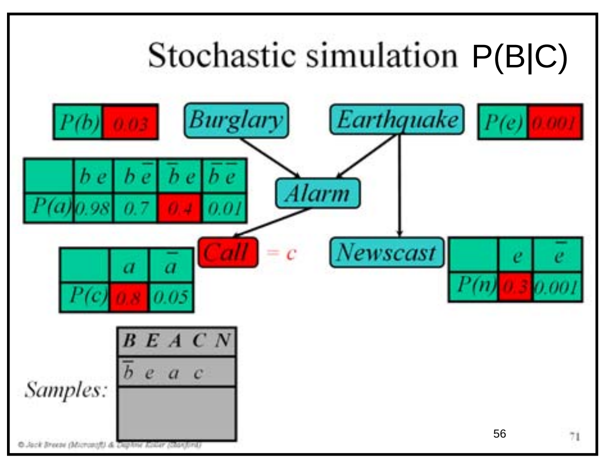

Stochastic simulation P(B|C)
P(b) 0.03
Burglary
Earthquake
P(e) 0.001
Alarm
b e | b $\bar{e}$ | $\bar{b}$ e | $\bar{b}$ $\bar{e}$
P(a) 0.98 | 0.7 | 0.4 | 0.01
Call = c
Newscast
a | $\bar{a}$
P(c) 0.8 | 0.05
e | $\bar{e}$
P(n) 0.3 | 0.001
B E A C N
$\bar{b}$ e
Samples:
51


Stochastic simulation P(B|C)
P(b) 0.03
Burglary
Earthquake
P(e) 0.001
Alarm
b e | b $\bar{e}$ | $\bar{b}$ e | $\bar{b}$ $\bar{e}$
P(a) 0.98 | 0.7 | 0.4 | 0.01
Call = c
Newscast
a | $\bar{a}$
P(c) 0.8 | 0.05
e | $\bar{e}$
P(n) 0.3 | 0.001
B E A C N
$\bar{b}$ e
Samples:
52

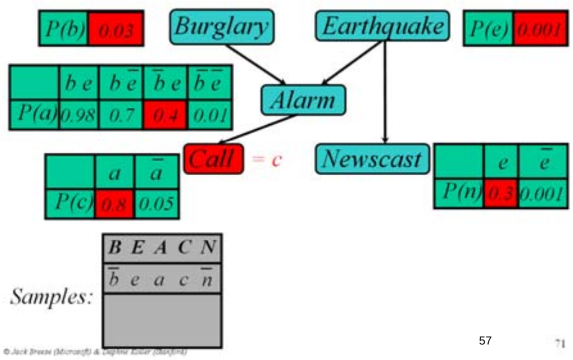

Stochastic simulation P(B|C)
P(b) 0.03
Burglary
Earthquake
P(e) 0.001
Alarm
b e | b $\bar{e}$ | $\bar{b}$ e | $\bar{b}$ $\bar{e}$
P(a) 0.98 | 0.7 | 0.4 | 0.01
Call = c
Newscast
a | $\bar{a}$
P(c) 0.8 | 0.05
e | $\bar{e}$
P(n) 0.3 | 0.001
B E A C N
$\bar{b}$ e a
Samples:
53


Stochastic simulation P(B|C)
P(b) 0.03
Burglary
Earthquake
P(e) 0.001
Alarm
b e | b $\bar{e}$ | $\bar{b}$ e | $\bar{b}$ $\bar{e}$
P(a) 0.98 | 0.7 | 0.4 | 0.01
Call = c
Newscast
a | $\bar{a}$
P(c) 0.8 | 0.05
e | $\bar{e}$
P(n) 0.3 | 0.001
B E A C N
$\bar{b}$ e a
Samples:
54


Stochastic simulation P(B|C)
P(b) 0.03
Burglary
Earthquake
P(e) 0.001
Alarm
b e | b $\bar{e}$ | $\bar{b}$ e | $\bar{b}$ $\bar{e}$
P(a) 0.98 | 0.7 | 0.4 | 0.01
Call = c
Newscast
a | $\bar{a}$
P(c) 0.8 | 0.05
e | $\bar{e}$
P(n) 0.3 | 0.001
B E A C N
$\bar{b}$ e a c
Samples:
55

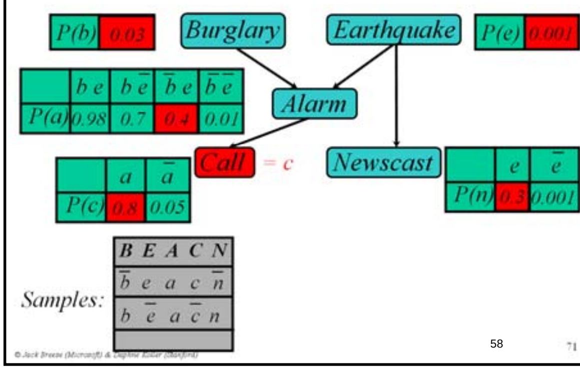

Stochastic simulation P(B|C)
P(b) 0.03
Burglary
Earthquake
P(e) 0.001
Alarm
b e | b $\bar{e}$ | $\bar{b}$ e | $\bar{b}$ $\bar{e}$
P(a) 0.98 | 0.7 | 0.4 | 0.01
Call = c
Newscast
a | $\bar{a}$
P(c) 0.8 | 0.05
e | $\bar{e}$
P(n) 0.3 | 0.001
B E A C N
$\bar{b}$ e a c
Samples:
56

Stochastic simulation P(B|C)



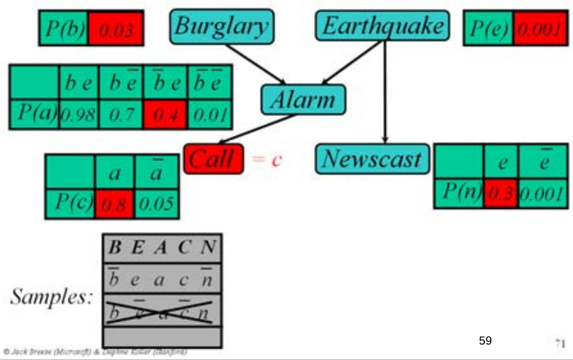
57

Stochastic simulation P(B|C)



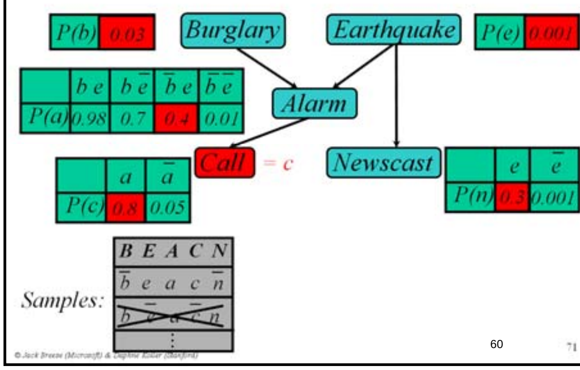
58

Stochastic simulation P(B|C)



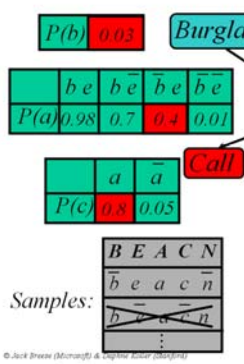
59

Stochastic simulation P(B|C)



60

Stochastic simulation P(B|C)



$$P(b|c) \sim \frac{\#\text{ of live samples with } B=b}{\text{total }\#\text{ of live samples}}$$

61

Rejection Sampling

- Sample from the prior
  - reject if do not match the evidence
- Returns consistent posterior estimates
- Hopelessly expensive if P(e) is small
  - P(e) drops off exponentially with no. of evidence vars

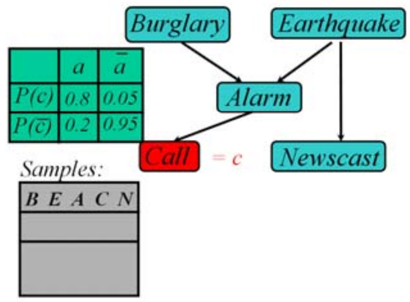
62

Likelihood Weighting

- Idea:
  - fix evidence variables
  - sample only non-evidence variables
  - weight each sample by the likelihood of evidence

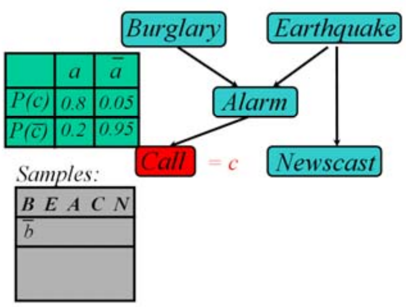
63

Likelihood weighting P(B|C)



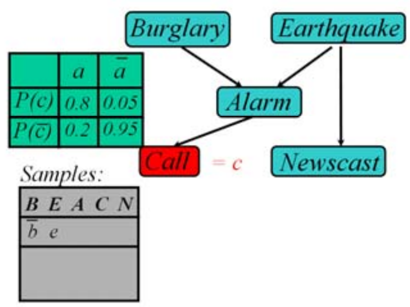
64 72

Likelihood weighting P(B|C)



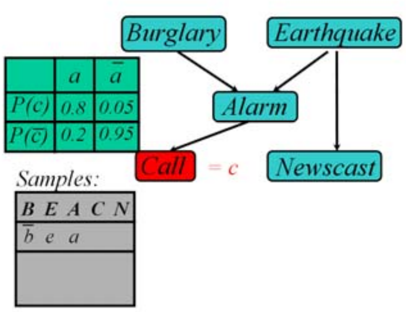
65 72

Likelihood weighting P(B|C)



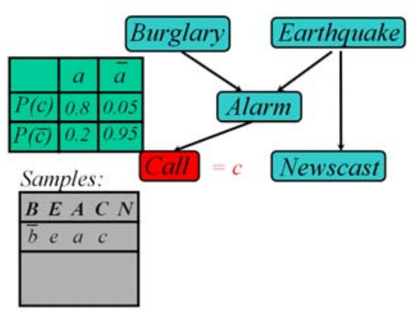
66 72

Likelihood weighting P(B|C)



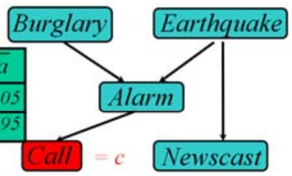
67 72

Likelihood weighting P(B|C)



68 72

Likelihood weighting P(B|C)

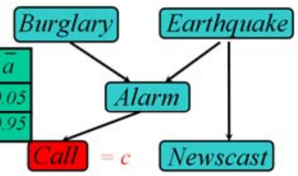


	a	$\bar{a}$
$P(c)$	0.8	0.05
$P(\bar{c})$	0.2	0.95

Samples:

B E A C N
$\bar{b}$ e a c $\bar{n}$

Likelihood weighting P(B|C)

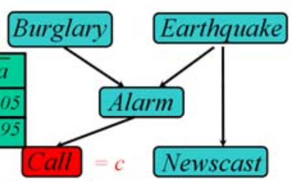


	a	$\bar{a}$
$P(c)$	0.8	0.05
$P(\bar{c})$	0.2	0.95

Samples:

B E A C N	weight
$\bar{b}$ e a c $\bar{n}$	0.8

Likelihood weighting P(B|C)

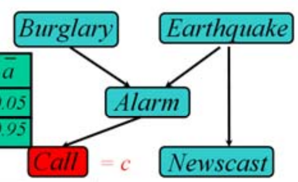


	a	$\bar{a}$
$P(c)$	0.8	0.05
$P(\bar{c})$	0.2	0.95

Samples:

B E A C N	weight
$\bar{b}$ e a c $\bar{n}$	0.8
b $\bar{e}$ $\bar{a}$ c n	0.05

Likelihood weighting P(B|C)

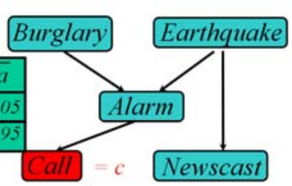


	a	$\bar{a}$
$P(c)$	0.8	0.05
$P(\bar{c})$	0.2	0.95

Samples:

B E A C N	weight
$\bar{b}$ e a c $\bar{n}$	0.8
b $\bar{e}$ $\bar{a}$ c n	0.05
⋮	

Likelihood weighting P(B|C)



	a	$\bar{a}$
$P(c)$	0.8	0.05
$P(\bar{c})$	0.2	0.95

Samples:

B E A C N	weight
$\bar{b}$ e a c $\bar{n}$	0.8
b $\bar{e}$ $\bar{a}$ c n	0.05
⋮	

$$P(b|c) = \frac{\text{weight of samples with } B=b}{\text{total weight of samples}}$$

Likelihood Weighting

- Sampling probability: S(z,e) = $\prod_i P(z_i \mid \text{Parents}(Z_i))$
  - Neither prior nor posterior
- Wt for a sample <z,e>: $w(z, e) = \prod_i P(e_i \mid \text{Parents}(E_i))$
- Weighted Sampling probability S(z,e)w(z,e)

$$= \prod_i P(z_i \mid \text{Parents}(Z_i)) \prod_i P(e_i \mid \text{Parents}(E_i))$$

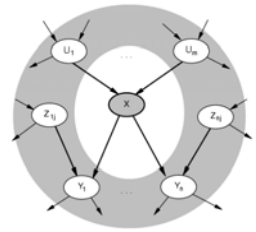
= P(z,e)

- → returns consistent estimates
- performance degrades w/ many evidence vars
  - but a few samples have nearly all the total weight
  - late occuring evidence vars do not guide sample generation

MCMC with Gibbs Sampling

- Fix the values of observed variables
- Set the values of all non-observed variables randomly
- Perform a random walk through the space of complete variable assignments. On each move:
  1. Pick a variable X
  2. Calculate Pr(X=true | all other variables)
  3. Set X to true with that probability
- Repeat many times. Frequency with which any variable X is true is it's posterior probability.
- Converges to true posterior when frequencies stop changing significantly
  - stable distribution, mixing

Given Markov Blanket, X is Independent of All Other Nodes



MB(X) = Par(X) ∪ Childs(X) ∪ Par(Childs(X))

Markov Blanket Sampling

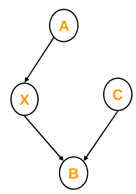
- How to calculate Pr(X=true | all other variables) ?
- Recall: a variable is independent of all others given it's Markov Blanket
  - parents
  - children
  - other parents of children

- So problem becomes calculating Pr(X=true | MB(X))
  - We solve this sub-problem exactly
  - Fortunately, it is easy to solve

$$P(X) = \alpha P(X \mid Parents(X)) \prod_{Y \in Children(X)} P(Y \mid Parents(Y))$$

Example

$$P(X) = \alpha P(X \mid Parents(X)) \prod_{Y \in Children(X)} P(Y \mid Parents(Y))$$



$$P(X \mid A,B,C) = \frac{P(X,A,B,C)}{P(A,B,C)}$$

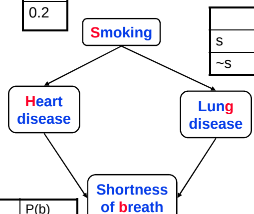
$$= \frac{P(A)P(X \mid A)P(C)P(B \mid X,C)}{P(A,B,C)}$$

$$= \left[\frac{P(A)P(C)}{P(A,B,C)}\right] P(X \mid A)P(B \mid X,C)$$

$$= \alpha P(X \mid A)P(B \mid X,C)$$

Example

P(s)
0.2



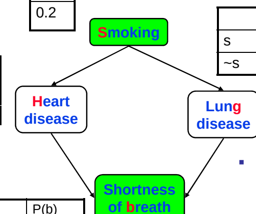
	P(h)
s	0.6
~s	0.1

	P(g)
s	0.8
~s	0.1

H	G	P(b)
h	g	0.9
h	~g	0.8
~h	g	0.7
~h	~g	0.1

Example

P(s)
0.2



	P(h)
s	0.6
~s	0.1

	P(g)
s	0.8
~s	0.1

- Evidence: s, b

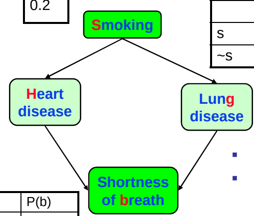
H	G	P(b)
h	g	0.9
h	~g	0.8
~h	g	0.7
~h	~g	0.1

Example

P(s)
0.2

	P(g)
s	0.8
~s	0.1

	P(h)
s	0.6
~s	0.1



H	G	P(b)
h	g	0.9
h	~g	0.8
~h	g	0.7
~h	~g	0.1

- Evidence: s, b
- Randomly set: h, b

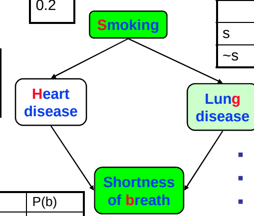
81

Example

P(s)
0.2

	P(g)
s	0.8
~s	0.1

	P(h)
s	0.6
~s	0.1



H	G	P(b)
h	g	0.9
h	~g	0.8
~h	g	0.7
~h	~g	0.1

- Evidence: s, b
- Randomly set: h, g
- Sample H using P(H|s,g,b)

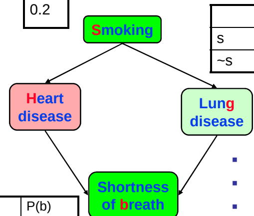
82

Example

P(s)
0.2

	P(g)
s	0.8
~s	0.1

	P(h)
s	0.6
~s	0.1



H	G	P(b)
h	g	0.9
h	~g	0.8
~h	g	0.7
~h	~g	0.1

- Evidence: s, b
- Randomly set: ~h, g
- Sample H using P(H|s,g,b)
- ⇒ Suppose result is ~h

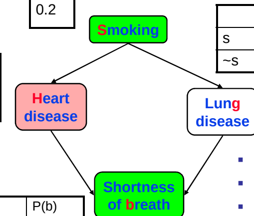
83

Example

P(s)
0.2

	P(g)
s	0.8
~s	0.1

	P(h)
s	0.6
~s	0.1



H	G	P(b)
h	g	0.9
h	~g	0.8
~h	g	0.7
~h	~g	0.1

- Evidence: s, b
- Randomly set: ~h, g
- Sample H using P(H|s,g,b)

Suppose result is ~h

Sample G using P(G|s,~h,b)

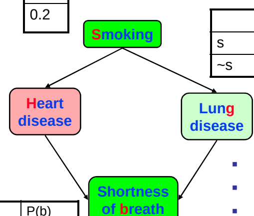
84

Example

P(s)
0.2

	P(g)
s	0.8
~s	0.1

	P(h)
s	0.6
~s	0.1



H	G	P(b)
h	g	0.9
h	~g	0.8
~h	g	0.7
~h	~g	0.1

- Evidence: s, b
- Randomly set: ~h, g
- Sample H using P(H|s,g,b)

Suppose result is ~h

Sample G using P(G|s,~h,b)

⇒Suppose result is g

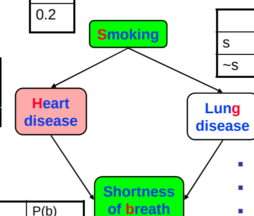
85

Example

P(s)
0.2

	P(g)
s	0.8
~s	0.1

	P(h)
s	0.6
~s	0.1



H	G	P(b)
h	g	0.9
h	~g	0.8
~h	g	0.7
~h	~g	0.1

- Evidence: s, b
- Randomly set: ~h, g
- Sample H using P(H|s,g,b)

Suppose result is ~h

Sample G using P(G|s,~h,b)

⇒Suppose result is g

Sample G using P(G|s,~h,b)

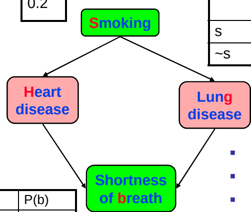
86

Example

P(s)
0.2

	P(g)
s	0.8
~s	0.1

	P(h)
s	0.6
~s	0.1



H	G	P(b)
h	g	0.9
h	~g	0.8
~h	g	0.7
~h	~g	0.1

- Evidence: s, b
- Randomly set: ~h, g
- Sample H using P(H|s,g,b)

Suppose result is ~h

Sample G using P(G|s,~h,b)

⇒Suppose result is g

Sample G using P(G|s,~h,b)

87

Gibbs MCMC Summary

$$P(X|E) = \frac{\text{number of samples with } X{=}x}{\text{total number of samples}}$$

- Advantages:
  - No samples are discarded
  - No problem with samples of low weight
  - Can be implemented very efficiently
    - 10K samples @ second
- Disadvantages:
  - Can get stuck if relationship between vars is *deterministic*
  - Many variations devised to make MCMC more robust

88

Other inference methods

- Exact inference
  - Junction tree
- Approximate inference
  - Belief Propagation
  - Variational Methods

89