CSE 573: Artificial Intelligence
Autumn 2012

Bayesian Networks

Dan Weld

Many slides adapted from Dan Klein, Stuart Russell, Andrew Moore & Luke Zettlemoyer

Outline

- Probabilistic models (and inference)
  - Bayesian Networks (BNs)
  - Independence in BNs
  - Efficient Inference in BNs
  - Learning
- Whirlwind, so…
  - Take CSE 515 (Statistical Methods)
  - Ben Taskar, Spring 2013

Bayes' Nets: Big Picture

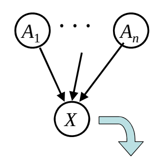
- Two problems with using full joint distribution tables as our probabilistic models:
  - Unless there are only a few variables, the joint is WAY too big to represent explicitly
  - Hard to learn (estimate) anything empirically about more than a few variables at a time
- Bayes' nets: a technique for describing complex joint distributions (models) using simple, local distributions (conditional probabilities)
  - Aka graphical model
  - We describe how variables locally interact
  - Local interactions chain together to give global, indirect interactions

Bayes' Net Semantics

Formally:

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

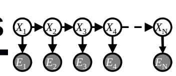
$$P(X|a_1 \ldots a_n)$$

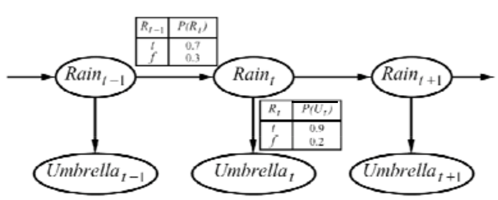


$$P(X|A_1 \ldots A_n)$$

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

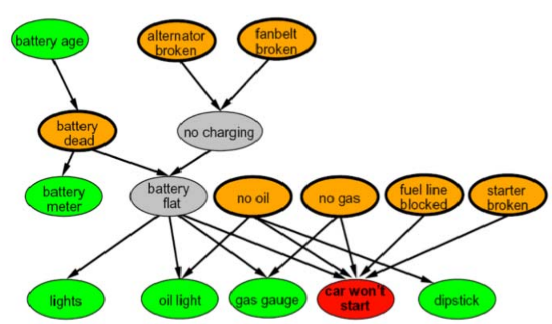
Hidden Markov Models





- An HMM is defined by:
  - Initial distribution: $P(X_1)$
  - Transitions: $P(X_t|X_{t-1})$
  - Emissions: $P(E|X)$

Example Bayes' Net: Car



Example: Car Insurance

 7

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:

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

- Does this always work? Why?
- Not every BN can represent every joint distribution
  - The topology enforces certain *independence* assumptions
  - Compare to the exact decomposition according to the chain rule!

Example: Independent Coin Flips

- N independent coin flips

X_1 X_2 $\cdots$ X_n

$P(X_1)$

h	0.5
t	0.5

$P(X_2)$

h	0.5
t	0.5

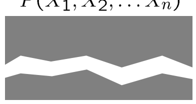
$\cdots$

$P(X_n)$

h	0.5
t	0.5

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

$2^n - 1$



- No interactions between variables $X \perp\!\!\!\perp Y$

Conditional Independence

- Unconditional (absolute) independence very rare (why?)
- *Conditional independence* is our most basic and robust form of knowledge about uncertain environments:

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

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

- What about fire, smoke, alarm?

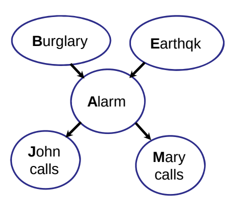
Example: Alarm Network

- Variables
  - B: Burglary
  - A: Alarm goes off
  - M: Mary calls
  - J: John calls
  - E: Earthquake!
- How big is joint distribution?
  - 2^n-1 = 31 parameters

Example: Alarm Network

Only 10 params

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

Example: Traffic II

- Let's build a graphical model
- Variables
  - T: Traffic
  - R: It rains
  - L: Low pressure
  - D: Roof drips
  - B: Ballgame
  - C: Cavity

Changing Bayes' Net Structure

- The same joint distribution can be encoded in many different Bayes' nets
- Analysis question: given some edges, what other edges do you need to add?
  - One answer: fully connect the graph
  - Better answer: don't make any false conditional independence assumptions

Example: Independence

- 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!

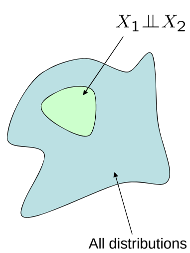
X_1 X_2

$P(X_1)$

h	0.5
t	0.5

$P(X_2)$

h	0.5
t	0.5



Example: Coins

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

X_1 X_2

$P(X_1)$

h	0.5
t	0.5

$P(X_2)$

h	0.5
t	0.5

$X_1 \rightarrow X_2$

$P(X_1)$

h	0.5
t	0.5

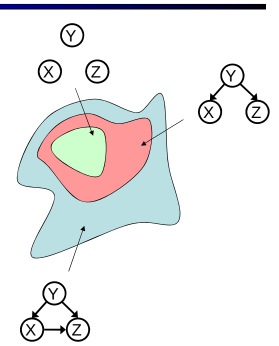
$P(X_2|X_1)$

h | h	0.5
t | h	0.5
h | t	0.5
t | t	0.5

- Adding unneeded arcs isn't wrong, it's just inefficient

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
- Full conditioning can encode any distribution



Independence in a BN

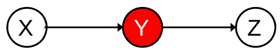
- Important question about a BN:
  - Are two nodes independent given certain evidence?
  - If yes, can prove using algebra (tedious in general)
  - If no, can prove with a counter example
  - Example:

$X \rightarrow Y \rightarrow Z$

- 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?

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?

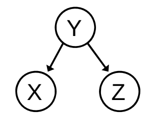
$$P(z|x,y) = \frac{P(x,y,z)}{P(x,y)} = \frac{P(x)P(y|x)P(z|y)}{P(x)P(y|x)}$$

$$= P(z|y)$$ *Yes!*

Evidence along the chain "blocks" the influence

Common Parent

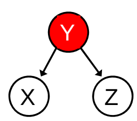
- Another basic configuration: two effects of the same parent
  - Are X and Z independent?



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

Common Parent

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



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

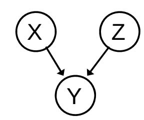
$$= P(z|y)$$ *Yes!*

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

- Observing the cause blocks influence between effects.

Common Effect

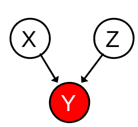
- Last configuration: two causes of one effect (v-structures)
  - Are X and Z independent?
    - Yes: the ballgame and the rain cause traffic, but they are not correlated
    - Still need to prove they must be (try it!)



X: Raining
Z: Ballgame
Y: Traffic

Common Effect

- Last configuration: two causes of one effect (v-structures)
  - Are X and Z independent?
    - Yes: the ballgame and the rain cause traffic, but they are not correlated
    - Still need to prove they must be (try it!)
  - Are X and Z independent given Y?
    - ***No:*** seeing traffic puts the rain and the ballgame in competition as explanation!
  - This is backwards from the other cases
    - Observing an effect ***activates*** influence between possible causes.



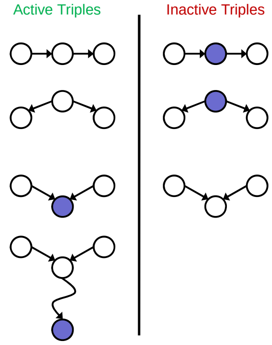
X: Raining
Z: Ballgame
Y: Traffic

The General Case

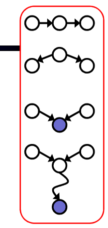
- Any complex example can be analyzed using these three canonical cases
- General question: in a given BN, are two variables independent (given evidence)?
- Solution: analyze the graph

Reachability (D-Separation)

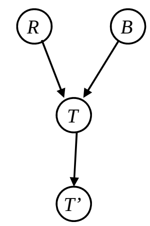
- Question: Are X and Y conditionally independent given evidence vars {Z}?
  - Yes, if X and Y "separated" by Z
  - Look for active paths from X to Y
  - No active paths = independence!
- A path is active if each triple is active:
  - Causal chain A → B → C where B is **unobserved** (either direction)
  - 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



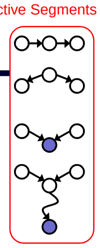
Example: Independent?



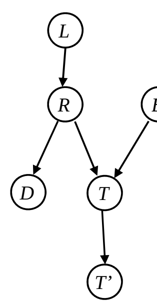
$R \perp\!\!\!\perp B$	*Yes*
$R \perp\!\!\!\perp B \mid T$	*No*
$R \perp\!\!\!\perp B \mid T'$	*No*



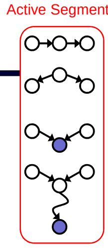
Example: Independent?



$L \perp\!\!\!\perp T' \mid T$	*Yes*
$L \perp\!\!\!\perp B$	*Yes*
$L \perp\!\!\!\perp B \mid T$	*No*
$L \perp\!\!\!\perp B \mid T'$	*No*
$L \perp\!\!\!\perp B \mid T, R$	*Yes*

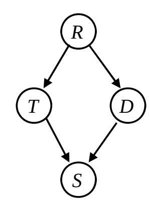


Example

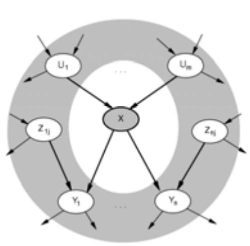


- Variables:
  - R: Raining
  - T: Traffic
  - D: Roof drips
  - S: I'm sad
- Questions:

$T \perp\!\!\!\perp D$	*No*
$T \perp\!\!\!\perp D \mid R$	*Yes*
$T \perp\!\!\!\perp D \mid R, S$	*No*

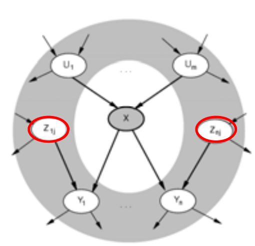


Given ***Markov Blanket***, X is Independent of All Other Nodes



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

Given ***Markov Blanket***, X is Independent of All Other Nodes



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

Summary

- Bayes nets compactly encode joint distributions (JDs)
  - Other graphical models too: factor graphs, CRFs, …
- Guaranteed independencies of distributions can be deduced from BN graph structure
- D-separation gives precise conditional independence guarantees from graph alone
- A Bayes' net's JD may have further (conditional) independence known only from specific CPTs

Outline

- Probabilistic models (and inference)
  - Bayesian Networks (BNs)
  - Independence in BNs
  - ***Efficient Inference in BNs***
  - Learning

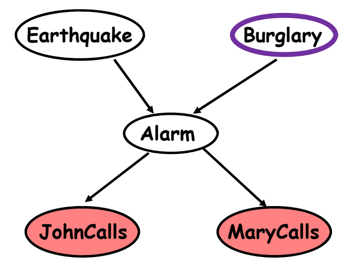
Inference in BNs

This 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

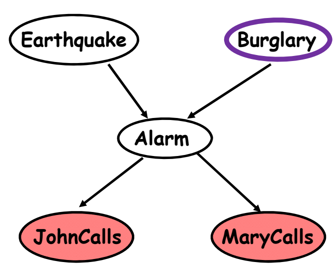
 54

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



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

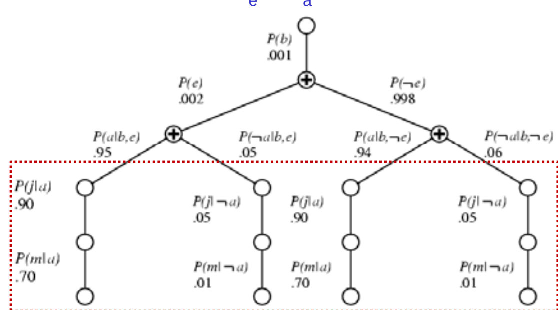
P(B | J=true, M=true)



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

Variable Elimination

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

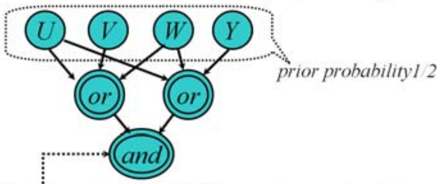


Repeated computations ➔ Dynamic Programming

Reducing 3-SAT to Bayes Nets

- **Theorem:** Inference in a multi-connected Bayesian network is NP-hard.

Boolean 3CNF formula $\phi = (u \vee \bar{v} \vee w) \wedge (\bar{u} \vee \bar{w} \vee y)$



Probability ( ) = $1/2^n$ · # satisfying assignments of ϕ

 70

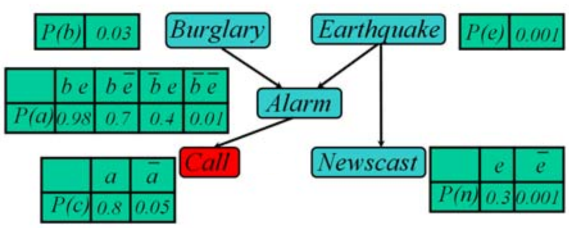
Approximate Inference in Bayes Nets
Sampling based methods

(Based on slides by Jack Breese and Daphne Koller)

67

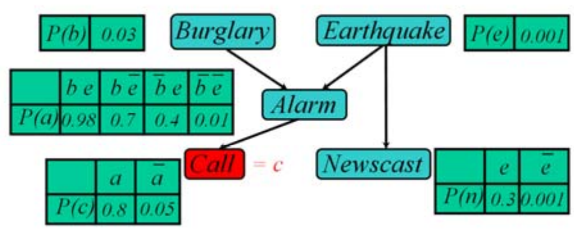
Bayes Net is a generative model

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



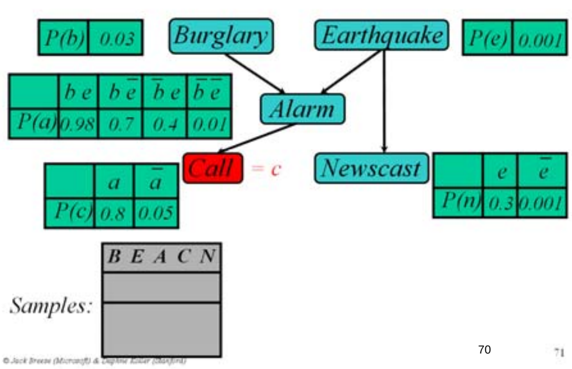
Use the samples to compute probabilities, say P(c) or P(n|c)

Stochastic simulation P(B|C)



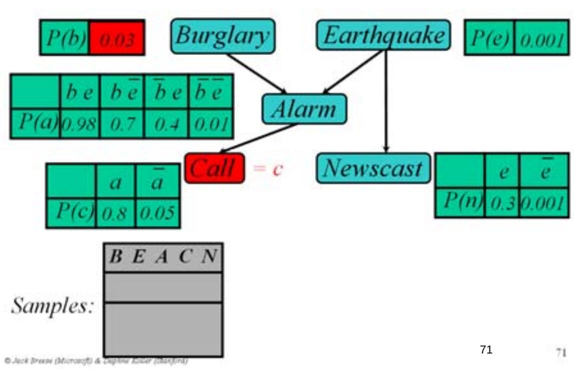
 69 71

Stochastic simulation P(B|C)



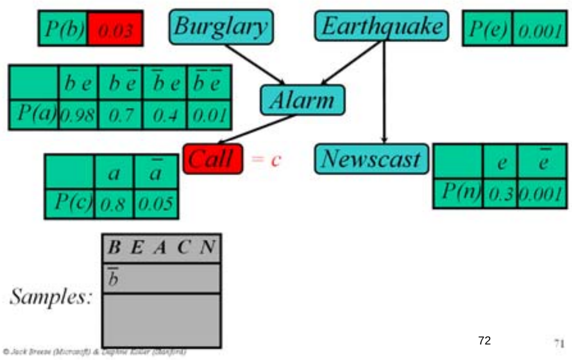
 70 71

Stochastic simulation P(B|C)



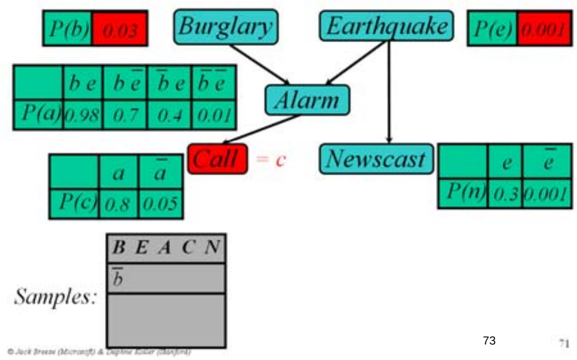
 71 71

Stochastic simulation P(B|C)



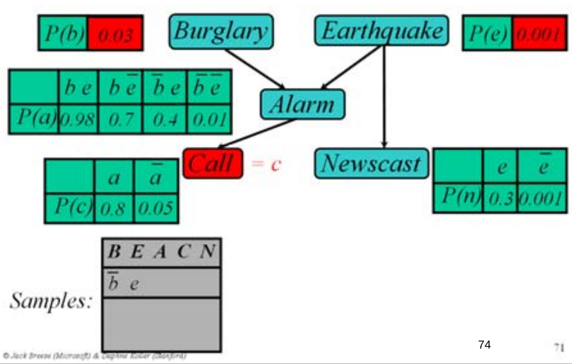
72

Stochastic simulation P(B|C)



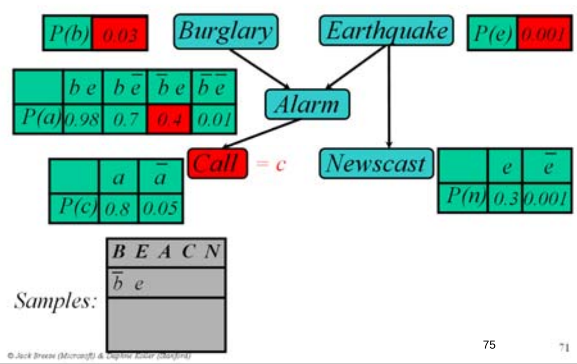
73

Stochastic simulation P(B|C)



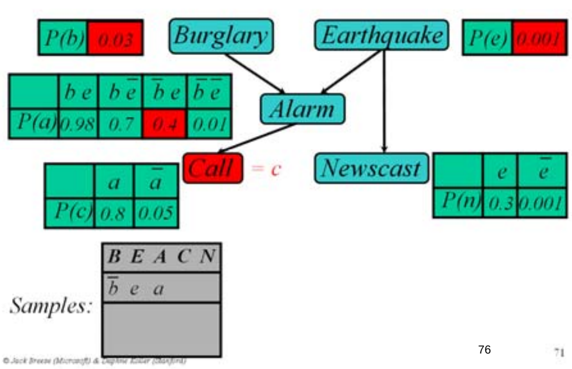
74

Stochastic simulation P(B|C)



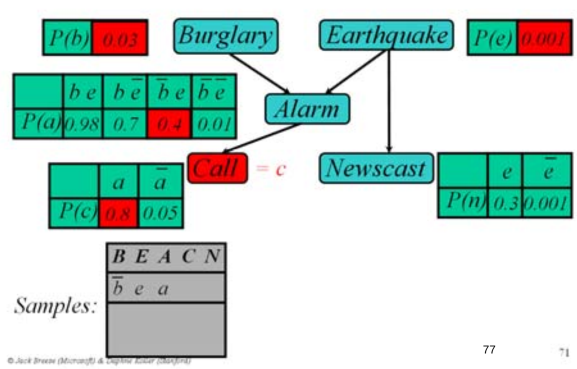
75

Stochastic simulation P(B|C)

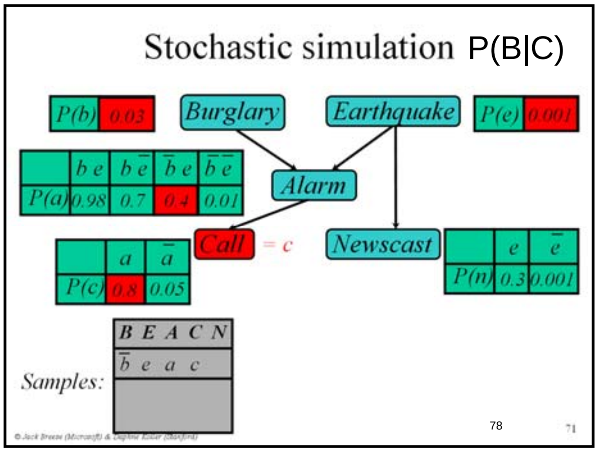


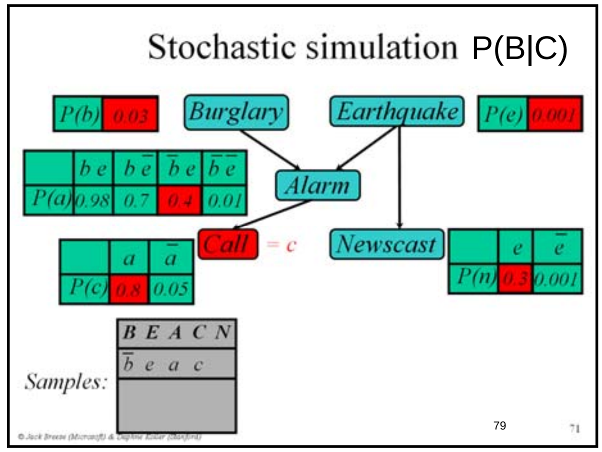
76

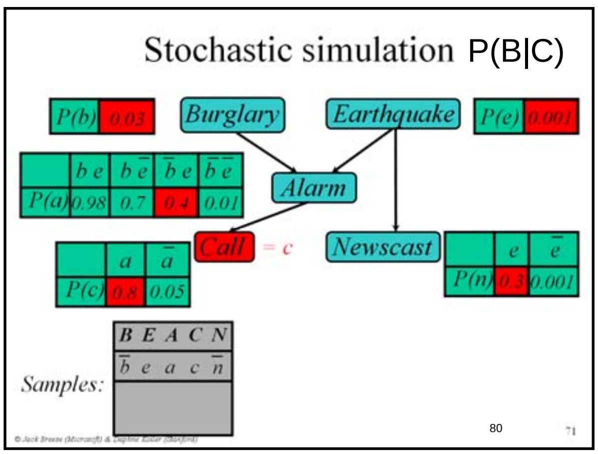
Stochastic simulation P(B|C)

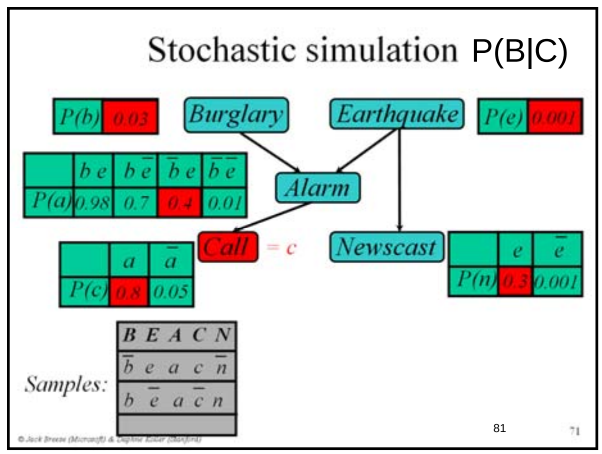


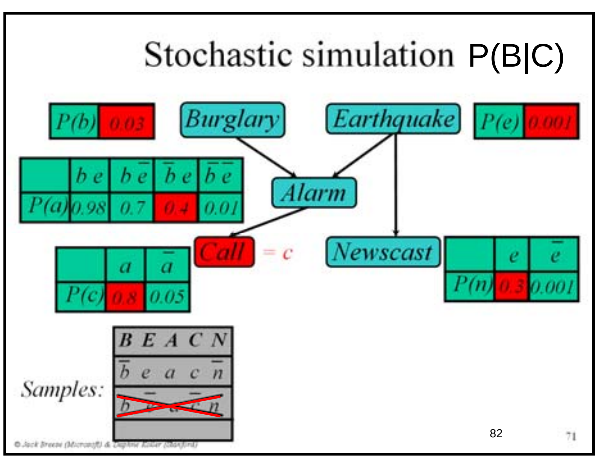
77

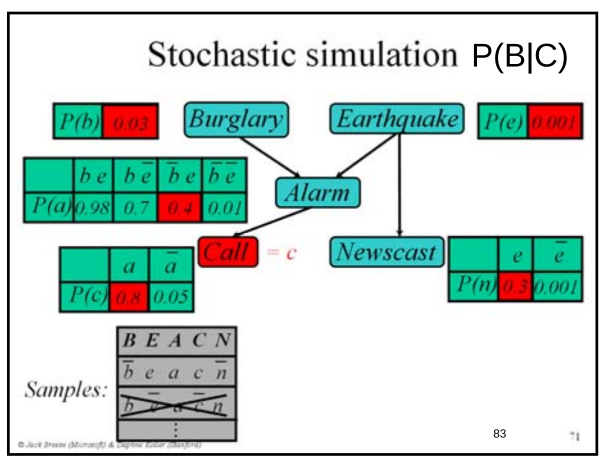

Stochastic simulation P(B|C)
P(b) 0.03
Burglary
Earthquake
P(e) 0.001
Alarm
Call = c
Newscast
P(a) 0.98 0.7 0.4 0.01
P(c) 0.8 0.05
P(n) 0.3 0.001
B E A C N
$\bar{b}$ e a c
Samples:
78


Stochastic simulation P(B|C)
P(b) 0.03
Burglary
Earthquake
P(e) 0.001
Alarm
Call = c
Newscast
P(a) 0.98 0.7 0.4 0.01
P(c) 0.8 0.05
P(n) 0.3 0.001
B E A C N
$\bar{b}$ e a c
Samples:
79

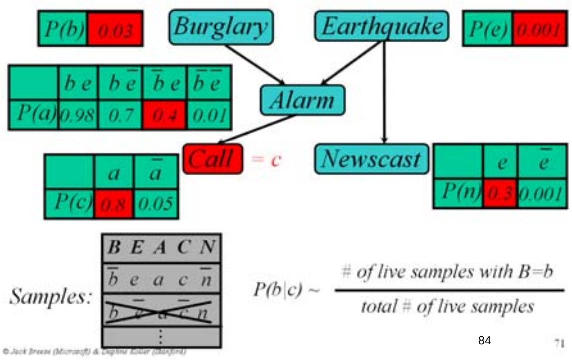

Stochastic simulation P(B|C)
P(b) 0.03
Burglary
Earthquake
P(e) 0.001
Alarm
Call = c
Newscast
P(a) 0.98 0.7 0.4 0.01
P(c) 0.8 0.05
P(n) 0.3 0.001
B E A C N
$\bar{b}$ e a c $\bar{n}$
Samples:
80


Stochastic simulation P(B|C)
P(b) 0.03
Burglary
Earthquake
P(e) 0.001
Alarm
Call = c
Newscast
P(a) 0.98 0.7 0.4 0.01
P(c) 0.8 0.05
P(n) 0.3 0.001
B E A C N
$\bar{b}$ e a c $\bar{n}$
b $\bar{e}$ a $\bar{c}$ n
Samples:
81


Stochastic simulation P(B|C)
P(b) 0.03
Burglary
Earthquake
P(e) 0.001
Alarm
Call = c
Newscast
P(a) 0.98 0.7 0.4 0.01
P(c) 0.8 0.05
P(n) 0.3 0.001
B E A C N
$\bar{b}$ e a c $\bar{n}$
$\cancel{b\ \bar{e}\ a\ \bar{c}\ n}$
Samples:
82


Stochastic simulation P(B|C)
P(b) 0.03
Burglary
Earthquake
P(e) 0.001
Alarm
Call = c
Newscast
P(a) 0.98 0.7 0.4 0.01
P(c) 0.8 0.05
P(n) 0.3 0.001
B E A C N
$\bar{b}$ e a c $\bar{n}$
$\cancel{b\ \bar{e}\ a\ \bar{c}\ n}$
Samples:
83

Stochastic simulation P(B|C)



84

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 exponentially with num of evidence vars

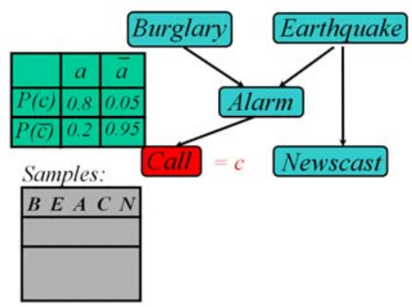
85

Likelihood Weighting

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

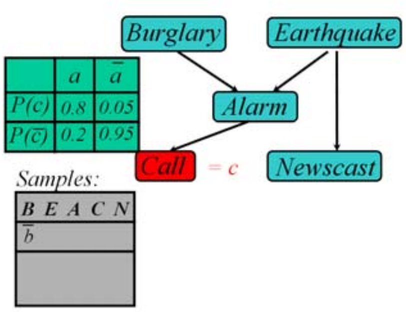
86

Likelihood weighting P(B|C)



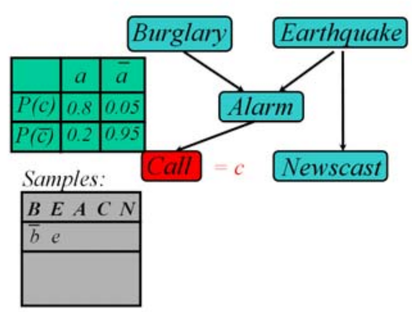
87

Likelihood weighting P(B|C)



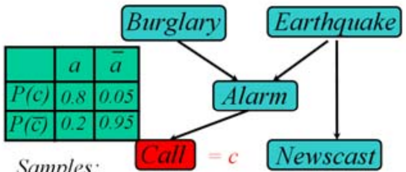
88

Likelihood weighting P(B|C)



89

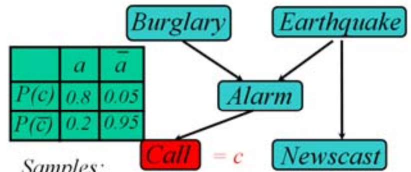
Likelihood weighting P(B|C)



Samples:

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

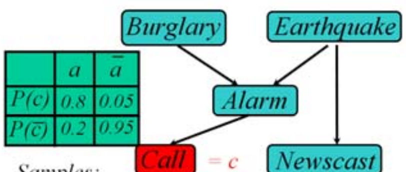
Likelihood weighting P(B|C)



Samples:

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

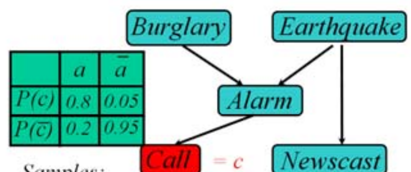
Likelihood weighting P(B|C)



Samples:

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

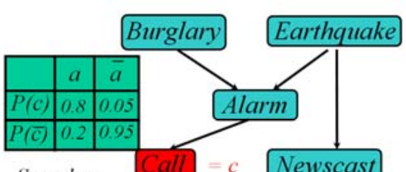
Likelihood weighting P(B|C)



Samples:

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

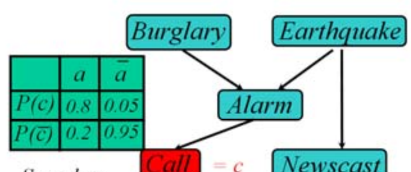
Likelihood weighting P(B|C)



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

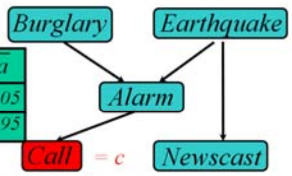
Likelihood weighting P(B|C)



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 n	0.8
b e $\bar{a}$ c n	0.05
⋮	

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

© Jack Breese (Microsoft) & Daphne Koller (Stanford)

96 72

Likelihood Weighting

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

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

= P(z,e)

➔ returns consistent estimates

performance degrades w/ many evidence vars

- a few samples get majority of the weight
- late occurring evidence vars don't guide sample generation

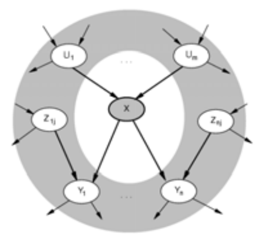
97

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 Y is true = its posterior probability.
- Converges to true posterior when frequencies stop changing significantly
  - stable distribution, mixing

98

Given Markov Blanket, X is Independent of All Other Nodes



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

© D. Weld and D. Fox

99

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

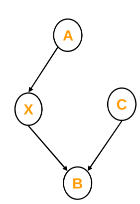
- So problem becomes calculating Pr(X=true | MB(X))
  - Fortunately, it is easy to solve exactly

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

100

Example

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



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

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

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

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

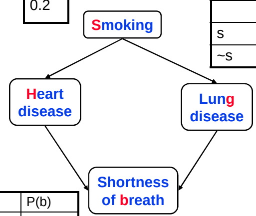
101

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

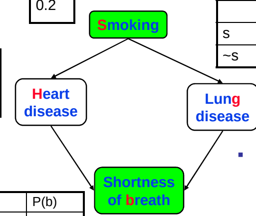
102

Example

P(s)
0.2

	P(g)
s	0.8
~s	0.1

	P(h)
s	0.6
~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

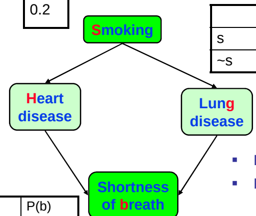
103

Example

P(s)
0.2

	P(g)
s	0.8
~s	0.1

	P(h)
s	0.6
~s	0.1



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

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

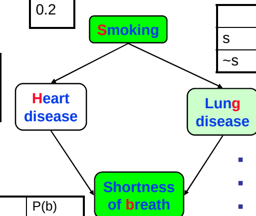
104

Example

P(s)
0.2

	P(g)
s	0.8
~s	0.1

	P(h)
s	0.6
~s	0.1



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

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

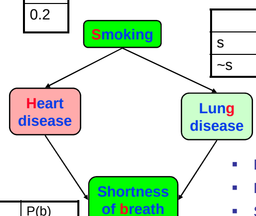
105

Example

P(s)
0.2

	P(g)
s	0.8
~s	0.1

	P(h)
s	0.6
~s	0.1



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

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

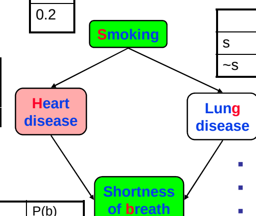
106

Example

P(s)
0.2

	P(g)
s	0.8
~s	0.1

	P(h)
s	0.6
~s	0.1



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

Suppose result is ~h

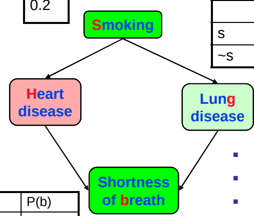
Sample G using P(G|s,~h,b)

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

107

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

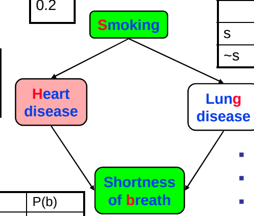
Sample G using P(G|s,~h,b)

⇒Suppose result is g

108

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)

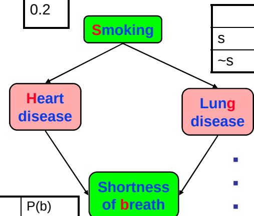
⇒Suppose result is g

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

109

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)

110

Gibbs MCMC Summary

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

- 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

111

Other inference methods

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

112

Outline

- Probabilistic models
  - Bayesian Networks (BNs)
  - Independence in BNs
  - Efficient Inference in BNs
  - ***Learning***