Approximate Inference in Bayes Nets
Sampling based methods

Mausam

(Based on slides by Jack Breese and Daphne Koller)

Intuition

- Suppose I have a coin whose p(heads) is unknown
- How could I estimate it?
- When will I get the correct probability?
- Bayes Net inference is not a learning problem
  - But similar intuitions apply
  - In particular, generate samples from a Bayes net
  - But the samples should be unbiased!

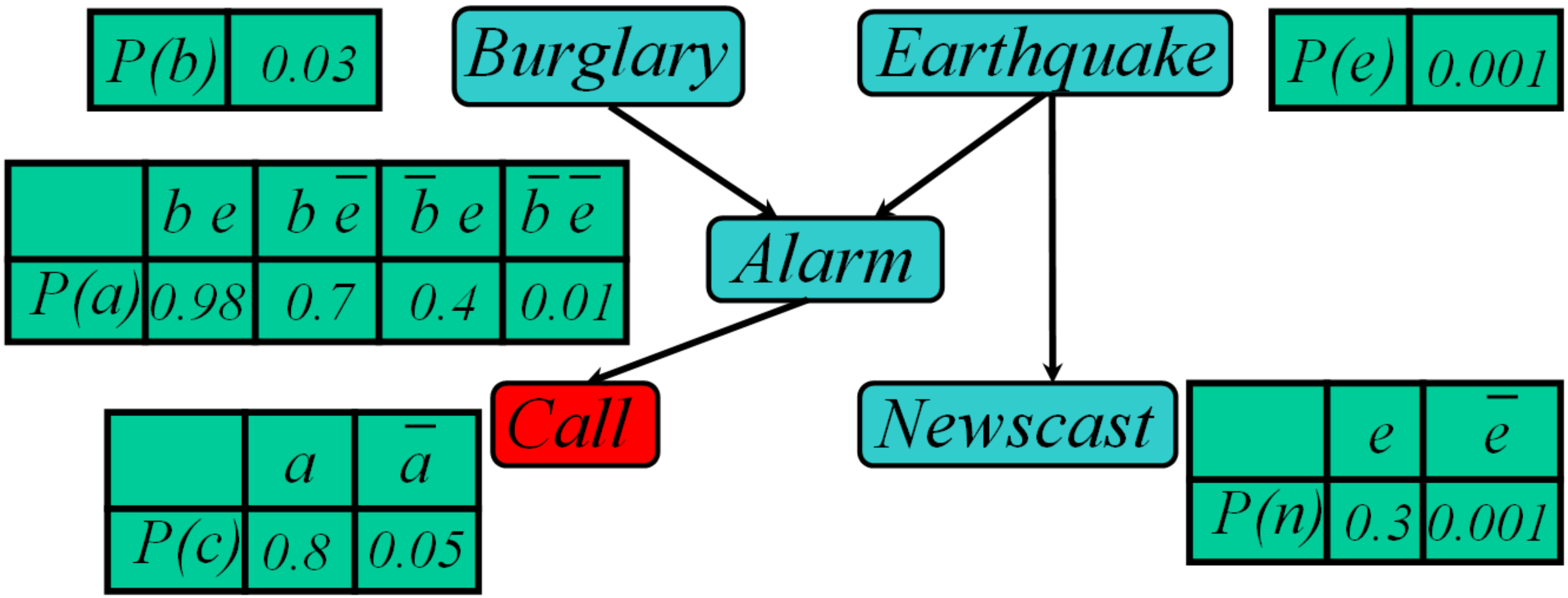
Sampling

- Samples should be representative of the world

- Samples: P(people > 60 yrs age in Delhi)
  - Computer Science class
  - Call on landline
  - Call on cellphone
  - Check facebook…
  - Count at election booth

Bayes Nets 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



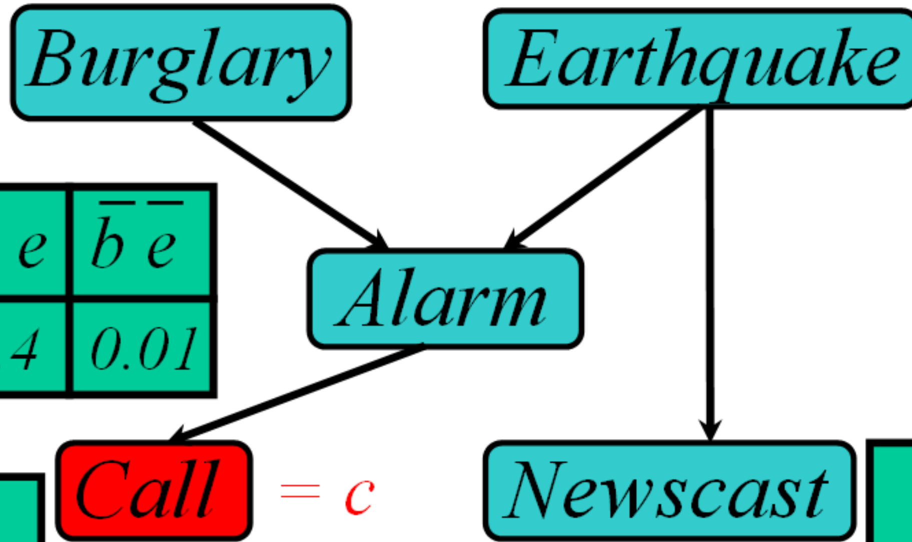
Use the samples to compute marginal probabilities, say P(c)

Stochastic simulation P(B|C)

$P(b)$	0.03

$P(e)$	0.001

	$b\ e$	$b\ \bar{e}$	$\bar{b}\ e$	$\bar{b}\ \bar{e}$
$P(a)$	0.98	0.7	0.4	0.01



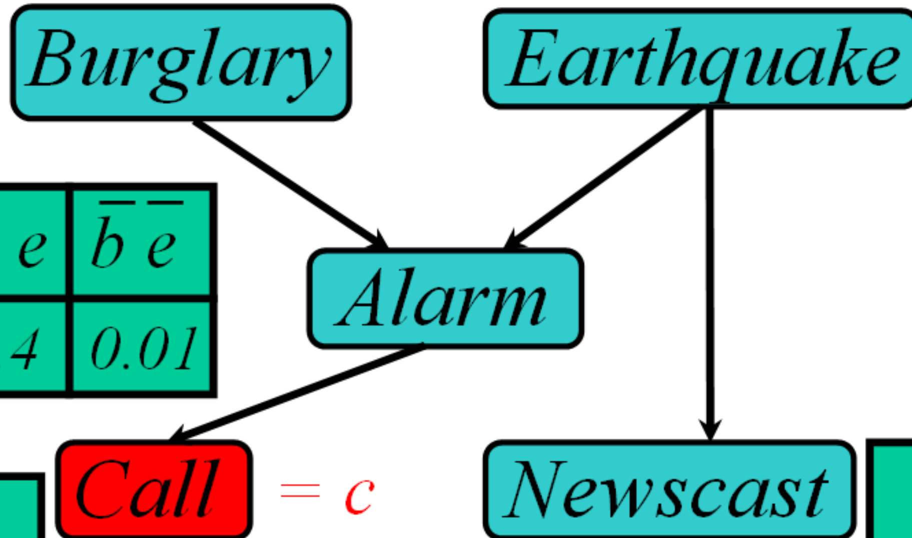
	a	$\bar{a}$
$P(c)$	0.8	0.05

	e	$\bar{e}$
$P(n)$	0.3	0.001

Stochastic simulation P(B|C)

$P(b)$	0.03

$P(e)$	0.001



	$b\ e$	$b\ \bar{e}$	$\bar{b}\ e$	$\bar{b}\ \bar{e}$
$P(a)$	0.98	0.7	0.4	0.01

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

	e	$\bar{e}$
$P(n)$	0.3	0.001

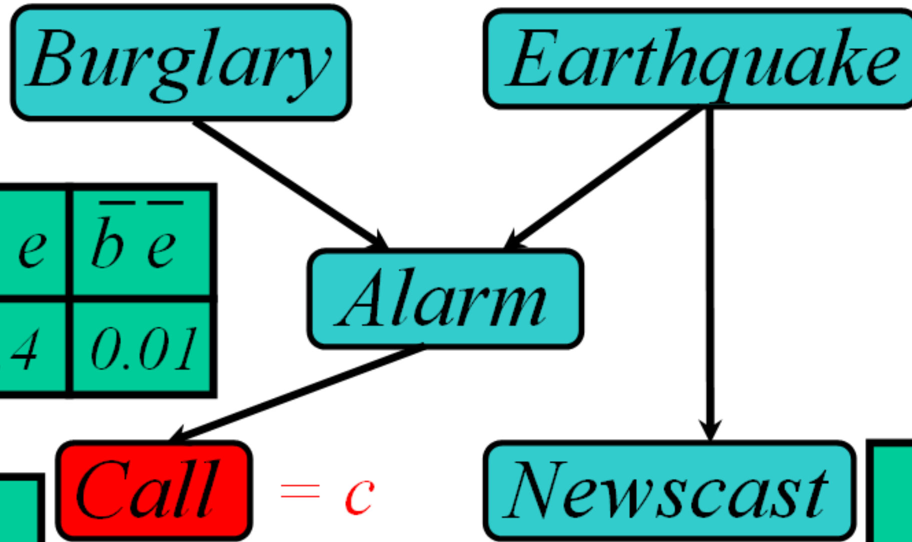
Samples:

B	E	A	C	N

Stochastic simulation P(B|C)

$P(b)$	0.03

$P(e)$	0.001



	$b\ e$	$b\ \bar{e}$	$\bar{b}\ e$	$\bar{b}\ \bar{e}$
$P(a)$	0.98	0.7	0.4	0.01

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

	e	$\bar{e}$
$P(n)$	0.3	0.001

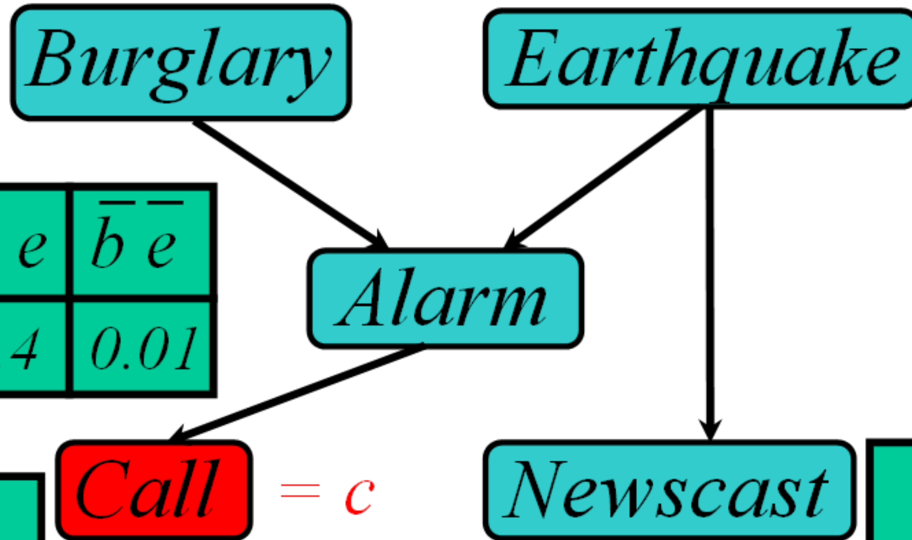
Samples:

B	*E*	*A*	*C*	*N*

Stochastic simulation P(B|C)

$P(b)$	0.03

$P(e)$	0.001



	$b\ e$	$b\ \bar{e}$	$\bar{b}\ e$	$\bar{b}\ \bar{e}$
$P(a)$	0.98	0.7	0.4	0.01

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

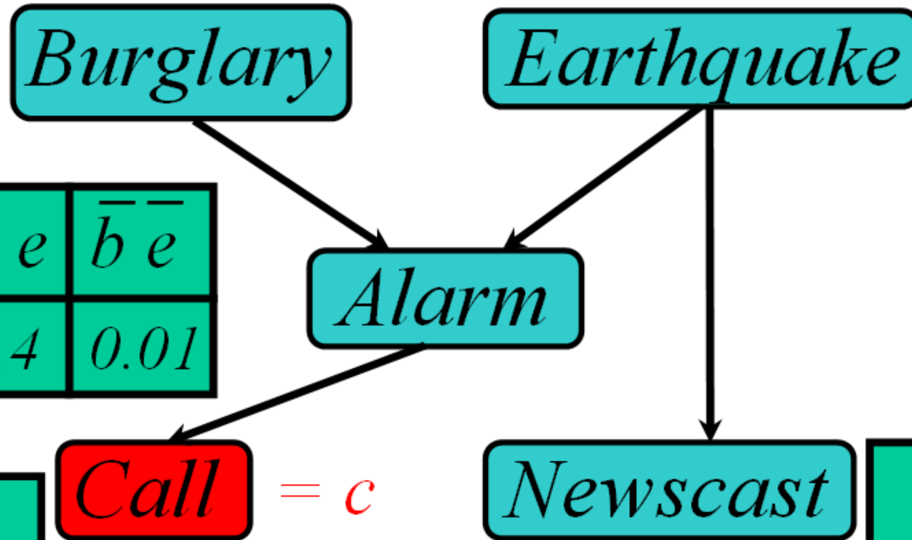
	e	$\bar{e}$
$P(n)$	0.3	0.001

Samples:

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

Stochastic simulation P(B|C)

$P(b)$	0.03



$P(e)$	0.001

	$b\ e$	$b\ \bar{e}$	$\bar{b}\ e$	$\bar{b}\ \bar{e}$
$P(a)$	0.98	0.7	0.4	0.01

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

	e	$\bar{e}$
$P(n)$	0.3	0.001

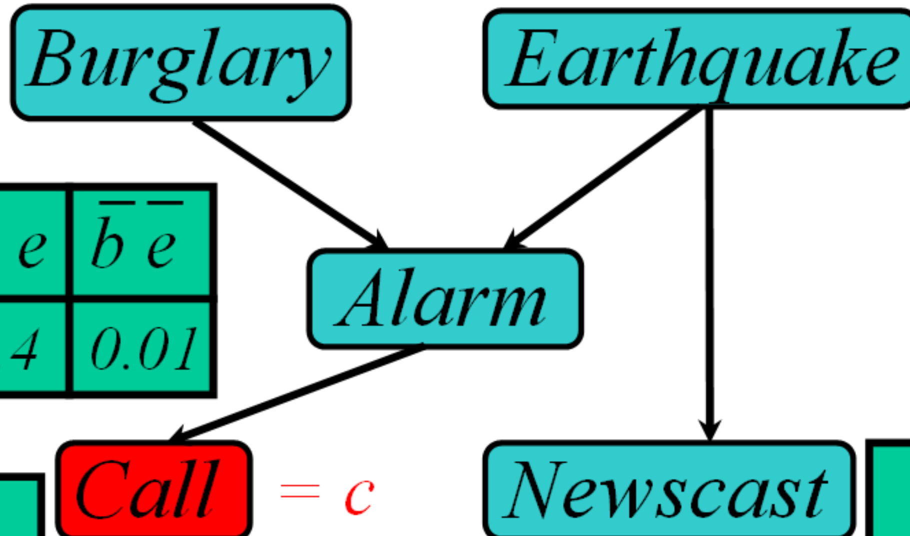
Samples:

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

Stochastic simulation P(B|C)

$P(b)$	0.03

$P(e)$	0.001



	$b\ e$	$b\ \bar{e}$	$\bar{b}\ e$	$\bar{b}\ \bar{e}$
$P(a)$	0.98	0.7	0.4	0.01

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

	e	$\bar{e}$
$P(n)$	0.3	0.001

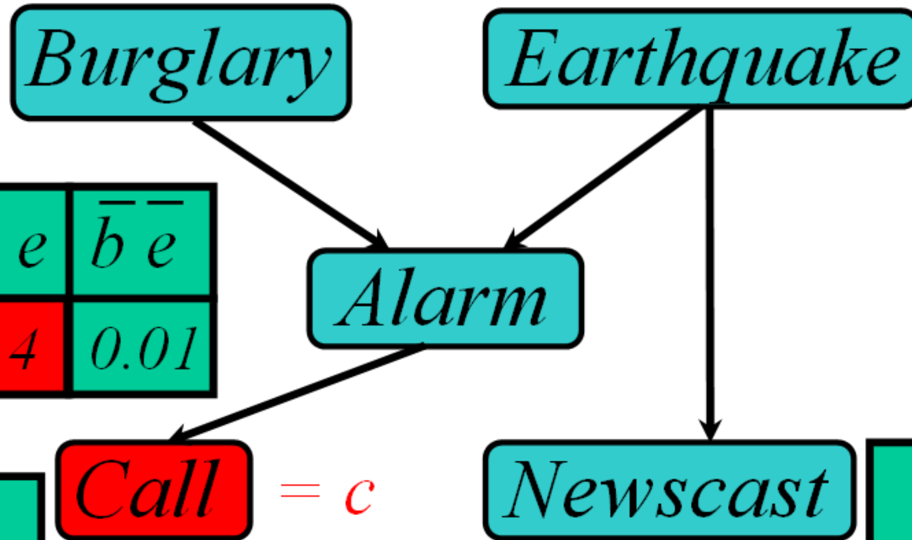
Samples:

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

Stochastic simulation P(B|C)

$P(b)$	0.03

$P(e)$	0.001



	$b\ e$	$b\ \bar{e}$	$\bar{b}\ e$	$\bar{b}\ \bar{e}$
$P(a)$	0.98	0.7	0.4	0.01

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

	e	$\bar{e}$
$P(n)$	0.3	0.001

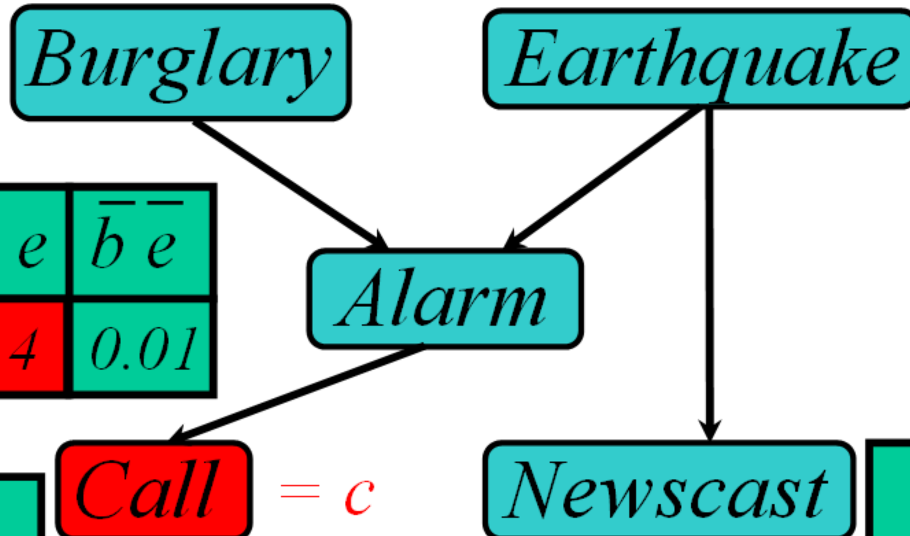
Samples:

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

Stochastic simulation P(B|C)

$P(b)$	0.03

$P(e)$	0.001



	$b\ e$	$b\ \bar{e}$	$\bar{b}\ e$	$\bar{b}\ \bar{e}$
$P(a)$	0.98	0.7	0.4	0.01

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

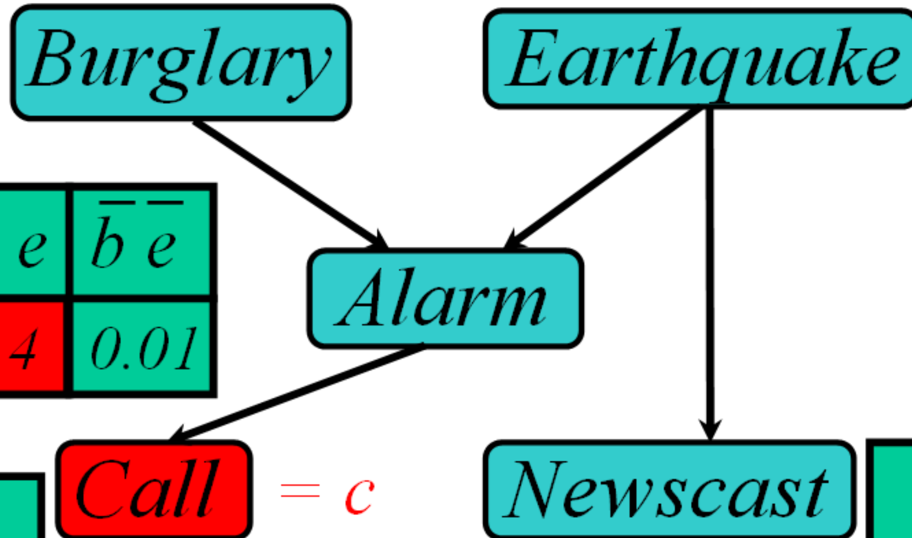
	e	$\bar{e}$
$P(n)$	0.3	0.001

Samples:

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

Stochastic simulation P(B|C)

$P(b)$	0.03



$P(e)$	0.001

	$b\ e$	$b\ \bar{e}$	$\bar{b}\ e$	$\bar{b}\ \bar{e}$
$P(a)$	0.98	0.7	0.4	0.01

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

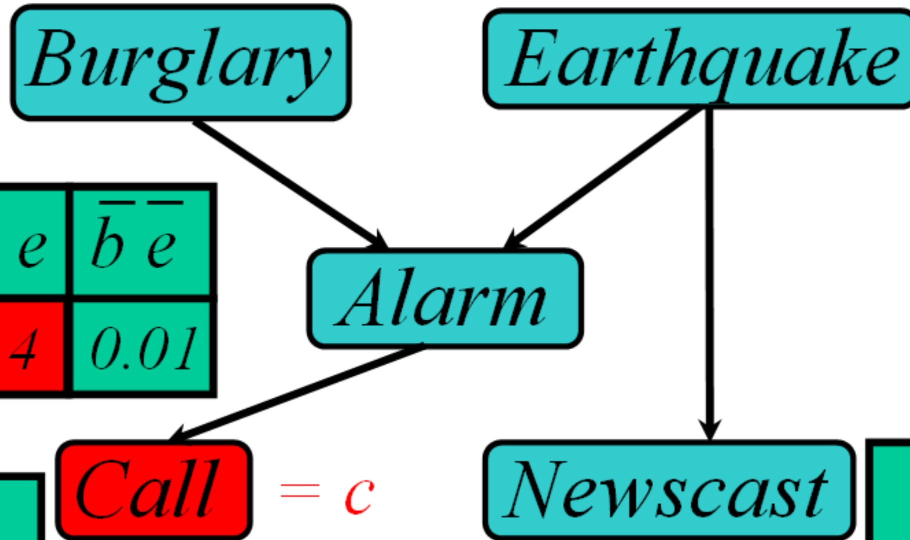
	e	$\bar{e}$
$P(n)$	0.3	0.001

Samples:

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

Stochastic simulation P(B|C)

$P(b)$	0.03



$P(e)$	0.001

	$b\ e$	$b\ \bar{e}$	$\bar{b}\ e$	$\bar{b}\ \bar{e}$
$P(a)$	0.98	0.7	0.4	0.01

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

	e	$\bar{e}$
$P(n)$	0.3	0.001

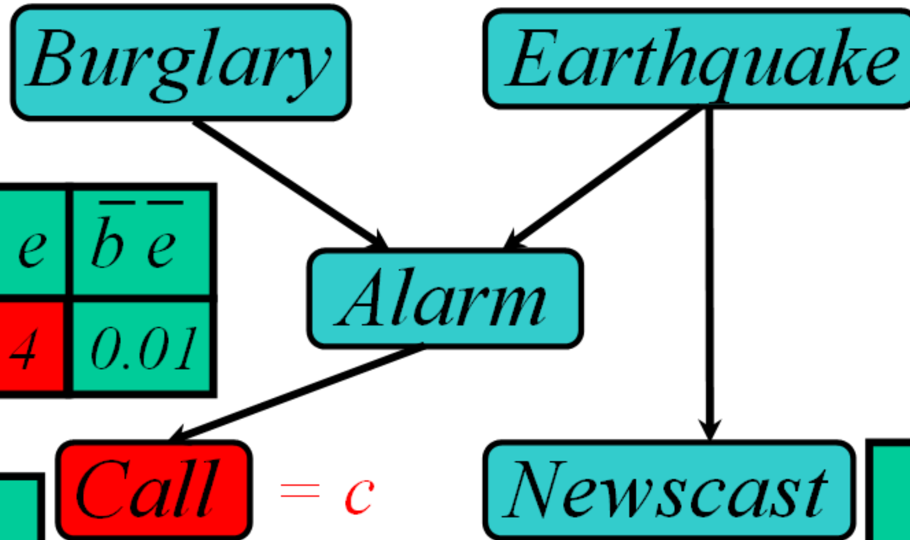
Samples:

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

Stochastic simulation P(B|C)

$P(b)$	0.03

$P(e)$	0.001



	$b\ e$	$b\ \bar{e}$	$\bar{b}\ e$	$\bar{b}\ \bar{e}$
$P(a)$	0.98	0.7	0.4	0.01

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

	e	$\bar{e}$
$P(n)$	0.3	0.001

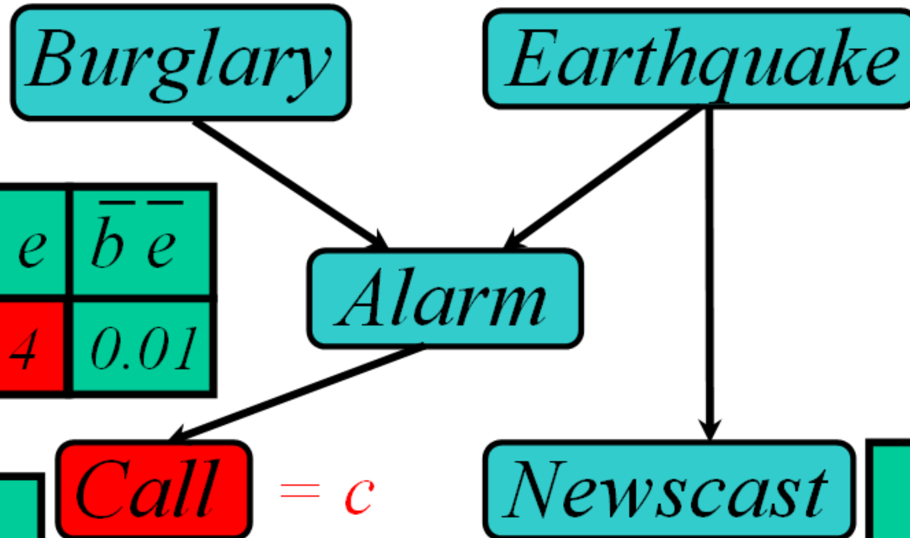
Samples:

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

Stochastic simulation P(B|C)

$P(b)$	0.03

$P(e)$	0.001



	$b\ e$	$b\ \bar{e}$	$\bar{b}\ e$	$\bar{b}\ \bar{e}$
$P(a)$	0.98	0.7	0.4	0.01

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

	e	$\bar{e}$
$P(n)$	0.3	0.001

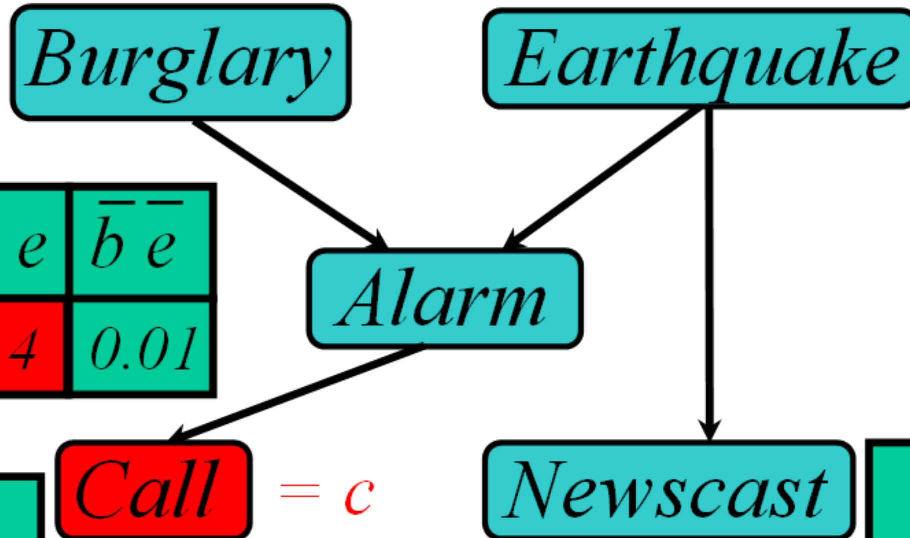
Samples:

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

Stochastic simulation P(B|C)

$P(b)$	0.03

$P(e)$	0.001



	$b\ e$	$b\ \bar{e}$	$\bar{b}\ e$	$\bar{b}\ \bar{e}$
$P(a)$	0.98	0.7	0.4	0.01

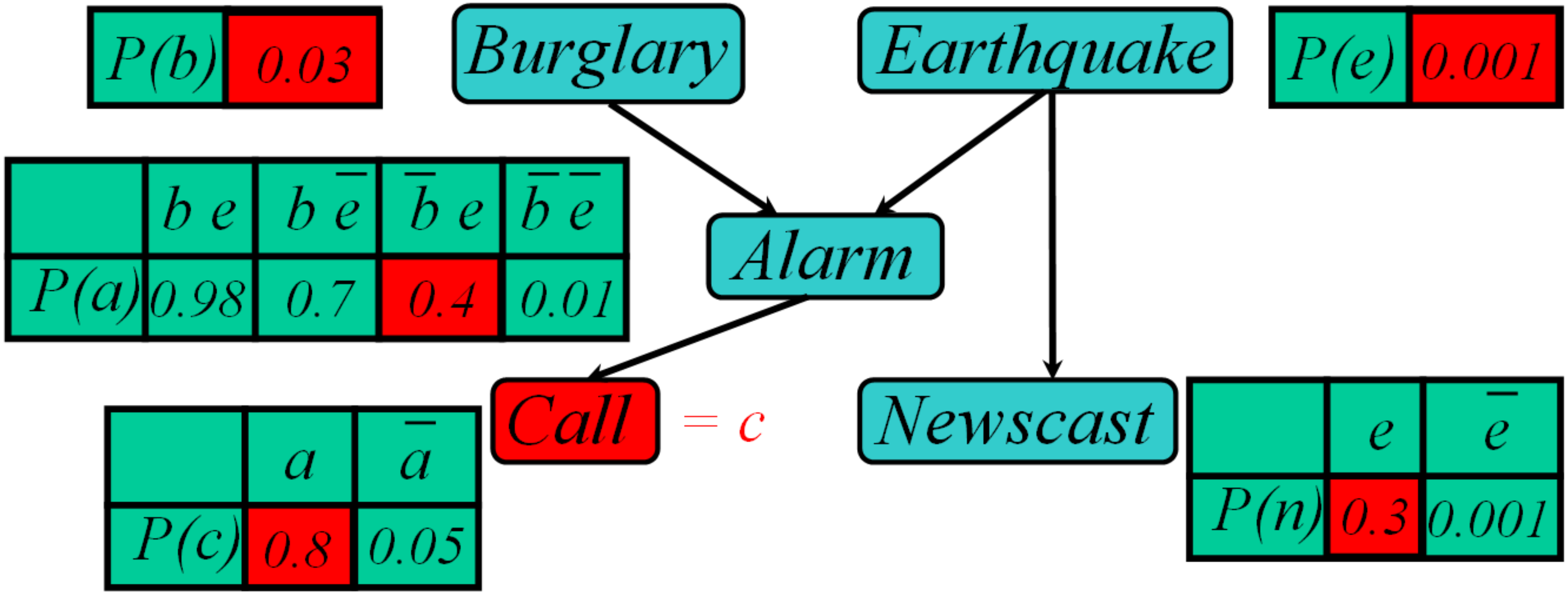
	a	$\bar{a}$
$P(c)$	0.8	0.05

	e	$\bar{e}$
$P(n)$	0.3	0.001

Samples:

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

Stochastic simulation P(B|C)



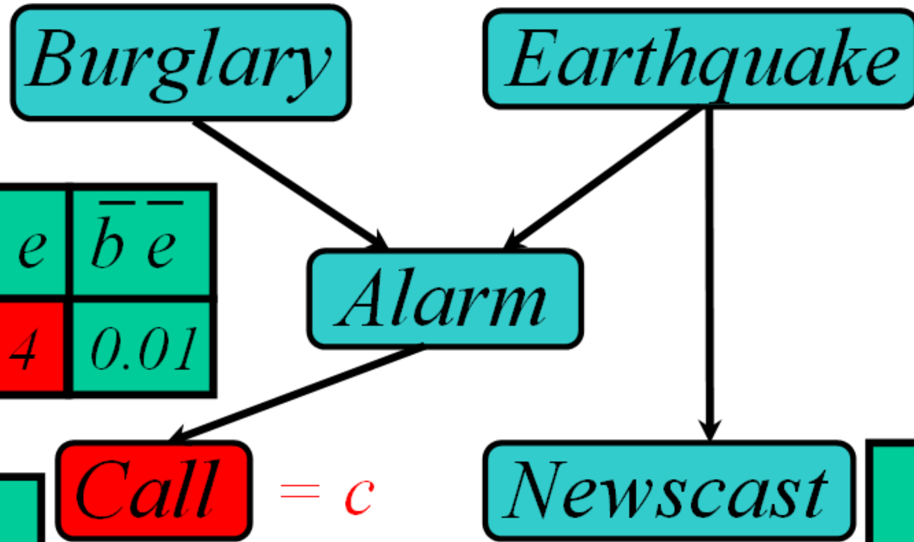
Samples:

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

Stochastic simulation P(B|C)

$P(b)$	0.03

$P(e)$	0.001



	$b\,e$	$b\,\bar{e}$	$\bar{b}\,e$	$\bar{b}\,\bar{e}$
$P(a)$	0.98	0.7	0.4	0.01

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

	e	$\bar{e}$
$P(n)$	0.3	0.001

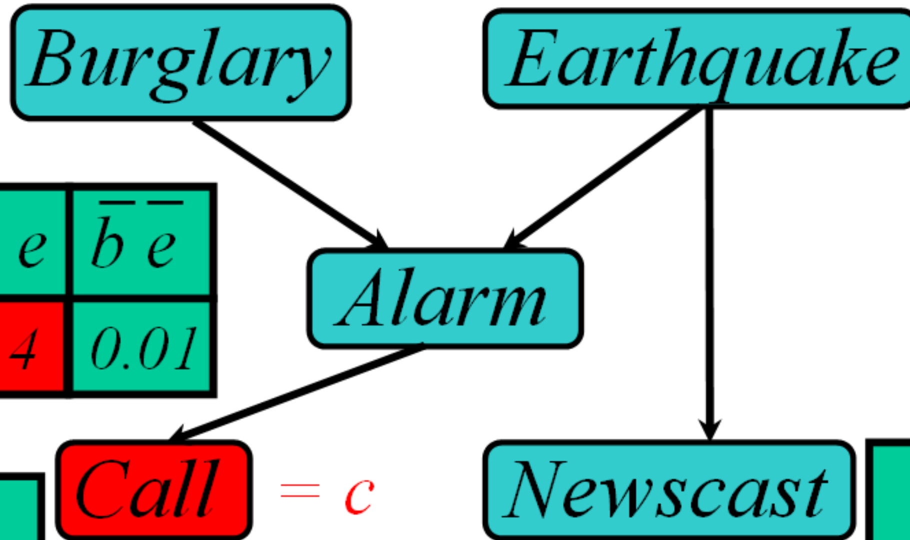
Samples:

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

Stochastic simulation P(B|C)

$P(b)$	0.03

$P(e)$	0.001



	$b\ e$	$b\ \bar{e}$	$\bar{b}\ e$	$\bar{b}\ \bar{e}$
$P(a)$	0.98	0.7	0.4	0.01

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

	e	$\bar{e}$
$P(n)$	0.3	0.001

Samples:

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

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

Rejection Sampling

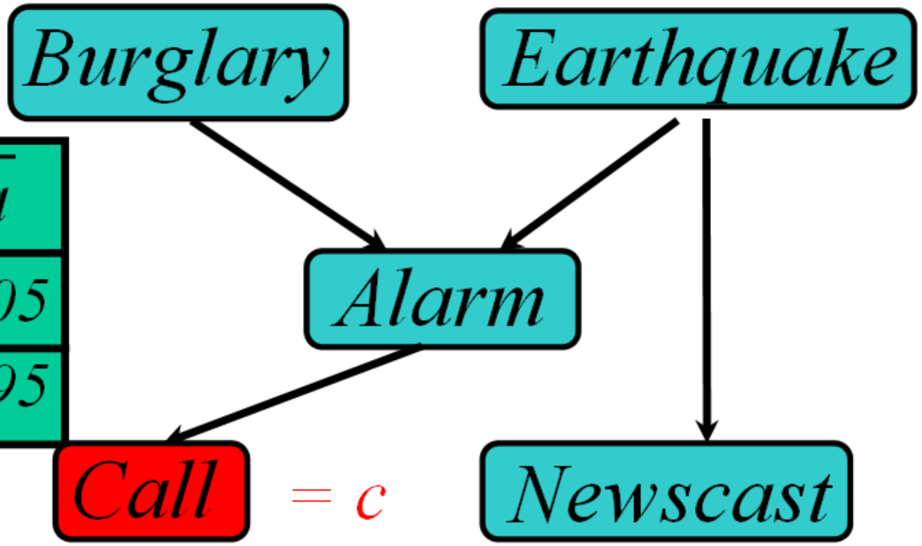
- Sample from the prior
  - reject if do not match the evidence
- Returns consistent posterior estimates
- Hopelessly expensive if P(e) is small
  - P(e) drops off exponentially with no. of evidence vars

Likelihood Weighting

- Idea
  - each sample agrees with evidence
  - pays some price for the agreement (weight)

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

Likelihood weighting P(B|C)



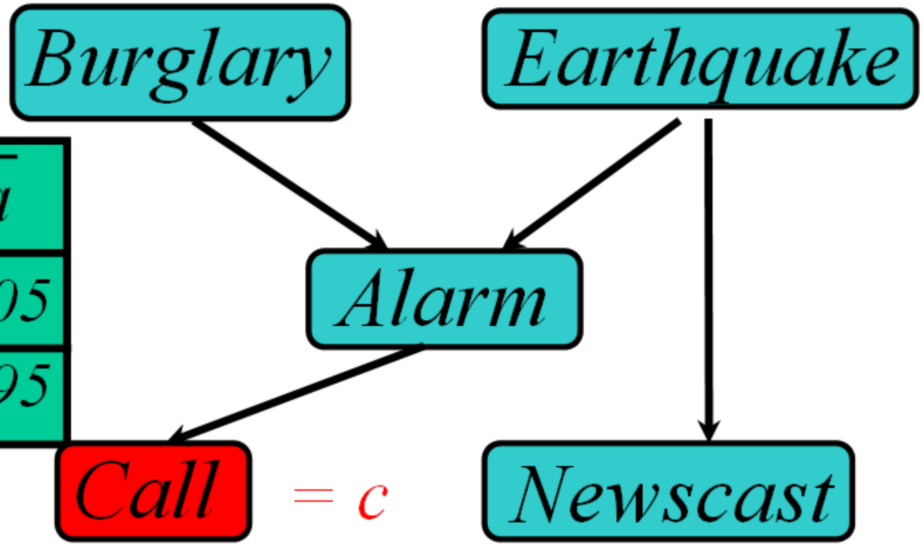
	a	$\bar{a}$
$P(c)$	0.8	0.05
$P(\bar{c})$	0.2	0.95

Samples:

B	***E***	***A***	***C***	***N***

Likelihood weighting P(B|C)

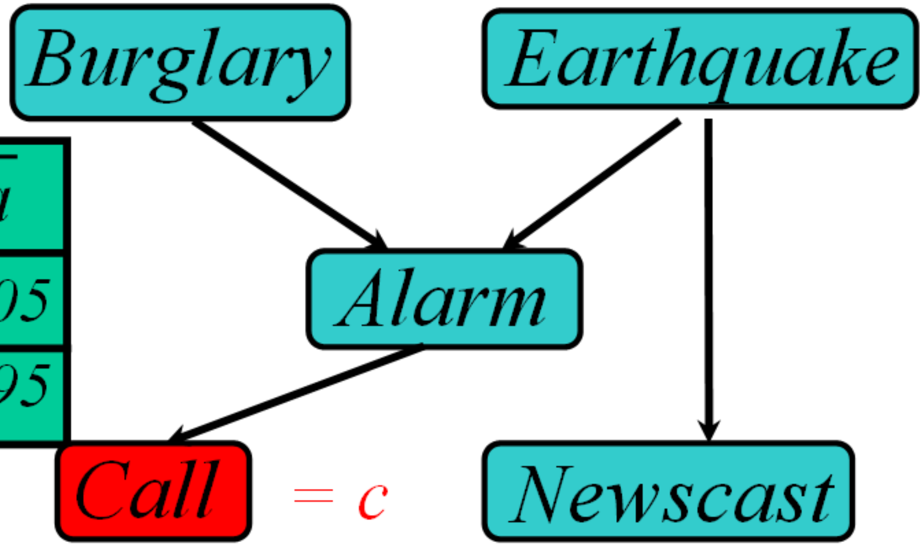
	a	$\bar{a}$
$P(c)$	0.8	0.05
$P(\bar{c})$	0.2	0.95



Samples:

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

Likelihood weighting P(B|C)



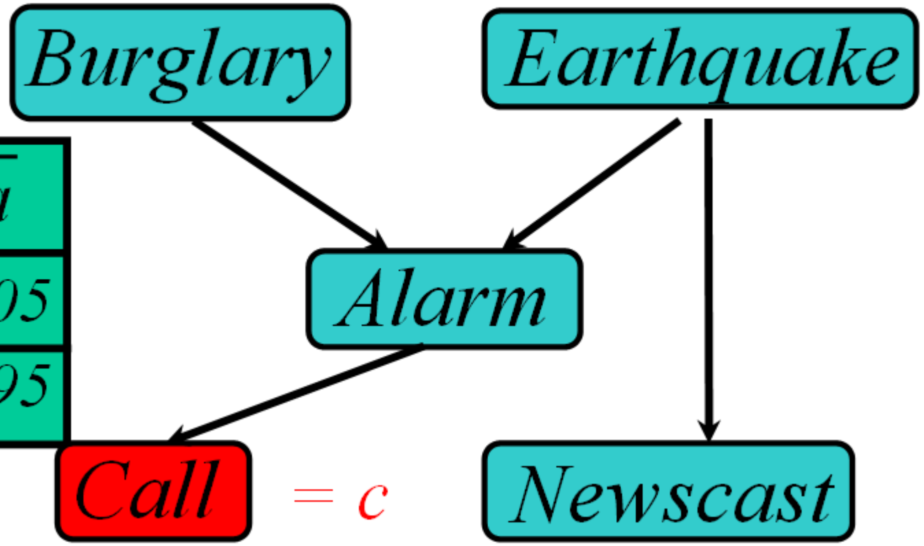
	a	$\bar{a}$
$P(c)$	0.8	0.05
$P(\bar{c})$	0.2	0.95

Samples:

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

Likelihood weighting P(B|C)

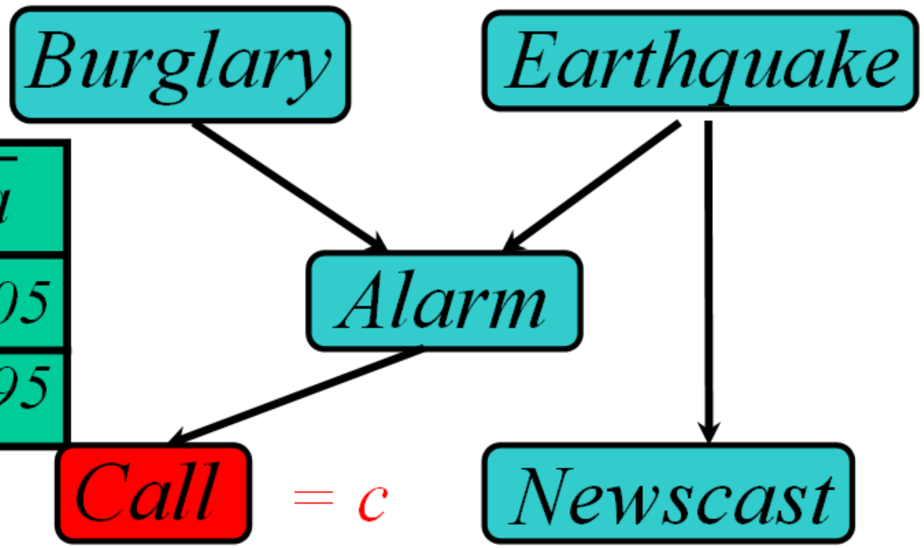
	a	$\bar{a}$
$P(c)$	0.8	0.05
$P(\bar{c})$	0.2	0.95



Samples:

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

Likelihood weighting P(B|C)

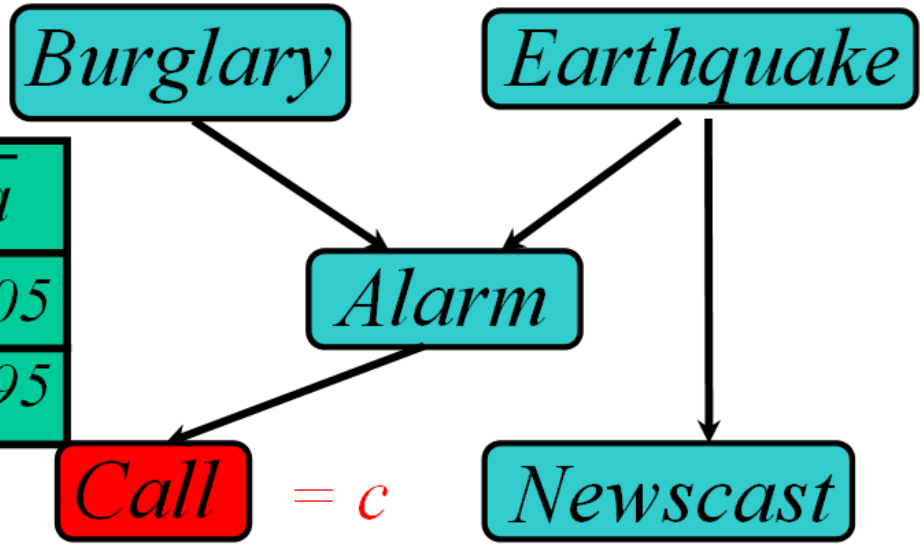


	a	$\bar{a}$
$P(c)$	0.8	0.05
$P(\bar{c})$	0.2	0.95

Samples:

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

Likelihood weighting P(B|C)



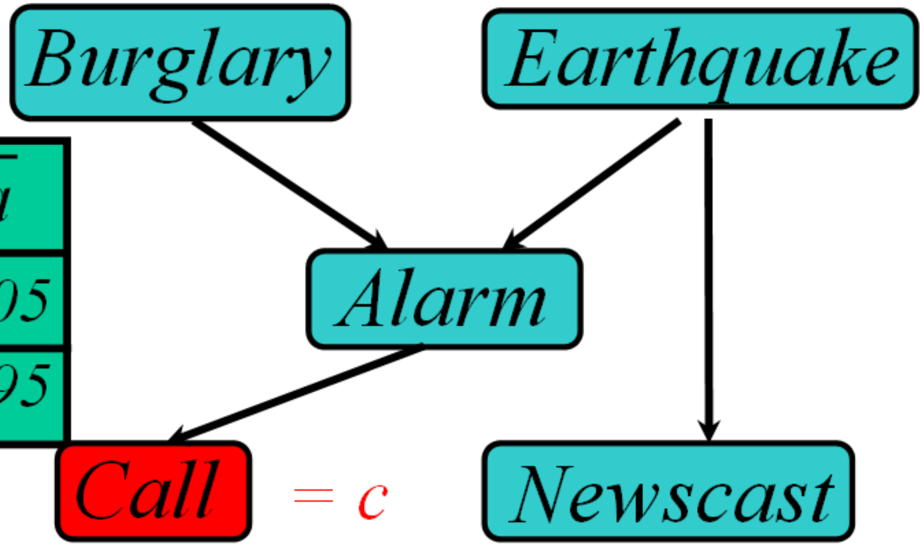
	a	$\bar{a}$
$P(c)$	0.8	0.05
$P(\bar{c})$	0.2	0.95

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

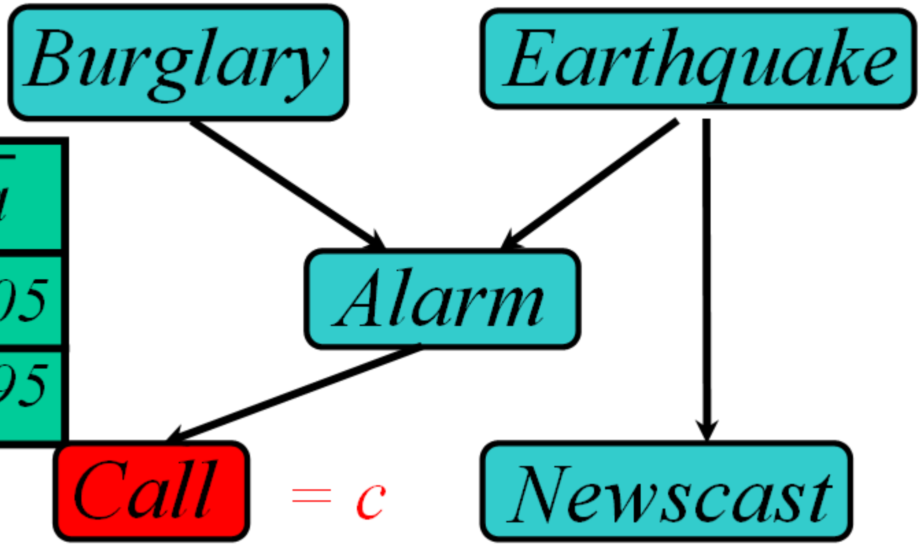


Samples:

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

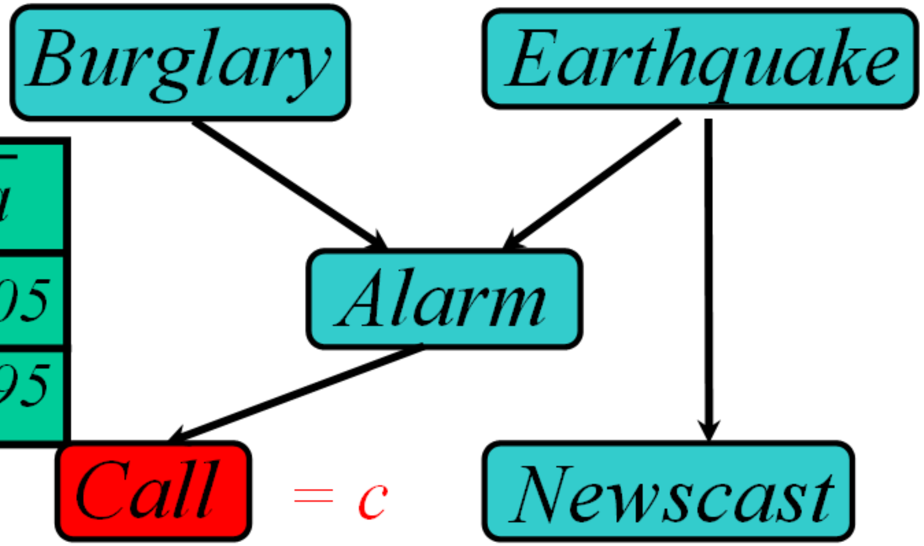


Samples:

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)

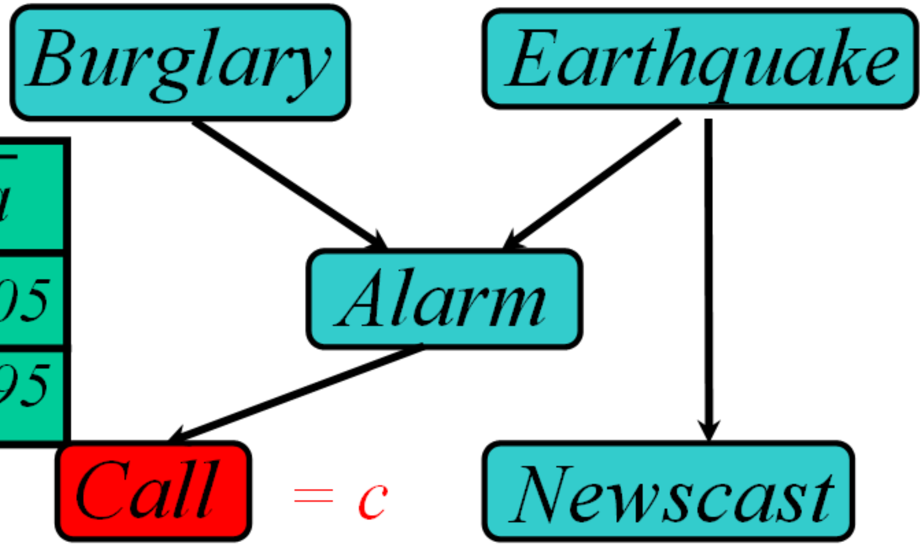
	a	$\overline{a}$
$P(c)$	0.8	0.05
$P(\overline{c})$	0.2	0.95



Samples:

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

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
		$\vdots$			

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

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
  - stationary distribution, mixing

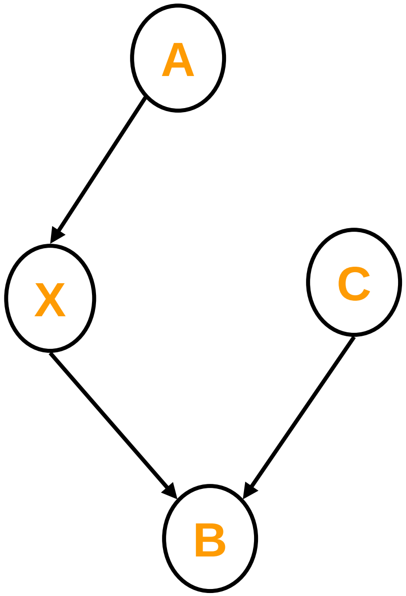
Markov Blanket Sampling

- 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

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



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

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

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

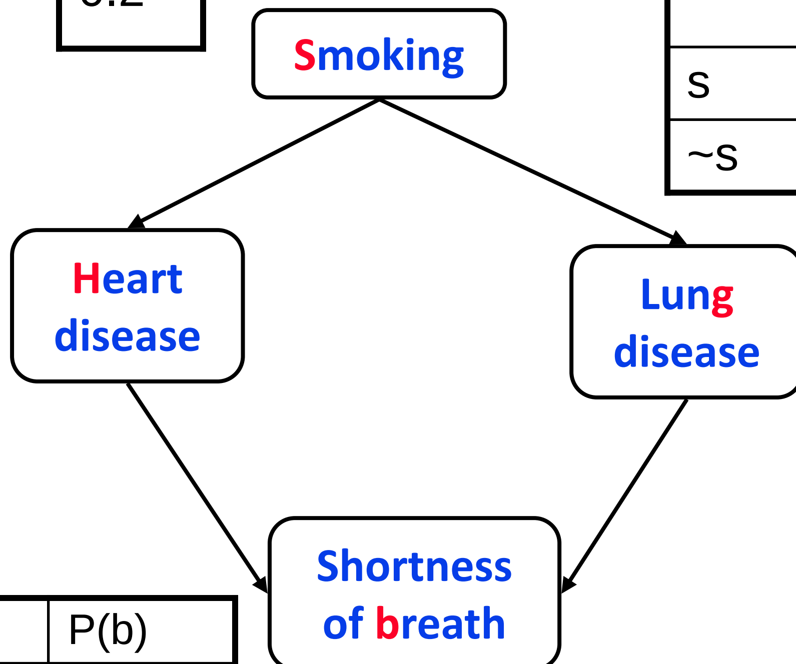
$$= \alpha P(X \mid A)P(B \mid X, C)$$

Example

P(s)
0.2

	P(g)
s	0.8
~s	0.1

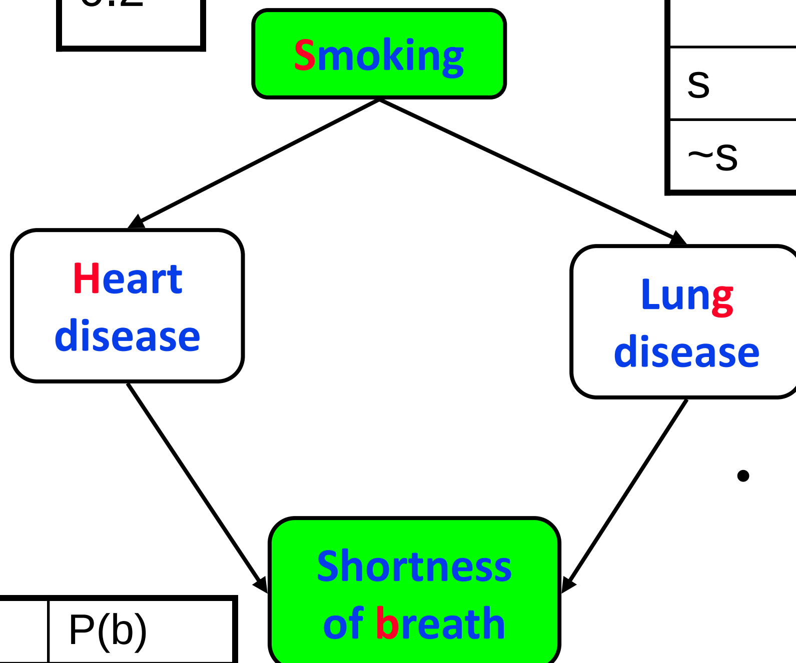
	P(h)
s	0.6
~s	0.1



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

Example

P(s)
0.2



	P(g)
s	0.8
~s	0.1

	P(h)
s	0.6
~s	0.1

- Evidence: s, b

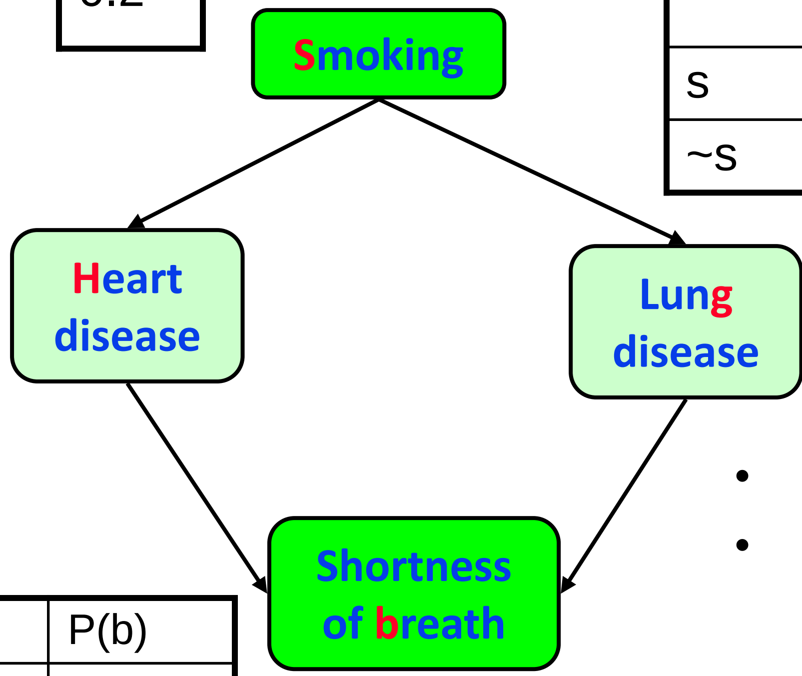
H	G	P(b)
h	g	0.9
h	~g	0.8
~h	g	0.7
~h	~g	0.1

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

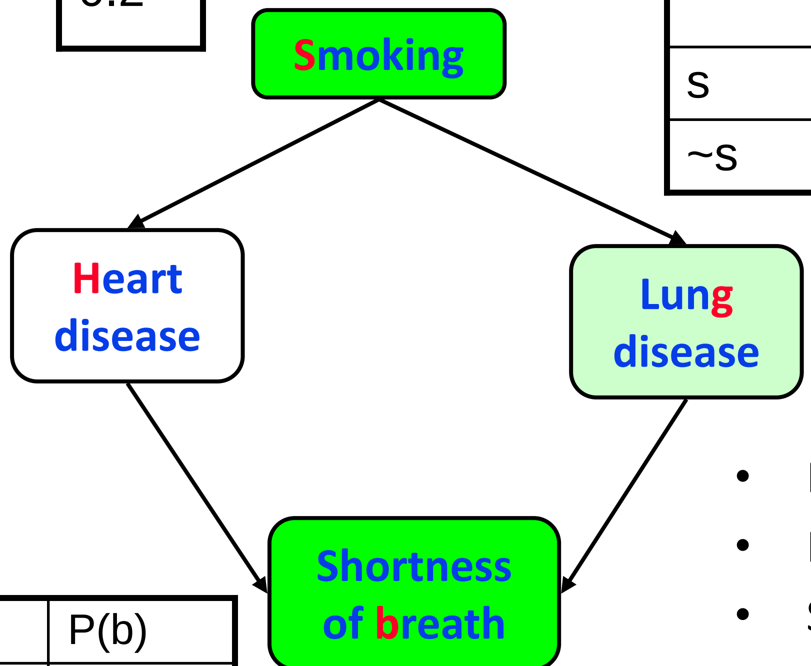
H	G	P(b)
h	g	0.9
h	~g	0.8
~h	g	0.7
~h	~g	0.1

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

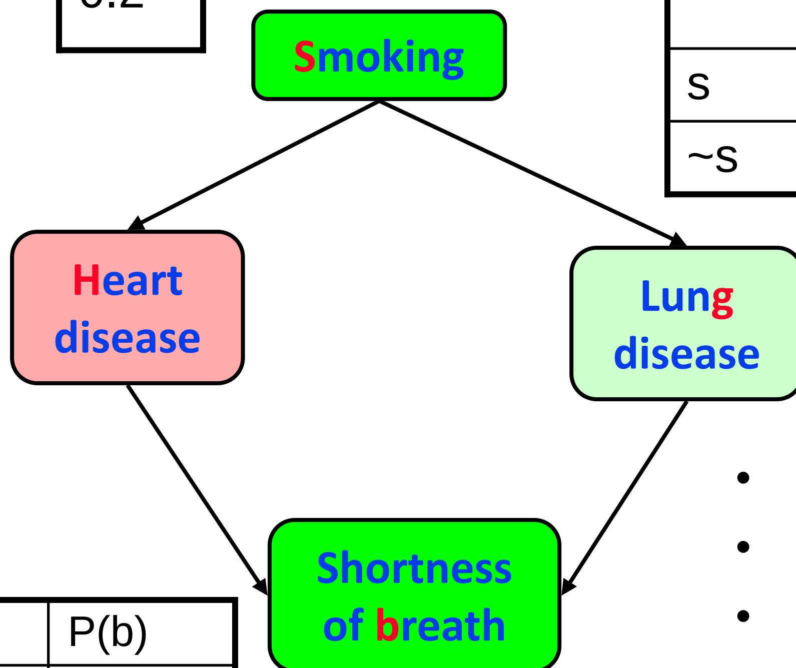
- Evidence: s, b
- Randomly set: h, g
- Sample H using P(H|s,g,b)

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

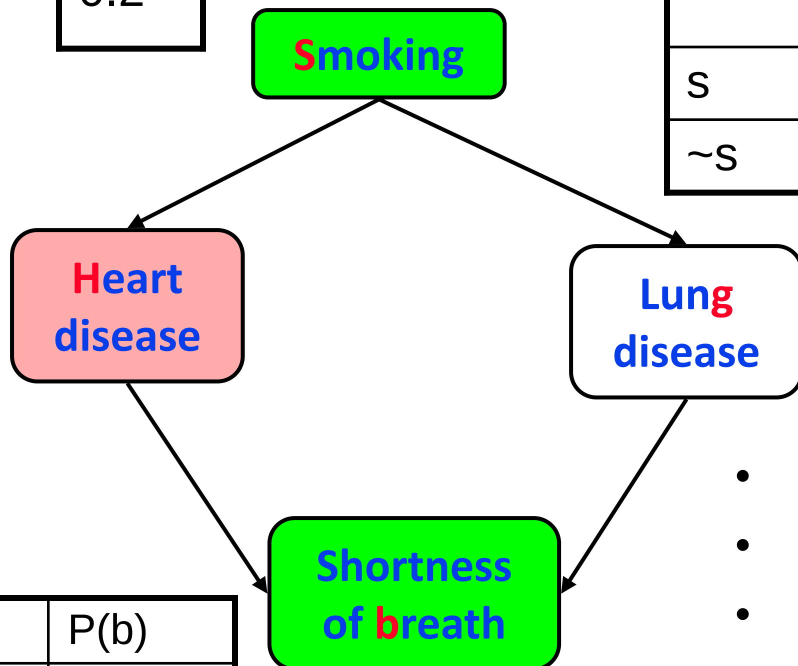
- Evidence: s, b
- Randomly set: ~h, g
- Sample H using P(H|s,g,b)
- $\Rightarrow$ Suppose result is ~h

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

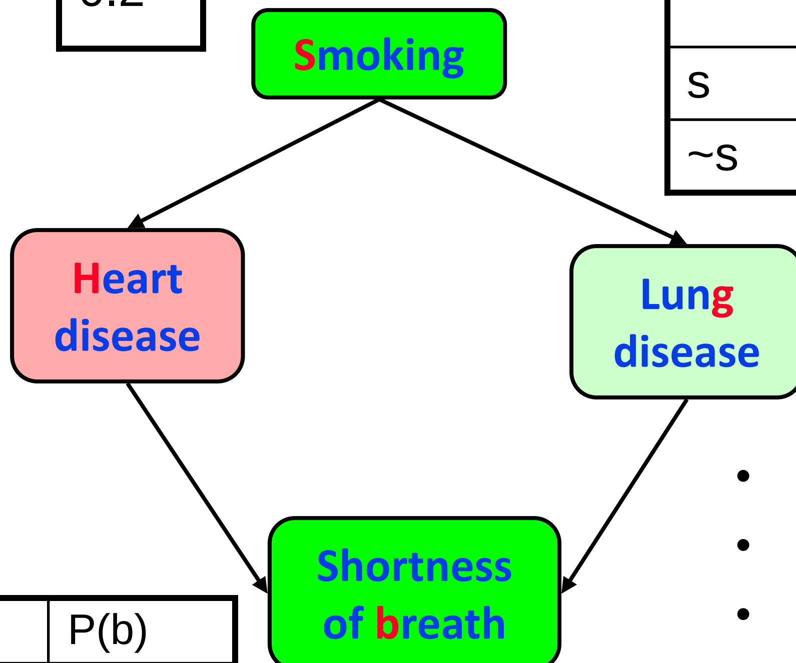
Sample G using P(G|s,~h,b)

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

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

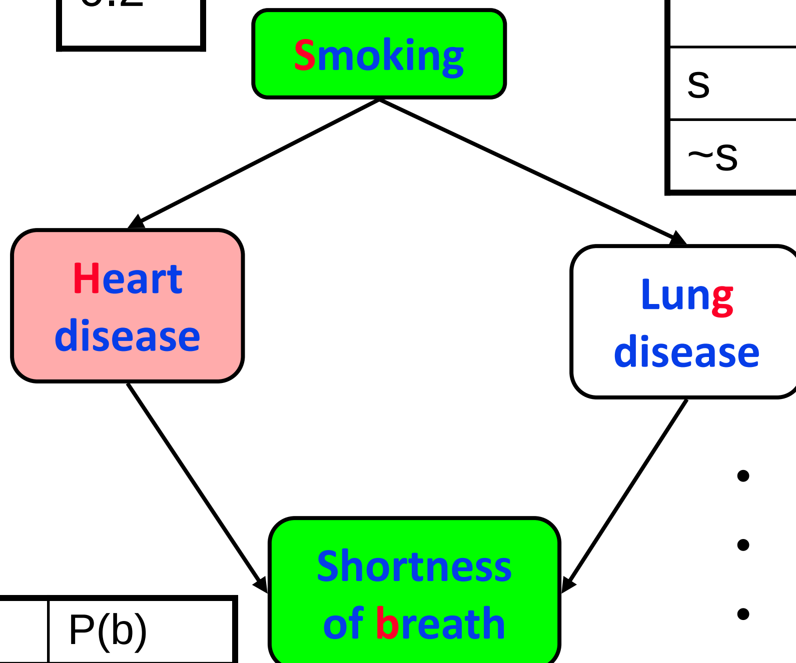
Suppose result is ~h

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

⇒Suppose result is g

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

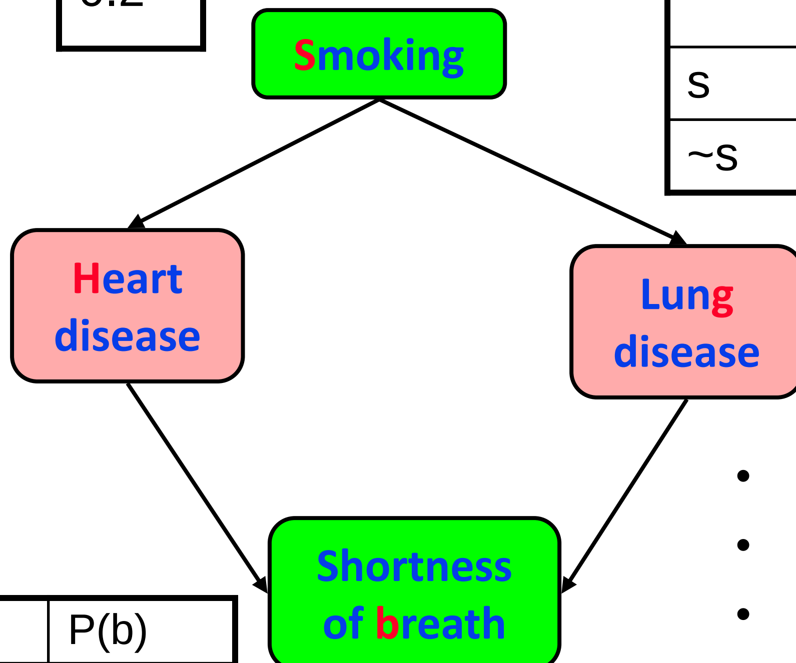
Sample G using P(G|s,~h,b)

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)

⇒Suppose result is ~g

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 two variables is *deterministic*
  - Many variations have been devised to make MCMC more robust

Other inference methods

- Exact inference
  - Junction tree

- Approximate inference
  - Belief Propagation
  - Variational Methods
  - Metropolis-Hastings

Programming Assignment

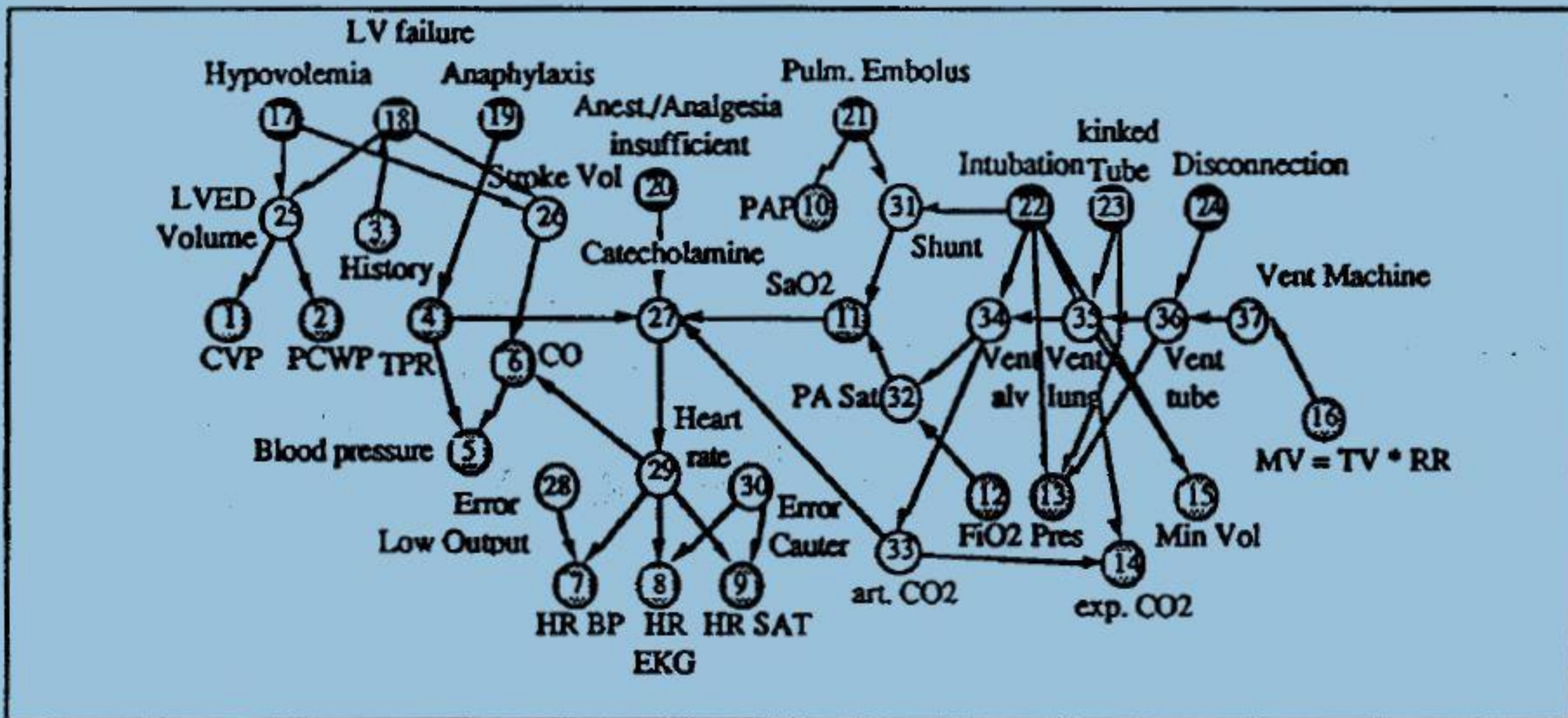


Fig. 1 The ALARM network representing causal relationships is shown with diagnostic (●), intermediate (○) and measurement (◉) nodes. CO: cardiac output, CVP: central venous pressure, LVED volume: left ventricular end-diastolic volume, LV failure: left ventricular failure, MV: minute ventilation, PA Sat: pulmonary artery oxygen saturation, PAP: pulmonary artery pressure, PCWP: pulmonary capillary wedge pressure, Pres: breathing pressure, RR: respiratory rate, TPR: total peripheral resistance, TV: tidal volume