Bayesian Networks

Read R&N Ch. 14.1-14.2

Next lecture: Read R&N 18.1-18.4

You will be expected to know

- Basic concepts and vocabulary of Bayesian networks.
  - Nodes represent random variables.
  - Directed arcs represent (informally) direct influences.
  - Conditional probability tables, P(Xi | Parents(Xi)).

- Given a Bayesian network:
  - Write down the full joint distribution it represents.

- Given a full joint distribution in factored form:
  - Draw the Bayesian network that represents it.

- Given a variable ordering and some background assertions of conditional independence among the variables:
  - Write down the factored form of the full joint distribution, as simplified by the conditional independence assertions.

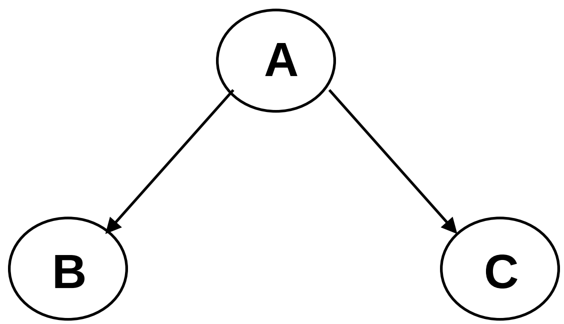
<u>“Faith, and it’s an uncertain world entirely.”</u>

--- Errol Flynn, “Captain Blood” (1935)

<u>= We need probability theory for our agents!! The world is chaos!!</u>

Could you design a logical agent that does the right thing below??

Extended example of 3-way Bayesian Networks



Conditionally independent effects:
$p(A,B,C) = p(B|A)p(C|A)p(A)$

B and C are conditionally independent Given A

E.g., A is a disease, and we model B and C as conditionally independent symptoms given A

E.g., A is Fire, B is Heat, C is Smoke.
“Where there’s Smoke, there’s Fire.”

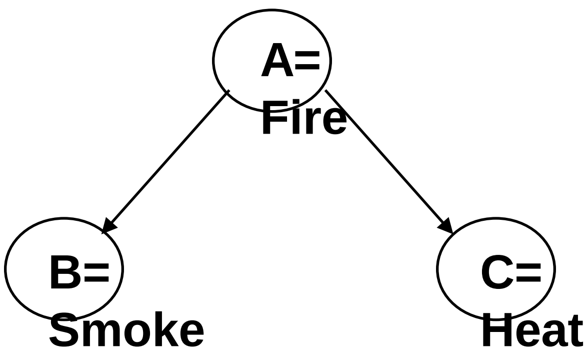
If we see Smoke, we can infer Fire.

If we see Smoke, observing Heat tells us very little additional information.

Extended example of 3-way Bayesian Networks

Suppose I build a fire in my fireplace about once every 10 days…

P(fire)
0.1



Fire	P(Smoke)
t	.90
f	.001

Fire	P(Heat)
t	.99
f	.0001

Conditionally independent effects:
P(A,B,C) = P(B|A)P(C|A)P(A)

Smoke and Heat are conditionally independent given Fire.

If we see B=Smoke, observing C=Heat tells us very little additional information.

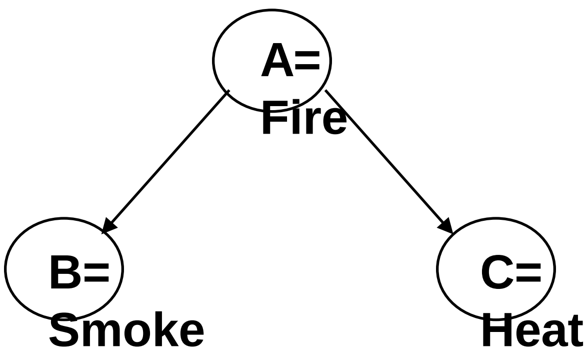
Extended example of 3-way Bayesian Networks

What is P(Fire=t | Smoke=t)?

P(Fire=t | Smoke=t)

=P(Fire=t & Smoke=t) / P(Smoke=t)

P(Fire)
0. 1

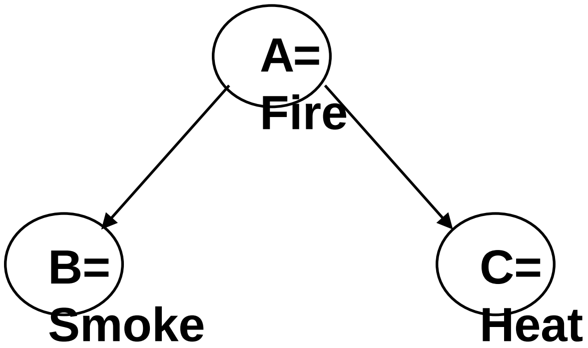


Fire	P(Smoke)
t	.90
f	.001

Fire	P(Heat)
t	.99
f	.0001

Extended example of 3-way Bayesian Networks

P(Fire)
0. 1



Fire	P(Smoke)
t	.90
f	.001

Fire	P(Heat)
t	.99
f	.0001

What is P(Fire=t & Smoke=t)?

P(Fire=t & Smoke=t)

=Σ_heat P(Fire=t&Smoke=t&heat)

=Σ_heat P(Smoke=t&heat|Fire=t)P(Fire=t)

=Σ_heat P(Smoke=t|Fire=t) P(heat|Fire=t)P(Fire=t)

=P(Smoke=t|Fire=t) P(heat=t|Fire=t)P(Fire=t)

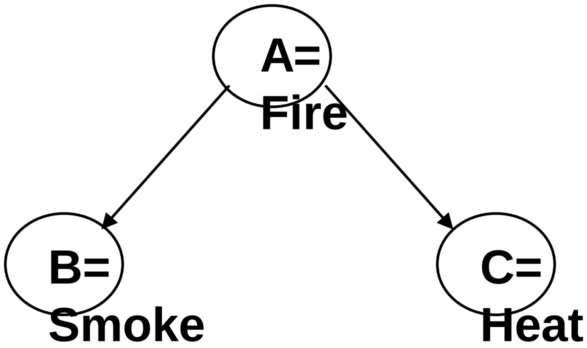
+P(Smoke=t|Fire=t)P(heat=f|Fire=t)P(Fire=t)

= (.90x.99x.1)+(.90x.01x.1)

= 0.09

Extended example of 3-way Bayesian Networks

P(Fire)
0. 1



Fire	P(Smoke)
t	.90
f	.001

Fire	P(Heat)
t	.99
f	.0001

What is P(Smoke=t)?

$P(Smoke=t)$

$=\Sigma\_fire\ \Sigma\_heat\ P(Smoke=t\&fire\&heat)$

$=\Sigma\_fire\ \Sigma\_heat\ P(Smoke=t\&heat|fire)P(fire)$

$=\Sigma\_fire\ \Sigma\_heat\ P(Smoke=t|fire)\ P(heat|fire)P(fire)$

$=P(Smoke=t|fire=t)\ P(heat=t|fire=t)P(fire=t)$

$+P(Smoke=t|fire=t)P(heat=f|fire=t)P(fire=t)$

$+P(Smoke=t|fire=f)\ P(heat=t|fire=f)P(fire=f)$

$+P(Smoke=t|fire=f)P(heat=f|fire=f)P(fire=f)$

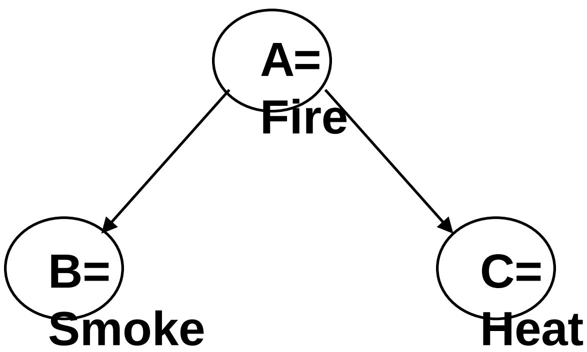
$= (.90x.99x.1)+(.90x.01x.1)$

$+(.001x.0001x.9)+(.001x.9999x.9)$

≈ 0.0909

Extended example of 3-way Bayesian Networks

P(Fire)
0. 1



Fire	P(Smoke)
t	.90
f	.001

Fire	P(Heat)
t	.99
f	.0001

What is P(Fire=t | Smoke=t)?

P(Fire=t | Smoke=t)

=P(Fire=t & Smoke=t) / P(Smoke=t)

≈ 0.09 / 0.0909

≈ 0.99

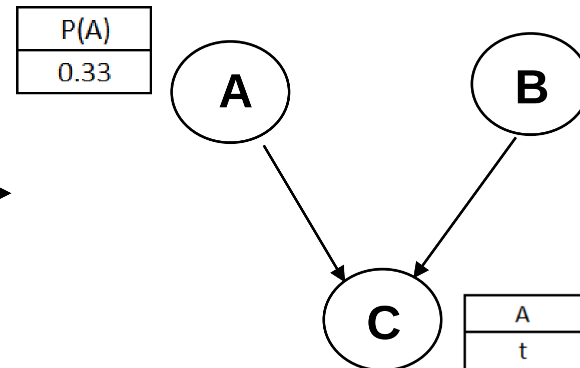
<u>So we've just proven that</u>

"Where there's smoke, there's (probably) fire."

Bayesian Network

- A Bayesian network specifies a joint distribution in a structured form:

$$p(A,B,C) = p(C|A,B)p(A)p(B)$$



A	B	P(C)
t	t	0.2
t	f	0.4
f	t	0.3
f	f	0.3

- Dependence/independence represented via a directed graph:
  - Node = random variable
  - Directed Edge = conditional dependence
  - Absence of Edge = conditional independence

- Allows concise view of joint distribution relationships:
  - Graph nodes and edges show conditional relationships between variables.
  - Tables provide probability data.

Bayesian Networks

- Structure of the graph $\Leftrightarrow$ Conditional independence relations

In general,

$$p(X_1, X_2, \ldots X_N) = \prod p(X_i \mid \text{parents}(X_i))$$

The full joint distribution

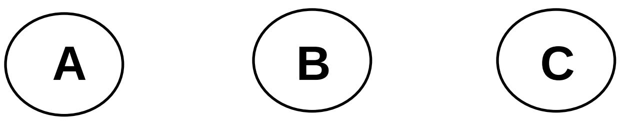
The graph-structured approximation

- Requires that graph is acyclic (no directed cycles)

- 2 components to a Bayesian network
  - The graph structure (conditional independence assumptions)
  - The numerical probabilities (for each variable given its parents)

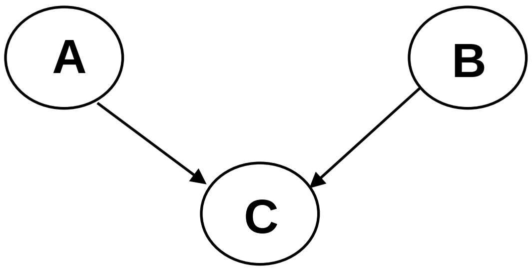
- Also known as belief networks, graphical models, causal networks

Examples of 3-way Bayesian Networks



Marginal Independence:
p(A,B,C) = p(A) p(B) p(C)

Examples of 3-way Bayesian Networks

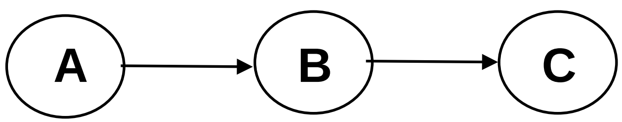


Independent Causes:
$p(A,B,C) = p(C|A,B)p(A)p(B)$

"Explaining away" effect:
Given C, observing A makes B less likely
e.g., earthquake/burglary/alarm example

A and B are (marginally) independent
but become dependent once C is known

Examples of 3-way Bayesian Networks



Markov dependence:
$p(A,B,C) = p(C|B)\, p(B|A) p(A)$

Burglar Alarm Example

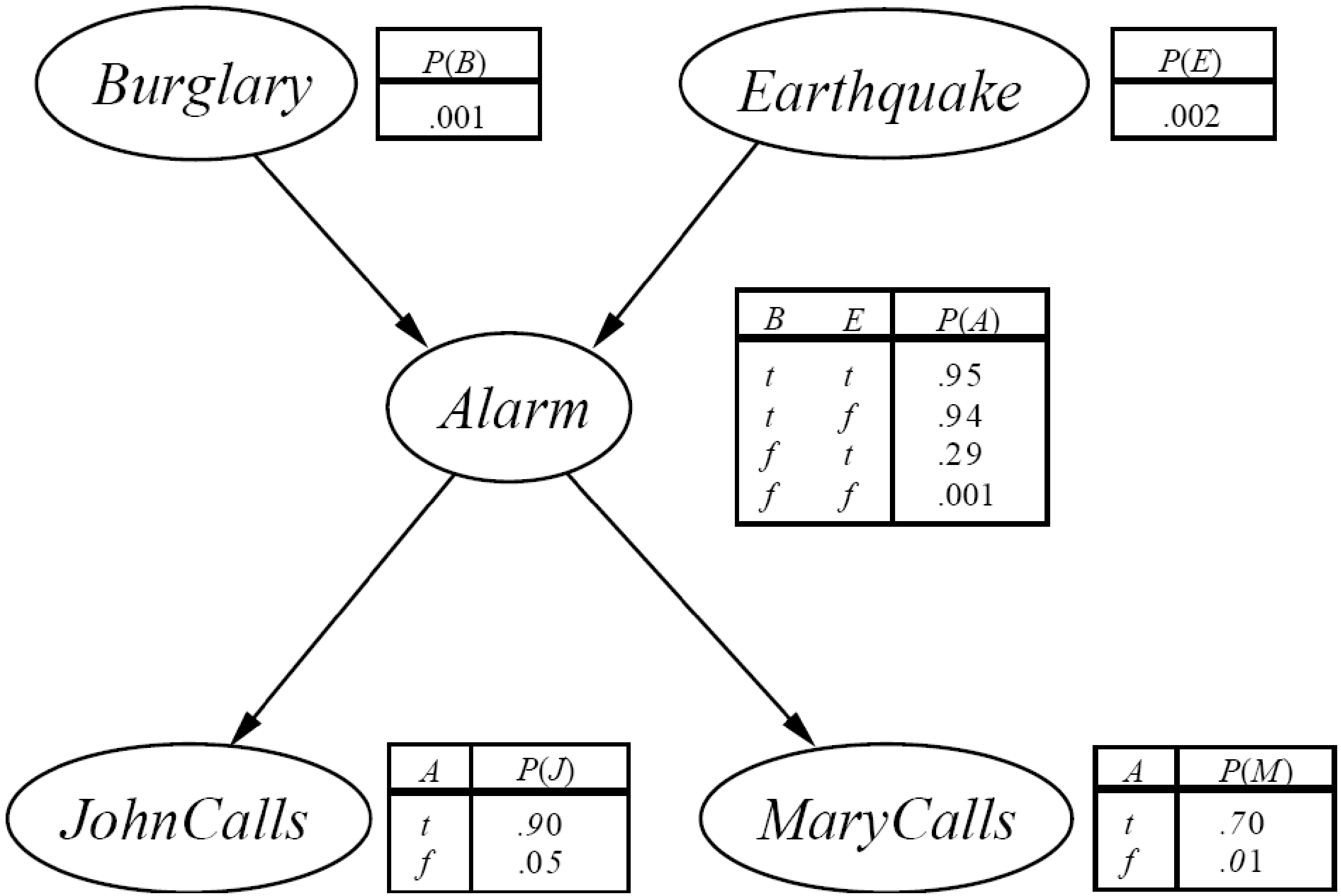
- Consider the following 5 binary variables:
  - B = a burglary occurs at your house
  - E = an earthquake occurs at your house
  - A = the alarm goes off
  - J = John calls to report the alarm
  - M = Mary calls to report the alarm

  - What is $P(B \mid M, J)$? (for example)

  - We can use the full joint distribution to answer this question
    - Requires $2^5 = 32$ probabilities

    - Can we use prior domain knowledge to come up with a Bayesian network that requires fewer probabilities?

The Desired Bayesian Network



Only requires 10 probabilities!

Constructing a Bayesian Network: Step 1

- Order the variables in terms of influence (may be a partial order)

  e.g., {E, B} -> {A} -> {J, M}

- $P(J, M, A, E, B) = P(J, M \mid A, E, B)\ P(A \mid E, B)\ P(E, B)$

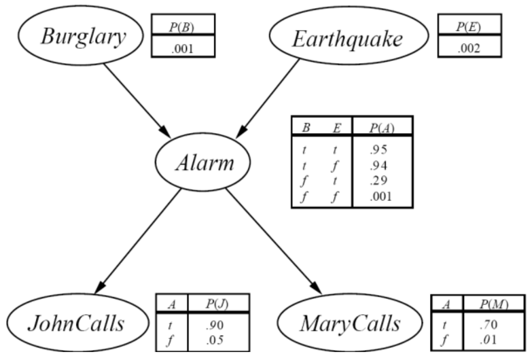
  $\approx P(J, M \mid A) \quad P(A \mid E, B)\ P(E)\ P(B)$

  $\approx P(J \mid A)\ P(M \mid A)\ P(A \mid E, B)\ P(E)\ P(B)$

These conditional independence assumptions are reflected in the graph structure of the Bayesian network

Constructing this Bayesian Network: Step 2

- P(J, M, A, E, B) =
  P(J | A) P(M | A) P(A | E, B) P(E) P(B)

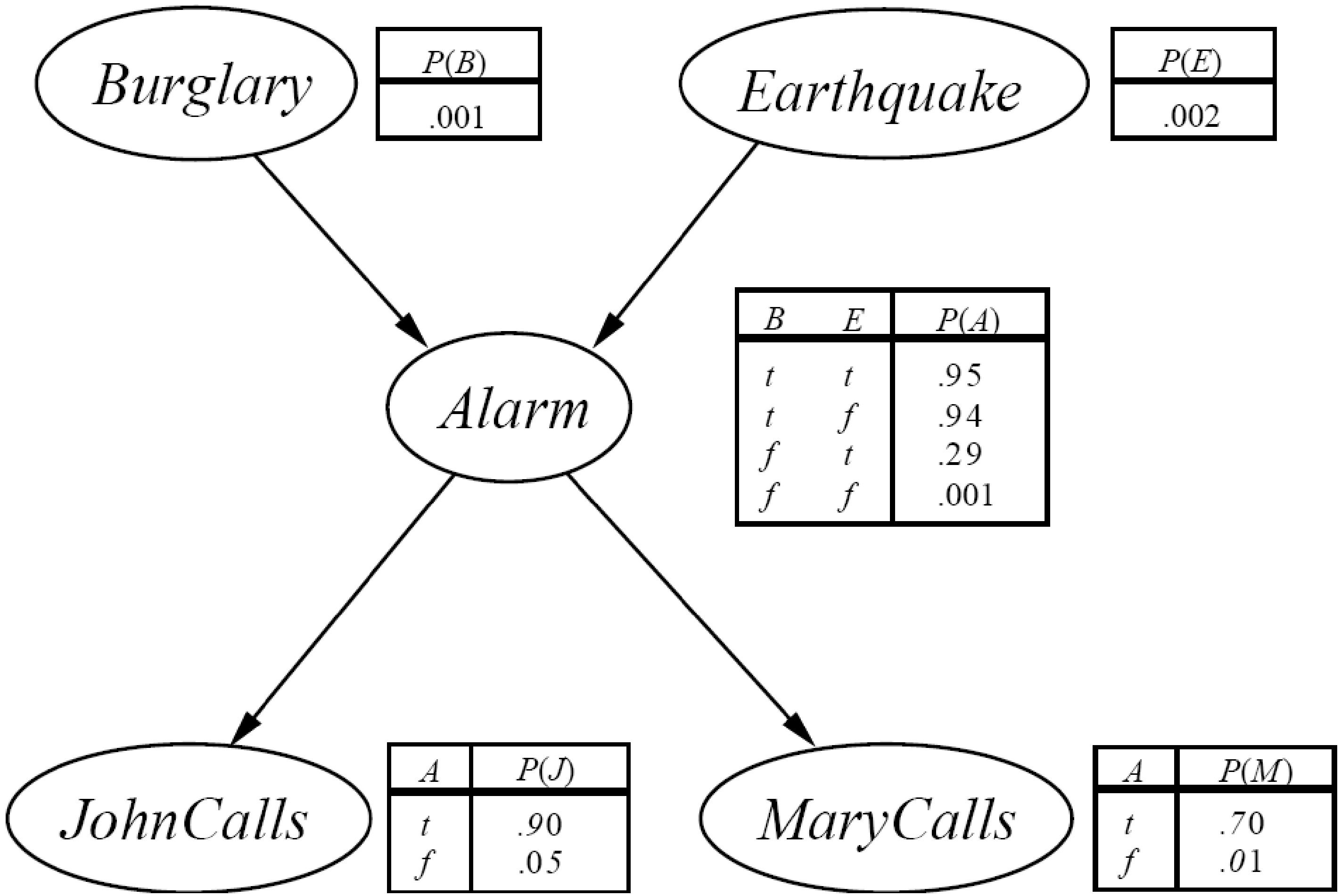


- There are 3 conditional probability tables (CPDs) to be determined:
  P(J | A), P(M | A), P(A | E, B)
  - Requiring 2 + 2 + 4 = 8 probabilities

- And 2 marginal probabilities P(E), P(B) -> 2 more probabilities

- Where do these probabilities come from?
  - Expert knowledge
  - From data (relative frequency estimates)
  - Or a combination of both - see discussion in Section 20.1 and 20.2 (optional)

The Resulting Bayesian Network



$P(B)$
.001

$P(E)$
.002

B	E	$P(A)$
t	t	.95
t	f	.94
f	t	.29
f	f	.001

A	$P(J)$
t	.90
f	.05

A	$P(M)$
t	.70
f	.01

Example of Answering a Probability Query

- So, what is $P(B \mid M, J)$?
  E.g., say, $P(b \mid m, \neg j)$, i.e., $P(B=true \mid M=true \wedge J=false)$

$P(b \mid m, \neg j) = P(b, m, \neg j) / P(m, \neg j)$;by definition

$P(b, m, \neg j) = \Sigma A \in \{a, \neg a\} \Sigma E \in \{e, \neg e\}\ P(\neg j, m, A, E, b)$;marginal

$P(J, M, A, E, B) \approx P(J \mid A)\ P(M \mid A)\ P(A \mid E, B)\ P(E)\ P(B)$; conditional indep.
$P(\neg j, m, A, E, b) \approx P(\neg j \mid A)\ P(m \mid A)\ P(A \mid E, b)\ P(E)\ P(b)$

Say, work the case $A=a \wedge E=\neg e$
$P(\neg j, m, a, \neg e, b) \approx P(\neg j \mid a)\ P(m \mid a)\ P(a \mid \neg e, b)\ P(\neg e)\ P(b)$
$\approx 0.10 \times 0.70 \times 0.94 \times 0.998 \times 0.001$
Similar for the cases of $a \wedge e$, $\neg a \wedge e$, $\neg a \wedge \neg e$.

Similar for $P(m, \neg j)$. Then just divide to get $P(b \mid m, \neg j)$.

Inference in Bayesian Networks

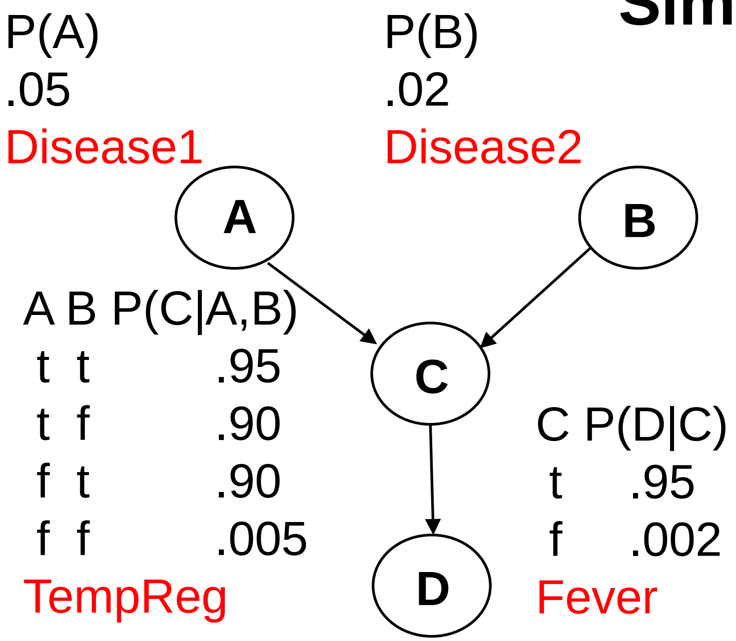
- **X** = *{ X1, X2, …, Xk }* = **query variables** of interest
- **E** = *{ E1, …, El }* = **evidence variables** that are observed
  - (**e**, an **event**)
- **Y** = *{ Y1, …, Ym }* = **hidden variables** (nonevidence, nonquery)

- **What is the posterior distribution of X, given E?**
- $\mathbf{P}(\mathbf{X} \mid \mathbf{e}) = \alpha \, \Sigma_{\mathbf{y}} \, \mathbf{P}(\mathbf{X}, \mathbf{y}, \mathbf{e})$

- **What is the most likely assignment of values to X, given E?**
- $\mathbf{argmax}_{\mathbf{x}} \, P(\mathbf{x} \mid \mathbf{e}) = \mathbf{argmax}_{\mathbf{x}} \, \Sigma_{\mathbf{y}} \, P(\mathbf{x}, \mathbf{y}, \mathbf{e})$

Inference in Bayesian Networks

Simple Example



Query Variables A, B

Hidden Variable C

Evidence Variable D

Note: Not an anatomically correct model of how diseases cause fever!

Suppose that two different diseases influence some imaginary internal body temperature regulator, which in turn influences whether fever is present.

Inference in Bayesian Networks

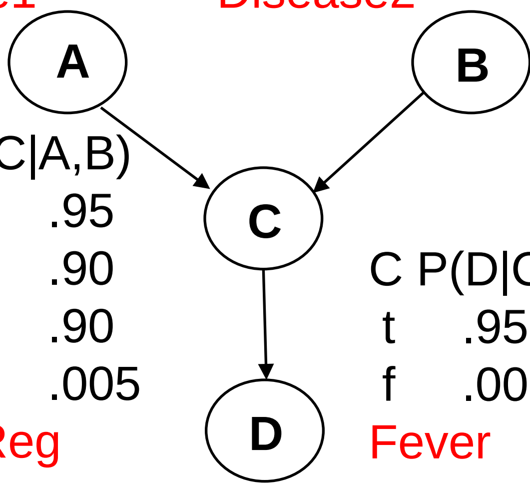
Simple Example

P(A)
.05

Disease1

P(B)
.02

Disease2



A	B	P(C|A,B)
t	t	.95
t	f	.90
f	t	.90
f	f	.005

TempReg

C	P(D|C)
t	.95
f	.002

Fever

What is the posterior conditional distribution of our query variables, given that fever was observed?

$P(A,B|d) = \alpha\ \Sigma_c\ P(A,B,c,d)$
$= \alpha\ \Sigma_c\ P(A)P(B)P(c|A,B)P(d|c)$
$= \alpha\ P(A)P(B)\ \Sigma_c\ P(c|A,B)P(d|c)$

$P(a,b|d) = \alpha\ P(a)P(b)\ \Sigma_c\ P(c|a,b)P(d|c) = \alpha\ P(a)P(b)\{\ P(c|a,b)P(d|c)+P(\neg c|a,b)P(d|\neg c)\ \}$
$= \alpha\ .05x.02x\{.95x.95+.05x.002\} \approx \alpha\ .000903 \approx .014$

$P(\neg a,b|d) = \alpha\ P(\neg a)P(b)\ \Sigma_c\ P(c|\neg a,b)P(d|c) = \alpha\ P(\neg a)P(b)\{\ P(c|\neg a,b)P(d|c)+P(\neg c|\neg a,b)P(d|\neg c)\ \}$
$= \alpha\ .95x.02x\{.90x.95+.10x.002\} \approx \alpha\ .0162 \approx .248$

$P(a,\neg b|d) = \alpha\ P(a)P(\neg b)\ \Sigma_c\ P(c|a,\neg b)P(d|c) = \alpha\ P(a)P(\neg b)\{\ P(c|a,\neg b)P(d|c)+P(\neg c|a,\neg b)P(d|\neg c)\ \}$
$= \alpha\ .05x.98x\{.90x.95+.10x.002\} \approx \alpha\ .0419 \approx .642$

$P(\neg a,\neg b|d) = \alpha\ P(\neg a)P(\neg b)\ \Sigma_c\ P(c|\neg a,\neg b)P(d|c) = \alpha\ P(\neg a)P(\neg b)\{\ P(c|\neg a,\neg b)P(d|c)+P(\neg c|\neg a,\neg b)P(d|\neg c)\ \}$
$= \alpha\ .95x.98x\{.005x.95+.995x.002\} \approx \alpha\ .00627 \approx .096$

$\alpha \approx 1\ /\ (.000903+.0162+.0419+.00627) \approx 1\ /\ .06527 \approx 15.32$

Inference in Bayesian Networks

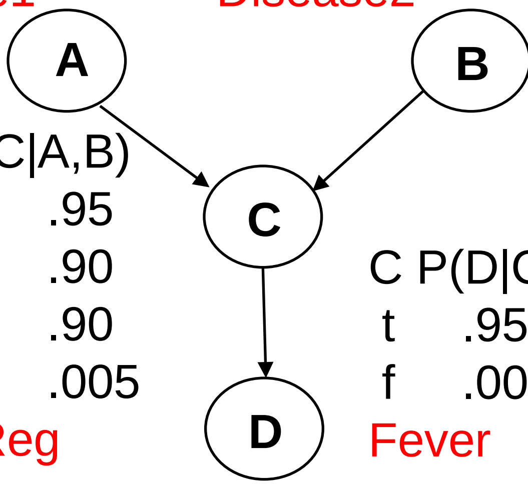
Simple Example

P(A)
.05

Disease1

P(B)
.02

Disease2



A	B	P(C|A,B)
t	t	.95
t	f	.90
f	t	.90
f	f	.005

TempReg

C	P(D|C)
t	.95
f	.002

Fever

What is the most likely posterior conditional assignment of values to our query variables, given that fever was observed?

$\text{argmax}_{\{a,b\}}\ P(a, b \mid d)$
$= \text{argmax}_{\{a,b\}}\ \Sigma_c\ P(a,b,c,d)$
$= \{ a, \neg b \}$

$P(a,b|d) = \alpha\, P(a)P(b)\ \Sigma_c\ P(c|a,b)P(d|c) = \alpha\, P(a)P(b)\{ P(c|a,b)P(d|c)+P(\neg c|a,b)P(d|\neg c) \}$
$= \alpha\ .05\text{x}.02\text{x}\{.95\text{x}.95+.05\text{x}.002\} \approx \alpha\ .000903 \approx .014$

$P(\neg a,b|d) = \alpha\, P(\neg a)P(b)\ \Sigma_c\ P(c|\neg a,b)P(d|c) = \alpha\, P(\neg a)P(b)\{ P(c|\neg a,b)P(d|c)+P(\neg c|\neg a,b)P(d|\neg c) \}$
$= \alpha\ .95\text{x}.02\text{x}\{.90\text{x}.95+.10\text{x}.002\} \approx \alpha\ .0162 \approx .248$

$\boxed{P(a,\neg b|d) = \alpha\, P(a)P(\neg b)\ \Sigma_c\ P(c|a,\neg b)P(d|c) = \alpha\, P(a)P(\neg b)\{ P(c|a,\neg b)P(d|c)+P(\neg c|a,\neg b)P(d|\neg c) \} = \alpha\ .05\text{x}.98\text{x}\{.90\text{x}.95+.10\text{x}.002\} \approx \alpha\ .0419 \approx .642}$

$P(\neg a,\neg b|d) = \alpha\, P(\neg a)P(\neg b)\ \Sigma_c\ P(c|\neg a,\neg b)P(d|c) = \alpha\, P(\neg a)P(\neg b)\{ P(c|\neg a,\neg b)P(d|c)+P(\neg c|\neg a,\neg b)P(d|\neg c) \}$
$= \alpha\ .95\text{x}.98\text{x}\{.005\text{x}.95+.995\text{x}.002\} \approx \alpha\ .00627 \approx .096$

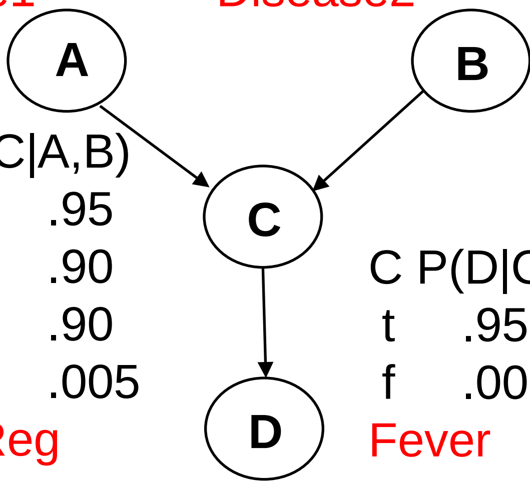
$\alpha \approx 1 / (.000903+.0162+.0419+.00627) \approx 1 / .06527 \approx 15.32$

Inference in Bayesian Networks

Simple Example

P(A)
.05
Disease1

P(B)
.02
Disease2



A	B	P(C|A,B)
t	t	.95
t	f	.90
f	t	.90
f	f	.005

TempReg

C	P(D|C)
t	.95
f	.002

Fever

What is the posterior conditional distribution of A, given that fever was observed? (I.e., temporarily make B into a hidden variable.)

We can use P(A,B|d) from above.

$P(A|d) = \alpha \; \Sigma_b \; P(A,b|d)$

$P(a|d) = \Sigma_b P(a,b|d) = P(a,b|d)+P(a,\neg b|d)$
$= (.014+.642) \approx .656$

$P(\neg a|d) = \Sigma_b P(\neg a,b|d) = P(\neg a,b|d)+P(\neg a,\neg b|d)$
$= (.248+.096) \approx .344$

This is a marginalization, so we expect from theory that α = 1; but check for round-off error.

A	B	P(A,B|d) from above
t	t	≈ .014
f	t	≈ .248
t	f	≈ .642
f	f	≈ .096

Inference in Bayesian Networks

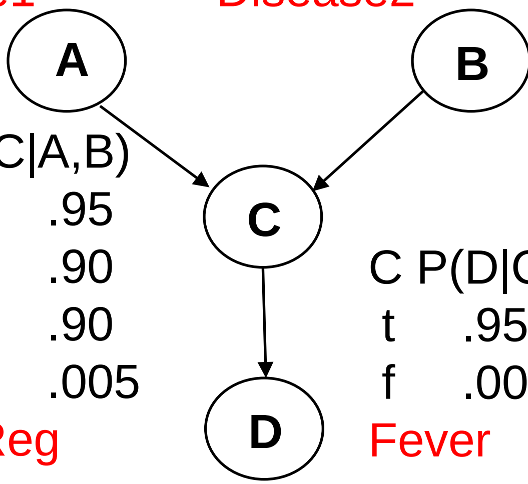
Simple Example

P(A)
.05

Disease1

P(B)
.02

Disease2



A	B	P(C|A,B)
t	t	.95
t	f	.90
f	t	.90
f	f	.005

TempReg

C	P(D|C)
t	.95
f	.002

Fever

What is the posterior conditional distribution of A, given that fever was observed, <u>and that further lab tests definitely rule out Disease2?</u> (I.e., temporarily make B into an evidence variable with B = false.)

$P(A|\neg b,d) = \alpha\ P(A,\neg b|d)$

$P(a|\neg b,d) = \alpha\ P(a,\neg b|d)$

$\approx \alpha\ .642 \approx .870$

$P(\neg a|\neg b,d) = \alpha\ P(\neg a,\neg b|d)$

$\approx \alpha\ .096 \approx .130$

$\alpha \approx 1\ /\ (.642+.096) \approx 1\ /\ .738 \approx 1.355$

A	B	P(A,B|d) from above
t	t	≈ .014
f	t	≈ .248
t	f	≈ .642
f	f	≈ .096

General Strategy for inference

- Want to compute P(q | e)

Step 1:

$P(q \mid e) = P(q,e)/P(e) = \alpha\, P(q,e)$, since P(e) is constant wrt Q

Step 2:

$P(q,e) = \Sigma_{a..z}\, P(q, e, a, b, \ldots.\, z)$, by the law of total probability

Step 3:

$\Sigma_{a..z}\, P(q, e, a, b, \ldots.\, z) = \Sigma_{a..z}\, \Pi_i\, P(\text{variable } i \mid \text{parents } i)$

(using Bayesian network factoring)

Step 4:

Distribute summations across product terms for efficient computation

Section 14.4 discusses exact inference in Bayesian Networks. The complexity depends strongly on the network structure. The general case is intractable, but there are things you can do. Section 14.5 discusses approximation by sampling.

Number of Probabilities in Bayesian Networks

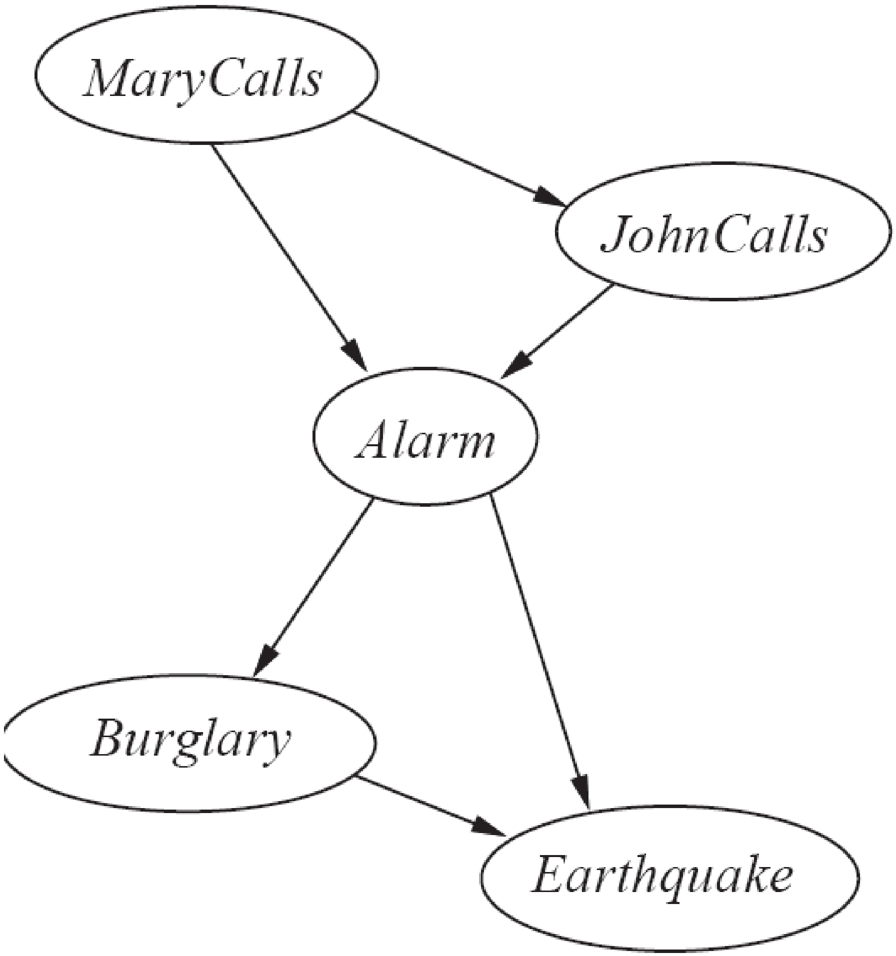
- Consider n binary variables

- Unconstrained joint distribution requires $O(2^n)$ probabilities

- If we have a Bayesian network, with a maximum of k parents for any node, then we need $O(n\ 2^k)$ probabilities

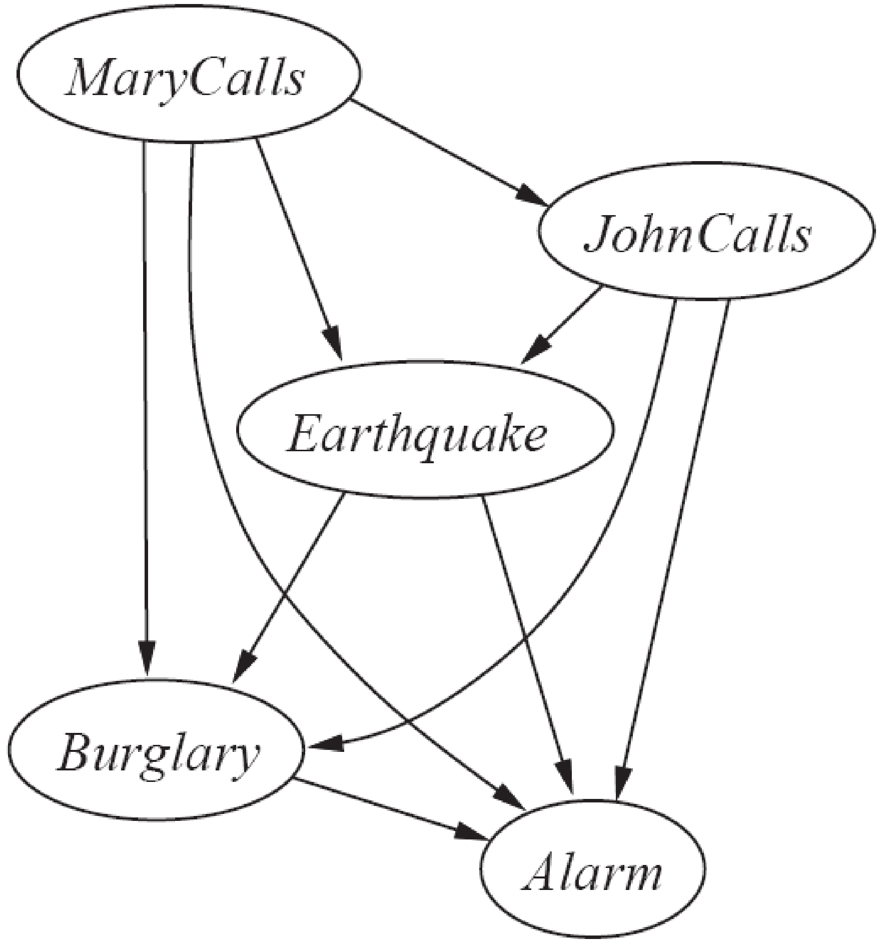
- Example
  - Full unconstrained joint distribution
    - $n = 30$, $k = 4$: need 10^9 probabilities for full joint distribution
  - Bayesian network
    - $n = 30$, $k = 4$: need 480 probabilities

The Bayesian Network from a different Variable Ordering



(a)

The Bayesian Network from a different Variable Ordering



(b)

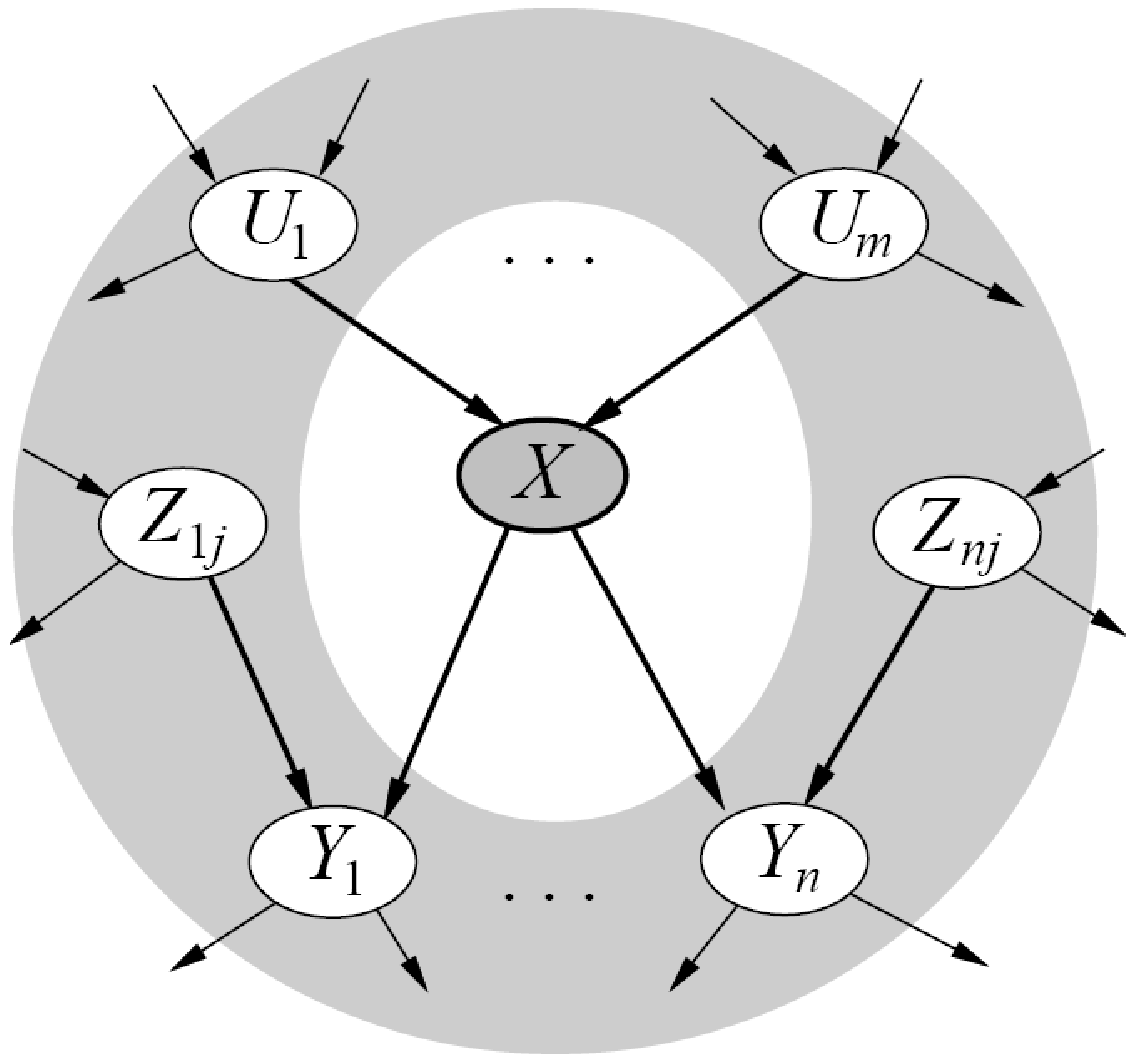
Given a graph, can we "read off" conditional independencies?

The "Markov Blanket" of X (the gray area in the figure)

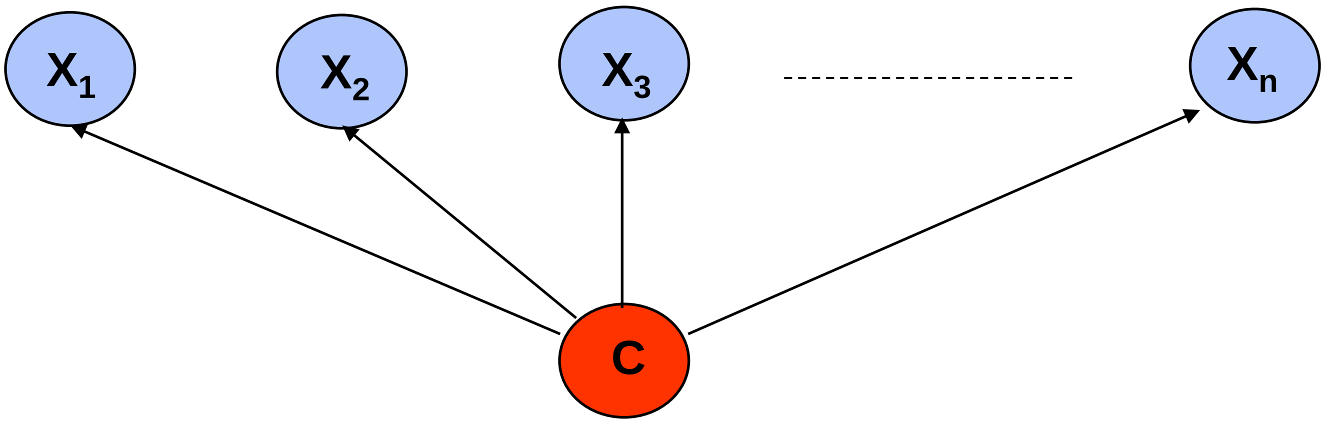
X is conditionally independent of everything else, GIVEN the values of:

* X's parents
* X's children
* X's children's parents

X is conditionally independent of its non-descendants, GIVEN the values of its parents.



Naïve Bayes Model



$$P(C \mid X_1,\ldots,X_n) = \alpha \prod P(X_i \mid C)\ P(C)$$

Features X are conditionally independent given the class variable C

Widely used in machine learning

e.g., spam email classification: X's = counts of words in emails

Probabilities P(C) and P(Xi | C) can easily be estimated from labeled data

Naïve Bayes Model (2)

$$P(C \mid X_1, \ldots X_n) = \alpha \prod P(X_i \mid C) \; P(C)$$

Probabilities P(C) and P(Xi | C) can easily be estimated from labeled data

P(C = cj) ≈ #(Examples with class label cj) / #(Examples)

P(Xi = xik | C = cj)
≈ #(Examples with Xi value xik and class label cj)
/ #(Examples with class label cj)

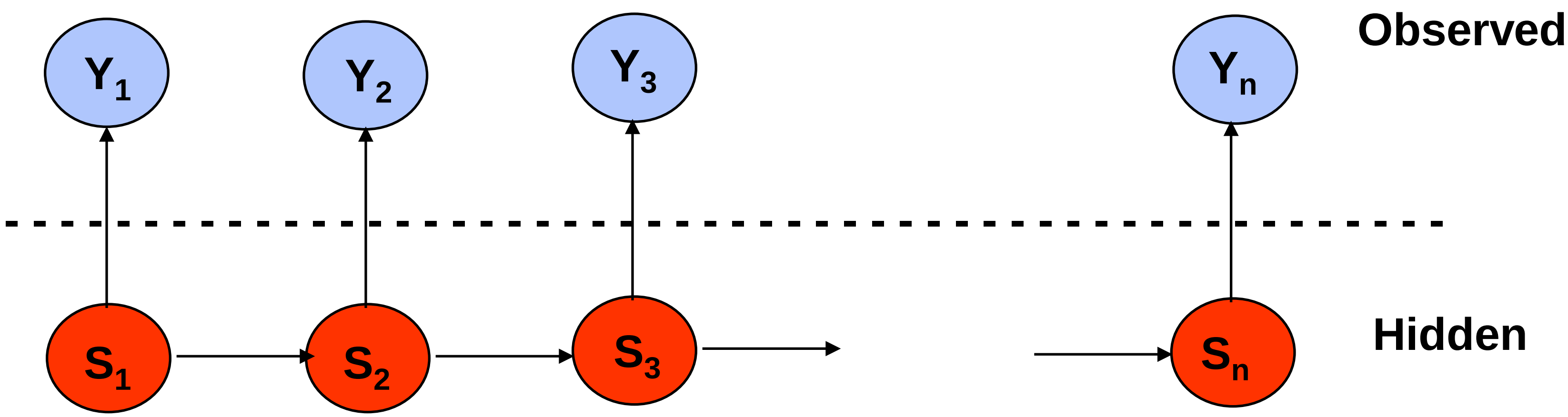
Usually easiest to work with logs

$$\log [\, P(C \mid X_1, \ldots X_n) \,] = \log \alpha + \sum [\, \log P(X_i \mid C) + \log P(C) \,]$$

DANGER: Suppose ZERO examples with Xi value xik and class label cj ?
An unseen example with Xi value xik will NEVER predict class label cj !

Practical solutions: Pseudocounts, e.g., add 1 to every #() , etc.
Theoretical solutions: Bayesian inference, beta distribution, etc.

Hidden Markov Model (HMM)



Two key assumptions:

1. hidden state sequence is Markov
2. observation Y_t is CI of all other variables given S_t

Widely used in speech recognition, protein sequence models

Since this is a Bayesian network polytree, inference is linear in n

Summary

- Bayesian networks represent a joint distribution using a graph
- The graph encodes a set of conditional independence assumptions
- Answering queries (or inference or reasoning) in a Bayesian network amounts to efficient computation of appropriate conditional probabilities
- Probabilistic inference is intractable in the general case
  - But can be carried out in linear time for certain classes of Bayesian networks