COL333/671: Introduction to AI

Semester I, 2022-23

Probabilistic Reasoning

Rohan Paul

Outline

- Last Class
  - Adversarial Search
- This Class
  - Probabilistic Reasoning
- Reference Material
  - AIMA Ch. 13 and 14

Acknowledgement

These slides are intended for teaching purposes only. Some material has been used/adapted from web sources and from slides by Doina Precup, Dorsa Sadigh, Percy Liang, Mausam, Dan Klein, Anca Dragan, Nicholas Roy and others.

Uncertainty in AI

- Uncertainty:
  - **Observed variables (evidence)**: Agent knows certain things about the state of the world (e.g., sensor measurements or symptoms)
  - **Unobserved variables**: Agent needs to reason about other aspects (e.g. what disease is present, is the car operational, location of the burglar)
  - **Model**: Agent knows something about how the known variables relate to the unknown variables
- Probabilistic reasoning gives us a framework for managing our beliefs and knowledge.

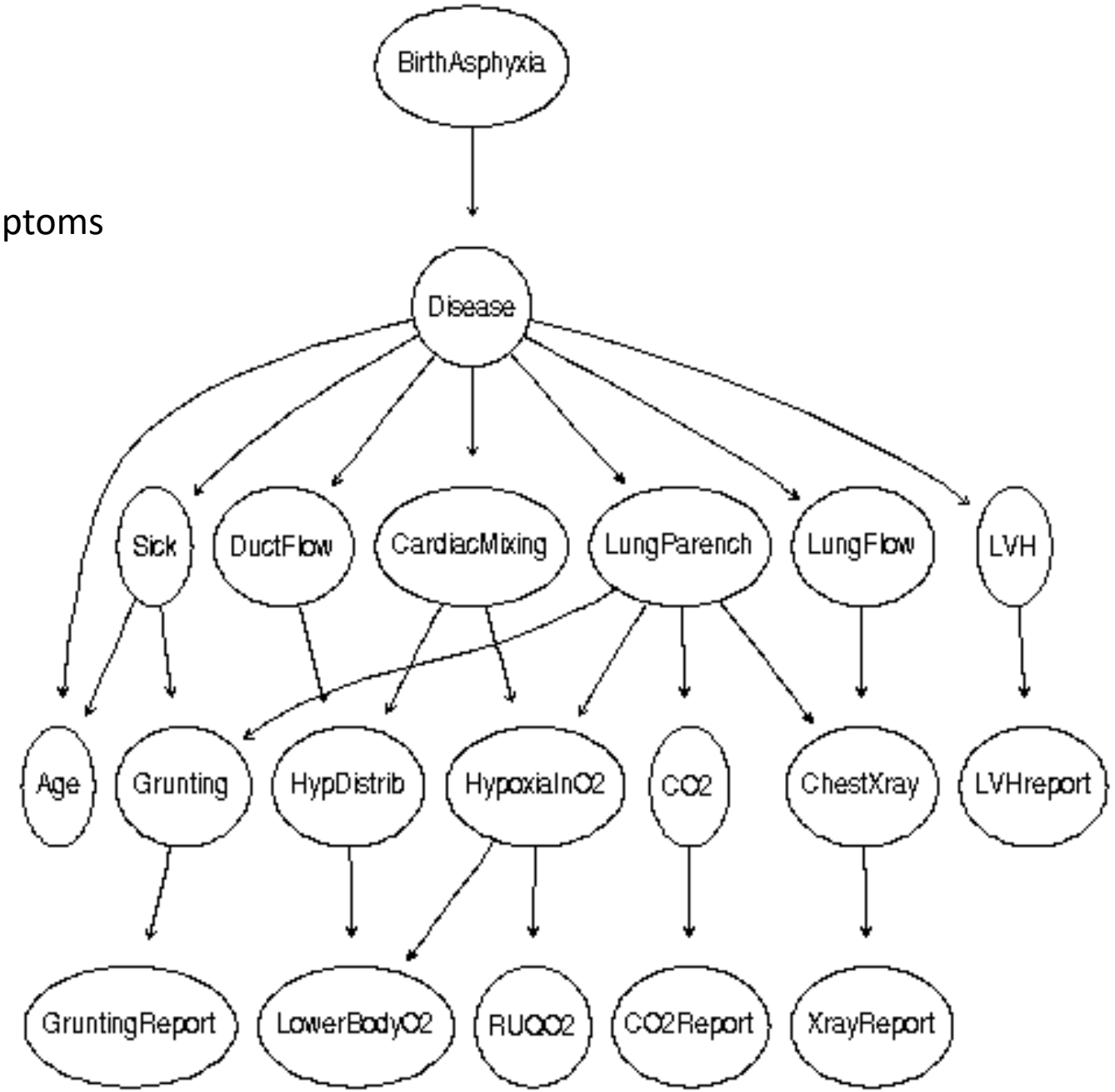
I hear an unusual sound and a burning smell in my car, what fault is there in my engine?

I have fever, loss of smell, loss of taste, do I have Covid?

I hear some footsteps in my house, where is the burglar?

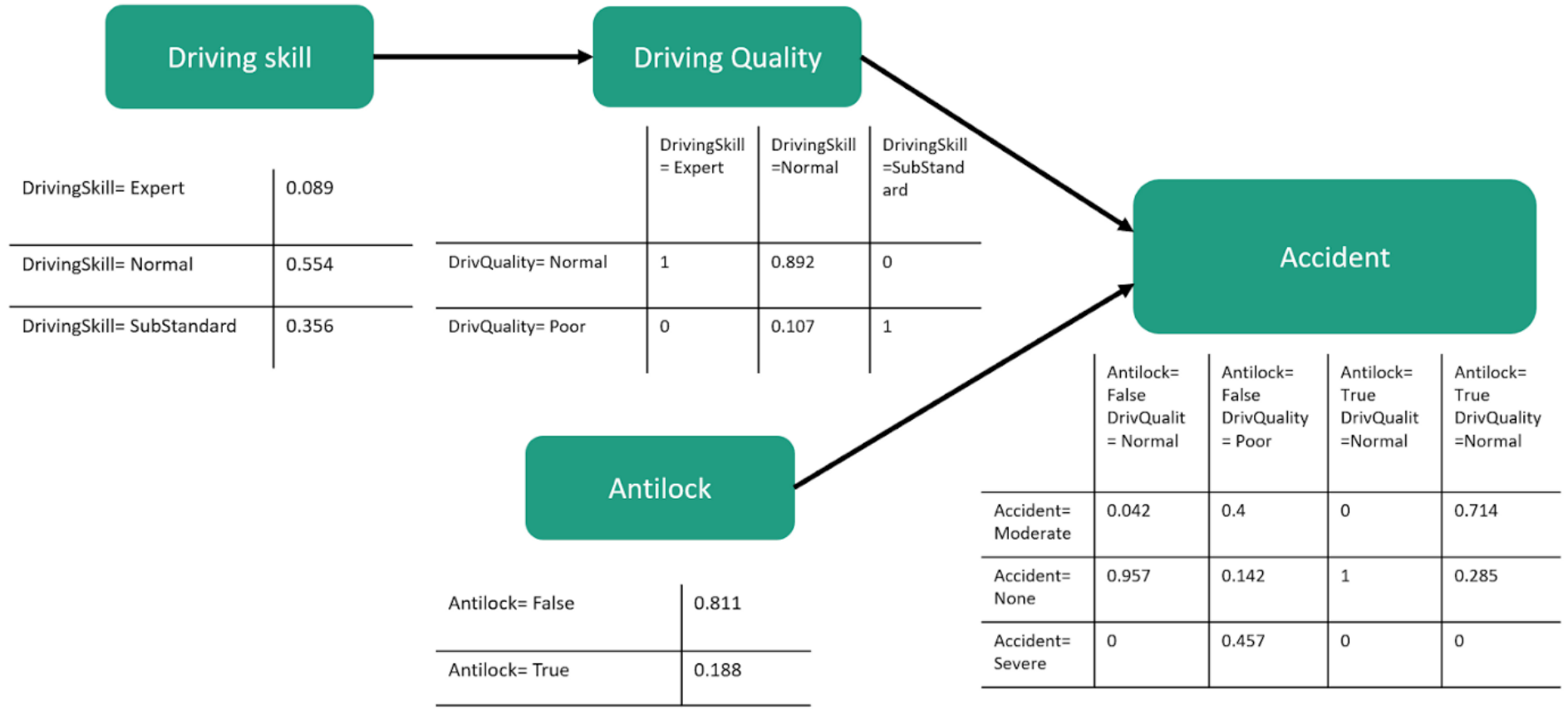
Examples

Inferring disease from symptoms



Examples

Accident – driving domain.



DrivingSkill= Expert	0.089
DrivingSkill= Normal	0.554
DrivingSkill= SubStandard	0.356

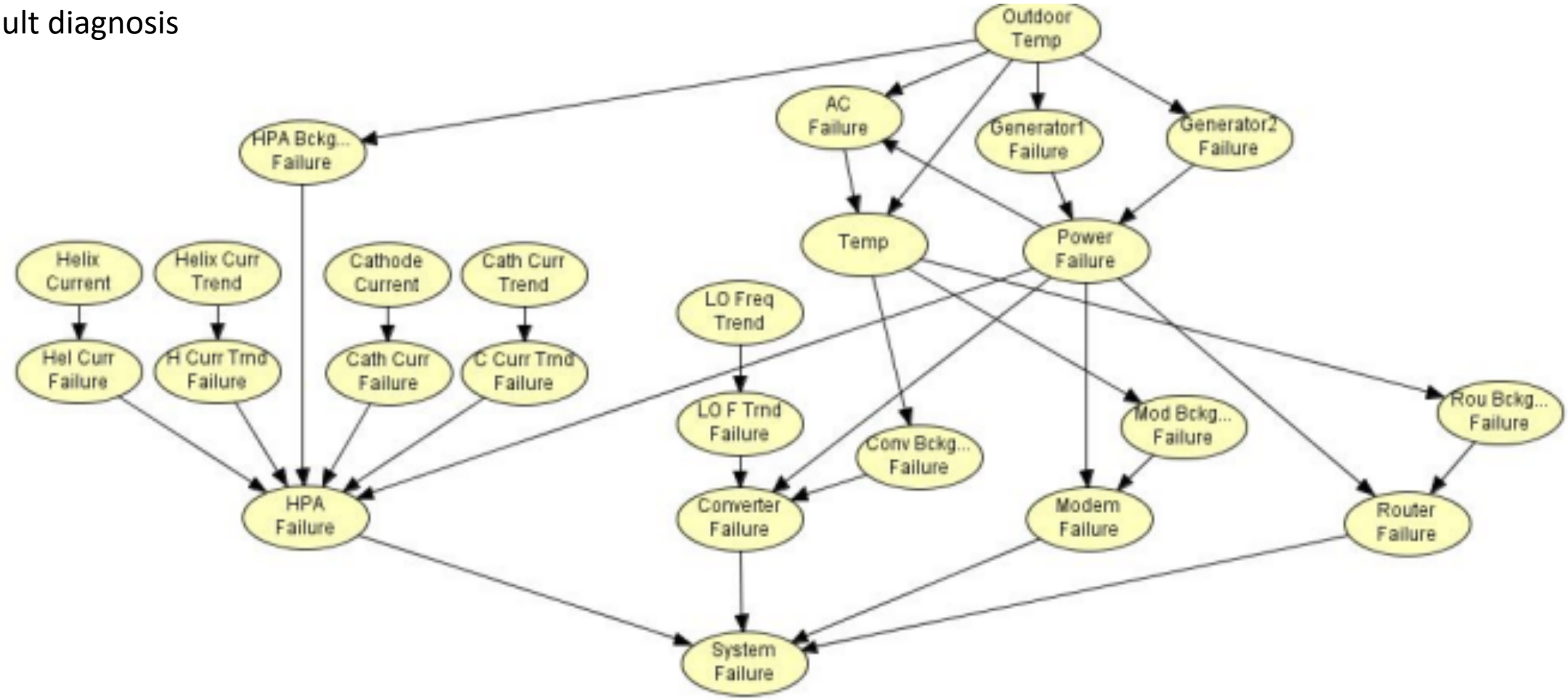
	DrivingSkill = Expert	DrivingSkill =Normal	DrivingSkill =SubStandard
DrivQuality= Normal	1	0.892	0
DrivQuality= Poor	0	0.107	1

Antilock= False	0.811
Antilock= True	0.188

	Antilock= False DrivQualit = Normal	Antilock= False DrivQuality = Poor	Antilock= True DrivQualit =Normal	Antilock= True DrivQuality =Normal
Accident= Moderate	0.042	0.4	0	0.714
Accident= None	0.957	0.142	1	0.285
Accident= Severe	0	0.457	0	0

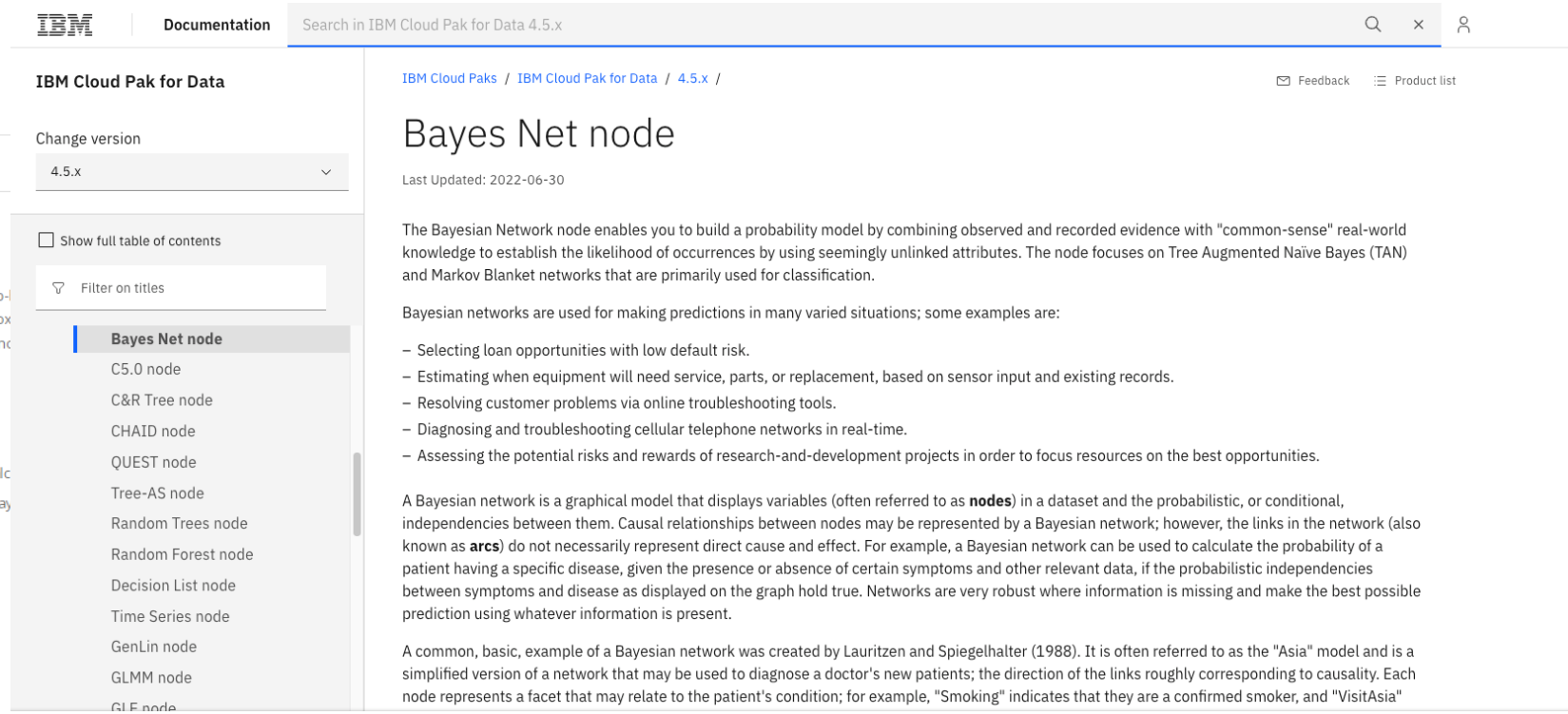
Examples

Fault diagnosis



Examples

Predictive analytics/expert systems



Outline

- Representation for Uncertainty (review)
- Bayes Nets:
  - Probabilistic reasoning gives us a framework for managing our beliefs and knowledge.
- Answering queries using Bayes Net
  - Inference methods
- Approximate methods for answering queries
- Use of learning

Random Variables

- A random variable is some aspect of the world about which we (may) have uncertainty
  - R = Do I have Covid?
  - T = Engine is faulty or working?
  - D = How long will it take to drive to IIT?
  - L = Where is the person?

- Domains
  - R in {true, false} (often write as {+r, -r})
  - T in {faulty, working}
  - D in [0, ∞)
  - L in possible locations in a grid {(0,0), (0,1), …}

I hear an unusual sound and a burning smell in my car, what fault is there in my engine?

I have fever, loss of smell, loss of taste, do I have Covid?

I hear some footsteps in my house, where is the burglar?

Joint Distributions

- A **joint distribution** over a set of random variables: $X_1, X_2, \ldots X_n$ specifies a real number for each assignment (or *outcome*):

$$P(X_1 = x_1, X_2 = x_2, \ldots X_n = x_n)$$

$$P(x_1, x_2, \ldots x_n)$$

  - Must obey:

$$P(x_1, x_2, \ldots x_n) \geq 0$$

$$\sum_{(x_1, x_2, \ldots x_n)} P(x_1, x_2, \ldots x_n) = 1$$

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

Note: Joint distribution can answer all probabilistic queries.

Problem: Table size is d^n.

Events

- An event is a set E of outcomes

$$P(E) = \sum_{(x_1 \ldots x_n) \in E} P(x_1 \ldots x_n)$$

- From a joint distribution, we can calculate the probability of any event
  - Probability that it's hot AND sunny? .4
  - Probability that it's hot? .4 + .1
  - Probability that it's hot OR sunny? .4 + .1 + .2

$$P(T, W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

Marginalization

- From a joint distribution (>1 variable) reduce it to a distribution over a smaller set of variables
- Called marginal distributions are sub-tables which eliminate variables
- Marginalization (summing out): Combine collapsed rows by adding likelihoods

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P(t) = \sum_s P(t, s)$$

$P(T)$

T	P
hot	0.5
cold	0.5

$$P(s) = \sum_t P(t, s)$$

$P(W)$

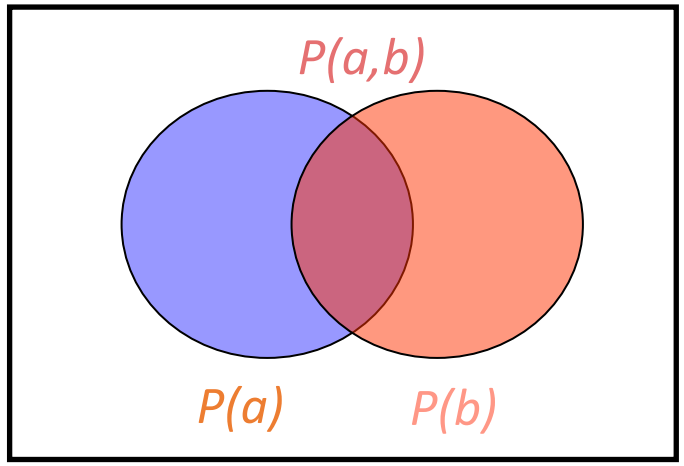
W	P
sun	0.6
rain	0.4

$$P(X_1 = x_1) = \sum_{x_2} P(X_1 = x_1, X_2 = x_2)$$

Conditioning

- Conditional distributions are probability distributions over some variables given fixed values of others

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



Joint Distribution

$$P(T,W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

Conditional Distributions

$$P(W|T)$$

$$P(W|T = cold)$$

W	P
sun	0.4
rain	0.6

$$P(W|T = hot)$$

W	P
sun	0.8
rain	0.2

Inference by Enumeration

- P(W)?

S	T	W	P
summer	hot	sun	0.30
summer	hot	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	hot	sun	0.10
winter	hot	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.20

Inference by Enumeration

- P(W)?

P(sun)=.3+.1+.1+.15=.65

S	T	W	P
summer	hot	sun	0.30
summer	**hot**	**rain**	**0.05**
summer	cold	sun	0.10
summer	**cold**	**rain**	**0.05**
winter	hot	sun	0.10
winter	**hot**	**rain**	**0.05**
winter	cold	sun	0.15
winter	**cold**	**rain**	**0.20**

Inference by Enumeration

- P(W)?

P(sun)=.3+.1+.1+.15=.65
P(rain)=1-.65=.35

S	T	W	P
summer	hot	sun	0.30
summer	hot	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	hot	sun	0.10
winter	hot	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.20

Inference by Enumeration

- P(W | winter, hot)?

S	T	W	P
summer	hot	sun	0.30
summer	hot	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	hot	sun	0.10
winter	hot	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.20

Inference by Enumeration

- P(W | winter, hot)?

P(sun|winter,hot)~.1
P(rain|winter,hot)~.05
P(sun|winter,hot)=2/3
P(rain|winter,hot)=1/3

S	T	W	P
summer	hot	sun	0.30
summer	hot	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	hot	sun	0.10
winter	hot	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.20

Product Rule

- Marginal and a conditional provides the joint distribution.

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

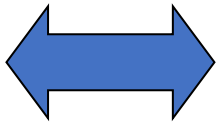
- Example:

$P(W)$

R	P
sun	0.8
rain	0.2

$P(D|W)$

D	W	P
wet	sun	0.1
dry	sun	0.9
wet	rain	0.7
dry	rain	0.3



$P(D,W)$

D	W	P
wet	sun	
dry	sun	
wet	rain	
dry	rain	

Chain Rule

Chain rule is derived by successive application of product rule:

$$
\begin{aligned}
& P(X_1, \ldots, X_n) = \\
= \quad & P(X_1, \ldots, X_{n-1}) P(X_n | X_1, \ldots, X_{n-1}) \\
= \quad & P(X_1, \ldots, X_{n-2}) P(X_{n-1} | X_1, \ldots, X_{n-2}) P(X_n | X_1, \ldots, X_{n-1}) \\
= \quad & \ldots \\
= \quad & \prod_{i=1}^{n} P(X_i | X_1, \ldots, X_{i-1})
\end{aligned}
$$

Constructing a larger distribution by simpler distribution.

Bayes Rule

- Two ways to factor a joint distribution over two variables:

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

- Dividing, we get:

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

- Example: Diagnostic probability from causal probability:

$$P(\text{cause}|\text{effect}) = \frac{P(\text{effect}|\text{cause})P(\text{cause})}{P(\text{effect})}$$

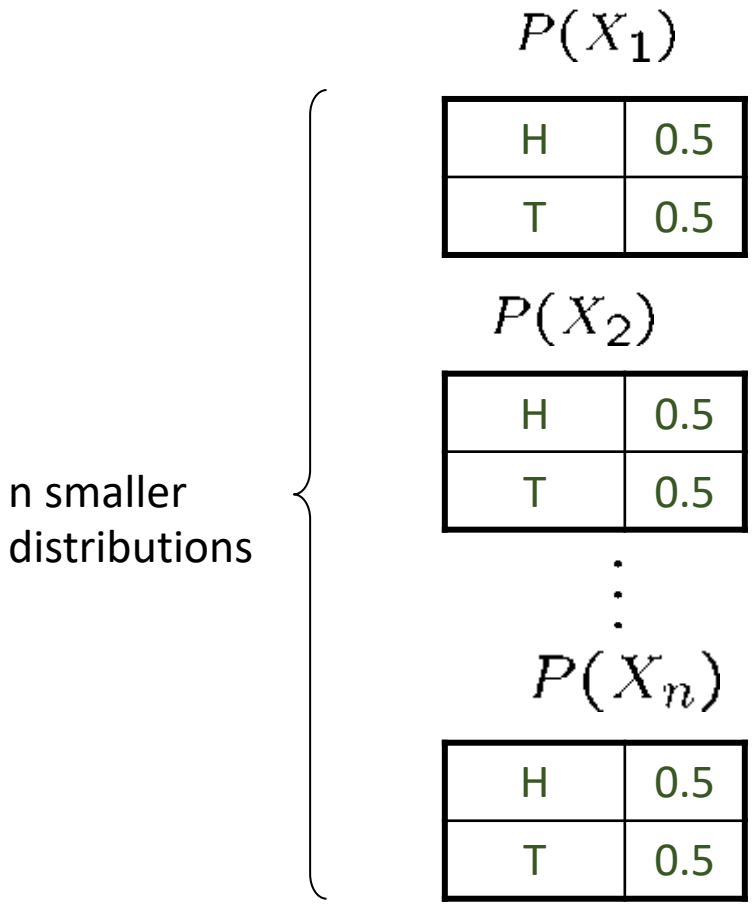
- Usefulness
  - Lets us build one conditional from its reverse.
  - Often one conditional is difficult to obtain but the other one is simple.

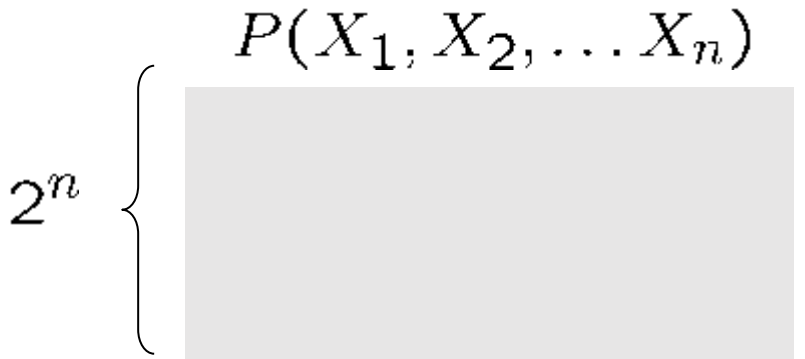
Independence

- Two variables are *independent* if:

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

  - This says that their joint distribution *factors* into a product two simpler distributions
  - Another form: $\forall x, y : P(x|y) = P(x)$
  - We write: $X \perp\!\!\!\perp Y$
  - Example
    - N-independent flips of a fair coin.



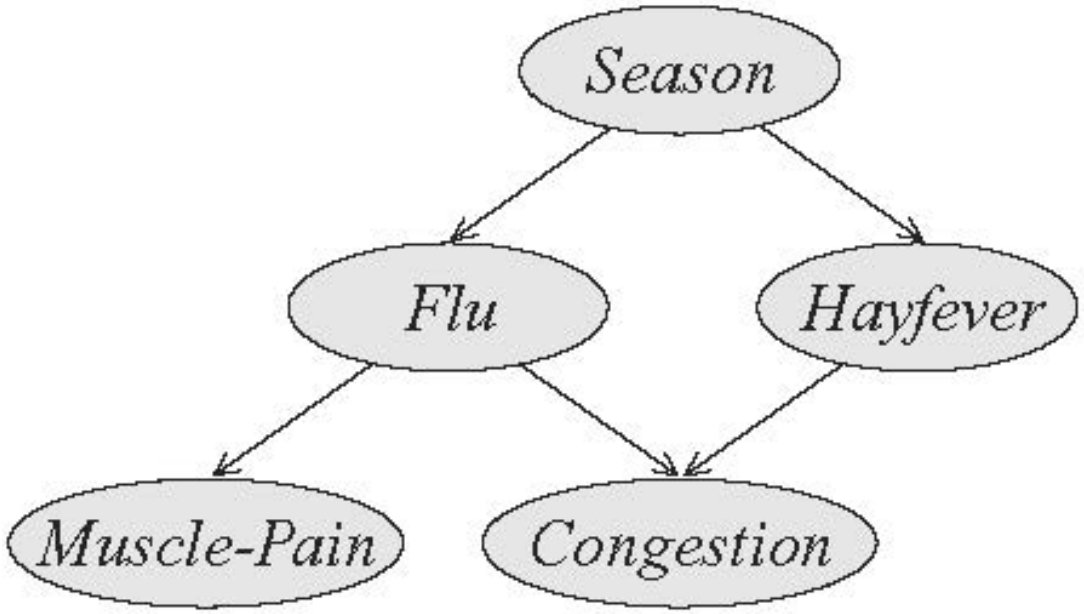


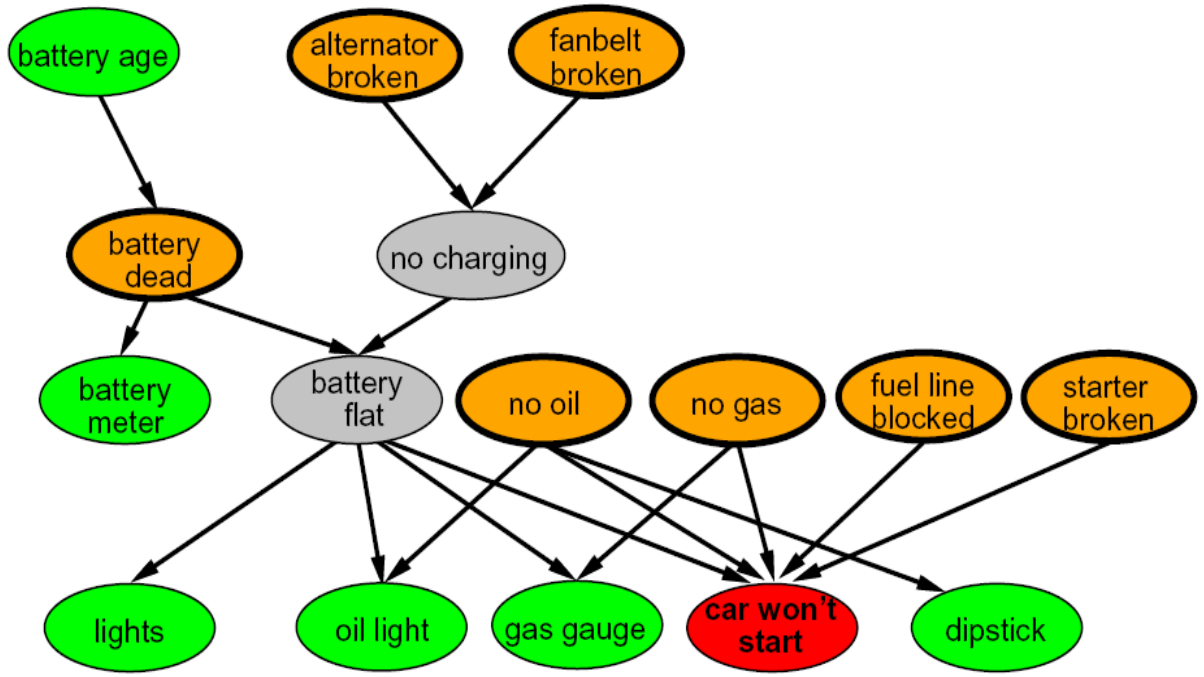
Bayesian Networks

- Problem with using full joint distribution tables as our probabilistic models:
  - Unless there are only a few variables, the joint is hard to represent explicitly.

- Bayesian Networks:
  - A technique for describing complex joint distributions (models) using simple, local distributions (conditional probabilities)
    - Also known as probabilistic graphical models
  - Encode how variables locally influence each other. Local interactions chain together to give global, indirect interactions

Examples


Season
Flu
Hayfever
Muscle-Pain
Congestion


battery age
alternator broken
fanbelt broken
battery dead
no charging
battery meter
battery flat
no oil
no gas
fuel line blocked
starter broken
lights
oil light
gas gauge
car won't start
dipstick

Bayesian Networks: Semantics

- A directed, acyclic graph, one node per random variable
- A conditional probability table (CPT) for each node
  - A collection of distributions over X, one for each combination of parents' values

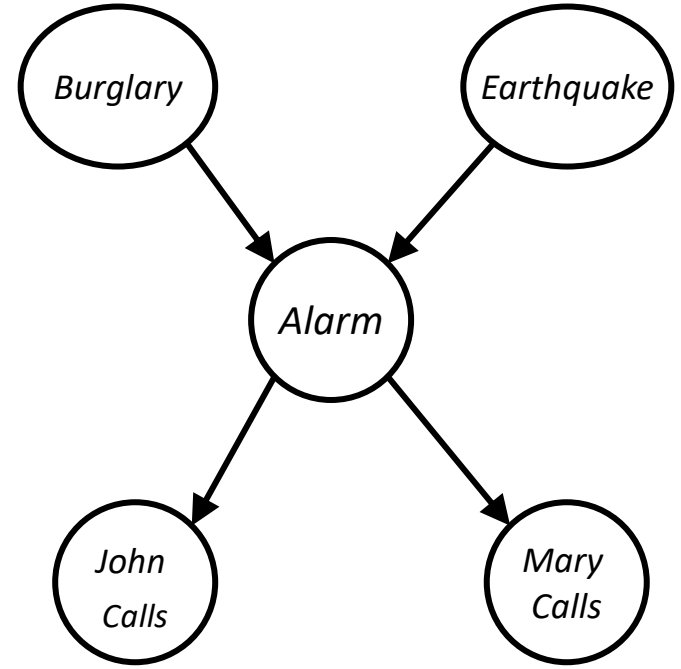
$$P(X|a_1 \ldots a_n)$$

- Bayesian Networks 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:

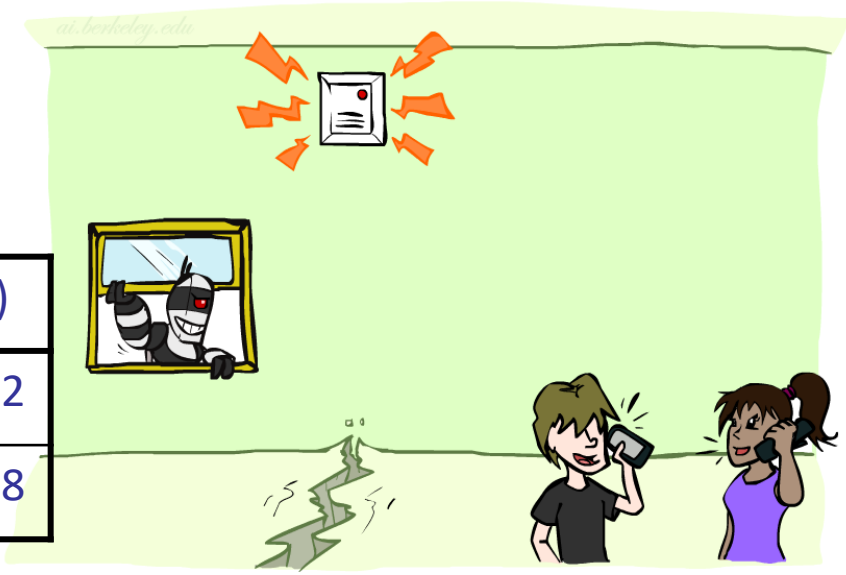
$$P(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} P(x_i|parents(X_i))$$

Example: The Alarm Network

B	P(B)
+b	0.001
-b	0.999



E	P(E)
+e	0.002
-e	0.998

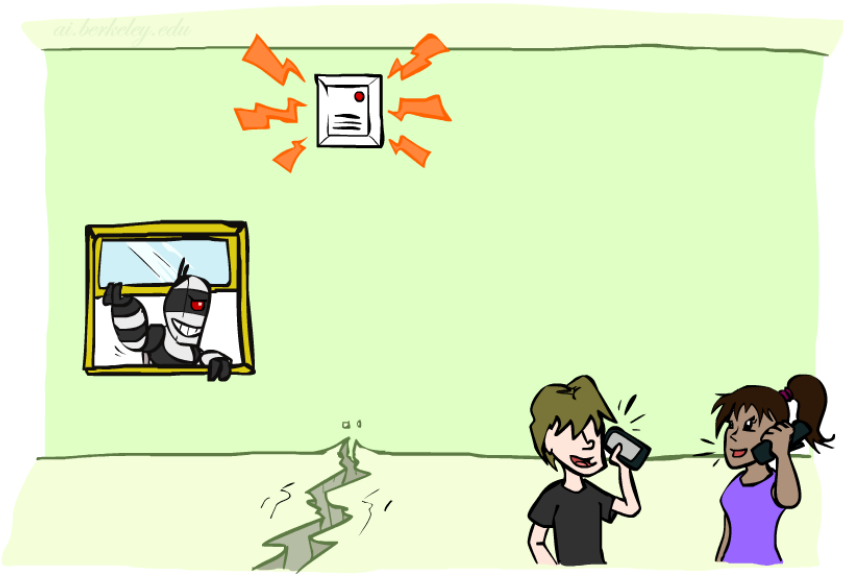


A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

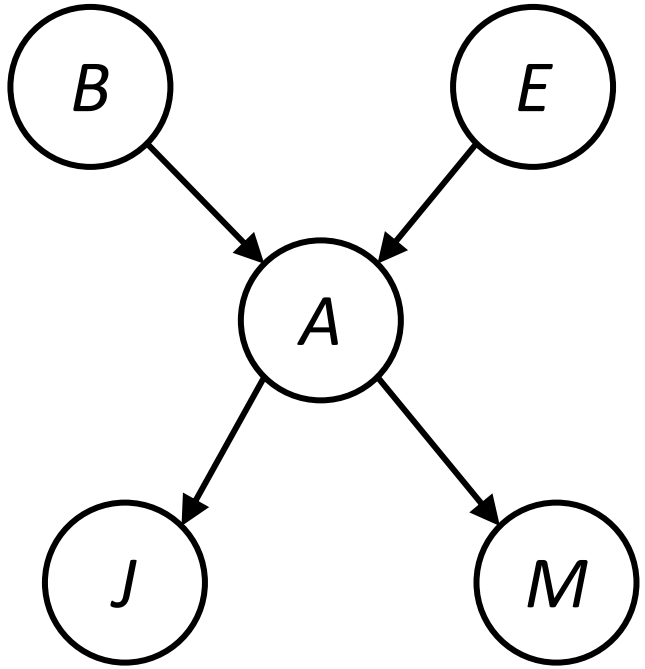
A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

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

Example: The Alarm Network



B	P(B)
+b	0.001
-b	0.999



E	P(E)
+e	0.002
-e	0.998

A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

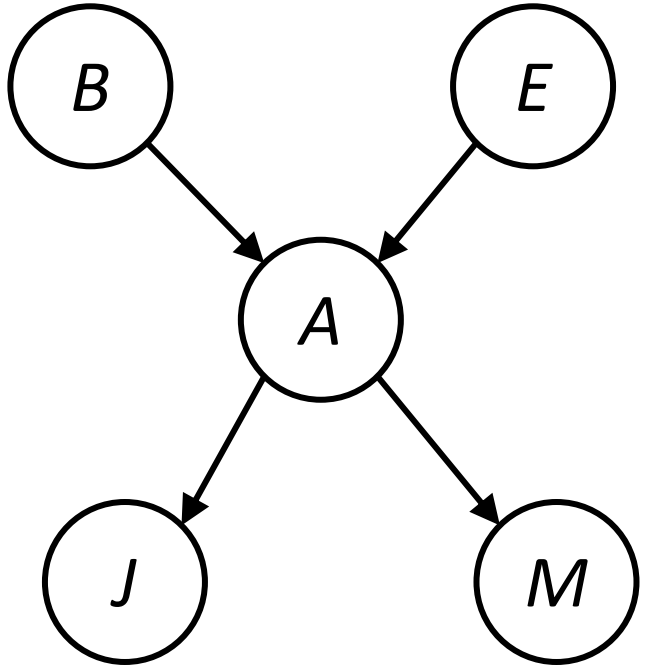
A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

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

$$P(+b, -e, +a, -j, +m) =$$

Example: The Alarm Network

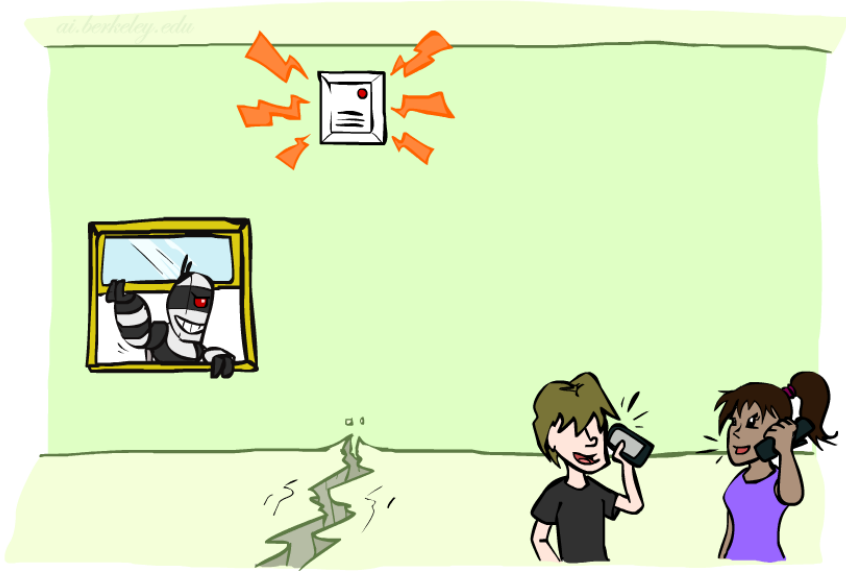
B	P(B)
+b	**0.001**
-b	0.999



E	P(E)
+e	0.002
-e	**0.998**

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



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

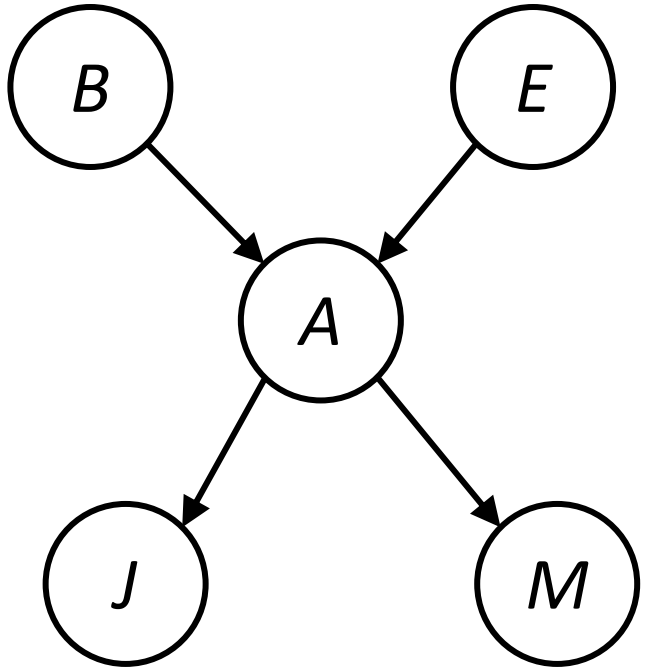
$$P(+b, -e, +a, -j, +m) =$$

$$P(+b)P(-e)P(+a|+b,-e)P(-j|+a)P(+m|+a) =$$

$$0.001 \times 0.998 \times 0.94 \times 0.1 \times 0.7$$

Inference by Enumeration in a Bayes Net

- Inference by enumeration is one way to perform inference in a Bayesian Network (Bayes Net).



$$P(B \mid +j, +m) \propto_B P(B, +j, +m)$$

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

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

$$= P(B)P(+e)P(+a|B,+e)P(+j|+a)P(+m|+a) + P(B)P(+e)P(-a|B,+e)P(+j|-a)P(+m|-a)$$
$$P(B)P(-e)P(+a|B,-e)P(+j|+a)P(+m|+a) + P(B)P(-e)P(-a|B,-e)P(+j|-a)P(+m|-a)$$

Bayesian Networks: Inference

- **Bayesian Networks**
  - Implicitly encode a probability distribution
  - As a product of local conditional distributions
- **Variables**
  - Query variables
  - Evidence variables
  - Hidden variables
- **Inference: What we want to estimate?**
  - Estimating some useful quantity from the joint distribution.
  - Posterior probability
  - Most likely explanation

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

$$\left.\begin{array}{l} Q \\ E_1 \ldots E_k = e_1 \ldots e_k \\ H_1 \ldots H_r \end{array}\right\} X_1, X_2, \ldots X_n$$

All variables

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

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

Inference by Enumeration

- **Setup:** A distribution over query variables (Q) given evidence variables (E)
  - Select entries consistent with the evidence.
  - E.g., Alarm rang, it is rainy, disease present
- Compute the **joint distribution**
- **Sum out** (eliminate) the hidden variables (H)
- **Normalize** the distribution
- **Next**
  - Introduce a notion called **factors**
  - Understand this computation using joining and marginalization of factors.

$$P(Q|e_1 \dots e_k)$$

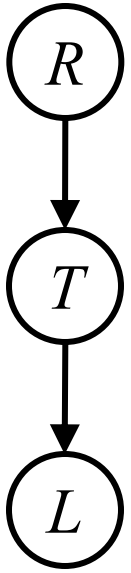
$$P(Q, e_1 \dots e_k) = \sum_{h_1 \dots h_r} P(Q, h_1 \dots h_r, e_1 \dots e_k)$$

$$Z = \sum_q P(Q, e_1 \cdots e_k)$$

$$P(Q|e_1 \cdots e_k) = \frac{1}{Z} P(Q, e_1 \cdots e_k)$$

Example: Traffic Domain

- **Random Variables**
  - R: Raining
  - T: Traffic
  - L: Late for class



$$P(L) = ?$$

$$= \sum_{r,t} P(r,t,L)$$

$$= \sum_{r,t} P(r)P(t|r)P(L|t)$$

$P(R)$

+r	0.1
-r	0.9

$P(T|R)$

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

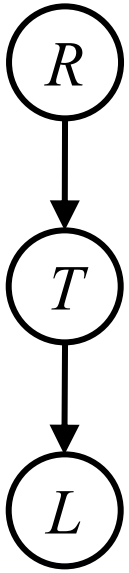
$P(L|T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

Example: Traffic Domain

- **What are factors?**
  - A factor is a function from some set of variables into a specific value.
- **Initial factors**
  - Conditional probability tables (one per node)
  - Select the values consistent with the evidence
- **Inference by Enumeration**
  - Procedure that joins all the factors and then sums out all the hidden variables.
  - Define “joining” and “summing” next.

Traffic domain



$P(R)$

+r	0.1
-r	0.9

$P(T|R)$

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

$P(L|T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

$L = +\ell$ applied to the initial factors

$P(R)$

+r	0.1
-r	0.9

$P(T|R)$

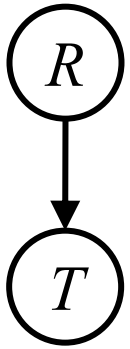
+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

$P(+\ell|T)$

+t	+l	0.3
-t	+l	0.1

Operation I: Joining Factors

- **Joining**
  - Get all the factors over the joining variables.
  - Build a new factor over the union of variables involved.
  - Computation for each entry: pointwise products



$$P(R) \times P(T|R) \Rightarrow P(R,T)$$

+r	0.1
-r	0.9

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

+r	+t	0.08
+r	-t	0.02
-r	+t	0.09
-r	-t	0.81

$$\forall r, t: \quad P(r,t) = P(r) \cdot P(t|r)$$

Joining Factors

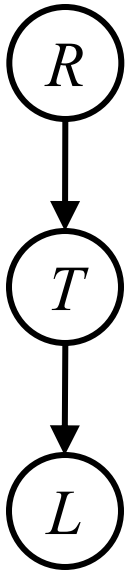
A	B	$\mathbf{f}_1(A,B)$	B	C	$\mathbf{f}_2(B,C)$	A	B	C	$\mathbf{f}_3(A,B,C)$
T	T	.3	T	T	.2	T	T	T	$.3 \times .2 = .06$
T	F	.7	T	F	.8	T	T	F	$.3 \times .8 = .24$
F	T	.9	F	T	.6	T	F	T	$.7 \times .6 = .42$
F	F	.1	F	F	.4	T	F	F	$.7 \times .4 = .28$
						F	T	T	$.9 \times .2 = .18$
						F	T	F	$.9 \times .8 = .72$
						F	F	T	$.1 \times .6 = .06$
						F	F	F	$.1 \times .4 = .04$

Figure 14.10 Illustrating pointwise multiplication: $\mathbf{f}_1(A,B) \times \mathbf{f}_2(B,C) = \mathbf{f}_3(A,B,C)$.

$$\mathbf{f}(X_1 \ldots X_j, Y_1 \ldots Y_k, Z_1 \ldots Z_l) = \mathbf{f}_1(X_1 \ldots X_j, Y_1 \ldots Y_k)\, \mathbf{f}_2(Y_1 \ldots Y_k, Z_, \ldots Z_l).$$

Source: AIMA Ch 14.

Joining Multiple Factors



$P(R)$

+r	0.1
-r	0.9

$P(T|R)$

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

$P(L|T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

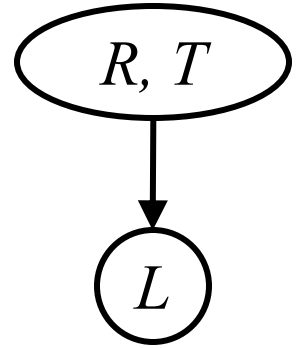
Join R

$P(R,T)$

+r	+t	0.08
+r	-t	0.02
-r	+t	0.09
-r	-t	0.81

$P(L|T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9



Join T

$P(R,T,L)$

+r	+t	+l	0.024
+r	+t	-l	0.056
+r	-t	+l	0.002
+r	-t	-l	0.018
-r	+t	+l	0.027
-r	+t	-l	0.063
-r	-t	+l	0.081
-r	-t	-l	0.729

Operation II: Eliminating Factors

- **Marginalization**
  - Take a factor and sum out a variable
  - Shrinks the factor to a smaller one

$P(R,T)$

+r	+t	0.08
+r	-t	0.02
-r	+t	0.09
-r	-t	0.81

Sum out R

$P(T)$

+t	0.17
-t	0.83

$P(R,T,L)$

R, T, L

+r	+t	+l	0.024
+r	+t	-l	0.056
+r	-t	+l	0.002
+r	-t	-l	0.018
-r	+t	+l	0.027
-r	+t	-l	0.063
-r	-t	+l	0.081
-r	-t	-l	0.729

Sum out R

T, L

$P(T,L)$

+t	+l	0.051
+t	-l	0.119
-t	+l	0.083
-t	-l	0.747

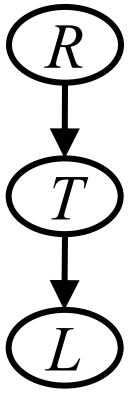
Sum out T

L

$P(L)$

+l	0.134
-l	0.866

Inference by Enumeration



$P(L) = ?$

Multiple join operations and multiple eliminate operations

$P(R,T,L)$

$P(L)$

$P(R)$

+r	0.1
-r	0.9

$P(T|R)$

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

$P(L|T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

Join R

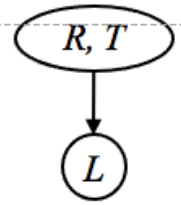
$P(R,T)$

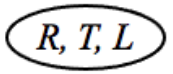
+r	+t	0.08
+r	-t	0.02
-r	+t	0.09
-r	-t	0.81

$P(L|T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

Join T

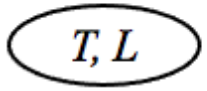




$P(R,T,L)$

+r	+t	+l	0.024
+r	+t	-l	0.056
+r	-t	+l	0.002
+r	-t	-l	0.018
-r	+t	+l	0.027
-r	+t	-l	0.063
-r	-t	+l	0.081
-r	-t	-l	0.729

Sum out R



$P(T,L)$

+t	+l	0.051
+t	-l	0.119
-t	+l	0.083
-t	-l	0.747

Sum out T

$P(L)$

+l	0.134
-l	0.866

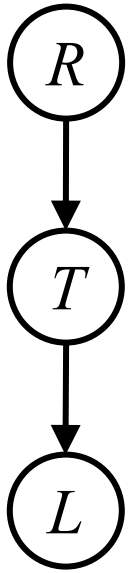
Variable Elimination

- **Inference by Enumeration**
  - Problem: the whole distribution is “joined up“ before “sum out” the hidden variables

- **Variable Elimination**
  - Interleaves joining and eliminating variables
  - Does not create the full joint distribution in one go
  - *Key Idea:*
    - Picks a variable ordering. Picks a variable.
    - Joins all factors containing that variable.
    - Sums out the influence of the variable on new factor.

  - Leverage the **structure** (topology) of the Bayesian Network
  - Marginalize **early** (avoid growing the full joint distribution)

Inference by Enumeration vs. Variable Elimination



$$P(L) = ?$$

Inference by Enumeration

$$= \sum_t \sum_r P(L|t)P(r)P(t|r)$$

Join on r

Join on t

Eliminate r

Eliminate t

Variable Elimination

$$= \sum_t P(L|t) \sum_r P(r)P(t|r)$$

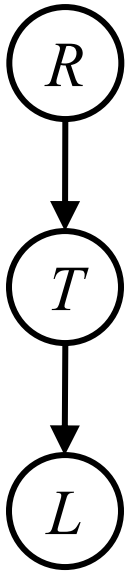
Join on r

Eliminate r

Join on t

Eliminate t

Variable Elimination



$P(R)$

+r	0.1
-r	0.9

$P(T|R)$

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

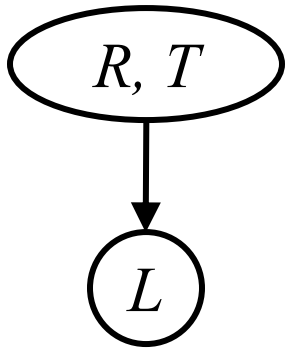
$P(L|T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

Join R

$P(R,T)$

+r	+t	0.08
+r	-t	0.02
-r	+t	0.09
-r	-t	0.81



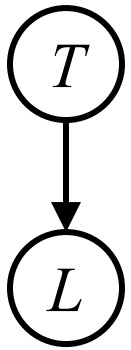
$P(L|T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

Sum out R

$P(T)$

+t	0.17
-t	0.83



$P(L|T)$

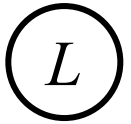
+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

Join T

$P(T,L)$

+t	+l	0.051
+t	-l	0.119
-t	+l	0.083
-t	-l	0.747

Sum out T



$P(L)$

+l	0.134
-l	0.866

Incorporating Evidence

- Till Now, we computed **P(Late)**?
- What happens when **P(Late| Rain)**?
- How to incorporate *evidence* in Variable Elimination.
- **Solution**
  - If evidence, then start with factors and select the evidence.
  - After selecting evidence, eliminate all variables other than query and evidence.

$P(R)$

+r	0.1
-r	0.9

$P(T|R)$

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

$P(L|T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

$P(+r)$

+r	0.1

$P(T|+r)$

+r	+t	0.8
+r	-t	0.2

$P(L|T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

Evidence incorporated in the initial factors

General Variable Elimination

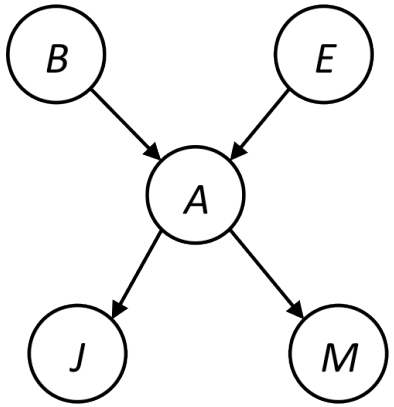
- **Query:** $P(Q|E_1 = e_1, \ldots E_k = e_k)$

- **Start with initial factors:**
  - Local conditional probability tables.
  - Evidence (known) variables are instantiated.

- **While there are still hidden variables (not Q or evidence):**
  - Pick a hidden variable H (from some ordering)
  - Join all factors mentioning H
  - Eliminate (sum out) H

- **Join all the remaining factors and normalize**

Example: Alarm Domain



$$P(B|j,m) \propto P(B,j,m)$$

$P(B)$	$P(E)$	$P(A\|B,E)$	$P(j\|A)$	$P(m\|A)$

$$P(B|j,m) \propto P(B,j,m)$$

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

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

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

$$= \sum_{e} P(B)P(e)f_1(j,m|B,e)$$

$$= P(B) \sum_{e} P(e)f_1(j,m|B,e)$$

$$= P(B)f_2(j,m|B)$$

marginal can be obtained from joint by summing out

use Bayes' net joint distribution expression

use x*(y+z) = xy + xz

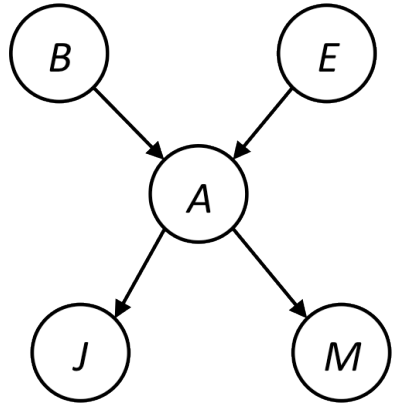
joining on a, and then summing out gives f_1

use x*(y+z) = xy + xz

joining on e, and then summing out gives f_2

Example: Alarm Domain

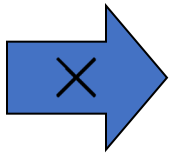
$$P(B|j,m) \propto P(B,j,m)$$



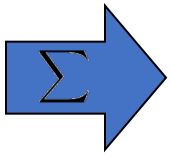
$P(B)$	$P(E)$	$P(A\|B,E)$	$P(j\|A)$	$P(m\|A)$

Choose A

$P(A|B,E)$
$P(j|A)$
$P(m|A)$



$P(j,m,A|B,E)$

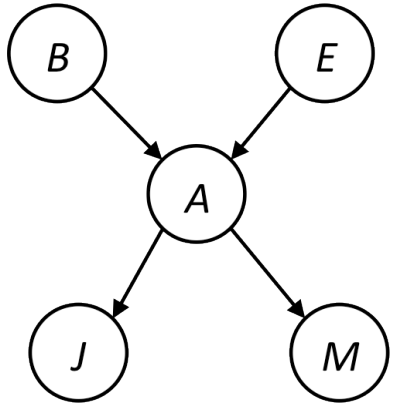


$P(j,m|B,E)$

$P(B)$	$P(E)$	$P(j,m\|B,E)$

Example: Alarm Domain

$P(B)$ $P(E)$ $P(j, m|B, E)$



Choose E

$P(E)$

$P(j, m|B, E)$

→ × → $P(j, m, E|B)$ → $\sum$ → $P(j, m|B)$

$P(B)$ $P(j, m|B)$

Finish with B

$P(B)$

$P(j, m|B)$

→ × → $P(j, m, B)$ → Normalize → $P(B|j, m)$

Variable Elimination: Structuring Computation

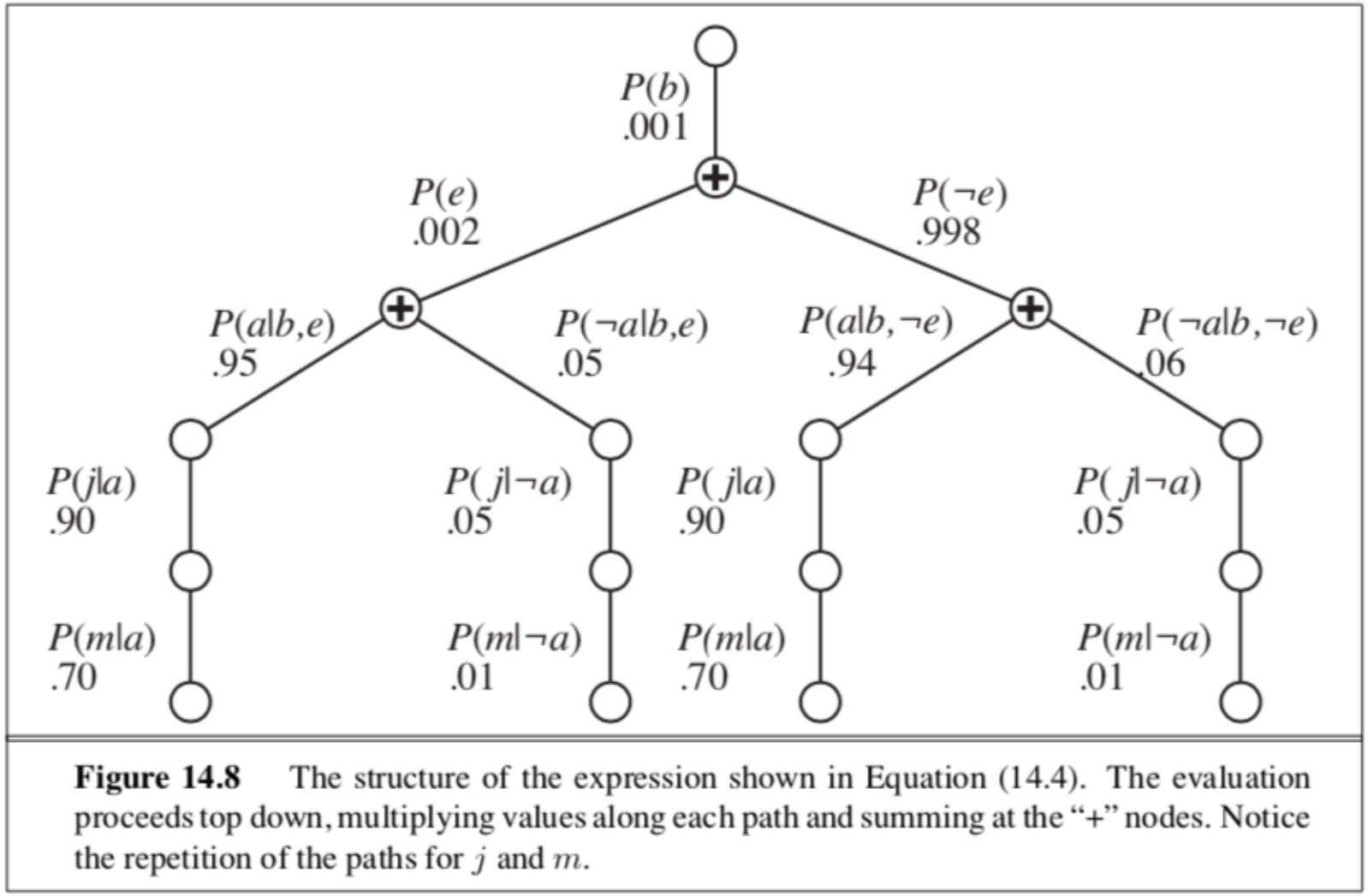


Figure 14.8 The structure of the expression shown in Equation (14.4). The evaluation proceeds top down, multiplying values along each path and summing at the "+" nodes. Notice the repetition of the paths for j and m.

Source: AIMA Ch 14.

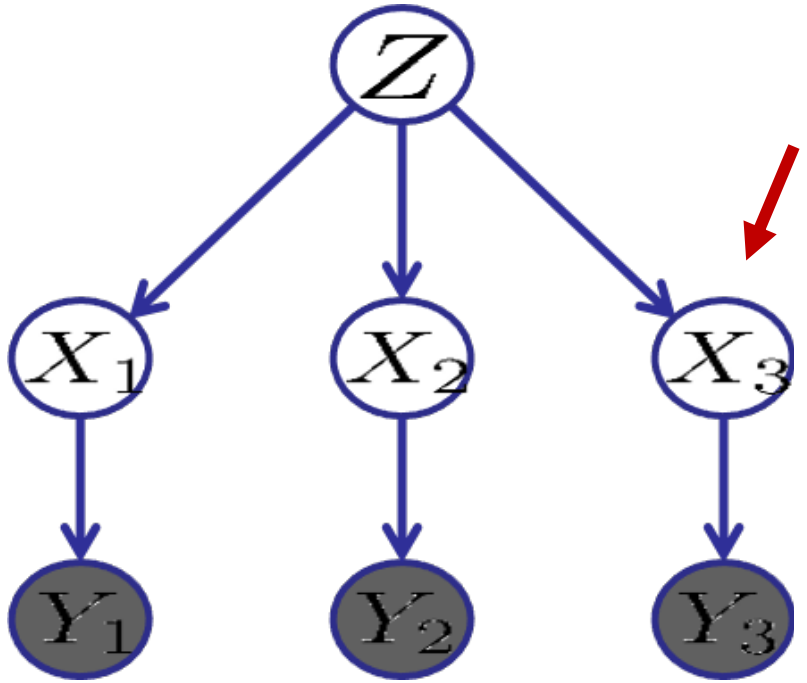
Example

Query: $P(X_3|Y_1 = y_1, Y_2 = y_2, Y_3 = y_3)$

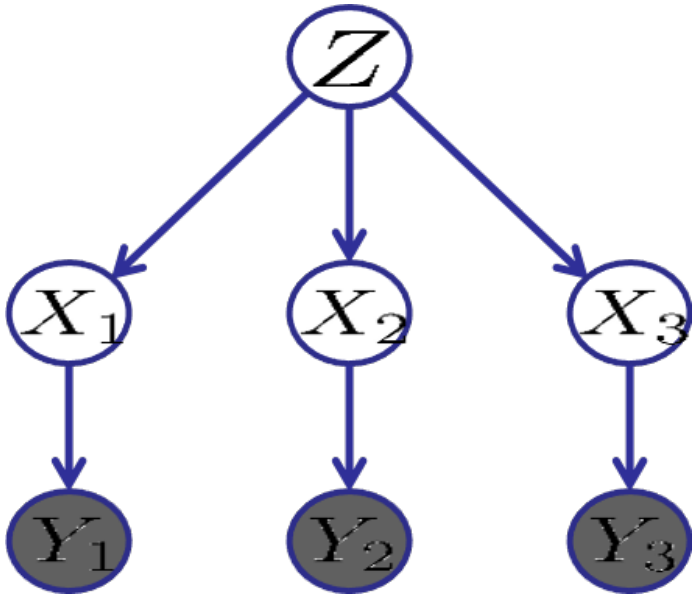
Start by inserting evidence, which gives the following initial factors:

$$P(Z), P(X_1|Z), P(X_2|Z), P(X_3|Z), P(y_1|X_1), P(y_2|X_2), P(y_3|X_3)$$

There are three variables to eliminate { X_1, X_2 and Z }. The Y variables are observed (instantiated).



Example



Query: $P(X_3|Y_1 = y_1, Y_2 = y_2, Y_3 = y_3)$

Start by inserting evidence, which gives the following initial factors:

$$P(Z), P(X_1|Z), P(X_2|Z), P(X_3|Z), P(y_1|X_1), P(y_2|X_2), P(y_3|X_3)$$

Eliminate X_1, this introduces the factor $f_1(y_1|Z) = \sum_{x_1} P(x_1|Z)P(y_1|x_1)$, and we are left with:

$$P(Z), P(X_2|Z), P(X_3|Z), P(y_2|X_2), P(y_3|X_3), f_1(y_1|Z)$$

Eliminate X_2, this introduces the factor $f_2(y_2|Z) = \sum_{x_2} P(x_2|Z)P(y_2|x_2)$, and we are left with:

$$P(Z), P(X_3|Z), P(y_3|X_3), f_1(y_1|Z), f_2(y_2|Z)$$

Eliminate Z, this introduces the factor $f_3(y_1, y_2, X_3) = \sum_z P(z)P(X_3|z)f_1(y_1|Z)f_2(y_2|Z)$, and we are left with:

$$P(y_3|X_3), f_3(y_1, y_2, X_3)$$

No hidden variables left. Join the remaining factors to get:

$$f_4(y_1, y_2, y_3, X_3) = P(y_3|X_3), f_3(y_1, y_2, X_3)$$

Normalizing over X_3 gives $P(X_3|y_1, y_2, y_3) = f_4(y_1, y_2, y_3, X_3)/\sum_{x_3} f_4(y_1, y_2, y_3, x_3)$

Example

Query: $P(X_3|Y_1 = y_1, Y_2 = y_2, Y_3 = y_3)$

Start by inserting evidence, which gives the following initial factors:

$$P(Z), P(X_1|Z), P(X_2|Z), P(X_3|Z), P(y_1|X_1), P(y_2|X_2), P(y_3|X_3)$$

Eliminate X_1, this introduces the factor $f_1(y_1|Z) = \sum_{x_1} P(x_1|Z)P(y_1|x_1)$, and we are left with:

$$P(Z), P(X_2|Z), P(X_3|Z), P(y_2|X_2), P(y_3|X_3), f_1(y_1|Z)$$

Eliminate X_2, this introduces the factor $f_2(y_2|Z) = \sum_{x_2} P(x_2|Z)P(y_2|x_2)$, and we are left with:

$$P(Z), P(X_3|Z), P(y_3|X_3), f_1(y_1|Z), f_2(y_2|Z)$$

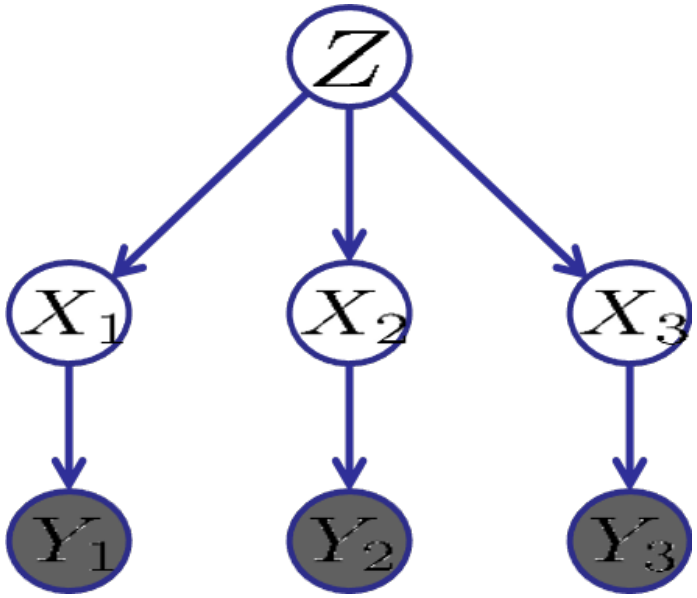
Eliminate Z, this introduces the factor $f_3(y_1, y_2, X_3) = \sum_z P(z)P(X_3|z)f_1(y_1|Z)f_2(y_2|Z)$, and we are left with:

$$P(y_3|X_3), f_3(y_1, y_2, X_3)$$

No hidden variables left. Join the remaining factors to get:

$$f_4(y_1, y_2, y_3, X_3) = P(y_3|X_3), f_3(y_1, y_2, X_3)$$

Normalizing over X_3 gives $P(X_3|y_1, y_2, y_3) = f_4(y_1, y_2, y_3, X_3)/\sum_{x_3} f_4(y_1, y_2, y_3, x_3)$

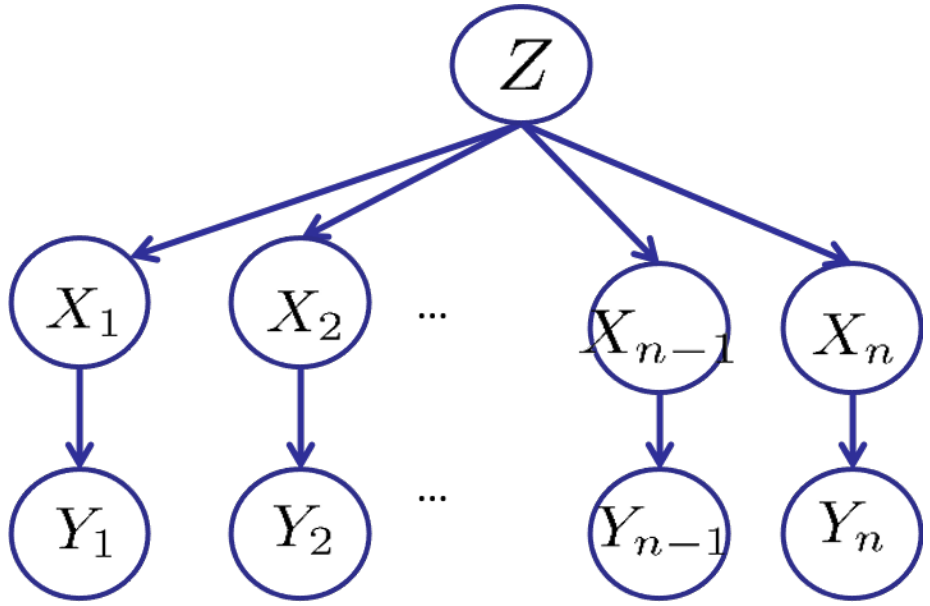


Computational complexity

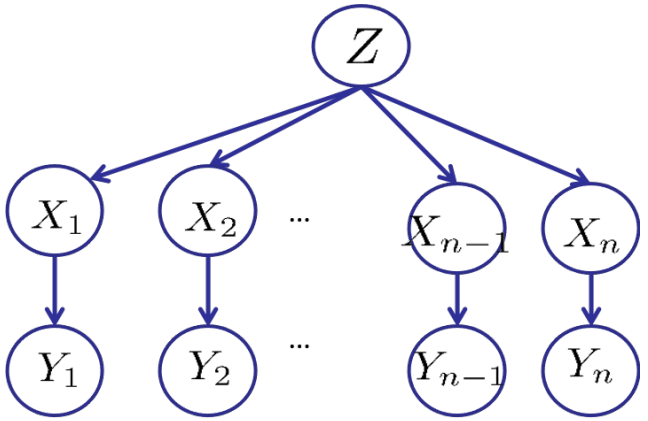
- Depends on the largest factor generated in VE.
- Factor size = number of entries in the table.
- In this example: each factor is of size 2 (only one variable). Note that y is observed.
- X_1, X_2, Z, X_3

Effect of Different Orderings

- For the query $P(X_n | y_1,\ldots,y_n)$
- Two different orderings as
  - Eliminate Z first. $Z, X_1, \ldots, X_{n-1}$
  - Eliminate Z last. $X_1, \ldots, X_{n-1}, Z$.
  - What is the size of the maximum factor generated for each of the orderings?



Example



Eliminate Z First

$$P(X_n, |y_1, y_2, \ldots, y_n) = \alpha P(Z)P(X_1|Z)P(X_2|Z), \ldots, P(X_n|Z)P(y_1|X_1)P(y_2|X_2), \ldots, P(y_n|X_n)$$

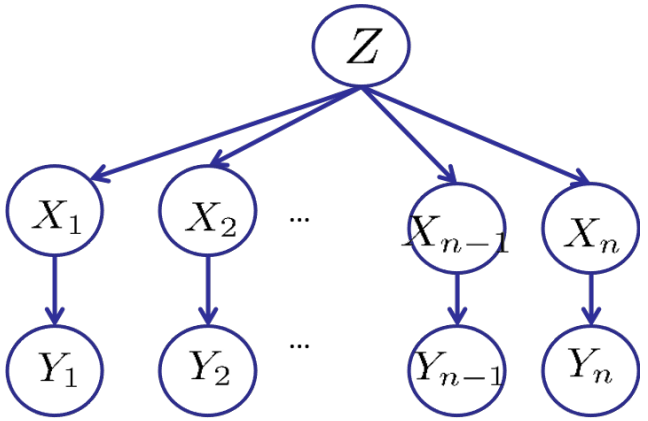
This factor is 2^n → $f_1(X_1, X_2, \ldots, X_n) = \sum_z P(z)P(X_1|z), P(X_1|z), \ldots, P(X_1|z)$

$$P(X_n, |y_1, y_2, \ldots, y_n) = \alpha f_1(X_1, X_2, \ldots, X_n)P(y_1|X_1)P(y_2|X_2), \ldots, P(y_n|X_n)$$

$$f_2(X_1, X_2, \ldots, X_{n-1}) = \sum_{x_n} f_1(X_1, X_2, \ldots, X_{n-1}, x_n)P(y_n|x_n)$$

$$P(X_n, |y_1, y_2, \ldots, y_n) = \alpha f_2(X_1, X_2, \ldots, X_{n-1})P(y_1|X_1)P(y_2|X_2), \ldots, P(y_{n-1}|X_{n-1})$$

Example



Eliminate Z Last

$$P(X_n, |y_1, y_2, \ldots, y_n) = \alpha P(Z)P(X_1|Z)P(X_2|Z), \ldots, P(X_n|Z)P(y_1|X_1)P(y_2|X_2), \ldots, P(y_n|X_n)$$

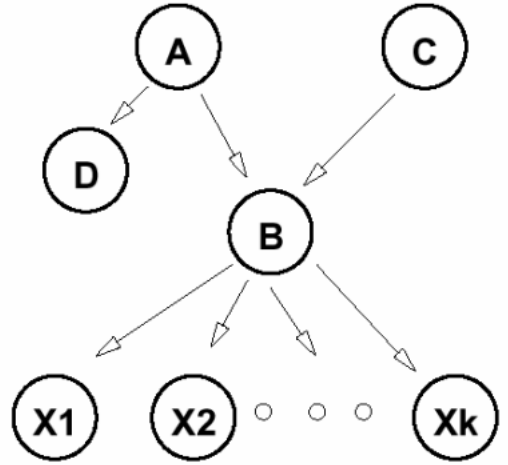
This factor is size 2 → $f_1(y_1|Z) = \sum_{x_1} P(y_1|x_1)P(x_1|Z)$

$$P(X_n, |y_1, y_2, \ldots, y_n) = \alpha P(Z)f_1(y_1|Z)P(X_2|Z), \ldots, P(X_n|Z)P(y_2|X_2), \ldots, P(y_n|X_n)$$

Other steps are like the previous example. Each factor is of size 2 consisting of one variable.
Variable ordering can have considerable impact.

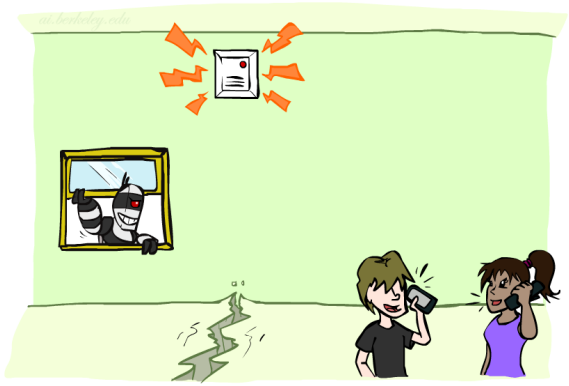
Properties

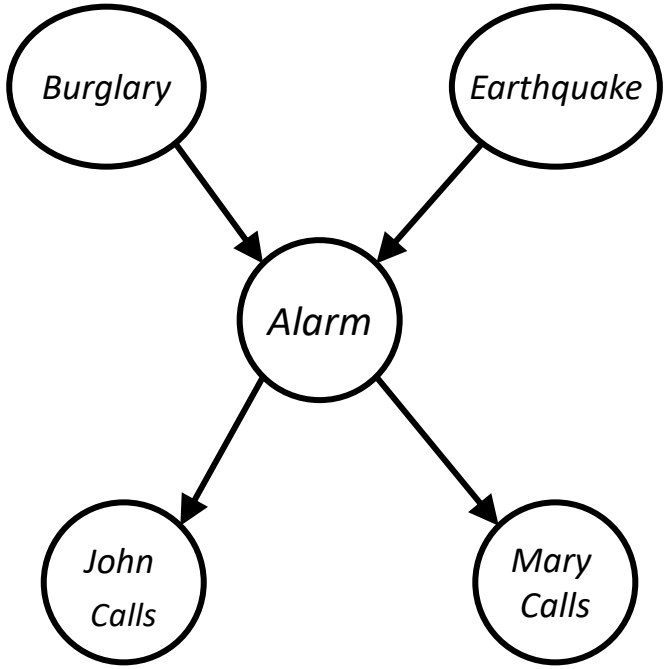
- Variable elimination is dominated by the size of the largest factor constructed during the operation of the algorithm.
- Depends on the structure of the network and order of elimination of the variables.
- Finding the optimal ordering is intractable.
  - Can pose the problem of finding good ordering as a search.
  - Use heuristics.
- Min-fill heuristic
  - Eliminate the variable that creates the smallest sized factor (greedy approach).
- Min-neighbors
  - Eliminate the variable that has the smallest number of neighbors in the current graph.



Rank A, B and D with the Min-Fill heuristic.

Irrelevant Variables





$$P(J)$$

$$= \sum_{M,A,B,E} P(J,M,A,B,E)$$

$$= \sum_{M,A,B,E} P(J|A)P(B)P(A|B,E)P(E)P(M|A)$$

$$= \sum_{A} P(J|A) \sum_{B} P(B) \sum_{E} P(A|B,E)P(E) \sum_{M} P(M|A)$$

Every variable that is not an ancestor of a query variable or evidence variable is irrelevant for the query.

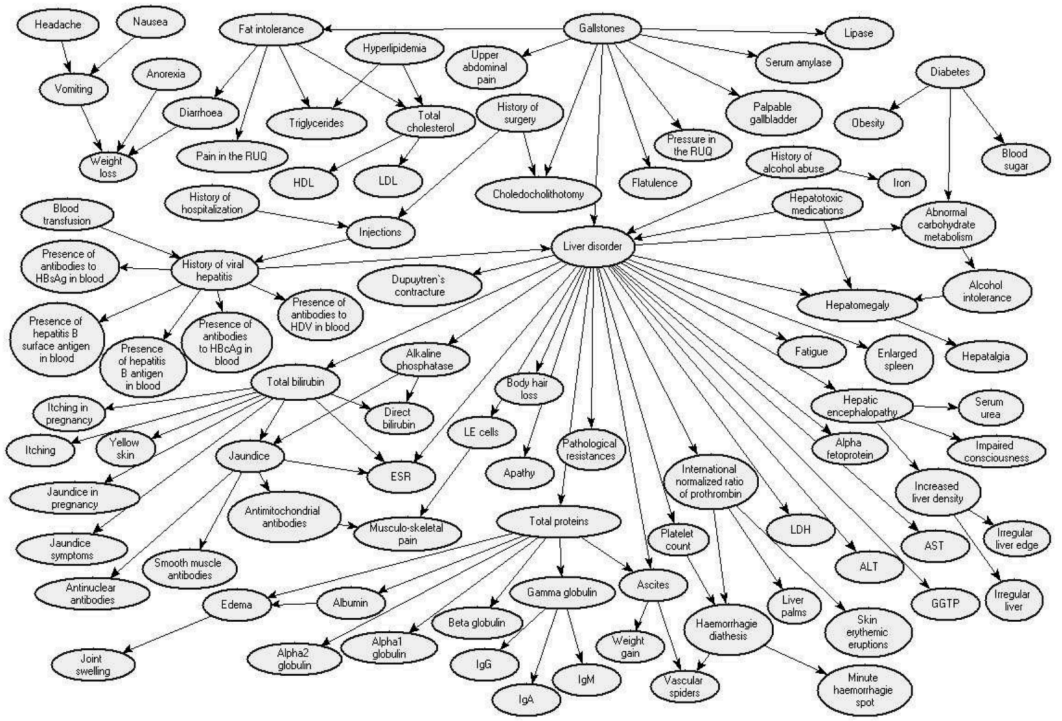
Slide adapted from Prof. Mausam

Bayesian Networks: Independence

- Bayesian Networks
  - Implicitly encode joint distributions
  - A collection of distributions over X, one for each combination of parents' values
  - Product of local conditional distributions
- Inference
  - Given a fixed BN, what is P(X | e)
  - Variable Elimination
- **Modeling**
  - Understanding the assumptions made when choosing a Bayes net graph

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

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



A Bayesian Network Model for Diagnosis of Liver Disorders. Onisko et al. 99.

Conditional Independence

- X and Y are **independent** if

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

- X and Y are **conditionally independent** given Z

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

- (Conditional) independence: Given Z, Y has no more information to convey about X or Y does not probabilistically influence X.

- Example: $Alarm \perp\!\!\!\perp Fire|Smoke$

Smoke causes the alarm to be triggered. Once there is smoke it does not matter what caused it (e.g., Fire or any other source).

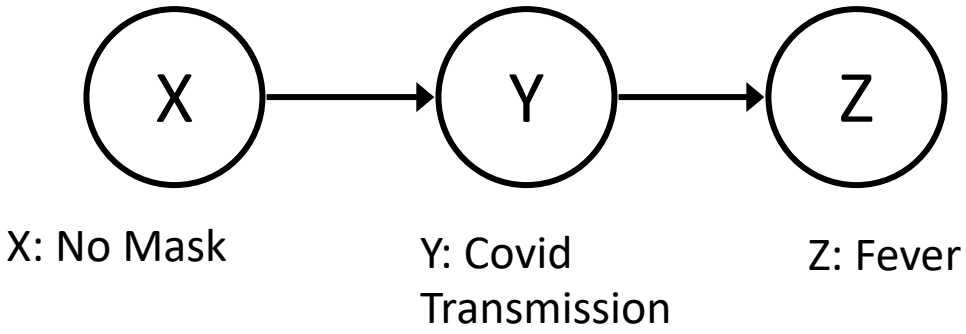
Bayesian Network: Independence Assumptions

- The conditional distributions defining the Bayesian Network (BN) assume conditional independences.

$$P(x_i|x_1 \cdots x_{i-1}) = P(x_i|parents(X_i))$$

- Often there are **additional conditional independences** that are implicit in the network.
- How to show if two variables (X and Y) are conditionally independent given evidence (say Z)?
  - **Yes.** Provide a proof by analyzing the probability expression.
  - **No.** Find a counter example. Instantiate a CPT for the BN such that X and Y are not independent given Z.

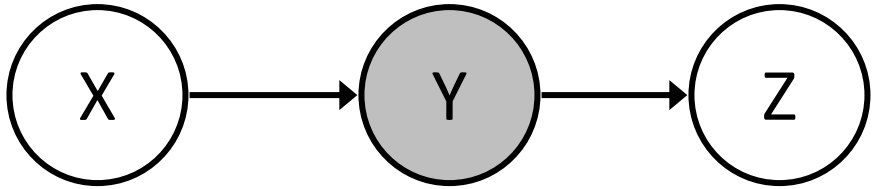
Causal Chains



- **Is X guaranteed to be independent of Z?**
  - **No**
- Intuitively
  - Wearing no masks causes virus transmission which causes fever.
  - Wearing masks causes no virus transmission causes no symptom.
  - Path between X and Z is active.
- Instantiate a CPT

$$P(+y \mid +x) = 1,\ P(-y \mid -x) = 1,$$
$$P(+z \mid +y) = 1,\ P(-z \mid -y) = 1$$

Causal Chains



X: No Mask

Y: Covid Transmission

Z: Fever

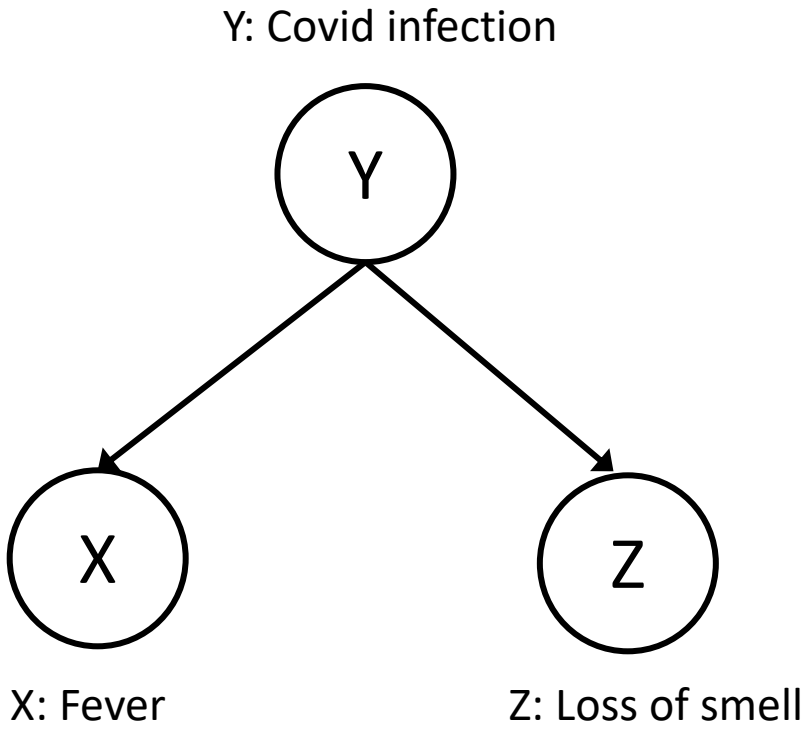
$$P(x, y, z) = P(x)P(y|x)P(z|y)$$

- **Is X guaranteed to be independent of Z given Y?**
  - **Yes**

$$\begin{aligned} P(z|x, y) &= \frac{P(x, y, z)}{P(x, y)} \\ &= \frac{P(x)P(y|x)P(z|y)}{P(x)P(y|x)} \\ &= P(z|y) \end{aligned}$$

- **Evidence along the chain blocks the influence (inactivates the path).**

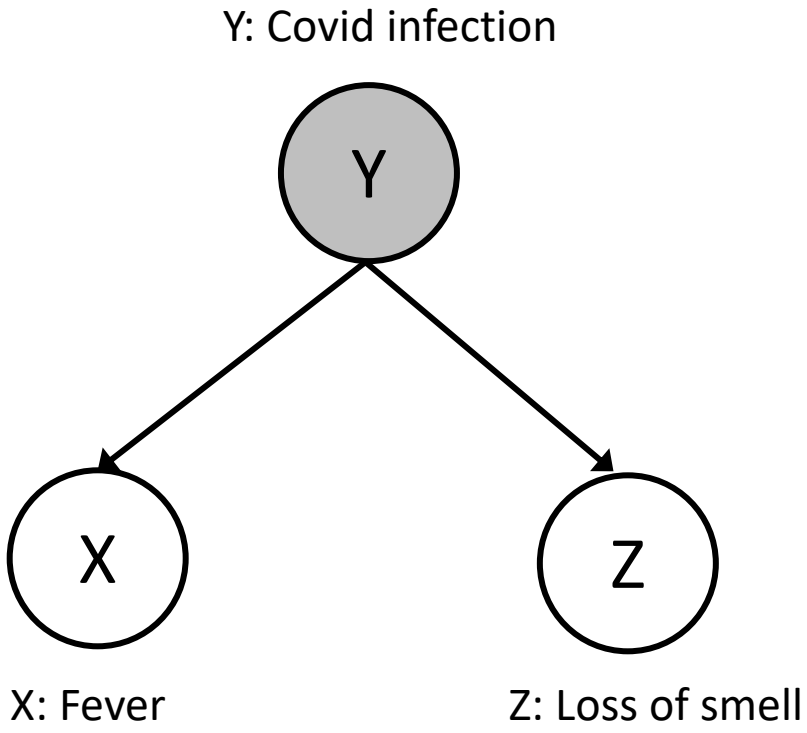
Common Cause



- **Is X guaranteed to be independent of Z?**
  - **No**
- Intuitively
  - Covid infection causes both Fever and Loss of Smell.
  - Path between X and Z is active.
- Instantiate a CPT

$$P(+x \mid +y) = 1,\ P(-x \mid -y) = 1,$$
$$P(+z \mid +y) = 1,\ P(-z \mid -y) = 1$$

Common Cause



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

- **Is X independent of Z given Y?**
  - **Yes**

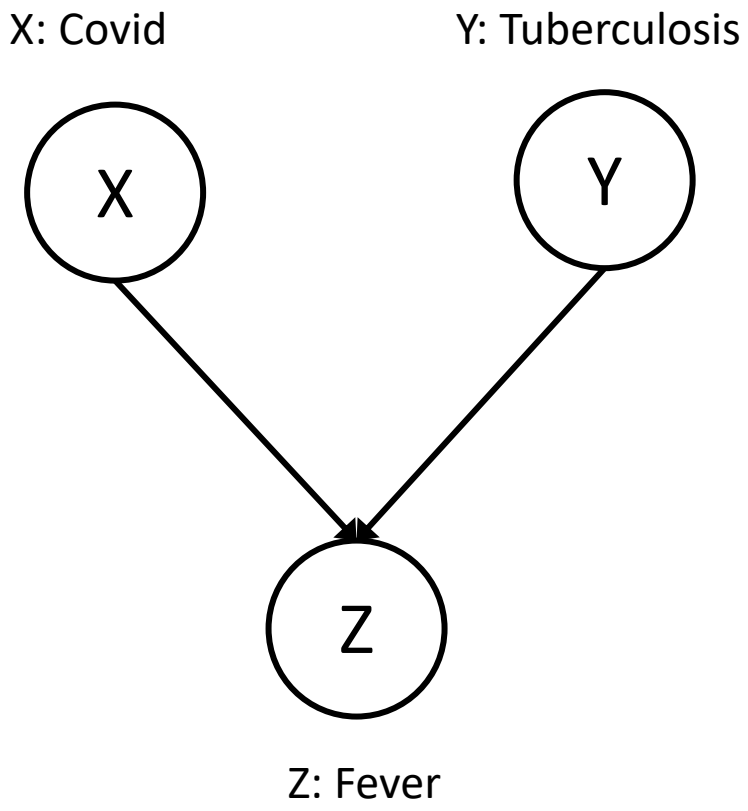
$$P(z|x, y) = \frac{P(x, y, z)}{P(x, y)}$$

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

$$= P(z|y)$$

- If you have Covid, then belief over the loss of smell is not affected by presence of fever
- **Observing the cause blocks the influence (inactivates the path).**

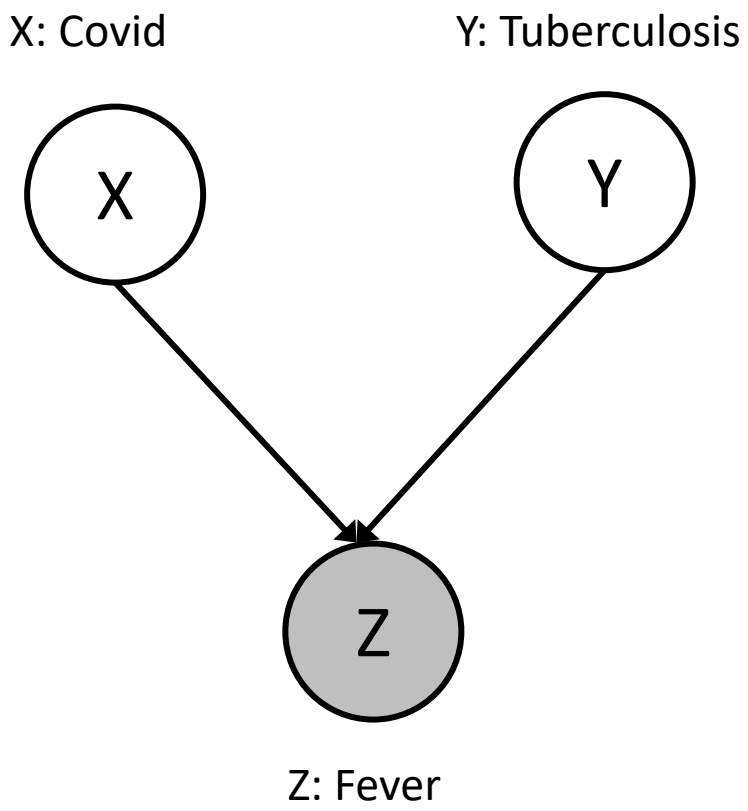
Common Effect



- **Are X and Y independent?**
  - **Yes**
  - Covid and TB both cause Fever. But can't say that if you have Covid then you are more or less likely to have TB (under this model)

$$
\begin{aligned}
P(x,y) &= \sum_z P(x,y,z) \\
&= \sum_z P(x)P(y)P(z|x,y) \\
&= P(x)P(y)\sum_z P(z|x,y) \\
&= P(x)P(y)
\end{aligned}
$$

Common Effect



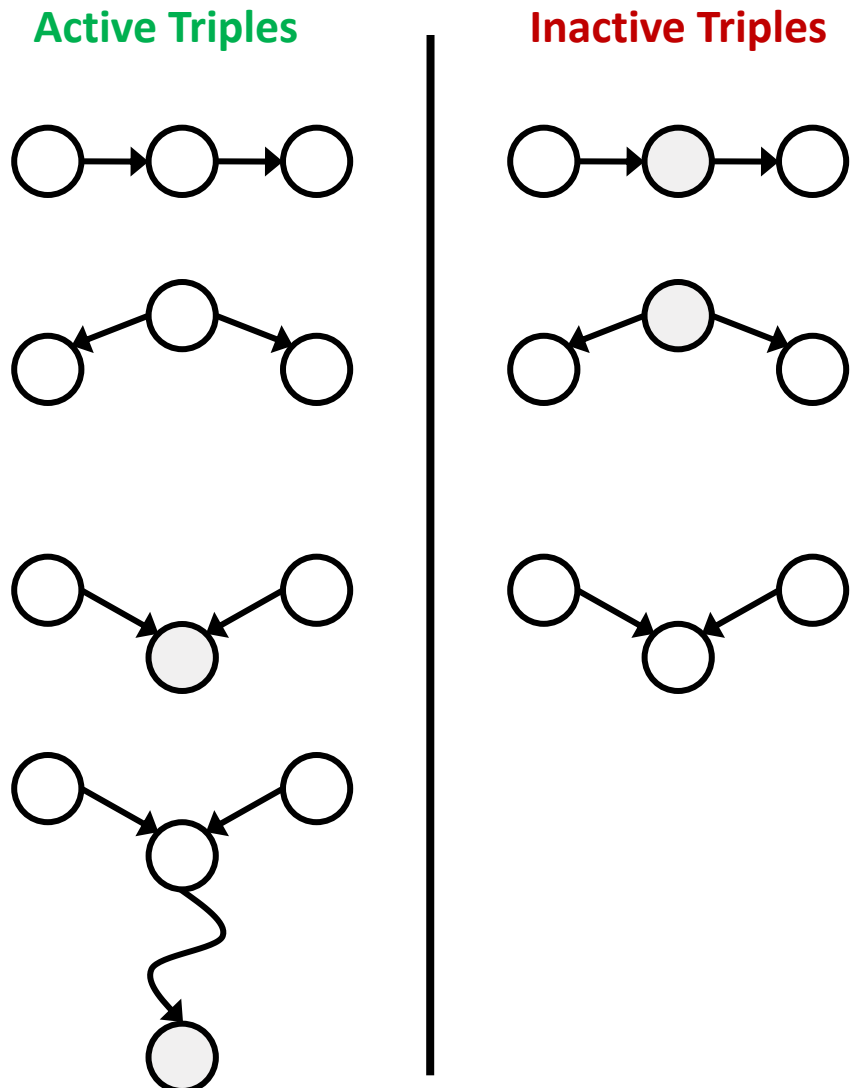
- **Is X independent of Y given Z?**
  - **No**
- Seeing the fever puts Covid and TB in competition as possible causal explanations.
- It is likely that one of them is the cause, rare for both. If Covid is present then the likelihood of TB being present is low (reduces its chances).
- **Observing the cause activates influence between possible causes.**

Active and Inactive Paths

- Question: Are X and Y conditionally independent given evidence variables {Z}?
  - Yes, if X and Y **"d-separated"** by Z
  - Consider all (**undirected**) 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 descendants* is observed

- A path is blocked with even a single inactive segment



D-Separation

- Query: $X_i \perp\!\!\!\perp X_j | \{X_{k_1}, ..., X_{k_n}\}$?
- Check all (undirected) paths between X_i and X_j
  - If one or more active, then independence not guaranteed

$$X_i \not\perp\!\!\!\perp X_j | \{X_{k_1}, ..., X_{k_n}\}$$

  - Otherwise (i.e. if all paths are inactive), then independence is guaranteed

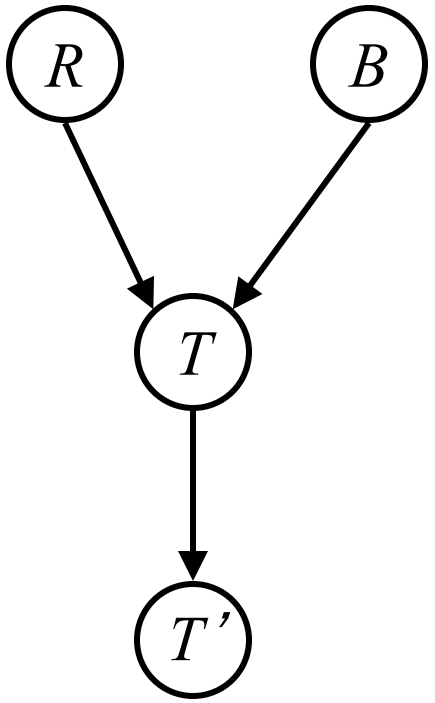
$$X_i \perp\!\!\!\perp X_j | \{X_{k_1}, ..., X_{k_n}\}$$

D-Separation: Examples

$R \perp\!\!\!\perp B$ *Yes*

$R \perp\!\!\!\perp B | T$

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



D-Separation: Examples

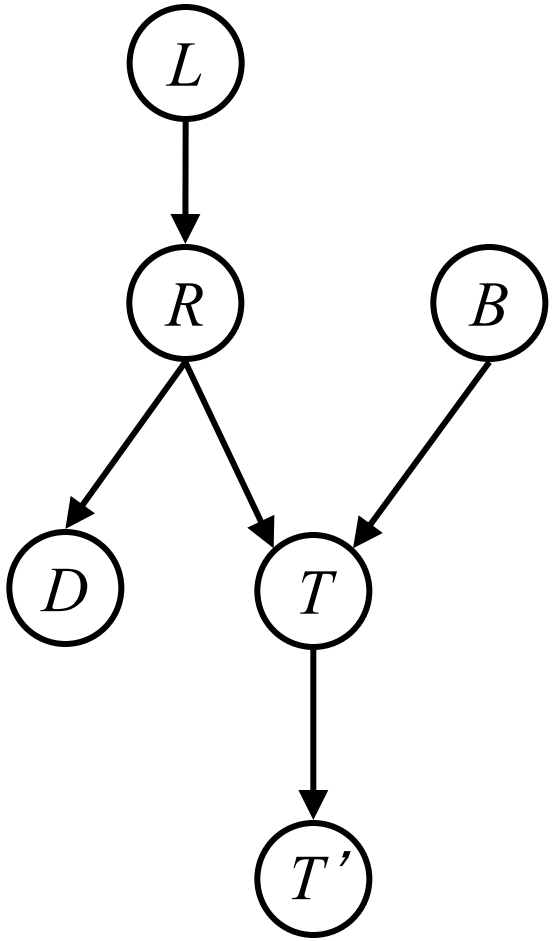
$L \perp\!\!\!\perp T' | T$ *Yes*

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

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

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

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

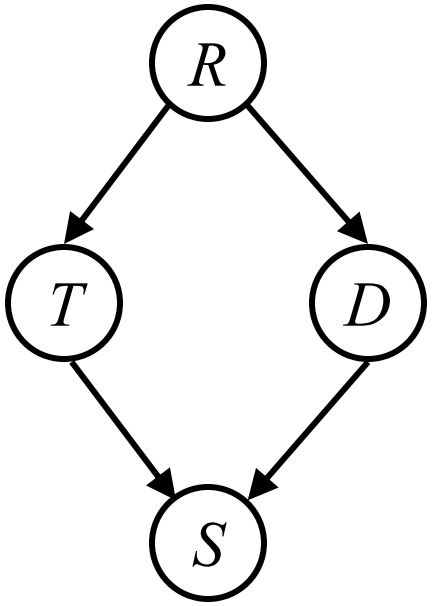


D-Separation: Examples

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

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

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



Encoding a Generative Process

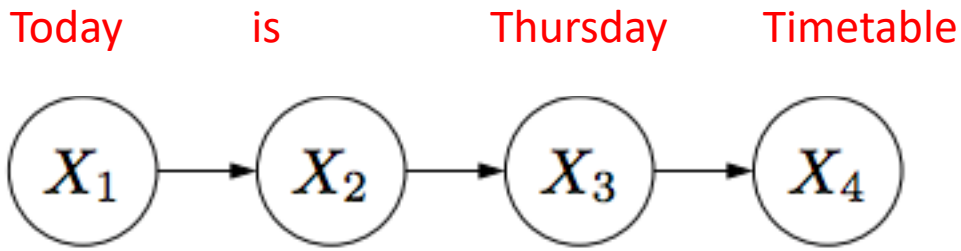
- Directed Graphical Models
  - Encode a generative structure
  - The model our assumptions about how the data was generated.
  - Include modeling assumptions that impact inference.
  - Can also understand as “probabilistic programs”

Examples

Naïve Language Modeling

For each position $i = 1, 2, \ldots, n$:

Generate word $X_i \sim p(X_i \mid X_{i-1})$

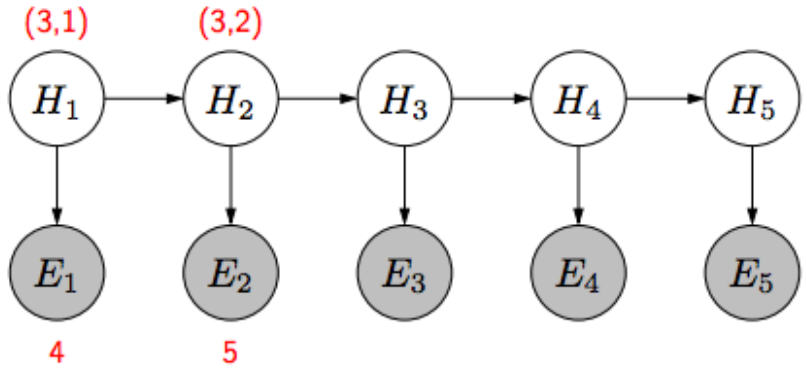


Basic Tracking

For each time step $t = 1, \ldots, T$:

Generate object location $H_t \sim p(H_t \mid H_{t-1})$

Generate sensor reading $E_t \sim p(E_t \mid H_t)$



Adapted from Dorsa Sadigh and Percy Liang

Given noisy observations where is the agent?

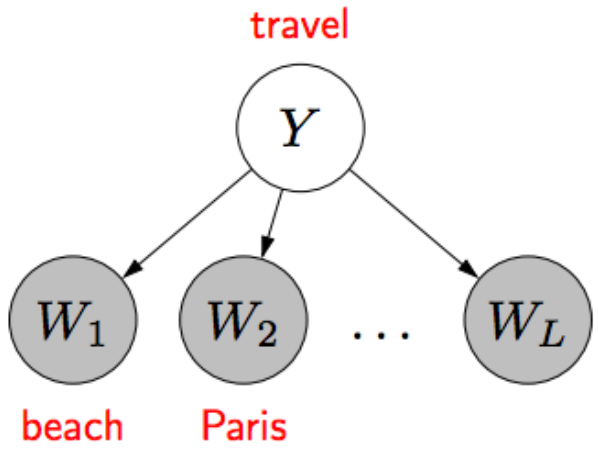
Examples

Document Analysis

Generate label $Y \sim p(Y)$

For each position $i = 1, \ldots, L$:

Generate word $W_i \sim p(W_i \mid Y)$



Given a text document what is it about?

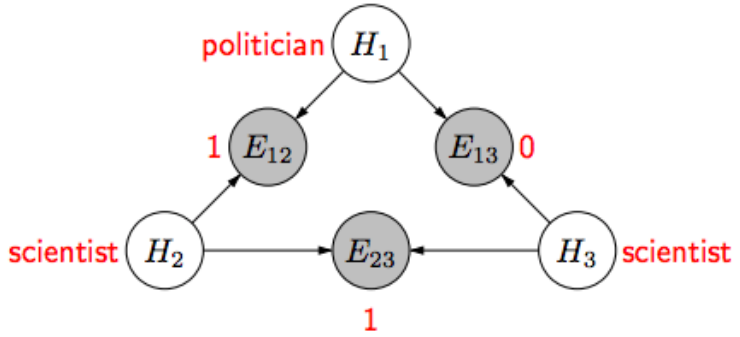
Social Network Analysis

For each person $i = 1, \ldots, n$:

Generate person type $H_i \sim p(H_i)$

For each pair of people $i \neq j$:

Generate connectedness $E_{ij} \sim p(E_{ij} \mid H_i, H_j)$



Given a social network (graph over n people), what types of people are there?

Adapted from Dorsa Sadigh and Percy Liang

Approximate Inference in Probabilistic Models

- **Exact Inference**
  - Inference by enumeration (Variable elimination)
  - Exact likelihood for probabilistic queries.
    - Exact Marginal likelihood P(Late = Yes)
    - Exact Conditional likelihood (posterior probability)
      - P(Late = True | Rain = Yes, Traffic = High) etc.
  - Problem:
    - In many practical applications variable elimination can be intractable. Variable elimination may need to create a large table.

- **Approximate Inference**
  - Compute an "approximate" posterior probability
  - Principle
    - Generate samples from the distribution.
    - Use the samples to construct an approximate estimate of the probabilistic query. P'(Late) or P'(Late| Rain, Traffic).
  - Advantage
    - Generating samples and constructing the approximate distribution is often faster.
    - Note that the estimate is approximate, not exact.

Methods

- Prior Sampling
- Rejection Sampling
- Likelihood Weighting
- Gibbs Sampling

Need the ability to sample from a distribution

- Sampling from given distribution
  - Step 1: Get sample *u* from uniform distribution over [0, 1)
  - Step 2: Convert this sample *u* into an outcome for the given distribution by having each outcome associated with a sub-interval of [0,1) with sub-interval size equal to probability of the outcome
- Utility? We should be able to sample from a CPT defining the probabilistic model.
- Next, we look at approaches for sampling from a Bayes Net.

Example:

C	P(C)
red	0.6
green	0.1
blue	0.3

$$0 \leq u < 0.6, \rightarrow C = red$$
$$0.6 \leq u < 0.7, \rightarrow C = green$$
$$0.7 \leq u < 1, \rightarrow C = blue$$

- If random() returns *u* = 0.83, then our sample is *C* = blue
- E.g, after sampling 8 times:

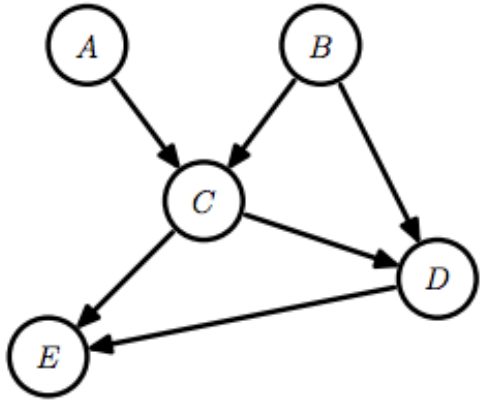
Prior Sampling

- For i=1, 2, …, n
  - **Sample x_i from $P(X_i \mid \text{Parents}(X_i))$**
- Return $(x_1, x_2, \ldots, x_n)$

Ancestral pass for directed graphical models:

— sample each top level variable from its marginal

— sample each other node from its conditional once its parents have been sampled



Sample:

$$A \sim P(A)$$
$$B \sim P(B)$$
$$C \sim P(C \mid A, B)$$
$$D \sim P(D \mid B, C)$$
$$E \sim P(D \mid C, D)$$

$$P(A, B, C, D, E) = P(A)\, P(B)\, P(C \mid A, B)\, P(D \mid B, C)\, P(E \mid C, D)$$

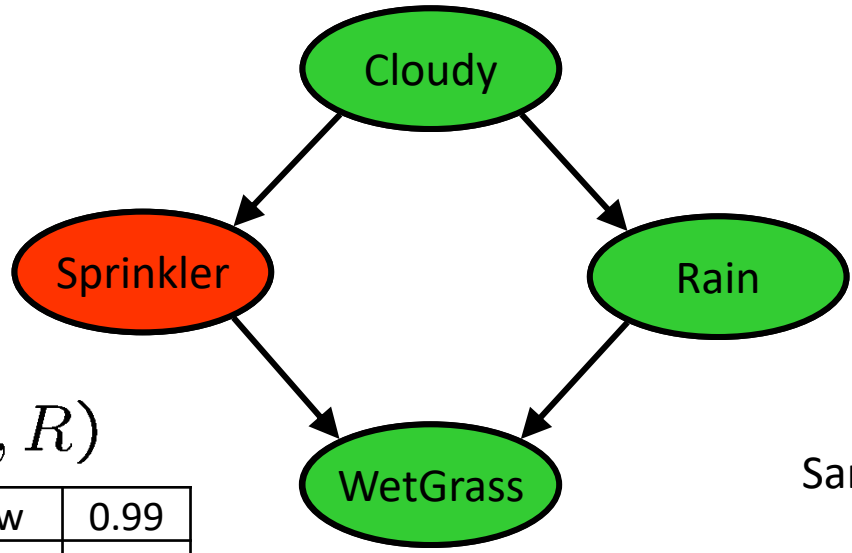
Source MLSS 2009. http://videolectures.net/mlss09uk_murray_mcmc/

Prior Sampling

- For i=1, 2, …, n
  - **Sample x_i from $P(X_i \mid Parents(X_i))$**
- Return $(x_1, x_2, \ldots, x_n)$

$P(C)$

+c	0.5
-c	0.5



$P(S|C)$

+c	+s	0.1
	-s	0.9
-c	+s	0.5
	-s	0.5

$P(R|C)$

+c	+r	0.8
	-r	0.2
-c	+r	0.2
	-r	0.8

$P(W|S,R)$

+s	+r	+w	0.99
		-w	0.01
	-r	+w	0.90
		-w	0.10
-s	+r	+w	0.90
		-w	0.10
	-r	+w	0.01
		-w	0.99

Samples:

+c, -s, +r, +w

-c, +s, -r, +w

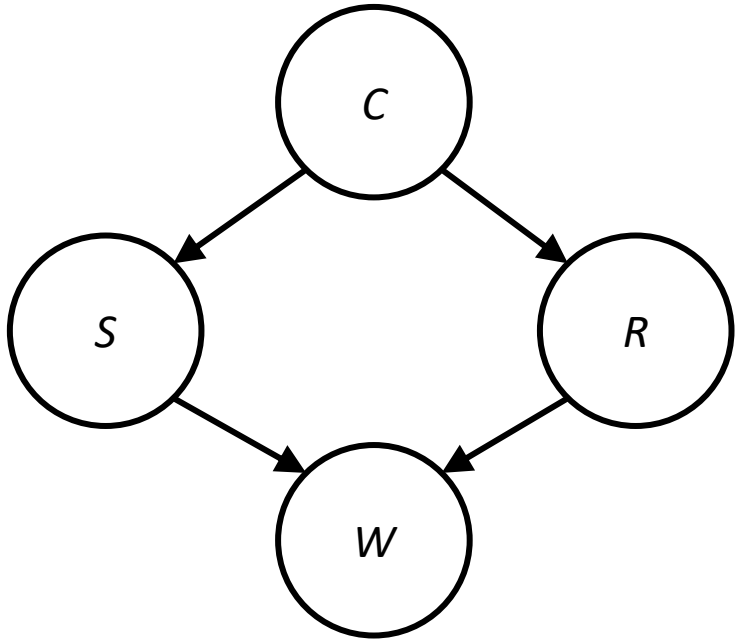
…

Approximate Probabilistic Queries with Samples

- Potential samples from the Bayes Net:

  +c, -s, +r, +w
  +c, +s, +r, +w
  -c, +s, +r, -w
  +c, -s, +r, +w
  -c, -s, -r, +w

- What can we do with these samples?
  - Can empirically estimate the probabilistic queries
- Estimating P(W)
  - We have counts <+w:4, -w:1>
  - Normalize to get P(W) = <+w:0.8, -w:0.2>
- Can estimate other probabilistic queries as well:
  - P(C| +w)? P(C| +r, +w)? P(C| -r, -w)?
  - Note: if some evidence is not observed then we cannot estimate it



Problem: Prior sampling is unaware of the types of probabilistic queries that will be asked later?
Can we be more efficient if we knew the queries from the Bayes Net?

Prior Sampling Convergence

- This process generates samples with probability:

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

- Let the number of samples of an event be $N_{PS}(x_1 \ldots x_n)$

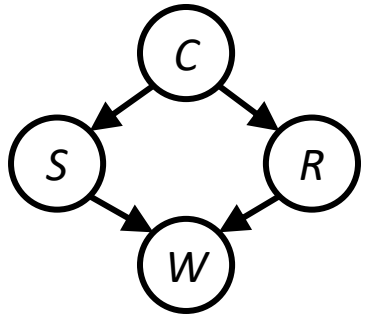
- Then,

$$\begin{aligned} \lim_{N\to\infty} \hat{P}(x_1, \ldots, x_n) &= \lim_{N\to\infty} N_{PS}(x_1, \ldots, x_n)/N \\ &= S_{PS}(x_1, \ldots, x_n) \\ &= P(x_1 \ldots x_n) \end{aligned}$$

- Prior sampling procedure is **consistent**. In the limit, the estimated distribution converges to the correct distribution.

Rejection Sampling

- IN: evidence instantiation
- For i=1, 2, …, n
  - Sample x_i from $P(X_i \mid \text{Parents}(X_i))$
  - **If x_i not consistent with evidence**
    - **Reject: Return, and no sample is generated in this cycle**
- Return $(x_1, x_2, \ldots, x_n)$



Rejection Sampling

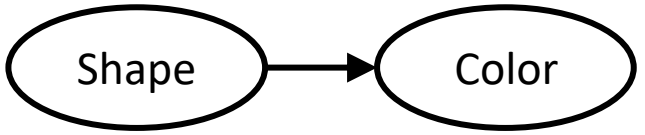
- Estimate P(C| +s)
- Tally the C outcomes, but reject samples which do not have S=+s.
- As you are sampling successively from the conditional probabilities, return if you find a sample inconsistent with the instantiated variables.
- **We can reject earlier, say, if there are 1000 variables and at the third variable we detect an inconsistency, we can reject the entire sample.**
- Rejection sampling is consistent for conditional probabilities in the limit.

+c, -s, +r, +w
+c, +s, +r, +w
-c, +s, +r, -w
+c, -s, +r, +w
-c, -s, -r, +w

Likelihood Weighting

- **Rejection Sampling**
  - Problem: iff the evidence is unlikely, the sampling will reject lots of samples.
  - Wait for a long time before you get a sample that is in agreement with the evidence.
  - Evidence is taken into account **after** you have sampled, not exploited as you sample.

Example: P(Shape|**blue**)

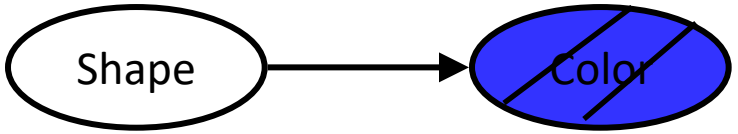


~~pyramid, green~~
~~pyramid, red~~
sphere, **blue**
~~cube, red~~
~~sphere, green~~

Sample and then throw away the inconsistent ones.

- **Likelihood Weighting**
  - Generate only events that are **consistent** with the evidence.
  - **Fix evidence variables** and **sample the rest**
  - **Problem: sample distribution not consistent.**
  - **Solution: weight by probability of evidence given parents.**

Example: P(Shape|**blue**)



pyramid, **blue**
pyramid, **blue**
sphere, **blue**
cube, **blue**
sphere, **blue**

Sample only the consistent values.

Likelihood Weighting

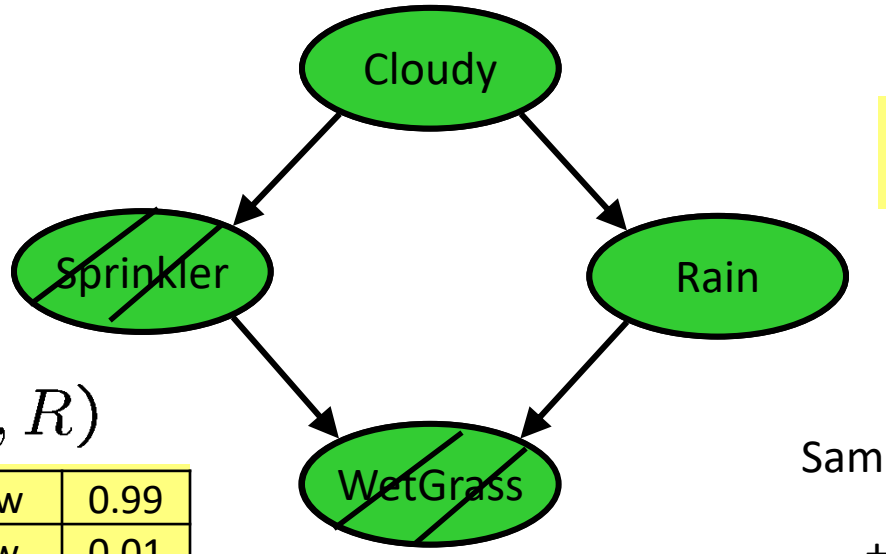
- IN: evidence instantiation
- **w = 1.0**
- for i=1, 2, …, n
  - **if X_i is an evidence variable**
    - **X_i = observation x_i for X_i**
    - **Set w = w * P(x_i | Parents(X_i))**
  - else
    - Sample x_i from P(X_i | Parents(X_i))
- return (x_1, x_2, …, x_n), **w**

$P(C)$

+c	0.5
-c	0.5

$P(S|C)$

+c	+s	0.1
	-s	0.9
-c	+s	0.5
	-s	0.5



$P(R|C)$

+c	+r	0.8
	-r	0.2
-c	+r	0.2
	-r	0.8

$P(W|S,R)$

+s	+r	+w	0.99
		-w	0.01
	-r	+w	0.90
		-w	0.10
-s	+r	+w	0.90
		-w	0.10
	-r	+w	0.01
		-w	0.99

Samples:

+c, +s, +r, +w

…

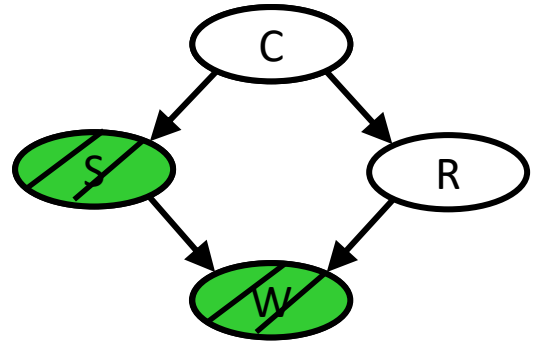
$$w = 1.0 \times 0.1 \times 0.99$$

When an evidence variable adjusted the weight.

Likelihood Weighting

- Sampling distribution if z sampled and e fixed evidence

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



- Now, samples have weights

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

- Together, weighted sampling distribution is consistent

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

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

Weighting corrects the distribution. It also represents the importance of the distribution.

Gibbs Sampling

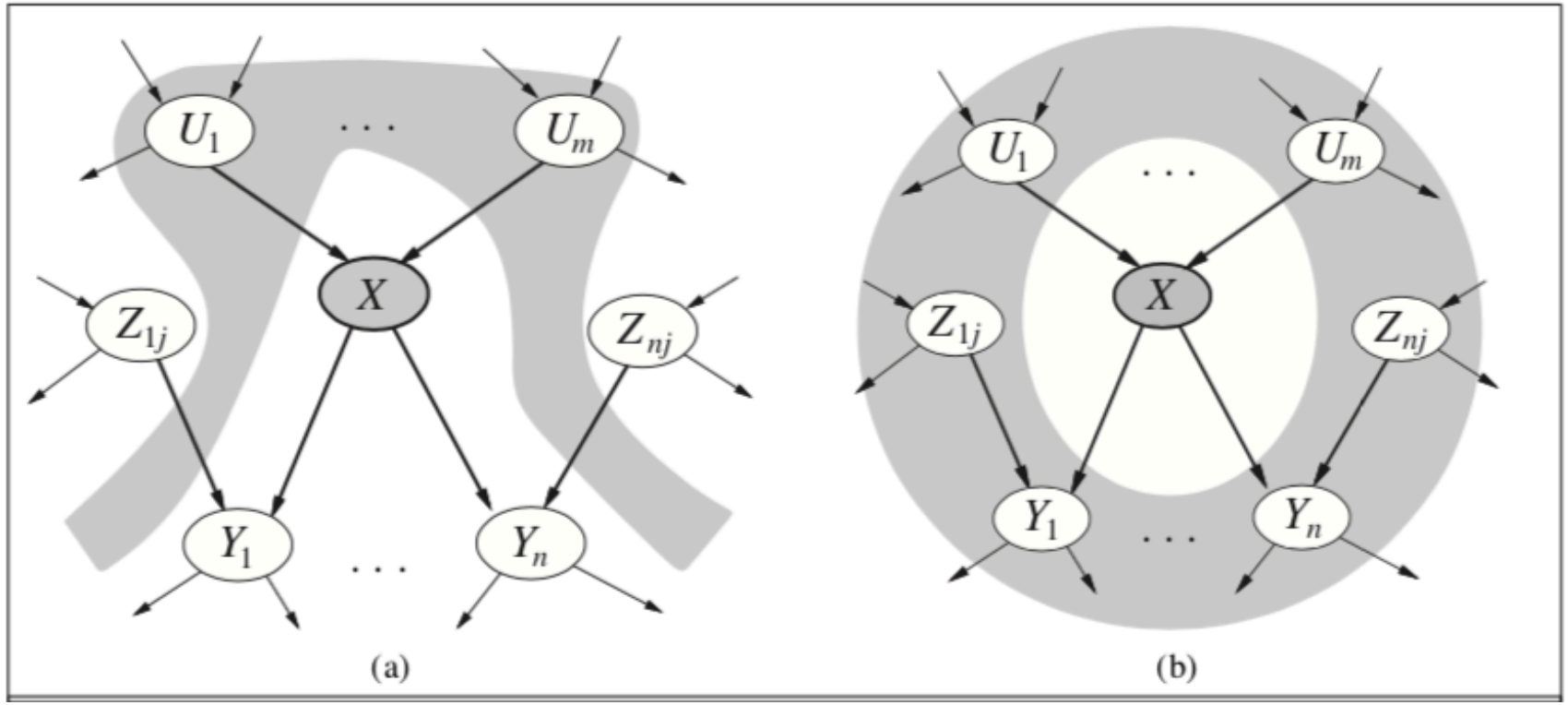
- Likelihood weighting
  - Evidence is taken into account during the sampling process.
  - Problem?
    - Evidence variables influence the choice of down stream variables and not the upstream ones.
    - During the sampling process there may be a low probability evidence encountered later, the sample will look promising for very long.

- Gibbs Sampling
  - Consider evidence when we sample *every* variable (both downstream and upstream)

Gibbs Sampling

- Procedure:
  - Track of a full instantiation $x_1, x_2, \ldots, x_n$.
  - Start with an arbitrary instantiation consistent with the evidence.
  - Sample one variable at a time, conditioned on all the rest, but keep evidence fixed.
  - Keep repeating this for a long time.
  - After repeating you get *one* sample from the distribution.
  - To get more samples: start again.
  - *Note: this is like local search.*
- Property: in the limit of repeating this infinitely many times the resulting sample is coming from the correct distribution.
- Rationale: both upstream and downstream variables condition on evidence.
- Note: Enough to sample from the Markov Blanket.

Conditional Independences and Markov Boundary



(a) A node is conditionally independent of its **non-descendants** given its parents.

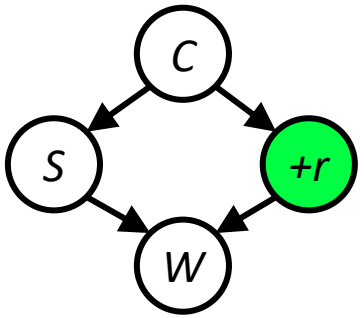
(b) A node is conditionally independent of all other nodes in the network given the **Markov blanket**, i.e., its parents, children and children's parents.

Gibbs Sampling: Example

Estimating P(S | +r)

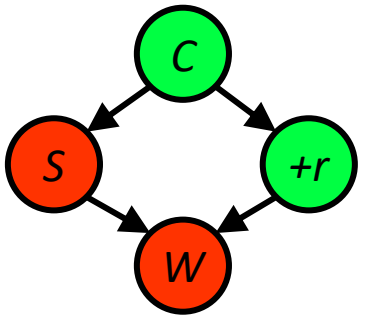
Step 1: Fix evidence

- R = +r



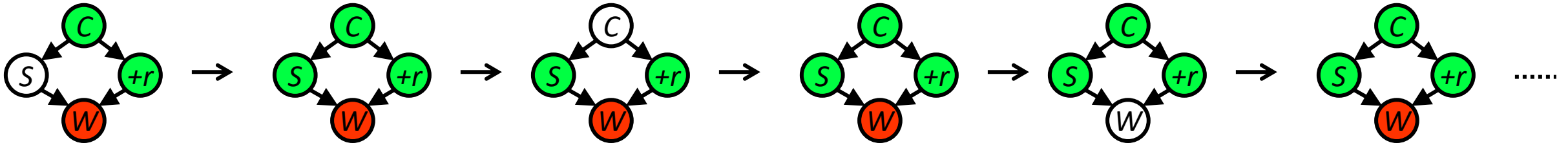
Step 2: Initialize other variables

- Randomly



Steps 3: Repeat

- Randomly select a non-evidence variable X
- Resample X from P(X | all other variables)

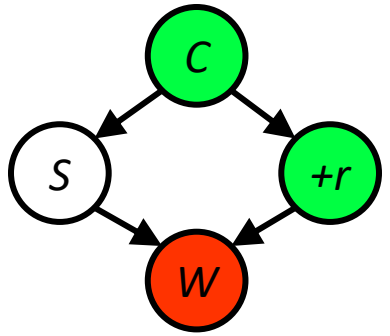


Sample from $P(S|+c,-w,+r)$ Sample from $P(C|+s,-w,+r)$ Sample from $P(W|+s,+c,+r)$

Sampling from the conditional

- Sample from P(S | +c, +r, -w)

$$
\begin{aligned}
P(S|+c,+r,-w) &= \frac{P(S,+c,+r,-w)}{P(+c,+r,-w)} \\
&= \frac{P(S,+c,+r,-w)}{\sum_s P(s,+c,+r,-w)} \\
&= \frac{P(+c)P(S|+c)P(+r|+c)P(-w|S,+r)}{\sum_s P(+c)P(s|+c)P(+r|+c)P(-w|s,+r)} \\
&= \frac{P(+c)P(S|+c)P(+r|+c)P(-w|S,+r)}{P(+c)P(+r|+c)\sum_s P(s|+c)P(-w|s,+r)} \\
&= \frac{P(S|+c)P(-w|S,+r)}{\sum_s P(s|+c)P(-w|s,+r)}
\end{aligned}
$$



Sampling from the conditional distribution is needed as a sub-routine for Gibbs sampling. It is typically easier to sample from. The expression is simpler due to instantiated variables, can even construct the probability table if needed.

Application: Topic Modeling

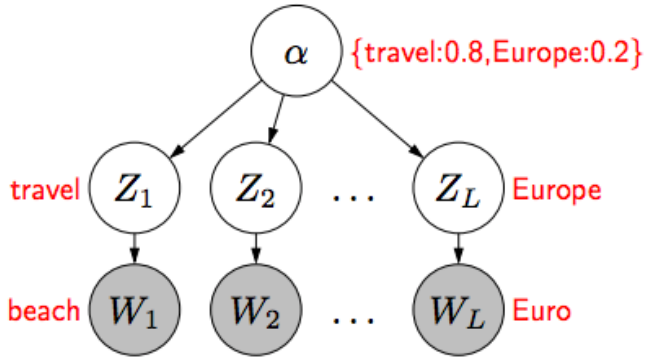
Topic Modeling

Generate a distribution over topics $\alpha \in \mathbb{R}^K$

For each position $i = 1, \ldots, L$:

Generate a topic $Z_i \sim p(Z_i \mid \alpha)$

Generate a word $W_i \sim p(W_i \mid Z_i)$



https://cacm.acm.org/magazines/2012/4/147361-probabilistic-topic-models/fulltext

Application: Topic Modeling

Gibbs Sampling and important technique for this application

Figure 1. The intuitions behind latent Dirichlet allocation. We assume that some number of "topics," which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.

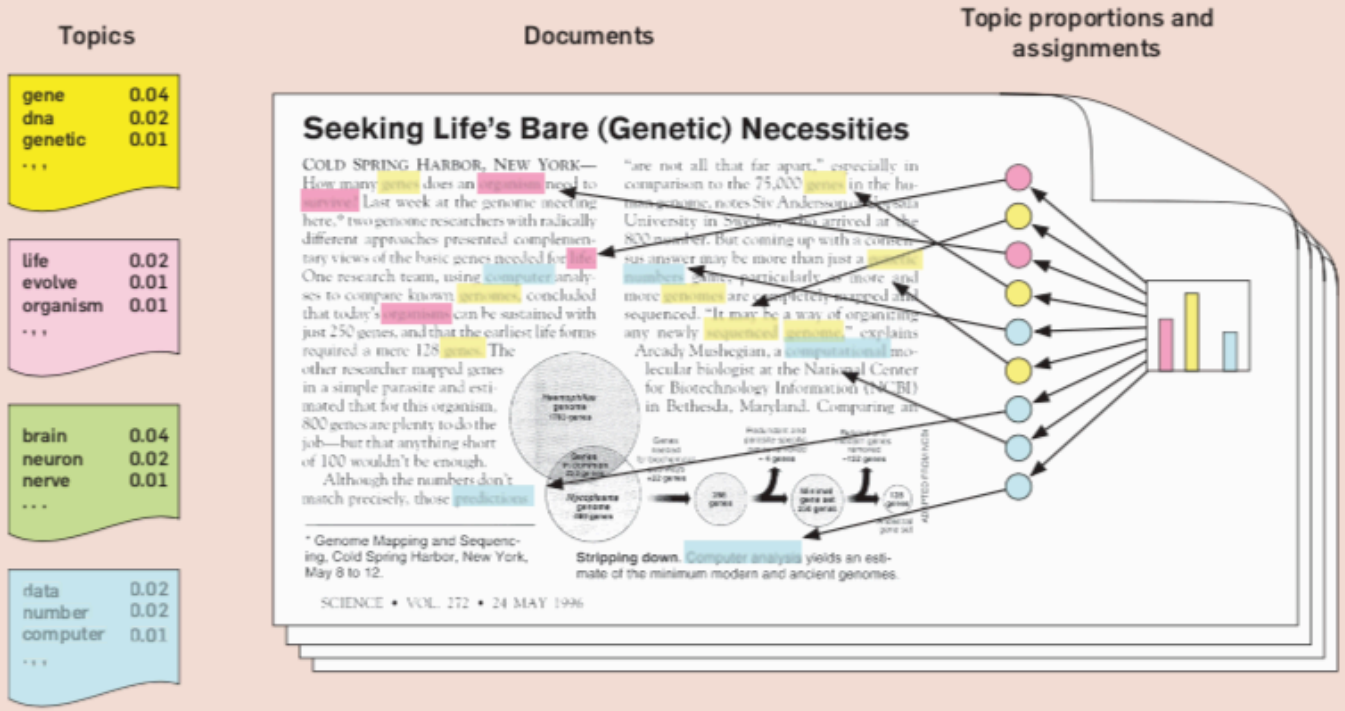
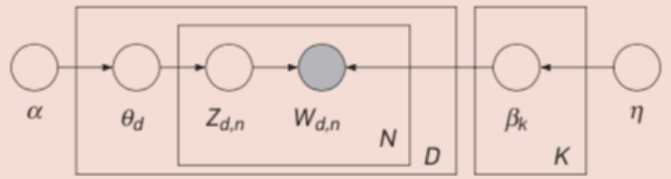


Figure 4. The graphical model for latent Dirichlet allocation. Each node is a random variable and is labeled according to its role in the generative process (see Figure 1). The hidden nodes—the topic proportions, assignments, and topics—are unshaded. The observed nodes—the words of the documents—are shaded. The rectangles are "plate" notation, which denotes replication. The *N* plate denotes the collection words within documents; the *D* plate denotes the collection of documents within the collection.



"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

https://cacm.acm.org/magazines/2012/4/147361-probabilistic-topic-models/fulltext