COL333/671: Introduction to AI

Semester I, 2024-25

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.

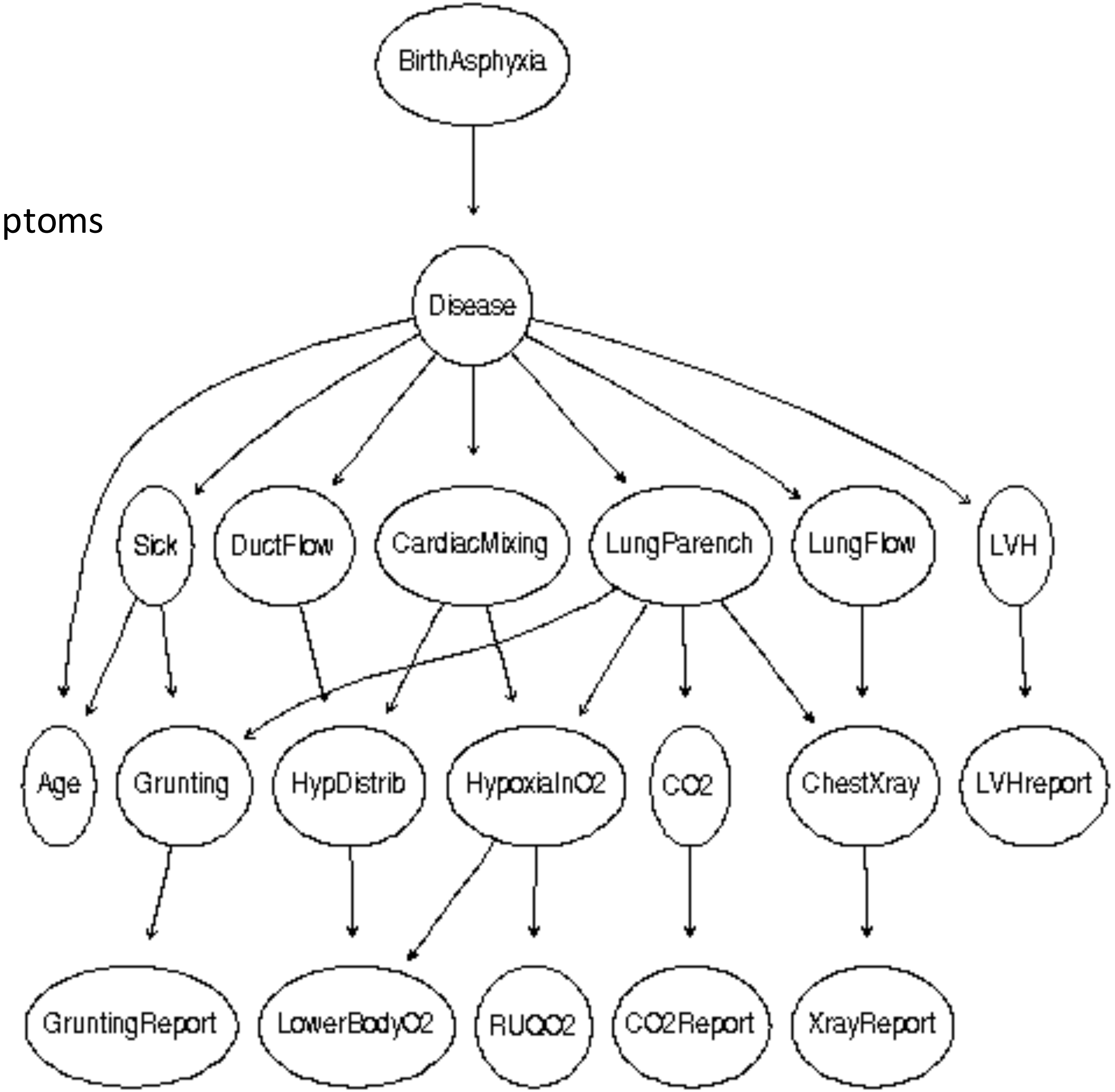
I hear an unusual sound and a burning smell in my car, what fault is there in my engine?

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

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

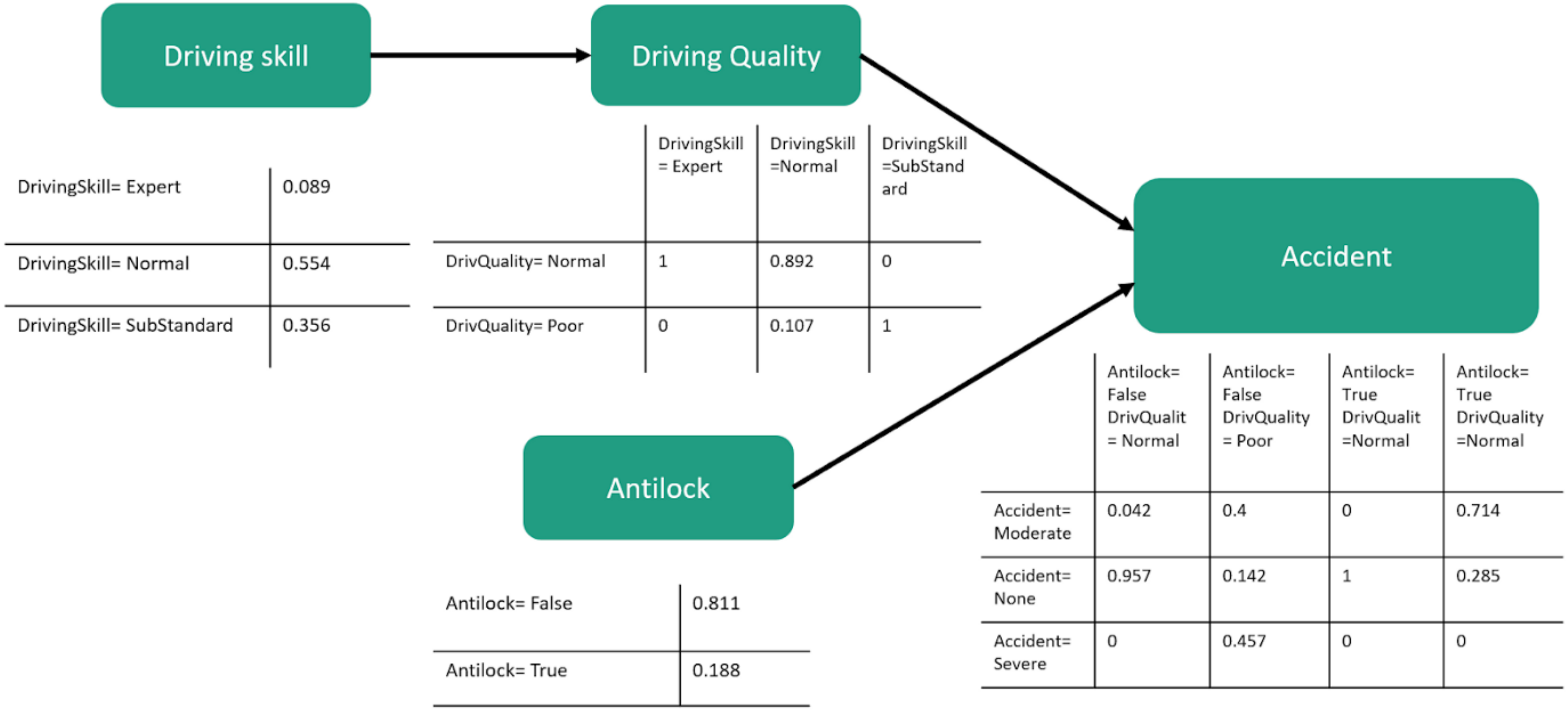
Examples

Inferring disease from symptoms



Examples

Accident – driving domain.



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

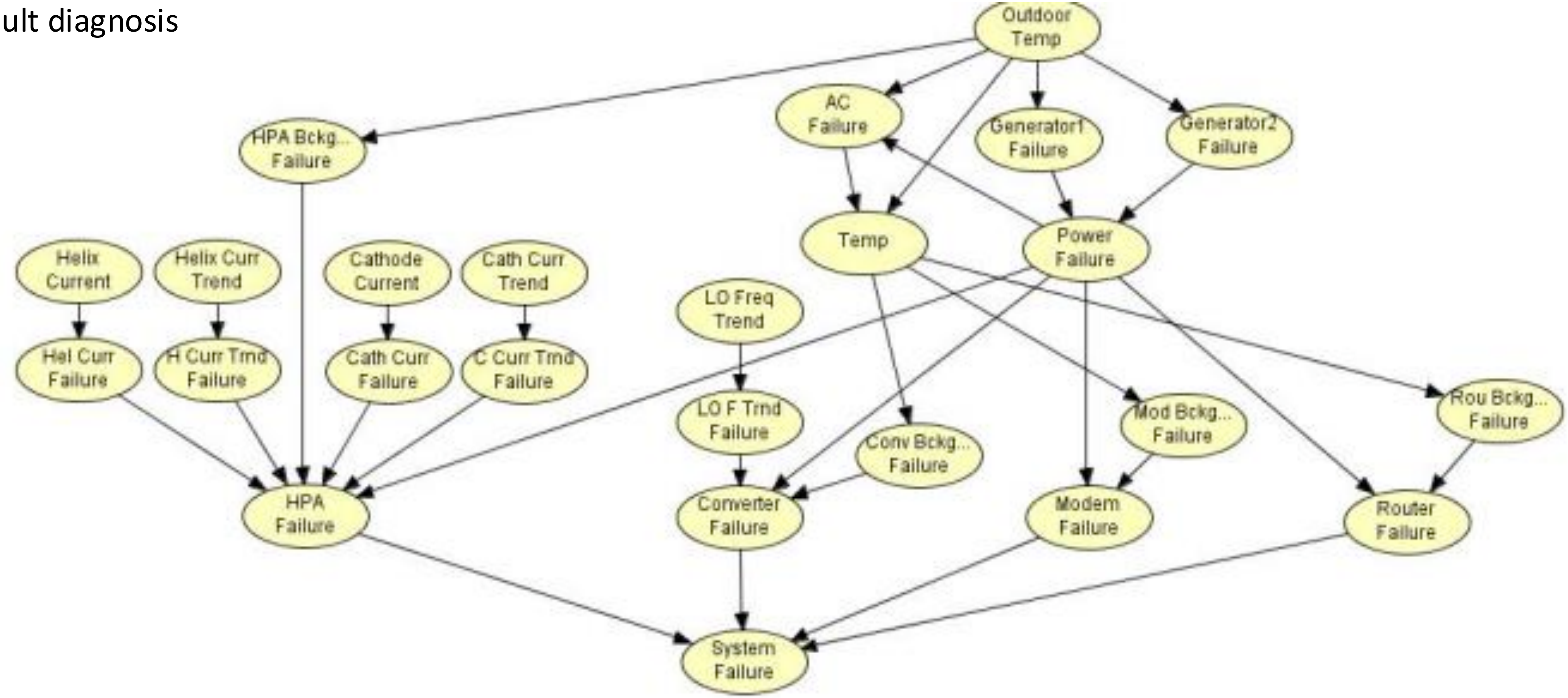
	DrivingSkill = Expert	DrivingSkill =Normal	DrivingSkill =SubStandard
DrivQuality= Normal	1	0.892	0
DrivQuality= Poor	0	0.107	1

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

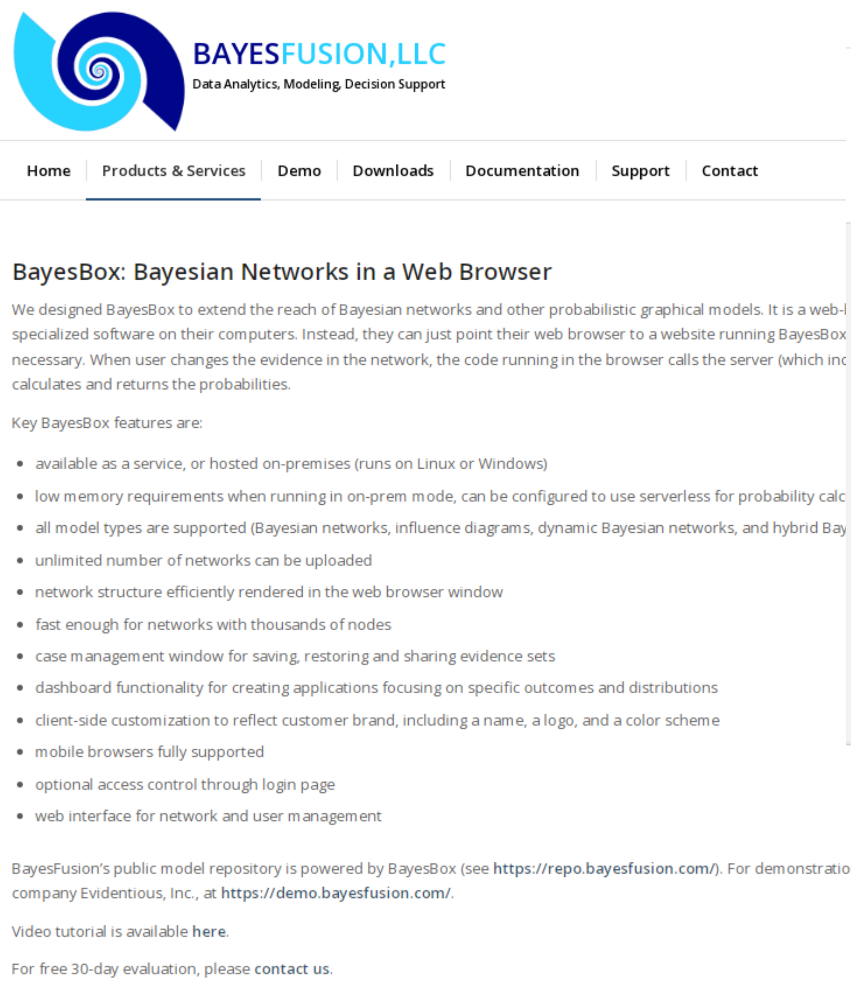
Examples

Fault diagnosis



Examples

Predictive analytics/expert systems



BAYESFUSION,LLC
Data Analytics, Modeling, Decision Support

BayesBox: Bayesian Networks in a Web Browser

We designed BayesBox to extend the reach of Bayesian networks and other probabilistic graphical models. It is a web-l specialized software on their computers. Instead, they can just point their web browser to a website running BayesBox necessary. When user changes the evidence in the network, the code running in the browser calls the server (which inc calculates and returns the probabilities.

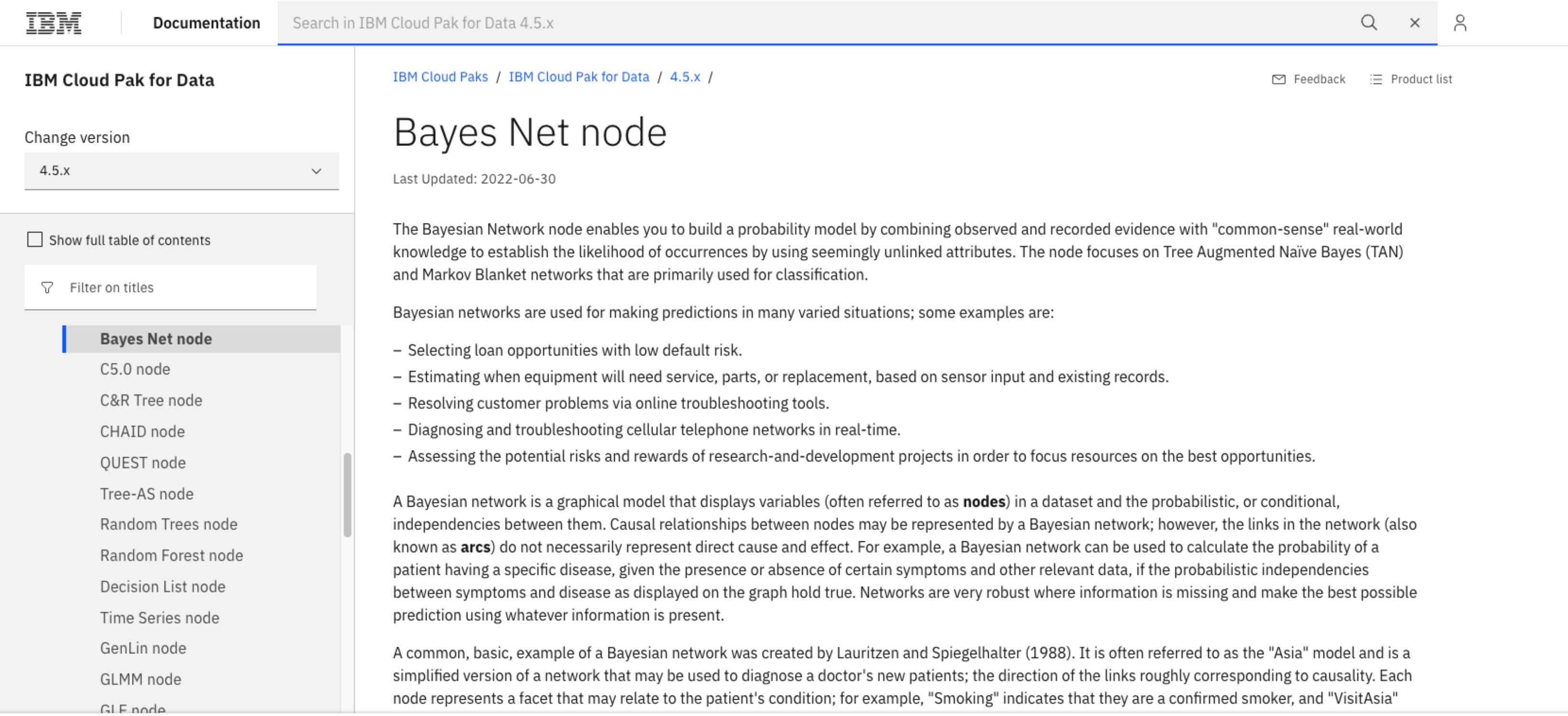
Key BayesBox features are:

- available as a service, or hosted on-premises (runs on Linux or Windows)
- low memory requirements when running in on-prem mode, can be configured to use serverless for probability calc
- all model types are supported (Bayesian networks, influence diagrams, dynamic Bayesian networks, and hybrid Bay
- unlimited number of networks can be uploaded
- network structure efficiently rendered in the web browser window
- fast enough for networks with thousands of nodes
- case management window for saving, restoring and sharing evidence sets
- dashboard functionality for creating applications focusing on specific outcomes and distributions
- client-side customization to reflect customer brand, including a name, a logo, and a color scheme
- mobile browsers fully supported
- optional access control through login page
- web interface for network and user management

BayesFusion's public model repository is powered by BayesBox (see https://repo.bayesfusion.com/). For demonstration purposes, we have also created a BayesBox-based web site of a fictitious company Evidentious, Inc., at https://demo.bayesfusion.com/.

Video tutorial is available here.

For free 30-day evaluation, please contact us.



Bayes Net node

Last Updated: 2022-06-30

The Bayesian Network node enables you to build a probability model by combining observed and recorded evidence with "common-sense" real-world knowledge to establish the likelihood of occurrences by using seemingly unlinked attributes. The node focuses on Tree Augmented Naïve Bayes (TAN) and Markov Blanket networks that are primarily used for classification.

Bayesian networks are used for making predictions in many varied situations; some examples are:

- Selecting loan opportunities with low default risk.
- Estimating when equipment will need service, parts, or replacement, based on sensor input and existing records.
- Resolving customer problems via online troubleshooting tools.
- Diagnosing and troubleshooting cellular telephone networks in real-time.
- Assessing the potential risks and rewards of research-and-development projects in order to focus resources on the best opportunities.

A Bayesian network is a graphical model that displays variables (often referred to as **nodes**) in a dataset and the probabilistic, or conditional, independencies between them. Causal relationships between nodes may be represented by a Bayesian network; however, the links in the network (also known as **arcs**) do not necessarily represent direct cause and effect. For example, a Bayesian network can be used to calculate the probability of a patient having a specific disease, given the presence or absence of certain symptoms and other relevant data, if the probabilistic independencies between symptoms and disease as displayed on the graph hold true. Networks are very robust where information is missing and make the best possible prediction using whatever information is present.

A common, basic, example of a Bayesian network was created by Lauritzen and Spiegelhalter (1988). It is often referred to as the "Asia" model and is a simplified version of a network that may be used to diagnose a doctor's new patients; the direction of the links roughly corresponding to causality. Each node represents a facet that may relate to the patient's condition; for example, "Smoking" indicates that they are a confirmed smoker, and "VisitAsia"

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

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

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

$P(T)$

T	P
hot	0.5
cold	0.5

$P(W)$

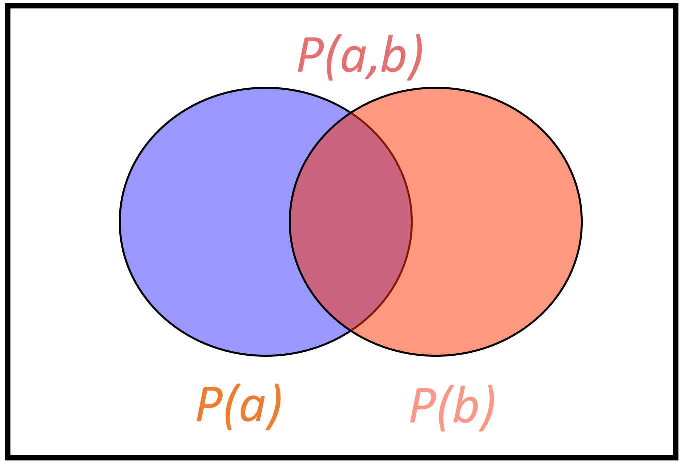
W	P
sun	0.6
rain	0.4

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

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

$$\overbrace{P(W|T = cold) \qquad P(W|T = hot)}^{P(W|T)}$$

$P(W|T = cold)$

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.

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

- Example:

$P(W)$

R	P
sun	0.8
rain	0.2

$P(D|W)$

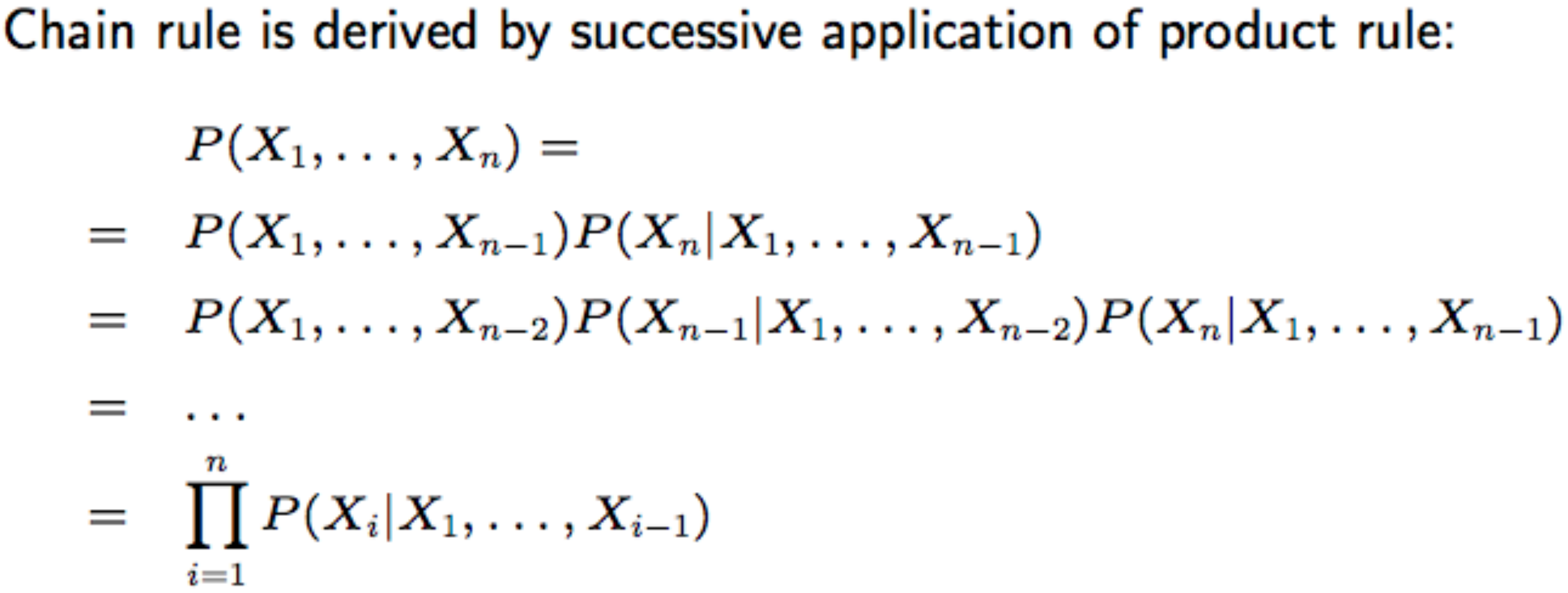
D	W	P
wet	sun	0.1
dry	sun	0.9
wet	rain	0.7
dry	rain	0.3

$\Longleftrightarrow$

$P(D,W)$

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.

n smaller distributions

$P(X_1)$

H	0.5
T	0.5

$P(X_2)$

H	0.5
T	0.5

$\vdots$

$P(X_n)$

H	0.5
T	0.5

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

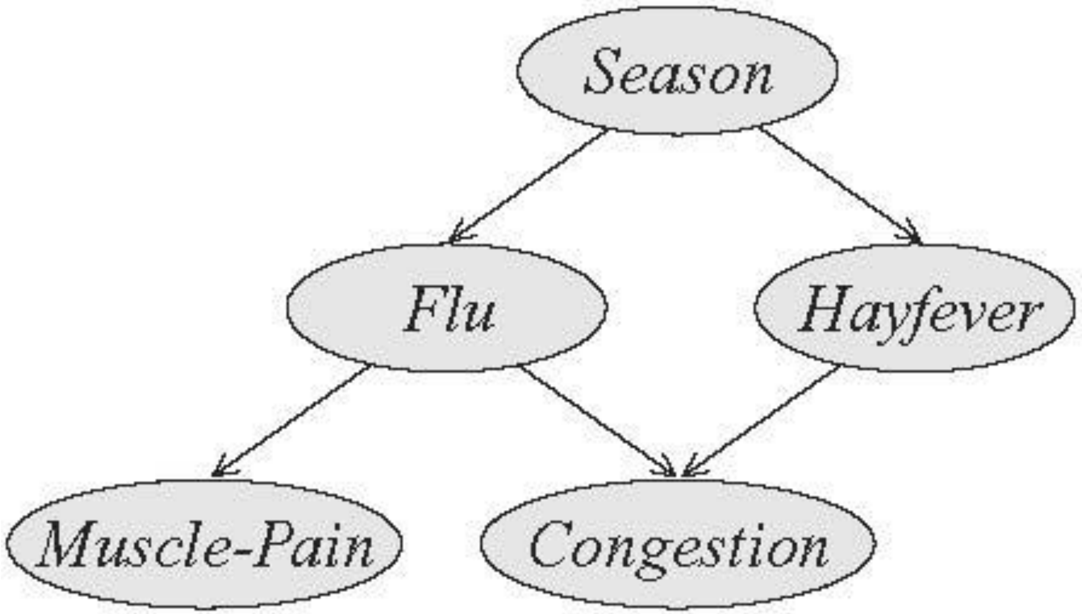
2^n

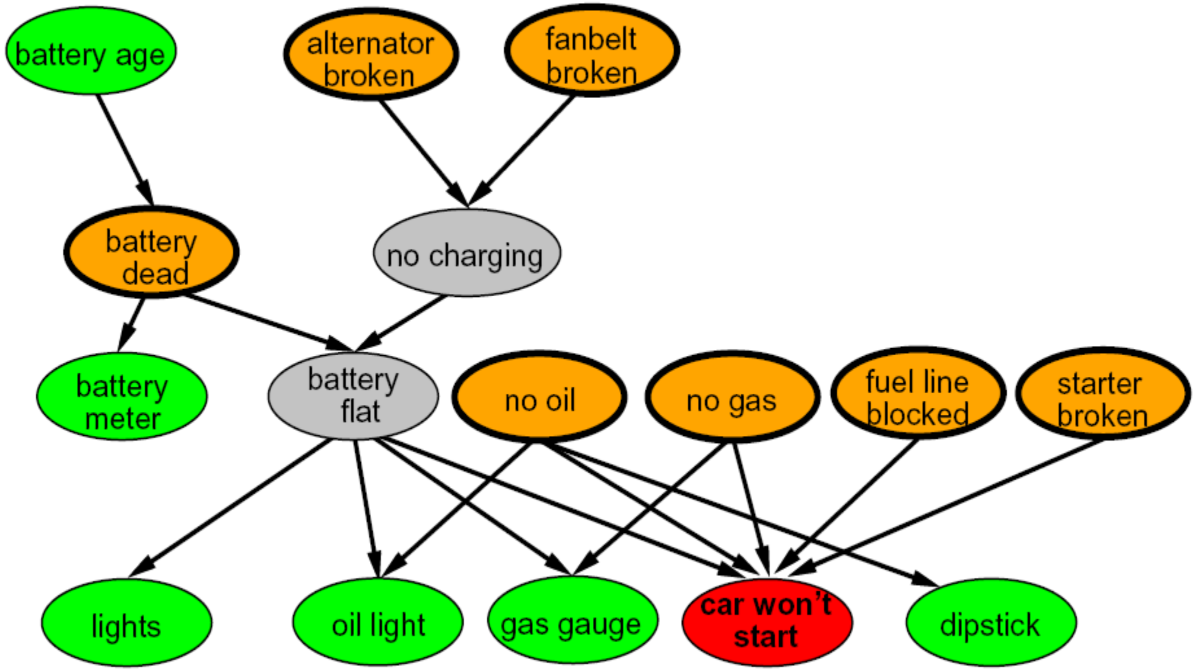
Bayesian Networks

- 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





Online tool: https://detsutut.shinyapps.io/shinyDBNet/

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

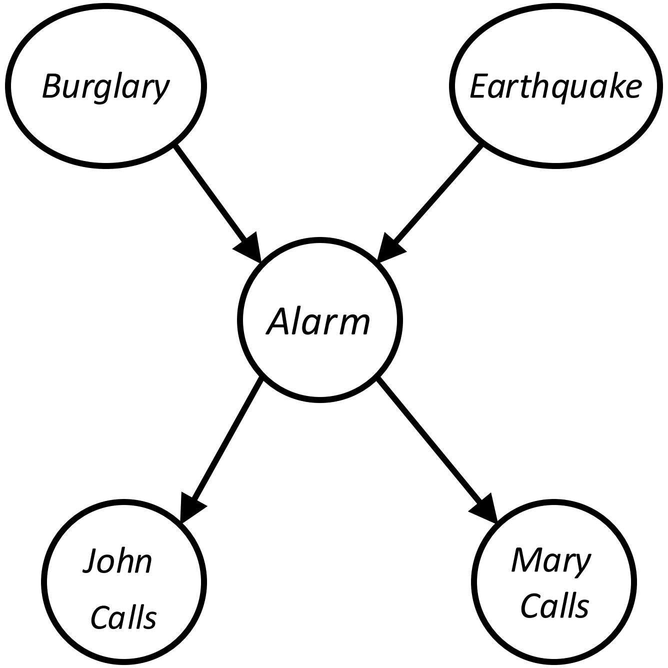
$$P(X|a_1 \ldots a_n)$$

- 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:

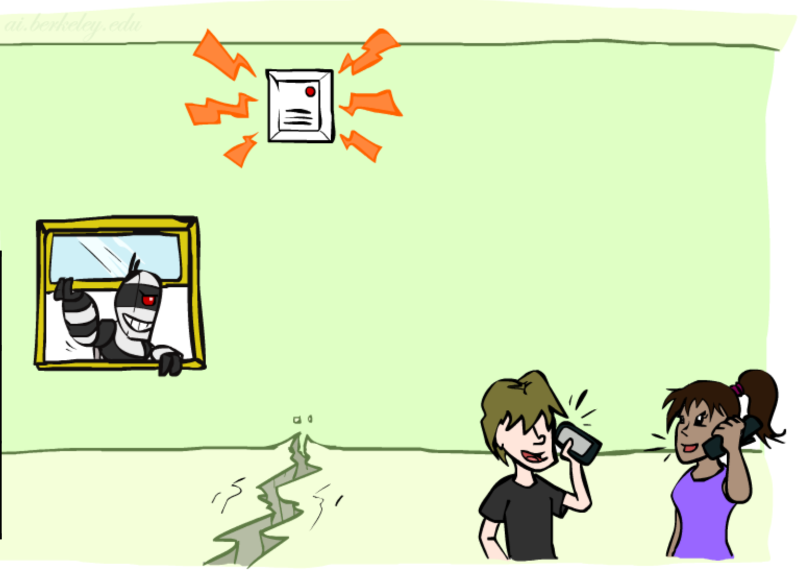
$$P(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} P(x_i|\mathit{parents}(X_i))$$

Example: The Alarm Network

B	P(B)
+b	0.001
-b	0.999



E	P(E)
+e	0.002
-e	0.998

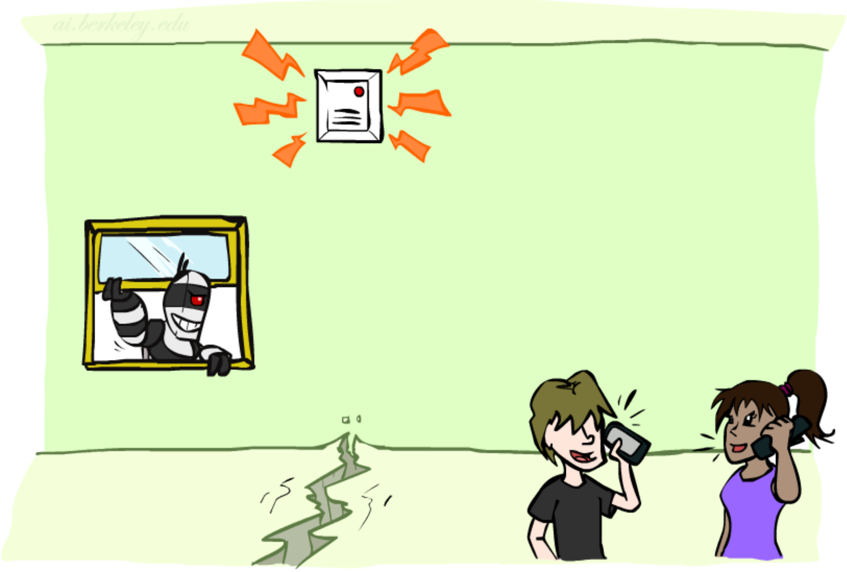


B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

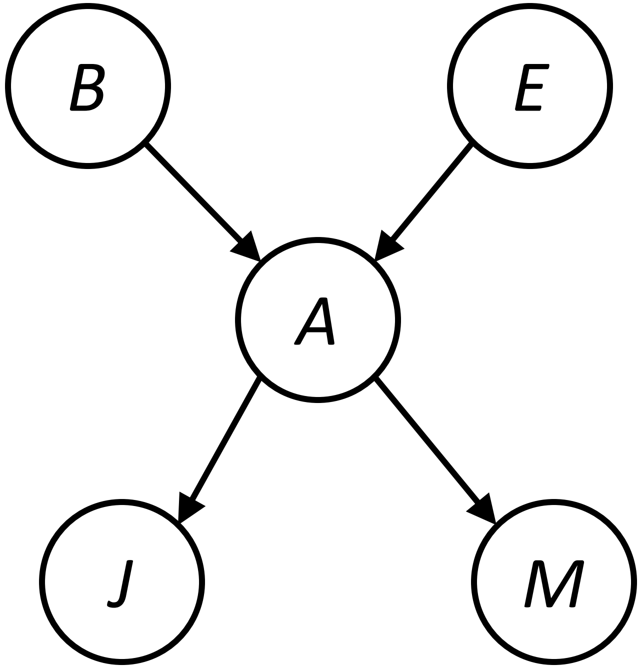
A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

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

Example: 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

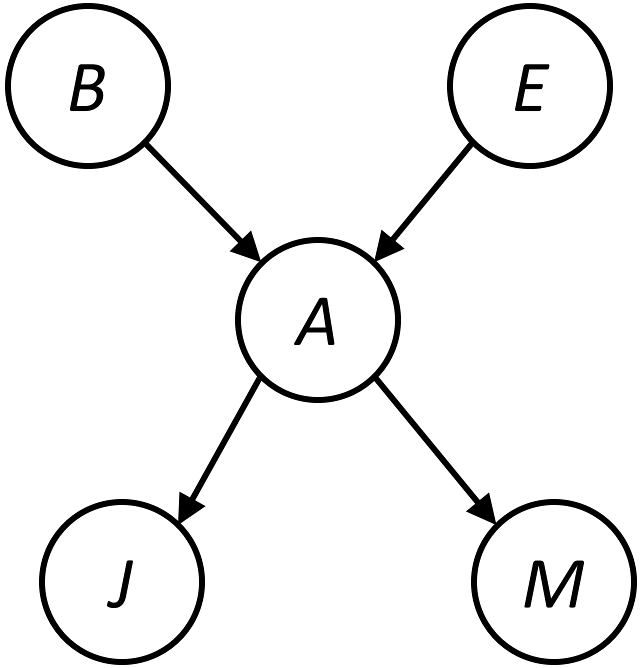
A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

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

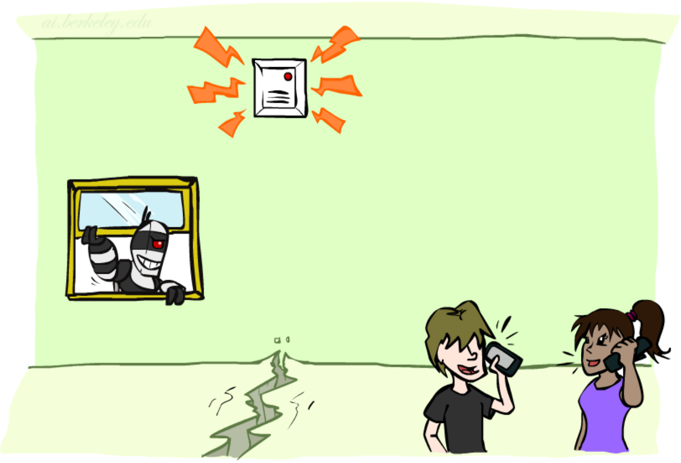
$$P(+b, -e, +a, -j, +m) =$$

Estimating likelihood of variables

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

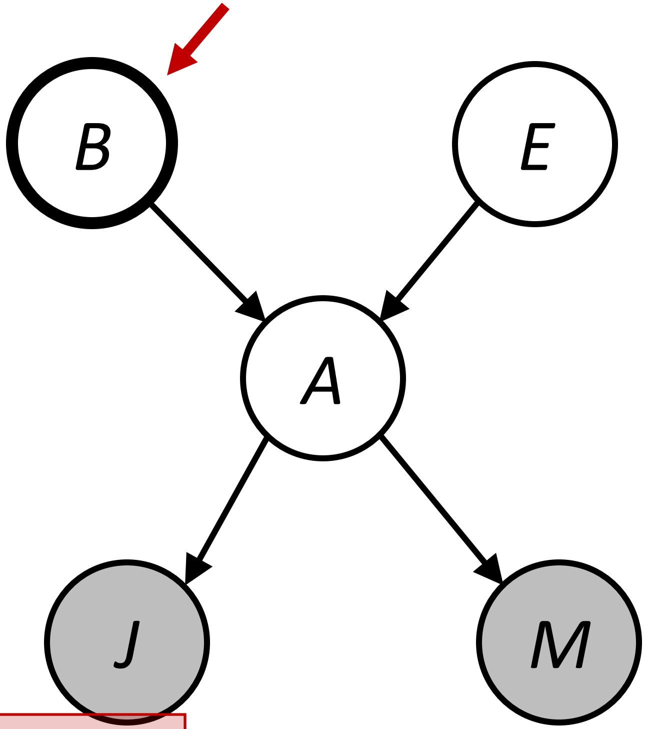
$$P(+b,-e,+a,-j,+m) =$$

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

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

Answering a general probabilistic query

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



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

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

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

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

Bayesian Networks: Inference

- **Bayesian Networks**
  - Implicitly encode a probability distribution
  - As a product of local conditional distributions

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

- **Variables**
  - Query variables
  - Evidence variables
  - Hidden variables

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

- **Inference: What we want to estimate?**
  - Estimating some useful quantity from the joint distribution.
  - Posterior probability
  - Most likely explanation

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

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

Inference by Enumeration: A way of answering probabilistic queries

- **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.

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

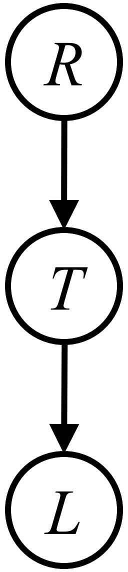
$$P(Q, e_1 \dots e_k) = \sum_{h_1 \dots h_r} P(Q, h_1 \dots h_r, e_1 \dots e_k)$$

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

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

Inference by Enumeration: Example

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



$$P(R)$$

+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

$$P(L) = ?$$

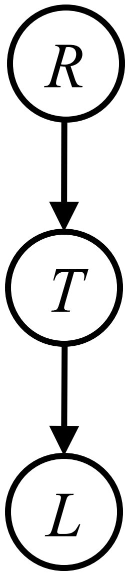
$$= \sum_{r,t} P(r, t, L)$$

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

Inference by Enumeration as Operations on Factors

- **Factors**
  - A factor is a function from some set of variables into a specific value.
  - Initial factos
    - Conditional probability tables (one per node)
    - Select the values consistent with the evidence
- **Inference by Enumeration**
  - Via factors, can be understood as a procedure that joins all the factors and then sums out all the hidden variables.
  - *Define two operations "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

If some variables are observed then apply that information to associated factor. Others are not affected.

$\boxed{L = +\ell}$

$P(R)$

+r	0.1
-r	0.9

$P(T|R)$

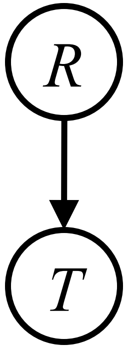
+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

$P(+\ell|T)$

+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*

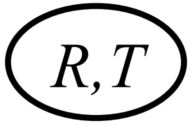


$$P(R) \quad \times \quad P(T|R) \quad \rightarrow \quad P(R,T)$$

+r	0.1
-r	0.9

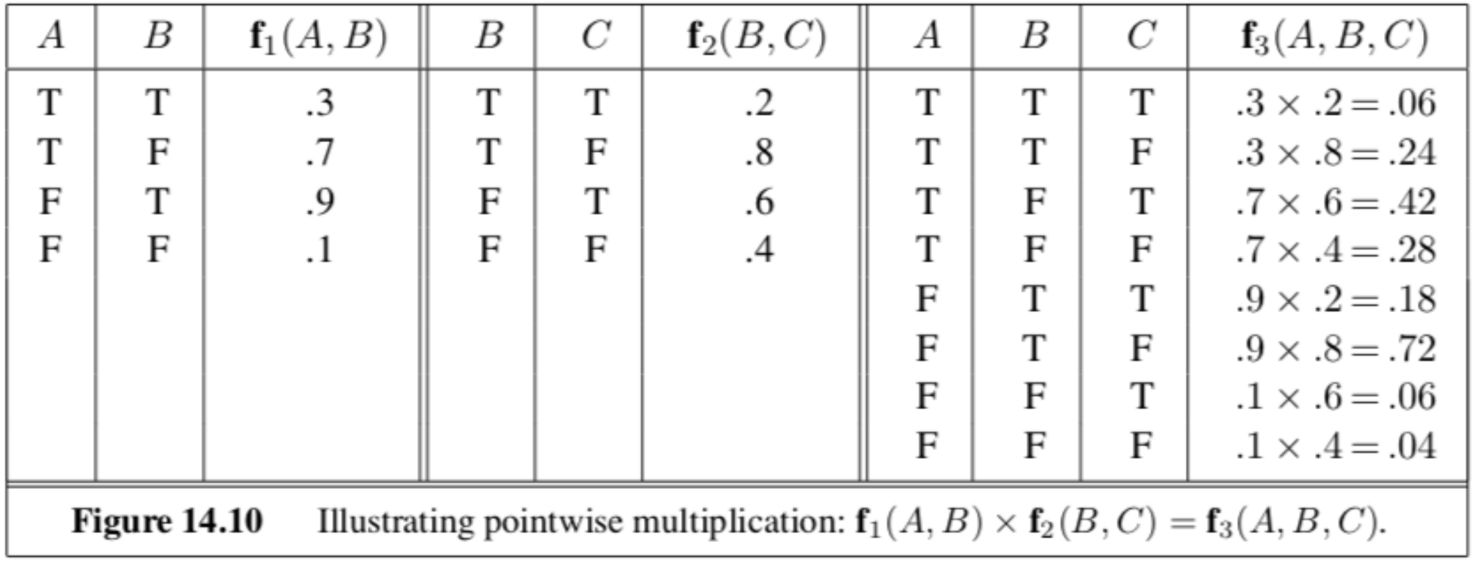
+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

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



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

Joining Factors



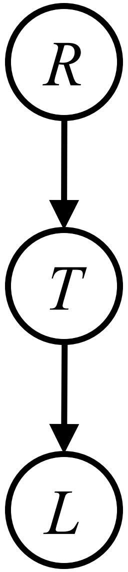
A	B	$\mathbf{f}_1(A,B)$	B	C	$\mathbf{f}_2(B,C)$	A	B	C	$\mathbf{f}_3(A,B,C)$
T	T	.3	T	T	.2	T	T	T	$.3 \times .2 = .06$
T	F	.7	T	F	.8	T	T	F	$.3 \times .8 = .24$
F	T	.9	F	T	.6	T	F	T	$.7 \times .6 = .42$
F	F	.1	F	F	.4	T	F	F	$.7 \times .4 = .28$
						F	T	T	$.9 \times .2 = .18$
						F	T	F	$.9 \times .8 =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

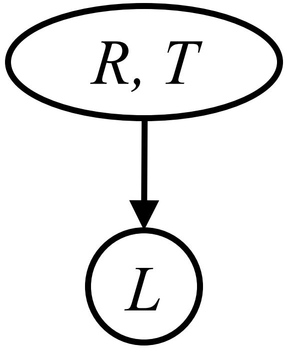
Join R

$P(R,T)$

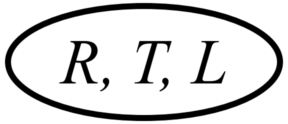
+r	+t	0.08
+r	-t	0.02
-r	+t	0.09
-r	-t	0.81

$P(L|T)$

+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	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 →

$P(T,L)$

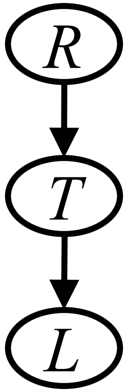
+t	+l	0.051
+t	-l	0.119
-t	+l	0.083
-t	-l	0.747

Sum out T →

$P(L)$

+l	0.134
-l	0.866

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

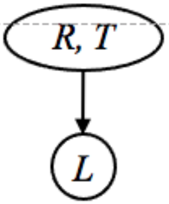
Join R

$P(R,T)$

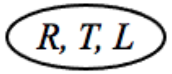
+r	+t	0.08
+r	-t	0.02
-r	+t	0.09
-r	-t	0.81

$P(L|T)$

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

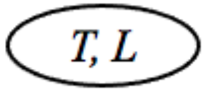


Join T



$P(R,T,L)$

+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

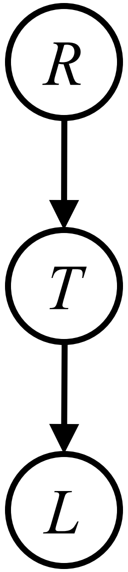
Variable Elimination

- **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



$$P(L) = ?$$

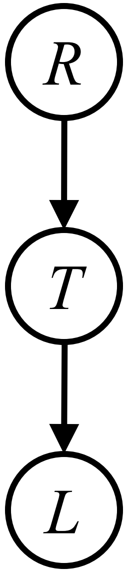
Inference by Enumeration

$$= \underbrace{\sum_t \underbrace{\sum_r \underbrace{P(L|t)\underbrace{P(r)P(t|r)}_{\text{Join on r}}}_{\text{Join on t}}}_{\text{Eliminate r}}}_{\text{Eliminate t}}$$

Variable Elimination

$$= \underbrace{\sum_t \underbrace{P(L|t)\underbrace{\sum_r \underbrace{P(r)P(t|r)}_{\text{Join on r}}}_{\text{Eliminate r}}}_{\text{Join on t}}}_{\text{Eliminate t}}$$

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

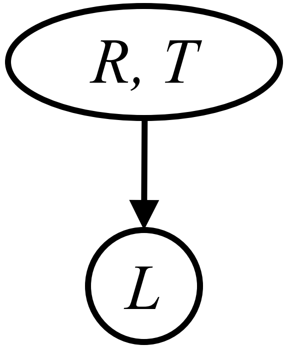
$P(L|T)$

+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



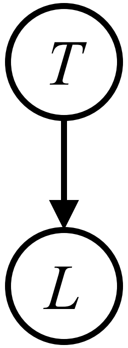
$P(L|T)$

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

Sum out R

$P(T)$

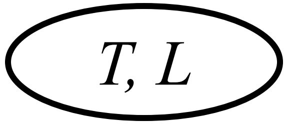
+t	0.17
-t	0.83



$P(L|T)$

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

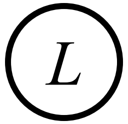
Join T



$P(T, L)$

+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

Variable Elimination (VE)

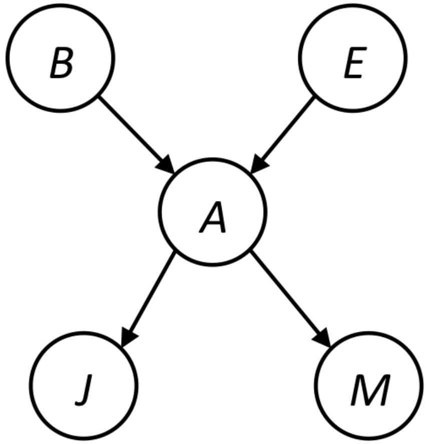
- **Query:** $P(Q|E_1 = e_1, \ldots E_k = e_k)$

- **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**

Variable Elimination: Alarm Domain



Model

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

Factors

$$\boxed{P(B) \qquad P(E) \qquad P(A|B,E) \qquad P(j|A) \qquad P(m|A)}$$

$$\begin{aligned}
P(B|j,m) &\propto P(B,j,m) \\
&= \sum_{e,a} P(B,j,m,e,a) \\
&= \sum_{e,a} P(B)P(e)P(a|B,e)P(j|a)P(m|a) \\
&= \sum_{e} P(B)P(e) \sum_{a} P(a|B,e)P(j|a)P(m|a) \\
&= \sum_{e} P(B)P(e)f_1(j,m|B,e) \\
&= P(B) \sum_{e} P(e)f_1(j,m|B,e) \\
&= P(B)f_2(j,m|B)
\end{aligned}$$

marginal can be obtained from joint by summing out

use Bayes' net joint distribution expression

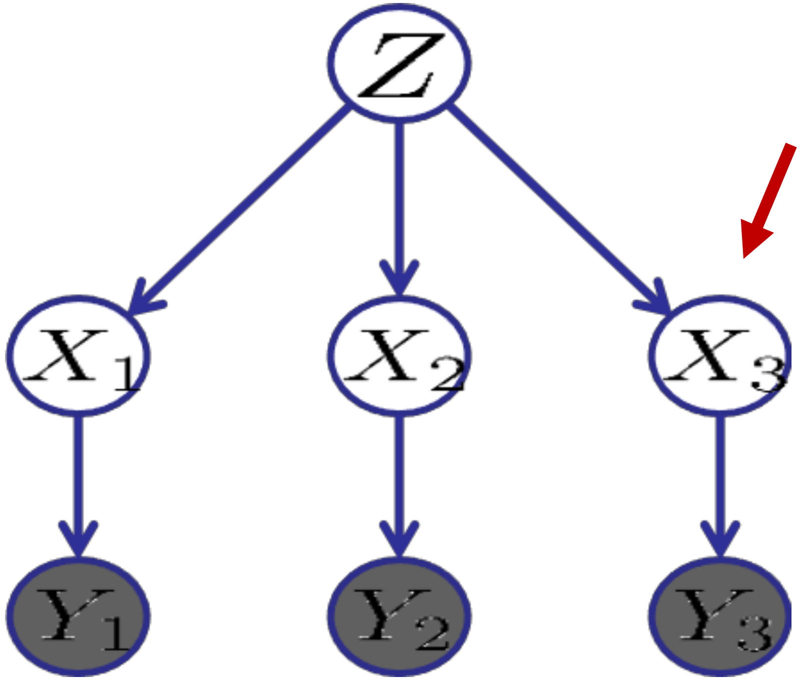
use x*(y+z) = xy + xz

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

Variable Elimination: Complexity



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:

$$P(Z), \boxed{P(X_1|Z),} P(X_2|Z), P(X_3|Z), \boxed{P(y_1|X_1),} P(y_2|X_2), P(y_3|X_3)$$

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:

$$P(Z), \boxed{P(X_2|Z),} P(X_3|Z), \boxed{P(y_2|X_2),} P(y_3|X_3), \boxed{f_1(y_1|Z)}$$

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:

$$\boxed{P(Z), P(X_3|Z),} P(y_3|X_3), \boxed{f_1(y_1|Z), \boxed{f_2(y_2|Z)}}$$

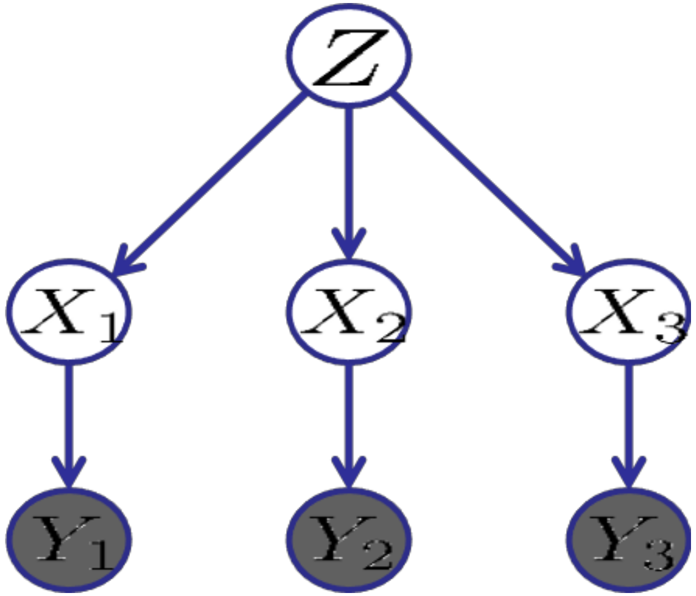
Eliminate Z, this introduces the factor $f_3(y_1, y_2, X_3) = \sum_z P(z)P(X_3|z)f_1(y_1|Z)f_2(y_2|Z)$, and we are left with:

$$P(y_3|X_3), \boxed{f_3(y_1, y_2, X_3)}$$

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



Variable Elimination: Efficient way to re-use factor computation.

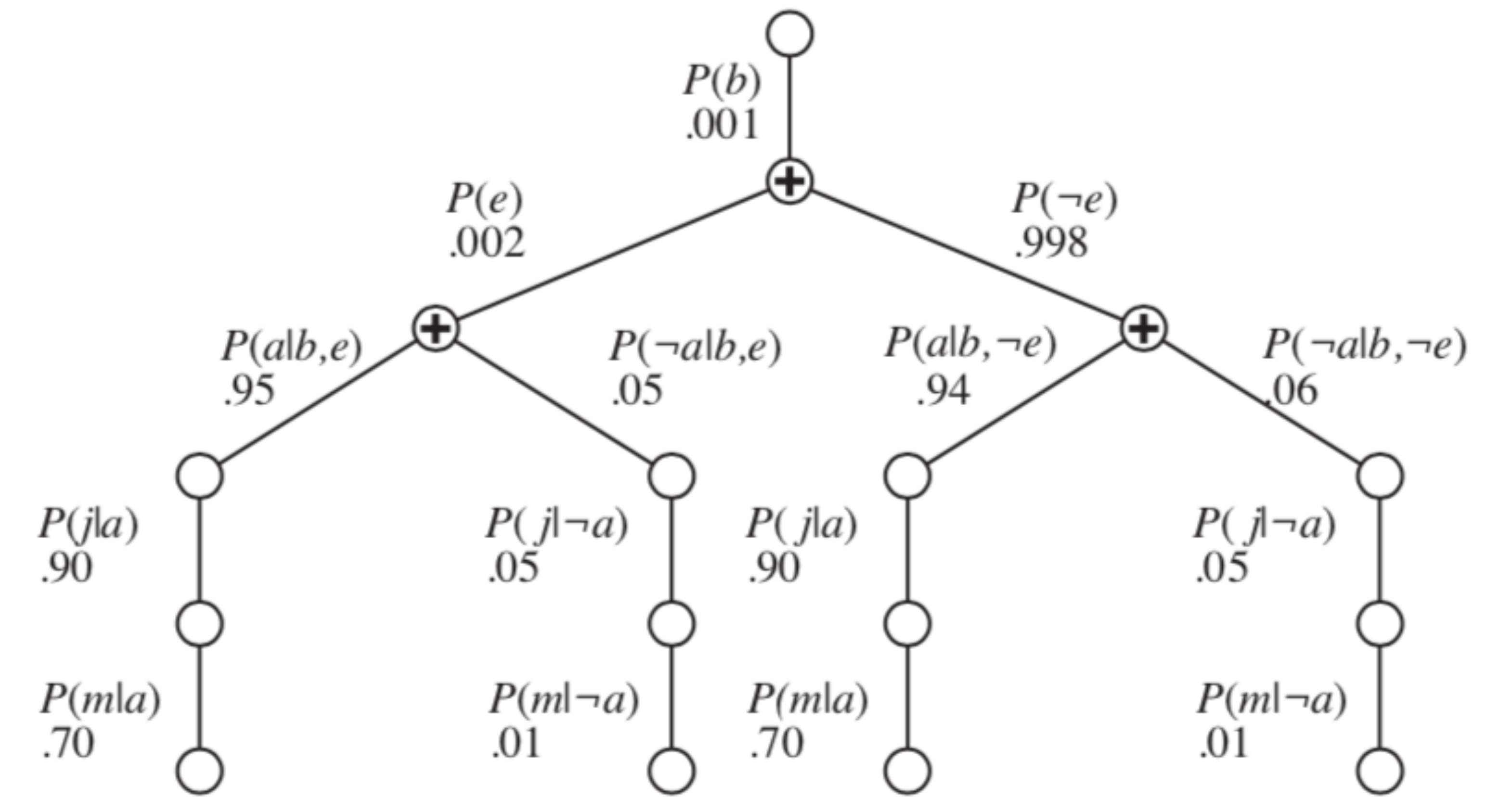


Figure 14.8 The structure of the expression shown in Equation (14.4). The evaluation proceeds top down, multiplying values along each path and summing at the "+" nodes. Notice the repetition of the paths for j and m.

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:

$$P(Z), \boxed{P(X_1|Z),} P(X_2|Z), P(X_3|Z), \boxed{P(y_1|X_1),} P(y_2|X_2), P(y_3|X_3)$$

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:

$$P(Z), \boxed{P(X_2|Z),} P(X_3|Z), \boxed{P(y_2|X_2),} P(y_3|X_3), \boxed{f_1(y_1|Z)}$$

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:

$$\boxed{P(Z), P(X_3|Z),} P(y_3|X_3), \boxed{f_1(y_1|Z), \boxed{f_2(y_2|Z)}}$$

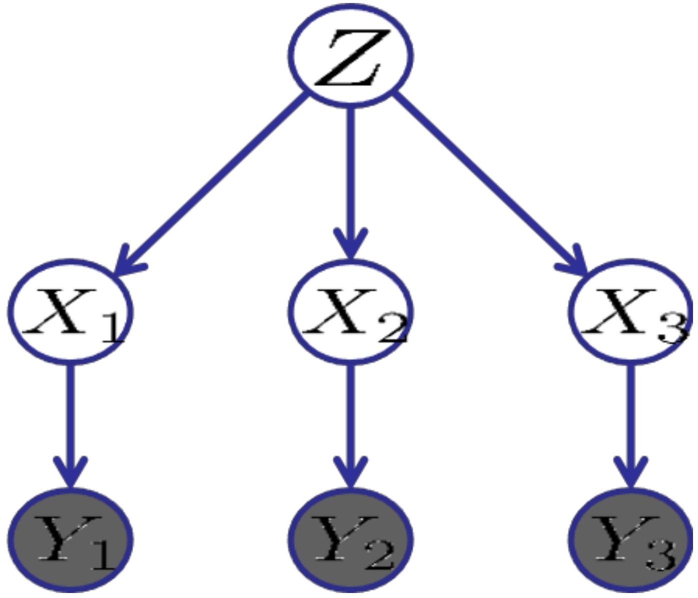
Eliminate Z, this introduces the factor $f_3(y_1, y_2, X_3) = \sum_z P(z)P(X_3|z)f_1(y_1|Z)f_2(y_2|Z)$, and we are left with:

$$P(y_3|X_3), \boxed{f_3(y_1, y_2, X_3)}$$

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

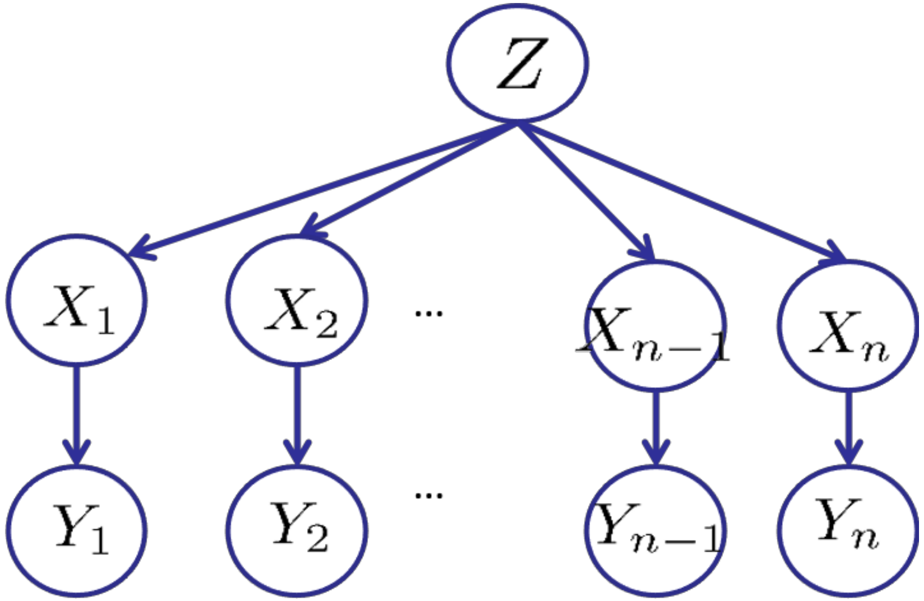


Computational complexity

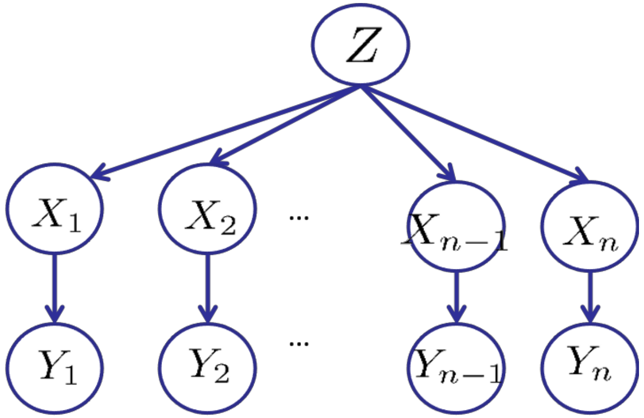
- **Depends on the largest factor generated in VE.**
- Factor size = number of entries in the table.
- In this example: each factor is of size 2 (only one variable). Note that y is observed.
- X_1, X_2, Z, X_3

How does variable ordering affect VE complexity?

- For the query P(X_n|y_1,…,y_n)
- There are n variables Z, X_1, …, $X_{n-1}$to eliminate.
- We need a way to order them (for joining and eliminating)
- Consider two different orderings as
  - Eliminate Z first. Z, X_1, …, X_{n-1}
  - Eliminate Z last. X_1, …, 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 →

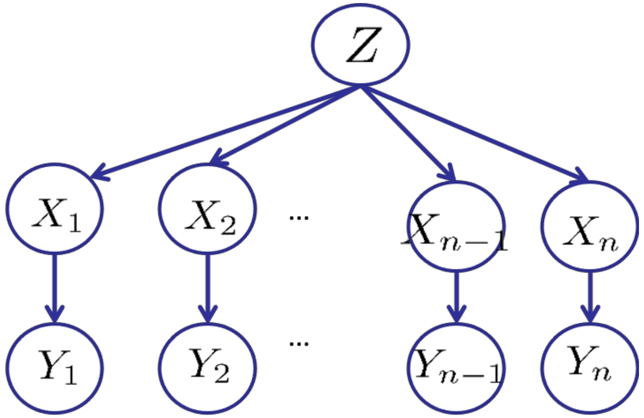
$$f_1(X_1, X_2, \ldots, X_n) = \textstyle\sum_z P(z)P(X_1|z), P(X_1|z), \ldots, P(X_1|z)$$

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

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

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

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 →

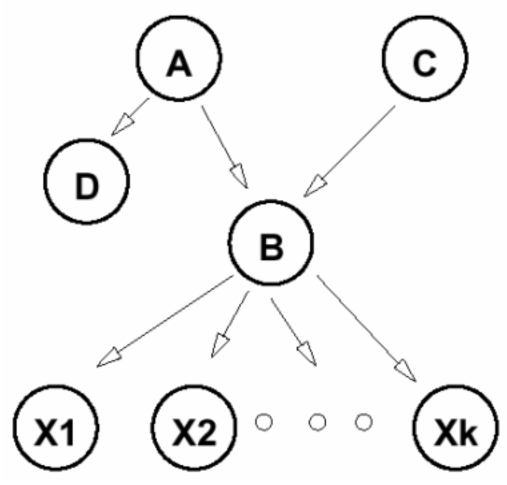
$$f_1(y_1|Z) = \sum_{x_1} P(y_1|x_1)P(x_1|Z)$$

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

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

Variable Ordering for VE

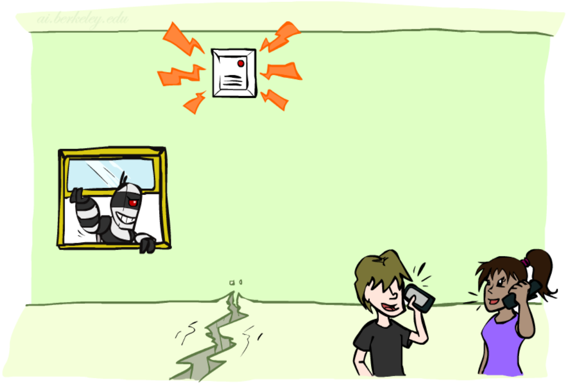
- Variable elimination is dominated by the size of the largest factor constructed during the operation of the algorithm.
- Depends on the structure of the network and order of elimination of the variables.
- Finding the optimal ordering is intractable.
  - Can pose the problem of finding good ordering 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.

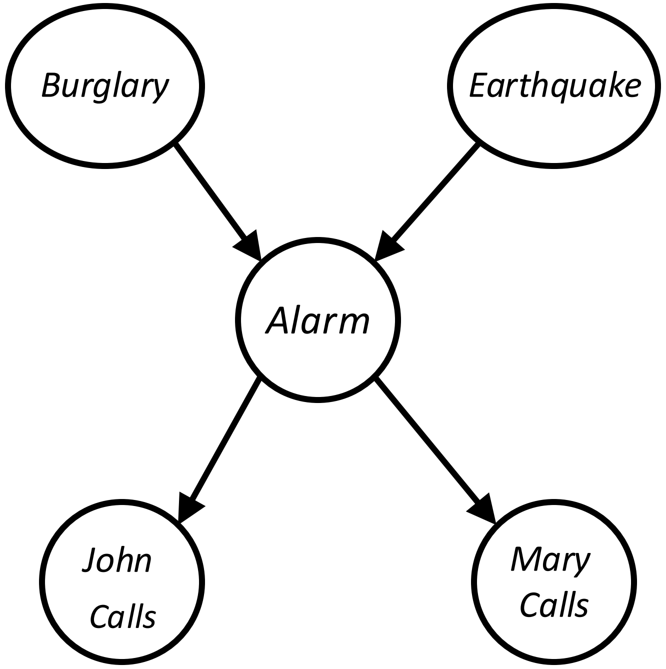


Rank A, B and D with the Min-Fill heuristic.

Some variables may be irrelevant for VE

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





$$P(J)$$

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

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

$$= \Sigma_A P(J|A)\ \Sigma_B P(B)\ \Sigma_E P(A|B,E)P(E) \boxed{\Sigma_M P(M|A)}$$

Variable can be eliminated.

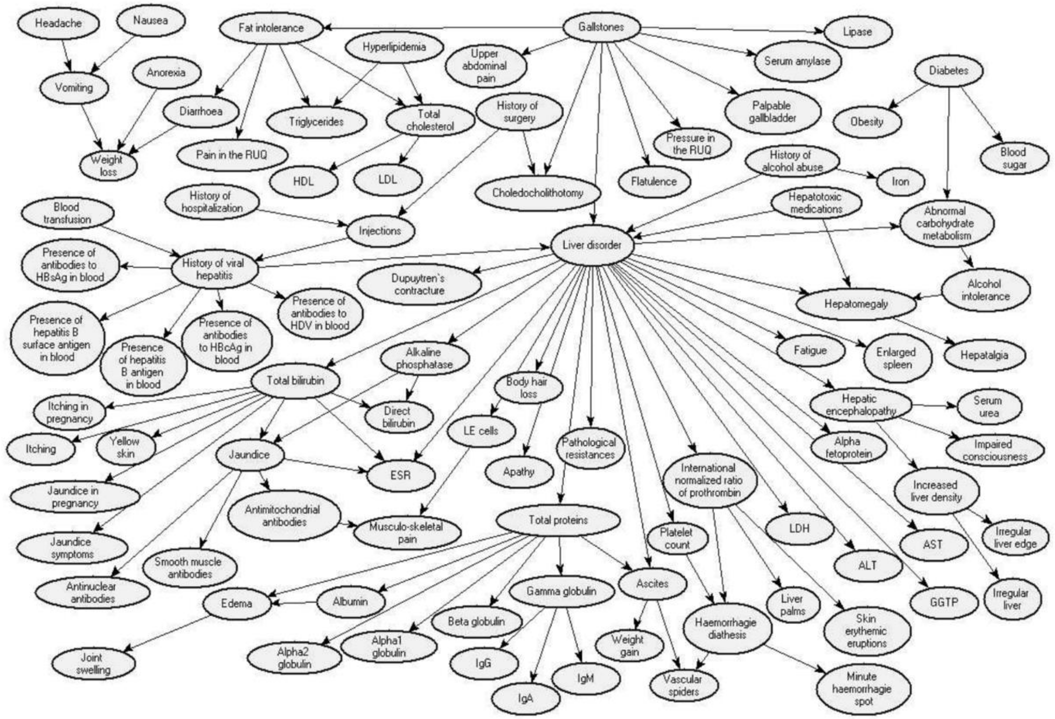
Slide adapted from Prof. Mausam

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

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

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



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

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

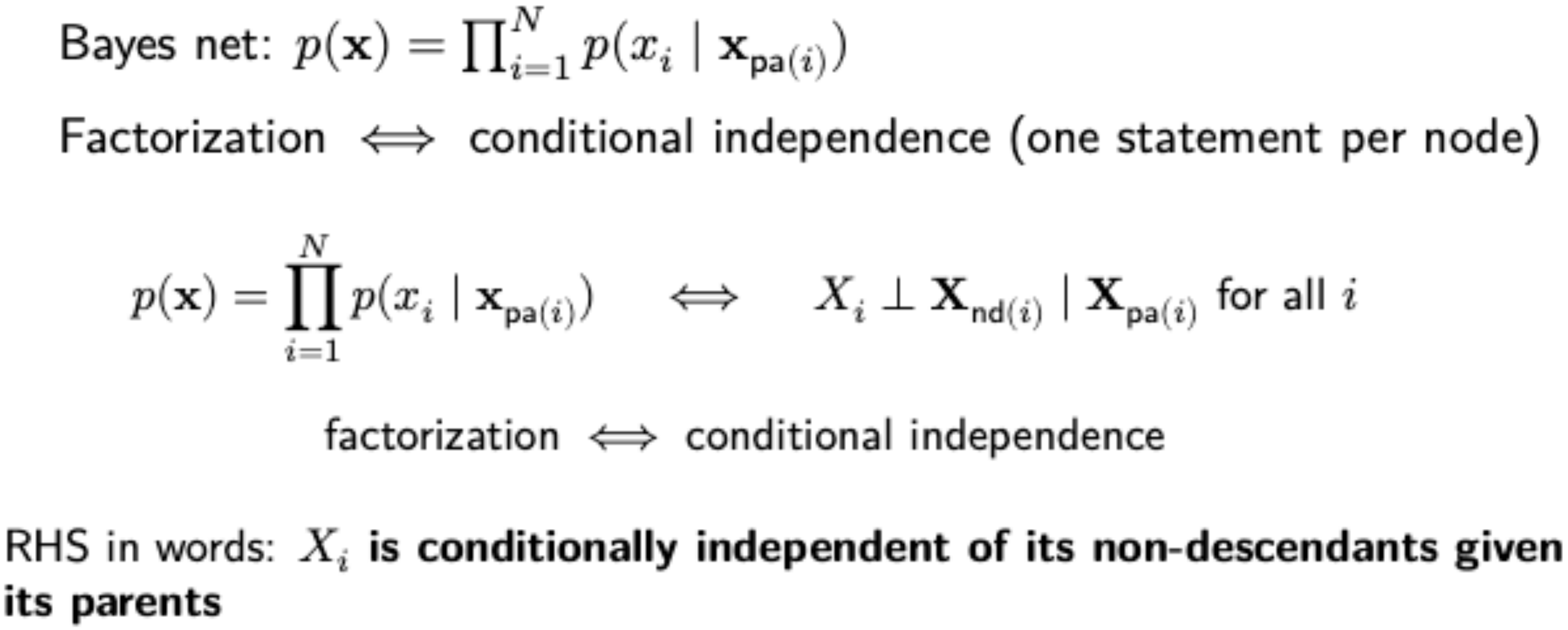
- (Conditional) independence: Given Z, Y has no more information to convey about X or Y does not probabilistically influence X.

- 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).

Graph structure encodes independence relations

- Conditional independence relations in a Bayes Net



Bayes net: $p(\mathbf{x}) = \prod_{i=1}^{N} p(x_i \mid \mathbf{x}_{\mathsf{pa}(i)})$

Factorization $\iff$ conditional independence (one statement per node)

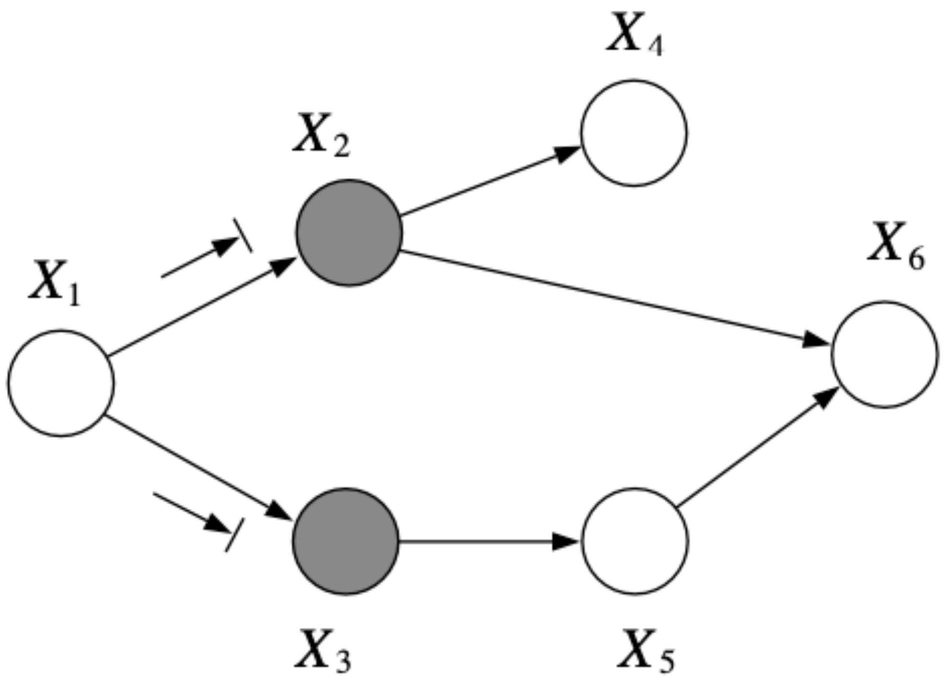
$$p(\mathbf{x}) = \prod_{i=1}^{N} p(x_i \mid \mathbf{x}_{\mathsf{pa}(i)}) \quad \iff \quad X_i \perp \mathbf{X}_{\mathsf{nd}(i)} \mid \mathbf{X}_{\mathsf{pa}(i)} \text{ for all } i$$

factorization $\iff$ conditional independence

RHS in words: X_i **is conditionally independent of its non-descendants given its parents**

Graph structure encodes independence relations

So far, we know $X_i \perp X_{\mathsf{nd}(i)} | X_{\mathsf{pa}(i)}$ for all i

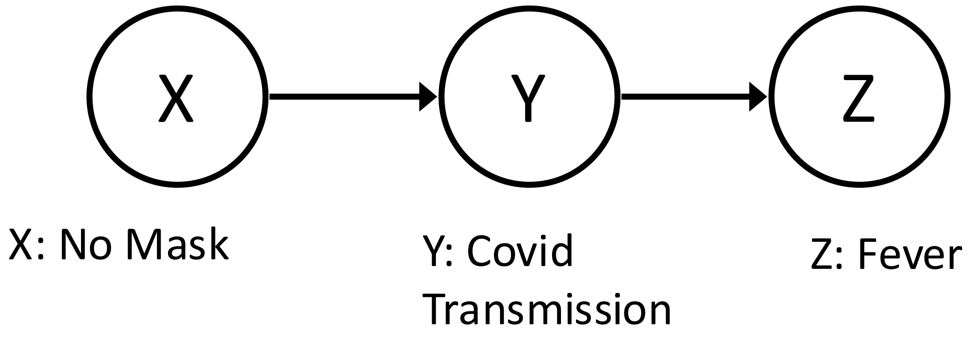


Is X1 conditionally independent of X6 given X3 and X4?

Bayesian Network: Independence Assumptions

- Often there are **additional conditional independences** that are implicit in the network.
- Core Idea: examine three node networks and then chain the ideas together.
- 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.

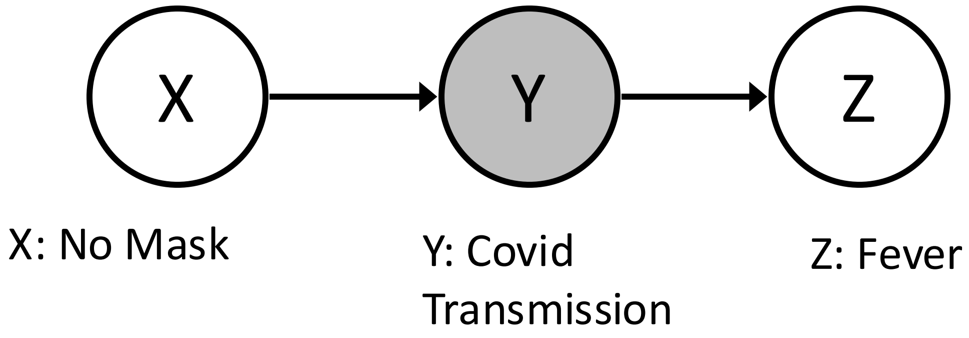
Causal Chains



- **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 | +x) = 1, P(-y | - x) = 1,
  P(+z | +y) = 1, P(-z | -y) = 1

Causal Chains



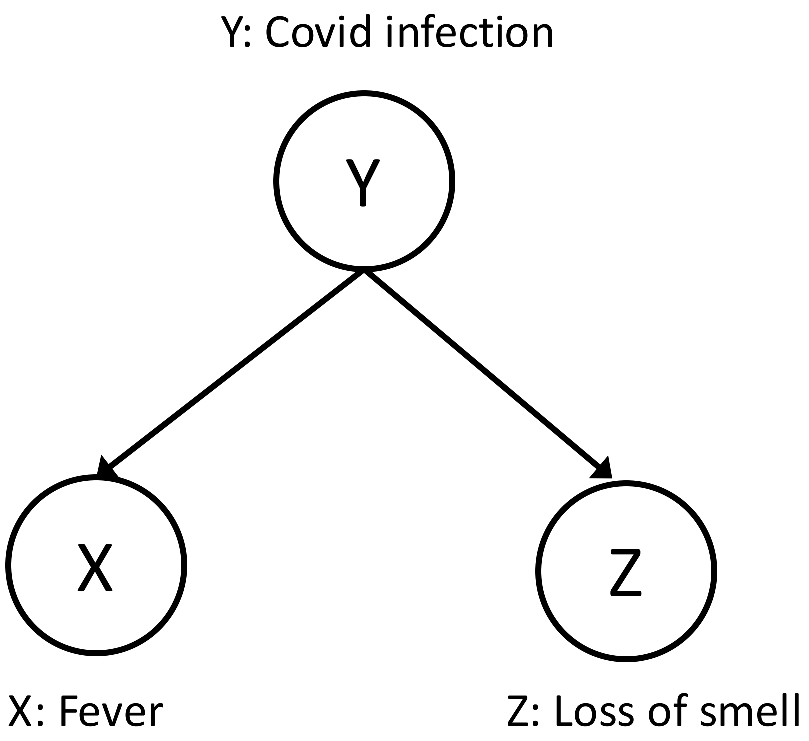
$$P(x, y, z) = P(x)P(y|x)P(z|y)$$

- 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).**

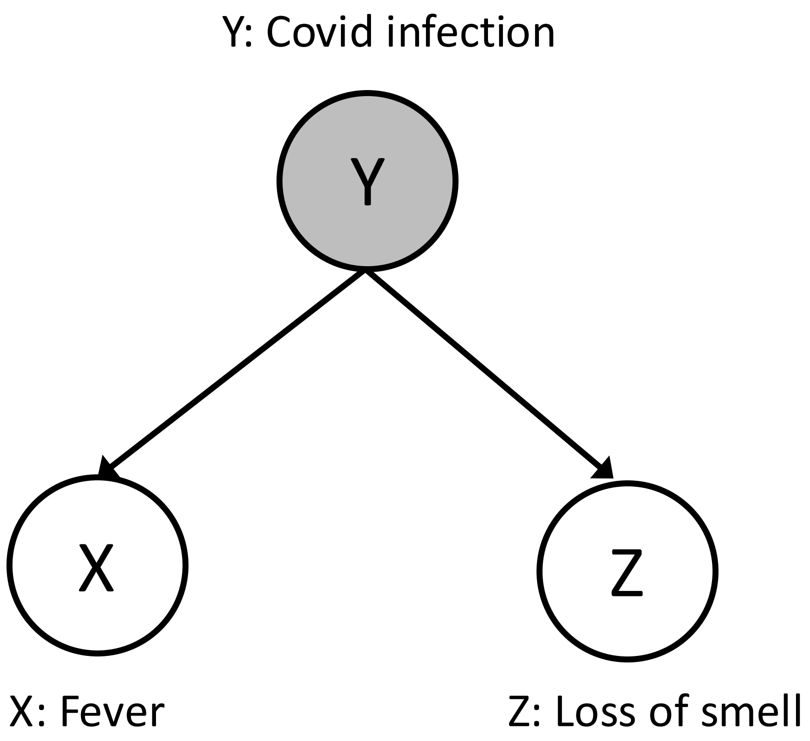
Common Cause



- **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 | +y) = 1, P(-x | -y) = 1,
  P(+z | +y) = 1, P(-z | -y) = 1

Common Cause



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

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

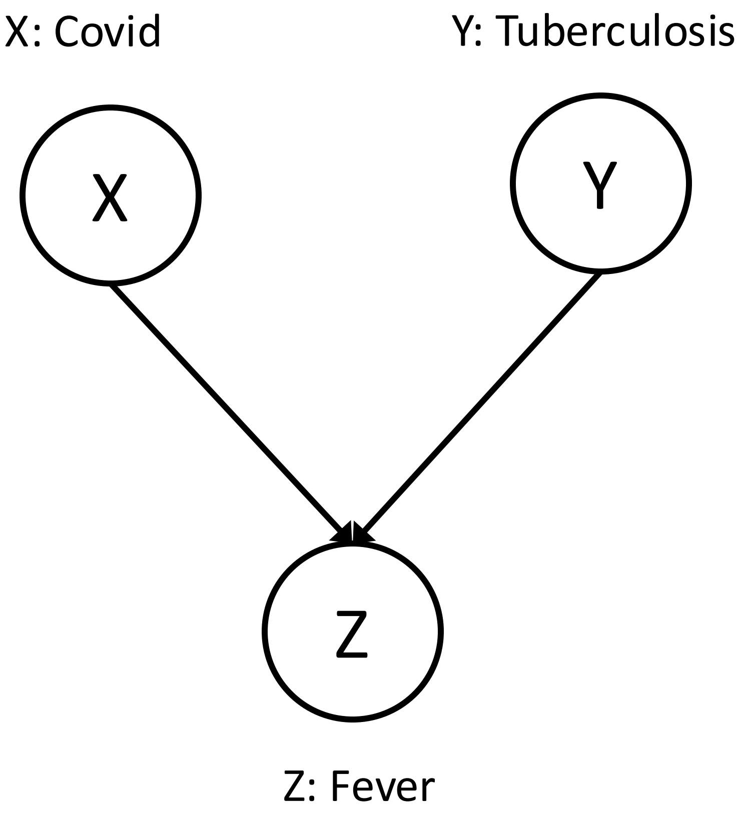
$$P(z|x, y) = \frac{P(x, y, z)}{P(x, y)}$$

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

$$= P(z|y)$$

- 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).**

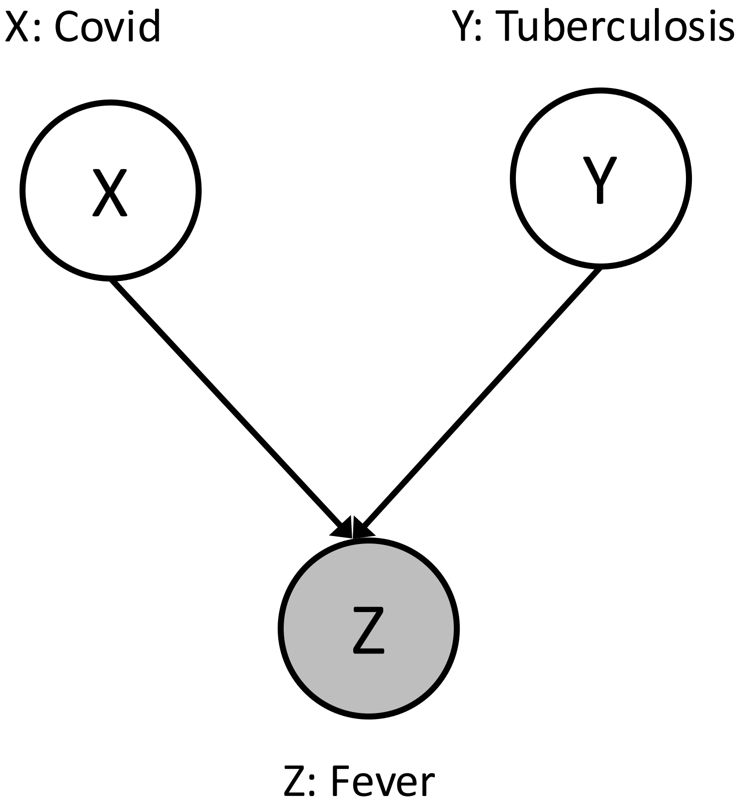
Common Effect



- **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)

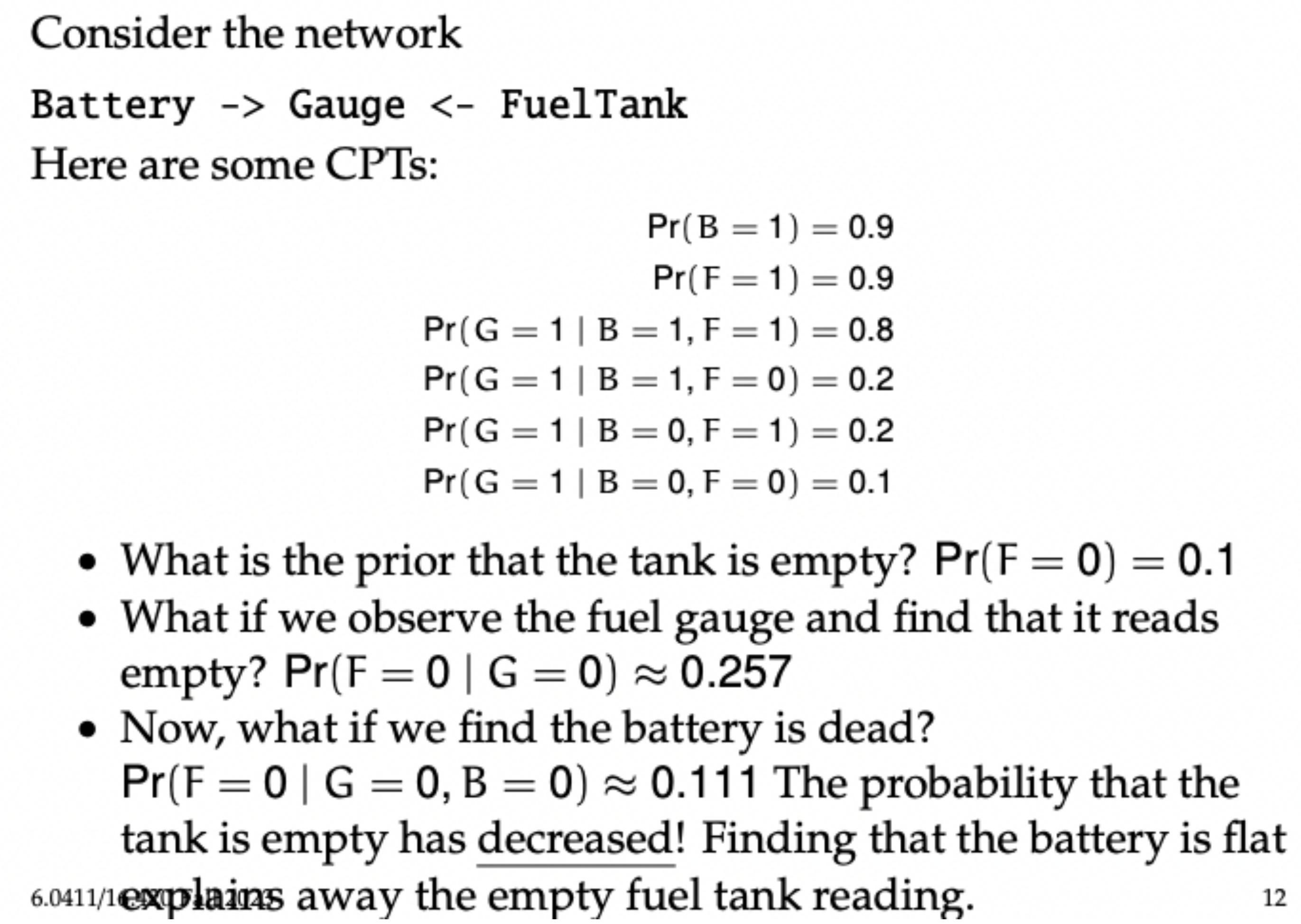
$$\begin{aligned} P(x,y) &= \sum_z P(x,y,z) \\ &= \sum_z P(x)P(y)P(z|x,y) \\ &= P(x)P(y)\sum_z P(z|x,y) \\ &= P(x)P(y) \end{aligned}$$

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.**

“Explaining Away”



Consider the network

`Battery -> Gauge <- FuelTank`

Here are some CPTs:

$$\Pr(B = 1) = 0.9$$
$$\Pr(F = 1) = 0.9$$
$$\Pr(G = 1 \mid B = 1, F = 1) = 0.8$$
$$\Pr(G = 1 \mid B = 1, F = 0) = 0.2$$
$$\Pr(G = 1 \mid B = 0, F = 1) = 0.2$$
$$\Pr(G = 1 \mid B = 0, F = 0) = 0.1$$

- What is the prior that the tank is empty? $\Pr(F = 0) = 0.1$
- What if we observe the fuel gauge and find that it reads empty? $\Pr(F = 0 \mid G = 0) \approx 0.257$
- Now, what if we find the battery is dead? $\Pr(F = 0 \mid G = 0, B = 0) \approx 0.111$ The probability that the tank is empty has decreased! Finding that the battery is flat explains away the empty fuel tank reading.

In words: if there are two possible causes for the observed evidence, knowing about one of the causes provides information about the other.

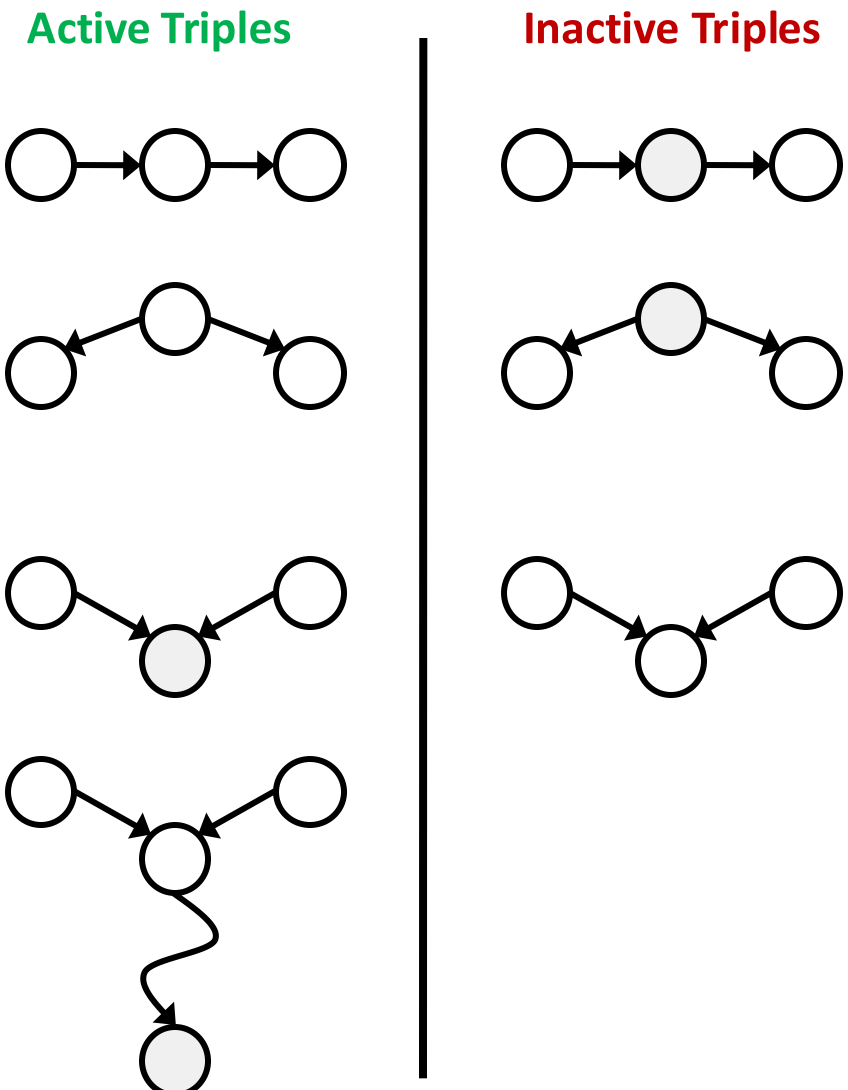
Example from Nicholas Roy

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

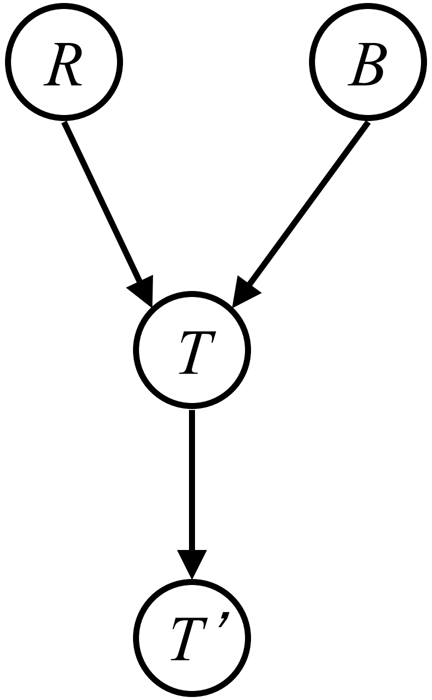
$$X_i \perp\!\!\!\perp X_j | \{X_{k_1}, ..., X_{k_n}\}$$

D-Separation: Examples

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

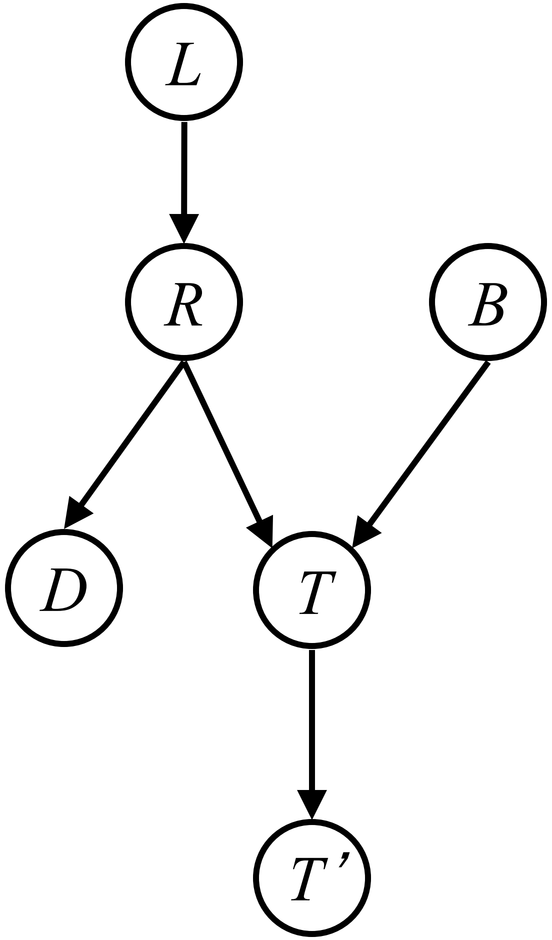
$R \perp\!\!\!\perp B | T$

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



D-Separation: Examples

$L \perp\!\!\!\perp T' \mid T$	*Yes*
$L \perp\!\!\!\perp B$	*Yes*
$L \perp\!\!\!\perp B \mid T$	
$L \perp\!\!\!\perp B \mid T'$	
$L \perp\!\!\!\perp B \mid T, R$	*Yes*



D-Separation: Examples

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

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

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