CSCI 5090/7090- Machine Learning

Spring 2018

Mehdi Allahyari
Georgia Southern University

Graphical Models

(slides borrowed from Tom Mitchell, Ali Borji)

Example

A patient comes into a doctor's office with a fever and a bad cough.

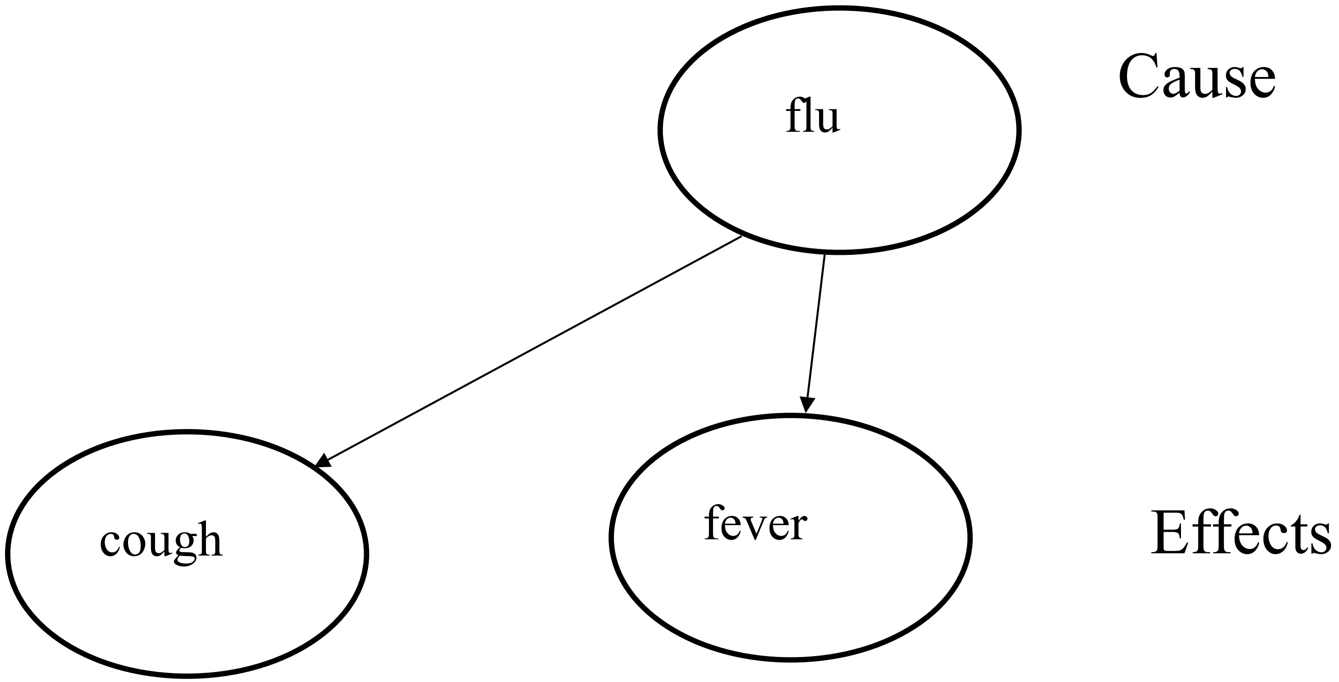
Hypothesis space *H*:

h_1: patient has flu

h_2: patient does not have flu

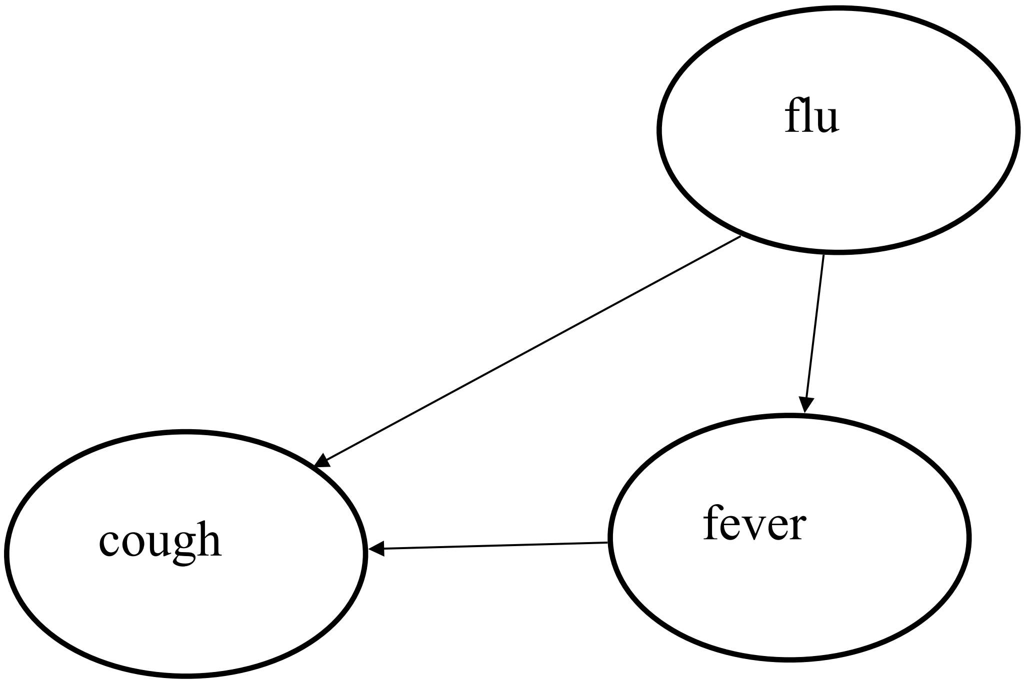
Data *D*:

coughing = true, *fever* = true, *smokes* = true

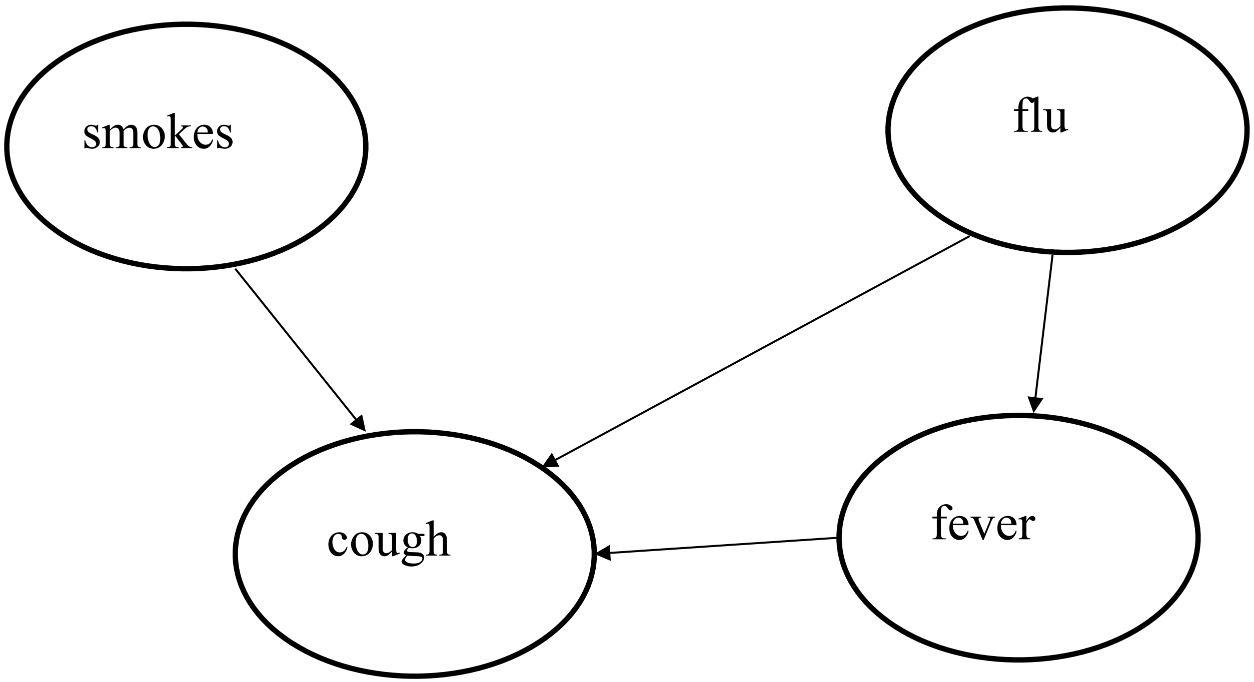


$$P(flu \mid cough, fever) \approx P(flu)P(cough \mid flu)P(fever \mid flu)$$

What if attributes are not independent?



What if more than one possible cause?



Full joint probability distribution

smokes				
	cough		*← cough*	
	Fever	*← Fever*	*Fever*	*← Fever*
flu	p_1	p_2	p_3	p_4
←flu	p_5	p_6	p_7	p_8

← smokes				
	cough		*← cough*	
	fever	*← fever*	*fever*	*← fever*
flu	p_9	p_{10}	p_{11}	p_{12}
←flu	p_{13}	p_{14}	p_{15}	p_{16}

Sum of all boxes is 1.

In principle, the full joint distribution can be used to answer any question about probabilities of these combined parameters.

However, size of full joint distribution scales exponentially with number of parameters so is expensive to store and to compute with.

Full joint probability distribution

	smokes			
	cough		*← cough*	
	Fever	*← Fever*	*Fever*	*← Fever*
flu	p_1	p_2	p_3	p_4
←flu	p_5	p_6	p_7	p_8

	← smokes			
	cough		*← cough*	
	fever	*← fever*	*fever*	*← fever*
flu	p_9	p_{10}	p_{11}	p_{12}
←flu	p_{13}	p_{14}	p_{15}	p_{16}

For example, what if we had another attribute, “allergies”?

How many probabilities would we need to specify?

Allergy				
smokes				
	cough		*← cough*	
	Fever	*←Fever*	*Fever*	*← Fever*
flu	p_1	p_2	p_3	p_4
←flu	p_5	p_6	p_7	p_8

←Allergy				
smokes				
	cough		*← cough*	
	Fever	*←Fever*	*Fever*	*← Fever*
flu	p_{17}	p_{18}	p_{19}	p_{20}
←flu	p_{21}	p_{22}	p_{23}	p_{24}

Allergy				
← smokes				
	cough		*← cough*	
	fever	*← fever*	*fever*	*← fever*
flu	p_9	p_{10}	p_{11}	p_{12}
←flu	p_{13}	p_{14}	p_{15}	p_{16}

←Allergy				
← smokes				
	cough		*← cough*	
	fever	*← fever*	*fever*	*← fever*
flu	p_{25}	p_{26}	p_{27}	p_{28}
←flu	p_{29}	p_{30}	p_{31}	p_{328}

Allergy				
smokes				
	cough		*← cough*	
	Fever	*←Fever*	*Fever*	*← Fever*
flu	p_1	p_2	p_3	p_4
←flu	p_5	p_6	p_7	p_8

←Allergy				
smokes				
	cough		*← cough*	
	Fever	*←Fever*	*Fever*	*← Fever*
flu	p_{17}	p_{18}	p_{19}	p_{20}
←flu	p_{21}	p_{22}	p_{23}	p_{24}

Allergy				
← smokes				
	cough		*← cough*	
	fever	*← fever*	*fever*	*← fever*
flu	p_9	p_{10}	p_{11}	p_{12}
←flu	p_{13}	p_{14}	p_{15}	p_{16}

←Allergy				
← smokes				
	cough		*← cough*	
	fever	*← fever*	*fever*	*← fever*
flu	p_{25}	p_{26}	p_{27}	p_{28}
←flu	p_{29}	p_{30}	p_{31}	p_{329}

But can reduce this if we know which variables are conditionally independent

Graphical Models

- Key Idea:
  - Conditional independence assumptions useful
  - but Naïve Bayes is extreme!
  - Graphical models express sets of conditional independence assumptions via graph structure
  - Graph structure plus associated parameters define *joint probability distribution over set of variables*

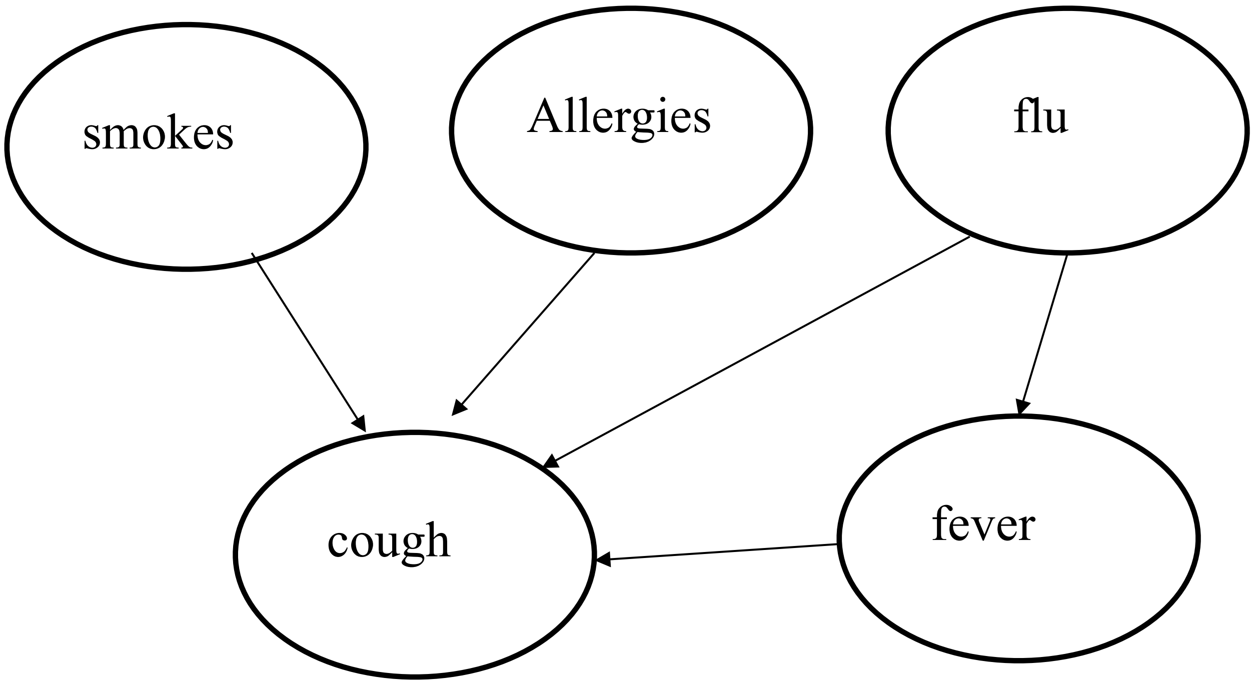
- Two types of graphical models:
  - Directed graphs (aka Bayesian Networks)
  - Undirected graphs (aka Markov Random Fields)

Graphical Models– Why Care?

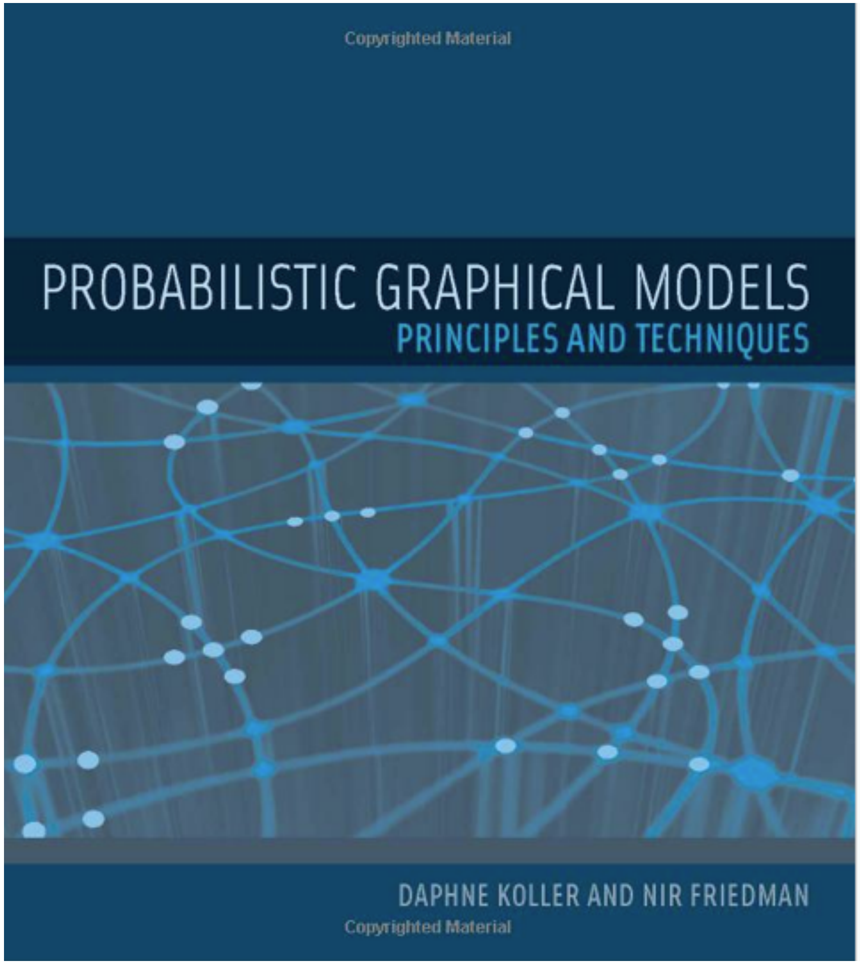
- Among most important ML developments of the decade
- Graphical models allow combining:
  - Prior knowledge in form of dependencies/independencies
  - Prior knowledge in form of priors over parameters
  - Observed training data
- Principled and ~general methods for
  - Probabilistic inference
  - Learning
- Useful in practice
  - Diagnosis, help systems, text analysis, time series models, ...

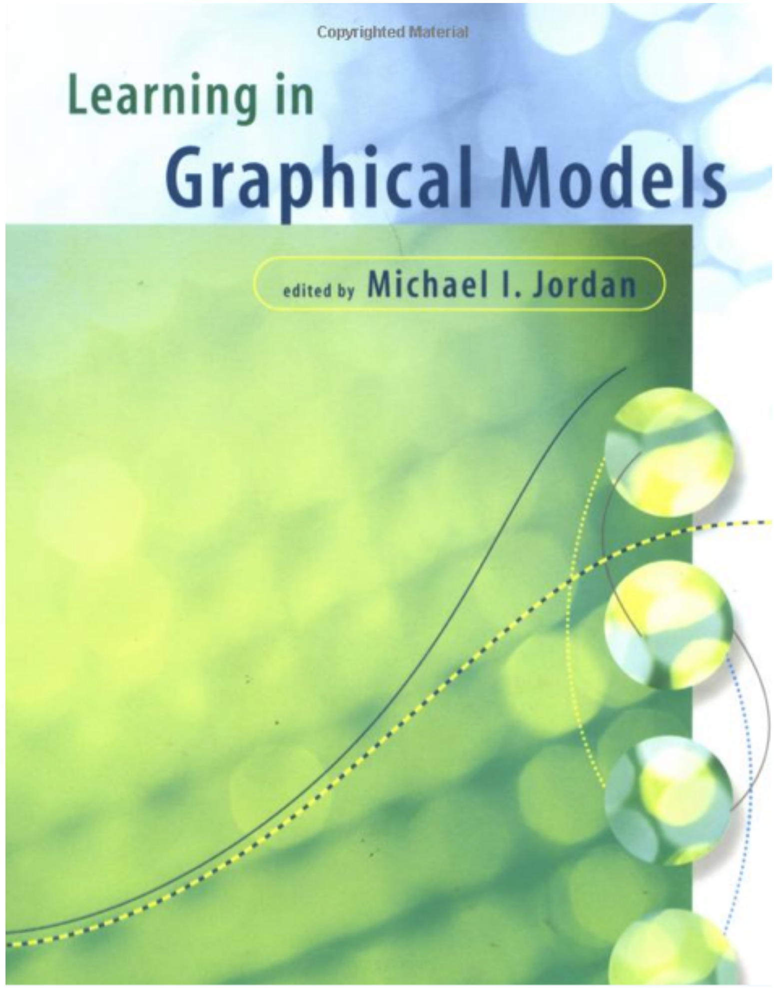
Bayesian networks

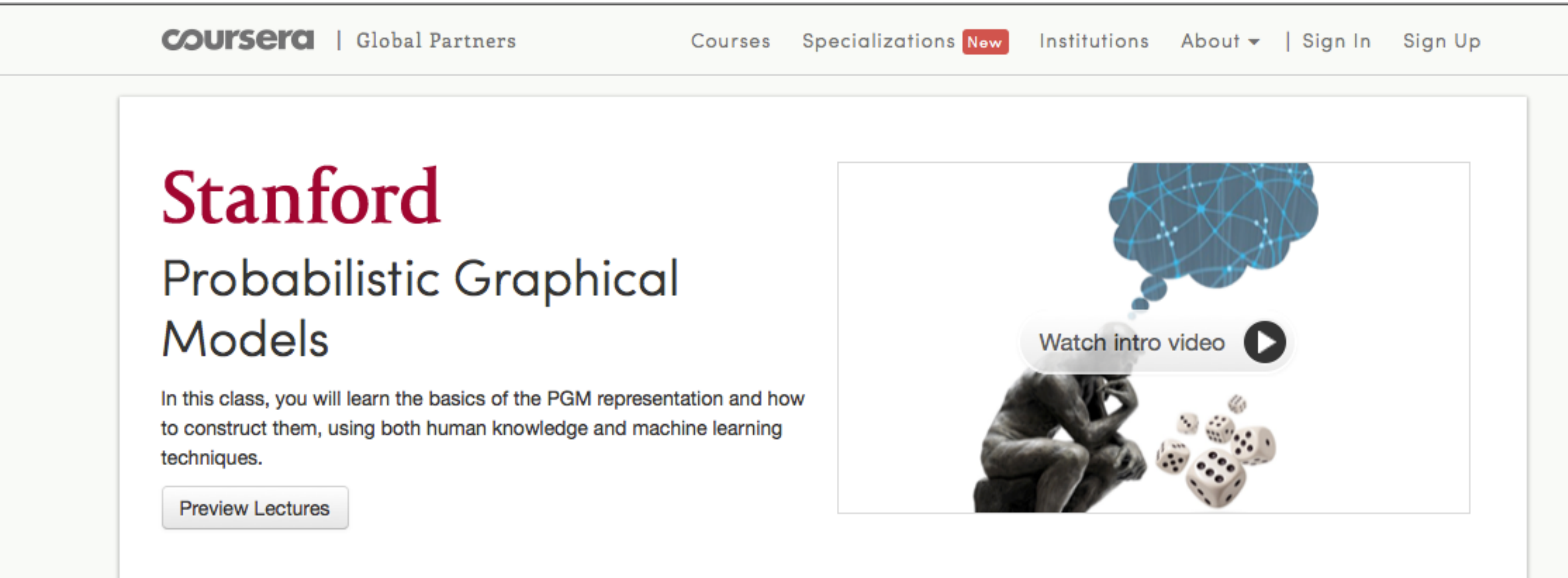
- Idea is to represent dependencies (or causal relations) for all the variables so that space and computation-time requirements are minimized.



“Graphical Models”


Copyrighted Material
PROBABILISTIC GRAPHICAL MODELS
PRINCIPLES AND TECHNIQUES
DAPHNE KOLLER AND NIR FRIEDMAN
Copyrighted Material


Copyrighted Material
Learning in
Graphical Models
edited by Michael I. Jordan


coursera | Global Partners
Courses
Specializations New
Institutions
About
Sign In
Sign Up
Stanford
Probabilistic Graphical Models
In this class, you will learn the basics of the PGM representation and how to construct them, using both human knowledge and machine learning techniques.
Preview Lectures
Watch intro video

Conditional Independence

Definition: X is conditionally independent of Y given Z, if the probability distribution governing X is independent of the value of Y, given the value of Z

$$(\forall i, j, k) P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z = z_k)$$

Which we often write $P(X|Y, Z) = P(X|Z)$

E.g., $P(Thunder|Rain, Lightning) = P(Thunder|Lightning)$

Marginal Independence

Definition: X is marginally independent of Y if

$$(\forall i,j)P(X=x_i,Y=y_j)=P(X=x_i)P(Y=y_j)$$

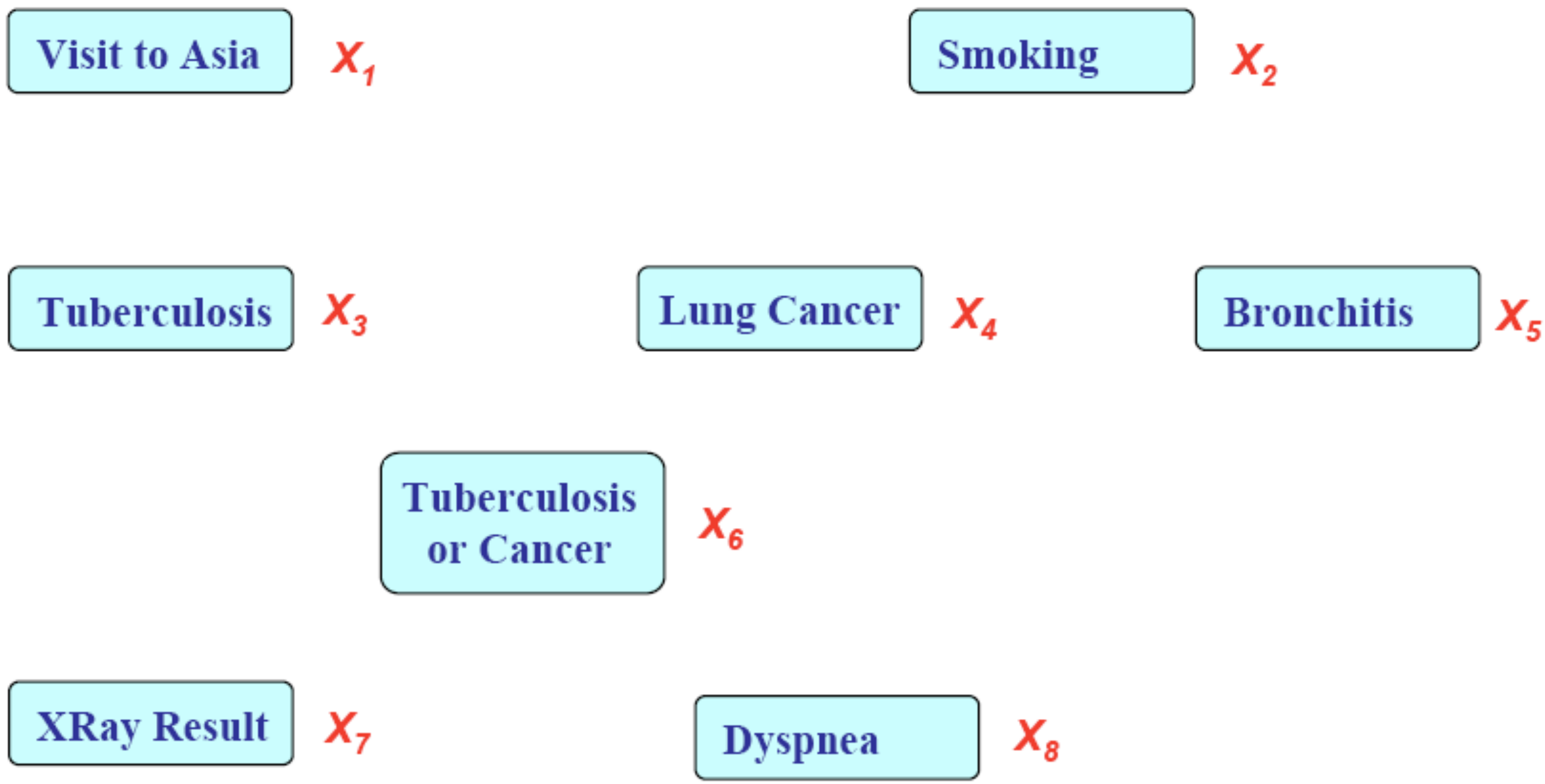
Equivalently, if

$$(\forall i,j)P(X=x_i|Y=y_j)=P(X=x_i)$$

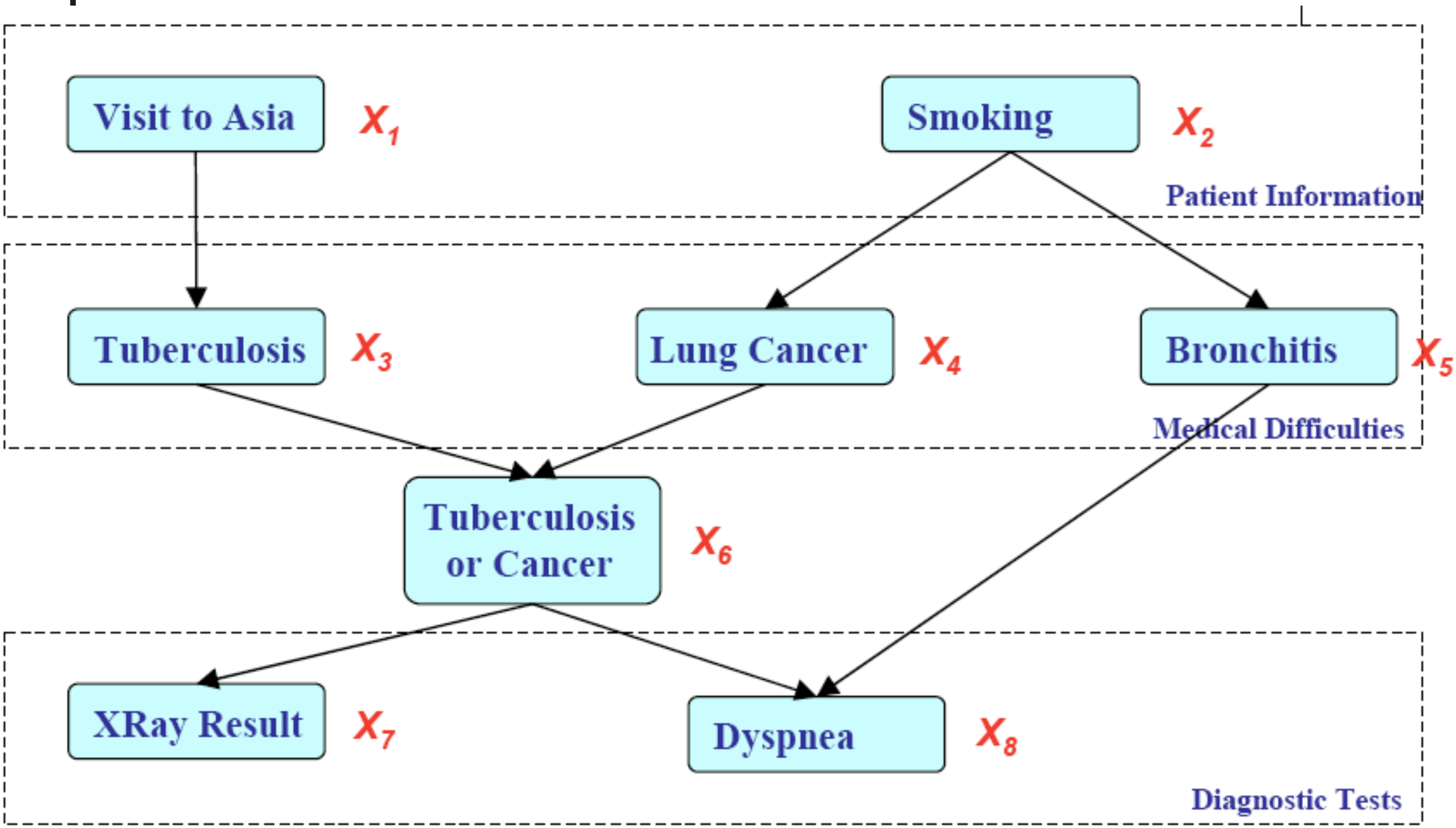
Equivalently, if

$$(\forall i,j)P(Y=y_i|X=x_j)=P(Y=y_i)$$

Represent Join Probability Distribution over Variables



Describe Network of Dependencies



Bayesian Networks

Bayesian Networks = Bayesian Belief Networks = Bayes Nets

Bayesian Network: Alternative representation for complete joint probability distribution

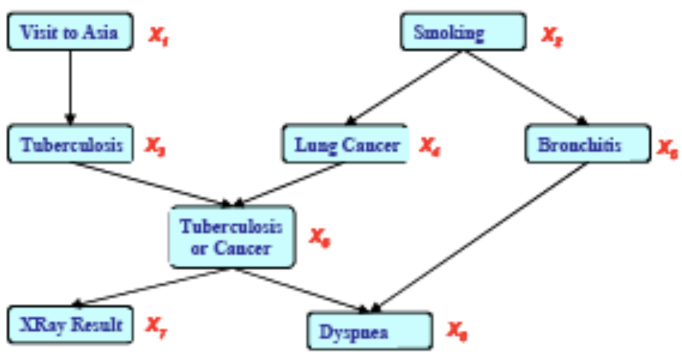
"Useful for making probabilistic inference about models domains characterized by inherent complexity and uncertainty"

Uncertainty can come from:

- incomplete knowledge of domain
- inherent randomness in behavior in domain

Bayesian Networks

Bayes Nets define Joint Probability Distribution in terms of this graph, plus parameters

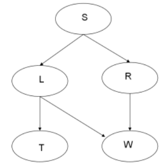


$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$
$$= P(X_1)\, P(X_2)\, P(X_3 | X_1)\, P(X_4 | X_2)\, P(X_5 | X_2)\, P(X_6 | X_3, X_4)\, P(X_7 | X_6)\, P(X_8 | X_5, X_6)$$

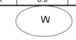
Benefits of Bayes Nets:

- Represent the full joint distribution in fewer parameters, using prior knowledge about dependencies
- Algorithms for inference and learning

Bayesian Networks Definition



Parents	P(W|Pa)	P(¬W|Pa)
L, R	0	1.0
L, ¬R	0	1.0
¬L, R	0.2	0.8
¬L, ¬R	0.9	0.1



A Bayes network represents the joint probability distribution over a collection of random variables

A Bayes network is a directed acyclic graph and a set of conditional probability distributions (CPD's)

- Each node denotes a random variable
- Edges denote dependencies
- For each node X_i its CPD defines $P(X_i \mid Pa(X_i))$
- The joint distribution over all variables is defined to be

$$P(X_1 \ldots X_n) = \prod_i P(X_i | Pa(X_i))$$

Pa(X) = immediate parents of X in the graph

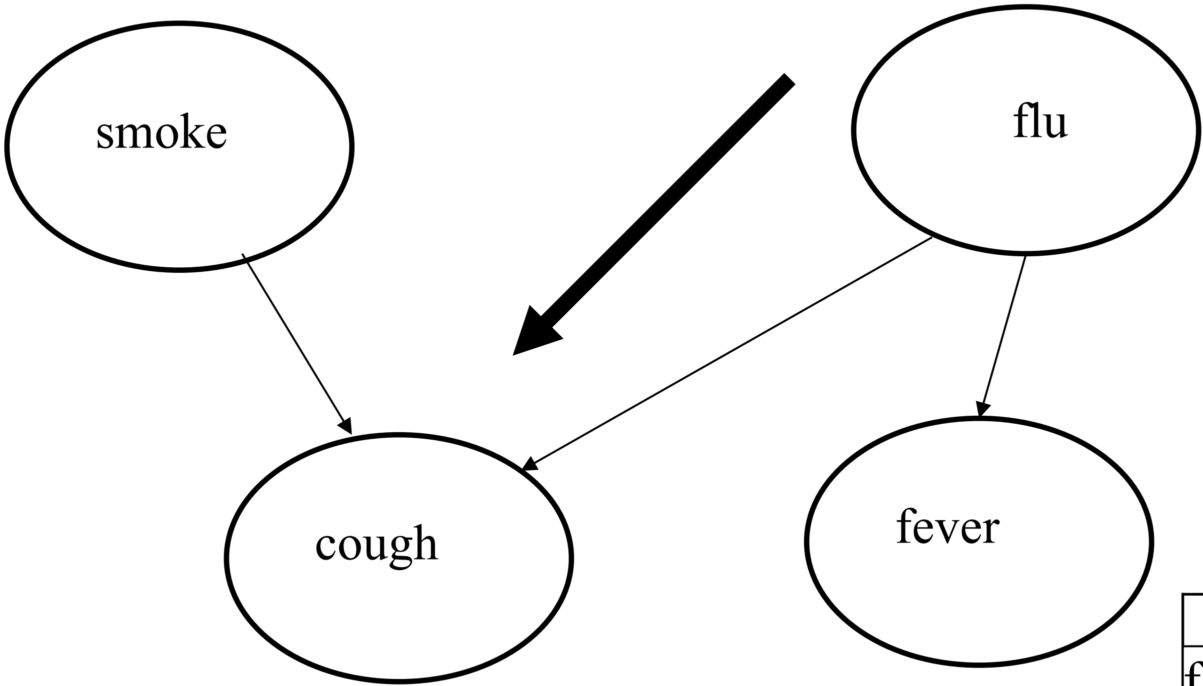
Example:

smoke	
true	*0.2*
false	*0.8*

		cough	
flu	smoke	*true*	*false*
True	*True*	*0.95*	*0.05*
True	*False*	*0.8*	*0.2*
False	*True*	*0.6*	*0.4*
false	*false*	*0.05*	*0.95*

Conditional probability tables for each node

flu	
true	*0.01*
false	*0.99*



	fever	
flu	*true*	*false*
true	*0.9*	*0.1*
false	*0.2*	*0.8*

Inference in Bayesian networks

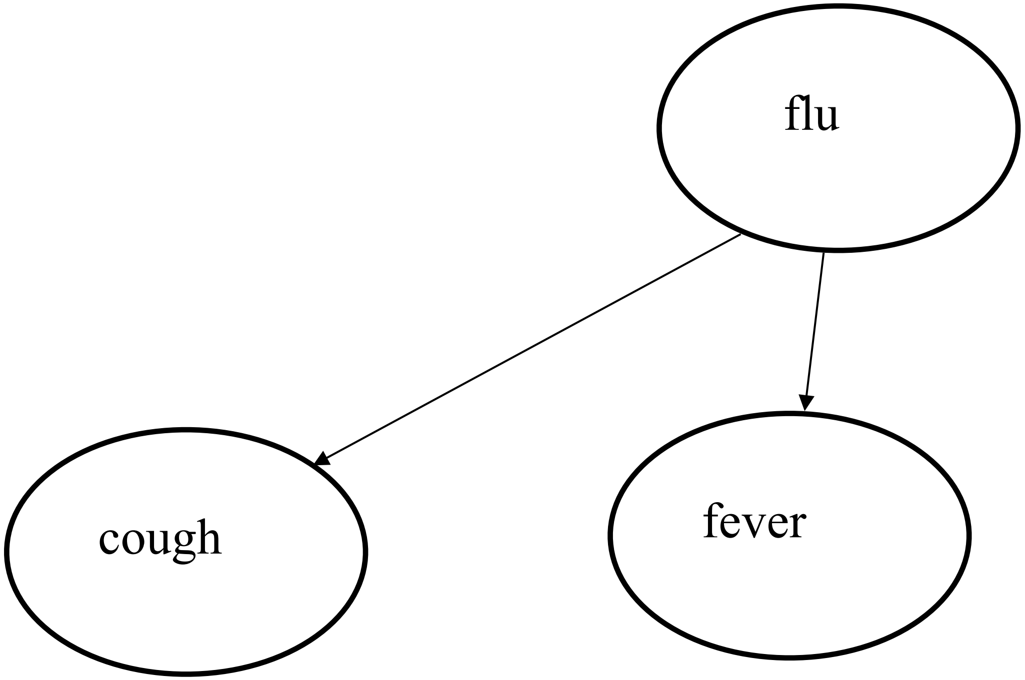
- If network is correct, can calculate full joint probability distribution from network.

$$P(X_1 \ldots X_n) = \prod_i P(X_i | Pa(X_i))$$

Pa(X) = immediate parents of X in the graph

where $pa(X_i)$ denotes specific values of parents of X_i.

Naïve Bayes Example



$$P(flu \mid cough, fever) \approx P(flu)P(cough \mid flu)P(fever \mid flu)$$

Example

- Calculate

$$P(cough = t \wedge fever = f \wedge flu = f \wedge smoke = f)$$

Example

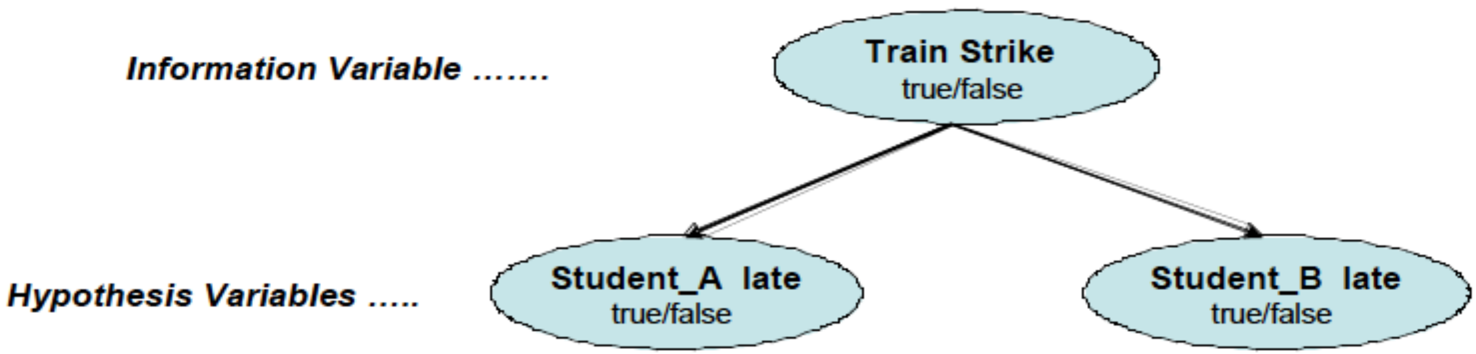
- Calculate

$$P(cough = t \wedge fever = f \wedge flu = f \wedge smoke = f)$$

$$= \prod_{i=1}^{n} P(X_i = x_i \mid parents(X_i))$$
$$= P(cough = t \mid flu = f \wedge smoke = f)$$
$$\times P(fever = f \mid flu = f)$$
$$\times P(flu = f)$$
$$\times P(smoke = f)$$
$$= .05 \times .8 \times .99 \times .8$$
$$= .032$$

Example

Train Strike	
true	0.1
false	0.9



	Train Strike	
Student_A late	*true*	*false*
true	0.8	0.1
false	0.2	0.9

	Train Strike	
Student_B late	*true*	*false*
true	0.6	0.5
false	0.4	0.5

Figure 1. BBN detailing the likely implications of a train strike on the arrival time of two different students (Student_A and Student_B)

Example

What is the probability that Student A is late?

What is the probability that Student B is late?

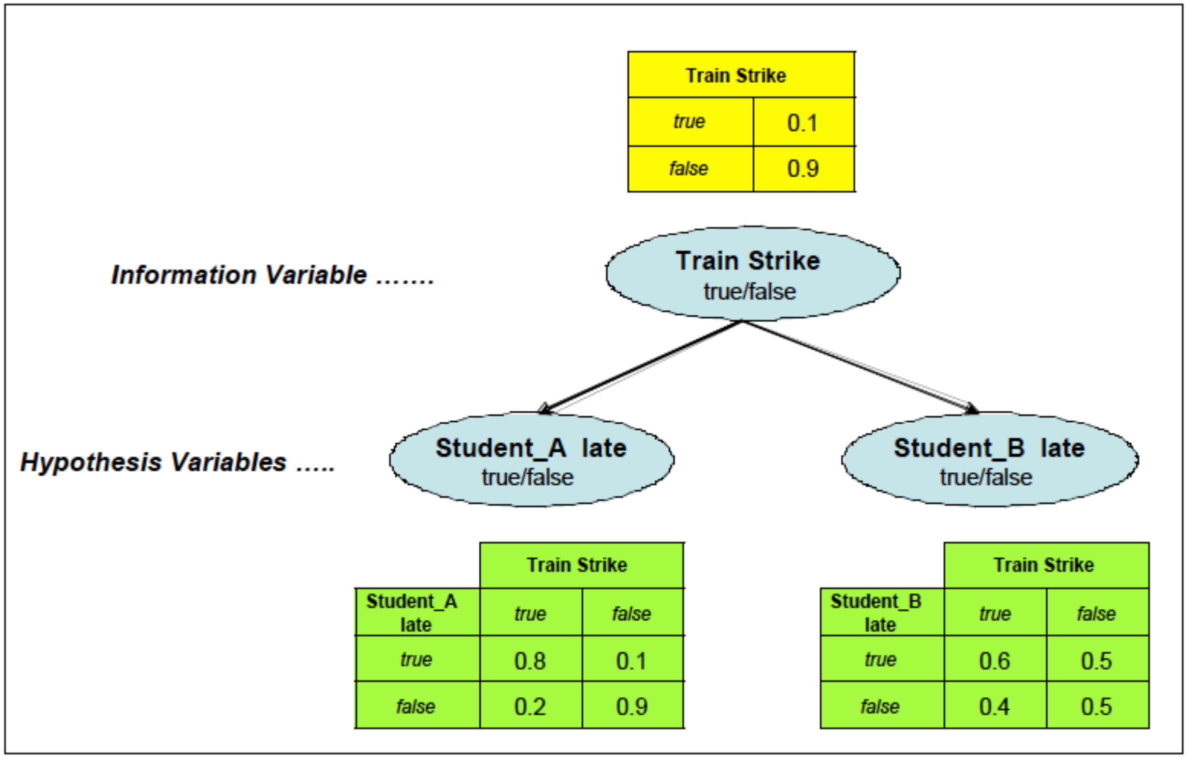


Figure 1. BBN detailing the likely implications of a train strike on the arrival time of two different students (Student_A and Student_B)

Example

What is the probability that Student A is late?

What is the probability that Student B is late?

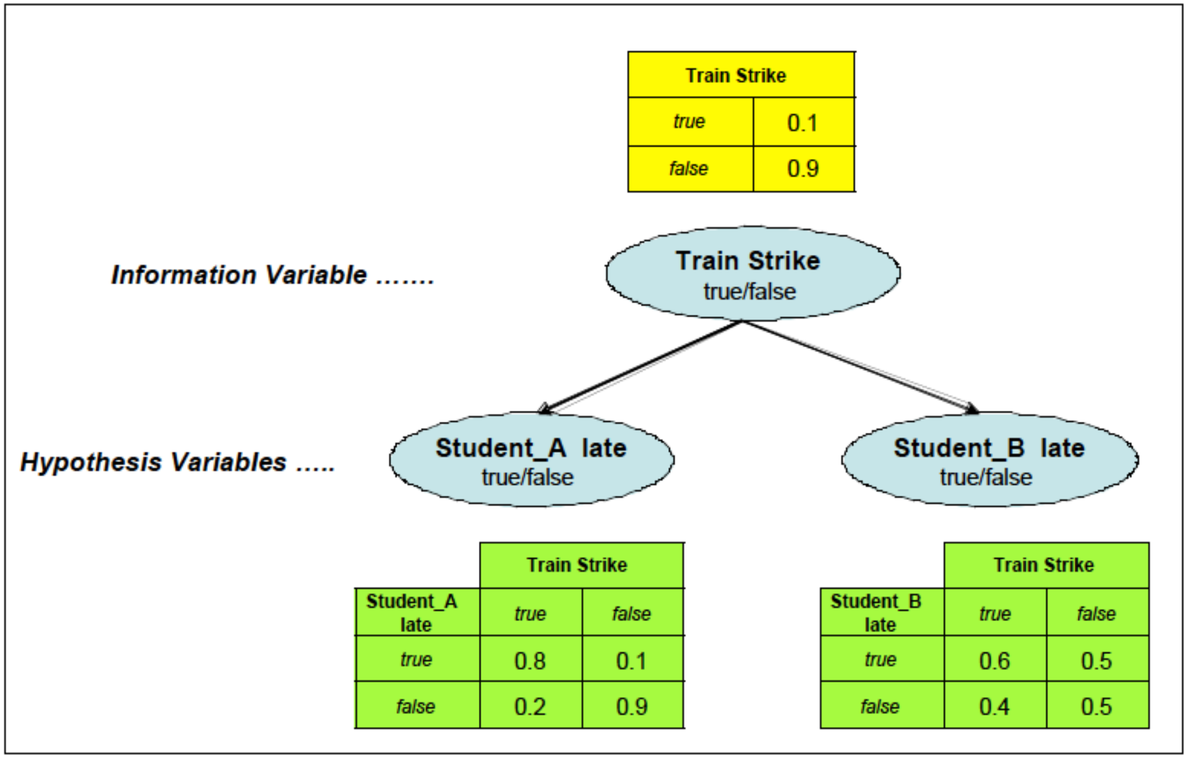


Figure 1. BBN detailing the likely implications of a train strike on the arrival time of two different students (Student_A and Student_B)

Unconditional ("marginal") probability. We don't know if there is a train strike.

What is the probability that Student A is late?

What is the probability that Student B is late?

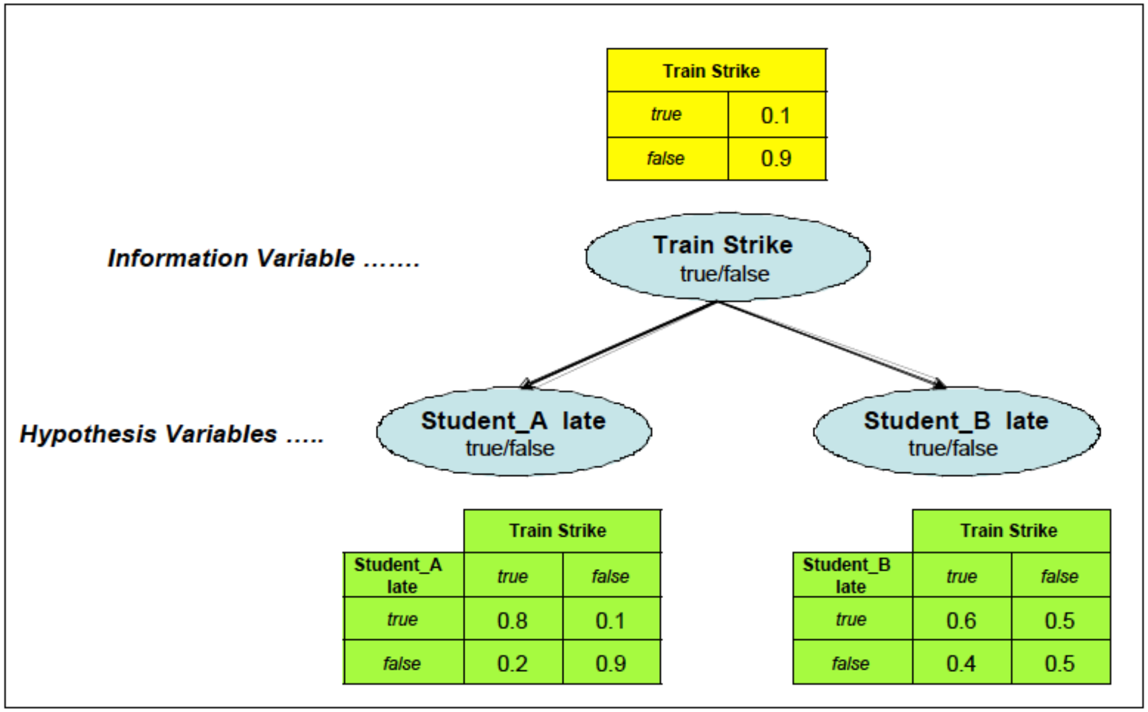


Figure 1. BBN detailing the likely implications of a train strike on the arrival time of two different students (Student_A and Student_B)

Unconditional ("marginal") probability. We don't know if there is a train strike.

$$P(StudentALate) = P(StudentALate \mid TrainStrike)P(TrainStrike)$$
$$+P(StudentALate \mid \neg TrainStrike)P(\neg TrainStrike)$$
$$= 0.8 \times 0.1 + 0.8 \times 0.9 = 0.17$$

$$P(StudentBLate) = P(StudentBLate \mid TrainStrike)P(TrainStrike)$$
$$+P(StudentBLate \mid \neg TrainStrike)P(\neg TrainStrike)$$
$$= 0.6 \times 0.1 + 0.5 \times 0.9 = 0.51$$

Now, suppose we know that there is a train strike. How does this revise the probability that the students are late?

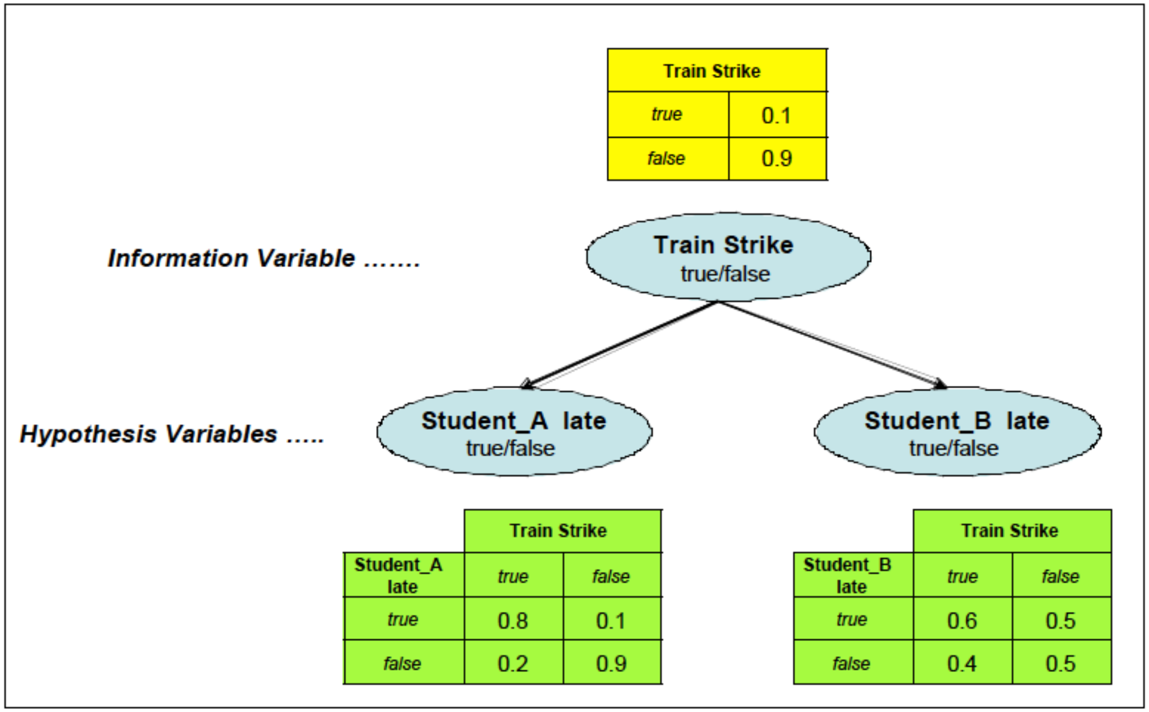


Figure 1. BBN detailing the likely implications of a train strike on the arrival time of two different students (Student_A and Student_B)

Now, suppose we know that there is a train strike. How does this revise the probability that the students are late?

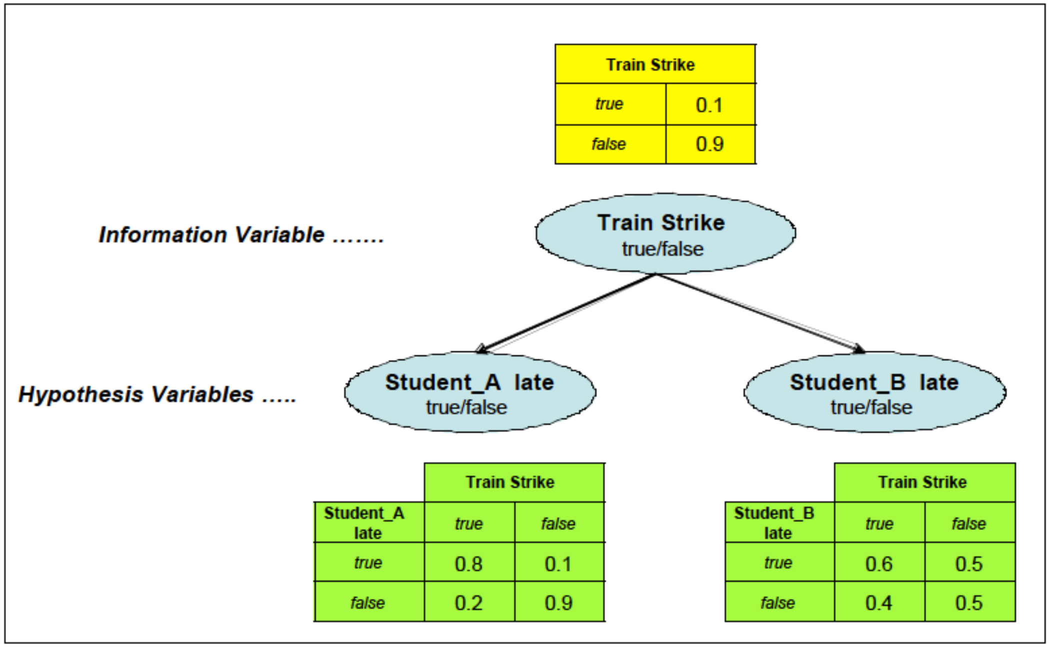


Figure 1. BBN detailing the likely implications of a train strike on the arrival time of two different students (Student_A and Student_B)

Evidence: There is a train strike.

$$P(StudentALate) = 0.8$$

$$P(StudentBLate) = 0.6$$

Now, suppose we know that Student A is late.

How does this revise the probability that there is a train strike?

How does this revise the probability that Student B is late?

Notion of "belief propagation".

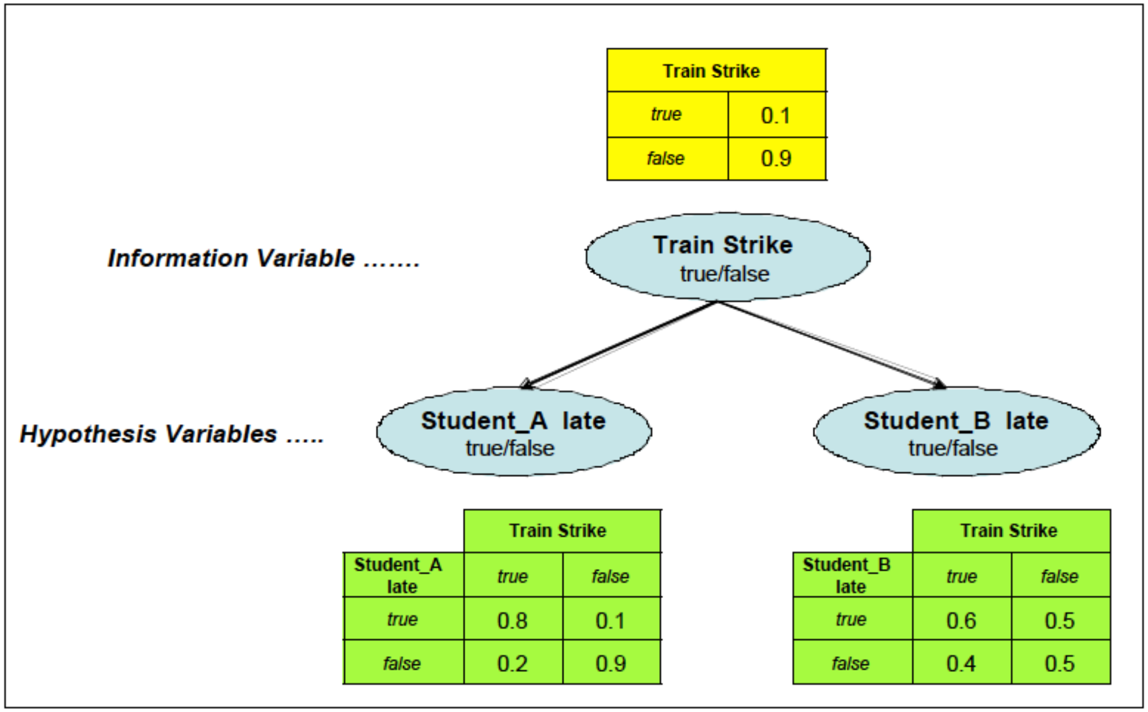


Figure 1. BBN detailing the likely implications of a train strike on the arrival time of two different students (Student_A and Student_B)

Evidence: Student A is late.

Now, suppose we know that Student A is late.

How does this revise the probability that there is a train strike?

How does this revise the probability that Student B is late?

Notion of "belief propagation".

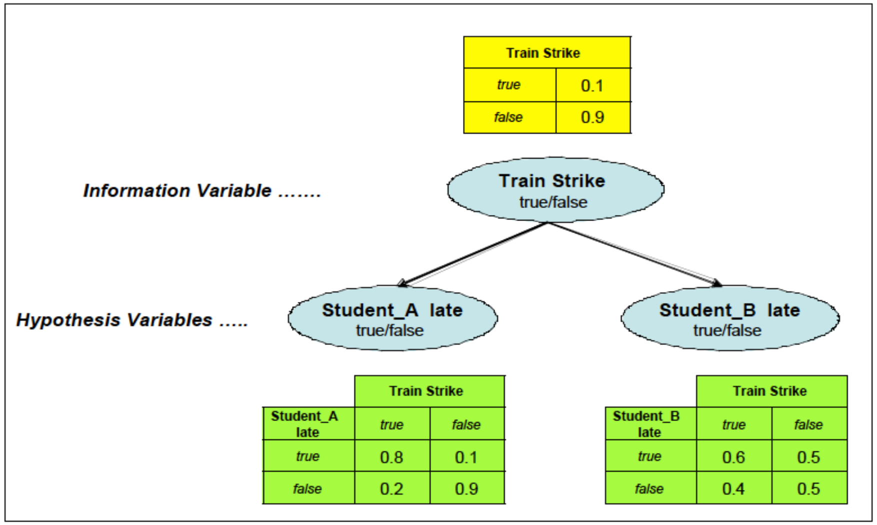


Figure 1. BBN detailing the likely implications of a train strike on the arrival time of two different students (Student_A and Student_B)

Evidence: Student A is late.

$$P(TrainStrike \mid StudentALate) = \frac{P(StudentALate \mid TrainStrike)\,P(TrainStrike)}{P(StudentALate)} \quad \text{by Bayes Theorem}$$

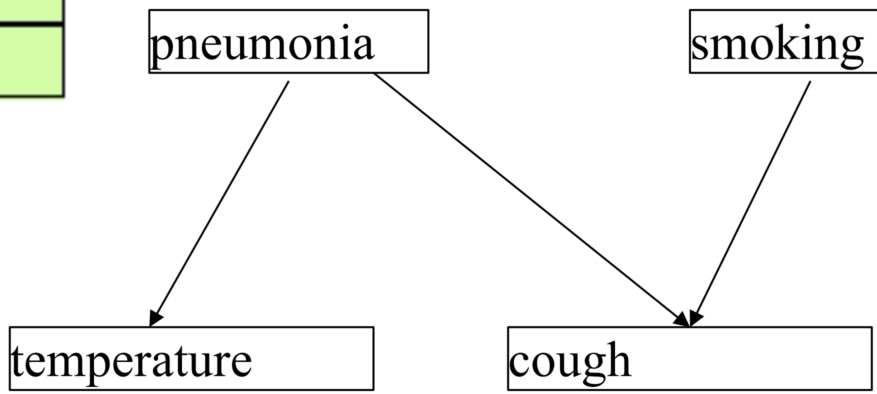
$$= \frac{0.8 \times 0.1}{0.17} = 0.47$$

$$P(StudentBLate) = P(StudentBLate \mid TrainStrike)\,P(TrainStrike)$$
$$+ P(StudentBLate \mid \neg TrainStrike)\,P(\neg TrainStrike)$$
$$= 0.6 \times 0.47 + 0.5 \times 0.53 = 0.55$$

Another example

pneumonia	
true	0.1
false	0.9

smoking	
yes	0.2
no	0.8

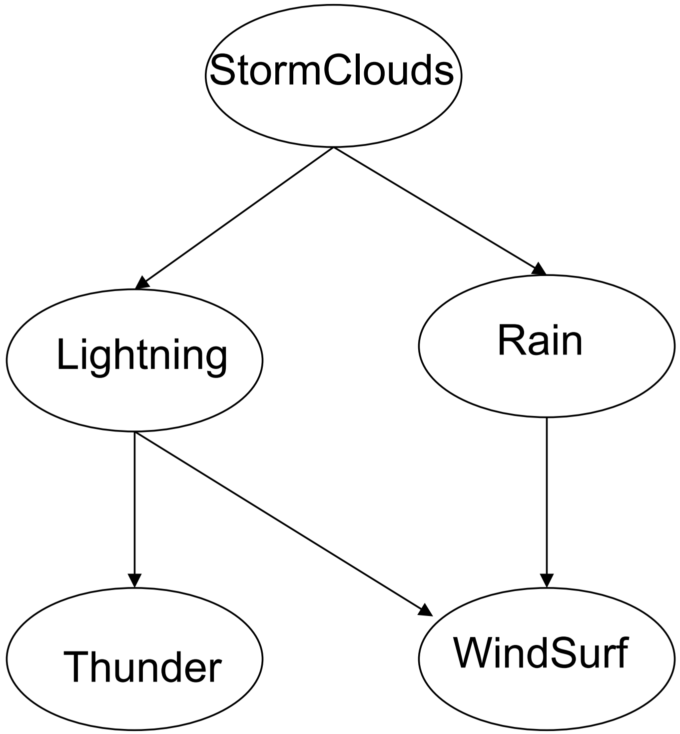


	temperature	
pneumonia	yes	no
yes	0.9	0.1
no	0.2	0.8

		cough	
pneumonia	smoking	true	false
true	yes	0.95	0.05
true	no	0.8	0.2
false	yes	0.6	0.4
false	no	0.05	0.95

What is $P(cough)$?

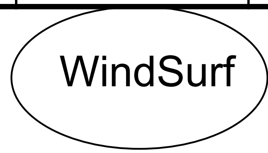
Bayesian Network - Example



Nodes = random variables

A conditional probability distribution (CPD) is associated with each node N, defining P(N | Parents(N))

Parents	P(W|Pa)	P(¬W|Pa)
L, R	0	1.0
L, ¬R	0	1.0
¬L, R	0.2	0.8
¬L, ¬R	0.9	0.1

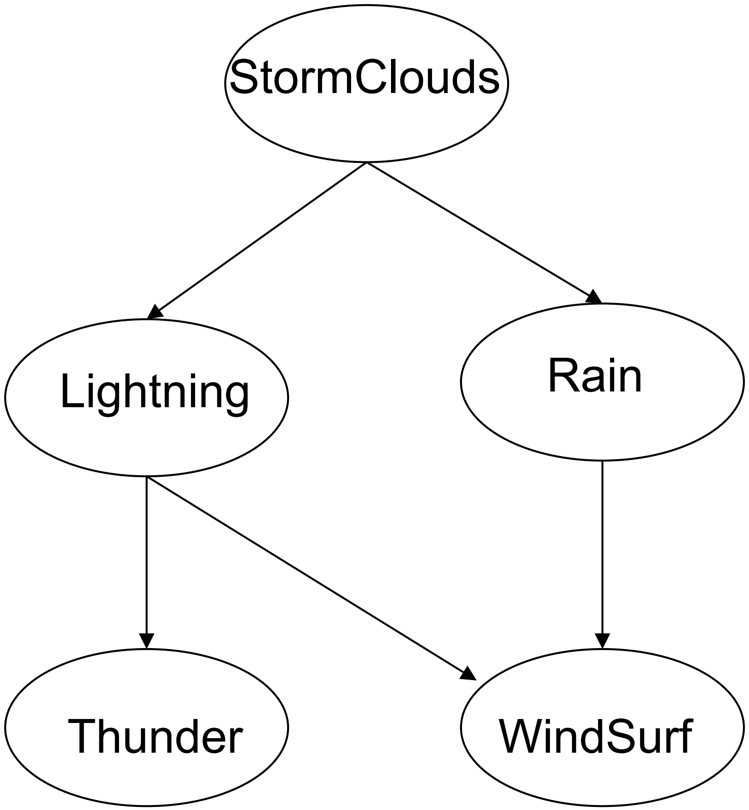


The joint distribution over all variables:

$$P(X_1 \ldots X_n) = \prod_i P(X_i | Pa(X_i))$$

Bayesian Network

Bayesian Network

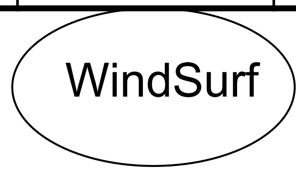


What can we say about conditional independencies in a Bayes Net?

One thing is this:

Each node is conditionally independent of its non-descendents, given only its immediate parents.

Parents	P(W|Pa)	P(¬W|Pa)
L, R	0	1.0
L, ¬R	0	1.0
¬L, R	0.2	0.8
¬L, ¬R	0.9	0.1



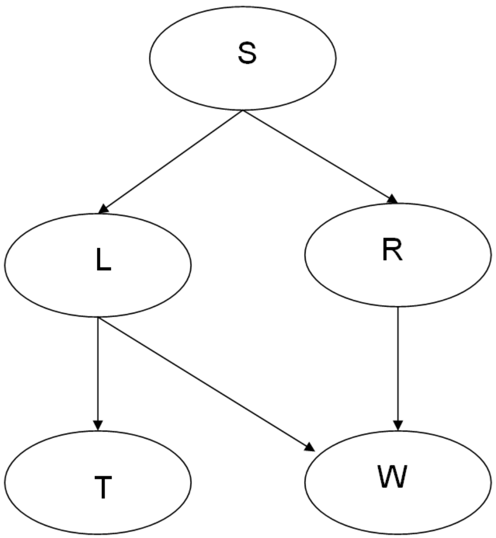
Some helpful terminology

Parents = Pa(X) = immediate parents

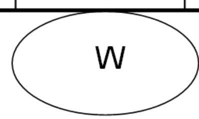
Antecedents = parents, parents of parents, ...

Children = immediate children

Descendents = children, children of children, ...

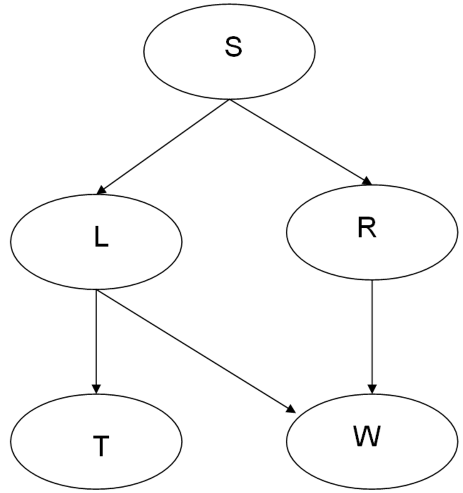


Parents	P(W|Pa)	P(¬W|Pa)
L, R	0	1.0
L, ¬R	0	1.0
¬L, R	0.2	0.8
¬L, ¬R	0.9	0.1

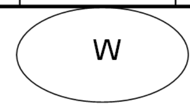


Bayesian Networks

- CPD for each node X_i describes $P(X_i \mid Pa(X_i))$



Parents	P(W|Pa)	P(¬W|Pa)
L, R	0	1.0
L, ¬R	0	1.0
¬L, R	0.2	0.8
¬L, ¬R	0.9	0.1

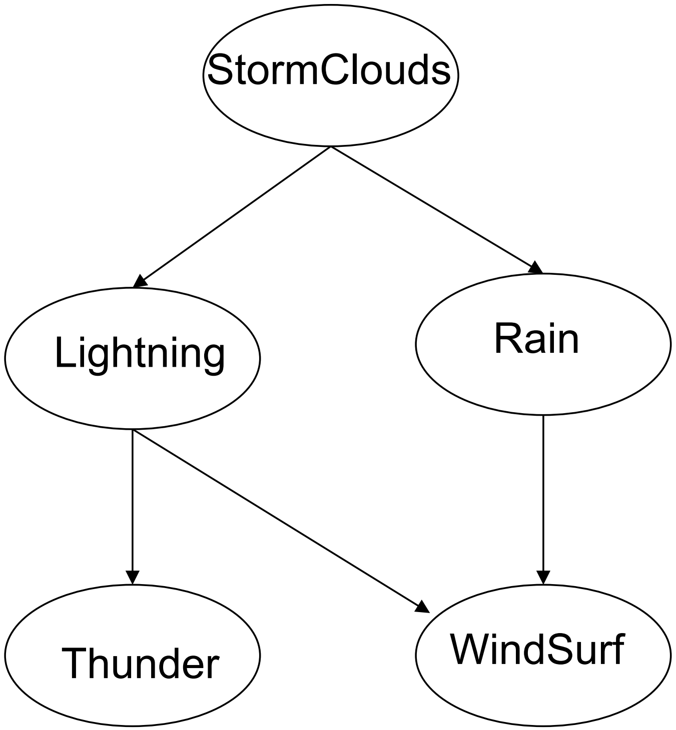


Chain rule of probability says that in general:

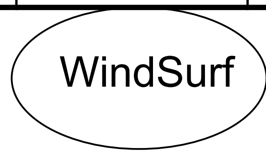
$$P(S,L,R,T,W) = P(S)P(L|S)P(R|S,L)P(T|S,L,R)P(W|S,L,R,T)$$

But in a Bayes net: $P(X_1 \ldots X_n) = \prod_i P(X_i|Pa(X_i))$

How many parameters?



Parents	P(W|Pa)	P(¬W|Pa)
L, R	0	1.0
L, ¬R	0	1.0
¬L, R	0.2	0.8
¬L, ¬R	0.9	0.1



To define joint distribution in general?

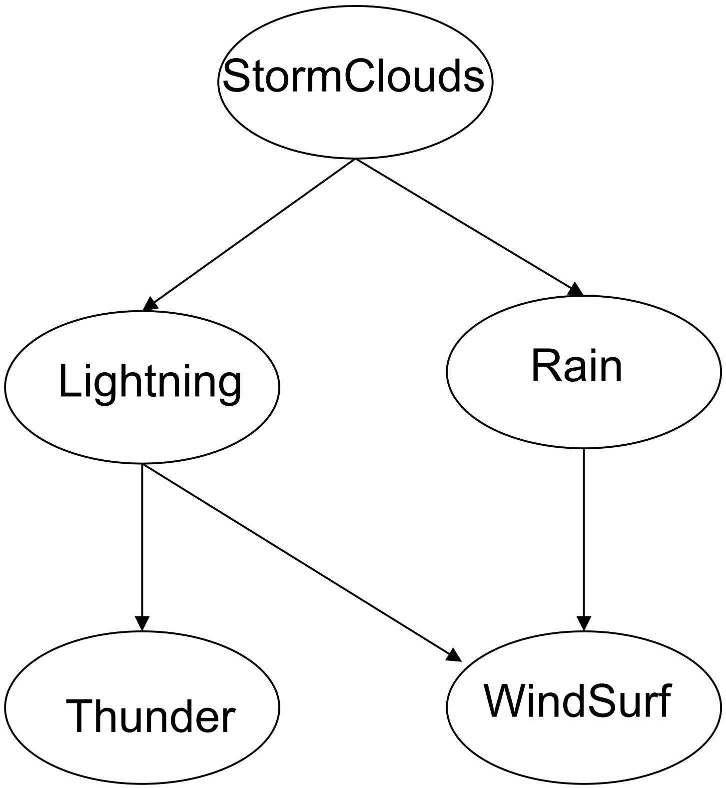
To define joint distribution for this Bayes Net?

Complexity of Bayesian Networks

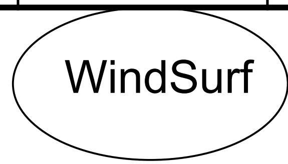
For n random Boolean variables:

- Full joint probability distribution: 2^n entries
- Bayesian network with at most k parents per node:
  - Each conditional probability table: at most 2^k entries
  - Entire network: $n\ 2^k$ entries

Inference in Bayes Nets

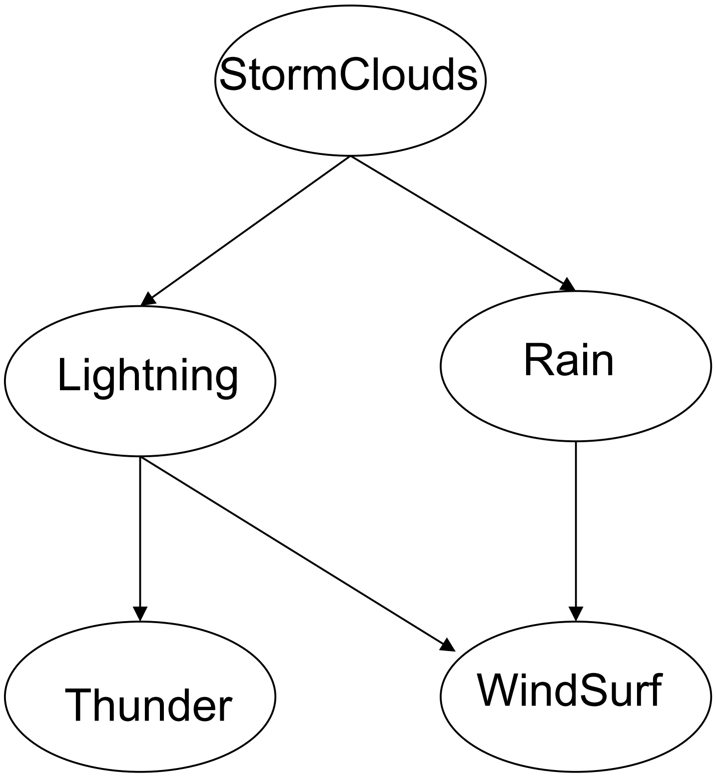


Parents	P(W|Pa)	P(¬W|Pa)
L, R	0	1.0
L, ¬R	0	1.0
¬L, R	0.2	0.8
¬L, ¬R	0.9	0.1

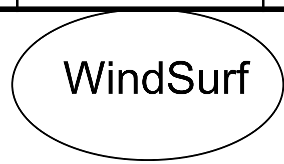


P(S=1, L=0, R=1, T=0, W=1) =

Learning a Bayes Net



Parents	P(W|Pa)	P(¬W|Pa)
L, R	0	1.0
L, ¬R	0	1.0
¬L, R	0.2	0.8
¬L, ¬R	0.9	0.1



Consider learning when graph structure is given, and data = { <s,l,r,t,w> }

What is the MLE solution? MAP?

Algorithm for Constructing Bayes Networks

- Choose an ordering over variables, e.g., $X_1, X_2, \ldots X_n$
- For i=1 to n
  - Add X_i to the network
  - Select parents $Pa(X_i)$ as minimal subset of $X_1 \ldots X_{i-1}$ such that

$$P(X_i | Pa(X_i)) = P(X_i | X_1, \ldots, X_{i-1})$$

Notice this choice of parents assures

$$P(X_1 \ldots X_n) = \prod_i P(X_i | X_1 \ldots X_{i-1}) \quad \text{(by chain rule)}$$

$$= \prod_i P(X_i | Pa(X_i)) \quad \text{(by construction)}$$

Example

- Bird flu and Allegies both cause Nasal problems
- Nasal problems cause Sneezes and Headaches

Example

- What is the Bayes Network for $X_1, \ldots, X_4$ with NO assumed conditional independencies?

What You Should Know

- Bayes nets are convenient representation for encoding dependencies / conditional independence
- BN = Graph plus parameters of CPD’s
  - Defines joint distribution over variables
  - Can calculate everything else from that
  - Though inference may be intractable
- Reading conditional independence relations from the graph
  - Each node is cond indep of non-descendents, given only its parents
  - ‘Explaining away’