10-708 Probabilistic Graphical Models

Machine Learning Department
School of Computer Science
Carnegie Mellon University

Directed Graphical Models

Matt Gormley
Lecture 2
Feb. 03, 2021

ROADMAP

Roadmap by Contrasts

- **Model:**
  - locally normalized vs. globally normalized
  - generative vs. discriminative
  - treewidth: high vs. low
  - cyclic vs. acyclic graphical models
  - exponential family vs. neural
  - deep vs. shallow (when viewed as neural network)
- **Inference:**
  - exact vs. approximate (and which models admit which)
  - dynamic programming vs. sampling vs. optimization
- **Inference problems:**
  - MAP vs. marginal vs. partition function
- **Learning:**
  - fully-supervised vs. partially-supervised (latent variable models) vs. unsupervised
  - partially-supervised vs. semi-supervised (missing some variable values vs. missing labels for entire instances)
  - loss-aware vs. not
  - probabilistic vs. non-probabilistic
  - frequentist vs. Bayesian

Roadmap by Example

***Whiteboard*:**
- Starting point: fully supervised HMM
- modifications to the model, inference, and learning
- corresponding technical terms of the result

Bayesian Networks

DIRECTED GRAPHICAL MODELS

Example: Ryan Reynolds Voicemail

From https://www.adweek.com/brand-marketing/ryan-reynolds-left-voicemails-for-all-mint-mobile-subscribers/

Example: Ryan Reynolds Voicemail

Images from imdb.com

Example: Ryan Reynolds' Voicemail

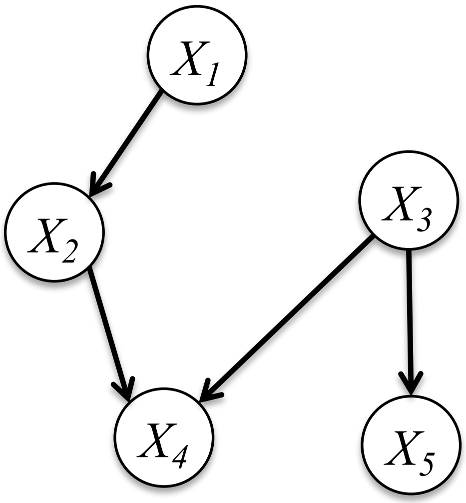
From https://www.adweek.com/brand-marketing/ryan-reynolds-left-voicemails-for-all-mint-mobile-subscribers/

Directed Graphical Models (Bayes Nets)

Whiteboard

- Example: Ryan Reynolds’ Voicemail
- Writing Joint Distributions
  - Idea #1: Giant Table
  - Idea #2: Rewrite using chain rule
  - Idea #3: Assume full independence
  - Idea #4: Drop variables from RHS of conditionals
- Definition: Bayesian Network

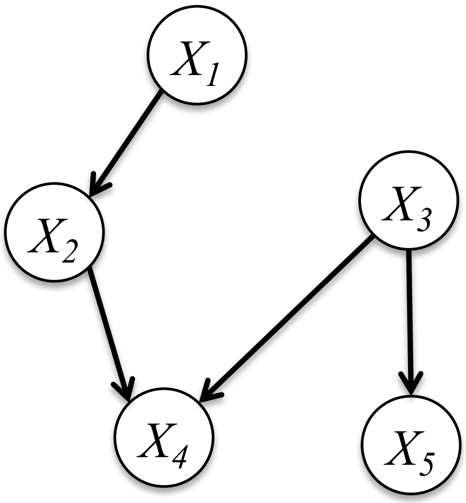
Bayesian Network



$$
\begin{aligned}
p(X_1, X_2, X_3, X_4, X_5) = \\
p(X_5|X_3)p(X_4|X_2, X_3) \\
p(X_3)p(X_2|X_1)p(X_1)
\end{aligned}
$$

Bayesian Network

Definition:



$$P(X_1 \ldots X_n) = \prod_{i=1}^{n} P(X_i \mid parents(X_i))$$

- A Bayesian Network is a **directed graphical model**
- It consists of a directed acyclic graph (DAG) **G** and the conditional probabilities **P**
- These two parts full specify the distribution:
  - Qualitative Specification: **G**
  - Quantitative Specification: **P**

Bayesian Networks & DAGs

Suppose we have an arbitrary directed graph G over T variables X_i and define the following product:

$$P_{\text{fact}}(\mathbf{X}) = \prod_{i=1}^{T} P(X_i | \text{parents}(X_i))$$

- **Proposition**: The function $P_{\text{fact}}(\mathbf{X})$ is a valid joint distribution when G is a DAG
- **Proof**: Let X_s be a leaf node. By our factorization we have that,

  $$P_{\text{fact}}(\mathbf{X}) = P(X_s | \text{parents}(X_s)) P_{\text{fact}}(\text{parents}(X_s))$$

  By induction, if $P_{\text{fact}}(\text{parents}(X_s))$ is a valid joint distribution then $P_{\text{fact}}(\mathbf{X})$ is a valid joint distrubution.

Qualitative Specification

- Where does the qualitative specification come from?
  - Prior knowledge of causal relationships
  - Prior knowledge of modular relationships
  - Assessment from experts
  - Learning from data (i.e. structure learning)
  - We simply prefer a certain architecture (e.g. a layered graph)
  - …

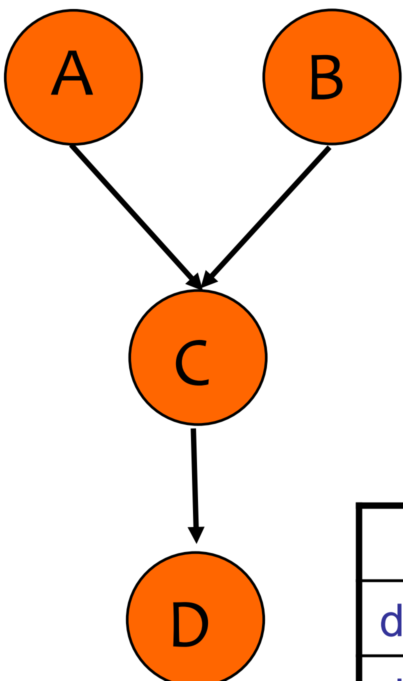
Quantitative Specification

Example: Conditional probability tables (CPTs) for discrete random variables

$$P(a,b,c.d) = P(a)P(b)P(c|a,b)P(d|c)$$

a^0	0.75
a^1	0.25

b^0	0.33
b^1	0.67



	a^0b^0	a^0b^1	a^1b^0	a^1b^1
c^0	0.45	1	0.9	0.7
c^1	0.55	0	0.1	0.3

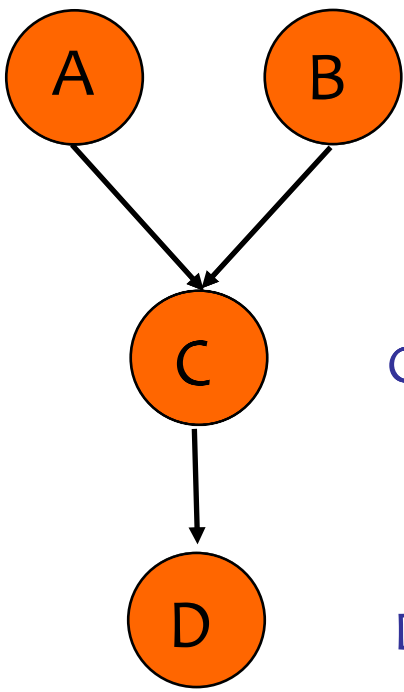
	c^0	c^1
d^0	0.3	0.5
d^1	07	0.5

Quantitative Specification

Example: Conditional probability density functions (CPDs) for continuous random variables

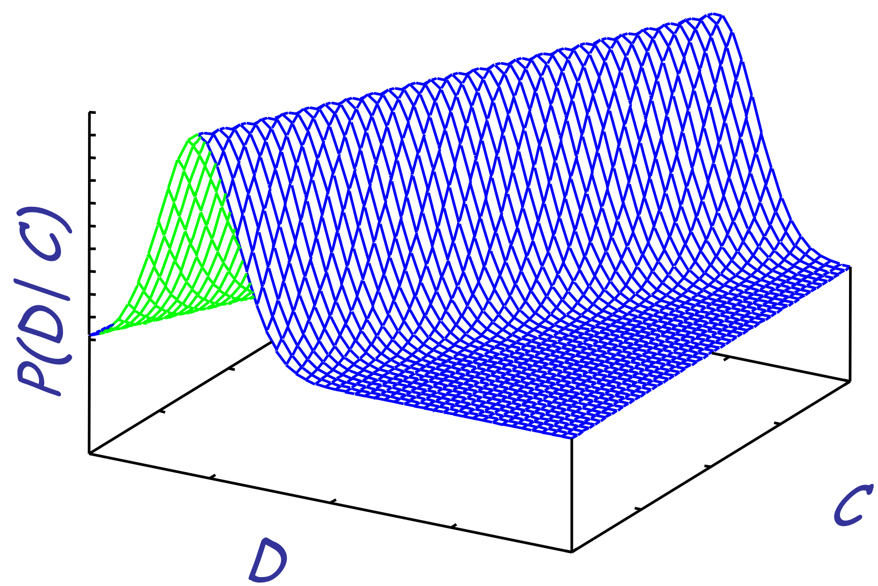
$A \sim N(\mu_a, \Sigma_a)$ $B \sim N(\mu_b, \Sigma_b)$

$$P(a,b,c.d) = P(a)P(b)P(c|a,b)P(d|c)$$



$C \sim N(A+B, \Sigma_c)$

$D \sim N(\mu_d + C, \Sigma_d)$



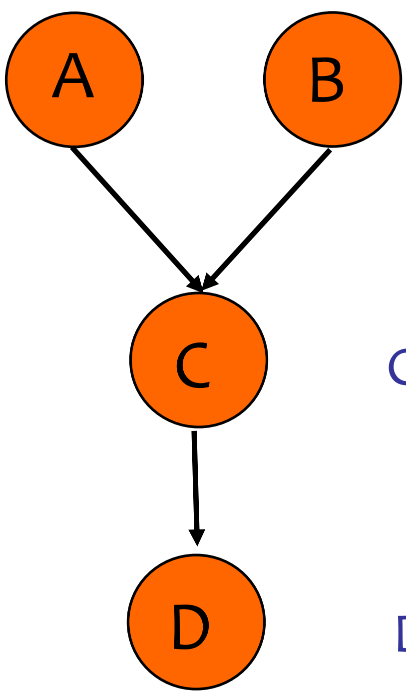
Quantitative Specification

Example: Combination of CPTs and CPDs
for a mix of discrete and continuous variables

a^0	0.75
a^1	0.25

b^0	0.33
b^1	0.67

$$P(a,b,c.d) = P(a)P(b)P(c|a,b)P(d|c)$$



$C \sim N(A+B, \Sigma_c)$

$D \sim N(\mu_d + C, \Sigma_d)$

Compactness of a BayesNet

Consider random variables $X_1, X_2, \dots, X_T$ where $X_i \in \mathcal{X}$, where $|\mathcal{X}| = R$

- To represent an arbitrary distribution $P(\mathbf{X})$ via a single joint probability table requires $R^T - 1$ values

Exponential in T

- If the distribution factors according to a graph G and $\max_{X_i} |\text{parents}(X_i)| \leq D$

  then each $P(X_i \mid \text{parents}(X_i))$ needs only $R^D(R - 1)$ values for a total of only $T(R^D(R - 1))$ values

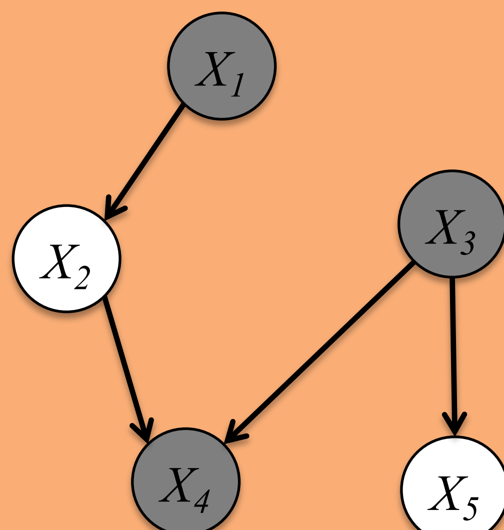
Linear in T

Observed Variables

- In a graphical model, **shaded nodes** are "**observed**", i.e. their values are given

Example:

$$P(X_2, X_5 \mid X_1 = 0, X_3 = 1, X_4 = 1)$$



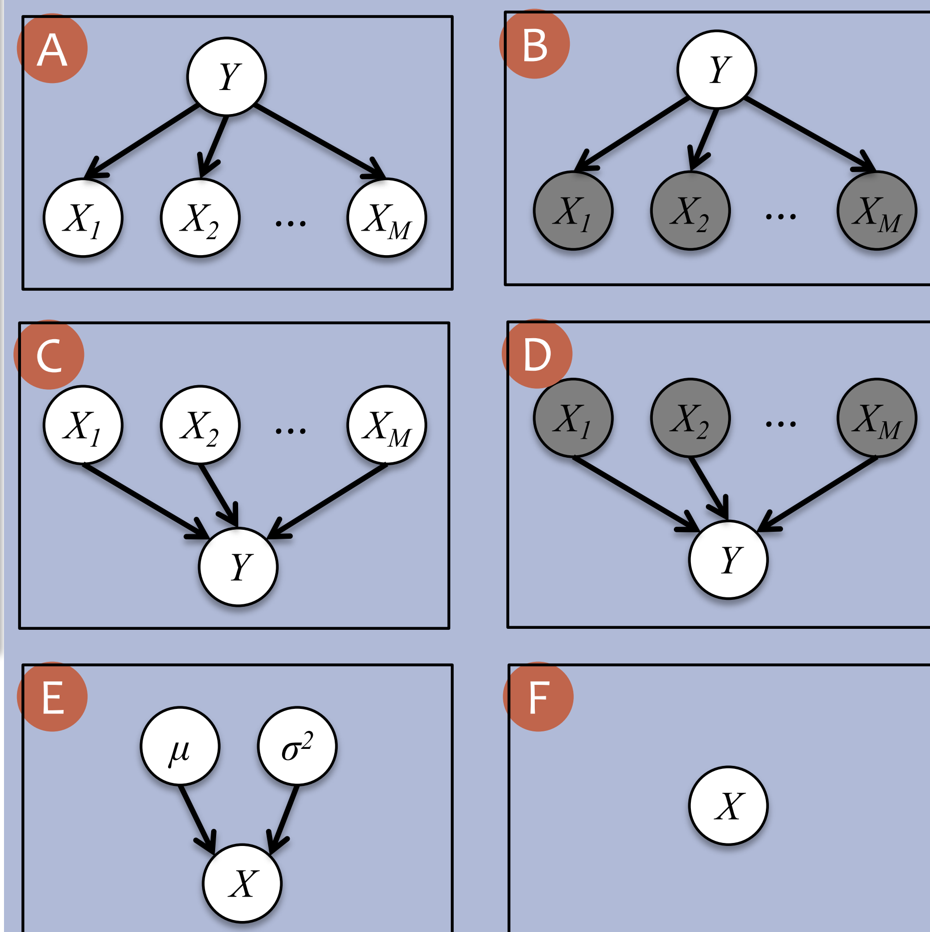
Familiar Models as Bayesian Networks

Question:

Match the model name to the corresponding Bayesian Network

1. Logistic Regression
2. Linear Regression
3. Bernoulli Naïve Bayes
4. Gaussian Naïve Bayes
5. 1D Gaussian

Answer:



GRAPHICAL MODELS: DETERMINING CONDITIONAL INDEPENDENCIES

What Independencies does a Bayes Net Model?

- In order for a Bayesian network to model a probability distribution, the following must be true:

  Each variable is conditionally independent of all its non-descendants in the graph given the value of all its parents.

- This follows from

$$P(X_1 \ldots X_n) = \prod_{i=1}^{n} P(X_i \mid parents(X_i))$$

$$= \prod_{i=1}^{n} P(X_i \mid X_1 \ldots X_{i-1})$$

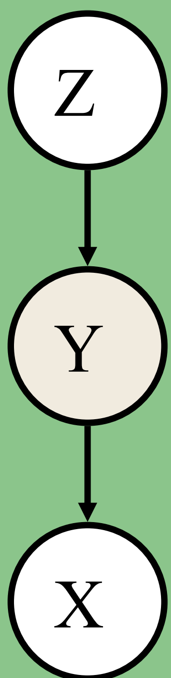
- But what else does it imply?

Slide from William Cohen

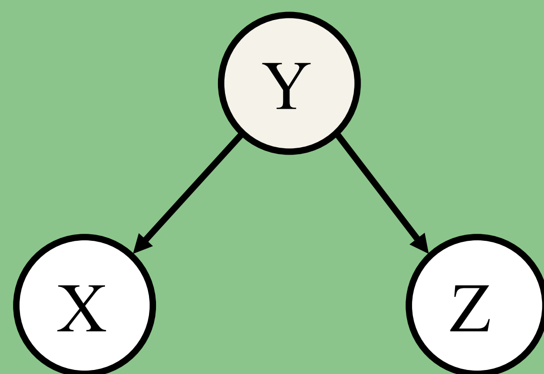
What Independencies does a Bayes Net Model?

Three cases of interest…

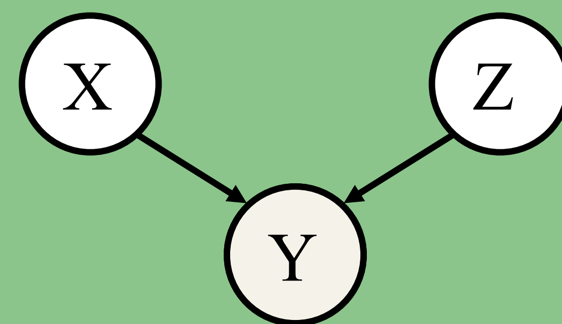
Cascade



Common Parent



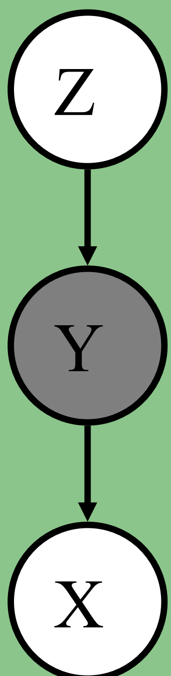
V-Structure



What Independencies does a Bayes Net Model?

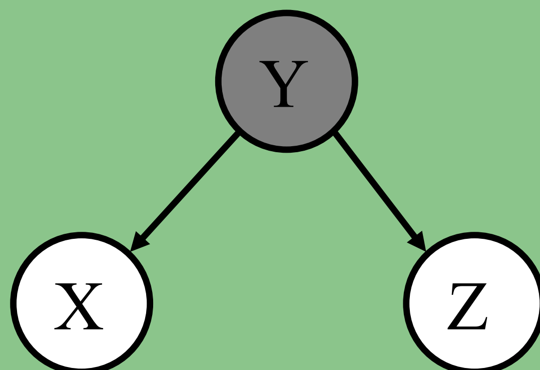
Three cases of interest…

Cascade



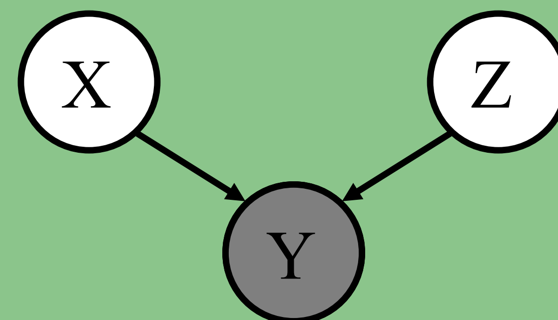
$$X \perp\!\!\!\perp Z \mid Y$$

Common Parent



$$X \perp\!\!\!\perp Z \mid Y$$

V-Structure



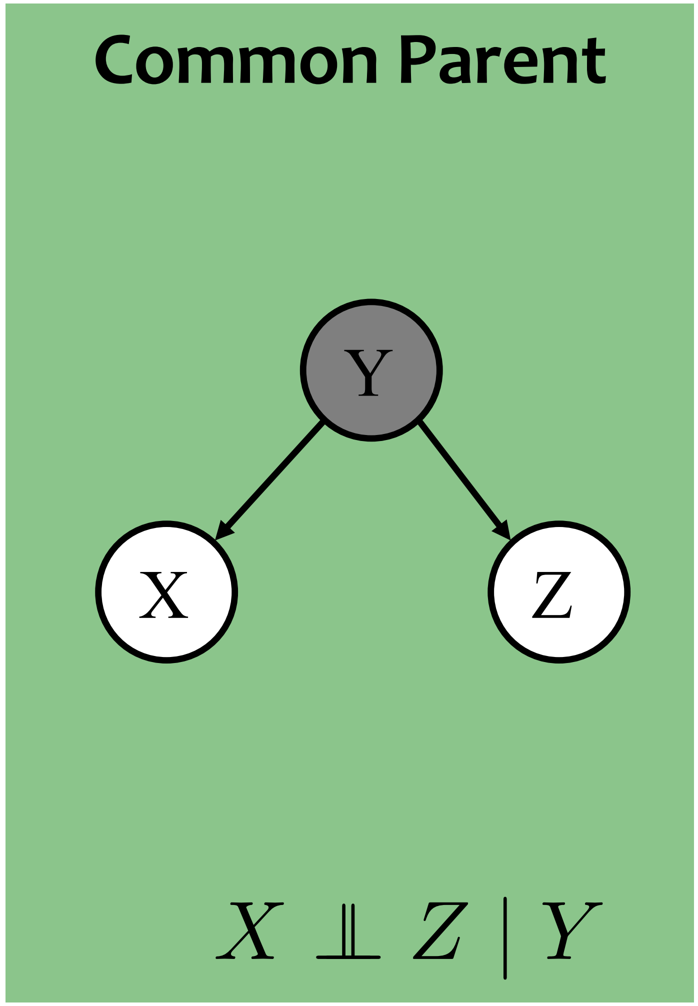
$$X \not\perp\!\!\!\perp Z \mid Y$$

Knowing Y
decouples X and Z

Knowing Y
couples X and Z

Whiteboard

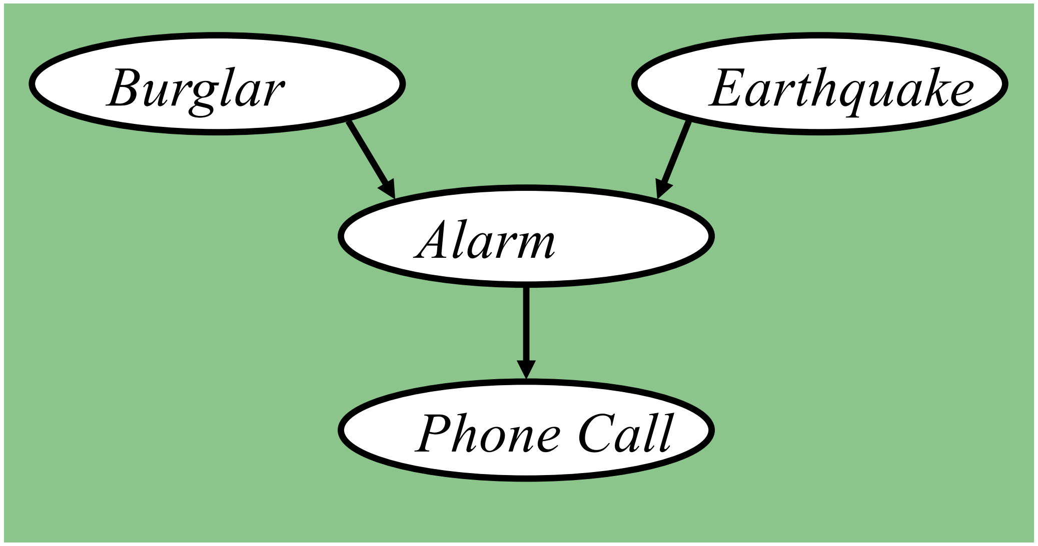
Proof of conditional independence



(The other two cases can be shown just as easily.)

The "Burglar Alarm" example

- Your house has a twitchy burglar alarm that is also sometimes triggered by earthquakes.
- Earth arguably doesn't care whether your house is currently being burgled
- While you are on vacation, one of your neighbors calls and tells you your home's burglar alarm is ringing. Uh oh!



Quiz: True or False?

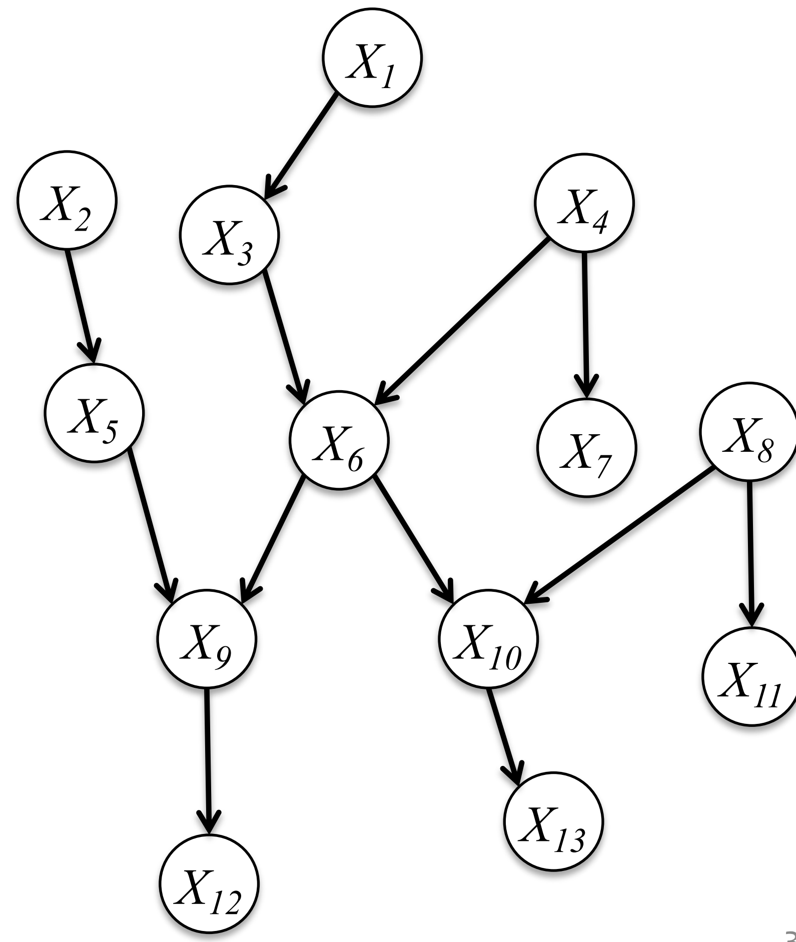
$$Burglar \perp\!\!\!\perp Earthquake \mid PhoneCall$$

Slide from William Cohen

Markov Blanket (Directed)

Def: the **co-parents** of a node are the parents of its children

Def: the **Markov Blanket** of a node in a directed graphical model is the set containing the node's parents, children, and co-parents.

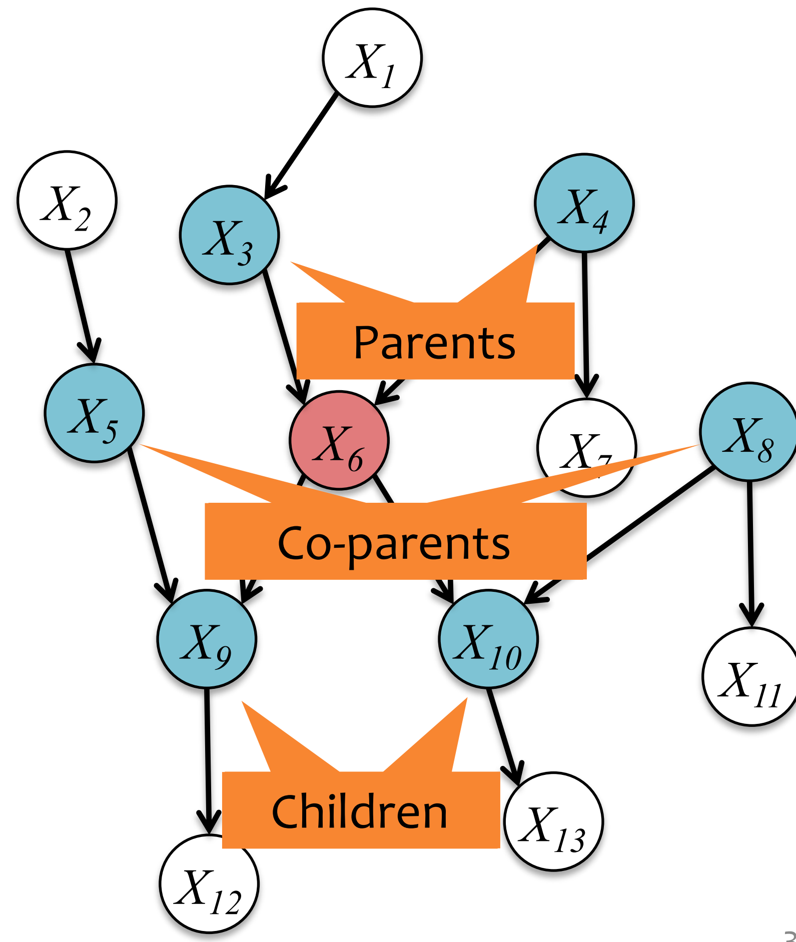


Markov Blanket (Directed)

Def: the **co-parents** of a node are the parents of its children

Def: the **Markov Blanket** of a node in a directed graphical model is the set containing the node's parents, children, and co-parents.

Example: The Markov Blanket of X_6 is $\{X_3, X_4, X_5, X_8, X_9, X_{10}\}$



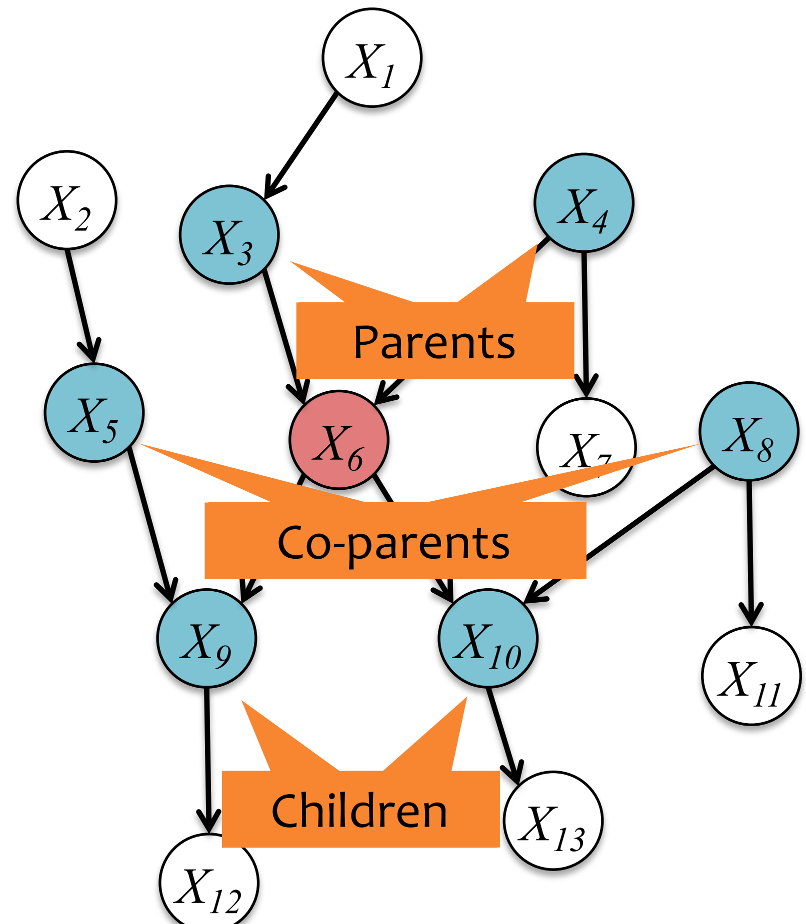
Markov Blanket (Directed)

Def: the **co-parents** of a node are the parents of its children

Def: the **Markov Blanket** of a node in a directed graphical model is the set containing the node's parents, children, and co-parents.

Theorem: a node is **conditionally independent** of every other node in the graph given its **Markov blanket**

Example: The Markov Blanket of X_6 is $\{X_3, X_4, X_5, X_8, X_9, X_{10}\}$



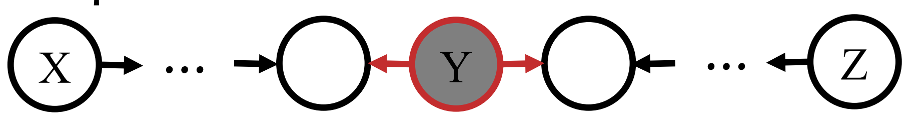
D-Separation

If variables X and Z are **d-separated** given a **set** of variables E
Then X and Z are **conditionally independent** given the **set** E

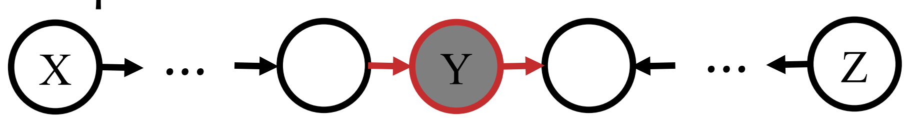
<u>Definition #1:</u>
Variables X and Z are **d-separated** given a **set** of evidence variables E iff every path from X to Z is "blocked".

A path is "blocked" whenever:

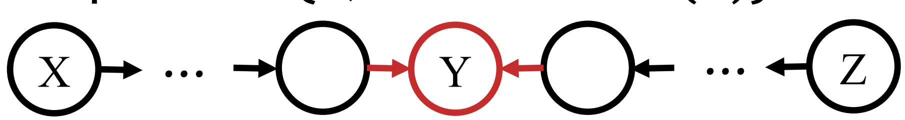
1. ∃Y on path s.t. Y ∈ E and Y is a "common parent"



2. ∃Y on path s.t. Y ∈ E and Y is in a "cascade"



3. ∃Y on path s.t. {Y, descendants(Y)} ∉ E and Y is in a "v-structure"



D-Separation

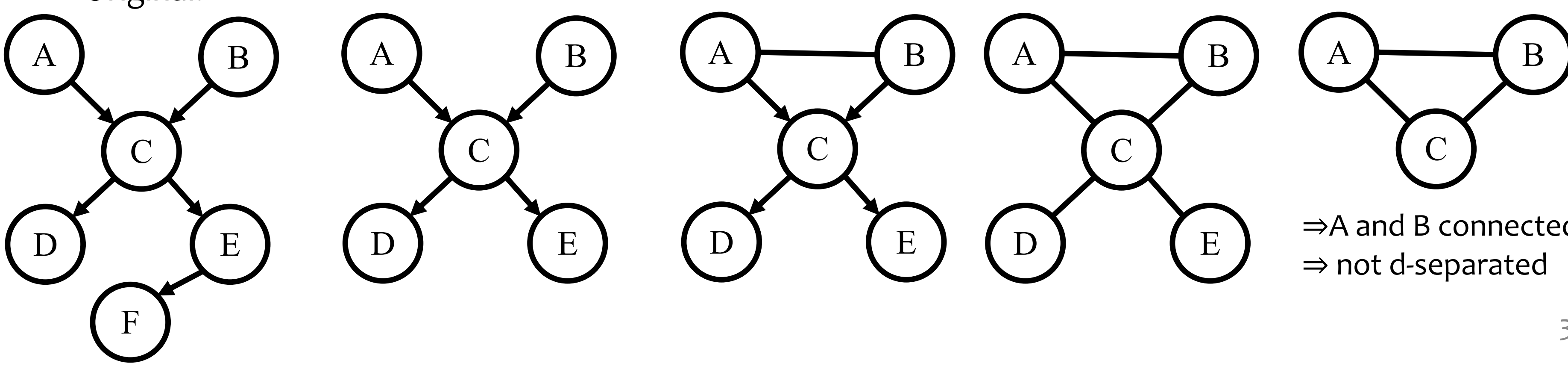
If variables X and Z are **d-separated** given a **set** of variables E
Then X and Z are **conditionally independent** given the **set** E

Definition #2:
Variables X and Z are **d-separated** given a **set** of evidence variables E iff there does **not** exist a path in the **undirected ancestral moral** graph **with E removed**.

1. **Ancestral graph**: keep only X, Z, E and their ancestors
2. **Moral graph**: add undirected edge between all pairs of each node's parents
3. **Undirected graph**: convert all directed edges to undirected
4. Givens Removed: delete any nodes in E

Example Query: $A \perp\!\!\!\perp B \mid \{D, E\}$

Original: Ancestral: Moral: Undirected: Givens Removed:



⇒A and B connected
⇒ not d-separated

SUPERVISED LEARNING FOR BAYES NETS

Recipe for Closed-form MLE

1. Assume data was generated i.i.d. from some model (i.e. write the generative story)

   $$x^{(i)} \sim p(x|\boldsymbol{\theta})$$

2. Write log-likelihood

   $$\ell(\boldsymbol{\theta}) = \log p(x^{(1)}|\boldsymbol{\theta}) + \ldots + \log p(x^{(N)}|\boldsymbol{\theta})$$

3. Compute partial derivatives (i.e. gradient)

   $$\partial\ell(\boldsymbol{\theta})/\partial\theta_1 = \ldots$$
   $$\partial\ell(\boldsymbol{\theta})/\partial\theta_2 = \ldots$$
   $$\ldots$$
   $$\partial\ell(\boldsymbol{\theta})/\partial\theta_M = \ldots$$

4. Set derivatives to zero and solve for $\boldsymbol{\theta}$

   $$\partial\ell(\boldsymbol{\theta})/\partial\theta_m = 0 \text{ for all } m \in \{1, \ldots, M\}$$
   $$\boldsymbol{\theta}^{MLE} = \text{solution to system of M equations and M variables}$$

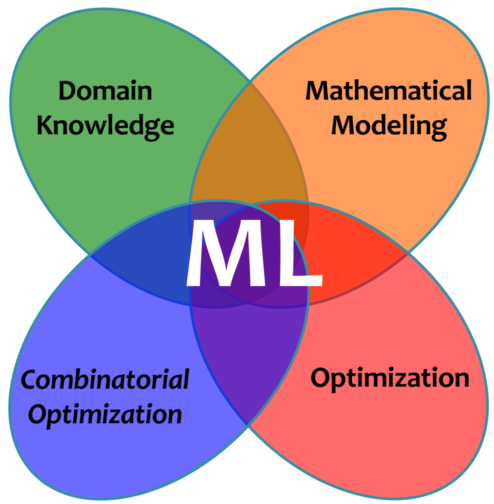
5. Compute the second derivative and check that $\ell(\boldsymbol{\theta})$ is concave down at $\boldsymbol{\theta}^{MLE}$

Machine Learning

The **data** inspires the structures we want to predict

Our **model** defines a score for each structure

It also tells us what to optimize



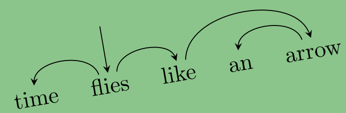
Inference finds {best structure, marginals, partition function} for a new observation

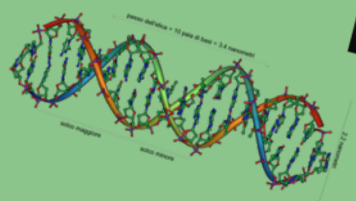
(**Inference** is usually called as a subroutine in learning)

Learning tunes the parameters of the model

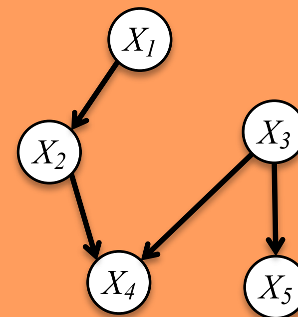
Machine Learning

Data

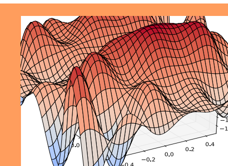




Model



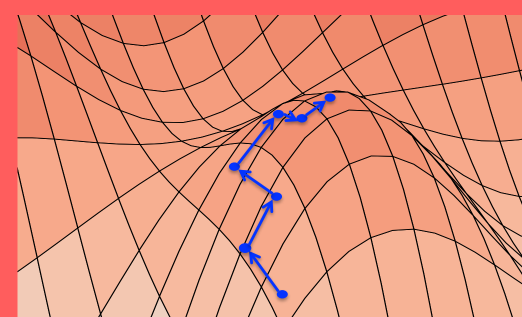
Objective



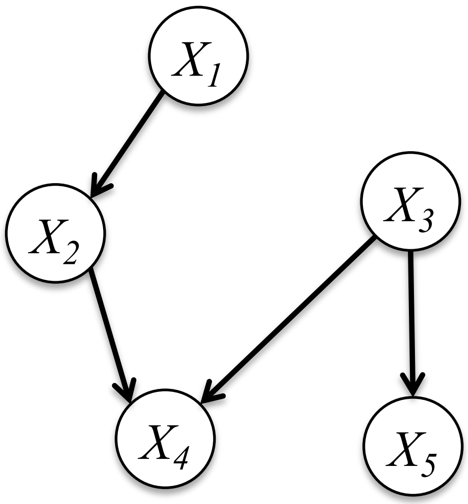
Inference

(**Inference** is usually called as a subroutine in learning)

Learning

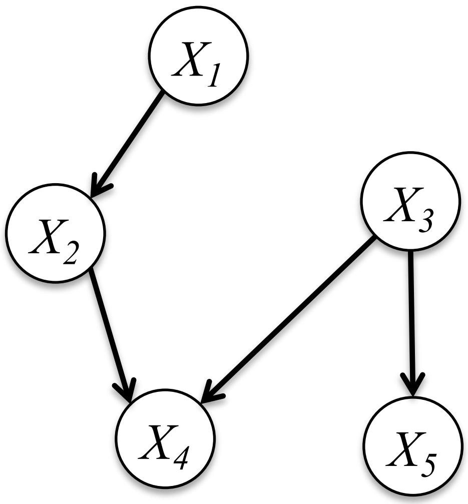


Learning Fully Observed BNs



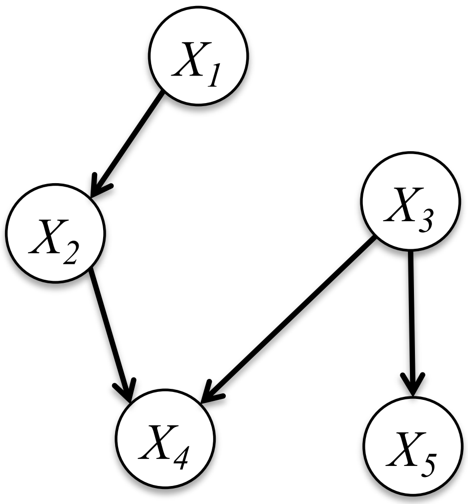
$$p(X_1, X_2, X_3, X_4, X_5) =$$
$$p(X_5|X_3)p(X_4|X_2, X_3)$$
$$p(X_3)p(X_2|X_1)p(X_1)$$

Learning Fully Observed BNs



$$p(X_1, X_2, X_3, X_4, X_5) =$$
$$p(X_5|X_3)p(X_4|X_2, X_3)$$
$$p(X_3)p(X_2|X_1)p(X_1)$$

Learning Fully Observed BNs



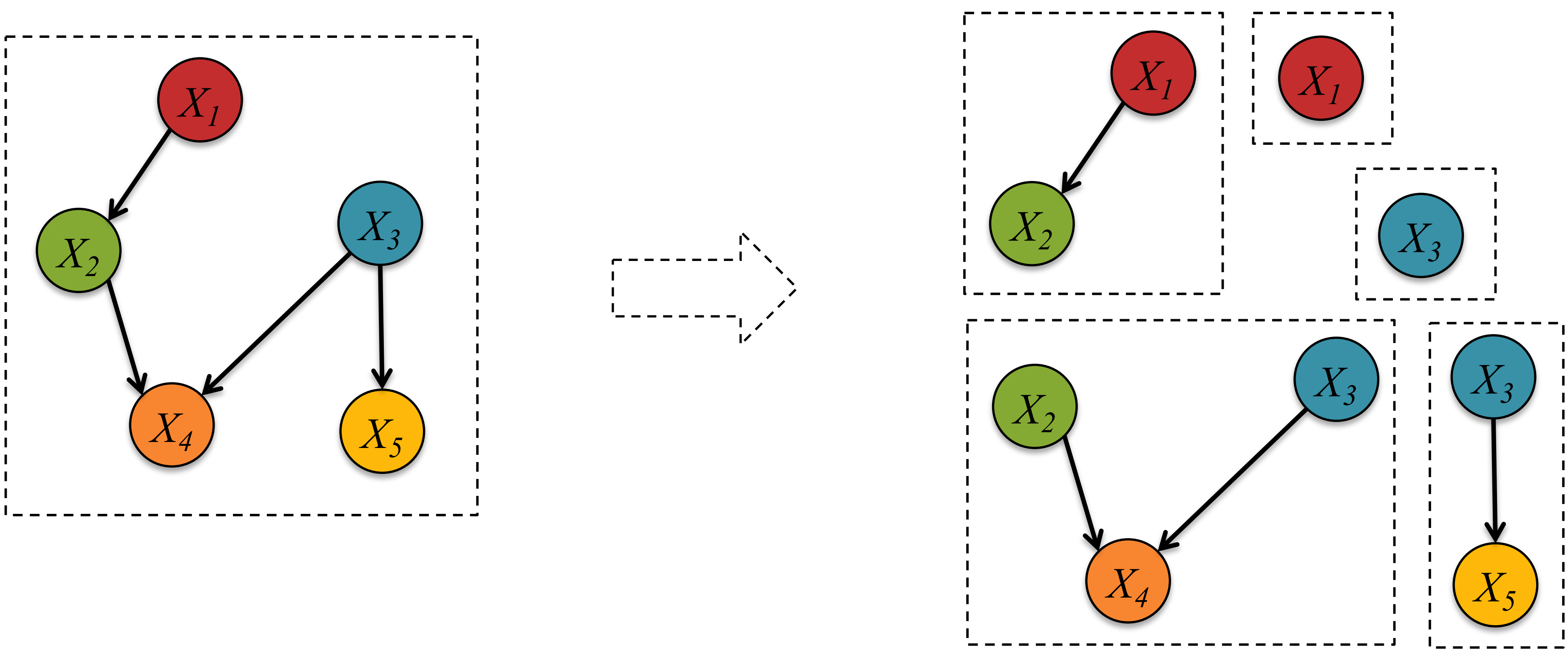
$$
\begin{aligned}
p(X_1, X_2, X_3, X_4, X_5) = \\
& p(X_5|X_3)p(X_4|X_2, X_3) \\
& p(X_3)p(X_2|X_1)p(X_1)
\end{aligned}
$$

How do we learn these conditional and marginal distributions for a Bayes Net?

Learning Fully Observed BNs

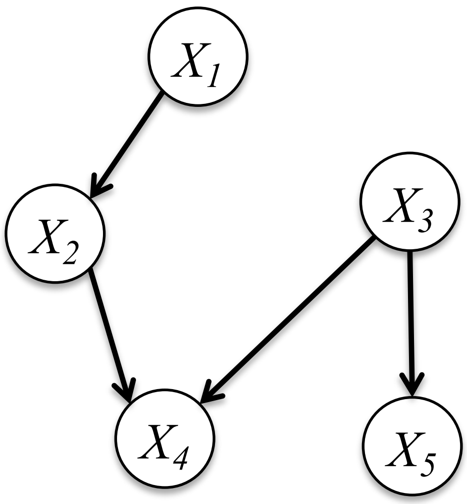
Learning this fully observed Bayesian Network is **equivalent** to learning five (small / simple) independent networks from the same data

$$p(X_1, X_2, X_3, X_4, X_5) = p(X_5|X_3)p(X_4|X_2, X_3) p(X_3)p(X_2|X_1)p(X_1)$$



Learning Fully Observed BNs

How do we **learn** these conditional and marginal distributions for a Bayes Net?



$$\begin{aligned}\boldsymbol{\theta}^* &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log p(X_1, X_2, X_3, X_4, X_5) \\ &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log p(X_5|X_3, \theta_5) + \log p(X_4|X_2, X_3, \theta_4) \\ &\qquad + \log p(X_3|\theta_3) + \log p(X_2|X_1, \theta_2) \\ &\qquad + \log p(X_1|\theta_1)\end{aligned}$$

$$\theta_1^* = \underset{\theta_1}{\operatorname{argmax}} \log p(X_1|\theta_1)$$

$$\theta_2^* = \underset{\theta_2}{\operatorname{argmax}} \log p(X_2|X_1, \theta_2)$$

$$\theta_3^* = \underset{\theta_3}{\operatorname{argmax}} \log p(X_3|\theta_3)$$

$$\theta_4^* = \underset{\theta_4}{\operatorname{argmax}} \log p(X_4|X_2, X_3, \theta_4)$$

$$\theta_5^* = \underset{\theta_5}{\operatorname{argmax}} \log p(X_5|X_3, \theta_5)$$

Example: Tornado Alarms

1. Imagine that you work at the 911 call center in Dallas
2. You receive six calls informing you that the Emergency Weather Sirens are going off
3. What do you conclude?

Example: Tornado Alarms

Hacking Attack Woke Up Dallas With Emergency Sirens, Officials Say

By ELI ROSENBERG and MAYA SALAM APRIL 8, 2017

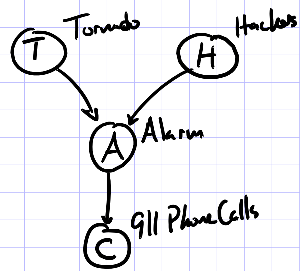
Warning sirens in Dallas, meant to alert the public to emergencies like severe weather, started sounding around 11:40 p.m. Friday, and were not shut off until 1:20 a.m. Rex C. Curry for The New York Times

1. Imagine that you work at the 911 call center in Dallas
2. You receive six calls informing you that the Emergency Weather Sirens are going off
3. What do you conclude?

Figure from https://www.nytimes.com/2017/04/08/us/dallas-emergency-sirens-hacking.html

Learning Fully Observed BNs

Ex: Tornado Alarms



$H \sim \text{Bernoulli}(\eta)$
$T \sim \text{Bernoulli}(\tau)$
$A \sim \text{Bernoulli}(\alpha_{H,T})$
$C \sim \text{Uniform}(\{1,\dots,6\}) + A * \text{Uniform}(\{1,\dots,6\})$

parameters

no parameters

integer

Dataset

i	T	H	A	C
1	0	0	0	2
2	0	0	0	6
3	0	0	0	4
⋮	1	0	0	3
	1	0	0	1
	1	0	1	10
	1	0	1	7
	0	1	0	2
	0	1	1	12
	0	1	0	5
⋮	1	1	1	10
12	1	0	0	2

MLEs in Closed Form

$$\ell(\eta,\tau,\alpha) = \log \prod_{i=1}^{12} p(t^{(i)}, h^{(i)}, a^{(i)}, c^{(i)} \mid \eta, \tau, \alpha)$$

$$= \sum_{i=1}^{12} \log p(t^{(i)} \mid \tau) + \log p(h^{(i)} \mid \eta) + \log p(a^{(i)} \mid t^{(i)}, h^{(i)}, \alpha) + \log p(c^{(i)} \mid a^{(i)})$$

$$\hat{\eta}, \hat{\tau}, \hat{\alpha} = \operatorname{argmax} \ell(\eta,\tau,\alpha)$$

$$\hat{\eta} = \operatorname*{argmax}_{\eta} \sum_{i=1}^{12} \log p(h^{(i)} \mid \eta) = \#(T=1)/N$$

$$\hat{\tau} = \operatorname*{argmax}_{\tau} \sum_{i=1}^{12} \log p(t^{(i)} \mid \tau) = \#(H=1)/N$$

$$\hat{\alpha} = \operatorname*{argmax}_{\alpha} \sum_{i=1}^{12} \log p(a^{(i)} \mid t^{(i)}, h^{(i)}, \alpha)$$

$$\hat{\alpha}_{t,h} = \frac{\#(A=1, T=t, H=h)}{\#(T=t, H=h)}$$

What are the MLEs?

$\hat{\eta} = 1/3$

$\hat{\tau} = 1/2$

$\hat{\alpha} =$

	H=0	H=1
T=0	0	1/3
T=1	2/3	1

INFERENCE FOR BAYESIAN NETWORKS

A Few Problems for Bayes Nets

Suppose we already have the parameters of a Bayesian Network…

1. How do we compute the probability of a specific assignment to the variables?
   $P(T=t, H=h, A=a, C=c)$

2. How do we draw a sample from the joint distribution?
   $t,h,a,c \sim P(T, H, A, C)$

3. How do we compute marginal probabilities?
   $P(A) = \ldots$

4. How do we draw samples from a conditional distribution?
   $t,h,a \sim P(T, H, A \mid C = c)$

5. How do we compute conditional marginal probabilities?
   $P(H \mid C = c) = \ldots$

Learning Objectives

Bayesian Networks

You should be able to…

1. Identify the conditional independence assumptions given by a generative story or a specification of a joint distribution
2. Draw a Bayesian network given a set of conditional independence assumptions
3. Define the joint distribution specified by a Bayesian network
4. User domain knowledge to construct a (simple) Bayesian network for a real-world modeling problem
5. Depict familiar models as Bayesian networks
6. Use d-separation to prove the existence of conditional independencies in a Bayesian network
7. Employ a Markov blanket to identify conditional independence assumptions of a graphical model
8. Develop a supervised learning algorithm for a Bayesian network