10-418 / 10-618 Machine Learning for Structured Data

Machine Learning Department
School of Computer Science
Carnegie Mellon University

Directed Graphical Models
+
Undirected Graphical Models

Matt Gormley
Lecture 7
Sep. 18, 2019

Q&A

Q: How will I earn the 5% participation points?

A: Very gradually. There will be a few aspects of the course (polls, surveys, meetings with the course staff) that we will attach participation points to.

That said, we might not actually use the whole 5% that is being held out.

Q&A

Q: When should I prefer a directed graphical model to an undirected graphical model?

A: As we'll see today, the primary differences between them are:

1. the conditional independence assumptions they define
2. the normalization assumptions they make (Bayes Nets are locally normalized)

(That said, we'll also tie them together via a single framework: factor graphs.)

There are also some practical differences (e.g. ease of learning) that result from the locally vs. globally normalized difference.

Reminders

- **Homework 1: DAgger for seq2seq**
  - Out: Thu, Sep. 12
  - Due: Thu, Sep. 26 at 11:59pm

SUPERVISED LEARNING FOR BAYES NETS

Recipe for Closed-form MLE

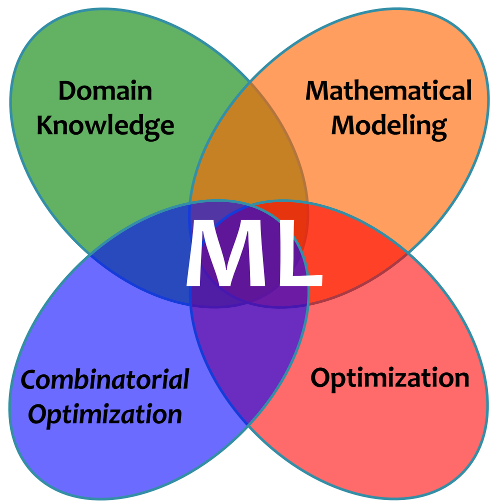
1. Assume data was generated i.i.d. from some model (i.e. write the generative story)
   $$x^{(i)} \sim p(x|\boldsymbol{\theta})$$
2. Write log-likelihood
   $$\ell(\boldsymbol{\theta}) = \log p(x^{(1)}|\boldsymbol{\theta}) + \ldots + \log p(x^{(N)}|\boldsymbol{\theta})$$
3. Compute partial derivatives (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 $\boldsymbol{\theta}$
   $$\partial\ell(\boldsymbol{\theta})/\partial\theta_m = 0 \text{ for all } m \in \{1, \ldots, M\}$$
   $\boldsymbol{\theta}^{MLE}$ = 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



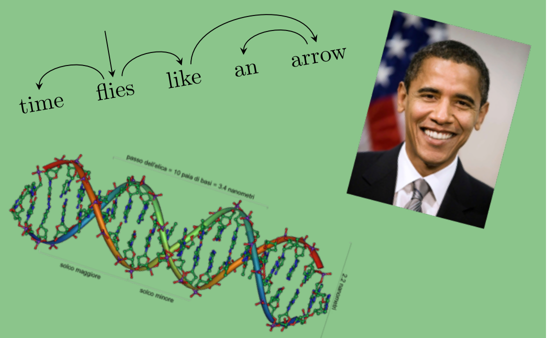
Learning tunes the parameters of the model

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

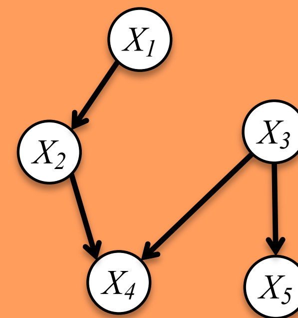
(**Inference** is usually called as a subroutine in learning)

Machine Learning

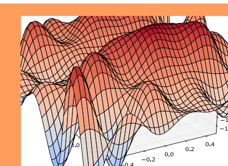
Data



Model



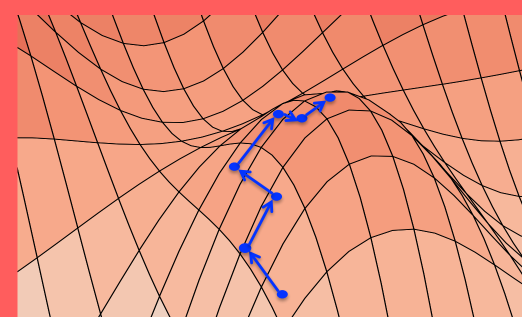
Objective



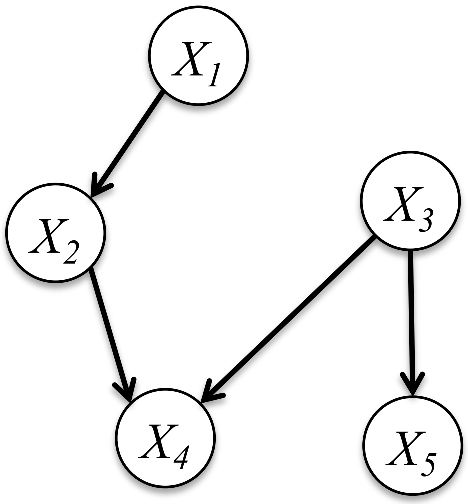
Inference

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

Learning

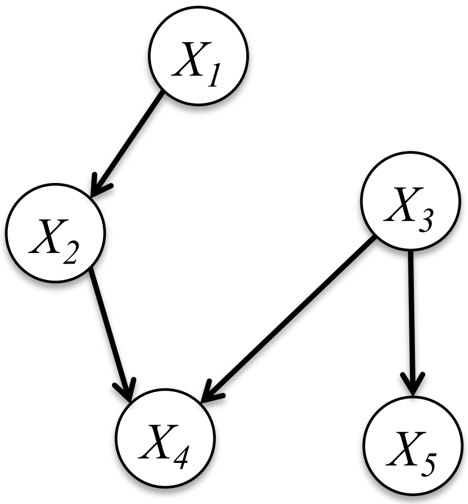


Learning Fully Observed BNs



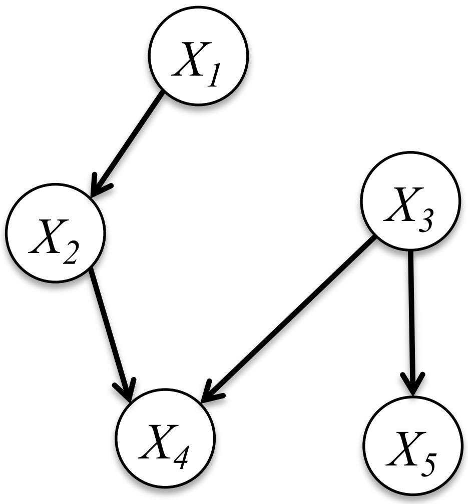
$$p(X_1, X_2, X_3, X_4, X_5) = \\ p(X_5|X_3)p(X_4|X_2, X_3) \\ p(X_3)p(X_2|X_1)p(X_1)$$

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



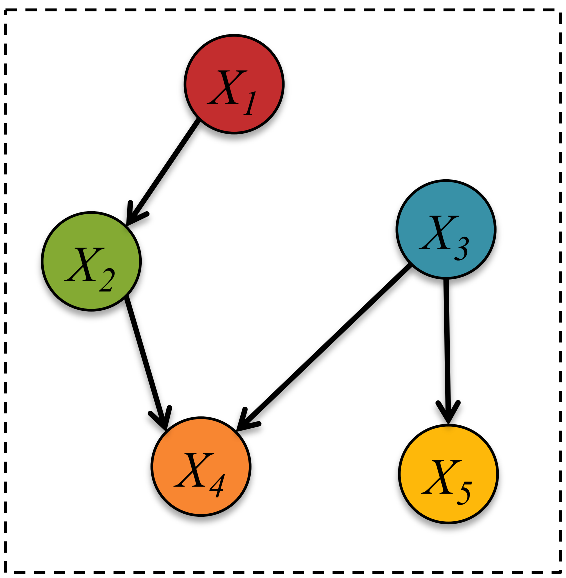
$$p(X_1, X_2, X_3, X_4, X_5) =$$
$$p(X_5|X_3)p(X_4|X_2, X_3)$$
$$p(X_3)p(X_2|X_1)p(X_1)$$

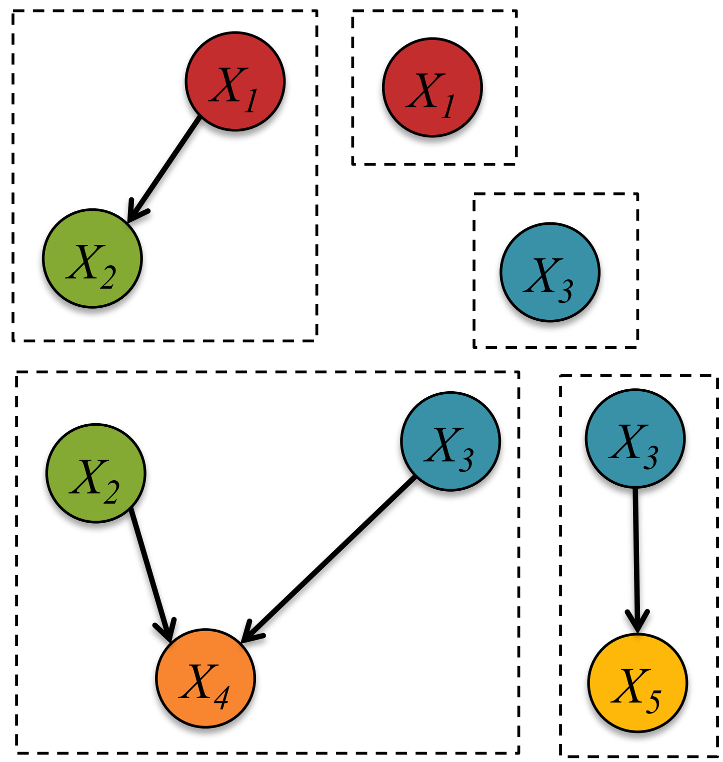
How do we learn these conditional and marginal distributions for a Bayes Net?

Learning Fully Observed BNs

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

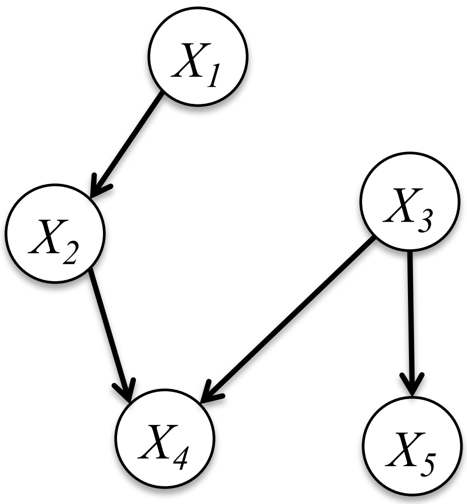
$$p(X_1, X_2, X_3, X_4, X_5) =$$
$$p(X_5|X_3)p(X_4|X_2, X_3)$$
$$p(X_3)p(X_2|X_1)p(X_1)$$





Learning Fully Observed BNs

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



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

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

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

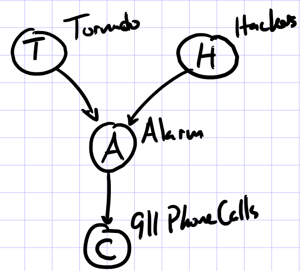
$$\theta_3^* = \underset{\theta_3}{\operatorname{argmax}} \log p(X_3|\theta_3)$$

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

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

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?
   $P(H \mid C = c) = \ldots$

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

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

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

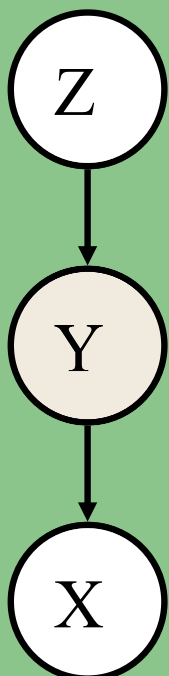
- But what else does it imply?

Slide from William Cohen

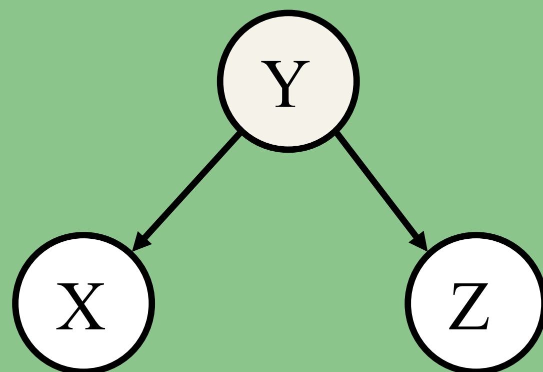
What Independencies does a Bayes Net Model?

Three cases of interest…

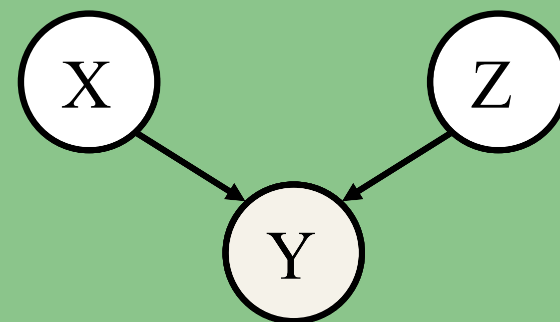
Cascade



Common Parent



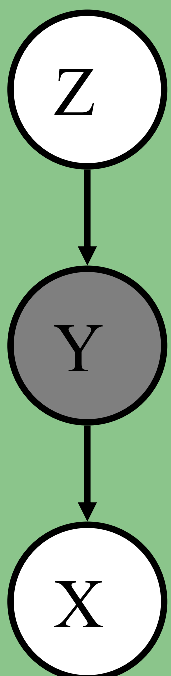
V-Structure



What Independencies does a Bayes Net Model?

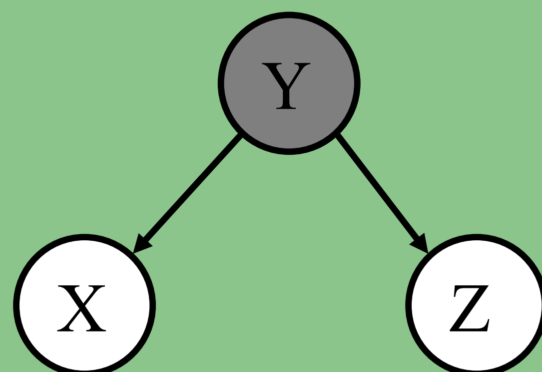
Three cases of interest...

Cascade



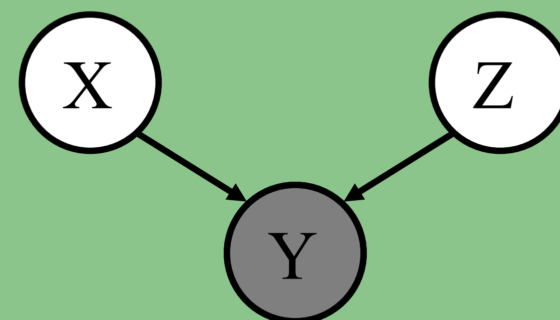
$$X \perp\!\!\!\perp Z \mid Y$$

Common Parent



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

V-Structure



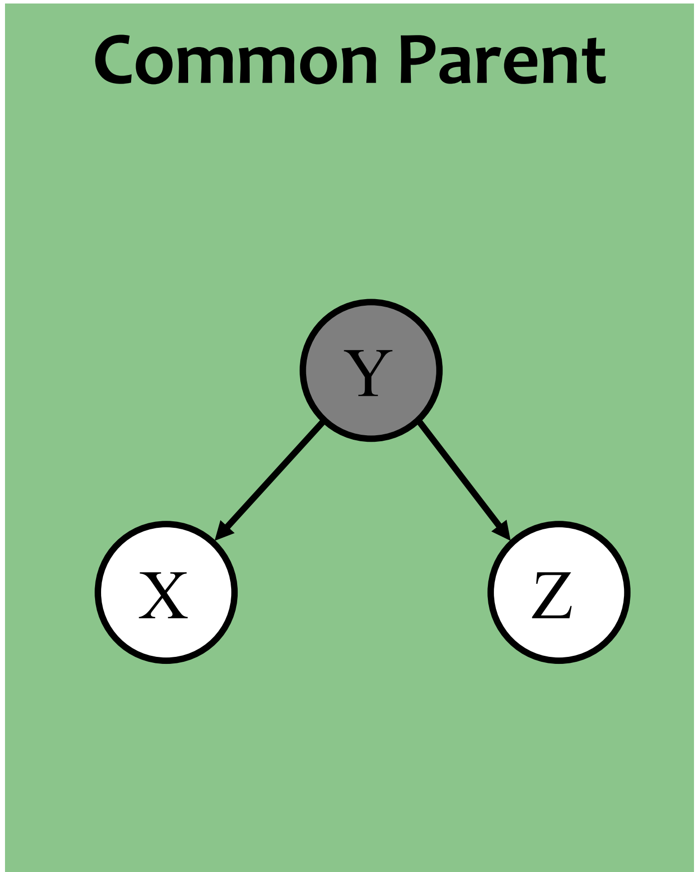
$$X \not\perp\!\!\!\perp Z \mid Y$$

Knowing Y **decouples** X and Z

Knowing Y **couples** X and Z

Whiteboard

Proof of conditional independence

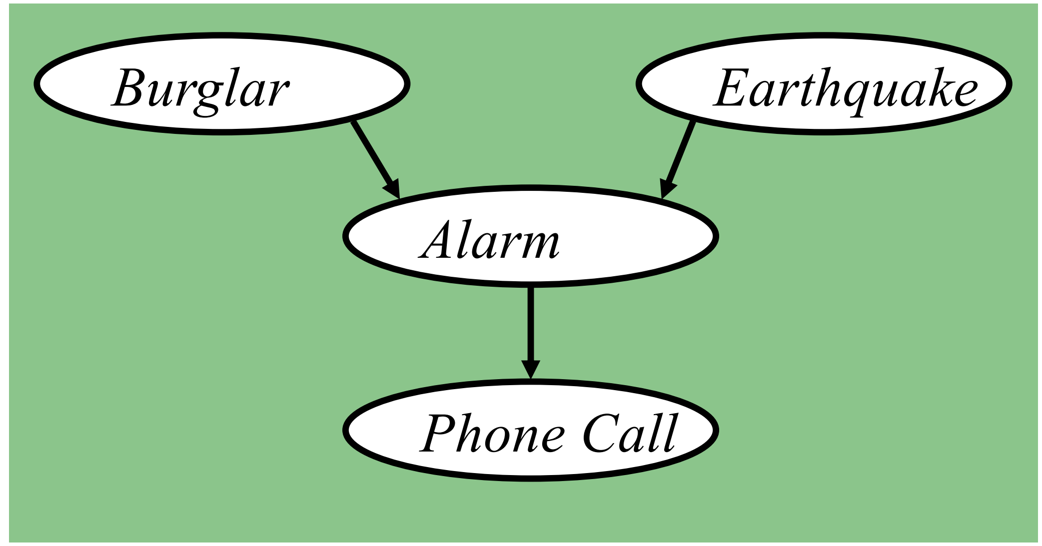


$$X \perp\!\!\!\perp Z \mid Y$$

(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?

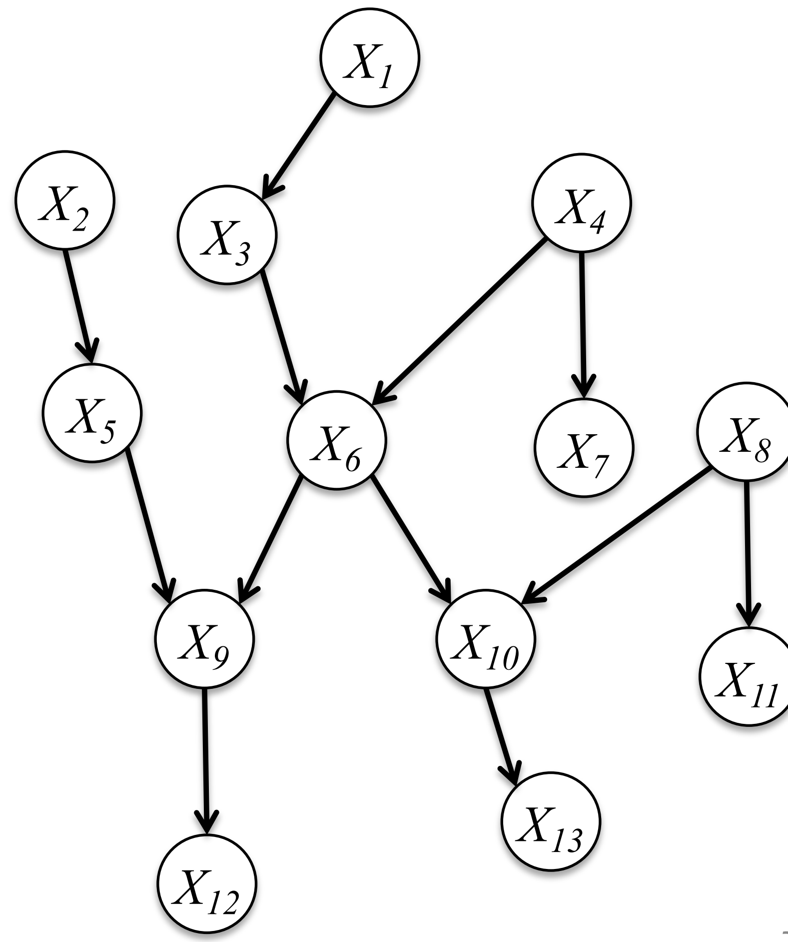
$$Burglar \perp\!\!\!\perp Earthquake \mid PhoneCall$$

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.

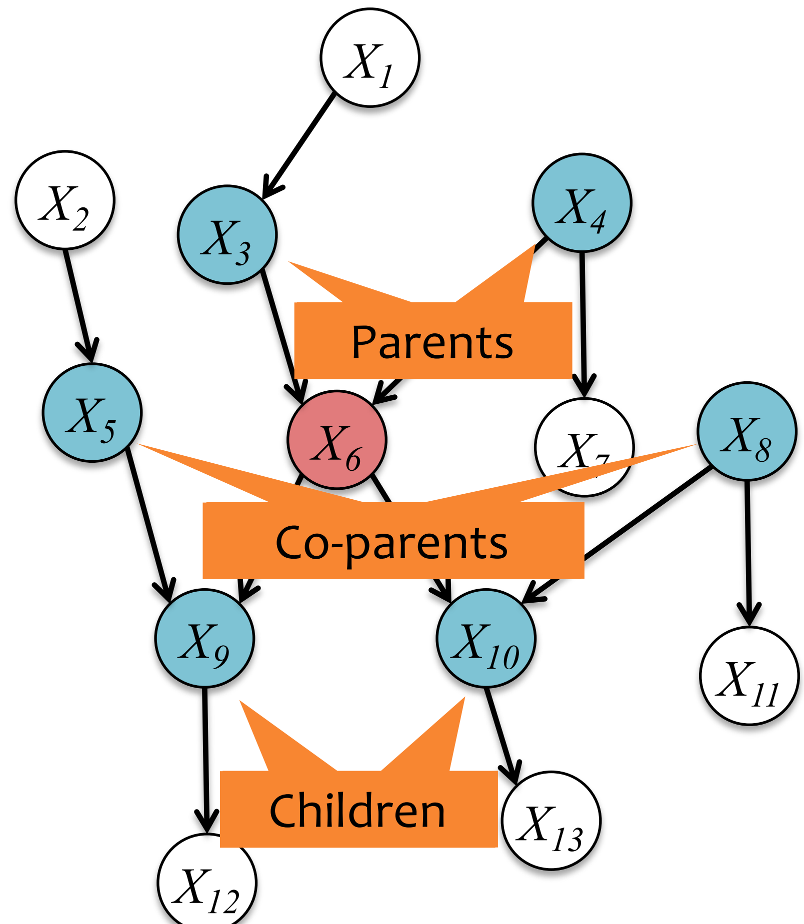


Markov Blanket (Directed)

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 $\{X_3, X_4, X_5, X_8, X_9, X_{10}\}$



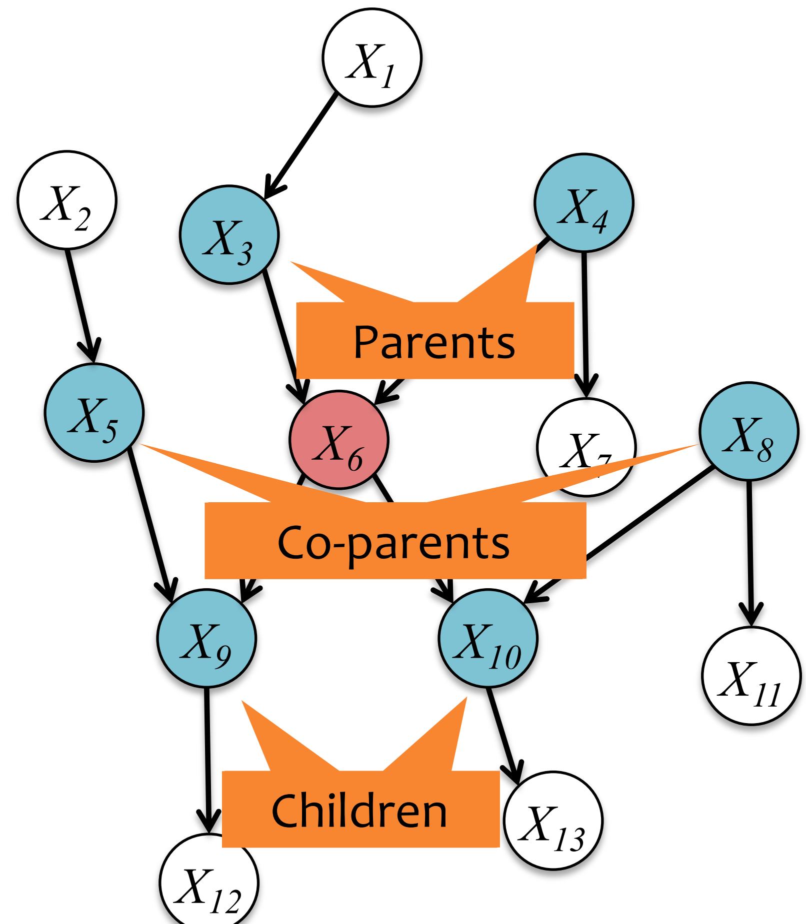
Markov Blanket (Directed)

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 $\{X_3, X_4, X_5, X_8, X_9, X_{10}\}$



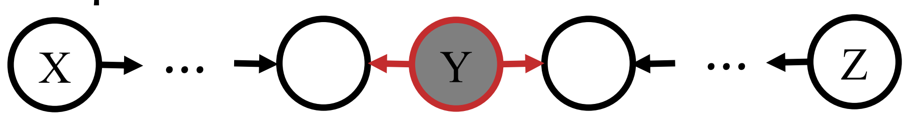
D-Separation

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

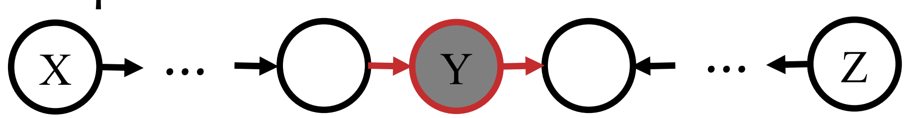
Definition #1:
Variables X and Z are **d-separated** given a **set** of evidence variables E iff every path from X to Z is “blocked”.

A path is “blocked” whenever:

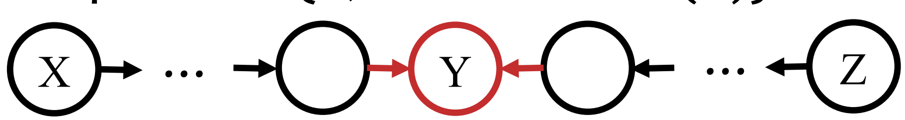
1. ∃Y on path s.t. Y ∈ E and Y is a “common parent”



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”



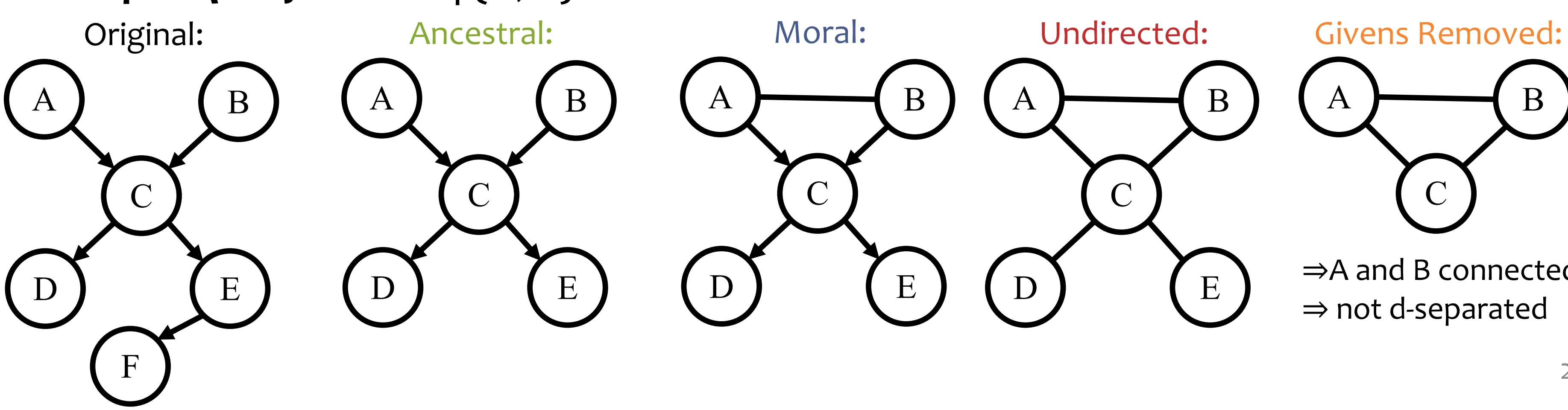
D-Separation

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



⇒A and B connected
⇒ not d-separated

Learning Objectives

Bayesian Networks

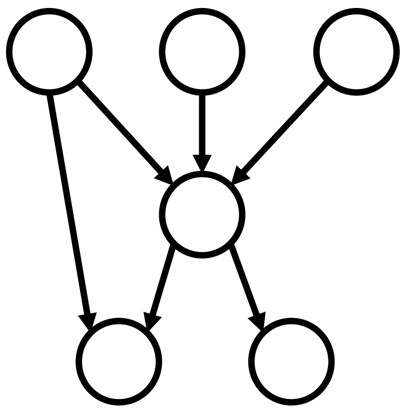
You should be able to…

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

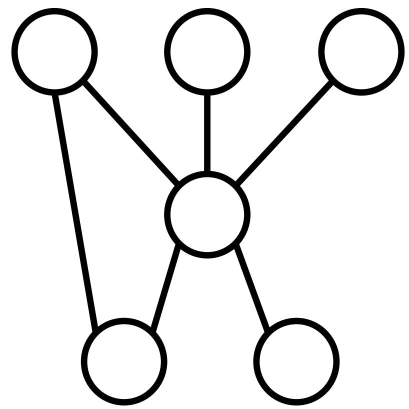
TYPES OF GRAPHICAL MODELS

Three Types of Graphical Models

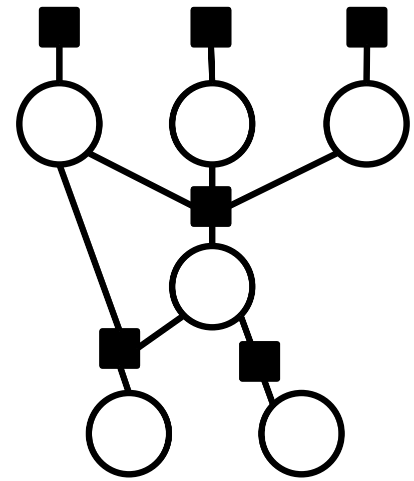
Directed Graphical Model



Undirected Graphical Model



Factor Graph



Key Concepts for Graphical Models

Graphical Models in General

1. A graphical model defines a **family of probability distributions**
2. That family shares in common a set of **conditional independence assumptions**
3. By choosing a **parameterization** of the graphical model, we obtain a single **model** from the family
4. The model may be either **locally or globally normalized**

Ex: Directed G.M.

1. Family:
2. Conditional Independencies:
3. Example parameterization:
4. Normalization:

Key Concepts for Graphical Models

Graphical Models in General

1. A graphical model defines a **family of probability distributions**
2. That family shares in common a set of **conditional independence assumptions**
3. By choosing a **parameterization** of the graphical model, we obtain a single **model** from the family
4. The model may be either **locally or globally normalized**

Ex: Undirected G.M.

1. Family:
2. Conditional Independencies:
3. Example parameterization:
4. Normalization:

Key Concepts for Graphical Models

Graphical Models in General

1. A graphical model defines a **family of probability distributions**
2. That family shares in common a set of **conditional independence assumptions**
3. By choosing a **parameterization** of the graphical model, we obtain a single **model** from the family
4. The model may be either **locally or globally normalized**

Ex: Factor Graph

1. Family:
2. Conditional Independencies:
3. Example parameterization:
4. Normalization:

Markov Random Fields

UNDIRECTED GRAPHICAL MODELS

Undirected Graphical Models

Whiteboard

- Conditional independence assumptions for undirected graphical model (graph separation)
- Definition: clique
- Definition: maximal clique
- Cliques and potential functions
- Non-negativity of potential functions
- Definition of model family (i.e. joint distribution)
- Global normalization and the partition function
- Example: Binary Variables for MRF

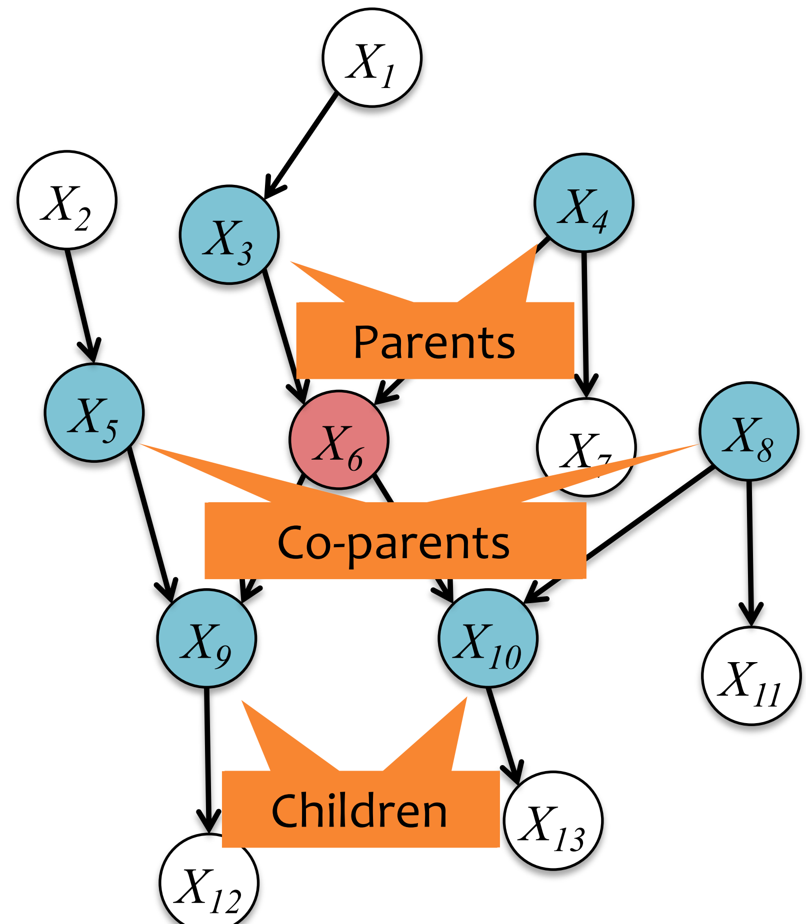
Markov Blanket (Directed)

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 $\{X_3, X_4, X_5, X_8, X_9, X_{10}\}$

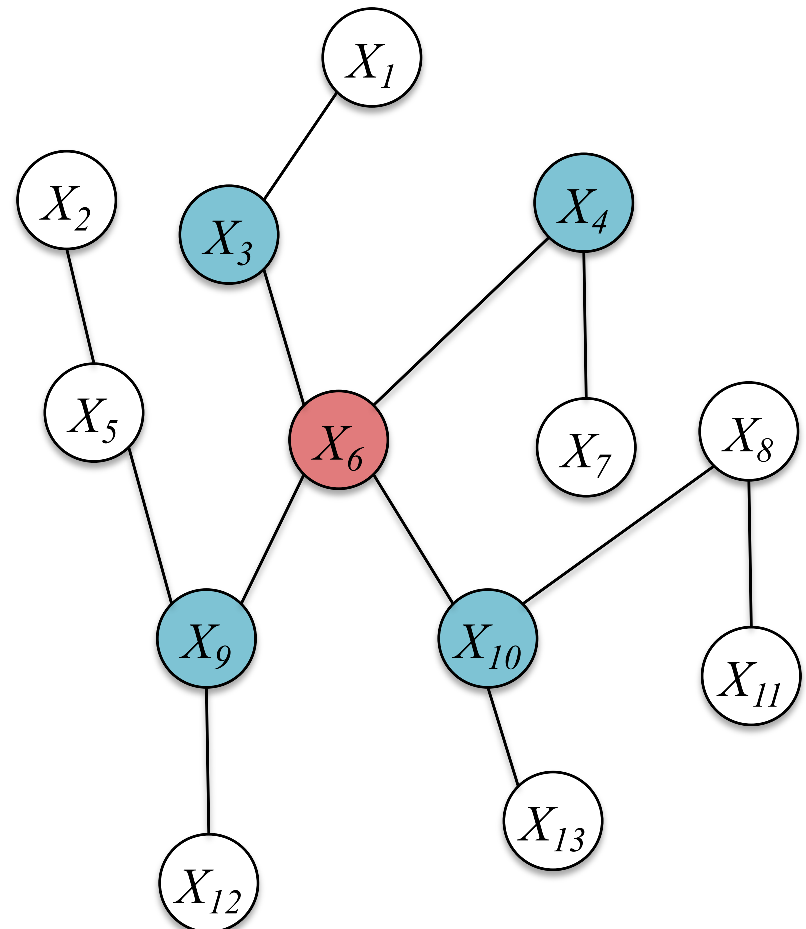


Markov Blanket (Undirected)

Def: the **Markov Blanket** of a node in an **undirected** graphical model is the set containing the node's neighbors.

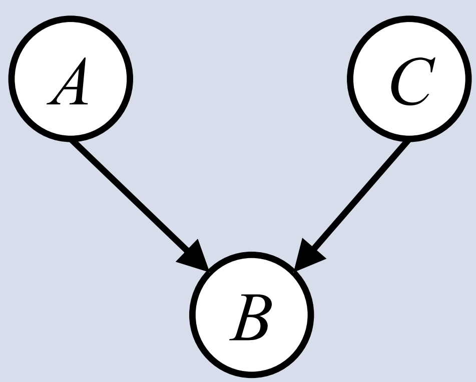
Theorem: a node is **conditionally independent** of every other node in the graph given its **Markov blanket**

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

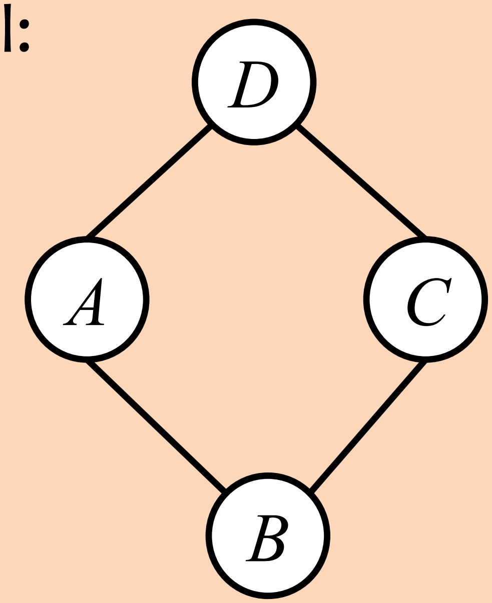


Non-equivalence of Directed / Undirected Graphical Models

There does **not** exist an **undirected** graphical model that can capture the conditional independence assumptions of this **directed** graphical model:



There does **not** exist a **directed** graphical model that can capture the conditional independence assumptions of this **undirected** graphical model:



Undirected Graphical Models

Whiteboard

- Parameterization (e.g. tabular vs. log-linear)
- Pairwise Markov Random Field (MRF)