Machine Learning Department
School of Computer Science
Carnegie Mellon University

Directed Graphical Models
+
Undirected Graphical Models

Matt Gormley
Lecture 6
Sep. 19, 2022

Reminders

- **Lecture 5.5: required video lecture**
- **Homework 2: Learning to Search for RNNs**
  - **Out: Sun, Sep 18**
  - **Due: Thu, Sep 29 at 11:59pm**
- **Poll Questions 0a and 0b about HW1**

Representation of both directed and undirected graphical models

INTUITION FOR FACTOR GRAPHS

Joint Modeling

After we come up with a way to decompose our structure into variables, what comes next?

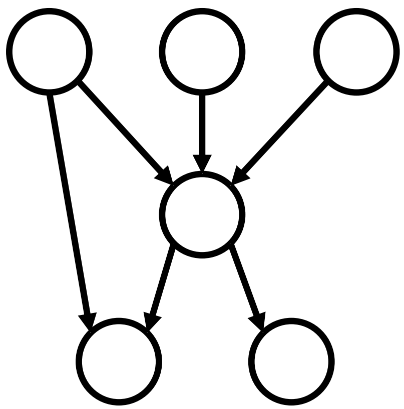
- We can define a **joint model** over those variables
- The joint model defines a **score for each possible structure** allowed by our decomposition
- The model should give high scores to "good" structures and low scores to "bad" structures
  - in probability terms: **high scores for likely structures** and **low scores for unlikely structures**
  - "likely structures" could be defined as those appearing in your **training dataset**
- (Hopefully, the joint model is also able to capture interesting interactions between pairs, triples, quadruples, … of variables)

How do we write down a joint model?

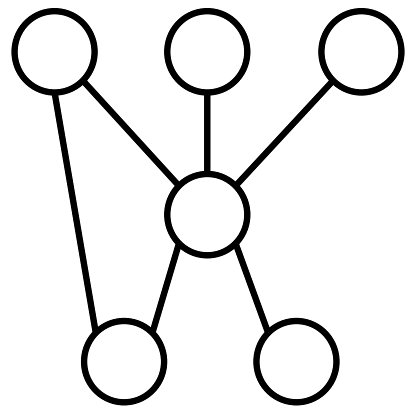
(Factor Graphs)

Three Types of Graphical Models

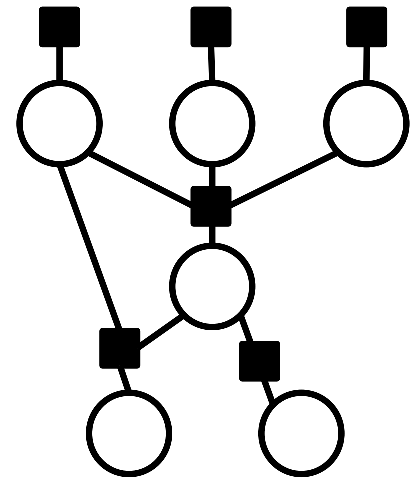
Directed Graphical Model

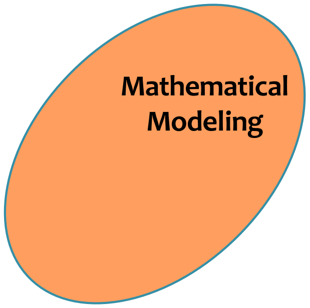


Undirected Graphical Model



Factor Graph



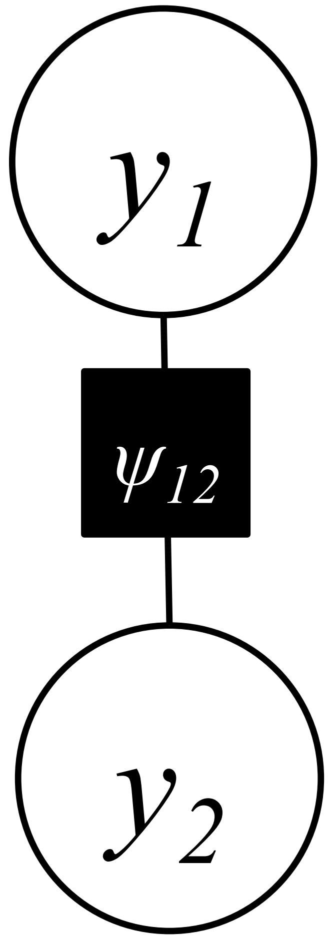


Factor Graphs

Factor Graph

(bipartite graph)

- variables (circles)
- factors (squares)



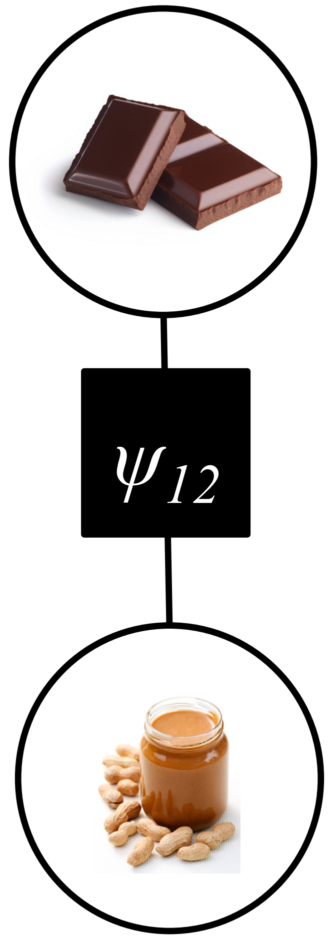
Mathematical Modeling

Factor Graphs

Factor Graph

(bipartite graph)

- variables (circles)
- factors (squares)



Each **random variable** can be assigned a **value**

The collection of values for all the random variables is called an **assignment.**

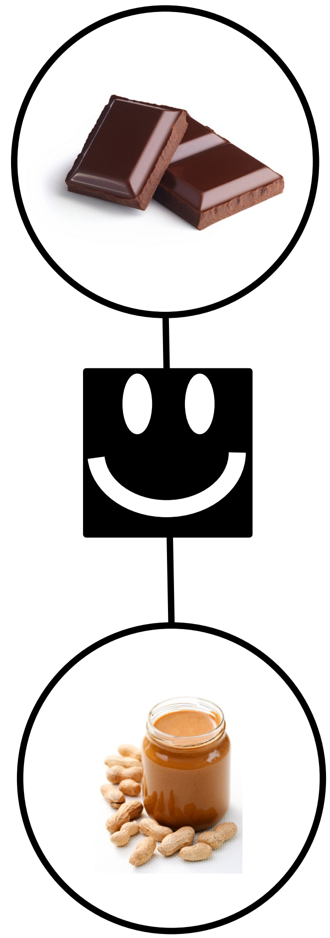
Mathematical Modeling

Factor Graphs

Factor Graph

(bipartite graph)

- variables (circles)
- factors (squares)



Factors have local opinions about the assignments of their neighboring variables

Mathematical Modeling

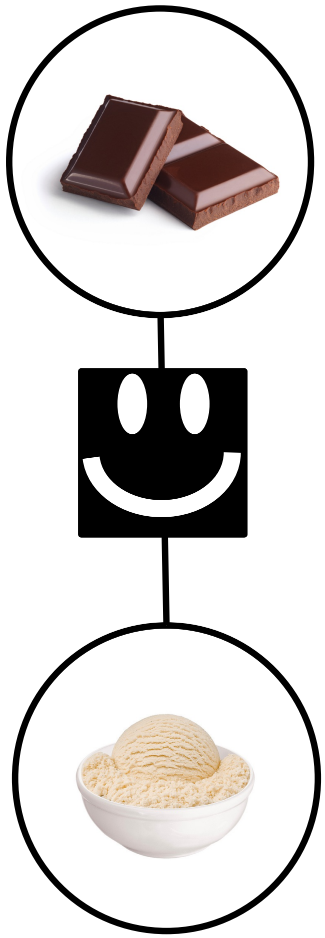
Factor Graphs

Factor Graph

(bipartite graph)

- variables (circles)
- factors (squares)

Factors have local opinions about the assignments of their neighboring variables



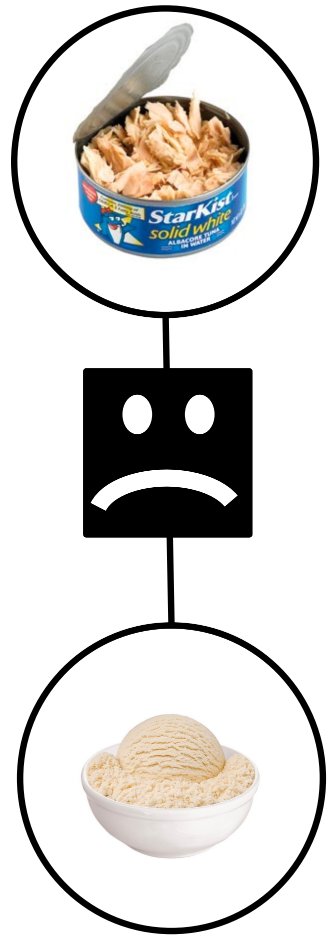
Factor Graphs

Factor Graph

(bipartite graph)

- variables (circles)
- factors (squares)

Factors have local opinions about the assignments of their neighboring variables



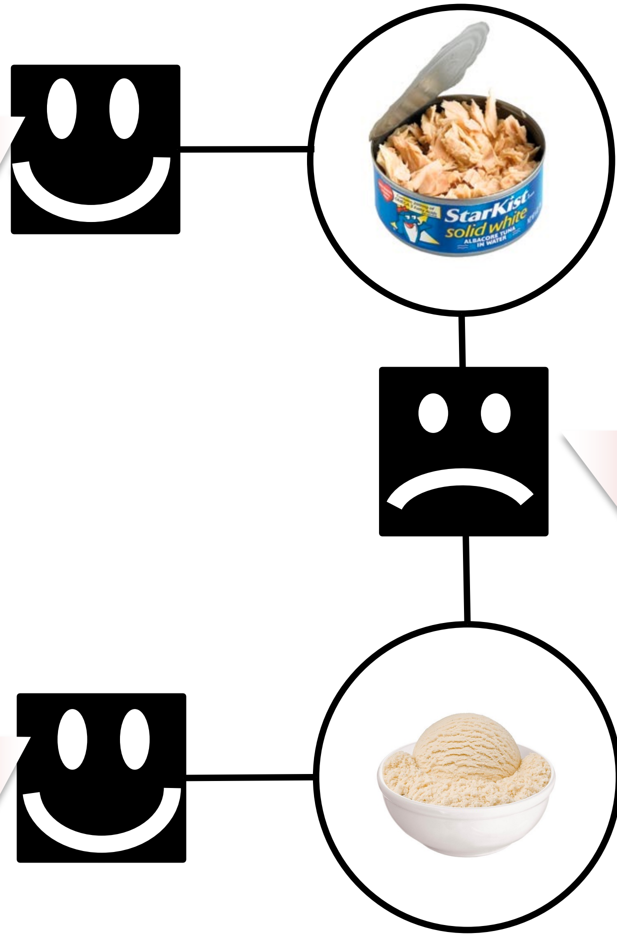
Mathematical Modeling

Factor Graphs

$P(\text{tuna, ice cream}) \; = \; ?$

Those opinions are expressed through potential tables

chocolate	0.1
peanut butter	5
ice cream	1
tuna	6
…	



	chocolate	peanut butter	ice cream	tuna	⋮
chocolate	2	9	7	0.1	
peanut butter	4	2	3	0.2	
ice cream	7	3	2	0.1	
tuna	0.1	0.2	0.1	2	
…					

chocolate	4
peanut butter	8
ice cream	7
tuna	3
…	

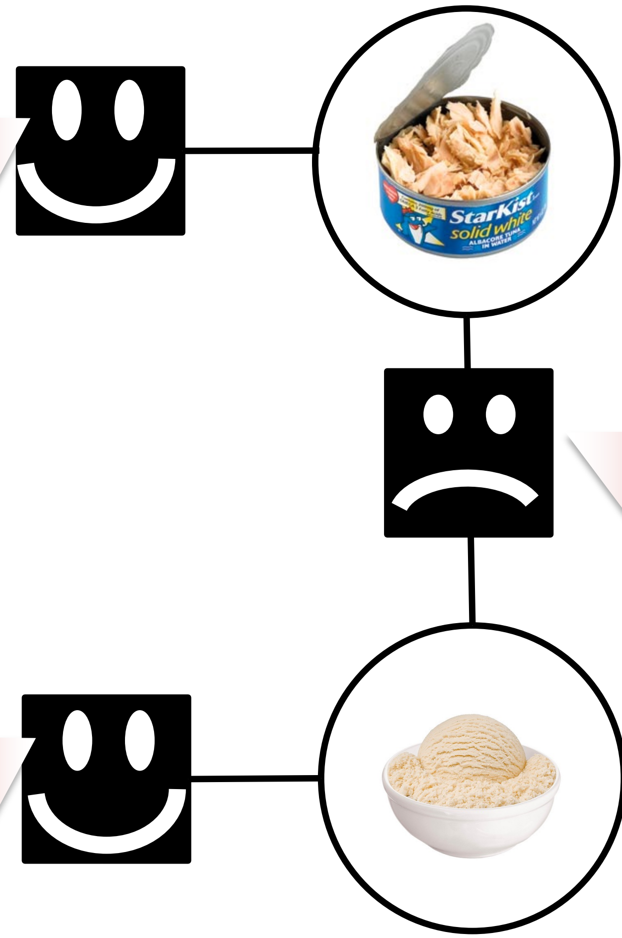
Mathematical Modeling

Factor Graphs

$$P(\text{tuna, ice cream}) = \frac{1}{Z}(6 * 7 * 0.1)$$

chocolate	0.1
peanut butter	5
ice cream	1
tuna	6
…	

chocolate	4
peanut butter	8
ice cream	7
tuna	3
…	



Uh-oh! The probabilities of the various assignments sum up to Z > 1.

So divide them all by Z.

	cho…	peanu…	ice c…	tuna	⋮
chocolate	2	9	7	0.1	
peanut butter	4	2	3	0.2	
ice cream	7	3	2	0.1	
tuna	0.1	0.2	0.1	2	
…					

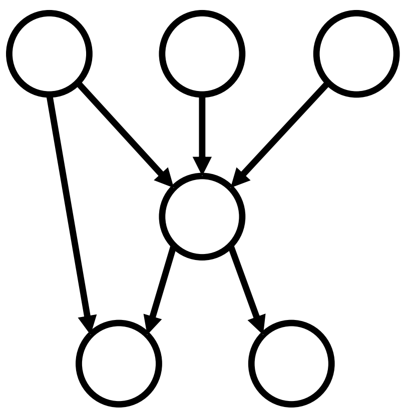
The combined potential tables of all factors defines the probability of an assignment

Bayesian Networks

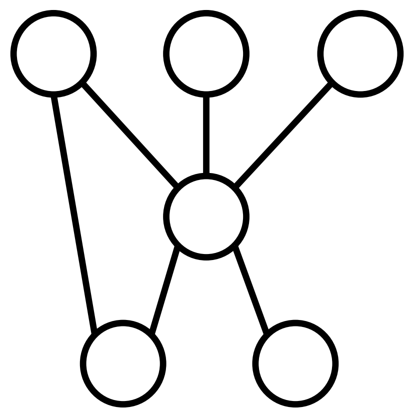
DIRECTED GRAPHICAL MODELS

Three Types of Graphical Models

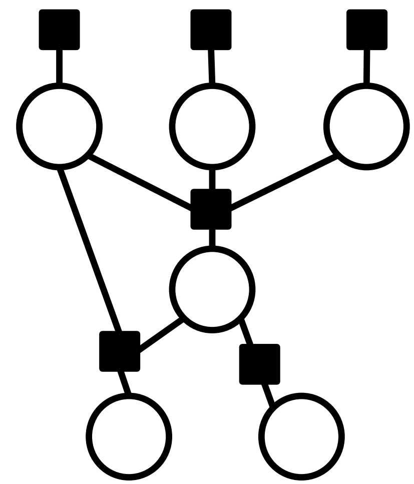
Directed Graphical Model



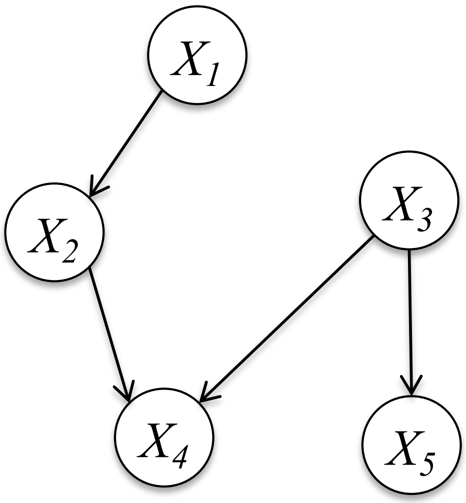
Undirected Graphical Model



Factor Graph



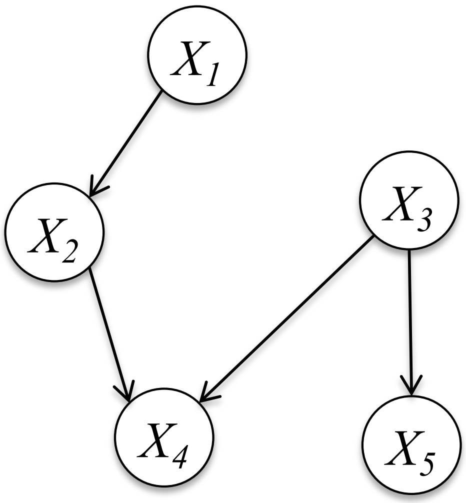
Bayesian Network



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

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)
  - …

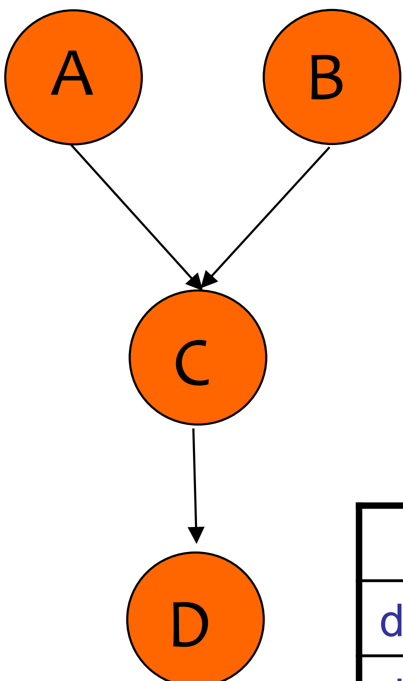
Quantitative Specification

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

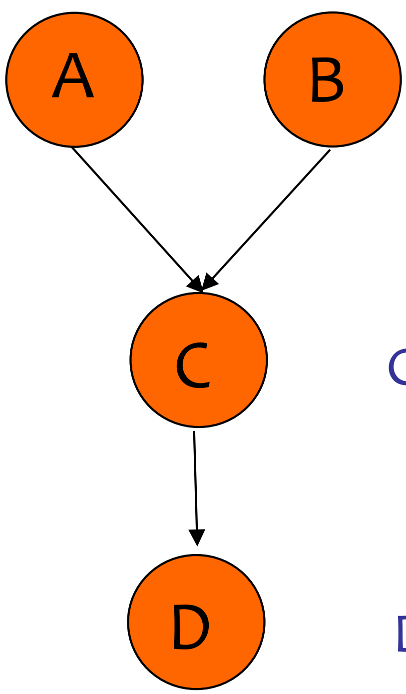
	c^0	c^1
d^0	0.3	0.5
d^1	07	0.5

Quantitative Specification

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

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

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



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

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

P(D| C)

D

C

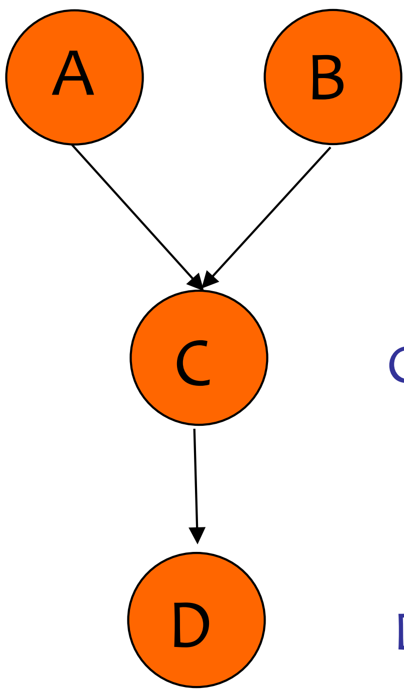
Quantitative Specification

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

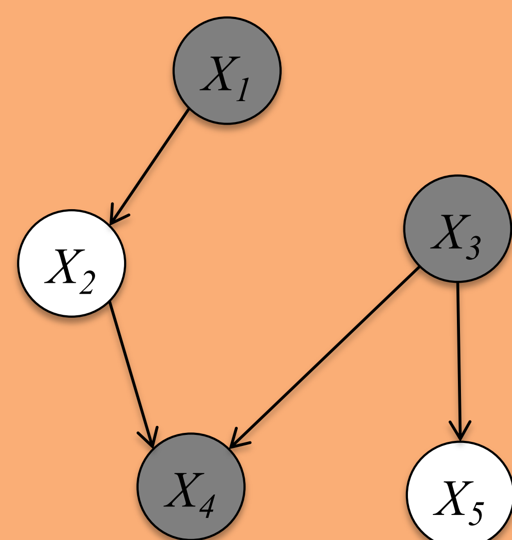
Polynomial in T

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 BayesNets

Question: Describe in words the directed graphical model that you would draw to represent an RNN-LM.

Answer:

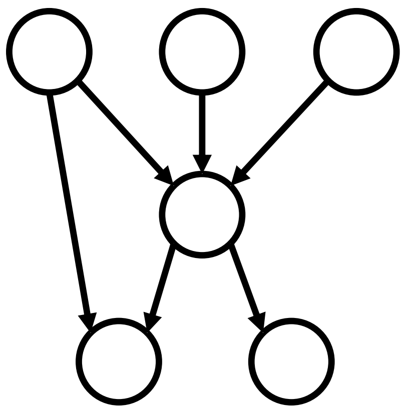
Question: Describe in words the directed graphical model that you would draw to represent a seq2seq model.

Answer:

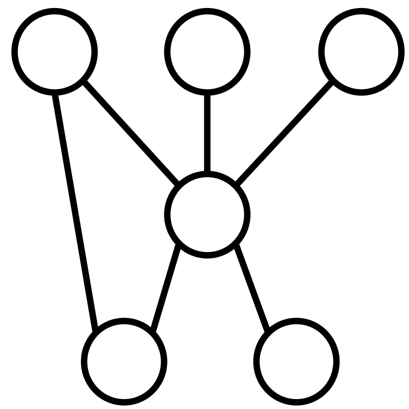
UNDIRECTED GRAPHICAL MODELS

Three Types of Graphical Models

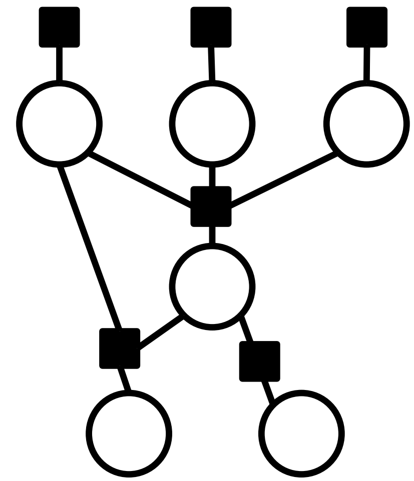
Directed Graphical Model



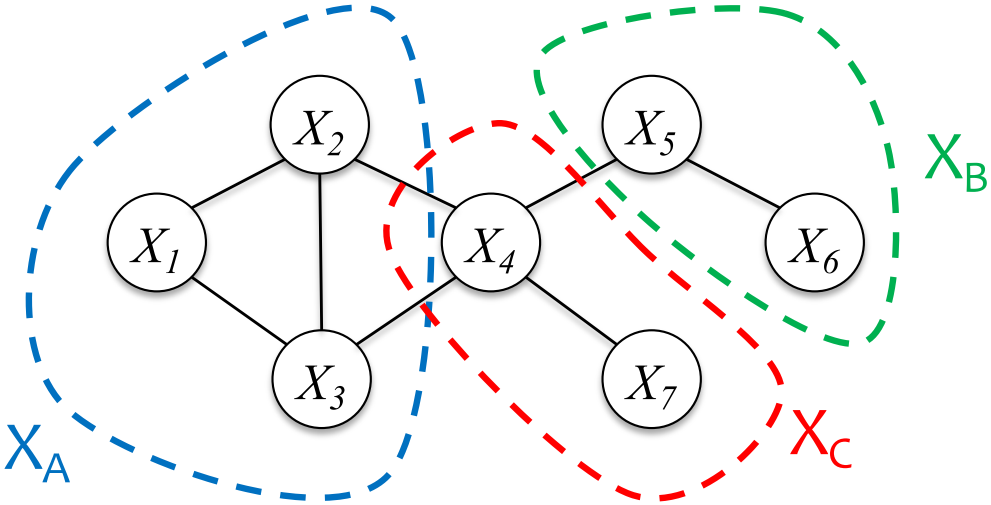
Undirected Graphical Model



Factor Graph



Undirected Graphical Models

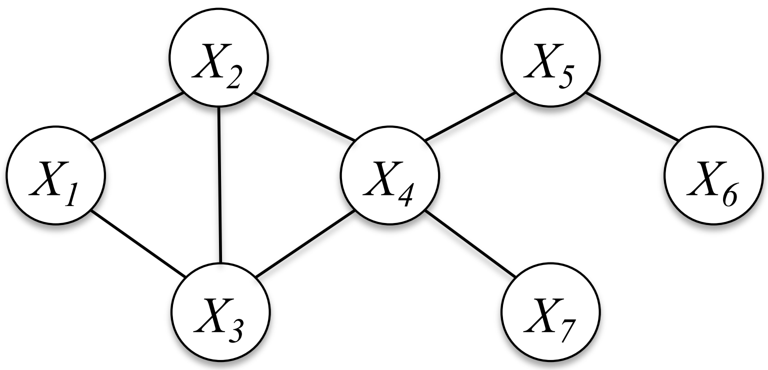


Notation: Let X_S denote all the variables with indices in the set $S \subset \mathbb{Z}^+$

Undirected Graph Terminology

- *Definition*: a **clique** is a set of fully connected nodes (e.g. $\{X_1, X_2\}$ or $\{X_1, X_2, X_3\}$)
- *Definition*: a **maximal clique** is a clique to which adding any node makes it no longer a clique (e.g. $\{X_1, X_2, X_3\}$ but not $\{X_1, X_2\}$)
- *Definition*: a set of nodes X_C **separates** sets X_A and X_B if removing X_C leaves no path from a node in X_A to one in X_B. (e.g. $\{X_4, X_7\}$ separates $\{X_1, X_2, X_3\}$ and $\{X_5, X_6\}$)

Undirected Graphical Models

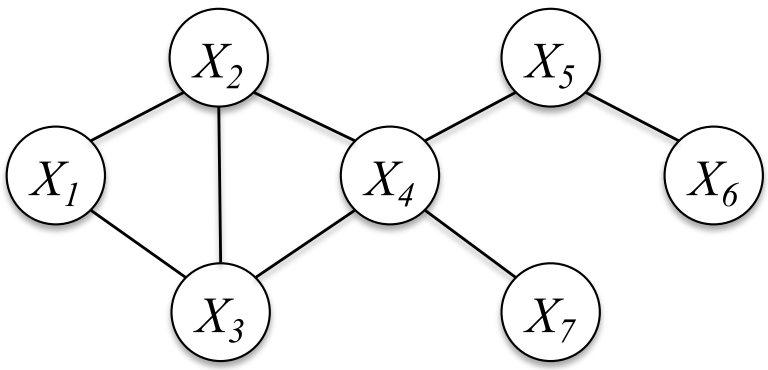


also called a Markov Random Field (MRF)

Def: an **undirected graphical model (UGM)** consists of a graph G (qualitative specification) and potential functions ψ (quantitative specification)

- The graph G is an undirected graph over random variables $X_1, \ldots, X_T$ and cycles are permitted
- The potential functions ψ, also called "factors", are used to define the joint probability

Undirected Graphical Models



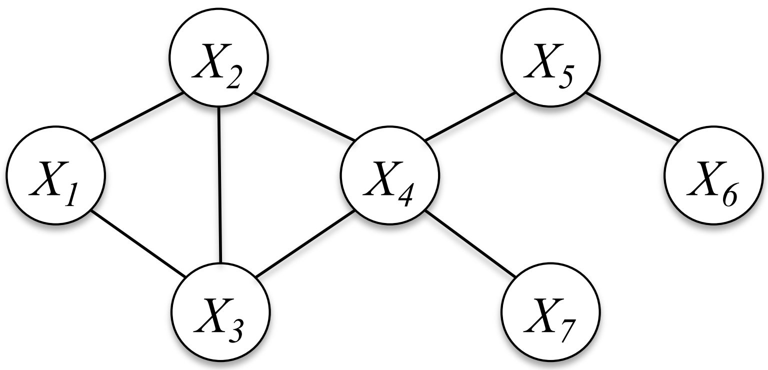
1. we have one potential function (aka. factor) per clique
2. potential functions must be non-negative
   $$\psi_C(x_C) \geq 0, \forall C, x_C$$
3. Z is the partition function → globally normalized model
   $$Z = \sum_{\mathbf{x} \in \mathcal{X}} \prod_{C \in \mathcal{C}} \psi_C(X_c)$$
   $$= \sum_{\mathbf{x} \in \mathcal{X}} s(\mathbf{x})$$

<u>*Def*</u>: Joint probability of a UGM

$$p(x_1, \ldots, x_T) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C)$$

where $\mathcal{C}$ is the set of all cliques and
$C \in \mathcal{C}$ is an index set $\Rightarrow C \subseteq \{1, \ldots, T\}$

Undirected Graphical Models



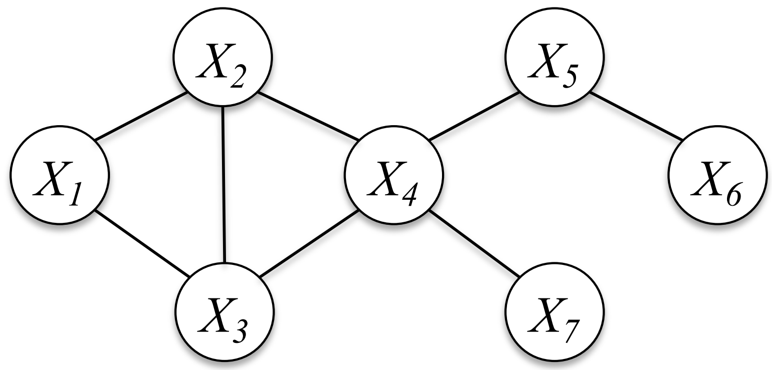
<u>Def</u>: A distribution is said to **factor according to the graph** G if it can be written as

$$p(x_1, \ldots, x_T) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C)$$

where $\mathcal{C}$ is the set of all cliques and
$C \in \mathcal{C}$ is an index set $\Rightarrow C \subseteq \{1, \ldots, T\}$

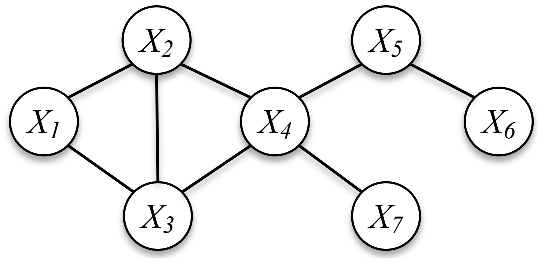
Undirected Graphical Models

Ex: Joint probability of UGM



$$p(\mathbf{x}) = \frac{1}{Z}\Big(\psi_{1,2,3}(x_1, x_2, x_3)\, \psi_{2,3,4}(x_2, x_3, x_4)$$
$$\psi_{4,5}(x_4, x_5)\, \psi_{5,6}(x_5, x_6)\, \psi_{4,7}(x_4, x_7)\Big)$$

Potential Functions for UGM



$$p(\mathbf{x}) = \frac{1}{Z}\Big(\psi_{1,2,3}(x_1, x_2, x_3)\, \psi_{2,3,4}(x_2, x_3, x_4)$$
$$\psi_{4,5}(x_4, x_5)\, \psi_{5,6}(x_5, x_6)\, \psi_{4,7}(x_4, x_7)\Big)$$

How should we **interpret** the potential functions in a UGM?

- *Idea #1*: Maybe as **marginals** of the distribution? In general, no.

$$p(\mathbf{x}) \neq \frac{1}{Z}\big(p(x_1, x_2, x_3)p(x_2, x_3, x_4)$$
$$p(x_4, x_5)p(x_5, x_6)p(x_4, x_7)\big)$$

- *Idea #2*: Maybe as **conditionals** of the distribution? In general, no.

$$p(\mathbf{x}) \neq \frac{1}{Z}\big(p(x_1|x_2, x_3)p(x_2, x_3|x_4)$$
$$p(x_4|x_5)p(x_5|x_6)p(x_7|x_4)\big)$$

Potential Functions for UGM

Whiteboard

– Simple example of potential functions as tables

Undirected Graphical Models

Whiteboard

– Alternate definition using maximal cliques
– Pairwise Markov Random Field (MRF)

Compactness of a UGM

Consider random variables $X_1, X_2, \ldots, X_T$ 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 $\max_{C \in \mathcal{C}} |C| \leq D$

  then each $\psi_C(X_C)$ needs only R^D values for a total of only $O(T(R^D))$ values

Polynomial in T

CONDITIONAL INDEPENDENCIES OF DGMS

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

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

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

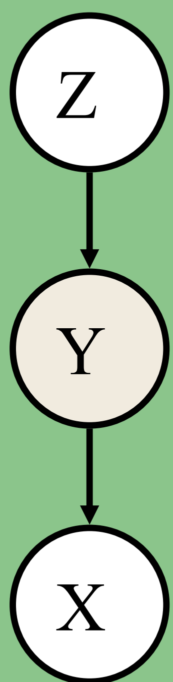
- But what else does it imply?

Slide from William Cohen

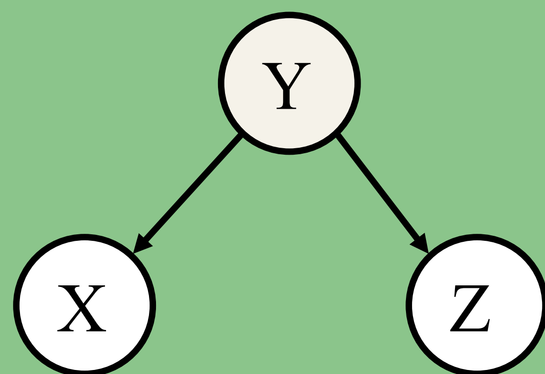
What Independencies does a Bayes Net Model?

Three cases of interest…

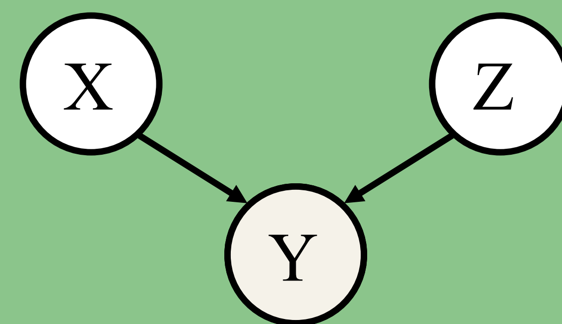
Cascade



Common Parent



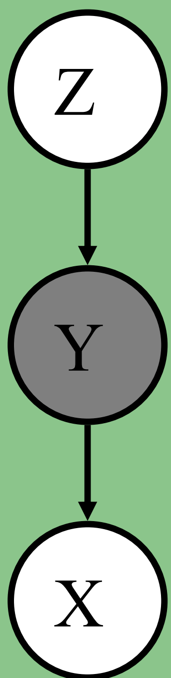
V-Structure



What Independencies does a Bayes Net Model?

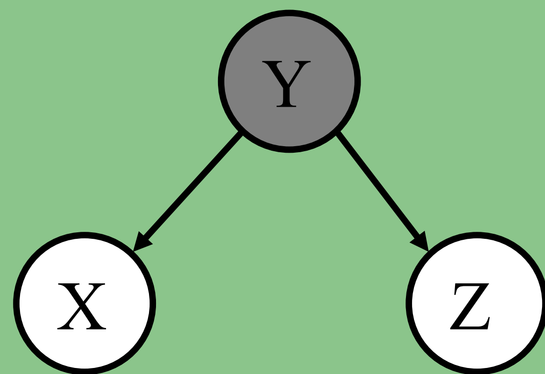
Three cases of interest…

Cascade



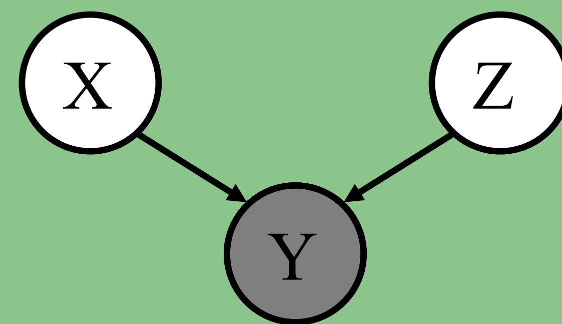
$X \perp\!\!\!\perp Z \mid Y$

Common Parent



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

V-Structure



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

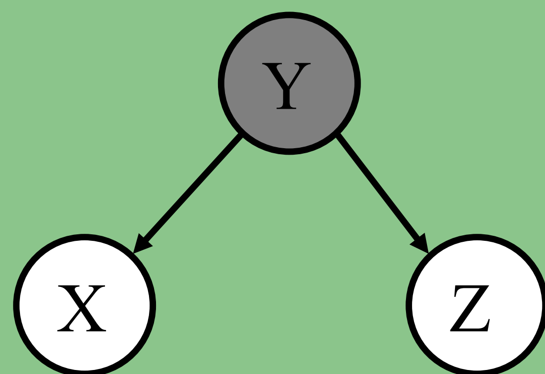
Knowing Y
decouples X and Z

Knowing Y
couples X and Z

Whiteboard

Proof of conditional independence

Common Parent

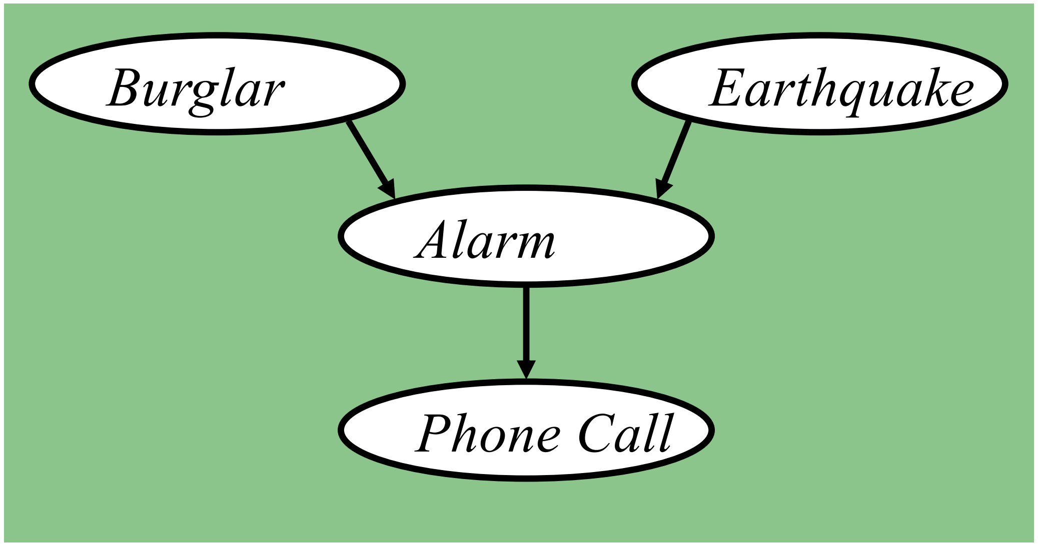


$$X \perp\!\!\!\perp Z \mid Y$$

(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?

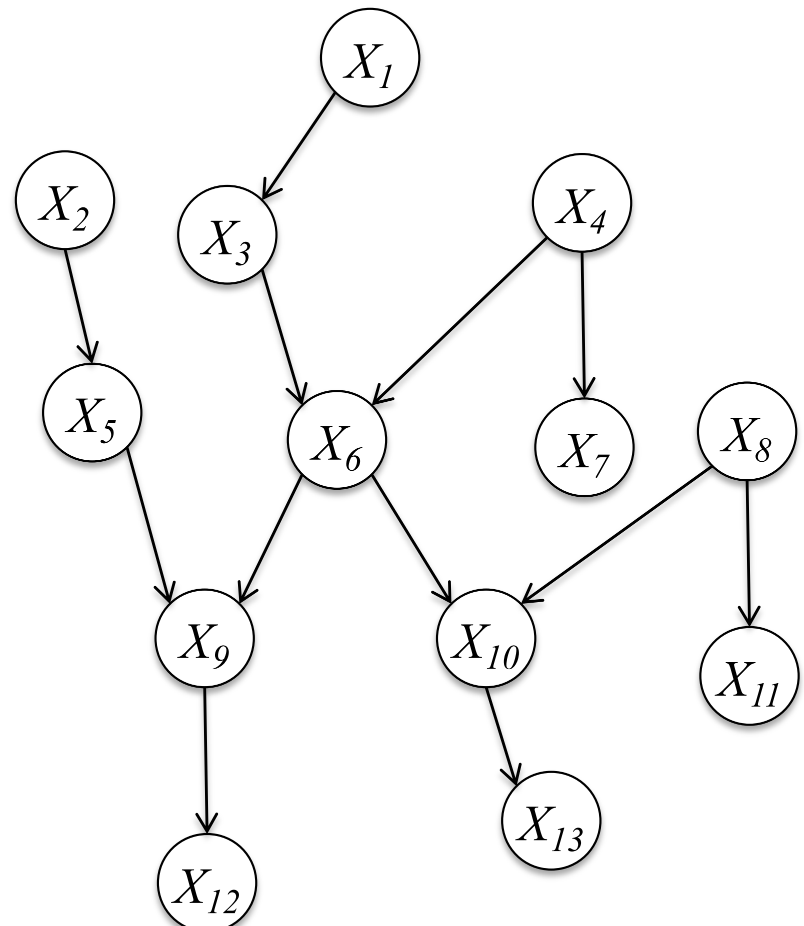
$$Burglar \perp\!\!\!\perp Earthquake \mid PhoneCall$$

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.

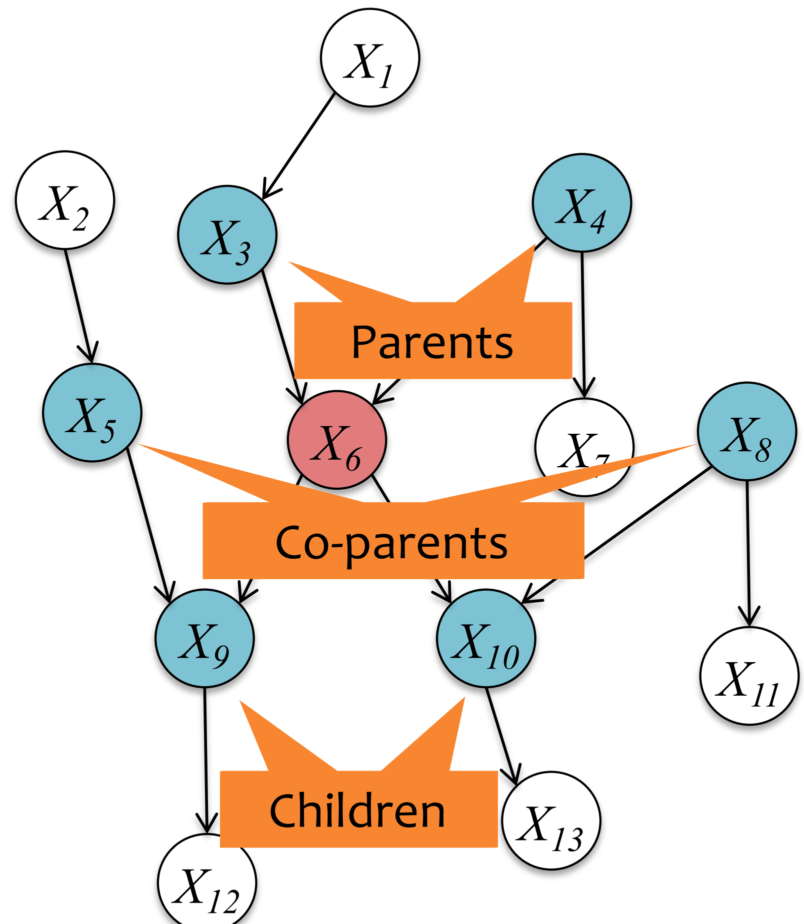


Markov Blanket (Directed)

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



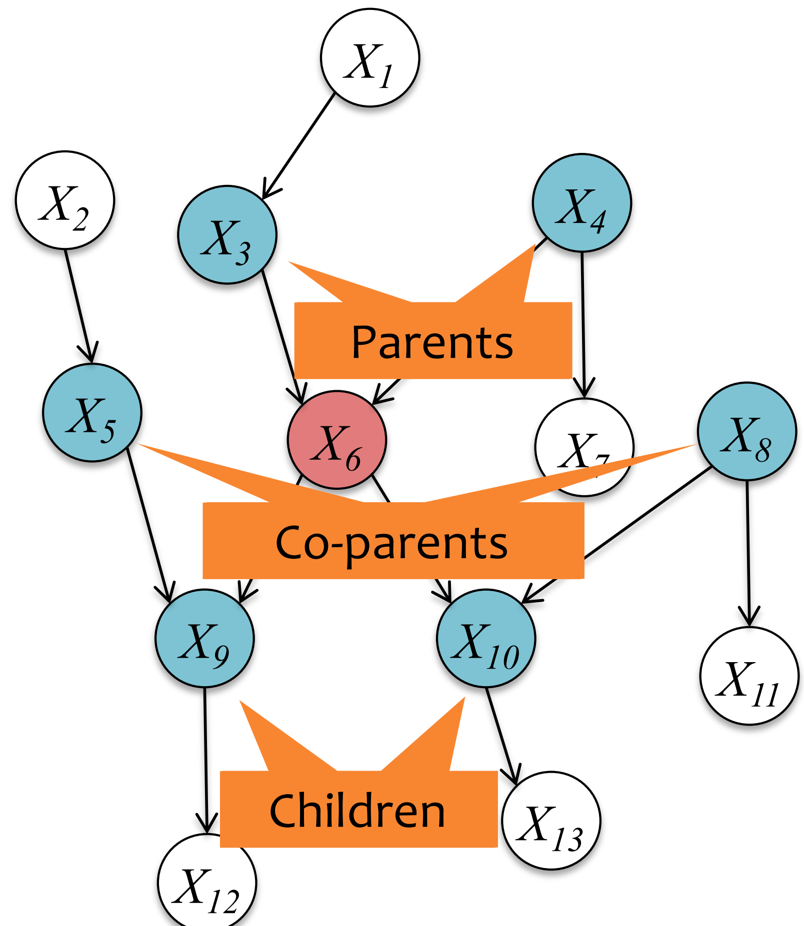
Markov Blanket (Directed)

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



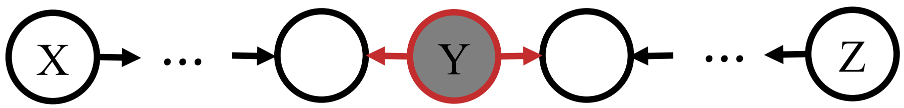
D-Separation

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

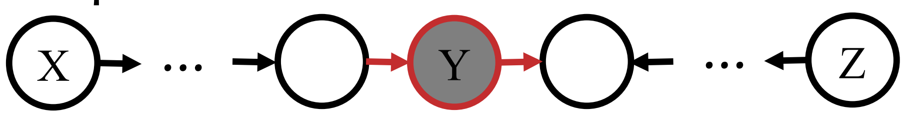
<u>Definition #1:</u>
Variables X and Z are **d-separated** given a **set** of evidence variables E iff every path from X to Z is "blocked".

A path is "blocked" whenever:

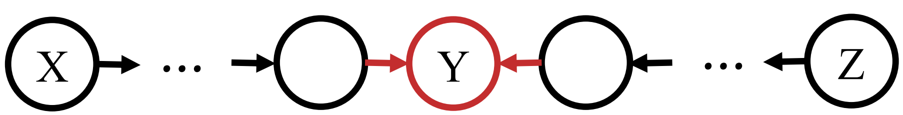
1. ∃Y on path s.t. Y ∈ E and Y is a "common parent"



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"



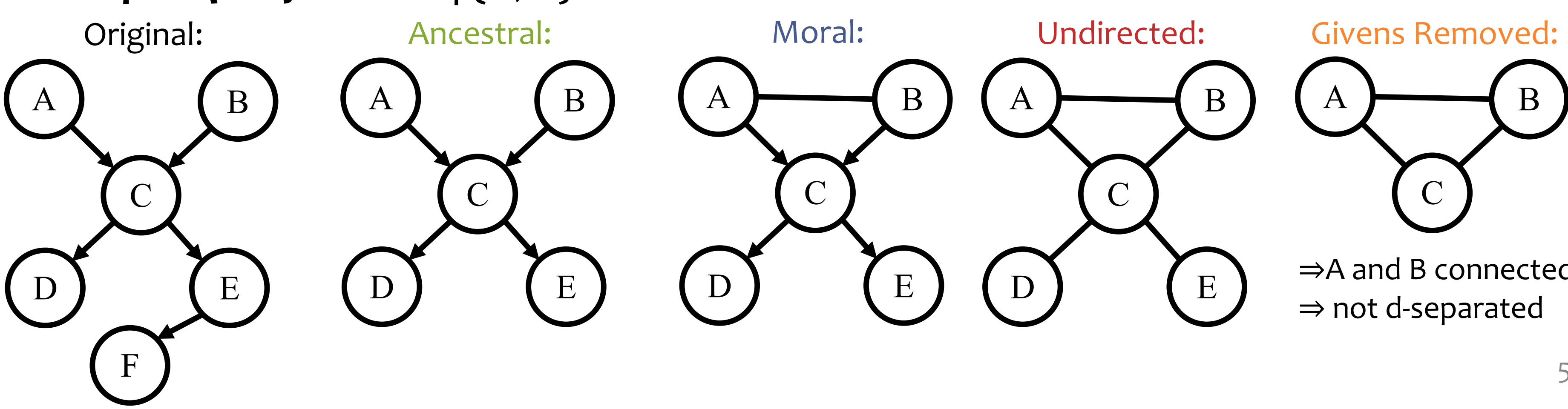
D-Separation

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



CONDITIONAL INDEPENDENCIES OF UGMS

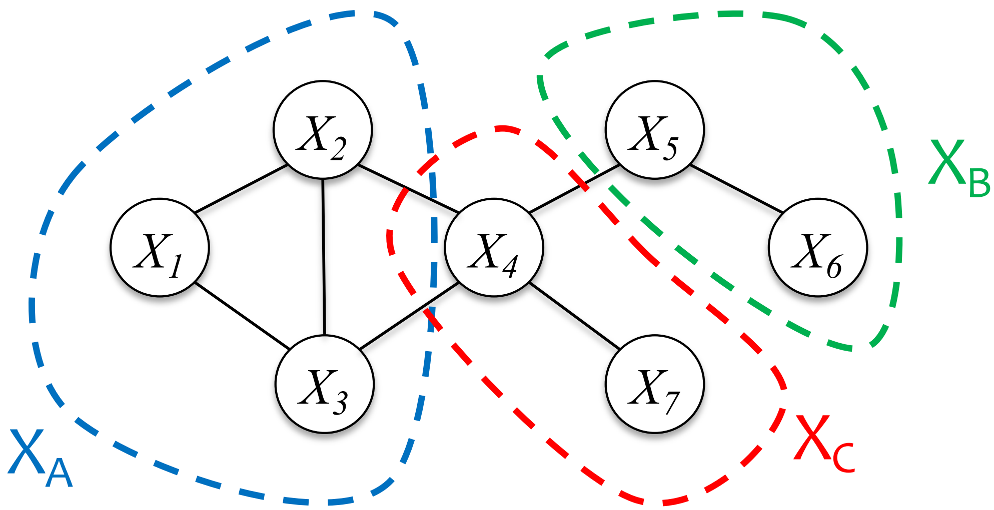
Undirected Graphical Models

Conditional Independence Semantics

Consider a distribution over r.v.s $X_1, \dots, X_T$

For a UGM and any disjoint index sets A, B, C,
(i.e., $A \subseteq \{1, \dots, T\}$, $B \subseteq \{1, \dots, T\}$, $C \subseteq \{1, \dots, T\}$)

X_A is **conditionally independent** of X_B given X_C iff X_C separates sets X_A and X_B



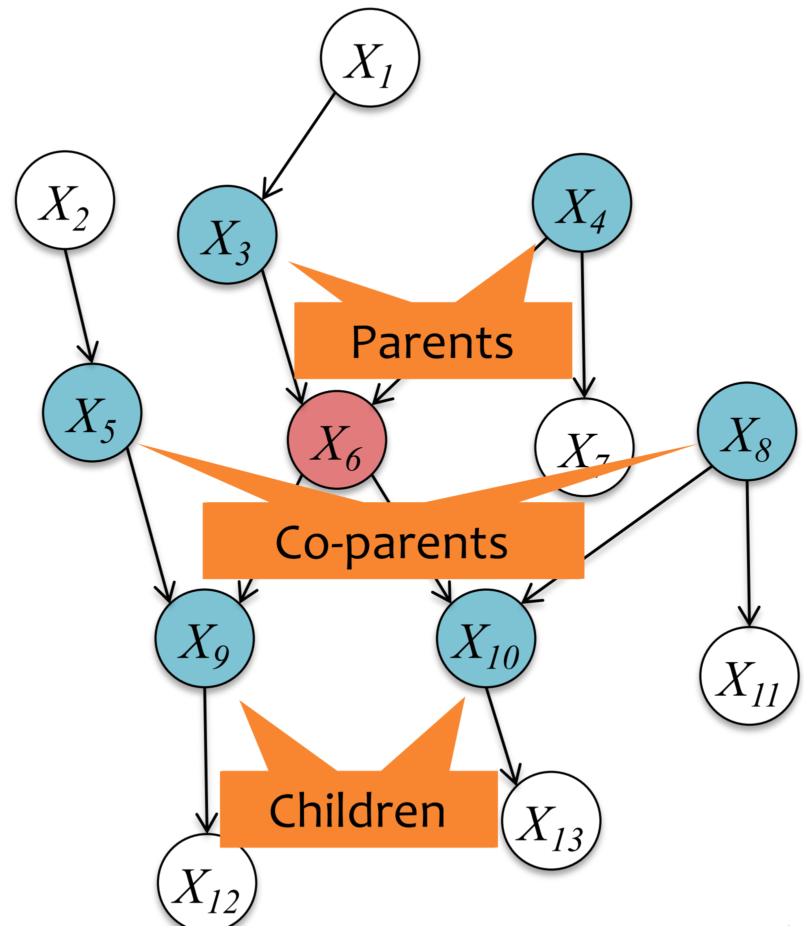
Markov Blanket (Directed)

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

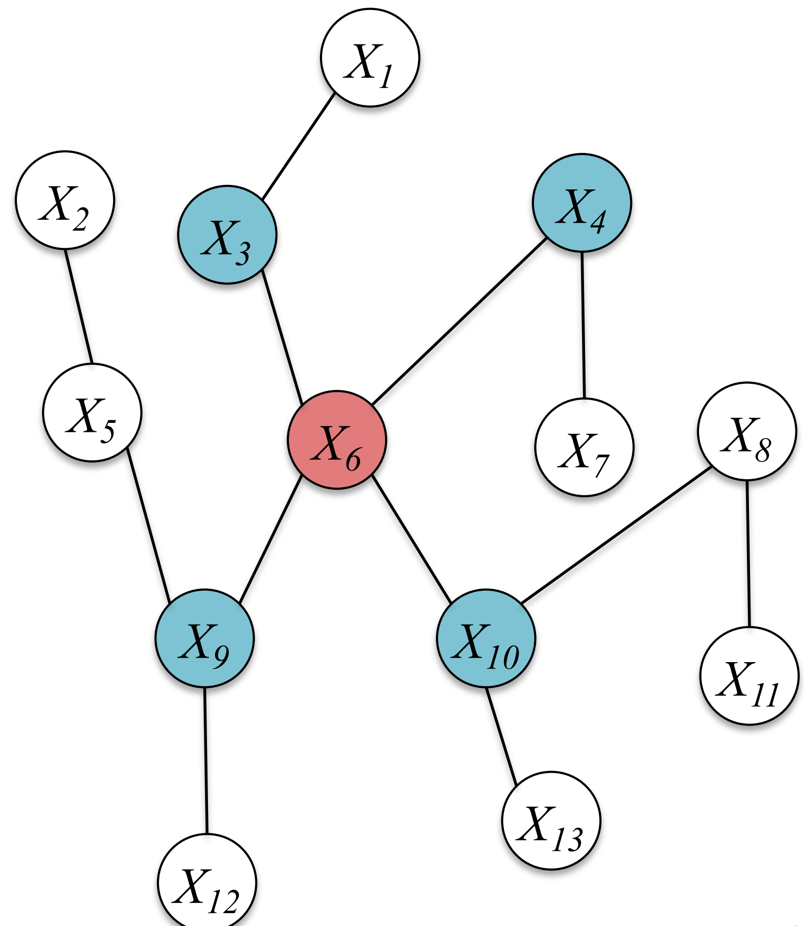


Markov Blanket (Undirected)

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



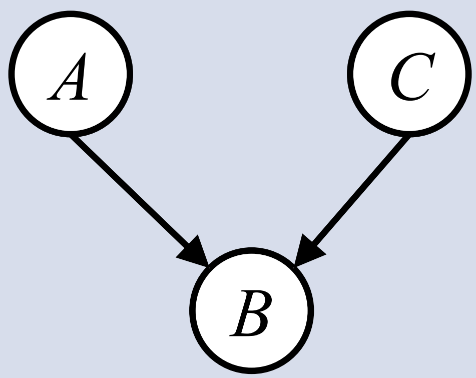
Undirected Graphical Models

Whiteboard

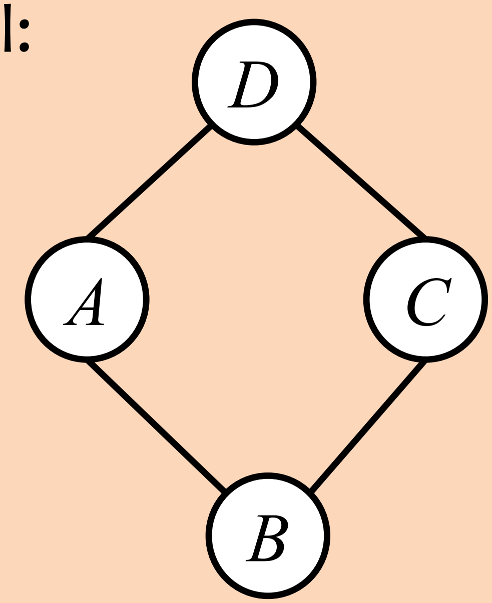
- Proof of independence by separation (simple case)

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:

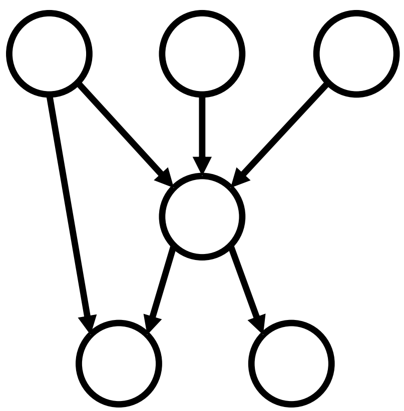


Representation of both directed and undirected graphical models

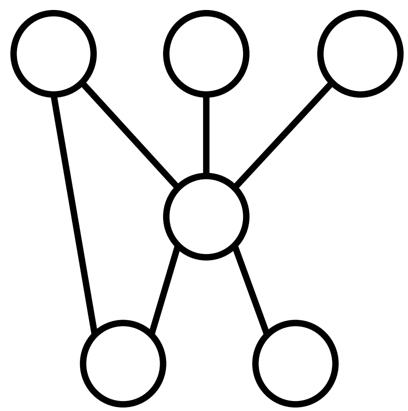
FACTOR GRAPHS

Three Types of Graphical Models

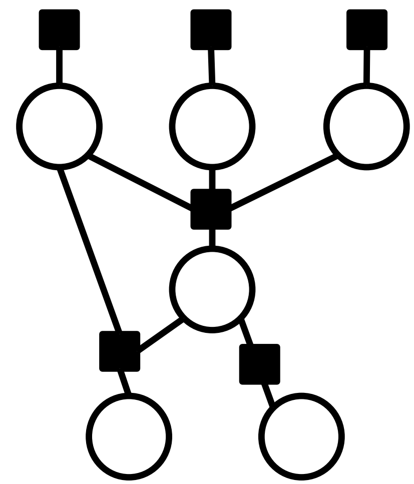
Directed Graphical Model



Undirected Graphical Model

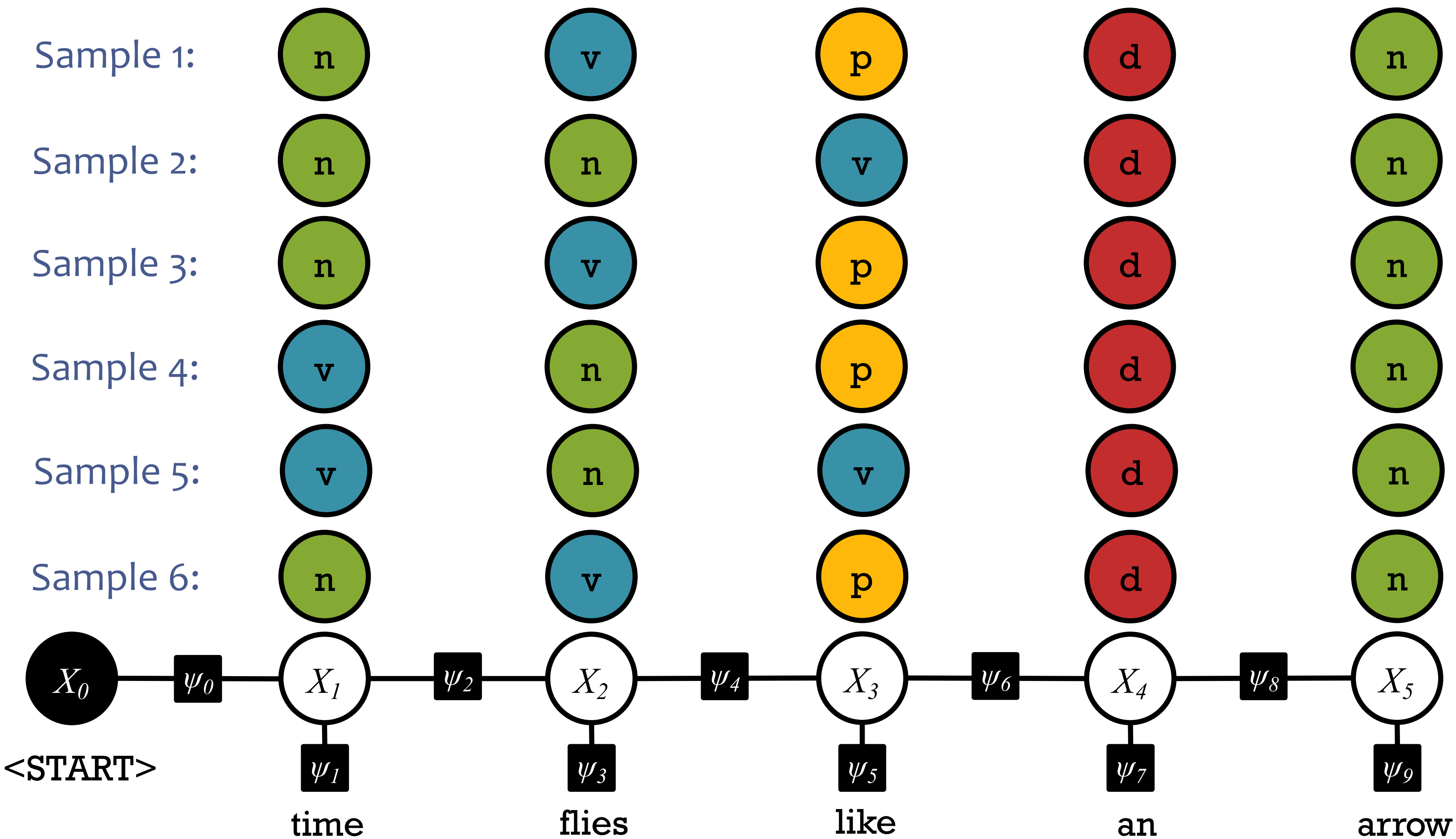


Factor Graph



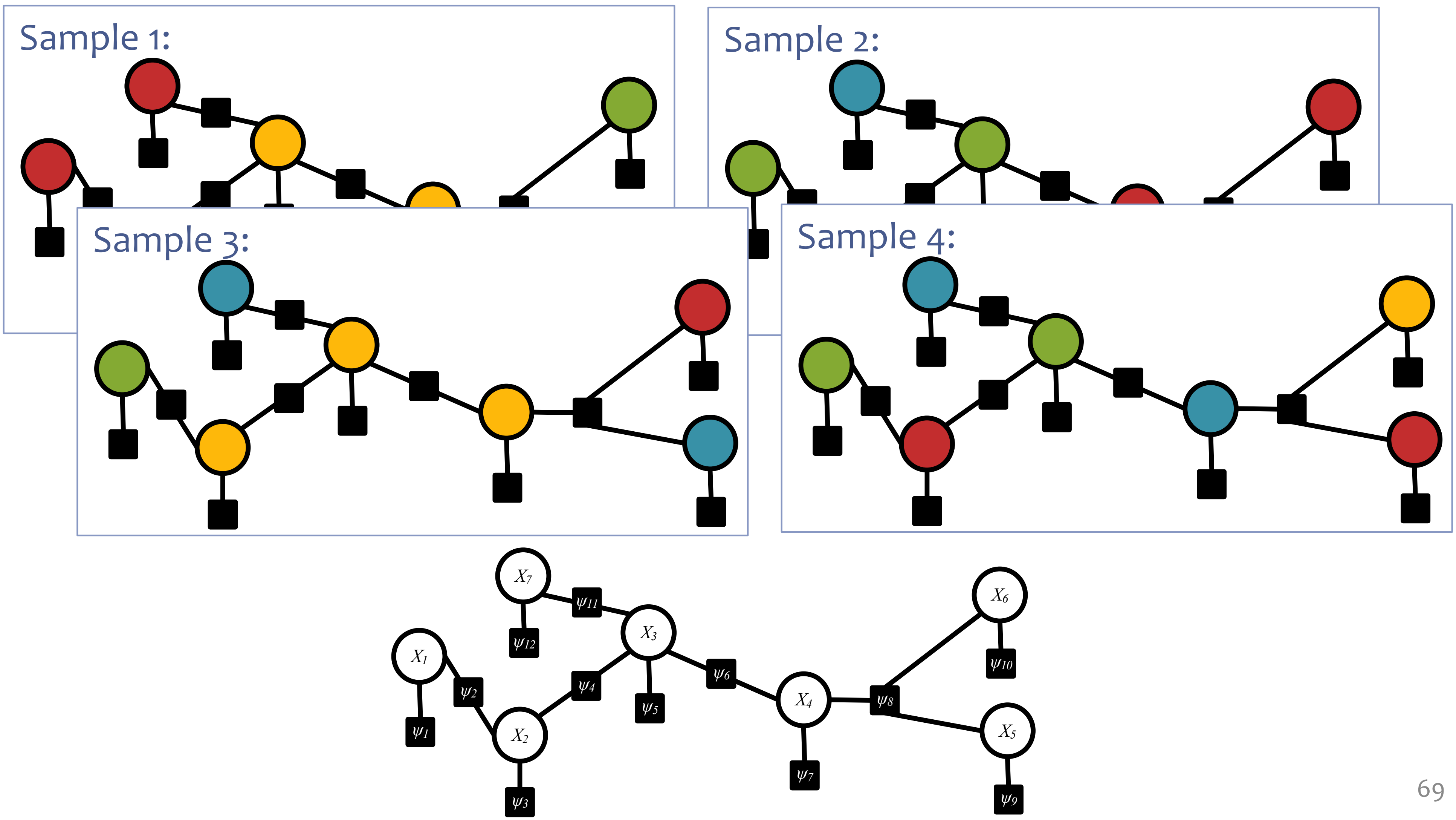
Sampling from a Joint Distribution

A **joint distribution** defines a probability $p(\boldsymbol{x})$ for each assignment of values $\boldsymbol{x}$ to variables $\boldsymbol{X}$. This gives the **proportion** of samples that will equal $\boldsymbol{x}$.



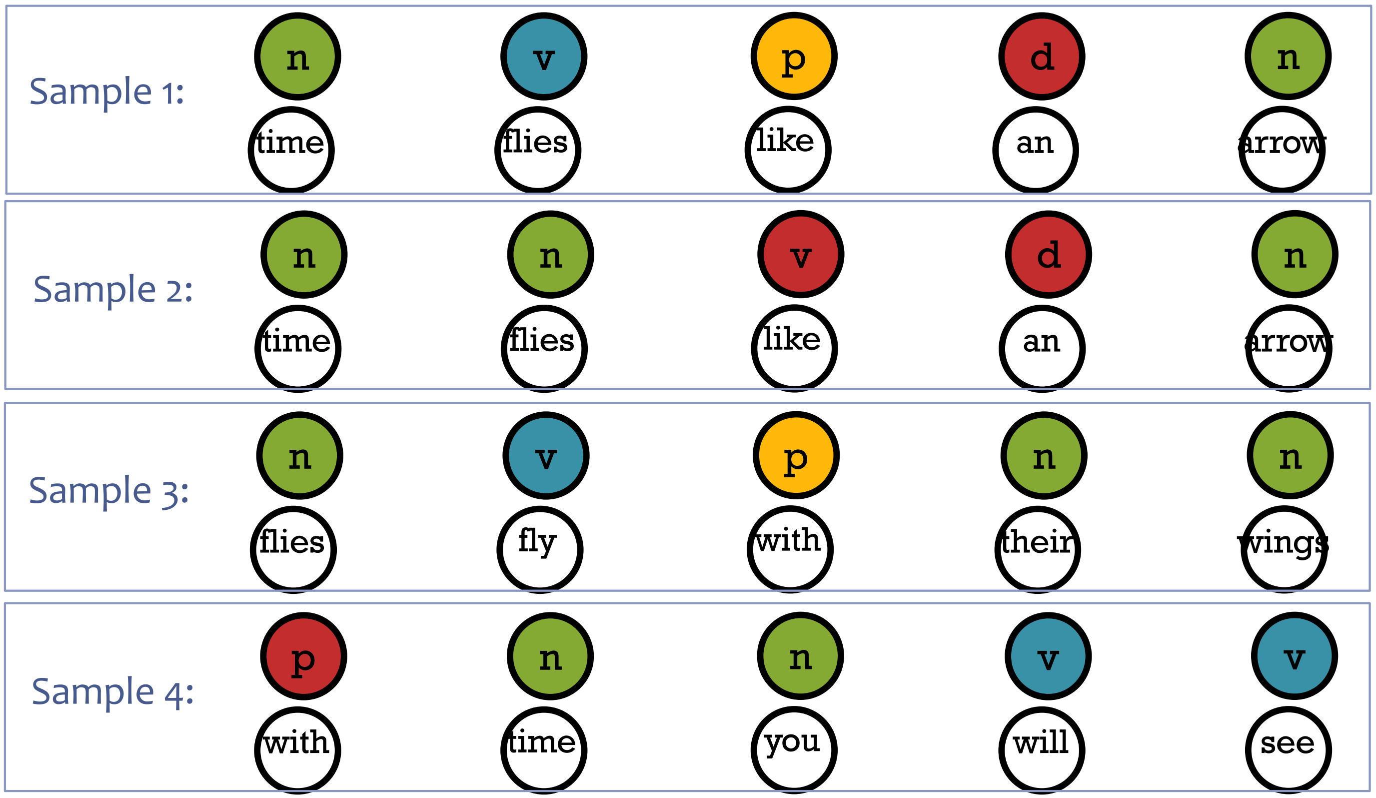
Sampling from a Joint Distribution

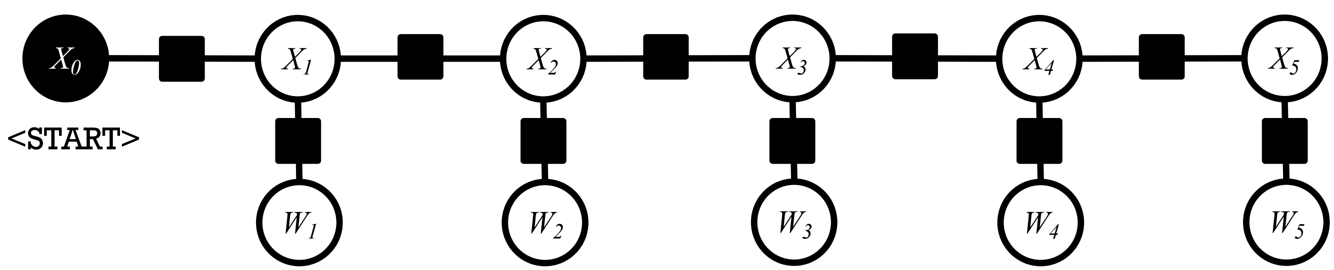
A **joint distribution** defines a probability $p(\boldsymbol{x})$ for each assignment of values $\boldsymbol{x}$ to variables $\boldsymbol{X}$. This gives the **proportion** of samples that will equal $\boldsymbol{x}$.



Sampling from a Joint Distribution

A **joint distribution** defines a probability $p(\boldsymbol{x})$ for each assignment of values $\boldsymbol{x}$ to variables $\boldsymbol{X}$. This gives the **proportion** of samples that will equal $\boldsymbol{x}$.

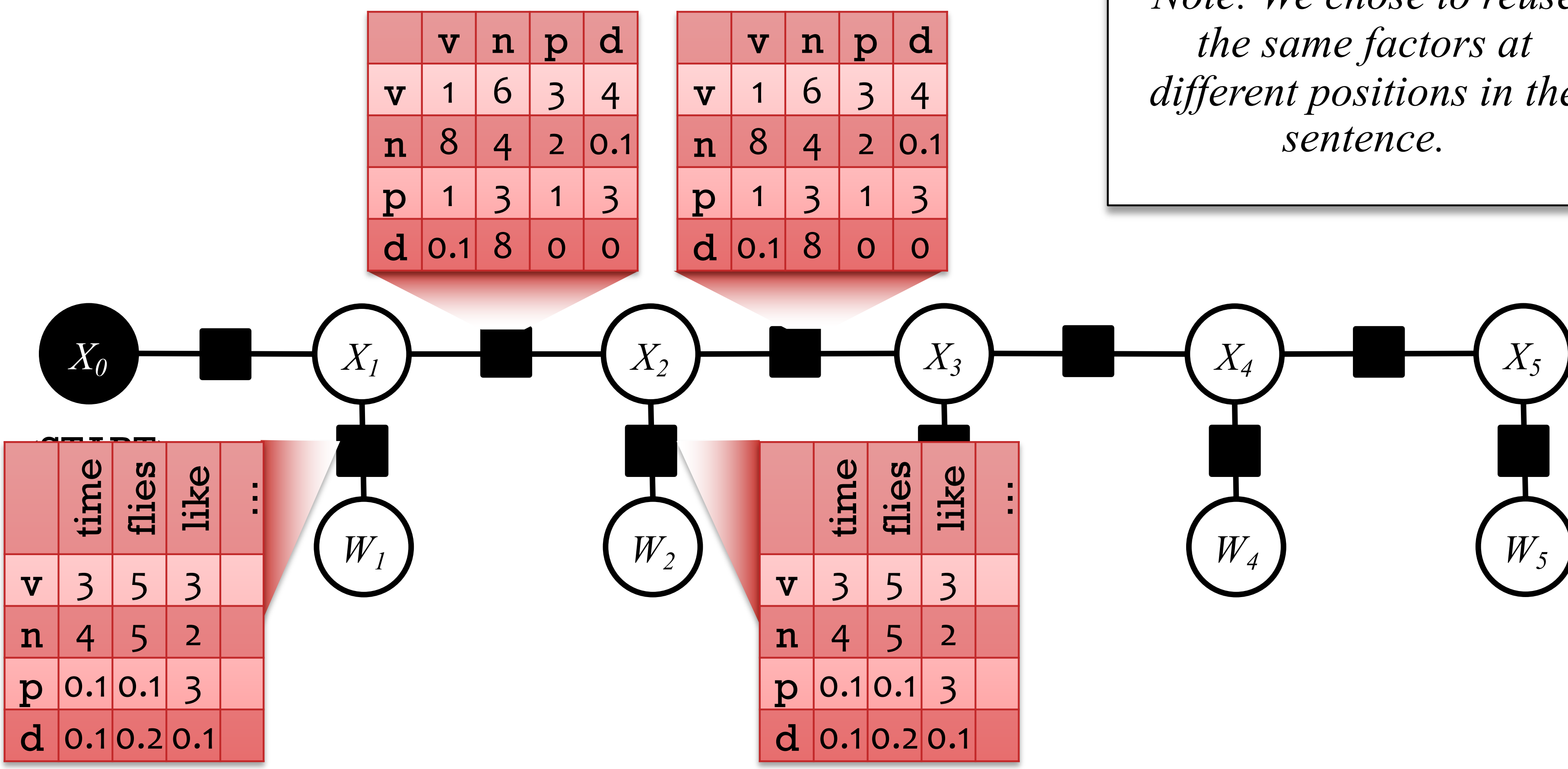




Factors have local opinions (≥ 0)

Each black box looks at *some* of the tags X_i and words W_i

Note: We chose to reuse the same factors at different positions in the sentence.



	v	n	p	d
v	1	6	3	4
n	8	4	2	0.1
p	1	3	1	3
d	0.1	8	0	0

	v	n	p	d
v	1	6	3	4
n	8	4	2	0.1
p	1	3	1	3
d	0.1	8	0	0

	time	flies	like	…
v	3	5	3	
n	4	5	2	
p	0.1	0.1	3	
d	0.1	0.2	0.1	

	time	flies	like	…
v	3	5	3	
n	4	5	2	
p	0.1	0.1	3	
d	0.1	0.2	0.1	

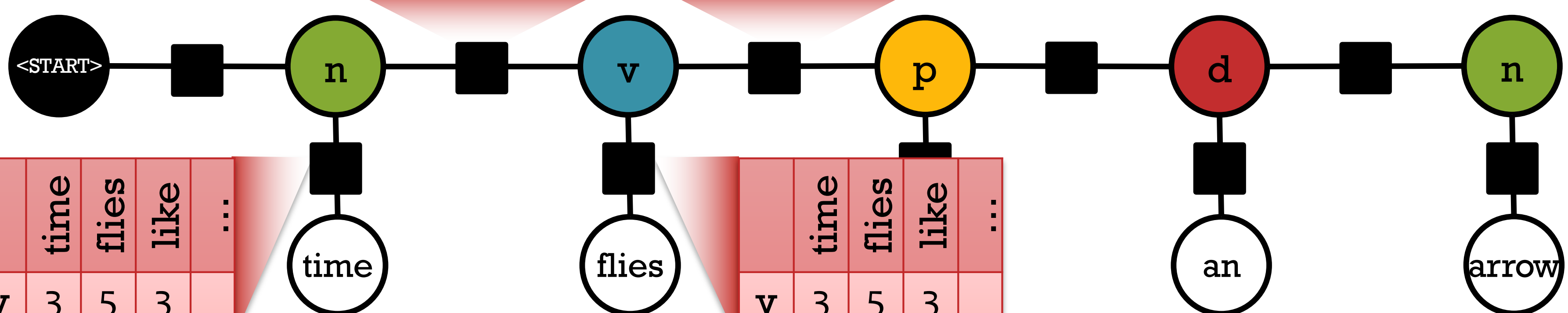
Factors have local opinions (≥ 0)

Each black box looks at *some* of the tags X_i and words W_i

$$p(\text{n, v, p, d, n, time, flies, like, an, arrow}) \quad = \quad ?$$

	v	n	p	d
v	1	6	3	4
n	8	4	2	0.1
p	1	3	1	3
d	0.1	8	0	0

	v	n	p	d
v	1	6	3	4
n	8	4	2	0.1
p	1	3	1	3
d	0.1	8	0	0



	time	flies	like	…
v	3	5	3	
n	4	5	2	
p	0.1	0.1	3	
d	0.1	0.2	0.1	

	time	flies	like	…
v	3	5	3	
n	4	5	2	
p	0.1	0.1	3	
d	0.1	0.2	0.1	

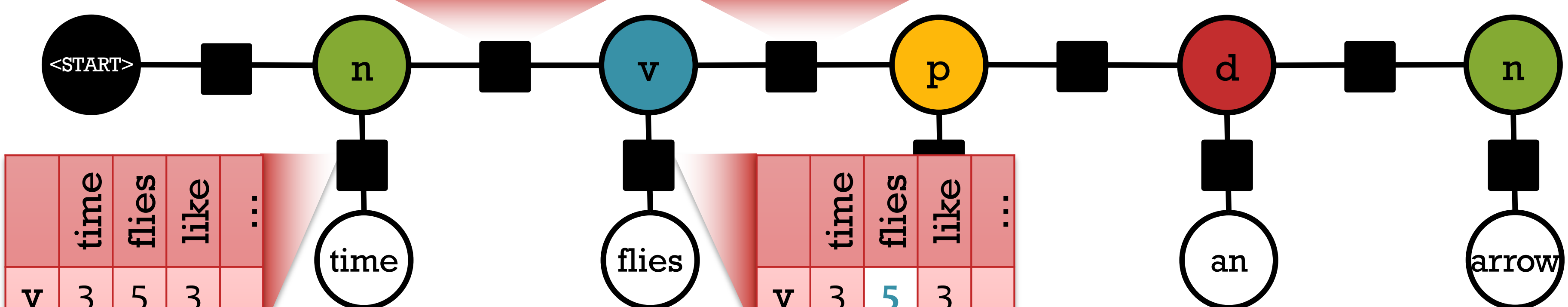
Global probability = product of local opinions

Each black box looks at *some* of the tags X_i and words W_i

$$p(\text{n, v, p, d, n, time, flies, like, an, arrow}) = \frac{1}{Z}(4 * 8 * 5 * 3 * \ldots)$$

Uh-oh! The probabilities of the various assignments sum up to Z > 1.

So divide them all by Z.



	v	n	p	d
v	1	6	3	4
n	8	4	2	0.1
p	1	3	1	3
d	0.1	8	0	0

	v	n	p	d
v	1	6	3	4
n	8	4	2	0.1
p	1	3	1	3
d	0.1	8	0	0

	time	flies	like	…
v	3	5	3	
n	4	5	2	
p	0.1	0.1	3	
d	0.1	0.2	0.1	

	time	flies	like	…
v	3	5	3	
n	4	5	2	
p	0.1	0.1	3	
d	0.1	0.2	0.1	

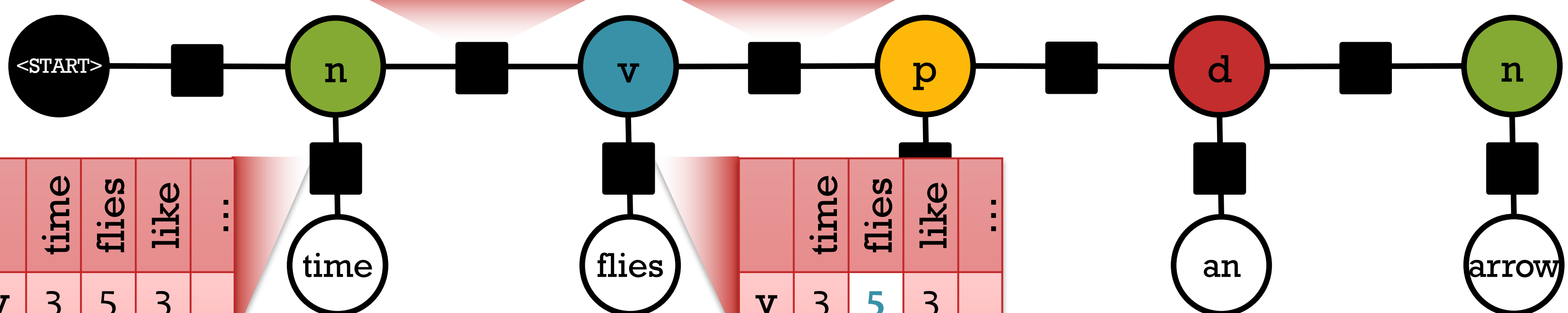
Markov Random Field (MRF)

Joint distribution over tags X_i <u>and</u> words W_i
The individual factors aren't *necessarily* probabilities.

$$p(\text{n, v, p, d, n, time, flies, like, an, arrow}) = \frac{1}{Z}(4 * 8 * 5 * 3 * \ldots)$$

	v	n	p	d
v	1	6	3	4
n	8	4	2	0.1
p	1	3	1	3
d	0.1	8	0	0

	v	n	p	d
v	1	6	3	4
n	8	4	2	0.1
p	1	3	1	3
d	0.1	8	0	0



	time	flies	like	…
v	3	5	3	
n	4	5	2	
p	0.1	0.1	3	
d	0.1	0.2	0.1	

	time	flies	like	…
v	3	5	3	
n	4	5	2	
p	0.1	0.1	3	
d	0.1	0.2	0.1	

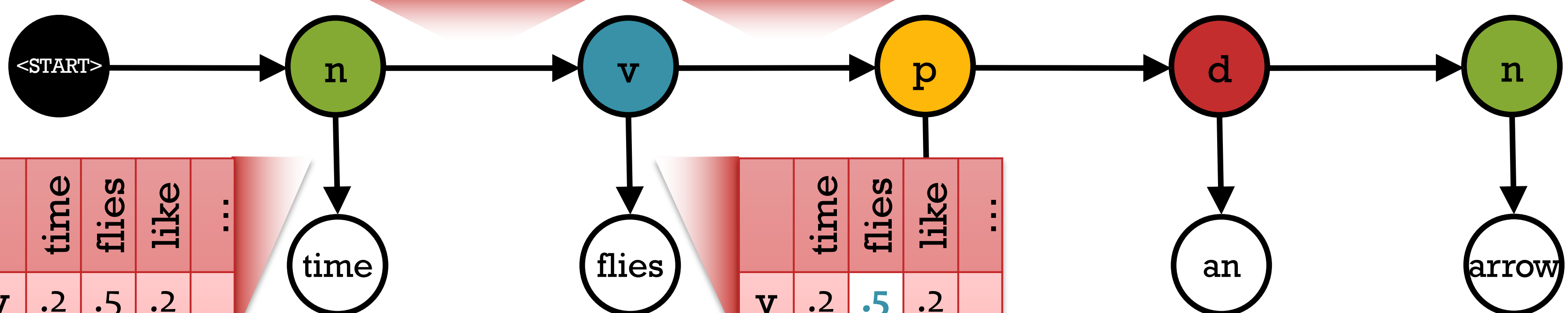
Bayesian Networks

But sometimes we *choose* to make them probabilities.
Constrain each row of a factor to sum to one. Now $Z = 1$.

$$p(\text{n, v, p, d, n, time, flies, like, an, arrow}) = \cancel{\frac{1}{Z}}(.3 * .8 * .2 * .5 * \dots)$$

	v	n	p	d
v	.1	.4	.2	.3
n	.8	.1	.1	0
p	.2	.3	.2	.3
d	.2	.8	0	0

	v	n	p	d
v	.1	.4	.2	.3
n	.8	.1	.1	0
p	.2	.3	.2	.3
d	.2	.8	0	0



	time	flies	like	…
v	.2	.5	.2	
n	.3	.4	.2	
p	.1	.1	.3	
d	.1	.2	.1	

	time	flies	like	…
v	.2	.5	.2	
n	.3	.4	.2	
p	.1	.1	.3	
d	.1	.2	.1	

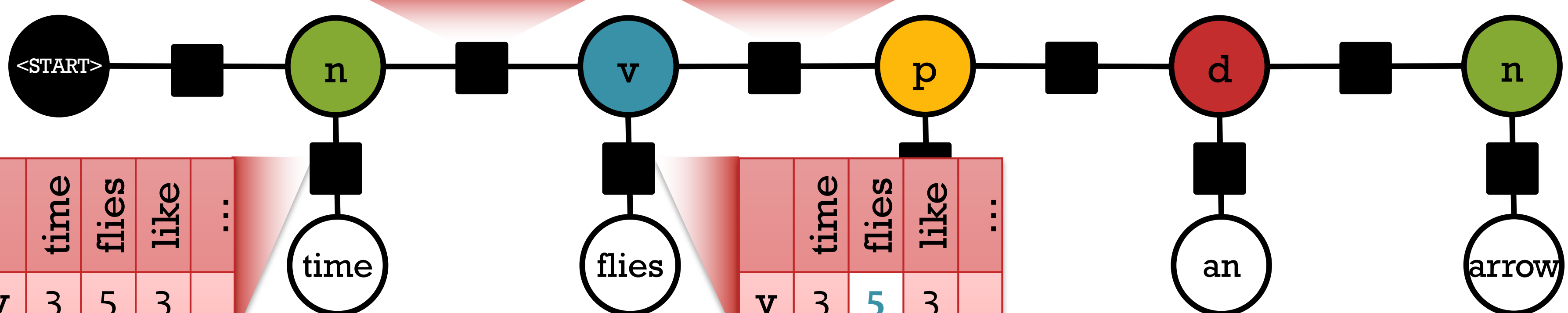
Markov Random Field (MRF)

Joint distribution over tags X_i and words W_i

$$p(\text{n, v, p, d, n, time, flies, like, an, arrow}) = \frac{1}{Z}(4 * 8 * 5 * 3 * \ldots)$$

	v	n	p	d
v	1	6	3	4
n	8	4	2	0.1
p	1	3	1	3
d	0.1	8	0	0

	v	n	p	d
v	1	6	3	4
n	8	4	2	0.1
p	1	3	1	3
d	0.1	8	0	0



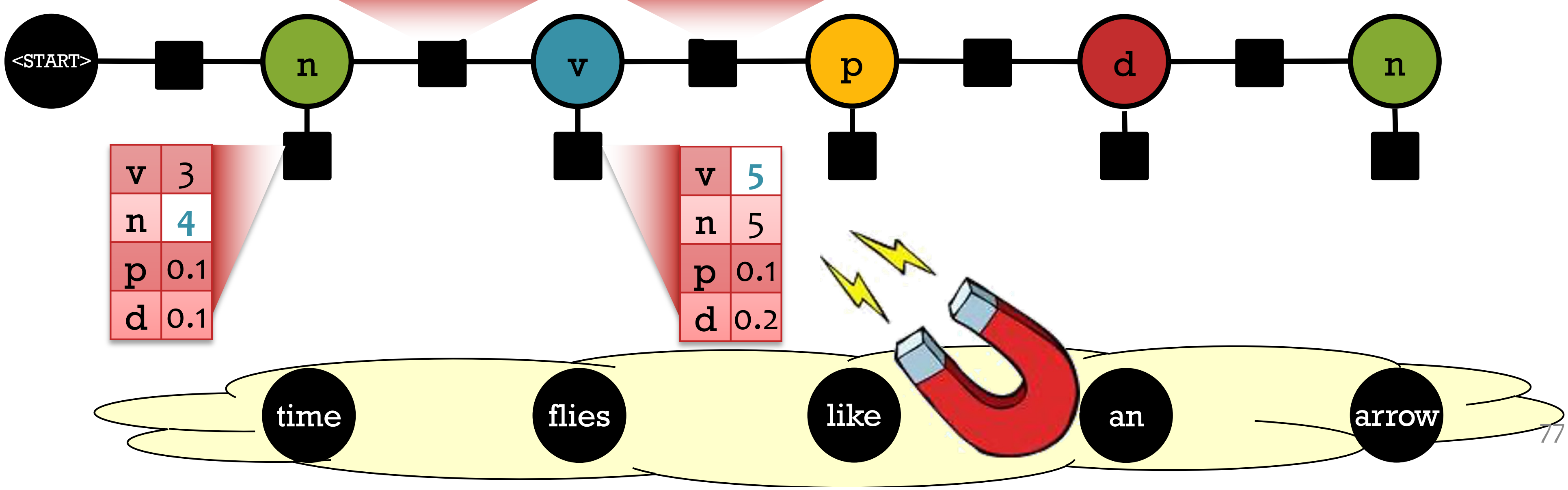
	time	flies	like	…
v	3	5	3	
n	4	5	2	
p	0.1	0.1	3	
d	0.1	0.2	0.1	

	time	flies	like	…
v	3	5	3	
n	4	5	2	
p	0.1	0.1	3	
d	0.1	0.2	0.1	

Conditional Random Field (CRF)

Conditional distribution over tags X_i given words w_i.
The factors and Z are now specific to the sentence w.

$$p(\text{n, v, p, d, n} \mid \text{time, flies, like, an, arrow}) = \frac{1}{Z}(4 * 8 * 5 * 3 * \ldots)$$



	v	n	p	d
v	1	6	3	4
n	8	4	2	0.1
p	1	3	1	3
d	0.1	8	0	0

	v	n	p	d
v	1	6	3	4
n	8	4	2	0.1
p	1	3	1	3
d	0.1	8	0	0

v	3
n	4
p	0.1
d	0.1

v	5
n	5
p	0.1
d	0.2

How General Are Factor Graphs?

- Factor graphs can be used to describe
  - Markov Random Fields (undirected graphical models)
    - i.e., log-linear models over a tuple of variables
  - Conditional Random Fields
  - Bayesian Networks (directed graphical models)

- *Inference* treats all of these interchangeably.
  - Convert your model to a factor graph first.
  - Pearl (1988) gave key strategies for *exact* inference:
    - **Belief propagation**, for inference on *acyclic* graphs
    - **Junction tree algorithm**, for making *any* graph acyclic (by merging variables and factors: blows up the runtime)

Factor Graph Notation

- Variables:

$$\mathcal{X} = \{X_1, \ldots, X_i, \ldots, X_n\}$$

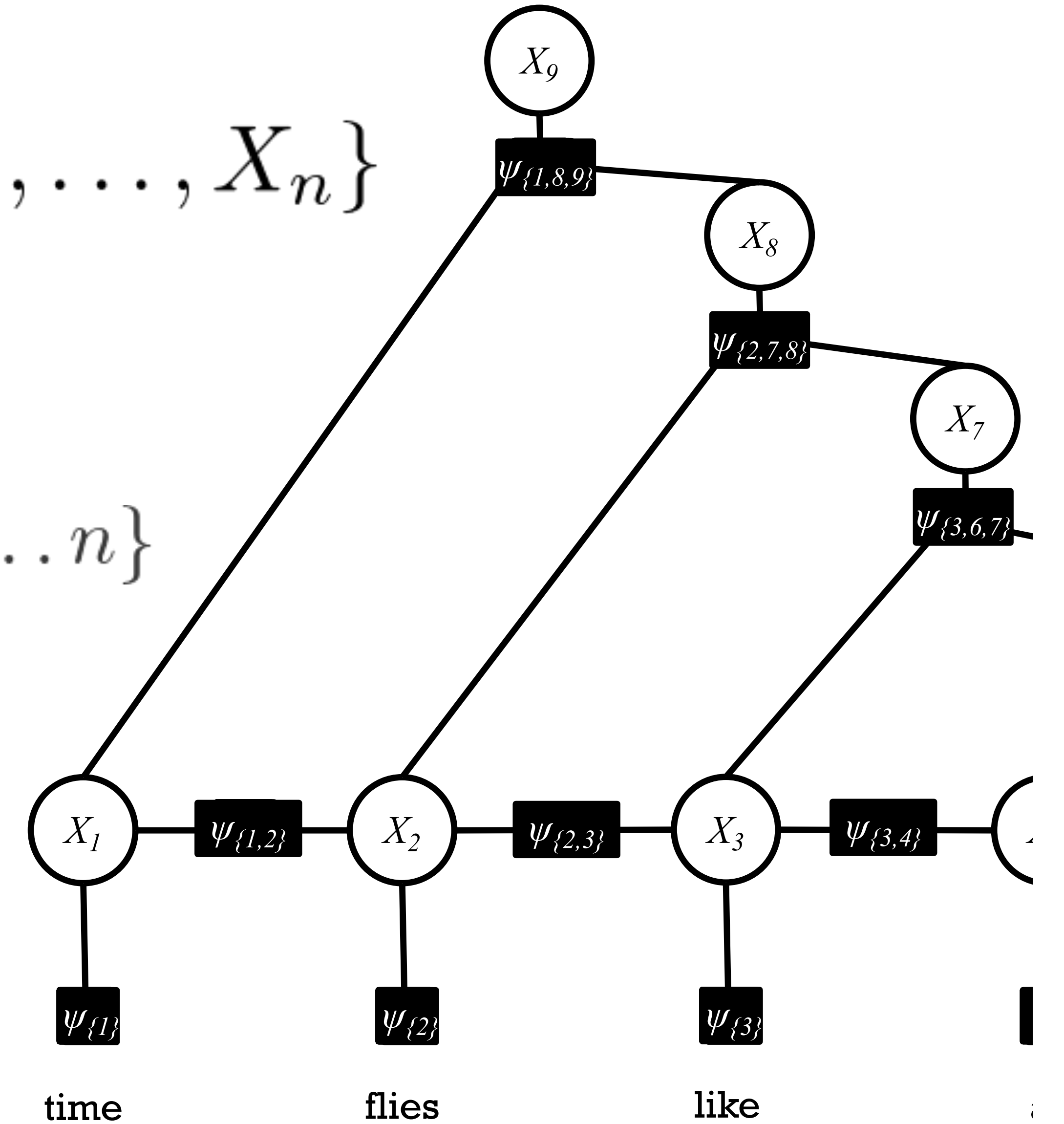
- Factors:

$$\psi_\alpha, \psi_\beta, \psi_\gamma, \ldots$$

where $\alpha, \beta, \gamma, \ldots \subseteq \{1, \ldots n\}$

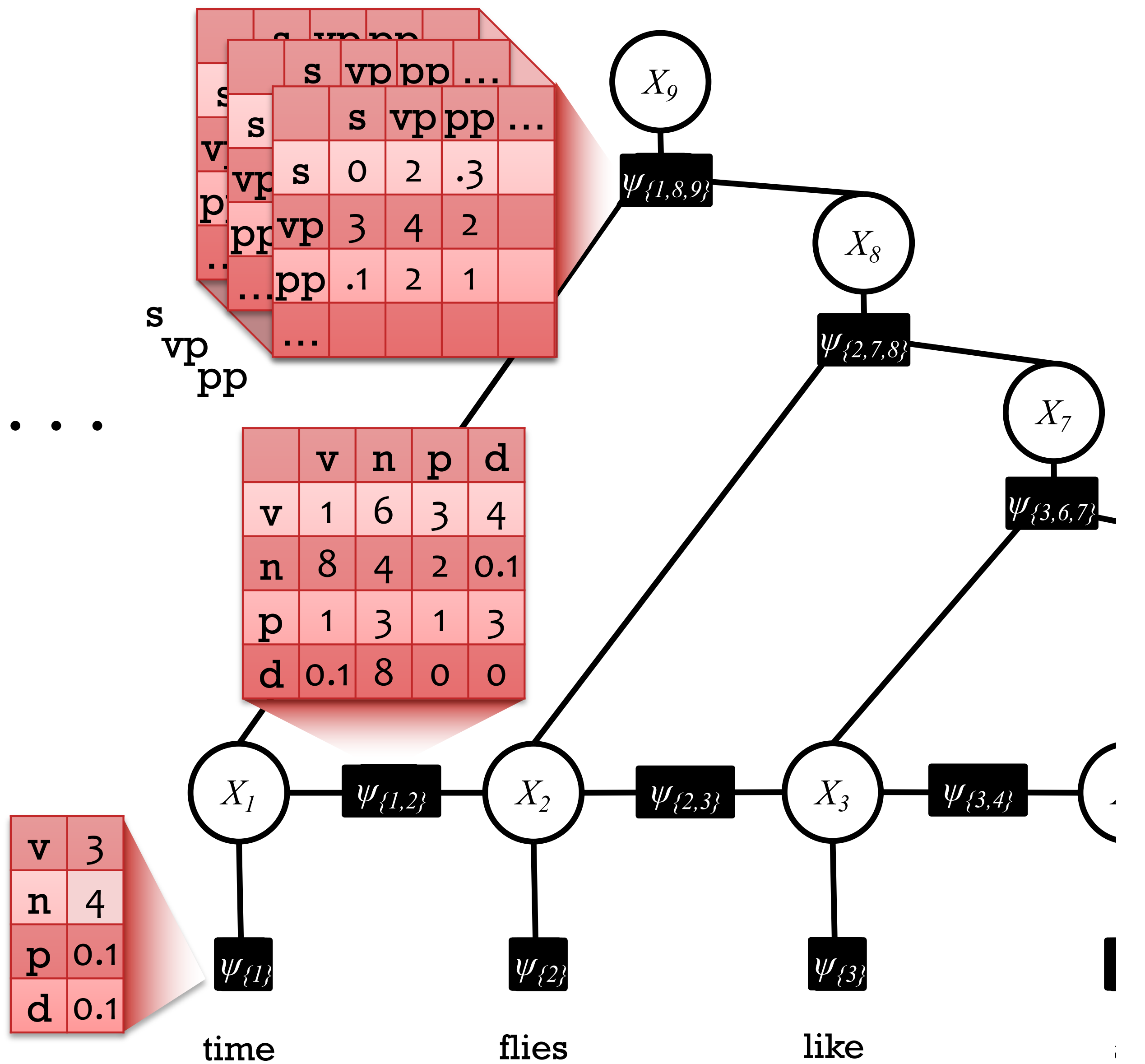
Joint Distribution

$$p(\boldsymbol{x}) = \frac{1}{Z} \prod_\alpha \psi_\alpha(\boldsymbol{x}_{\boldsymbol{\alpha}})$$



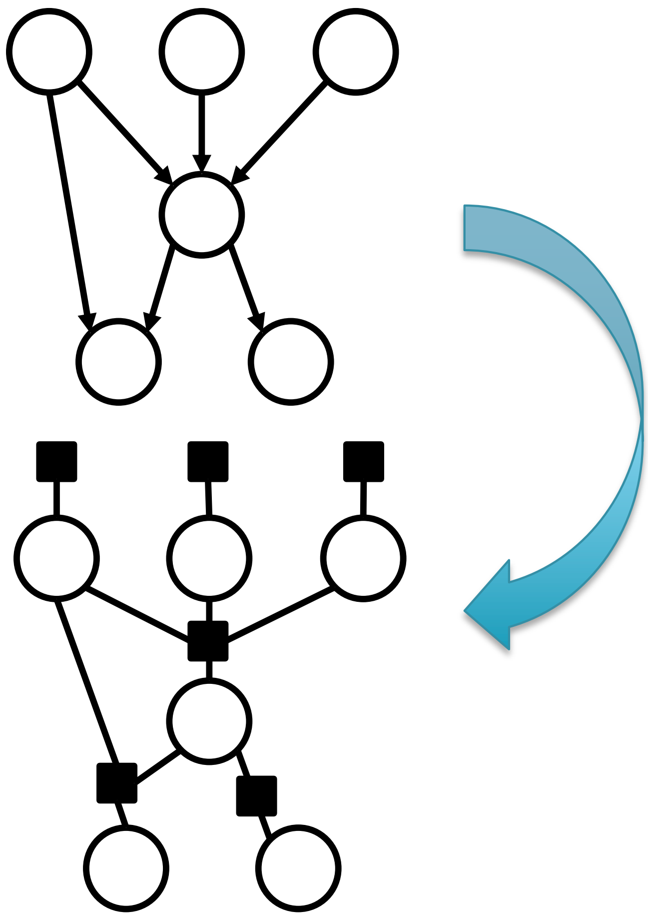
Factors are Tensors

- Factors:
  $\psi_\alpha, \psi_\beta, \psi_\gamma, \ldots$

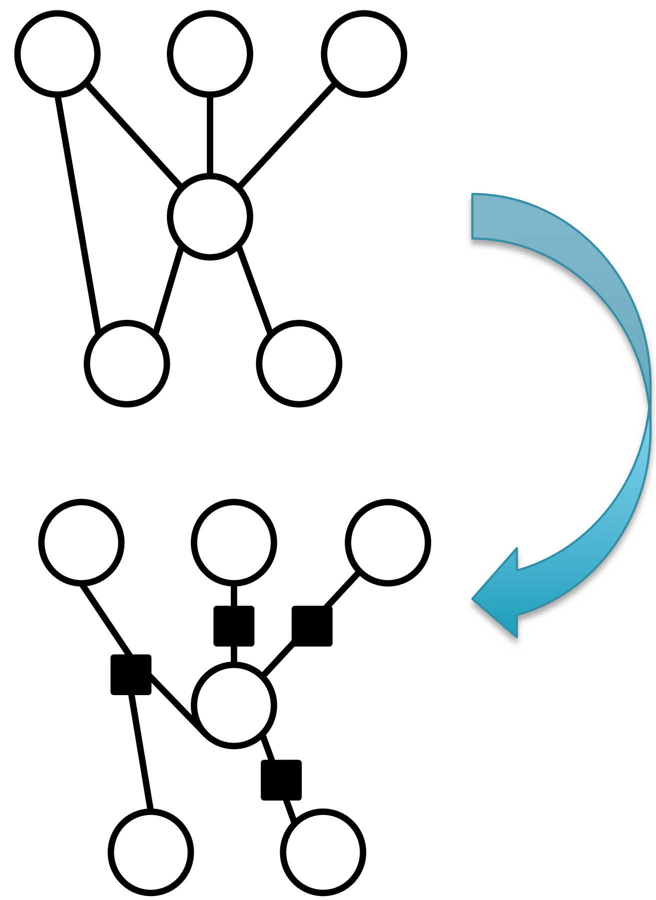


Converting to Factor Graphs

Each conditional and marginal distribution in a **directed GM** becomes a factor



Each maximal clique in an **undirected GM** becomes a factor



Equivalence of directed and undirected trees

- Any undirected tree can be converted to a directed tree by choosing a root node and directing all edges away from it
- A directed tree and the corresponding undirected tree make the same conditional independence assertions
- Parameterizations are essentially the same.
  - Undirected tree:
  - Directed tree:
  - Equivalence:

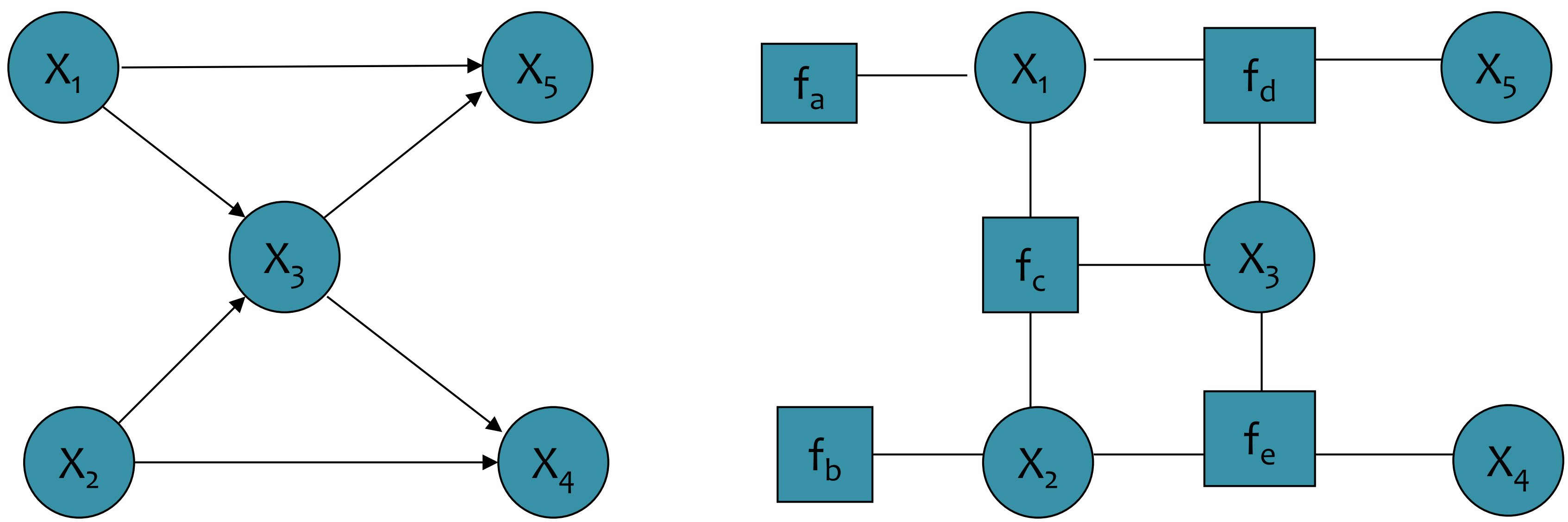
$$p(x) = \frac{1}{Z}\left(\prod_{i\in V}\psi(x_i)\prod_{(i,j)\in E}\psi(x_i,x_j)\right)$$

$$p(x) = p(x_r)\prod_{(i,j)\in E}p(x_j|x_i)$$

$$\psi(x_r) = p(x_r);\quad \psi(x_i,x_j) = p(x_j|x_i);$$
$$Z = 1,\quad \psi(x_i) = 1$$

Factor Graph Examples

- Example 1

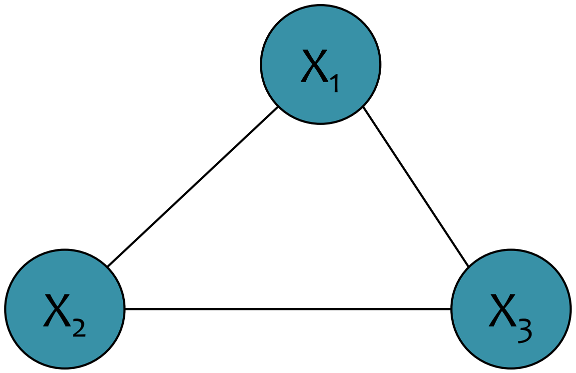


$P(X_1)$ $P(X_2)$ $P(X_3|X_1,X_2)$ $P(X_5|X_1,X_3)$ $P(X_4|X_2,X_3)$

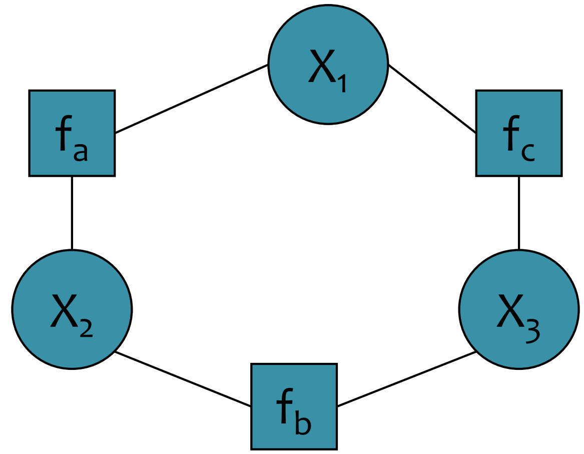
$f_a(X_1)$ $f_b(X_2)$ $f_c(X_3,X_1,X_2)$ $f_d(X_5,X_1,X_3)$ $f_e(X_4,X_2,X_3)$

Factor Graph Examples

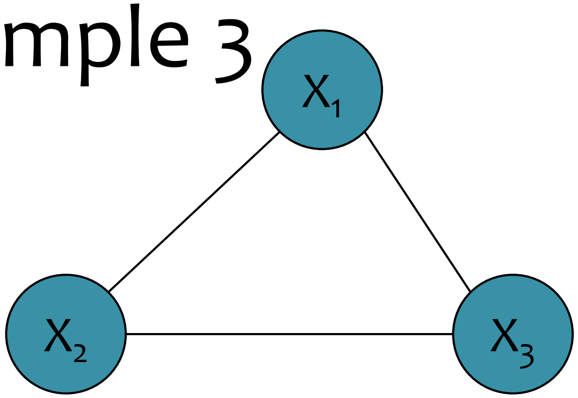
- Example 2



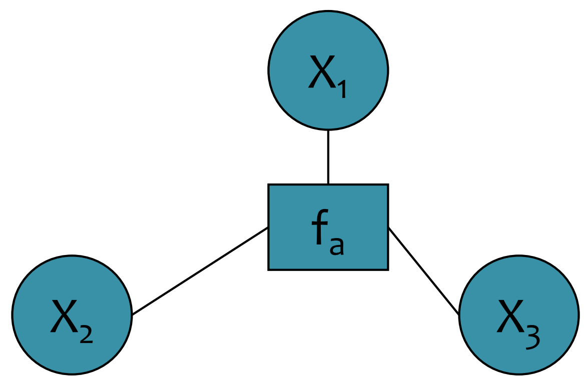
$\psi(x_1,x_2,x_3) = f_a(x_1,x_2)f_b(x_2,x_3)f_c(x_3,x_1)$



- Example 3

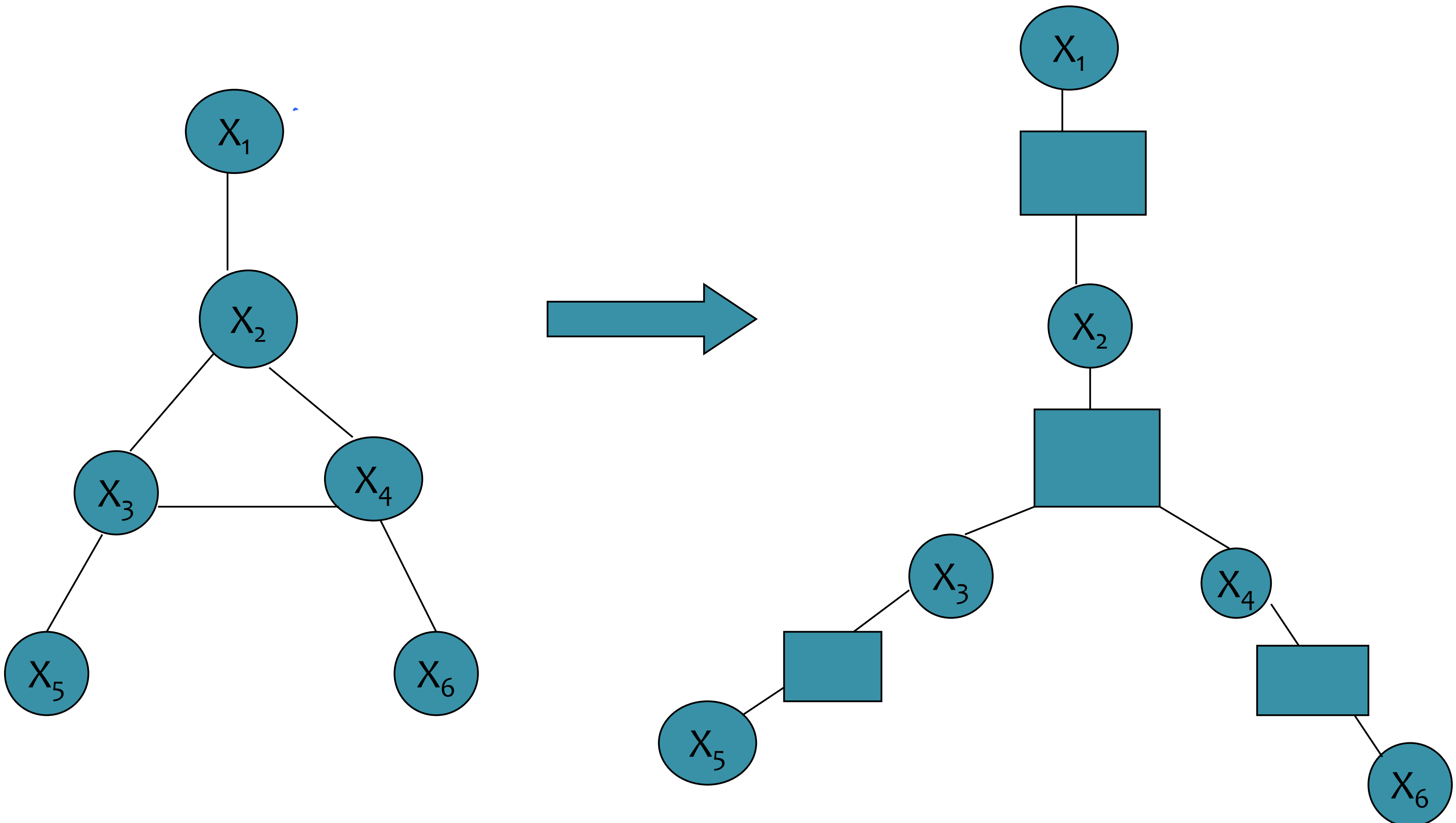


$\psi(x_1,x_2,x_3) = f_a(x_1,x_2,x_3)$



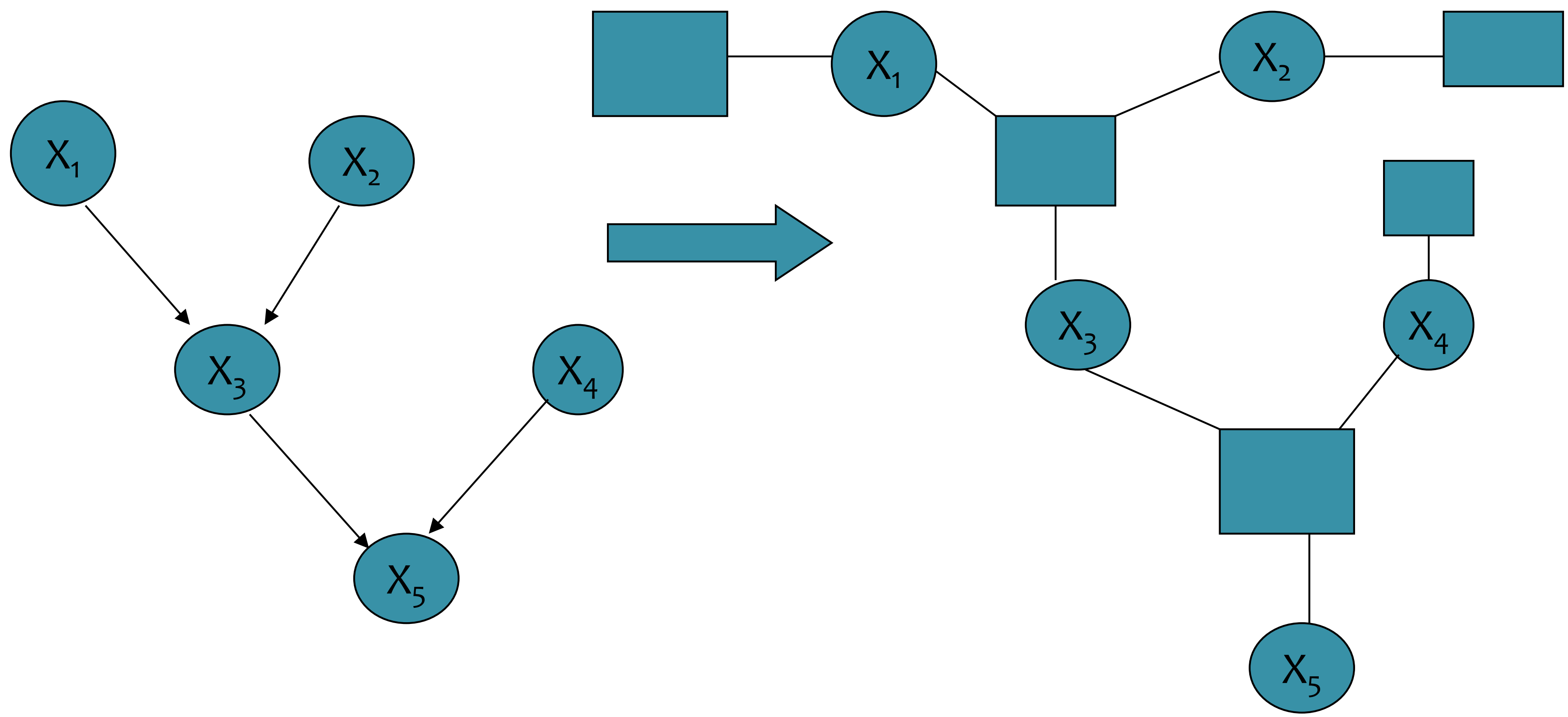
Tree-like Undirected GMs to Factor Trees

- Example 4

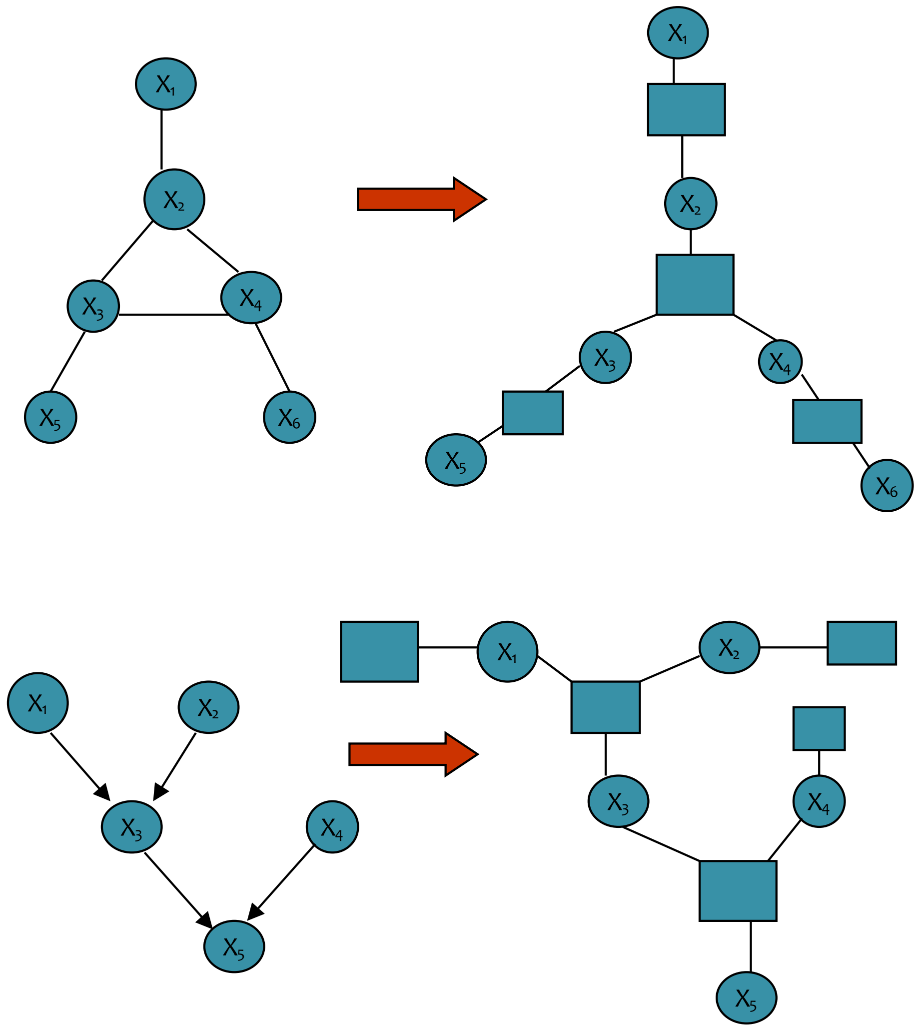


Poly-trees to Factor trees

- Example 5



Why factor graphs?



- Because FG turns tree-like graphs to factor trees,
- Trees are a data-structure that guarantees correctness of BP !

MRF VS. CRF

MRF vs. CRF

Markov Random Field (MRF):

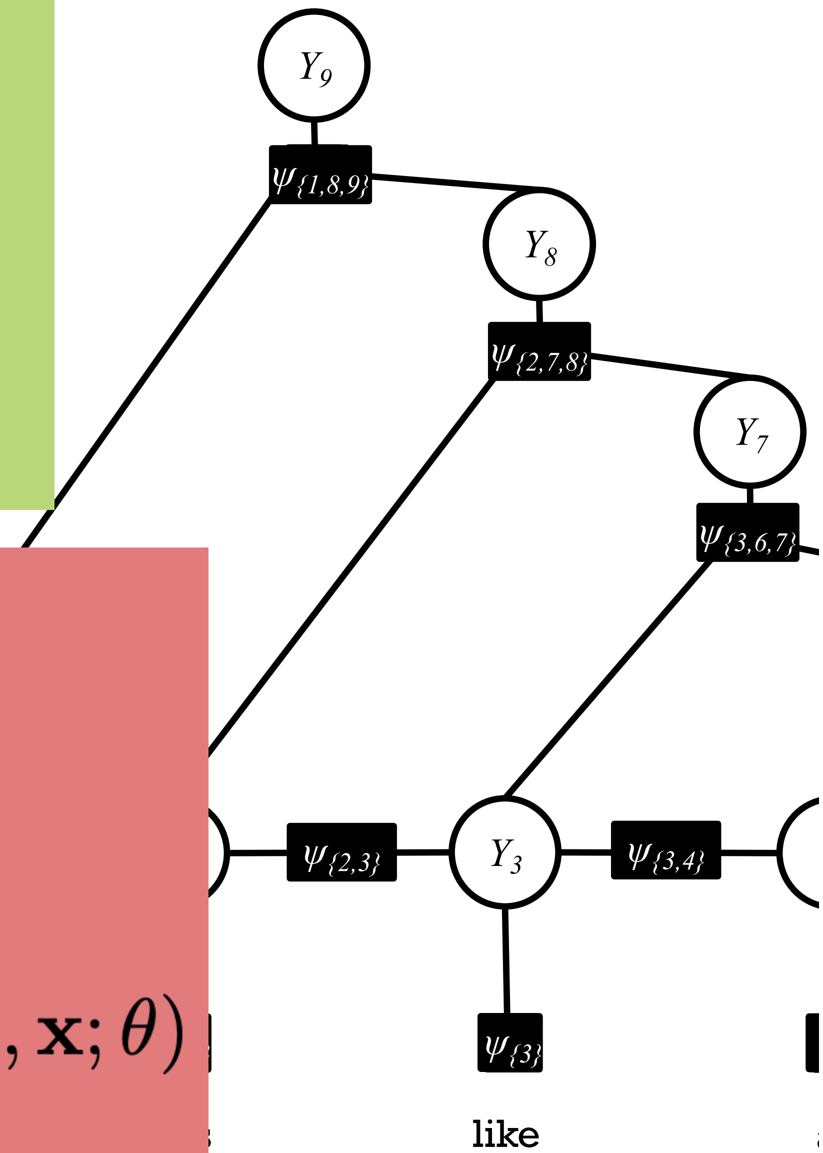
- just a distribution over variables **y**
- partition function Z is just a function of the parameters

$$p_{\boldsymbol{\theta}}(\mathbf{y}) = \frac{1}{Z(\theta)} \prod_{\alpha} \psi_{\alpha}(\mathbf{y}_{\alpha}; \theta)$$

Conditional Random Field (CRF):

- conditions on some additional observed variables **x**
- partition function Z is a function of **x** as well

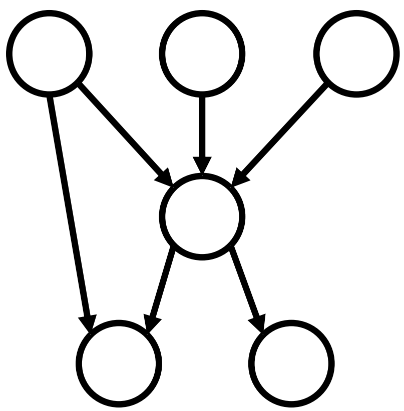
$$p_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z(\mathbf{x}; \theta)} \prod_{\alpha} \psi_{\alpha}(\mathbf{y}_{\alpha}, \mathbf{x}; \theta)$$



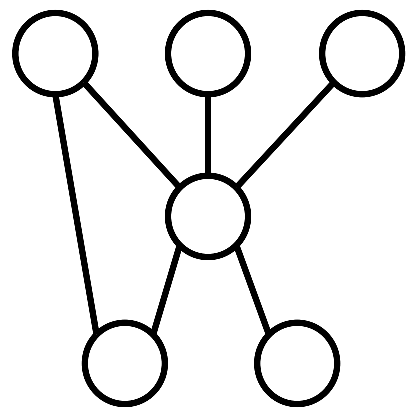
TYPES OF GRAPHICAL MODELS

Three Types of Graphical Models

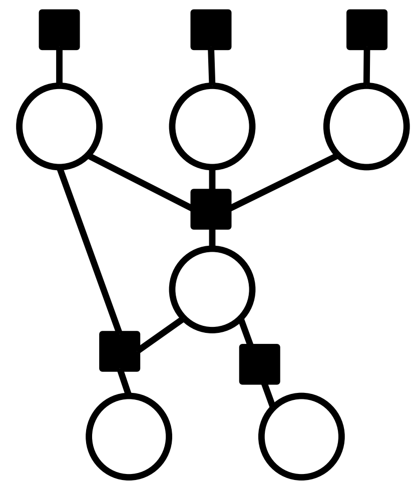
Directed Graphical Model



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**: directed graphs with locally normalized conditional probabilities
2. **Conditional Independencies**: d-separation, Markov blanket
3. **Example parameterization**: conditional probability tables (CPTs) for discrete var.s, conditional probability densities for continuous var.s
4. **Normalization**: locally normalized, partition function is always 1.0

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:** undirected graphs with unormalized potentials
2. **Conditional Independencies:** independence by separation, Markov blanket
3. **Example parameterization:** Markov random field (MRF), conditional random field (CRF), neural potentials
4. **Normalization:** globally normalized

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:** bipartite graph over variables and factors
2. **Conditional Independencies:** independence by separation, inferable from underlying DGM or UGM
3. **Example parameterization:** any DGM parameterization, any UGM parameterization
4. **Normalization:** locally normalized if based on DGM, globally normalized if based on UGM