Machine Learning Department
School of Computer Science
Carnegie Mellon University

Factor Graphs
+
Exact Inference

Matt Gormley
Lecture 8
Sep. 23, 2019

Q&A

Reminders

- **Homework 1: DAgger for seq2seq**
  - Out: Thu, Sep. 12
  - Due: Thu, Sep. 26 at 11:59pm
- **Homework 2: Semantic Segmentation**
  - Out: Thu, Sep. 26
  - Due: Thu, Oct. 10 at 11:59pm

Markov Random Fields

UNDIRECTED GRAPHICAL MODELS

Undirected Graphical Models

Whiteboard

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

Example MRFs

- Pairwise MRF
- Ising model
- Hopfield network
- Potts model

Pairwise Markov Random Field

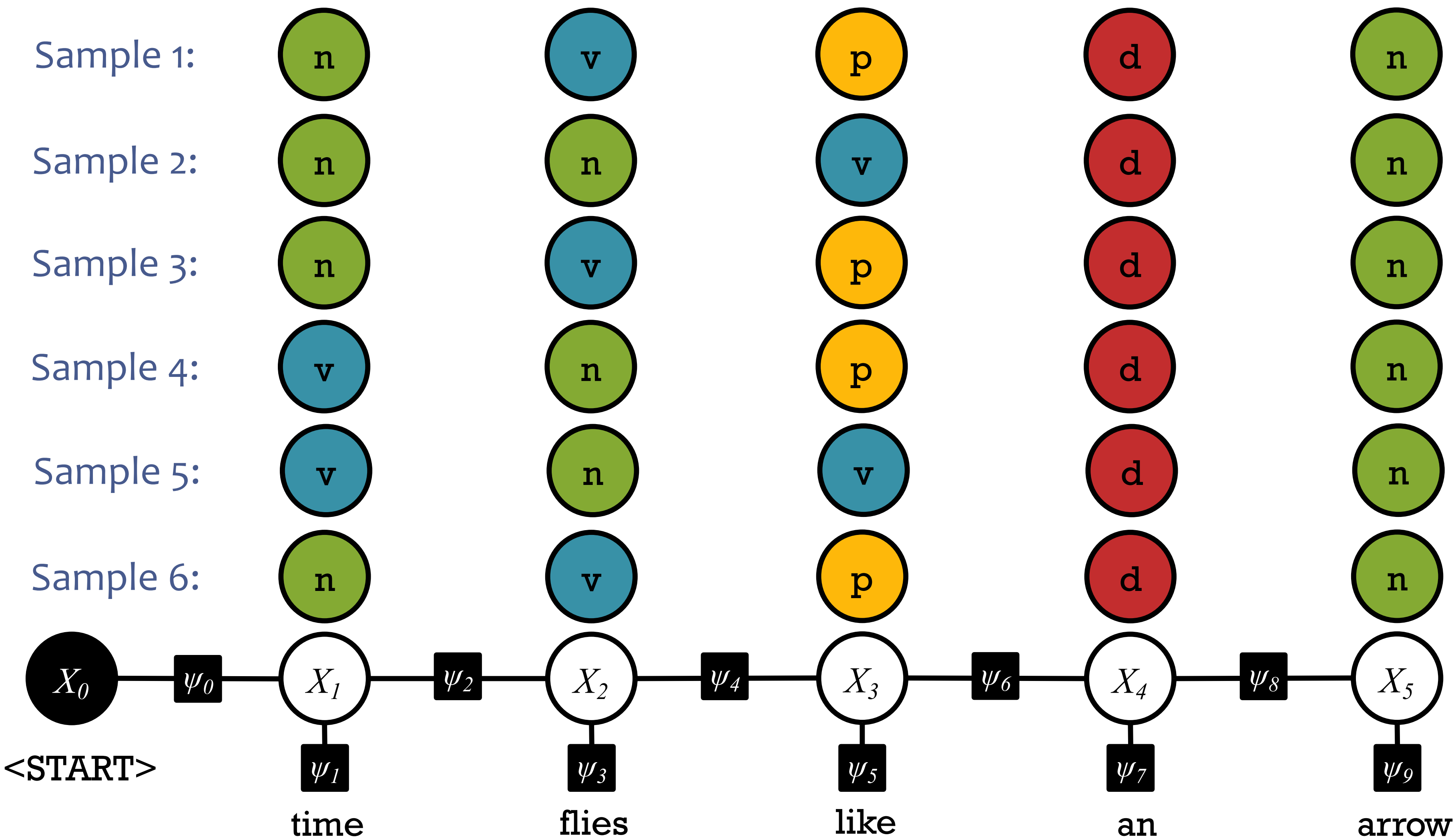
In a **pairwise MRF**, we define potential functions on the edges and the nodes, but not necessarily on maximal cliques

Representation of both directed and undirected graphical models

FACTOR GRAPHS

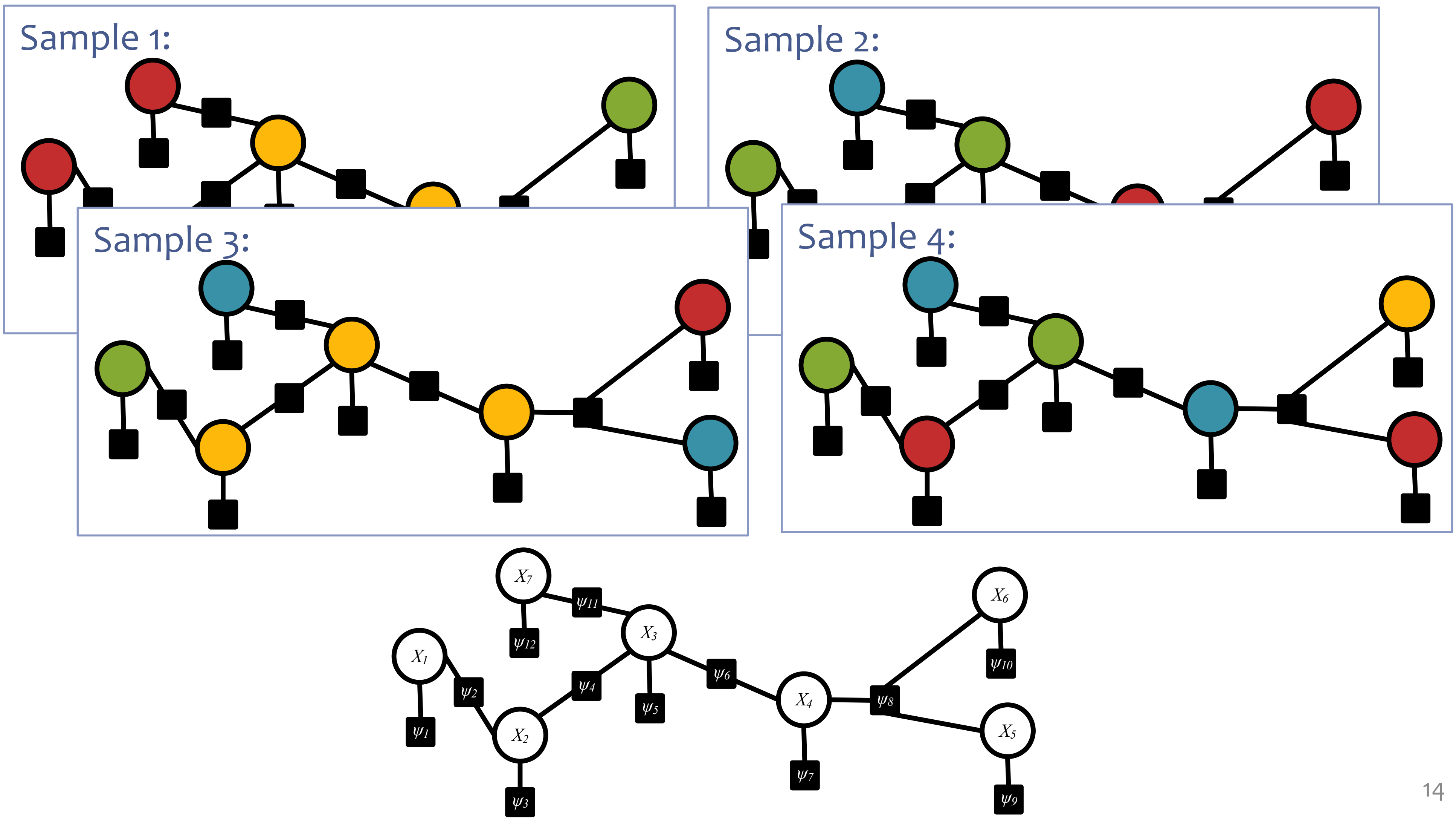
Sampling from a Joint Distribution

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 $\boldsymbol{x}$.



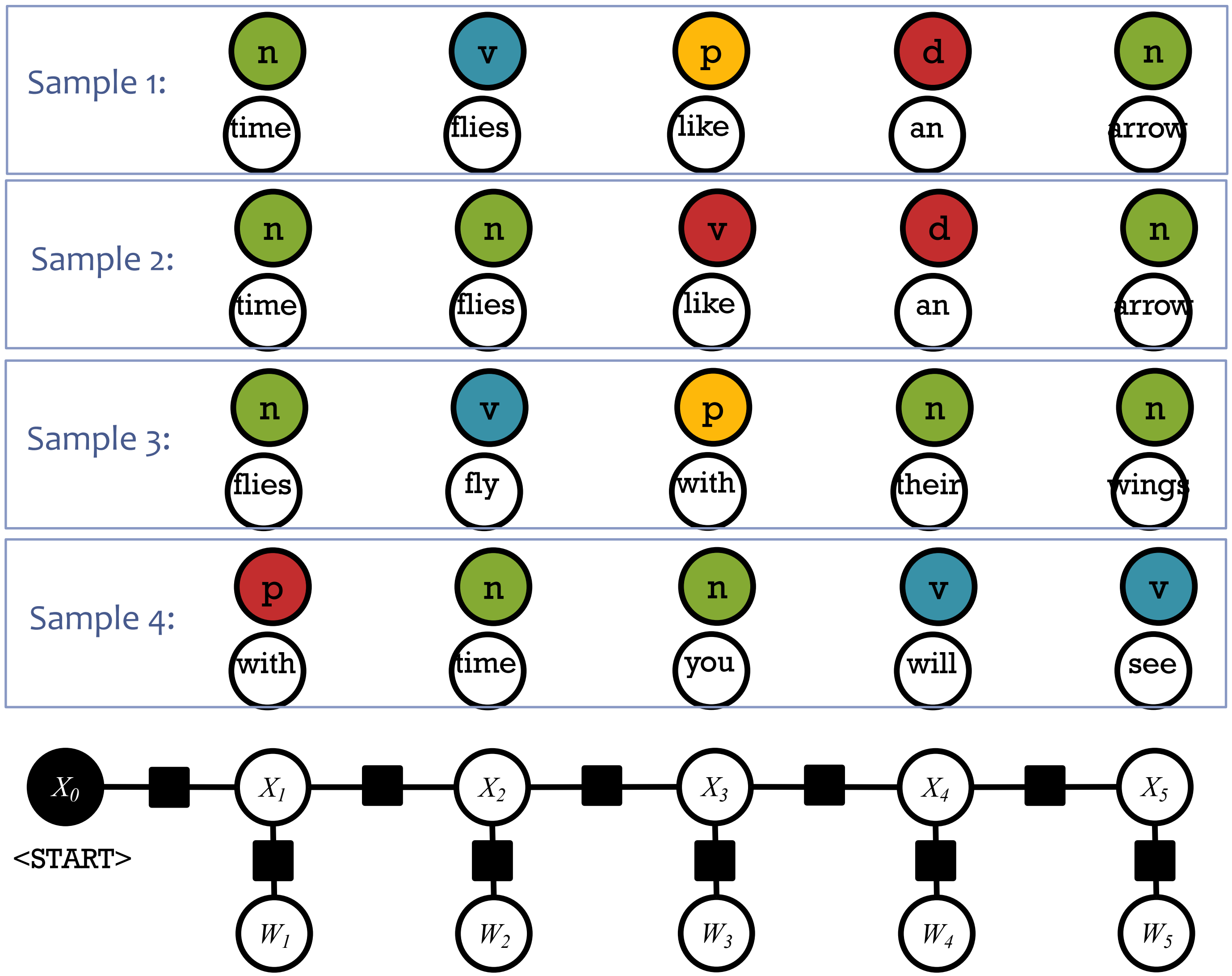
Sampling from a Joint Distribution

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 $\boldsymbol{x}$.



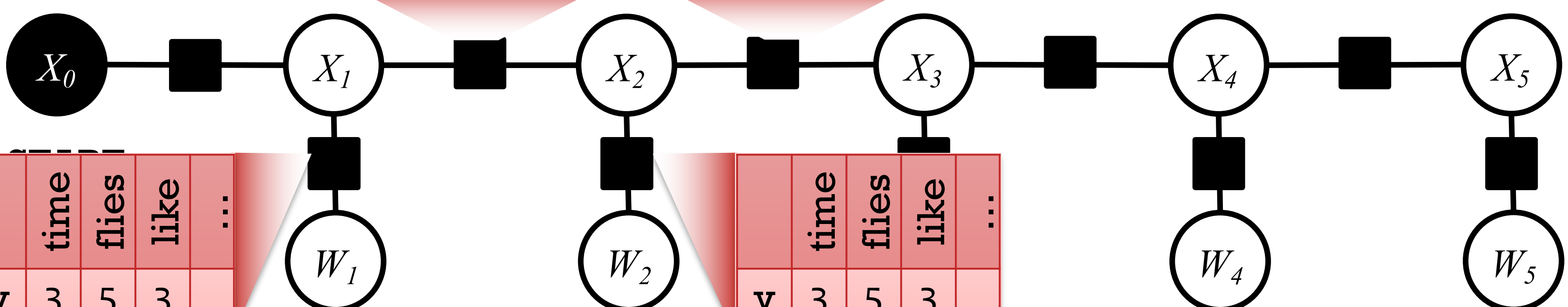
Factors have local opinions (≥ 0)

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	

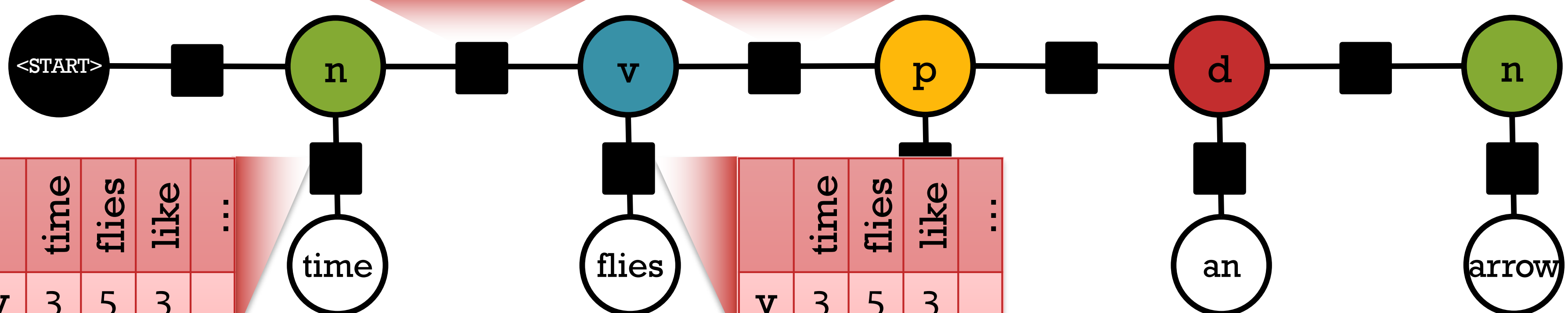
Factors have local opinions (≥ 0)

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

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

	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

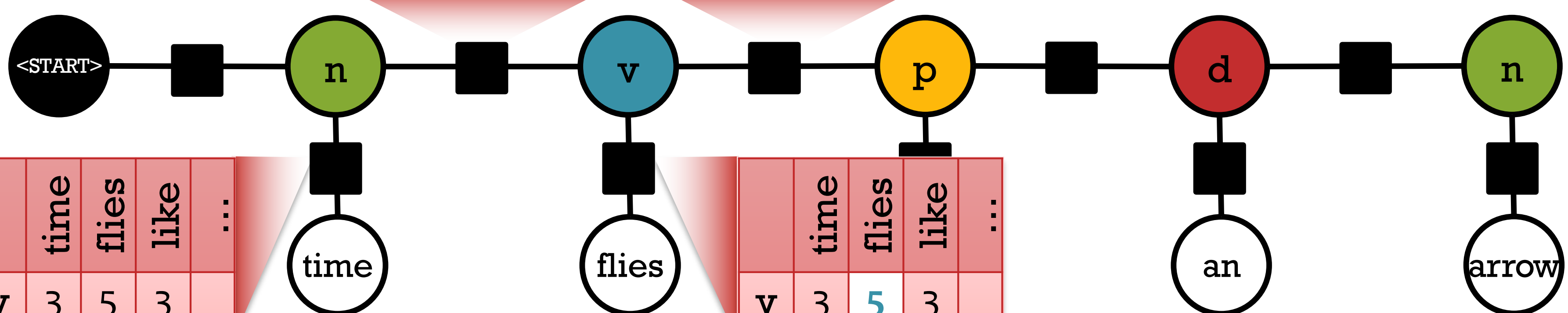
$$p(\text{n, v, p, d, n, time, flies, like, an, arrow}) = \frac{1}{Z}(4 * 8 * 5 * 3 * \ldots)$$

	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

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

So divide them all by Z.



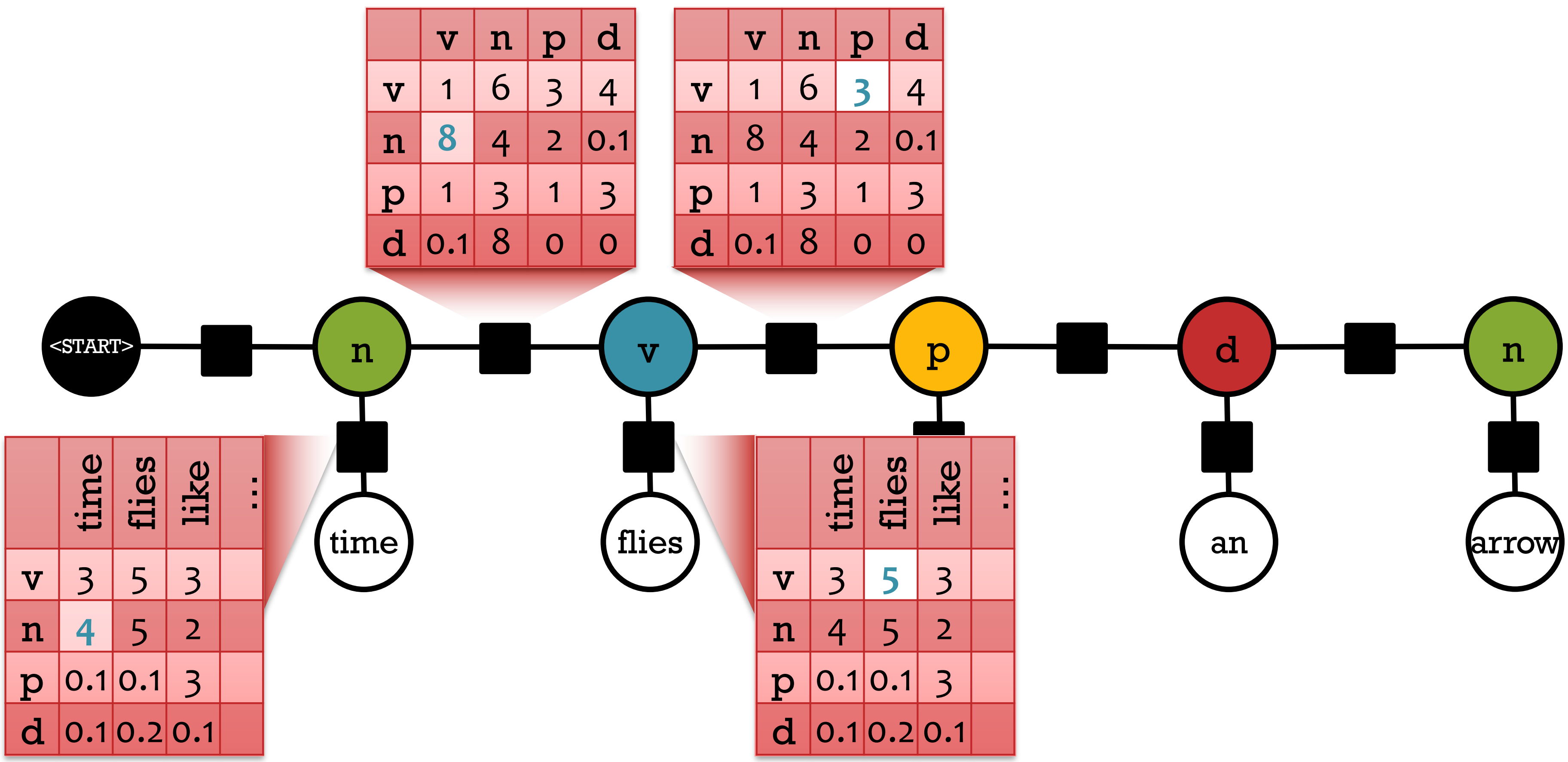
	time	flies	like	…
v	3	5	3	
n	4	5	2	
p	0.1	0.1	3	
d	0.1	0.2	0.1	

	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 and 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

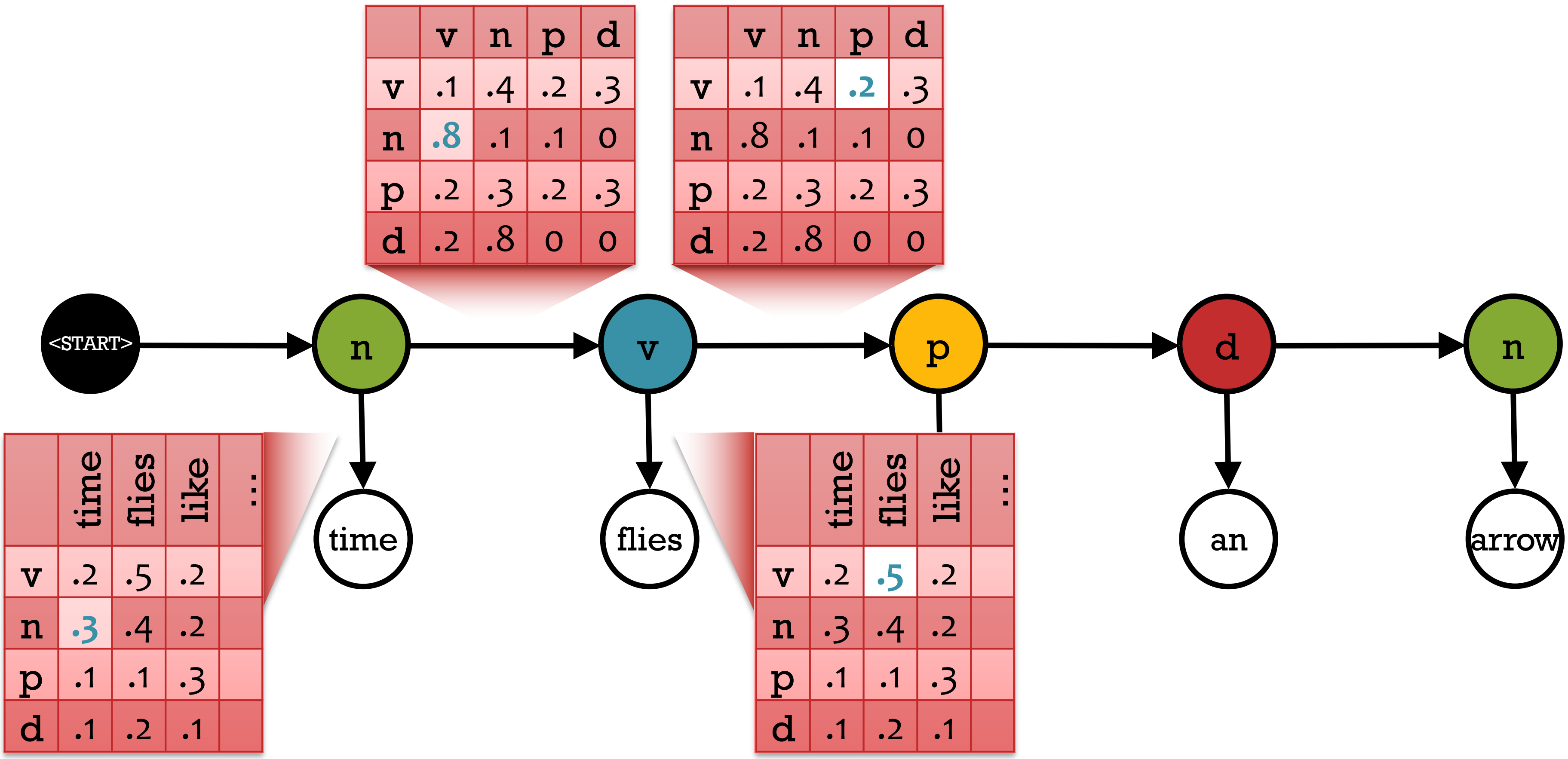
	time	flies	like	…
v	3	5	3	
n	4	5	2	
p	0.1	0.1	3	
d	0.1	0.2	0.1	

	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 $Z = 1$.

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



	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

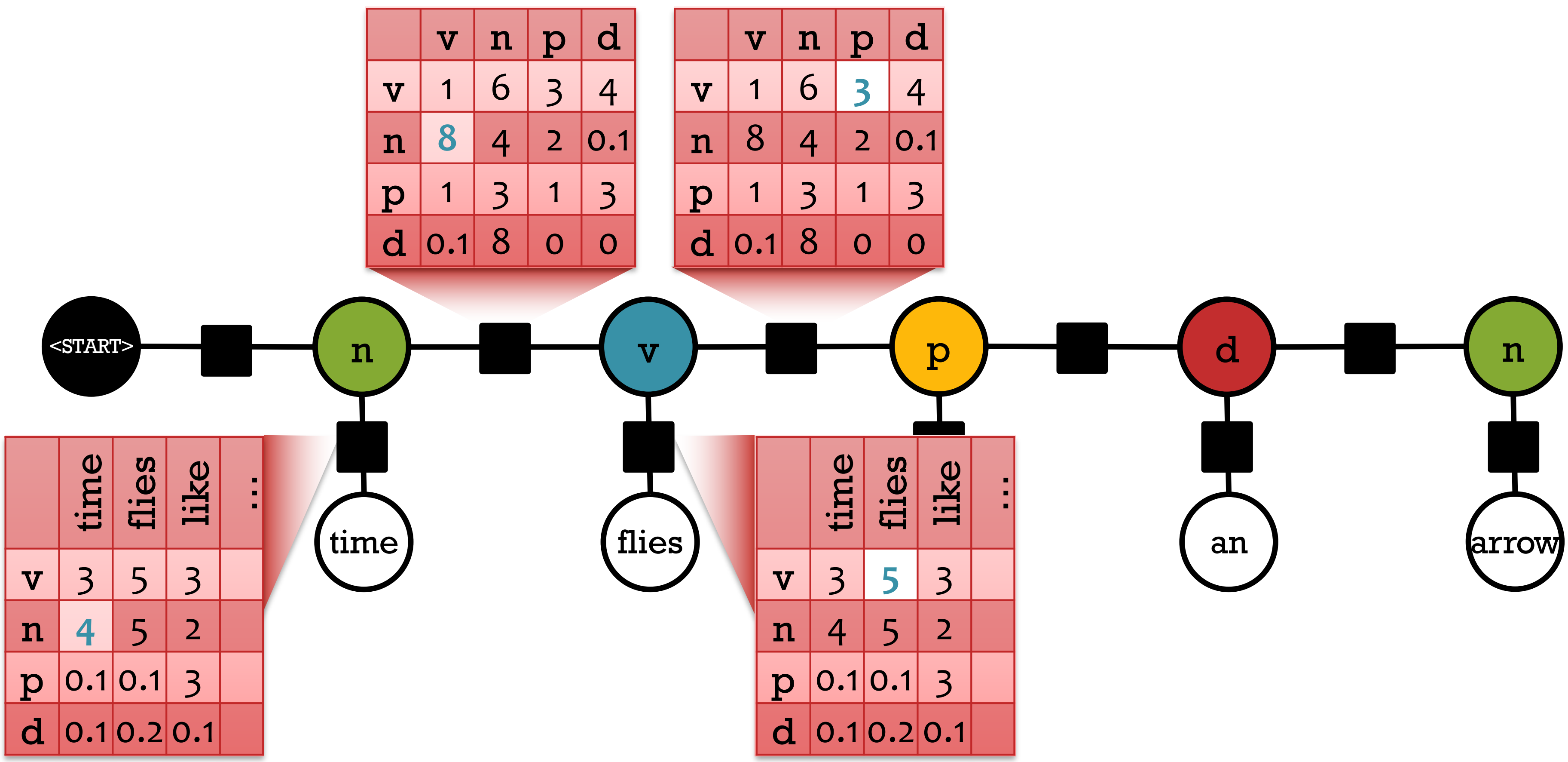
	time	flies	like	…
v	.2	.5	.2	
n	.3	.4	.2	
p	.1	.1	.3	
d	.1	.2	.1	

	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

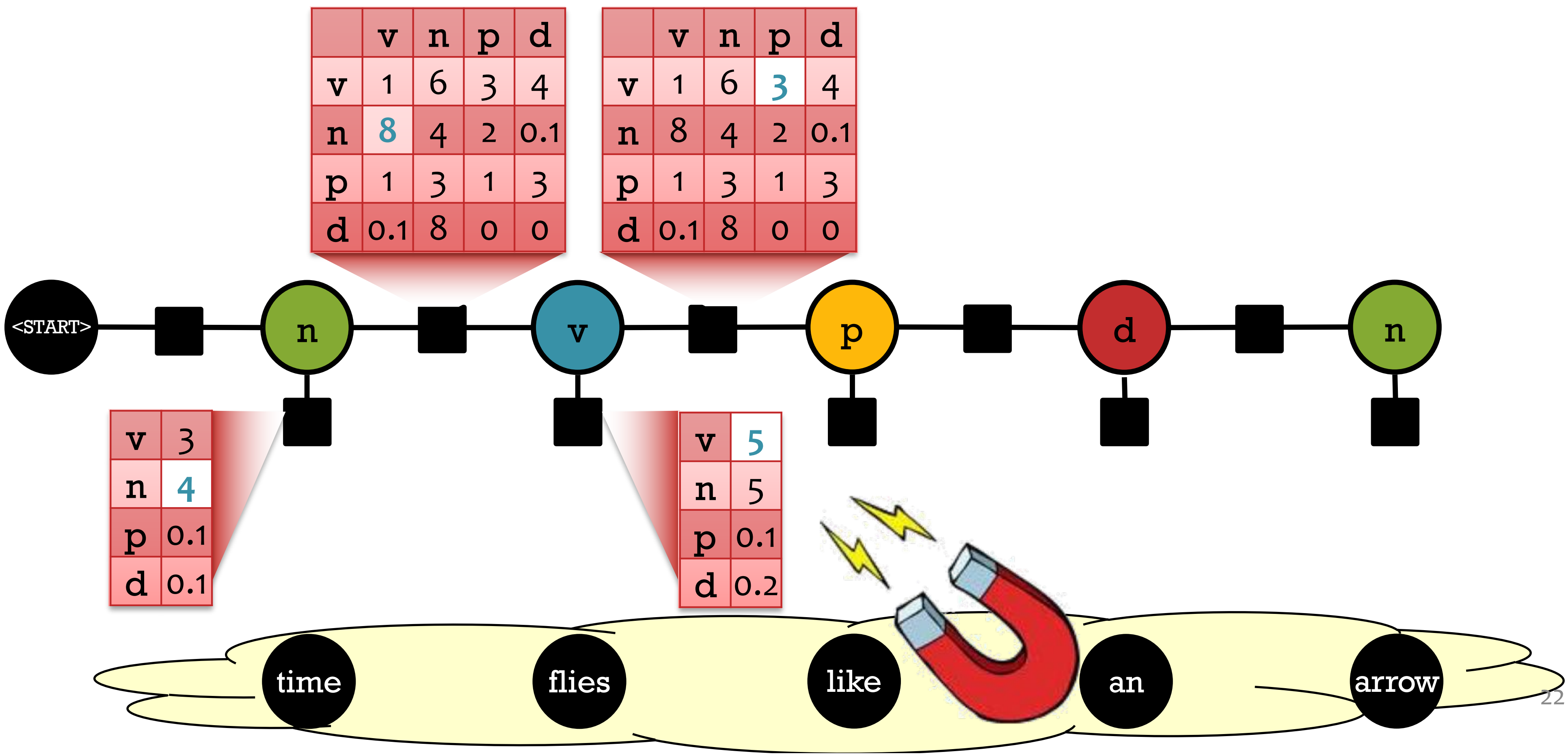
	time	flies	like	…
v	3	5	3	
n	4	5	2	
p	0.1	0.1	3	
d	0.1	0.2	0.1	

	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:**

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

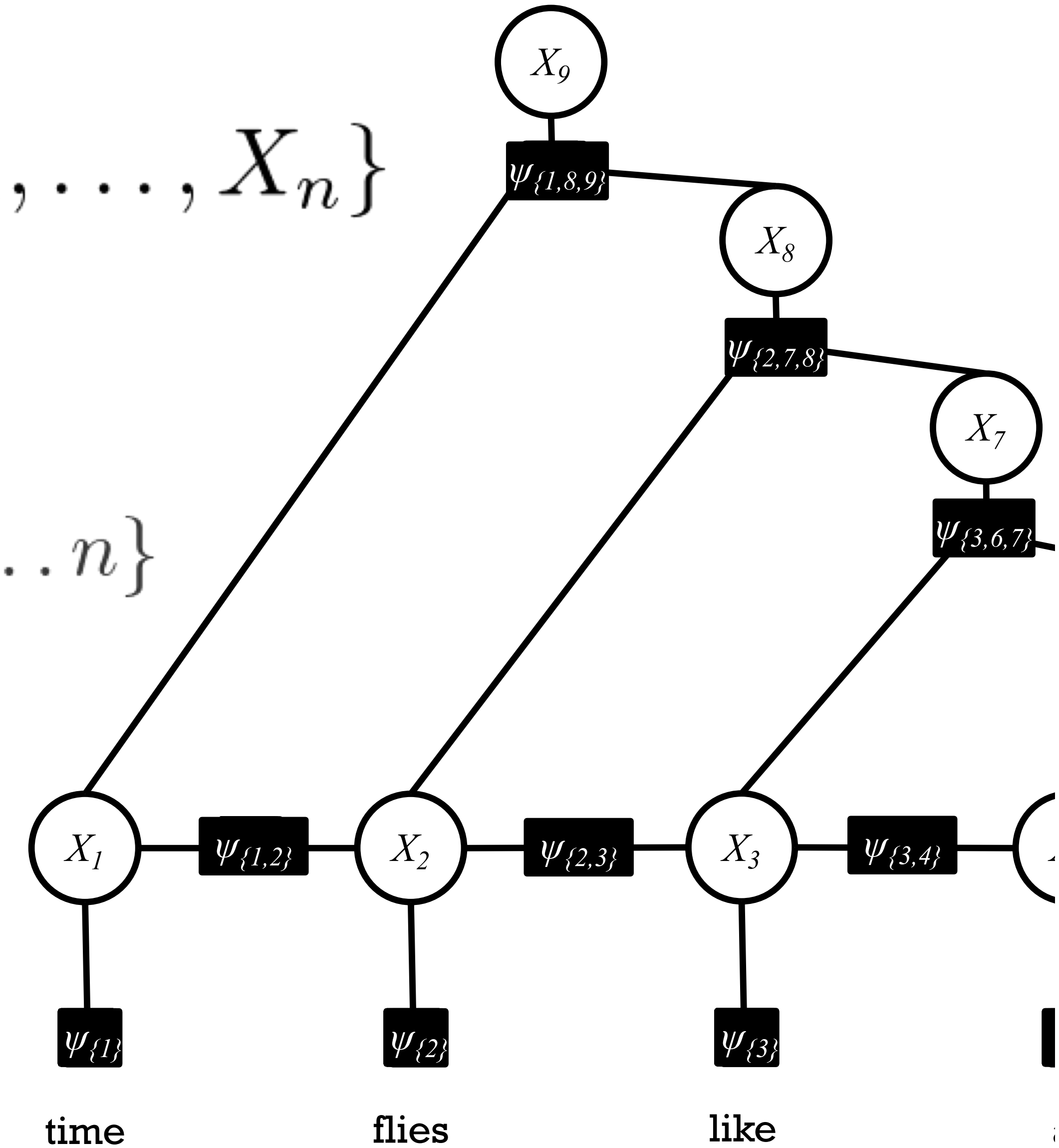
- **Factors:**

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

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

Joint Distribution

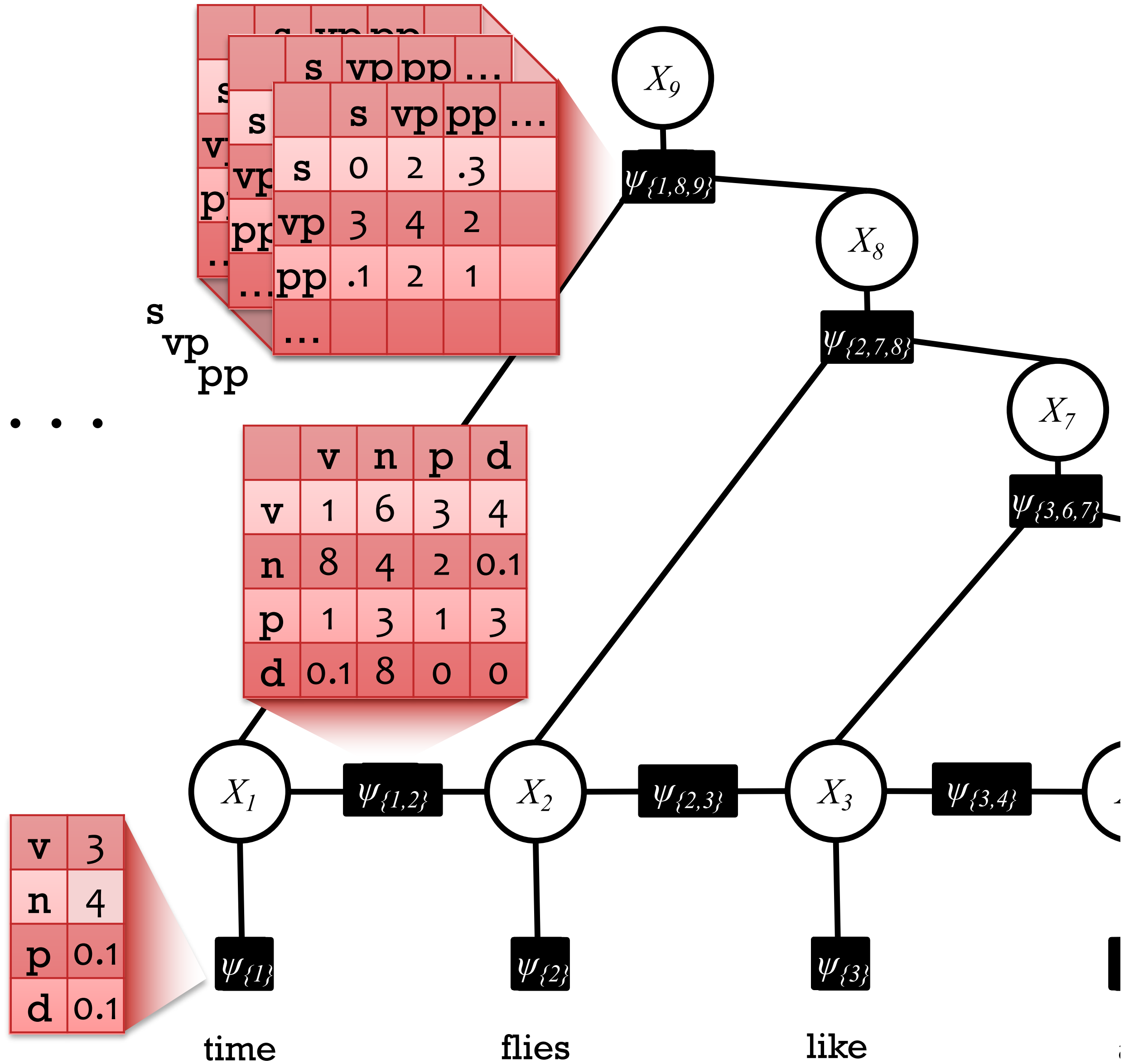
$$p(\boldsymbol{x}) = \frac{1}{Z} \prod_\alpha \psi_\alpha(\boldsymbol{x}_\alpha)$$



Factors are Tensors

- Factors:

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



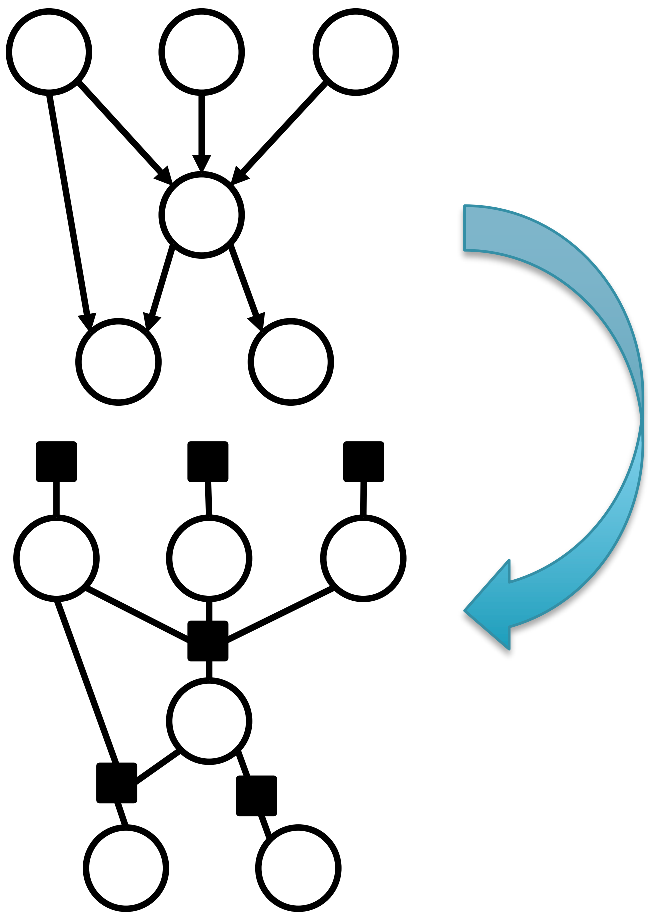
	s	vp	pp	…
s	0	2	.3	
vp	3	4	2	
pp	.1	2	1	
…				

	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

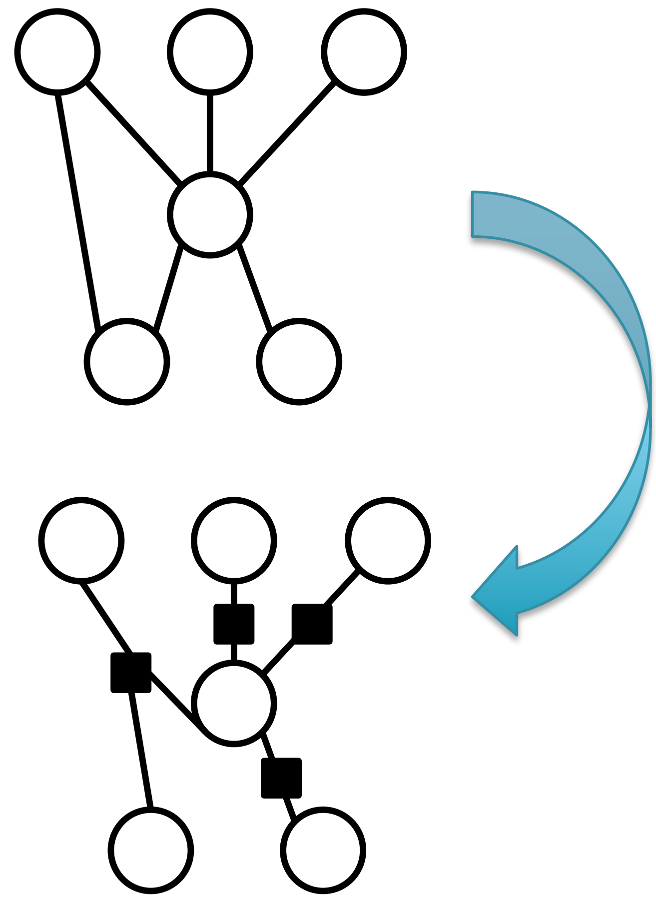
v	3
n	4
p	0.1
d	0.1

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:

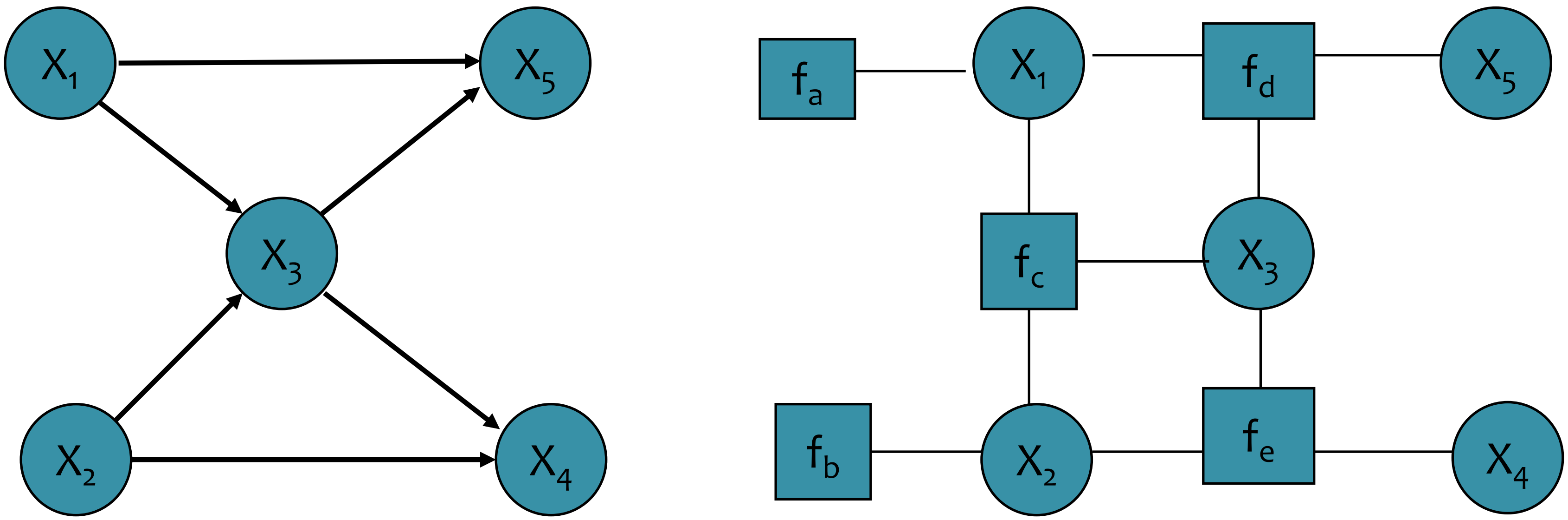
$$p(x) = \frac{1}{Z}\left(\prod_{i\in V}\psi(x_i)\prod_{(i,j)\in E}\psi(x_i,x_j)\right)$$

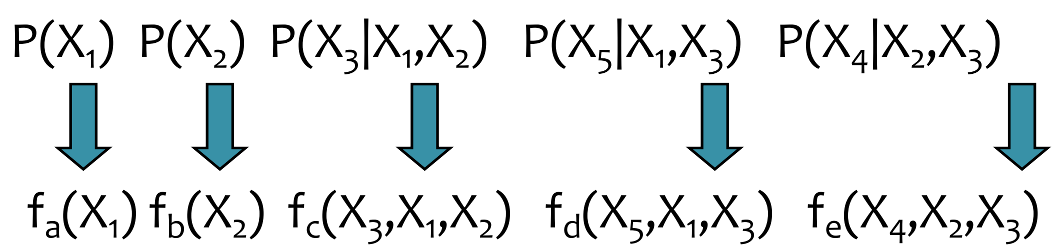
$$p(x) = p(x_r)\prod_{(i,j)\in E}p(x_j|x_i)$$

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

Factor Graph Examples

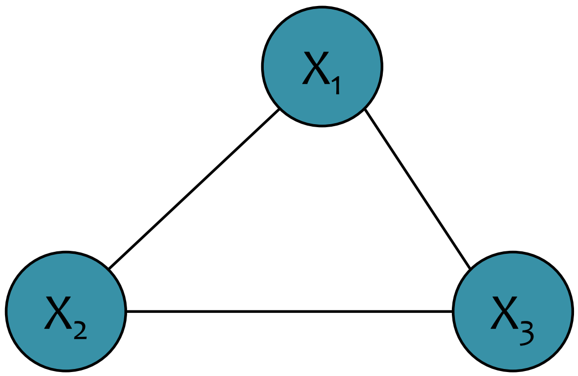
- Example 1



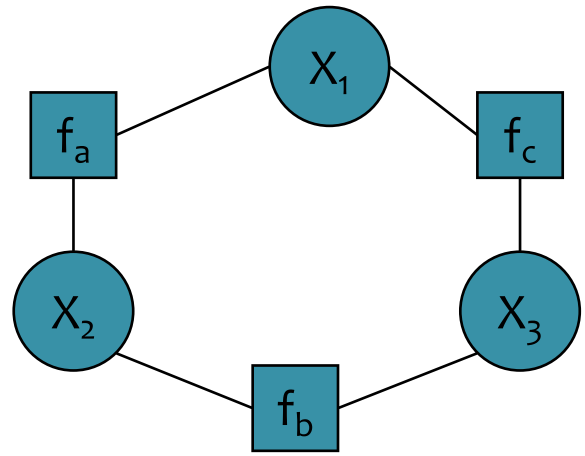


Factor Graph Examples

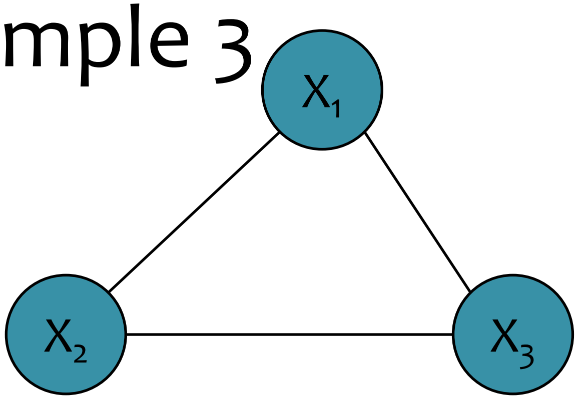
- Example 2



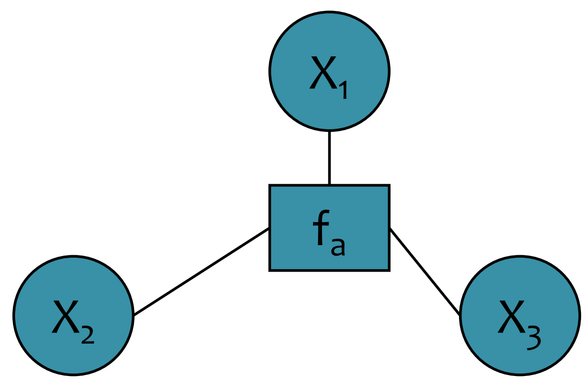
$$\psi(x_1,x_2,x_3) = f_a(x_1,x_2)f_b(x_2,x_3)f_c(x_3,x_1)$$



- Example 3

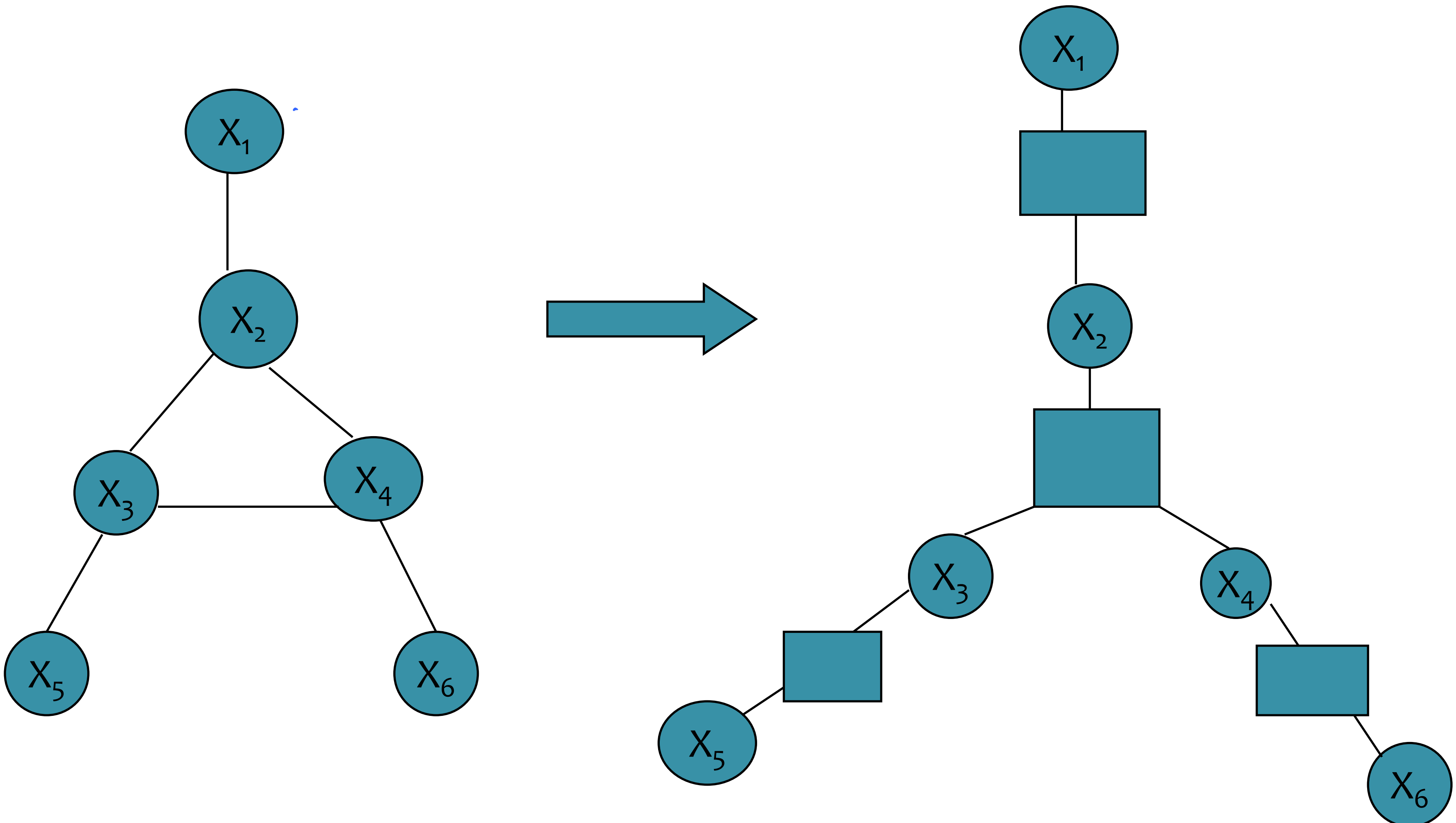


$$\psi(x_1,x_2,x_3) = f_a(x_1,x_2,x_3)$$



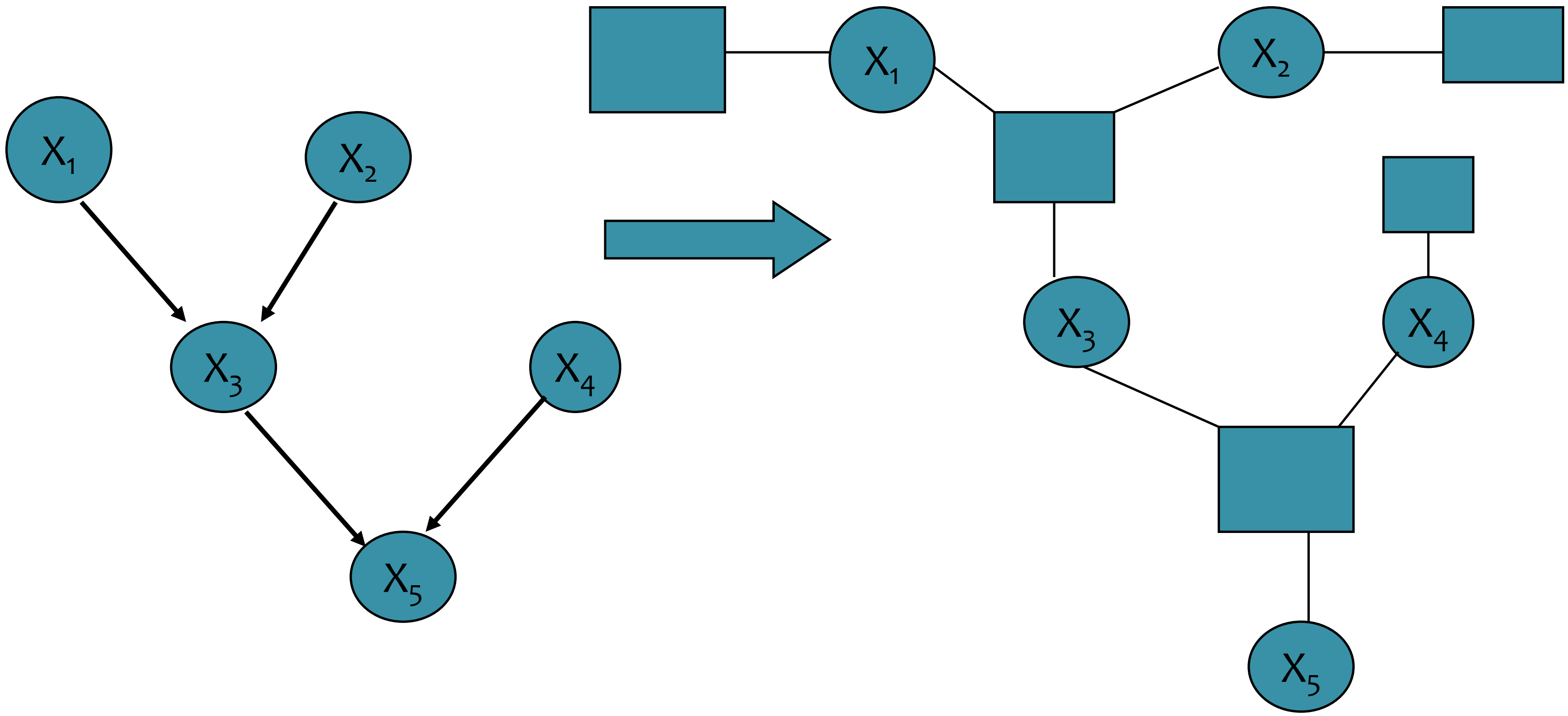
Tree-like Undirected GMs to Factor Trees

- Example 4

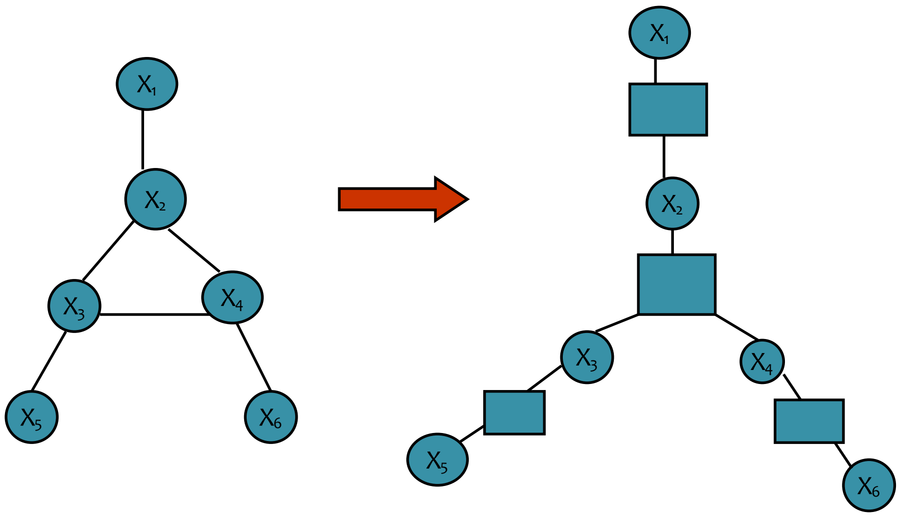


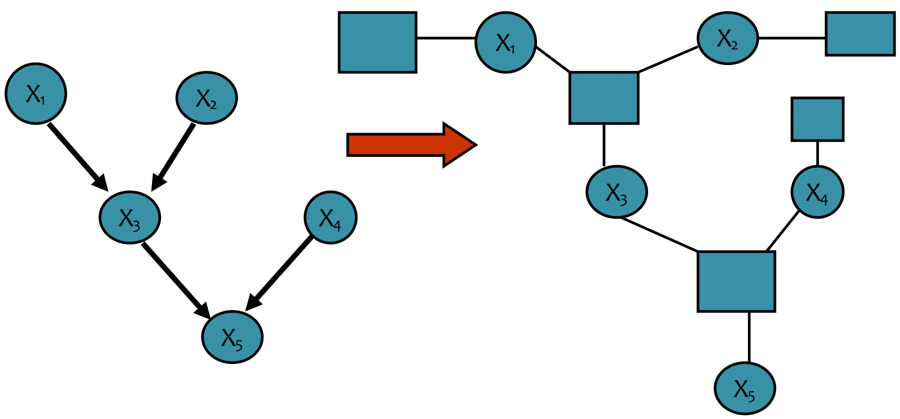
Poly-trees to Factor trees

- Example 5



Why factor graphs?





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

EXACT INFERENCE

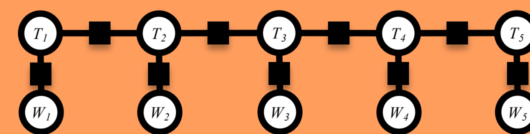
Exact Inference

1. Data

$$\mathcal{D} = \{\boldsymbol{x}^{(n)}\}_{n=1}^{N}$$

2. Model

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \prod_{C \in \mathcal{C}} \psi_C(\boldsymbol{x}_C)$$



3. Objective

$$\ell(\theta; \mathcal{D}) = \sum_{n=1}^{N} \log p(\boldsymbol{x}^{(n)} \mid \boldsymbol{\theta})$$

5. Inference

1. Marginal Inference

$$p(\boldsymbol{x}_C) = \sum_{\boldsymbol{x}' : \boldsymbol{x}'_C = \boldsymbol{x}_C} p(\boldsymbol{x}' \mid \boldsymbol{\theta})$$

2. Partition Function

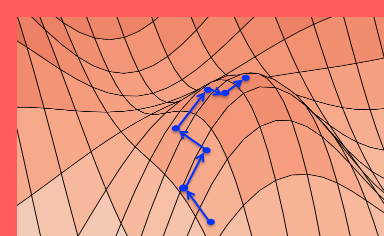
$$Z(\boldsymbol{\theta}) = \sum_{\boldsymbol{x}} \prod_{C \in \mathcal{C}} \psi_C(\boldsymbol{x}_C)$$

3. MAP Inference

$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{x}}{\operatorname{argmax}} \; p(\boldsymbol{x} \mid \boldsymbol{\theta})$$

4. Learning

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \, \ell(\boldsymbol{\theta}; \mathcal{D})$$



5. Inference

Three Tasks:

1. Marginal Inference (#P-Hard)
Compute marginals of variables and cliques

$$p(x_i) = \sum_{\boldsymbol{x}' : x'_i = x_i} p(\boldsymbol{x}' \mid \boldsymbol{\theta}) \qquad \Big| \qquad p(\boldsymbol{x}_C) = \sum_{\boldsymbol{x}' : \boldsymbol{x}'_C = \boldsymbol{x}_C} p(\boldsymbol{x}' \mid \boldsymbol{\theta})$$

2. Partition Function (#P-Hard)
Compute the normalization constant

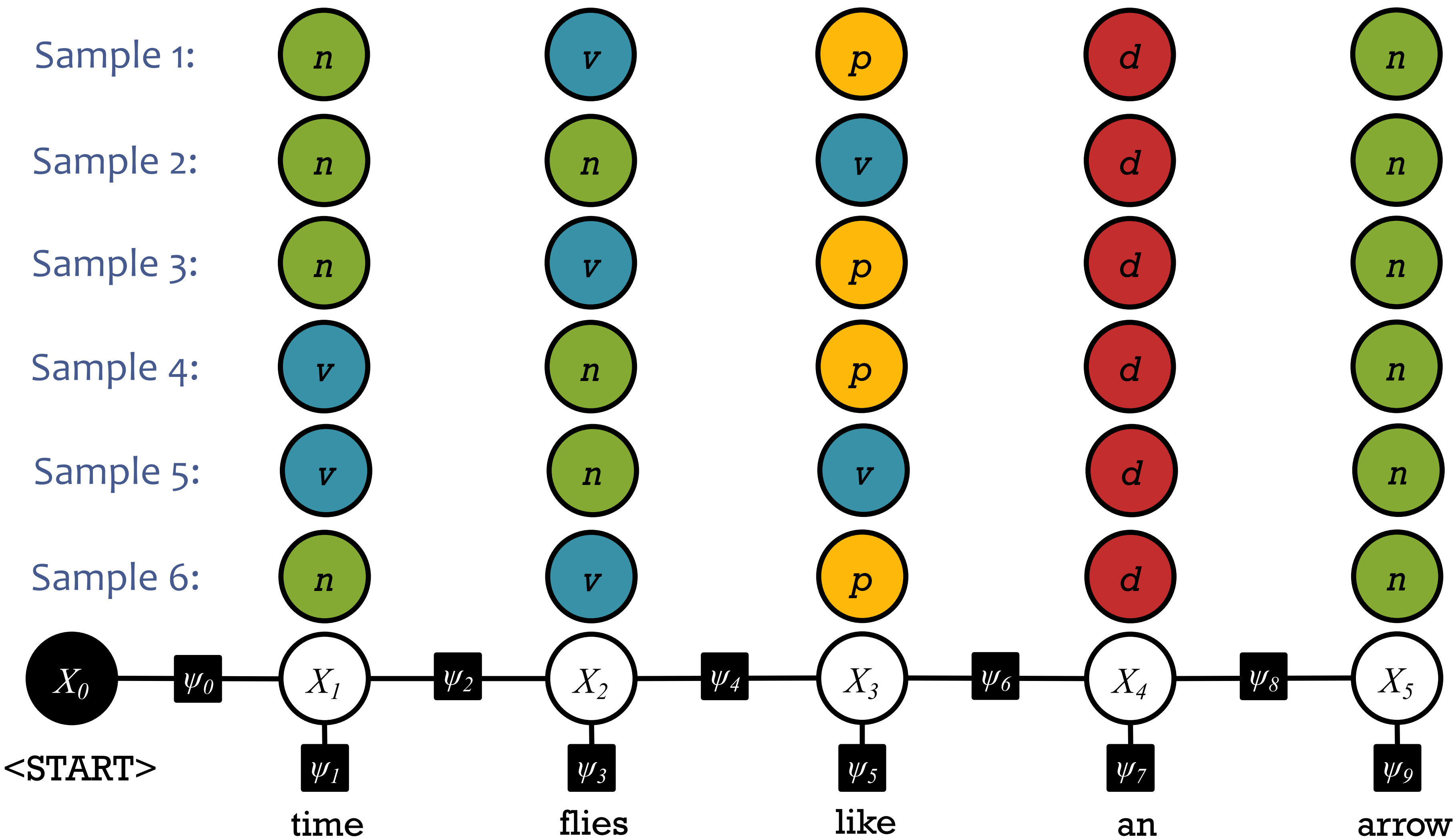
$$Z(\boldsymbol{\theta}) = \sum_{\boldsymbol{x}} \prod_{C \in \mathcal{C}} \psi_C(\boldsymbol{x}_C)$$

3. MAP Inference (NP-Hard)
Compute variable assignment with highest probability

$$\hat{\boldsymbol{x}} = \operatorname*{argmax}_{\boldsymbol{x}} \; p(\boldsymbol{x} \mid \boldsymbol{\theta})$$

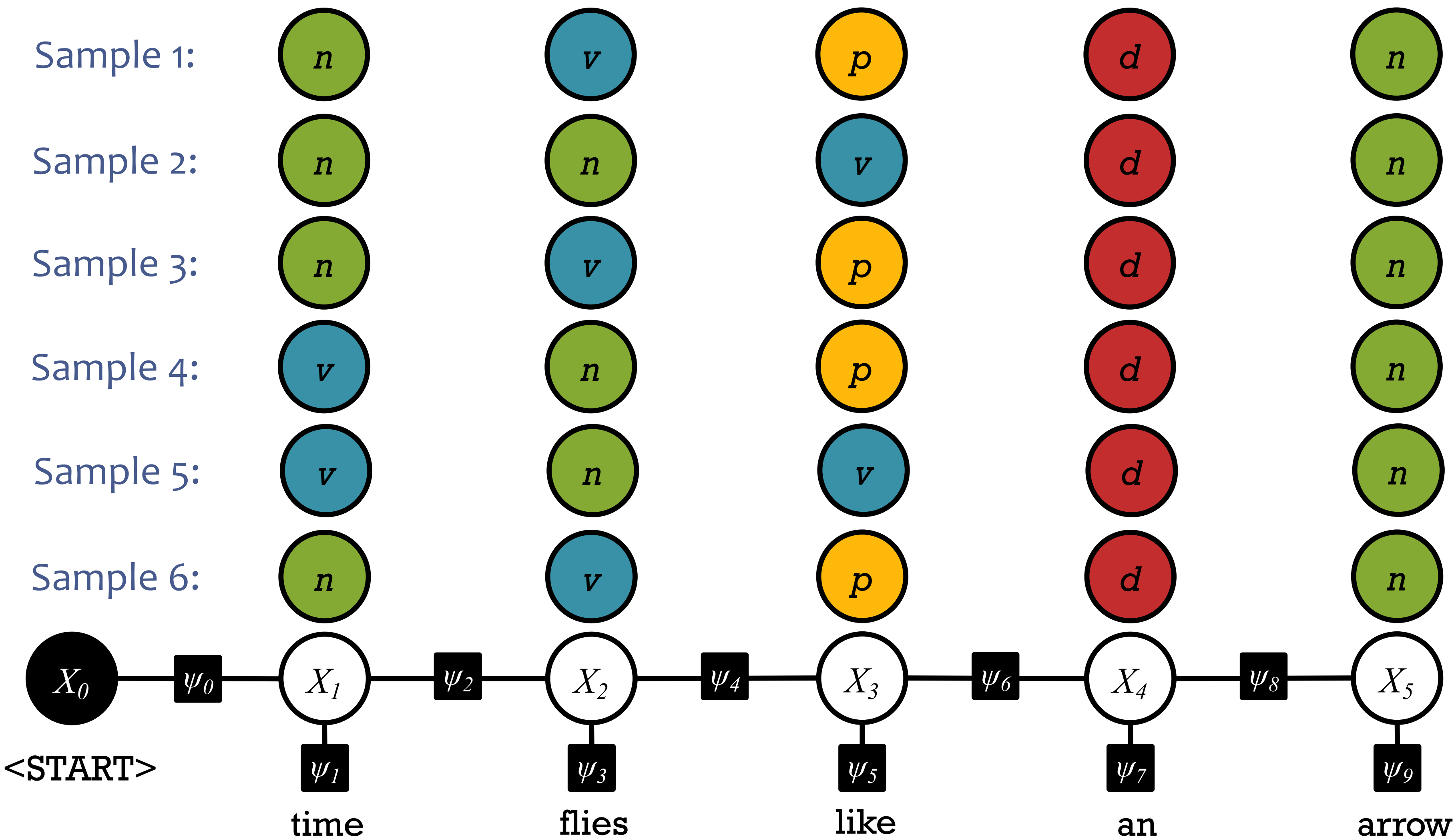
Marginals by Sampling on Factor Graph

Suppose we took many samples from the distribution over taggings: $p(\boldsymbol{x}) = \frac{1}{Z}\prod_{\alpha}\psi_\alpha(\boldsymbol{x_\alpha})$

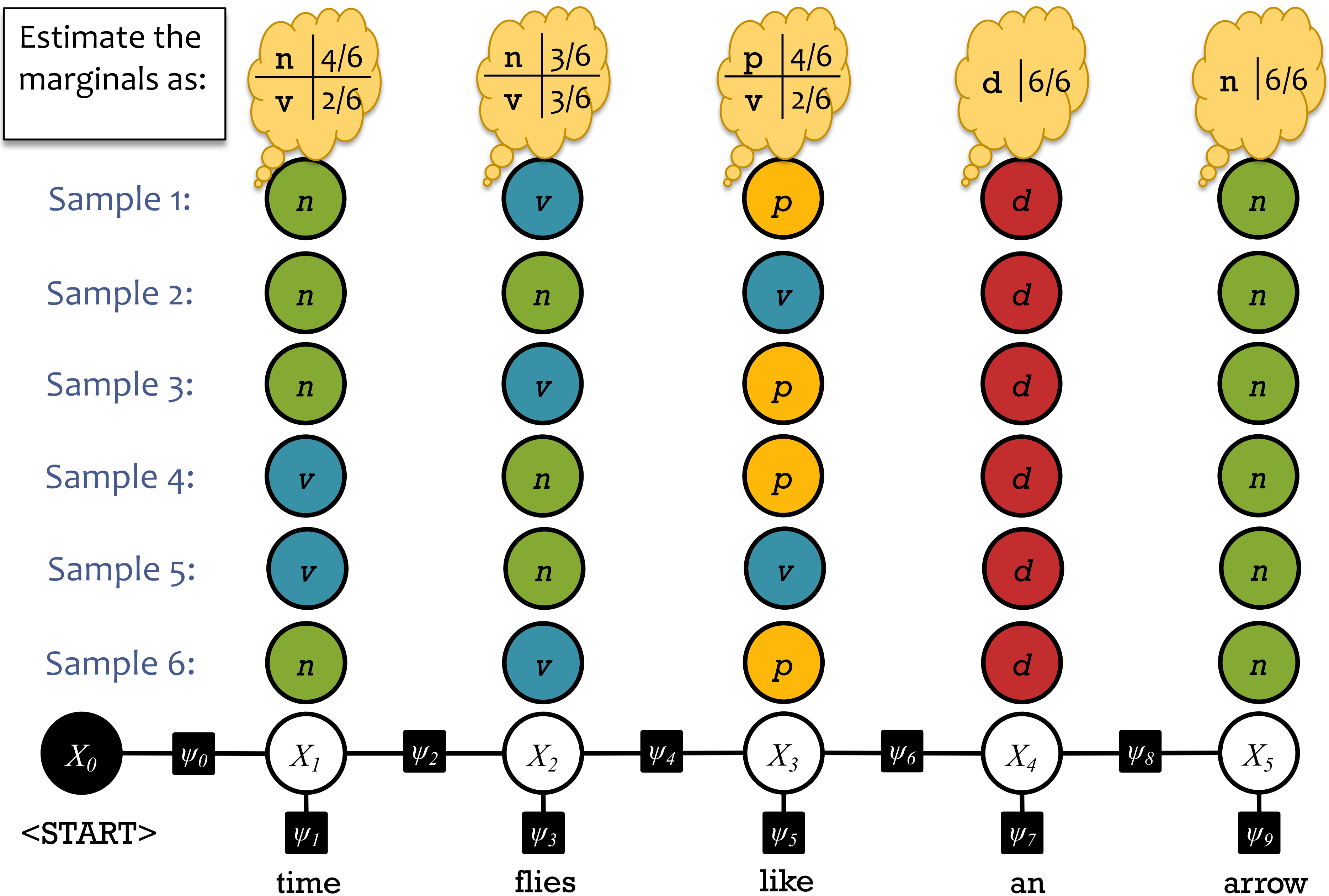


Marginals by Sampling on Factor Graph

The marginal $p(X_i = x_i)$ gives the probability that variable X_i takes value x_i in a random sample



Marginals by Sampling on Factor Graph



Simple and general exact inference for graphical models

VARIABLE ELIMINATION

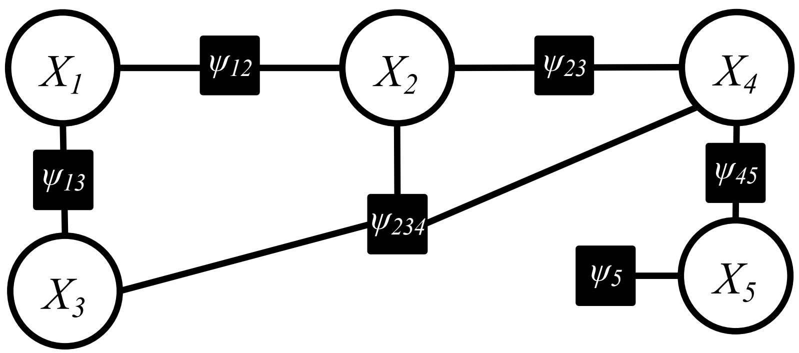
Brute Force (Naïve) Inference

For all i, suppose the **range** of X_i is $\{0, 1, 2\}$.

Let $k=3$ denote the **size of the range.**

The distribution **factorizes** as:

$$s(\boldsymbol{x}) = \psi_{12}(x_1, x_2)\psi_{13}(x_1, x_3)\psi_{24}(x_2, x_4)$$
$$\psi_{234}(x_2, x_3, x_4)\psi_{45}(x_4, x_5)\psi_5(x_5)$$



Naively, we compute the **partition function** as:

$$Z = \sum_{x_1}\sum_{x_2}\sum_{x_3}\sum_{x_4}\sum_{x_5} s(\boldsymbol{x})$$

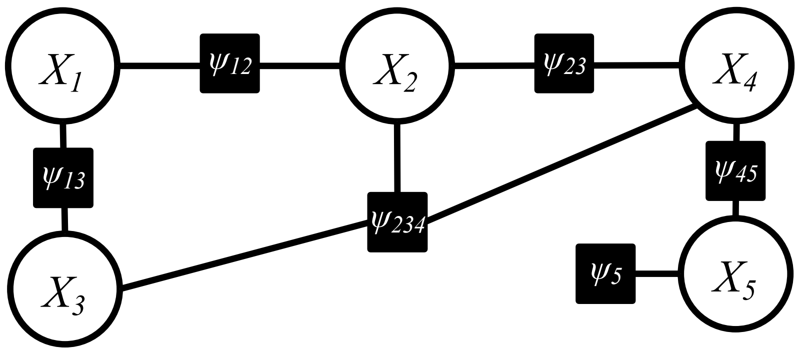
Brute Force (Naïve) Inference

For all i, suppose the **range** of X_i is $\{0, 1, 2\}$.

Let $k=3$ denote the **size of the range.**

The distribution **factorizes** as:

$$s(\boldsymbol{x}) = \psi_{12}(x_1, x_2)\psi_{13}(x_1, x_3)\psi_{24}(x_2, x_4)\psi_{234}(x_2, x_3, x_4)\psi_{45}(x_4, x_5)\psi_5(x_5)$$



Naively, we compute the **partition function** as:

$$Z = \sum_{x_1}\sum_{x_2}\sum_{x_3}\sum_{x_4}\sum_{x_5} s(\boldsymbol{x})$$

$s(\boldsymbol{x})$ can be represented as a joint probability table with 3^5 entries:

x_1	x_2	x_3	x_4	x_5	$s(\boldsymbol{x})$
0	0	0	0	0	0.019517693
0	0	0	0	1	0.017090249
0	0	0	0	2	0.014885825
0	0	0	1	0	0.024117638
0	0	0	1	1	0.000925849
0	0	0	1	2	0.028112576
0	0	0	2	0	0.028050205
0	0	0	2	1	0.004812689
0	0	0	2	2	0.007987737
0	0	1	0	0	0.028433687
0	0	1	0	1	0.037073469
0	0	1	0	2	0.013558227
0	0	1	1	0	0.019479016
0	0	1	1	1	0.012312901
0	0	1	1	2	0.023439775
0	0	1	2	0	0.038206131
0	0	1	2	1	0.038996005
0	0	1	2	2	0.041458783
0	0	2	0	0	0.044616806
0	0	2	0	1	0.020846989
0	0	2	0	2	0.03006475
0	0	2	1	0	0.048436964
0	0	2	1	1	0.02854376
0	0	2	1	2	0.029191506
0	0	2	2	0	0.031531118
0	0	2	2	1	0.005132392
0	0	2	2	2	0.032027091
…	…	…	…	…	…

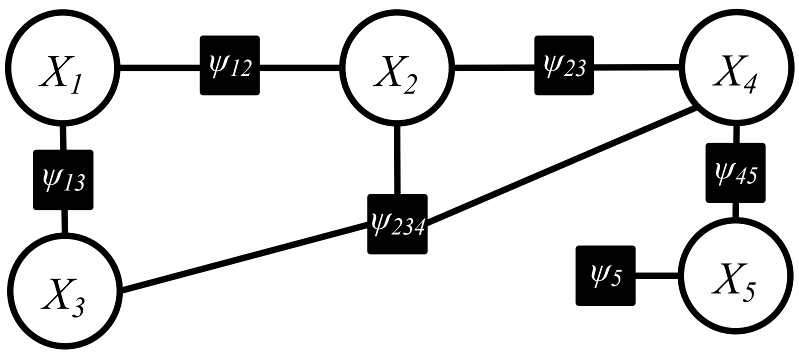
Brute Force (Naïve) Inference

For all i, suppose the **range** of X_i is $\{0, 1, 2\}$.

Let $k=3$ denote the **size of the range.**

The distribution **factorizes** as:

$$s(\boldsymbol{x}) = \psi_{12}(x_1, x_2)\psi_{13}(x_1, x_3)\psi_{24}(x_2, x_4)\psi_{234}(x_2, x_3, x_4)\psi_{45}(x_4, x_5)\psi_5(x_5)$$



Naively, we compute the **partition function** as:

$$Z = \sum_{x_1}\sum_{x_2}\sum_{x_3}\sum_{x_4}\sum_{x_5} s(\boldsymbol{x})$$

$s(\boldsymbol{x})$ can be represented as a joint probability table with 3^5 entries:

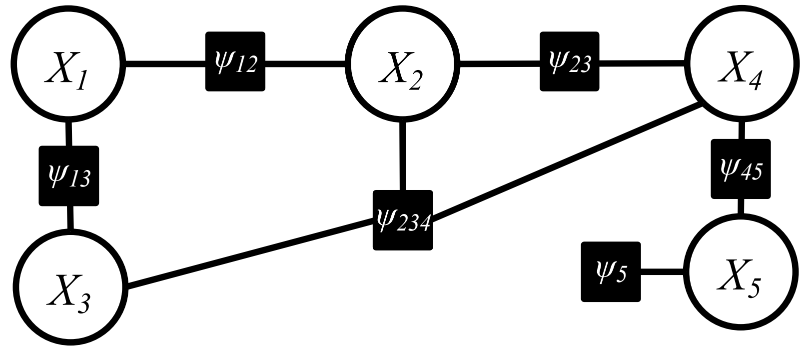
x_1	x_2	x_3	x_4	x_5	$s(\boldsymbol{x})$
0	0	0	0	0	0.019517693
0	0	0	0	1	0.017090249
0	0	0	0	2	0.014885825
0	0	0	1	0	0.024117638
0	0	0	1	1	0.000925849
0	0	0	1	2	0.028112576
0	0	0	2	0	0.028050205
0	0	0	2	1	0.004812689
0	0	0	2	2	0.007987737
0	0	1	0	0	0.028433687
0	0	1	0	1	0.037073469
0	0	1	0	2	0.013558227
0	0	1	1	0	0.019479016
0	0	1	1	1	0.012312901
0	0	1	1	2	0.023439775
0	0	1	2	0	0.038206131
0	0	1	2	1	0.038996005
0	0	1	2	2	0.041458783
0	0	2	0	0	0.044616806
0	0	2	0	1	0.020846989
0	0	2	0	2	0.03006475
0	0	2	1	0	0.048436964
0	0	2	1	1	0.02854376

Naïve computation of Z requires 3^5 additions.

Can we do better?

The Variable Elimination Algorithm

Instead, capitalize on the factorization of $s(\boldsymbol{x})$.



$$Z = \sum_{x_1}\sum_{x_2}\sum_{x_3}\sum_{x_4}\sum_{x_5} \psi_{12}(x_1,x_2)\psi_{13}(x_1,x_3)\psi_{24}(x_2,x_4)\psi_{234}(x_2,x_3,x_4)\psi_{45}(x_4,x_5)\psi_5(x_5)$$

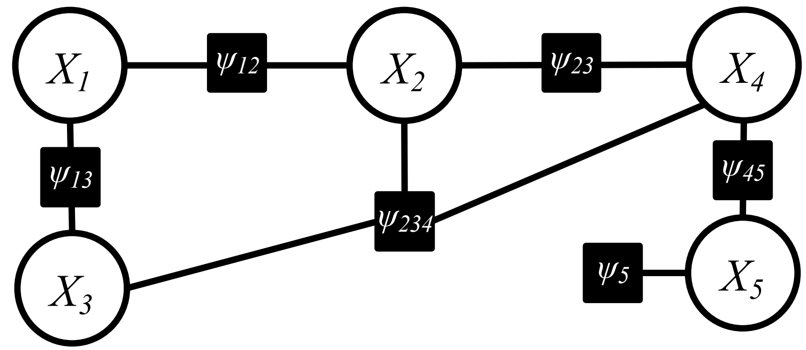
$$= \sum_{x_1}\sum_{x_2}\sum_{x_3}\sum_{x_4} \psi_{12}(x_1,x_2)\psi_{13}(x_1,x_3)\psi_{24}(x_2,x_4)\psi_{234}(x_2,x_3,x_4)\sum_{x_5}\psi_{45}(x_4,x_5)\psi_5(x_5)$$

This "factor" is a much smaller table with 3^2 entries:

x_4	x_5	$s(x_4, x_5)$
0	0	0.019517693
0	1	0.017090249
0	2	0.014885825
1	0	0.024117638
1	1	0.000925849
1	2	0.028112576
2	0	0.028050205
2	1	0.004812689
2	2	0.007987737

The Variable Elimination Algorithm

Instead, capitalize on the factorization of $s(\boldsymbol{x})$.



$$Z = \sum_{x_1}\sum_{x_2}\sum_{x_3}\sum_{x_4}\sum_{x_5} \psi_{12}(x_1,x_2)\psi_{13}(x_1,x_3)\psi_{24}(x_2,x_4)\psi_{234}(x_2,x_3,x_4)\psi_{45}(x_4,x_5)\psi_5(x_5)$$

$$= \sum_{x_1}\sum_{x_2}\sum_{x_3}\sum_{x_4} \psi_{12}(x_1,x_2)\psi_{13}(x_1,x_3)\psi_{24}(x_2,x_4)\psi_{234}(x_2,x_3,x_4) \sum_{x_5} \psi_{45}(x_4,x_5)\psi_5(x_5)$$

Only 3^2 additions are needed to marginalize out x_5.

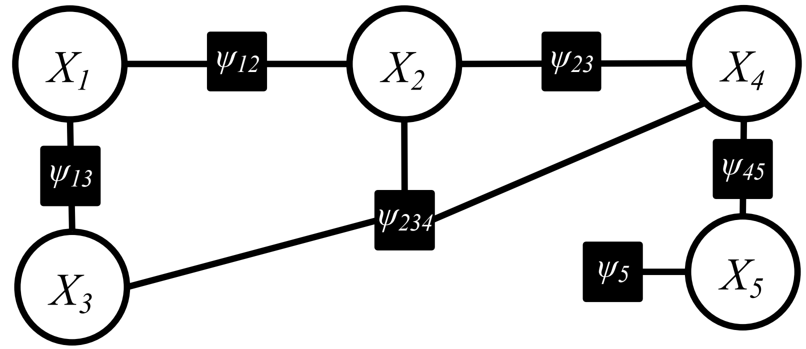
We denote the **marginal's table** by $m_5(x_4)$.

This "factor" is a much smaller table with 3 entries:

x_4	$m_5(x_4)$
0	0.019517693
1	0.017090249
2	0.014885825

The Variable Elimination Algorithm

Instead, capitalize on the factorization of *s(**x**)*.



$$Z = \sum_{x_1} \sum_{x_2} \sum_{x_3} \sum_{x_4} \sum_{x_5} \psi_{12}(x_1, x_2)\psi_{13}(x_1, x_3)\psi_{24}(x_2, x_4)\psi_{234}(x_2, x_3, x_4)\psi_{45}(x_4, x_5)\psi_5(x_5)$$

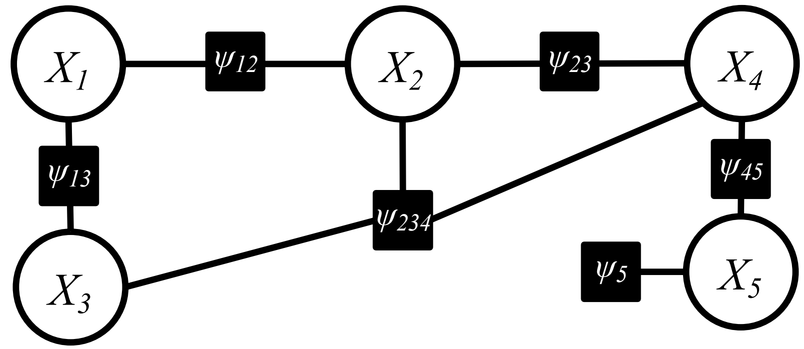
$$= \sum_{x_1} \sum_{x_2} \sum_{x_3} \sum_{x_4} \psi_{12}(x_1, x_2)\psi_{13}(x_1, x_3)\psi_{24}(x_2, x_4)\psi_{234}(x_2, x_3, x_4) \sum_{x_5} \psi_{45}(x_4, x_5)\psi_5(x_5)$$

$$= \sum_{x_1} \sum_{x_2} \sum_{x_3} \sum_{x_4} \psi_{12}(x_1, x_2)\psi_{13}(x_1, x_3)\psi_{24}(x_2, x_4)\psi_{234}(x_2, x_3, x_4) m_5(x_4)$$

$$m_5(x_4) \triangleq \sum_{x_5} \psi_{45}(x_4, x_5)\psi_5(x_5)$$

The Variable Elimination Algorithm

Instead, capitalize on the factorization of $s(\boldsymbol{x})$.



$$Z = \sum_{x_1} \sum_{x_2} \sum_{x_3} \sum_{x_4} \sum_{x_5} \psi_{12}(x_1, x_2)\psi_{13}(x_1, x_3)\psi_{24}(x_2, x_4)\psi_{234}(x_2, x_3, x_4)\psi_{45}(x_4, x_5)\psi_5(x_5)$$

$$= \sum_{x_1} \sum_{x_2} \sum_{x_3} \sum_{x_4} \psi_{12}(x_1, x_2)\psi_{13}(x_1, x_3)\psi_{24}(x_2, x_4)\psi_{234}(x_2, x_3, x_4) \sum_{x_5} \psi_{45}(x_4, x_5)\psi_5(x_5)$$

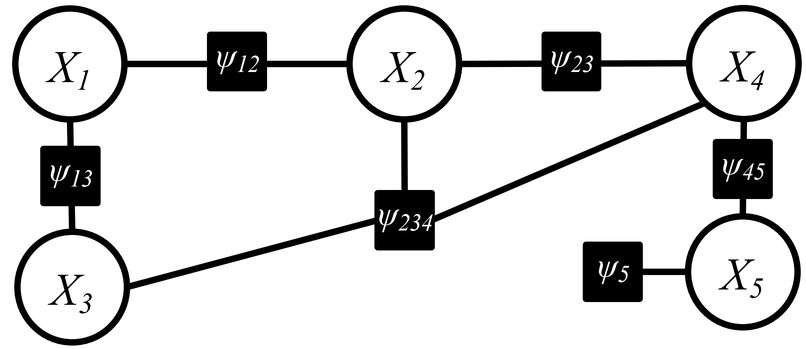
$$= \sum_{x_1} \sum_{x_2} \sum_{x_3} \sum_{x_4} \psi_{12}(x_1, x_2)\psi_{13}(x_1, x_3)\psi_{24}(x_2, x_4)\psi_{234}(x_2, x_3, x_4) m_5(x_4)$$

This "factor" is still a 3^4 table so apply the same trick again.

$$m_5(x_4) \triangleq \sum_{x_5} \psi_{45}(x_4, x_5)\psi_5(x_5)$$

The Variable Elimination Algorithm

Instead, capitalize on the factorization of $s(\boldsymbol{x})$.



$$Z = \sum_{x_1}\sum_{x_2}\psi_{12}(x_1,x_2)\sum_{x_3}\psi_{13}(x_1,x_3)\sum_{x_4}\psi_{24}(x_2,x_4)\psi_{234}(x_2,x_3,x_4)\sum_{x_5}\psi_{45}(x_4,x_5)\psi_5(x_5)$$

$$= \sum_{x_1}\sum_{x_2}\psi_{12}(x_1,x_2)\sum_{x_3}\psi_{13}(x_1,x_3)\sum_{x_4}\psi_{24}(x_2,x_4)\psi_{234}(x_2,x_3,x_4)m_5(x_4)$$

3^2 additions

$$= \sum_{x_1}\sum_{x_2}\psi_{12}(x_1,x_2)\sum_{x_3}\psi_{13}(x_1,x_3)m_4(x_2,x_3)$$

3^3 additions

$$= \sum_{x_1}\sum_{x_2}\psi_{12}(x_1,x_2)m_3(x_1,x_2)$$

3^3 additions

$$= \sum_{x_1} m_2(x_1)$$

3^2 additions

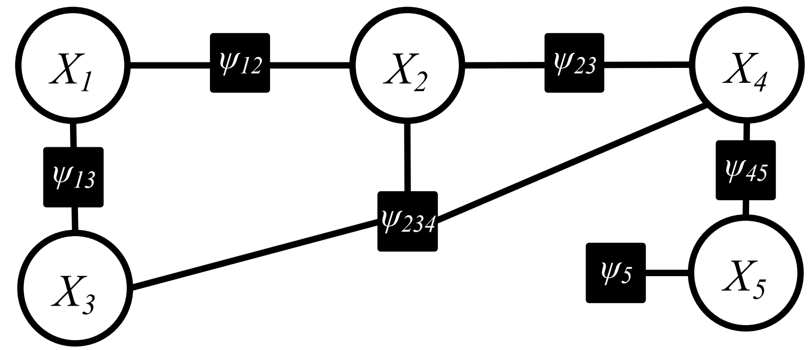
3 additions

Naïve solution requires $3^5 = 243$ additions.

Variable elimination only requires $3 + 3^2 + 3^3 + 3^3 + 3^2 = 75$ additions.

The Variable Elimination Algorithm

The same trick can be used to compute **marginal probabilities.** Just choose the variable elimination order such that the query variables are last.



$$p(x_1) = \frac{1}{Z} \sum_{x_2} \psi_{12}(x_1, x_2) \sum_{x_3} \psi_{13}(x_1, x_3) \sum_{x_4} \psi_{24}(x_2, x_4) \psi_{234}(x_2, x_3, x_4) \sum_{x_5} \psi_{45}(x_4, x_5) \psi_5(x_5)$$

$$= \frac{1}{Z} \sum_{x_2} \psi_{12}(x_1, x_2) \sum_{x_3} \psi_{13}(x_1, x_3) \sum_{x_4} \psi_{24}(x_2, x_4) \psi_{234}(x_2, x_3, x_4) m_5(x_4)$$

3^2 additions

$$= \frac{1}{Z} \sum_{x_2} \psi_{12}(x_1, x_2) \sum_{x_3} \psi_{13}(x_1, x_3) m_4(x_2, x_3)$$

3^3 additions

$$= \frac{1}{Z} \sum_{x_2} \psi_{12}(x_1, x_2) m_3(x_1, x_2)$$

3^3 additions

$$= \frac{1}{Z} m_2(x_1)$$

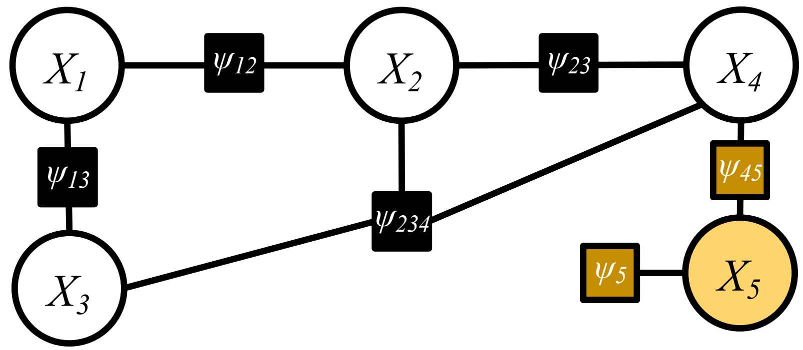
3^2 additions

3 different values on LHS

For directed graphs, Z = 1.

For undirected graphs, if we compute each (unnormalized) value on the LHS, we can sum them to get Z.

The Variable Elimination Algorithm



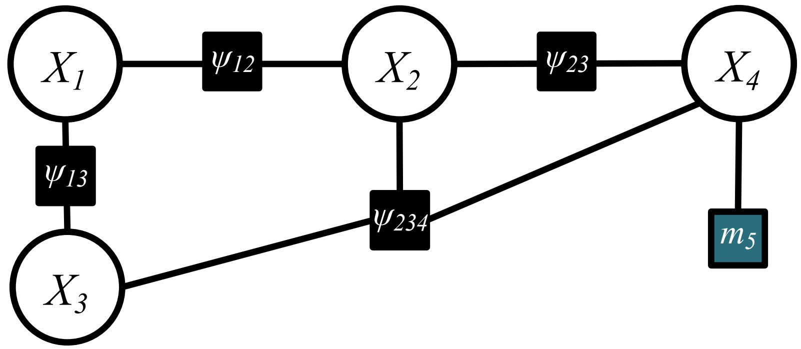
$$Z = \sum_{x_1} \sum_{x_2} \psi_{12}(x_1, x_2) \sum_{x_3} \psi_{13}(x_1, x_3) \sum_{x_4} \psi_{24}(x_2, x_4) \psi_{234}(x_2, x_3, x_4) \sum_{x_5} \psi_{45}(x_4, x_5) \psi_5(x_5)$$

$$= \sum_{x_1} \sum_{x_2} \psi_{12}(x_1, x_2) \sum_{x_3} \psi_{13}(x_1, x_3) \sum_{x_4} \psi_{24}(x_2, x_4) \psi_{234}(x_2, x_3, x_4) m_5(x_4)$$

$$= \sum_{x_1} \sum_{x_2} \psi_{12}(x_1, x_2) \sum_{x_3} \psi_{13}(x_1, x_3) m_4(x_2, x_3)$$

In a factor graph, variable **elimination** corresponds to replacement of a subgraph with a factor.

The Variable Elimination Algorithm



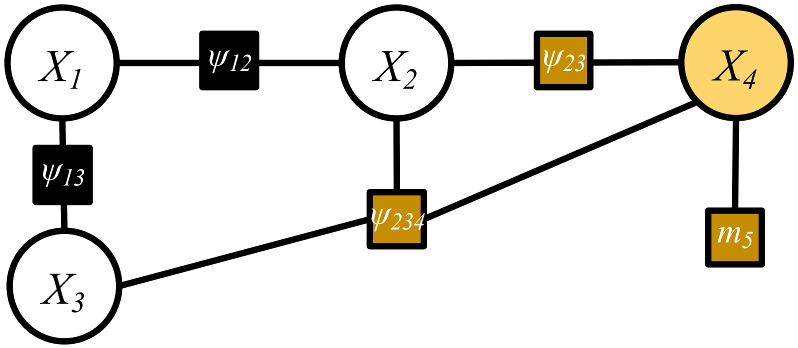
$$Z = \sum_{x_1} \sum_{x_2} \psi_{12}(x_1, x_2) \sum_{x_3} \psi_{13}(x_1, x_3) \sum_{x_4} \psi_{24}(x_2, x_4) \psi_{234}(x_2, x_3, x_4) \sum_{x_5} \psi_{45}(x_4, x_5) \psi_5(x_5)$$

$$= \sum_{x_1} \sum_{x_2} \psi_{12}(x_1, x_2) \sum_{x_3} \psi_{13}(x_1, x_3) \sum_{x_4} \psi_{24}(x_2, x_4) \psi_{234}(x_2, x_3, x_4) m_5(x_4)$$

$$= \sum_{x_1} \sum_{x_2} \psi_{12}(x_1, x_2) \sum_{x_3} \psi_{13}(x_1, x_3) m_4(x_2, x_3)$$

In a factor graph, variable **elimination** corresponds to replacement of a subgraph with a factor.

The Variable Elimination Algorithm



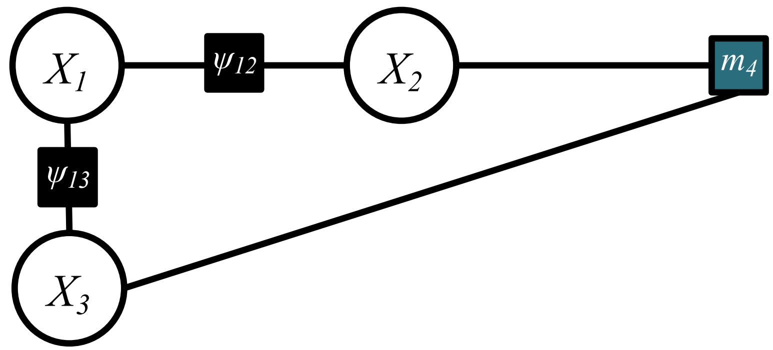
$$Z = \sum_{x_1} \sum_{x_2} \psi_{12}(x_1, x_2) \sum_{x_3} \psi_{13}(x_1, x_3) \sum_{x_4} \psi_{24}(x_2, x_4) \psi_{234}(x_2, x_3, x_4) \sum_{x_5} \psi_{45}(x_4, x_5) \psi_5(x_5)$$

$$= \sum_{x_1} \sum_{x_2} \psi_{12}(x_1, x_2) \sum_{x_3} \psi_{13}(x_1, x_3) \sum_{x_4} \psi_{24}(x_2, x_4) \psi_{234}(x_2, x_3, x_4) m_5(x_4)$$

$$= \sum_{x_1} \sum_{x_2} \psi_{12}(x_1, x_2) \sum_{x_3} \psi_{13}(x_1, x_3) m_4(x_2, x_3)$$

In a factor graph, variable **elimination** corresponds to replacement of a subgraph with a factor.

The Variable Elimination Algorithm



$$Z = \sum_{x_1} \sum_{x_2} \psi_{12}(x_1, x_2) \sum_{x_3} \psi_{13}(x_1, x_3) \sum_{x_4} \psi_{24}(x_2, x_4) \psi_{234}(x_2, x_3, x_4) \sum_{x_5} \psi_{45}(x_4, x_5) \psi_5(x_5)$$

$$= \sum_{x_1} \sum_{x_2} \psi_{12}(x_1, x_2) \sum_{x_3} \psi_{13}(x_1, x_3) \sum_{x_4} \psi_{24}(x_2, x_4) \psi_{234}(x_2, x_3, x_4) m_5(x_4)$$

$$= \sum_{x_1} \sum_{x_2} \psi_{12}(x_1, x_2) \sum_{x_3} \psi_{13}(x_1, x_3) m_4(x_2, x_3)$$

In a factor graph, variable **elimination** corresponds to replacement of a subgraph with a factor.

Variable Elimination for Marginal Inference

Algorithm 1: Variable Elimination for Marginal Inference

Input: the factor graph and the query variable

Output: the marginal distribution for the query variable

a. Run a breadth-first-search starting at the query variable to obtain an ordering of the variable nodes
b. Reverse that ordering
c. Eliminate each variable in the reversed ordering using Algorithm 2

Algorithm 2: Eliminate One Variable

Input: the variable to be eliminated

Output: new factor graph with the variable marginalized out

a. Find the input variable and its neighboring factors -- call this set the eliminated set
b. Replace the eliminated set with a new factor
    a. The neighbors of the new factor should be all the neighbors of all the factors in the eliminated set
    b. The new factor should assign a score to each possible assignment of its neighboring variables
    c. Said score should be identical to the product of the factors it is replacing, summing over the eliminated variable

Variable Elimination for Marginal Inference

Algorithm 3: Variable Elimination for the Partition Function

Input: the factor graph

Output: the partition function

a. Run a breadth-first-search starting at an arbitrary variable to obtain an ordering of the variable nodes
b. Eliminate each variable in the ordering using Algorithm 2

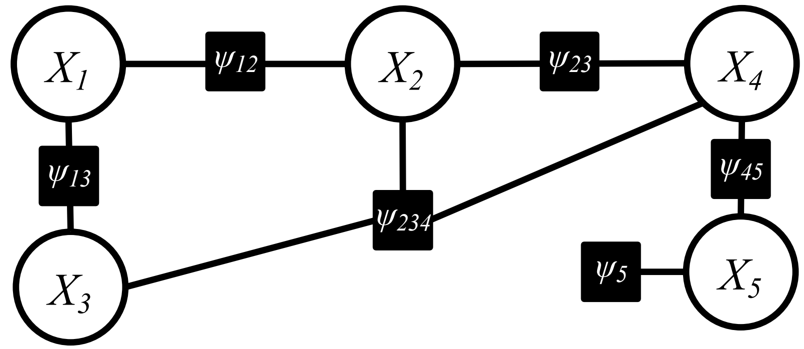
Algorithm 2: Eliminate One Variable

Input: the variable to be eliminated

Output: new factor graph with the variable marginalized out

a. Find the input variable and its neighboring factors -- call this set the eliminated set
b. Replace the eliminated set with a new factor
   a. The neighbors of the new factor should be all the neighbors of all the factors in the eliminated set
   b. The new factor should assign a score to each possible assignment of its neighboring variables
   c. Said score should be identical to the product of the factors it is replacing, summing over the eliminated variable

Variable Elimination Complexity



In-Class Exercise: *Fill in the blank*

Brute force, naïve, inference is O(____)

Variable elimination is O(____)

where
- n = # of variables
- k = max # values a variable can take
- r = # variables participating in largest "intermediate" table