ECE 6504: Advanced Topics in Machine Learning

Probabilistic Graphical Models and Large-Scale Learning

Topics

- Markov Random Fields: Representation
  - Conditional Random Fields
  - Log-Linear Models

Readings: KF 4.1-3; Barber 4.1-2

Dhruv Batra

Virginia Tech

Administrativia

- No class
  - Next week (Tue, Thu)

- Project Proposal
  - Due: ~~Mar 12~~, Mar 5, 11:59pm
  - <=2pages, NIPS format

- HW2
  - Out later today
  - Due: Mar 12, 11:59pm
  - Implementation: Variable Elimination in BNs

Recap of Last Time

Markov Nets

- Set of random variables

- Undirected graph
  - Encodes independence assumptions

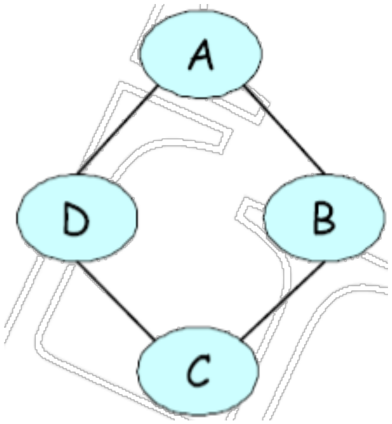
- Unnormalized Factor Tables

- Joint distribution:
  - Product of Factors

Pairwise MRFs

- Pairwise Factors
  - A function of 2 variables
    - Often unary terms are also allowed (although strictly speaking unnecessary)
  - On board

Pairwise MRF: Example



$\phi_1[A,B]$			$\phi_2[B,C]$			$\phi_3[C,D]$			$\phi_4[D,A]$		
a^0	b^0	30	b^0	c^0	100	c^0	d^0	1	d^0	a^0	100
a^0	b^1	5	b^0	c^1	1	c^0	d^1	100	d^0	a^1	1
a^1	b^0	1	b^1	c^0	1	c^1	d^0	100	d^1	a^0	1
a^1	b^1	10	b^1	c^1	100	c^1	d^1	1	d^1	a^1	100

Computing probabilities in Markov networks vs BNs

- In a BN, can compute prob. of an instantiation by multiplying CPTs

- In an Markov networks, can only compute ratio of probabilities directly

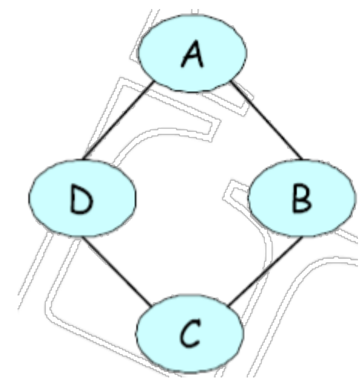
$\phi_1[A,B]$			$\phi_2[B,C]$			$\phi_3[C,D]$			$\phi_4[D,A]$		
a^0	b^0	30	b^0	c^0	100	c^0	d^0	1	d^0	a^0	100
a^0	b^1	5	b^0	c^1	1	c^0	d^1	100	d^0	a^1	1
a^1	b^0	1	b^1	c^0	1	c^1	d^0	100	d^1	a^0	1
a^1	b^1	10	b^1	c^1	100	c^1	d^1	1	d^1	a^1	100

Normalization for computing probabilities

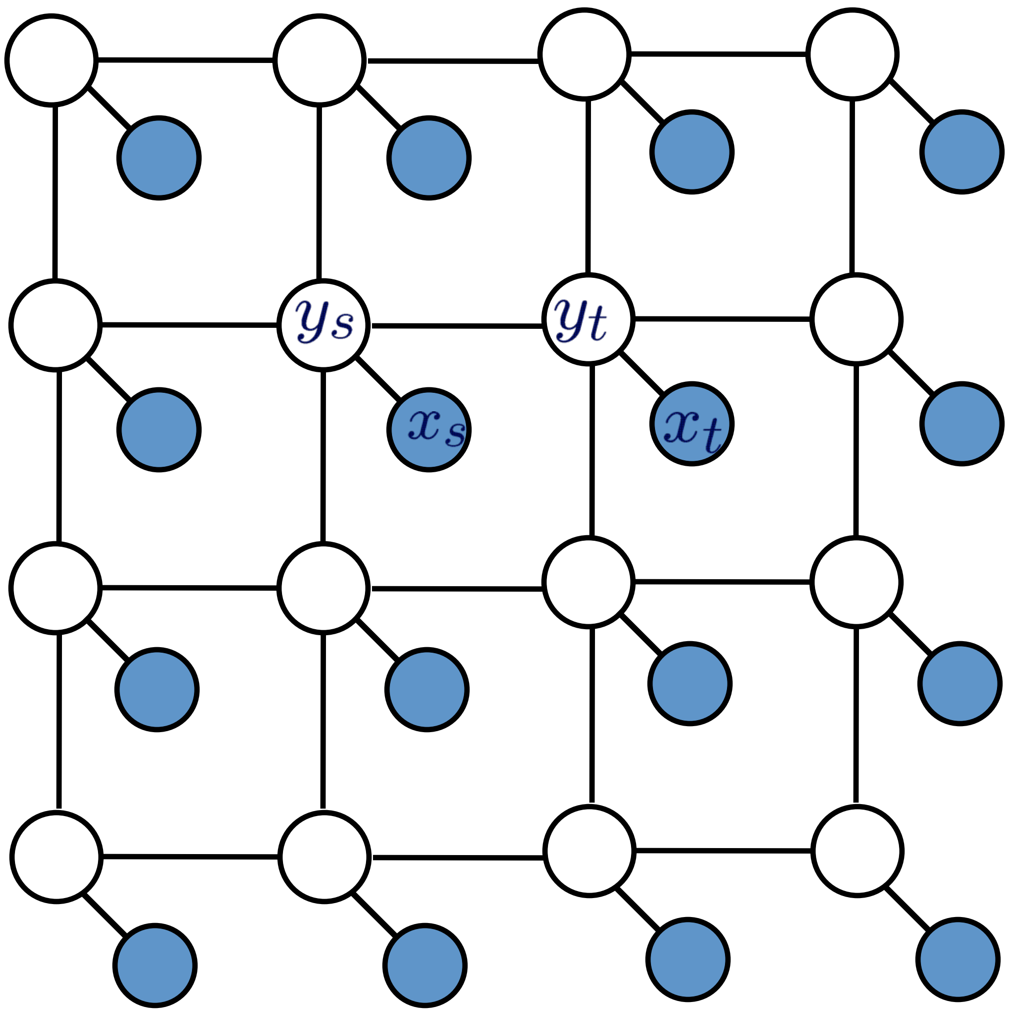
- To compute actual probabilities, must compute normalization constant (also called partition function)

Assignment				*Unnormalized*	*Normalized*
a^0	b^0	c^0	d^0	300000	0.04
a^0	b^0	c^0	d^1	300000	0.04
a^0	b^0	c^1	d^0	300000	0.04
a^0	b^0	c^1	d^1	30	$4.1 \cdot 10^{-6}$
a^0	b^1	c^0	d^0	500	$6.9 \cdot 10^{-5}$
a^0	b^1	c^0	d^1	500	$6.9 \cdot 10^{-5}$
a^0	b^1	c^1	d^0	5000000	0.69
a^0	b^1	c^1	d^1	500	$6.9 \cdot 10^{-5}$
a^1	b^0	c^0	d^0	100	$1.4 \cdot 10^{-5}$
a^1	b^0	c^0	d^1	1000000	0.14
a^1	b^0	c^1	d^0	100	$1.4 \cdot 10^{-5}$
a^1	b^0	c^1	d^1	100	$1.4 \cdot 10^{-5}$
a^1	b^1	c^0	d^0	10	$1.4 \cdot 10^{-6}$
a^1	b^1	c^0	d^1	100000	0.014
a^1	b^1	c^1	d^0	100000	0.014
a^1	b^1	c^1	d^1	100000	0.014

- Computing partition function is hard! Must sum over all possible assignments



Nearest-Neighbor Grids



Low Level Vision

- Image denoising
- Stereo
- Optical flow
- Shape from shading
- Superresolution
- Segmentation

y_s → unobserved or hidden variable

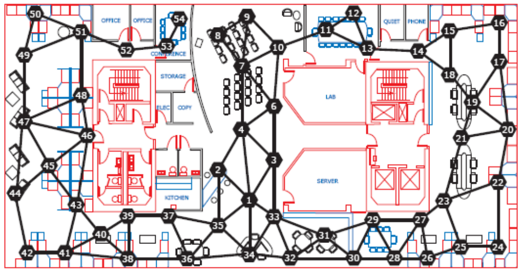
x_s → local observation

General Gibbs Distribution

- Arbitrary Factors

- “Induced” MRF Graph

Factorization in Markov networks



- Given an undirected graph *H* over variables $\mathbf{X}=\{X_1,...,X_n\}$

- A distribution *P* **factorizes** over *H* if there exist
  - subsets of variables $\mathbf{D}_1 \subseteq \mathbf{X}, \ldots, \mathbf{D}_m \subseteq \mathbf{X}$, such that $\mathbf{D}_i$ are *fully connected* in *H*
  - *non-negative potentials* (or factors) $\phi_1(\mathbf{D}_1), \ldots, \phi_m(\mathbf{D}_m)$
    - also known as clique potentials
  - such that

$$P(X_1, \ldots, X_n) = \frac{1}{Z} \prod_{i=1}^{m} \phi_i(\mathbf{D}_i)$$

- Also called Markov random field *H,* or Gibbs distribution over *H*

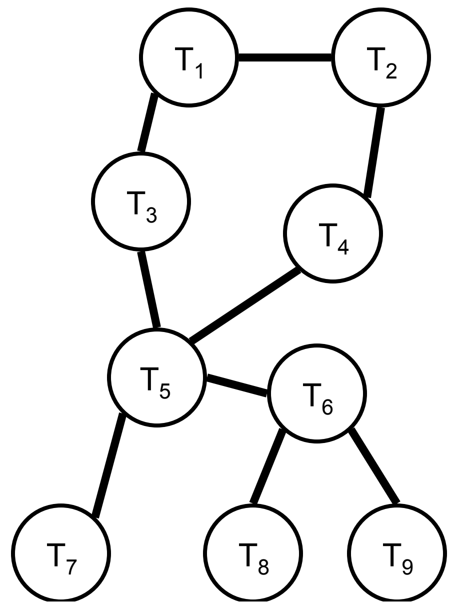
MRFs

- Given a graph H, are factors unique?

Active Trails and Separation

- A path X_1 – … – X_k is **active** when set of variables **Z** are observed
  - if none of $X_i \in \{X_1,\ldots,X_k\}$ are observed (are part of **Z**)

- Variables **X** are **separated** from **Y** given **Z** in graph
  - If no active path between any $X \in \mathbf{X}$ and any $Y \in \mathbf{Y}$ given **Z**



Markov networks representation Theorem 1

If joint probability distribution *P*: $$P(X_1, \ldots, X_n) = \frac{1}{Z} \prod_{i=1}^{m} \phi_i(\mathbf{D}_i)$$	Then	*H* is an I-map for *P*

- If
  - you can write distribution as a normalized product of factors
- Then
  - Can read independencies from graph

What about the other direction for Markov networks ?

If H is an I-map for P **Then** joint probability distribution P:

$$P(X_1, \ldots, X_n) = \frac{1}{Z} \prod_{i=1}^{m} \phi_i(\mathbf{D}_i)$$

- Counter-example: $X_1,\ldots,X_4$ are binary, and only eight assignments have positive probability:

(0,0,0,0)	(1,0,0,0)	(1,1,0,0)	~~(1,1,1,0)~~
(0,0,0,1)	(0,0,1,1)	(0,1,1,1)	~~(1,1,1,1)~~

- For example, $X_1 \perp X_3 | X_2, X_4$:
  - E.g., $P(X_1=0|X_2=0, X_4=0)$

- But distribution doesn't factorize!!

Representation Theorem for Markov Networks
Hammersley–Clifford theorem

If joint probability distribution *P*:

$$P(X_1, \ldots, X_n) = \frac{1}{Z} \prod_{i=1}^{m} \phi_i(\mathbf{D}_i)$$

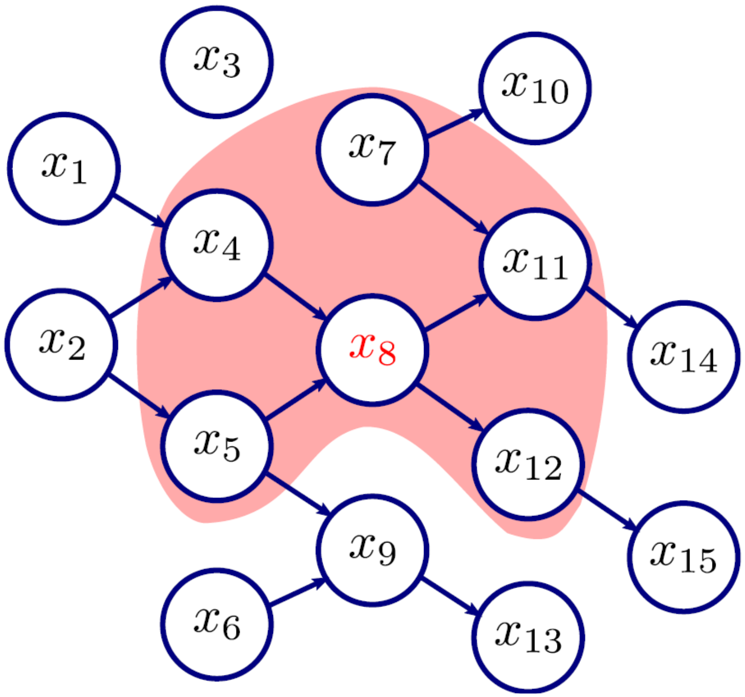
Then → *H* is an I-map for *P*

If *H* is an I-map for *P* and *P* is a positive distribution

Then → joint probability distribution *P*:

$$P(X_1, \ldots, X_n) = \frac{1}{Z} \prod_{i=1}^{m} \phi_i(\mathbf{D}_i)$$

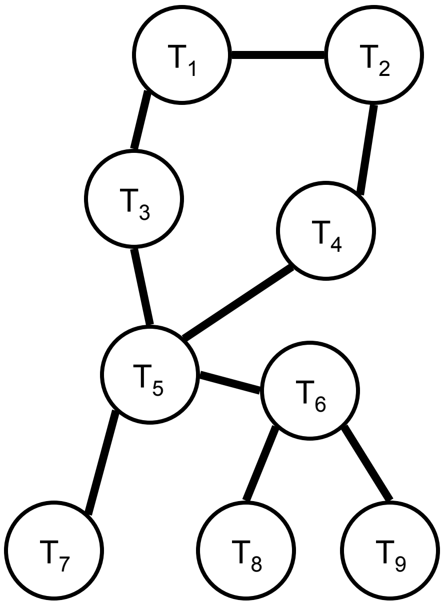
Markov Blanket



= Markov Blanket of variable x_8 – Parents, children and parents of children

Independence Assumptions in MNs

- **Separation** defines global independencies
- **Pairwise Markov Independence**:
  – Pairs of non-adjacent variables A,B are independent given all others
- **Markov Blanket**:
  – Variable A independent of rest given its neighbors



P-map

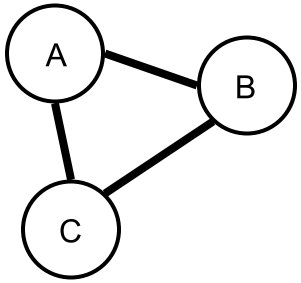
- Perfect map

- *G* is a **P-map** for *P* if
  - $I(P) = I(G)$

- Question: Does every distribution *P* have P-map?

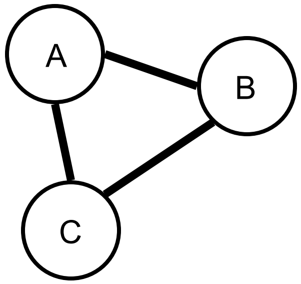
Structure in cliques

- Possible potentials for this graph:



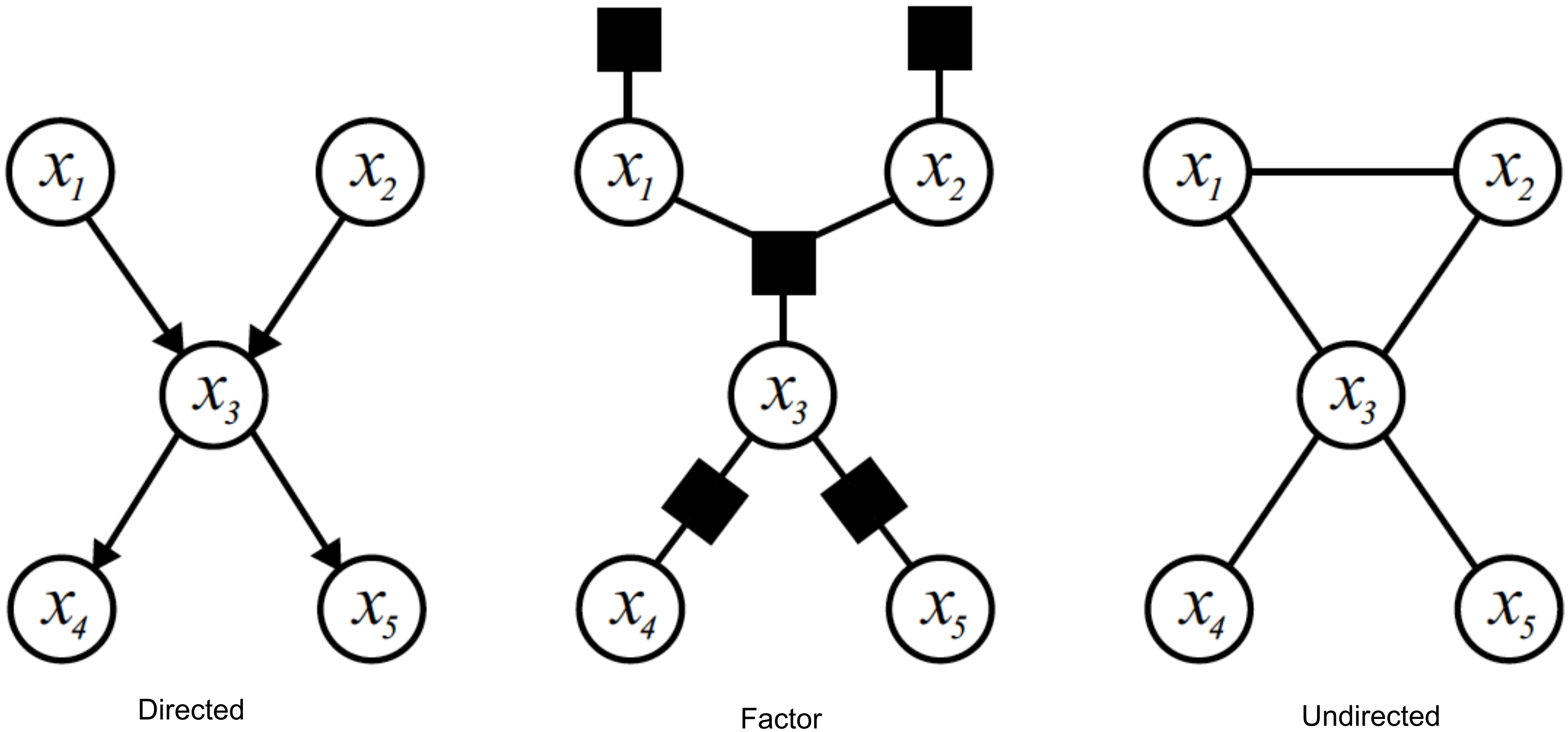
Factor graphs

- Bipartite graph:
  - variable nodes (ovals) for $X_1,\ldots,X_n$
  - factor nodes (squares) for $\phi_1,\ldots,\phi_m$
  - edge $X_i - \phi_j$ if $X_i \in Scope[\phi_j]$



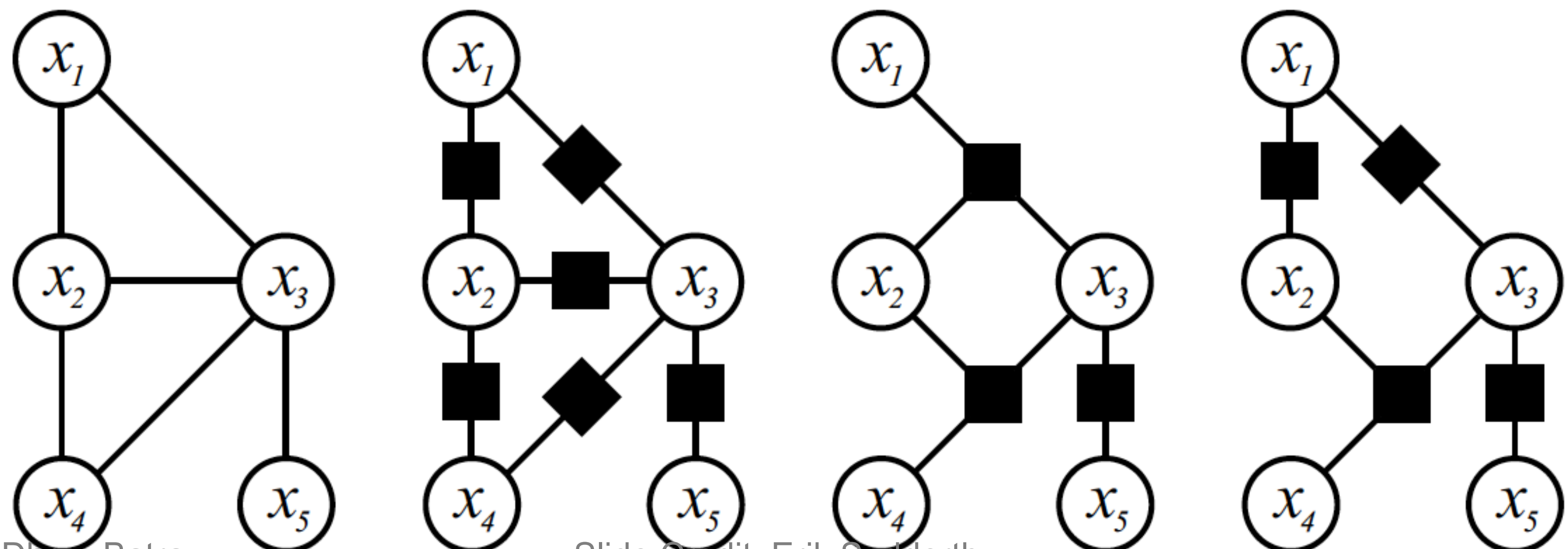
- Very useful for approximate inference
  - Make factor dependency explicit

Types of Graphical Models



Factor Graphs show Fine-grained Factorization

$$p(x) = \frac{1}{Z} \prod_{f \in \mathcal{F}} \psi_f(x_f)$$



Plan for today

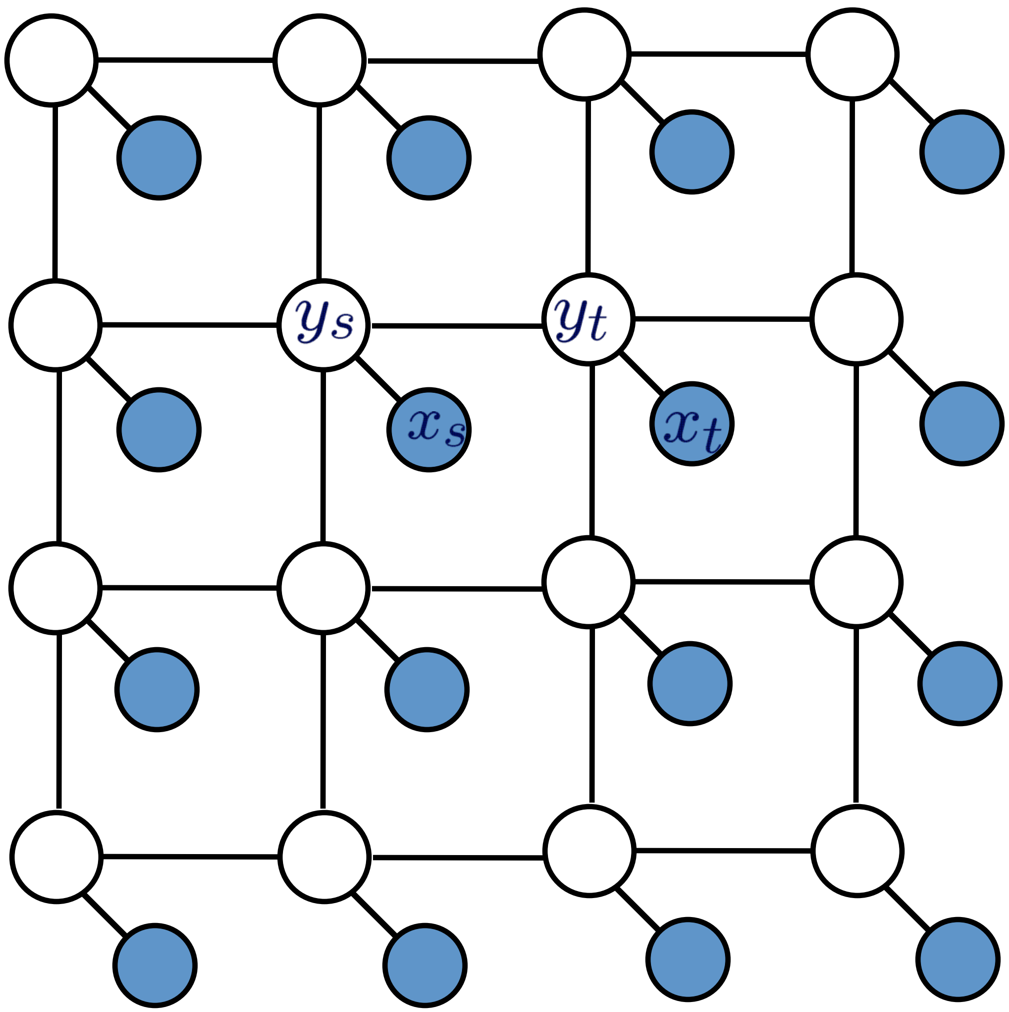
- Undirected Graphical Models: Representation
  - Conditional Random Fields
  - Log-Linear Models

- Undirected Graphical Models: Inference
  - Variable Elimination

Conditional Random Fields

- What's the difference between Naïve Bayes & Logistic Regression?

Nearest-Neighbor Grids



y_s → unobserved or hidden variable

x_s → local observation

Low Level Vision

- Image denoising
- Stereo
- Optical flow
- Shape from shading
- Superresolution
- Segmentation


Lazy Snapping
Siggraph 2004
Yin Li
Jian Sun
Chi-Keung Tang
Heung-Yeung Shum
Hong Kong Univercity of Science and Technology
Microsoft Research Asia

Logarithmic representation

- Standard model:

$$P(X_1, \ldots, X_n) = \frac{1}{Z} \prod_{i=1}^{m} \phi_i(\mathbf{D}_i)$$

- Log representation of potential (assuming positive potential):
  - also called the energy function

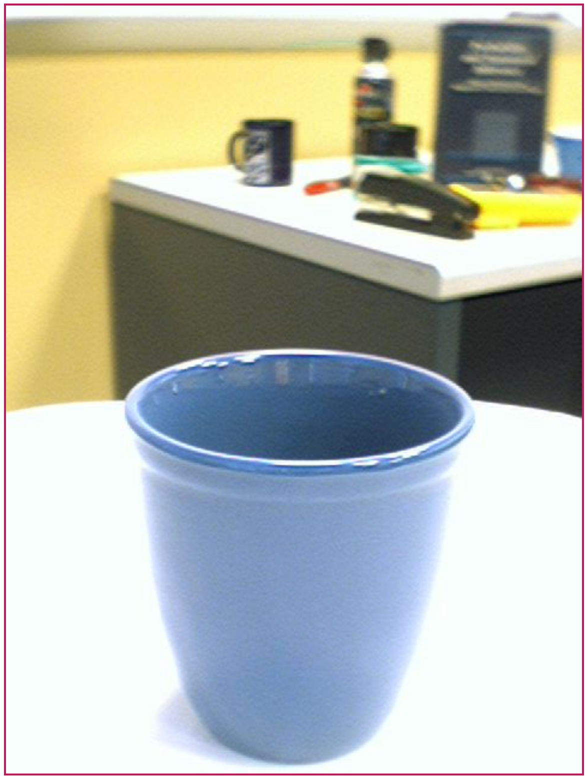
- Log representation of Markov net:

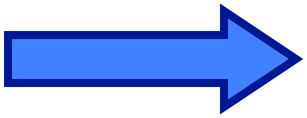
Log-linear Markov network (most common representation)

- **Feature (or Sufficient Statistic)** is some function f [**D**] for some subset of variables **D**
  - e.g., indicator function

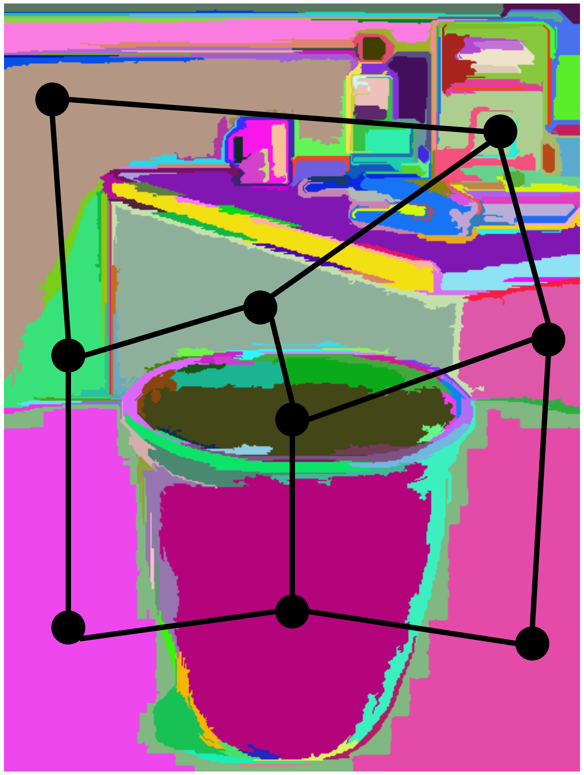
- **Log-linear model** over a Markov network *H*:
  - a set of features $f_1[\mathbf{D}_1],\ldots, f_k[\mathbf{D}_k]$
    - each $\mathbf{D}_i$ is a subset of a clique in *H*
    - two f's can be over the same variables
  - a set of weights $w_1,\ldots,w_k$
    - usually learned from data
  - $$P(X_1,\ldots,X_n) = \frac{1}{Z}\exp\left[\sum_{i=1}^{k} w_i f_i(\mathbf{D}_i)\right]$$

CRFs



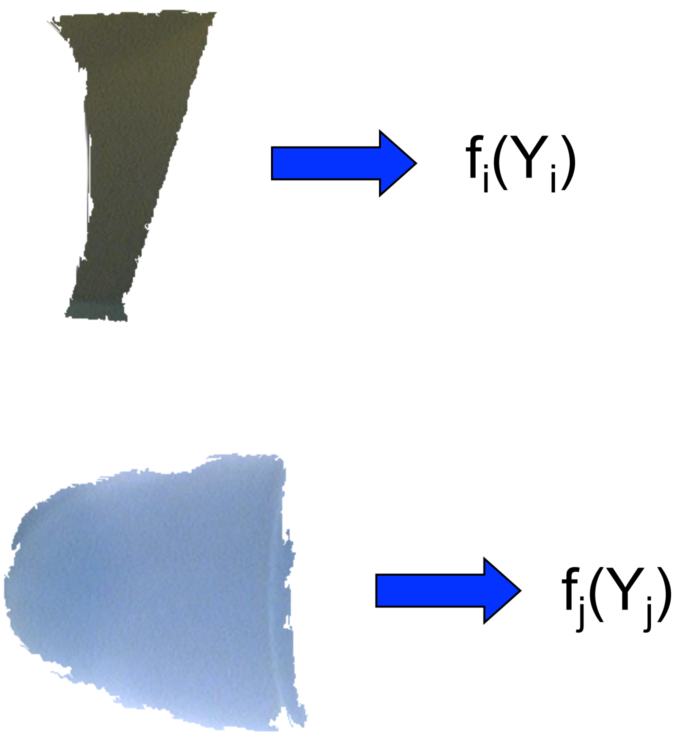


Felzenszwalb, Huttenlocher,
IJCV '04

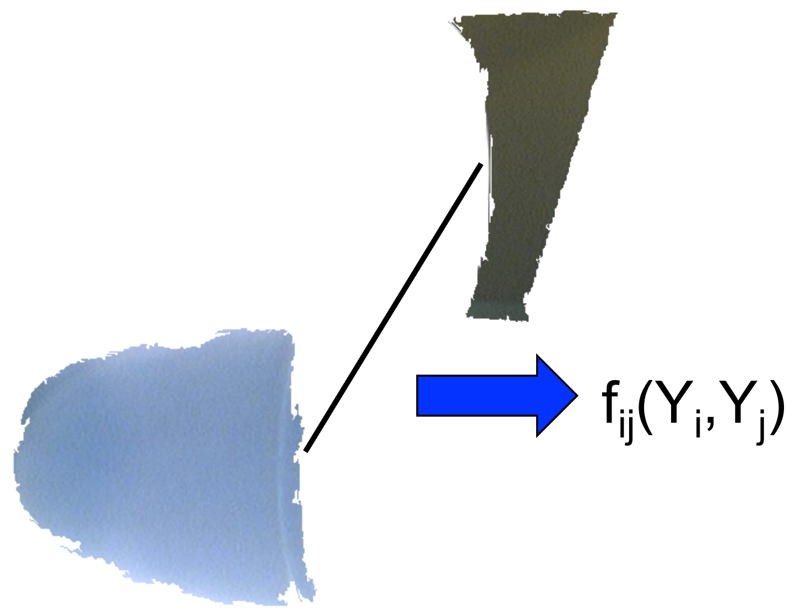


CRFs

Node Features



Edge Features



Node Feature -- Color

Superpixel

Feature Extraction
Step 1

Rmean

Gmean

Bmean

Hmean

Smean

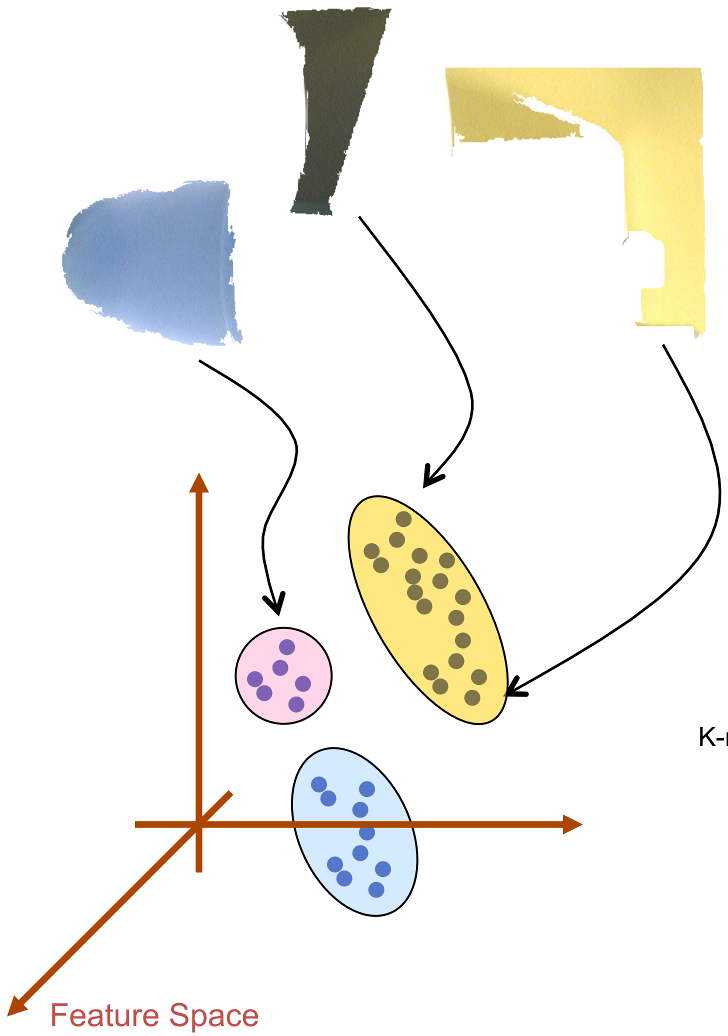
Vmean

5-dim
hist on H

3-dim
hist on S

Hoiem, Efros, Hebert, IJCV 2007

Node Feature – Color Clustering



$$\left\{ (\mu_1, \Sigma_1), (\mu_2, \Sigma_2), \ldots, (\mu_N, \Sigma_N) \right\}$$

K-means/X-means, Pelleg, Moore
Auton Lab implementation

Dictionary

Node Feature -- Color

Superpixel

Feature Extraction
Step 1

R_{mean}

G_{mean}

B_{mean}

H_{mean}

S_{mean}

V_{mean}

5-dim
hist on H

3-dim
hist on S

Feature Extraction
Step 2

Pr (Cluster | feature)

$\Pr(\text{Cluster } 1 \mid \phi_i)$

$\vdots$

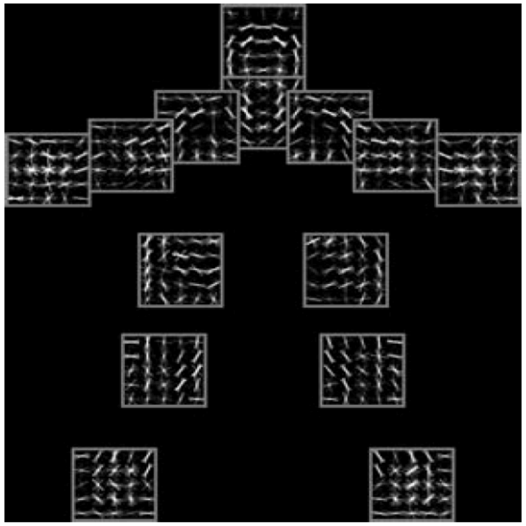
$\Pr(\text{Cluster } i \mid \phi_i)$

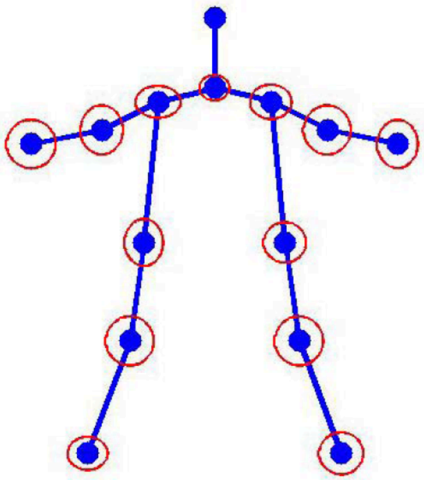
$\vdots$

$\Pr(\text{Cluster } N \mid \phi_i)$

Hoiem, Efros, Hebert, IJCV 2007

Conditional Random Fields





Summary of types of Markov nets

- Pairwise Markov networks
  - very common
  - potentials over nodes and edges

- General MRFs

- Factor graphs
  - explicit representation of factors
    - you know exactly what factors you have
  - very useful for approximate inference

- Log-linear models
  - log representation of potentials
  - linear coefficients learned from data
  - most common for learning MNs