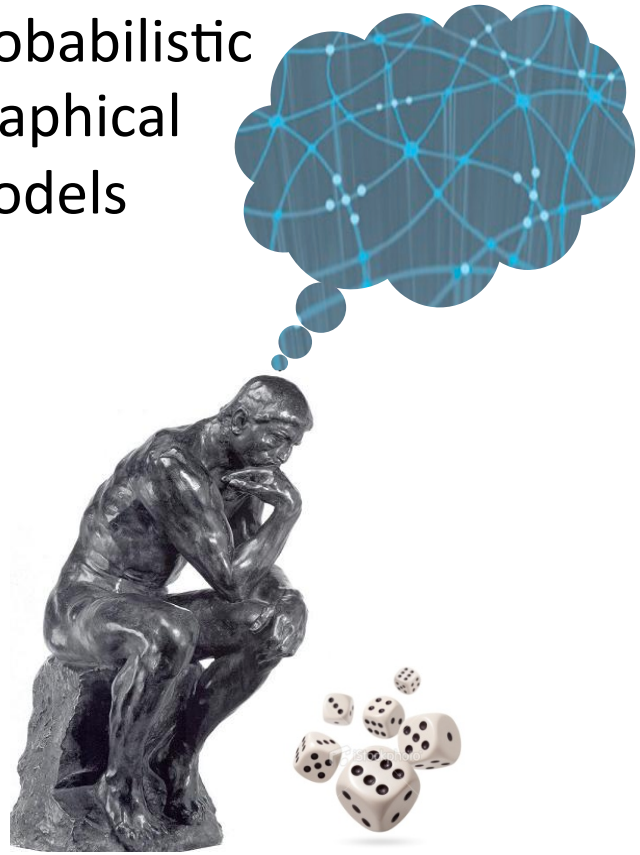
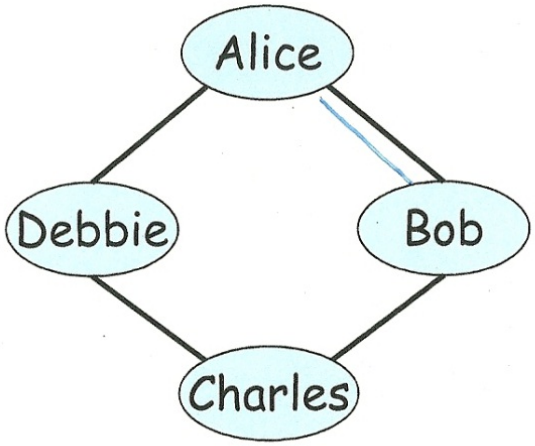
Probabilistic Graphical Models



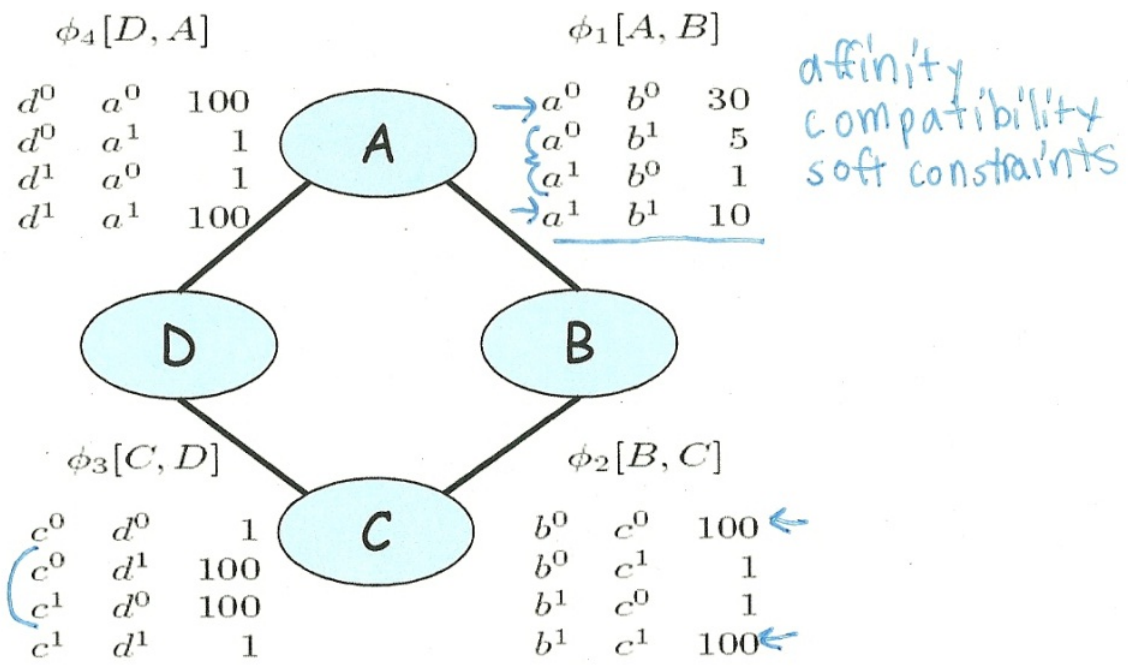
Representation

Markov Networks

Pairwise Markov Networks



Daphne Koller



Daphne Koller

$$\tilde{P}(A,B,C,D) = \phi_1(A,B) \times \phi_2(B,C) \times \phi_3(C,D) \times \phi_4(A,D)$$

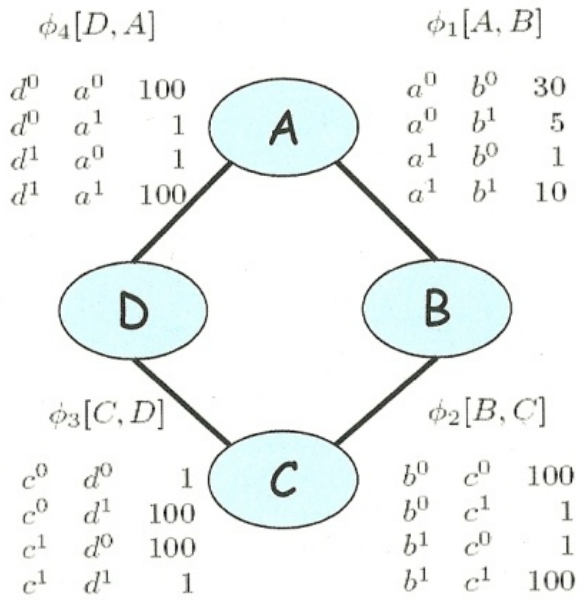
unnormalizedMeasure

$$P(A,B,C,D) = \frac{1}{Z}\tilde{P}(A,B,C,D)$$

partition function

Assignment				Unnormalized
a^0	b^0	c^0	d^0	300000
a^0	b^0	c^0	d^1	300000
a^0	b^0	c^1	d^0	300000
a^0	b^0	c^1	d^1	30
a^0	b^1	c^0	d^0	500
a^0	b^1	c^0	d^1	500
a^0	b^1	c^1	d^0	5000000
a^0	b^1	c^1	d^1	500
a^1	b^0	c^0	d^0	100
a^1	b^0	c^0	d^1	1000000
a^1	b^0	c^1	d^0	100
a^1	b^0	c^1	d^1	100
a^1	b^1	c^0	d^0	10
a^1	b^1	c^0	d^1	100000
a^1	b^1	c^1	d^0	100000
a^1	b^1	c^1	d^1	100000

Z



Daphne Koller

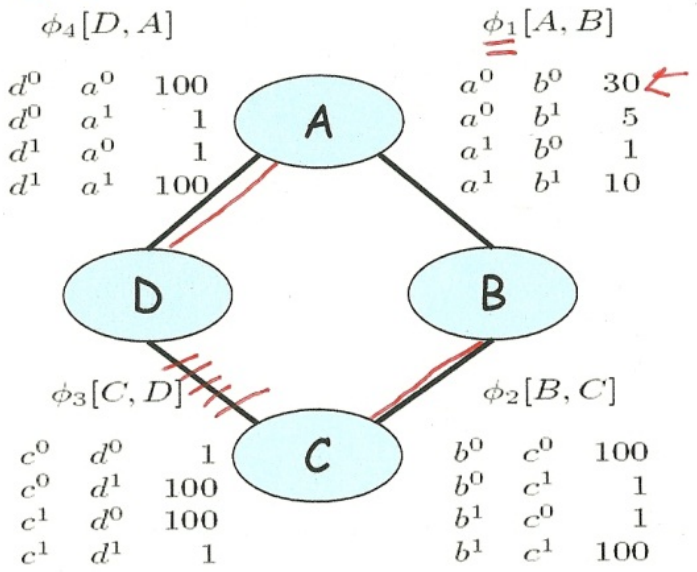
Consider the pairwise factor $\phi_1(A,B)$. That potential is proportional to:

- The marginal probability P(A,B)
- The conditional probability P(A | B)
- The conditional probability P(A, B | C, D)
- None of the above

$\Phi = \{\phi_1, \phi_2, \phi_3, \phi_4\}$

$P(A, B)$

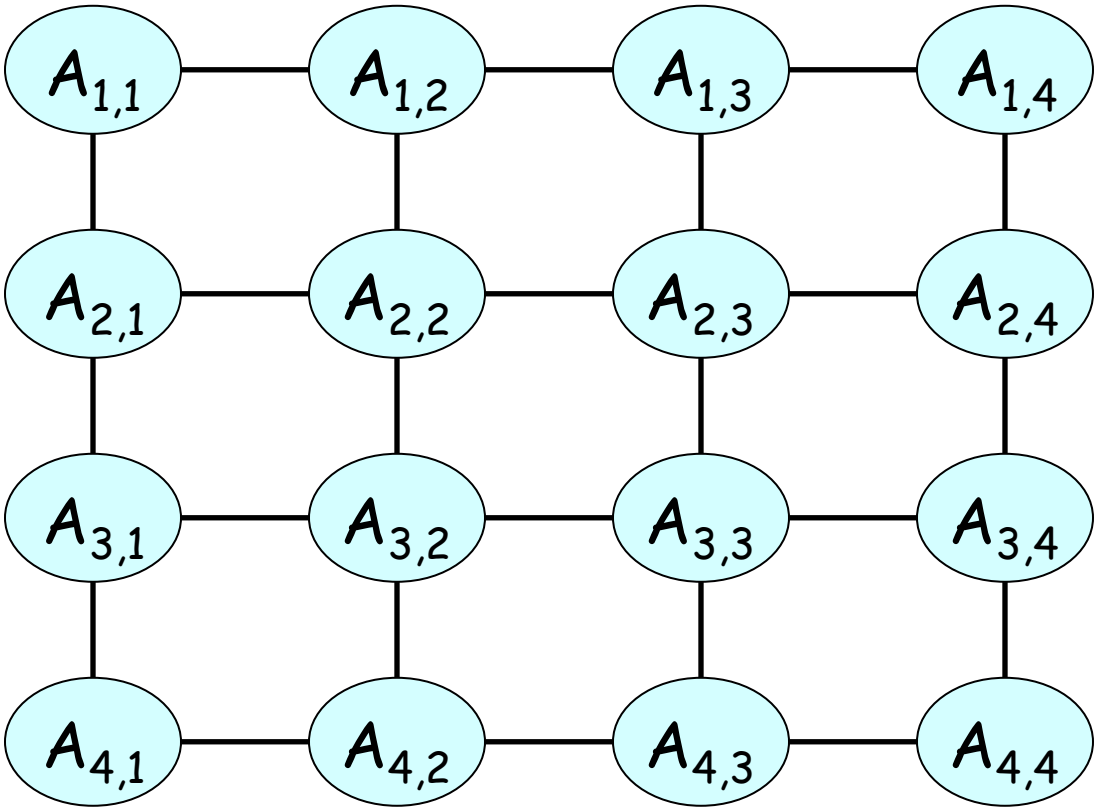
A	B	Prob.
a^0	b^0	0.13
a^0	b^1	0.69
a^1	b^0	0.14
a^1	b^1	0.04



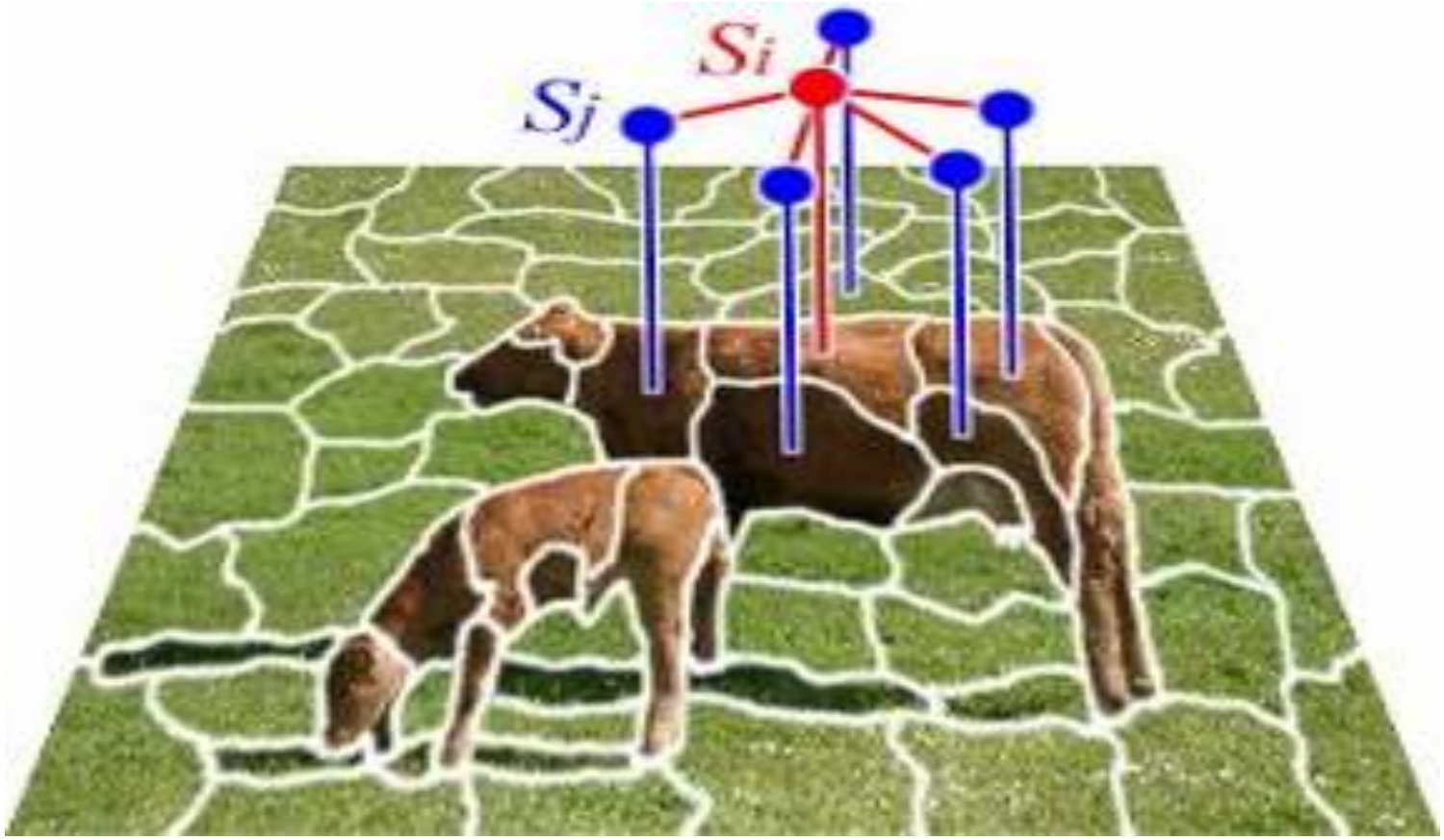
Daphne Koller

Pairwise Markov Networks

- A pairwise Markov network is an undirected graph whose nodes are $X_1,\ldots,X_n$ (random variables) and each edge $X_i - X_j$ is associated with a factor (potential) $\phi_{ij}(X_i, X_j)$



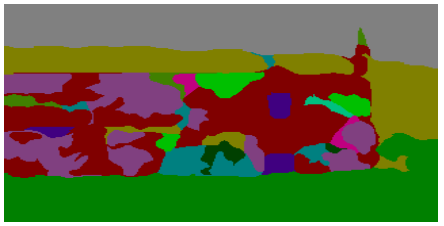
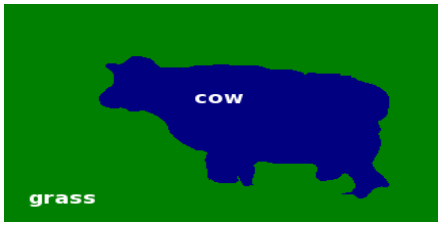
Daphne Koller



Daphne Koller



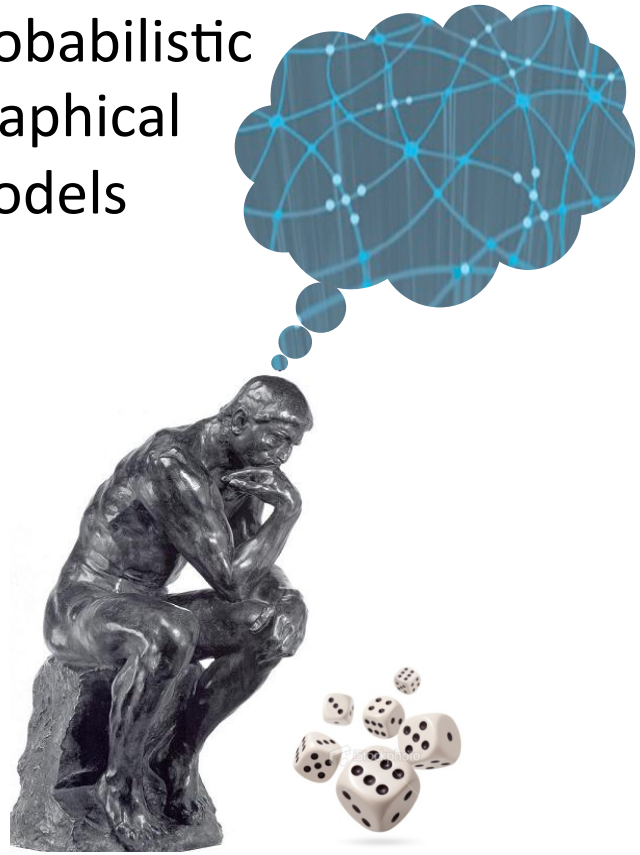


(a) (b) (c) (d)

Daphne Koller

Probabilistic Graphical Models

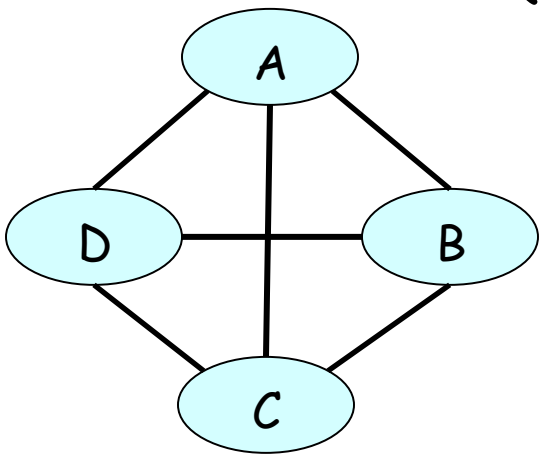


Representation

Markov Networks

General Gibbs Distribution

P(A,B,C,D)



Is this fully expressive?

Consider a fully connected pairwise Markov network over $X_1,\ldots,X_n$ where each X_i has d values. How many parameters does the network have?

- $O(d^n)$
- $O(n^d)$
- $O(n^2d^2)$
- $O(nd)$

$\binom{n}{2}$ edges $\times$ d^2 $\Rightarrow$ $O(n^2d^2)$ $\Rightarrow$ $O(d^n)$

Not every distribution can be represented as a pairwise Markov network

Gibbs Distribution

- Parameters:

  General factors $\phi_i(\mathbf{D}_i)$

  $\Phi = \{\phi_i(\mathbf{D}_i)\}$

a^1	b^1	c^1	0.25
a^1	b^1	c^2	0.35
a^1	b^2	c^1	0.08
a^1	b^2	c^2	0.16
a^2	b^1	c^1	0.05
a^2	b^1	c^2	0.07
a^2	b^2	c^1	0
a^2	b^2	c^2	0
a^3	b^1	c^1	0.15
a^3	b^1	c^2	0.21
a^3	b^2	c^1	0.09
a^3	b^2	c^2	0.18

Gibbs Distribution

Set of factors

$$\Phi = \{\phi_1(\boldsymbol{D}_1), \ldots, \phi_k(\boldsymbol{D}_k)\}$$

unnormalized measure

$$\tilde{P}_\Phi(X_1, \ldots, X_n) = \prod_{i=1} \phi_i(\boldsymbol{D}_i)$$

factor product

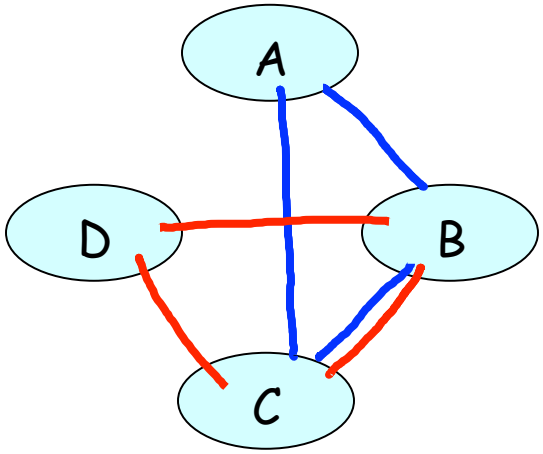
partition function

$$Z_\Phi = \sum_{X_1, \ldots, X_n} \tilde{P}_\Phi(X_1, \ldots, X_n)$$

$$\tilde{P}_\Phi(X_1, \ldots, X_n) = \frac{1}{Z_\Phi} \tilde{P}_\Phi(X_1, \ldots, X_n)$$

Induced Markov Network

$\phi_1(A, B, C), \phi_2(B, C, D)$



$$\Phi = \{\phi_1(\boldsymbol{D}_1), \ldots, \phi_k(\boldsymbol{D}_k)\}$$

Induced Markov network H_Φ has an edge X_i—X_j whenever there exists $\phi_m \in \Phi$ s.t. $X_i, X_j \in D_m$

Factorization

P factorizes over H if

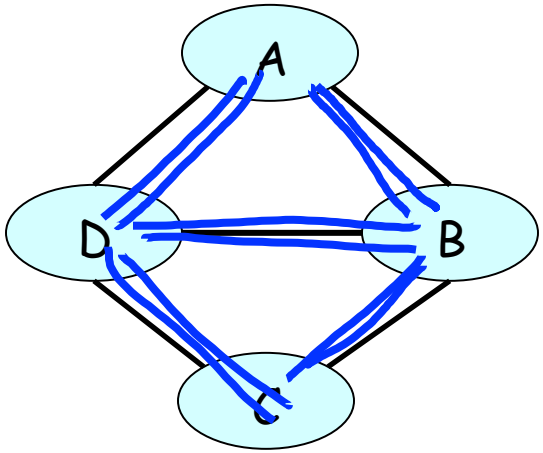
there exist $\Phi = \{\phi_1(\boldsymbol{D}_1), \ldots, \phi_k(\boldsymbol{D}_k)\}$

such that

$P = P_\Phi$ normalized product of factors in Φ

H is the induced graph for Φ

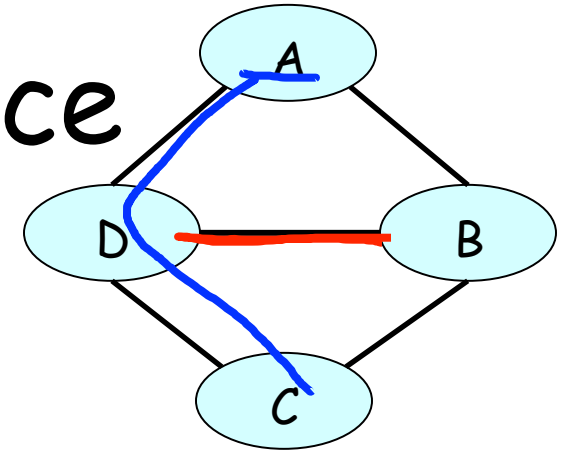
Which Gibbs distribution would induce the graph H?



- ☑ $\phi_1(A, B, D), \phi_2(B, C, D)$
- ☑ $\phi_1(A, B), \phi_2(B, C), \phi_3(C, D), \phi_4(A, D), \phi_5(B, D)$
- ☑ $\phi_1(A, B, D), \phi_2(B, C), \phi_3(C, D)$
- ○ All of the above

Cannot read the factorization from the graph

Flow of Influence



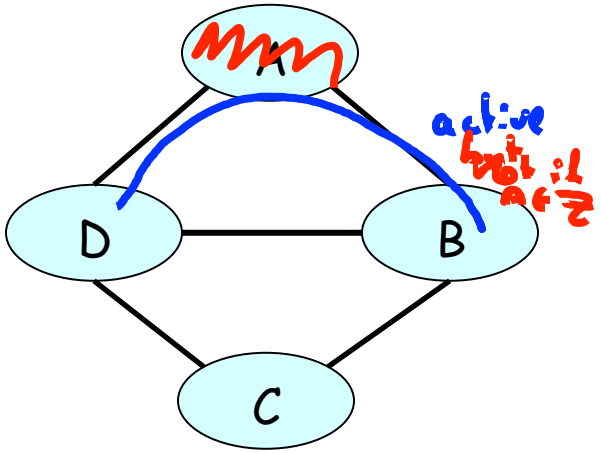
$\phi_1(A, B, D), \phi_2(B, C, D)$

$\phi_1(A, B), \phi_2(B, C), \phi_3(C, D), \phi_4(A, D), \phi_5(B, D)$

- Influence can flow along any trail, regardless of the form of the factors

Active Trails

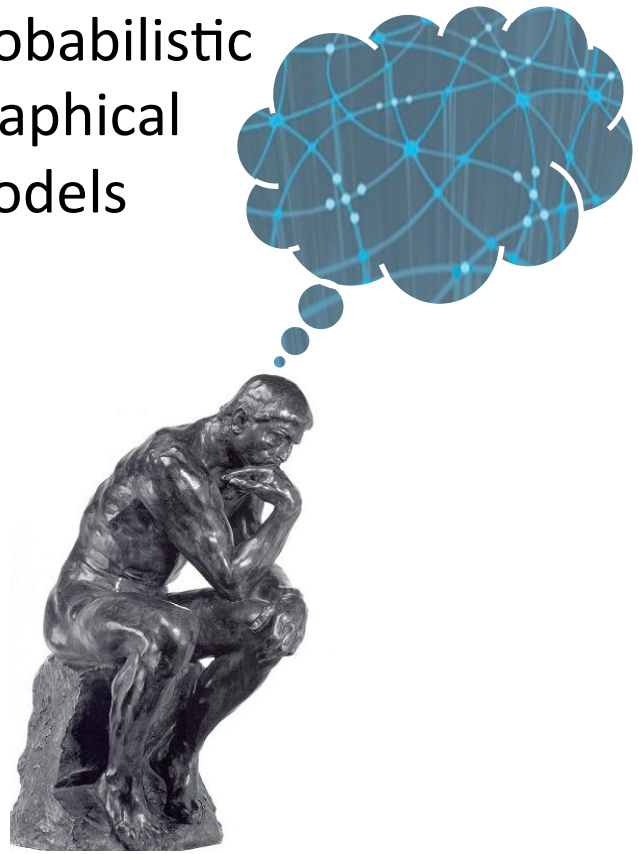
- A trail $X_1 - \ldots - X_n$ is active given **Z** if no X_i is in Z



Summary

- Gibbs distribution represents distribution as a product of factors
- Induced Markov network connects every pair of nodes that are in the same factor
- Markov network structure doesn't fully specify the factorization of P
- But active trails depend only on graph structure

Probabilistic Graphical Models



Representation

Markov Networks

Conditional Random Fields

Motivation

- Observed variables **X**
- Target variables **Y**

$X \longrightarrow Y$

$P(X,Y)$ joint

$P(Y|X)$ conditional

	X	Y
image segmentation	pixel values	pixel labels
text processing	words in a sentence	parts of speech

CRF Representation

$$\phi_1(\boldsymbol{D}_1), \ldots, \phi_k(\boldsymbol{D}_k)$$

$\tilde{P}(\bar{X}, \bar{Y}) = \prod_{i=1}^{k} \phi_i(\bar{D}_i)$ unnormalized measure

$Z(\bar{X}) = \sum_{\bar{Y}} \tilde{P}(\bar{X}, \bar{Y})$ different $Z(\bar{X})$ for every assignment to the obs. variables $\bar{X}$

$P(\bar{Y} \mid \bar{X}) = \frac{1}{Z(\bar{X})} \tilde{P}(\bar{X}, \bar{Y})$ $\left(\sum_{\bar{Y}} P(\bar{Y} \mid \bar{X}) = 1 \quad \text{for all } \bar{X} \right)$

Consider a new factor $\phi(\mathbf{D})$ such that $\mathbf{D} \subseteq X$. What is the effect of adding this factor ϕ on the distribution $P(Y|X)$?

- Introducing ϕ can change the distribution $P(Y|X)$
- ϕ multiplies both the unnormalized measure and the partition function and therefore cancels out, so $P(Y|X)$ remains unchanged
- ϕ affects only the partition function and therefore $P(Y|X)$ remains unchanged
- Introducing ϕ doesn't change $P(Y|X)$ only if $\mathbf{D}$ is a single variable

CRFs and Logistic Model

$$\phi_i(X_i, Y) = \exp\{w_i \mathbf{1}\{X_i = 1, Y = 1\}\}$$

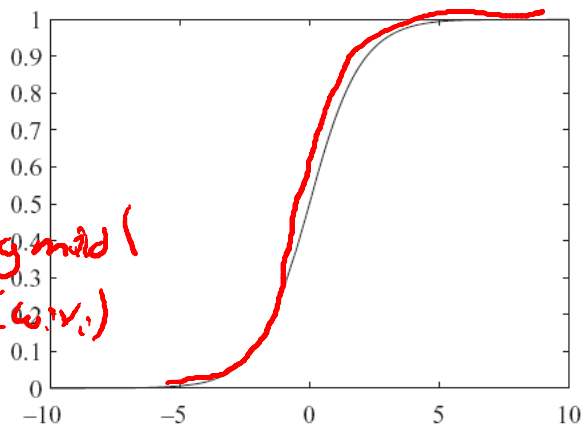
$\phi_i(x_i, Y=1) = \exp\{w_i x_i\}$

$\phi_i(x_i, Y=0) = 1$

$\tilde{p}(Y=1, x_1, \ldots, x_n) = \exp\left(\sum_i w_i x_i\right)$

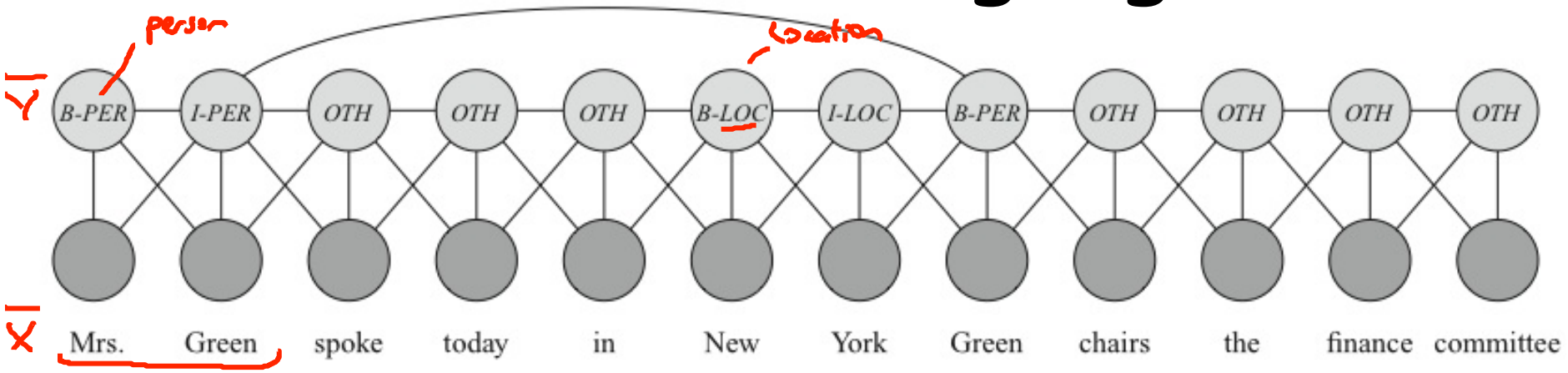
$\tilde{p}(Y=0, x_1, \ldots, x_n) = 1$

$$p(Y=1 \mid x_1, \ldots, x_n) = \frac{\exp\left(\sum_i w_i x_i\right)}{1 + \exp\left(\sum_i w_i x_i\right)} = \text{sigmoid}\left(\sum_i w_i x_i\right)$$



Daphne Koller

CRFs for Language

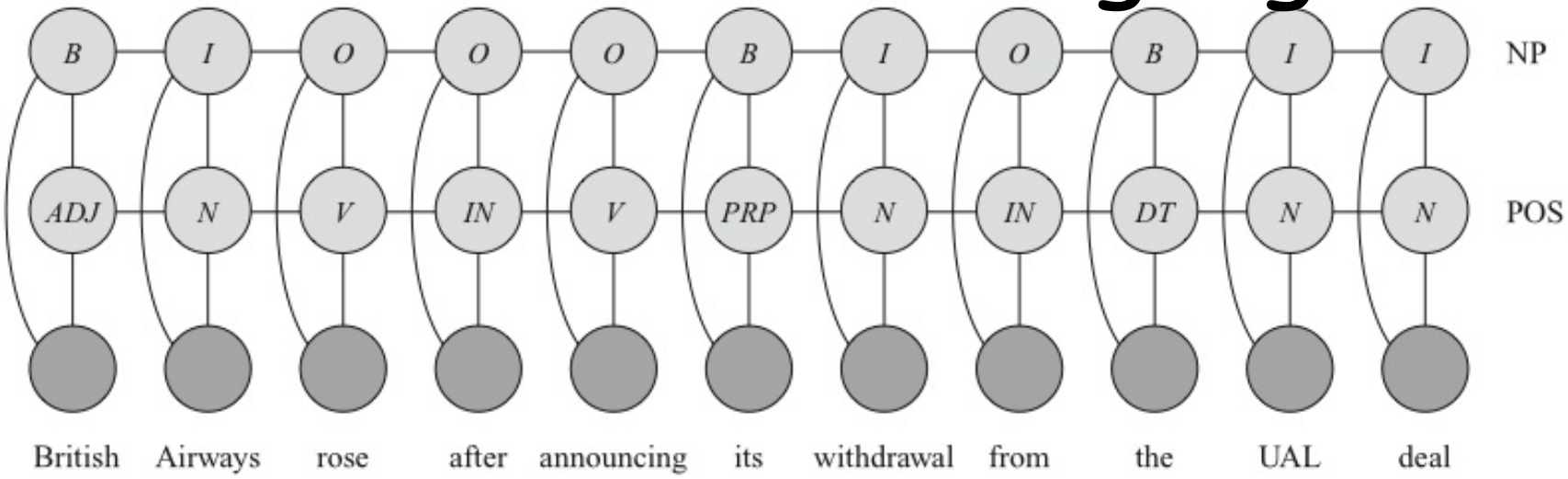


Features: word capitalized, word in atlas or name list, previous word is "Mrs", next word is "Times", …

$\tilde{P}(\bar{x},\bar{y})$

$\Rightarrow P(\bar{y}|\bar{x})$

More CRFs for Language



KEY

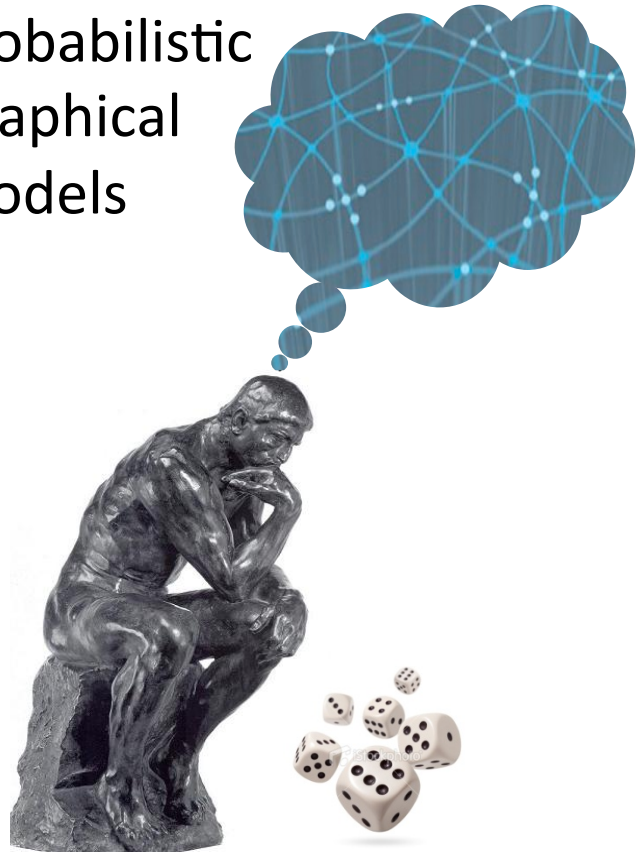
B	Begin noun phrase	*V*	Verb
I	Within noun phrase	*IN*	Preposition
O	Not a noun phrase	*PRP*	Possesive pronoun
N	Noun	*DT*	Determiner (e.g., a, an, the)
ADJ	Adiective		

Daphne Koller

Summary

- A CRF is parameterized the same as a Gibbs distribution, but normalized differently
- Don't need to model distribution over variables we don't care about
- Allows models with highly expressive features, without worrying about wrong independencies

Probabilistic Graphical Models



Representation

Independencies

Markov Networks

Separation in MNs

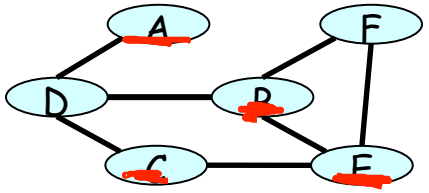
Definition:

X and **Y** are *separated* in H given **Z**

if there is no active trail in H

between **X** and **Y** given **Z**

no node along trail is in Z



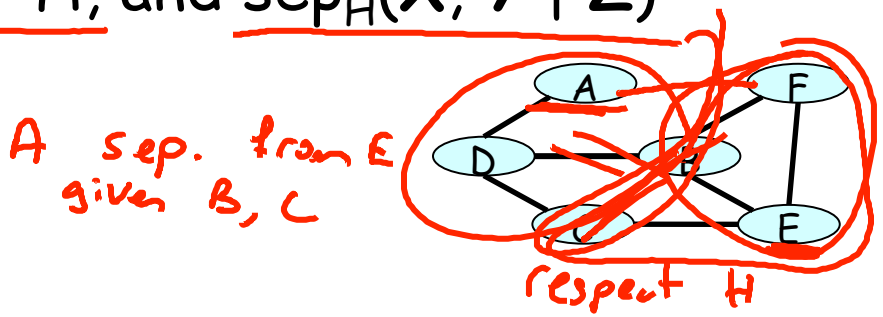
A, E separated given B, D

given D

given B, C

Factorization ⇨ Independence: MNS

Theorem: If P factorizes over H, and $sep_H(X, Y \mid Z)$ then P satisfies $(X \perp Y \mid Z)$



$\Pi\ \phi$ on A side $\cdot$ $\Pi\ \phi$ on E side $=$

cannot involve E

cannot involve A

Factorization ⇨ Independence: MNs

$$I(H) = \{(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}) : sep_H(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z})\}$$

If P satisfies I(H), we say that H is an I-map (independency map) of P

Theorem: If P factorizes over H, then H is an I-map of P

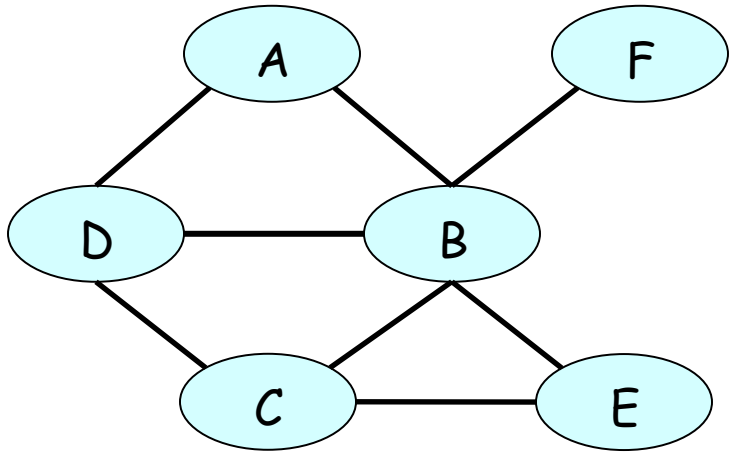
For the graph H shown below, which independence assertions are in I(H)? Mark all that apply.

$(A \perp C \mid B, D)$

$(A \perp E \mid B)$

$(C \perp F \mid B)$

$(A \perp F \mid B, D)$



Independence ⇨ Factorization

- **Theorem** (Hammersley Clifford):
  For a positive distribution P, if H is an I-map for P, then P factorizes over H

$P(\xi) > 0 \quad \forall \xi$

Summary

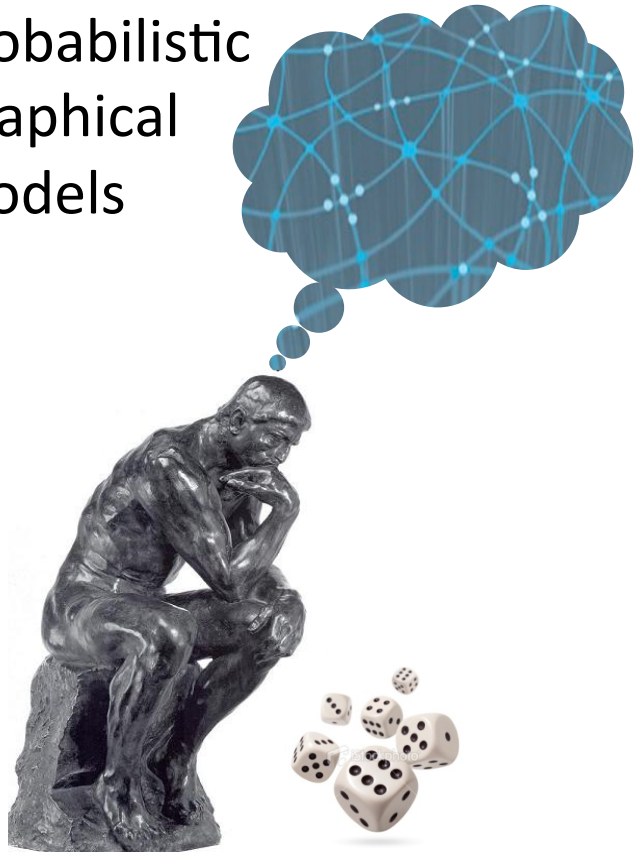
Two equivalent* views of graph structure:

- Factorization: H allows P to be represented
- I-map: Independencies encoded by H hold in P

If P factorizes over a graph H, we can read from the graph independencies that must hold in P (an independency map)

* for positive distributions

Probabilistic Graphical Models



Representation

Independencies

I-maps and Perfect Maps

Capturing Independencies in P

$$I(P) = \{(\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z}) : P \models (\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z})\}$$

- P factorizes over G ⇨ G is an I-map for P:

$$I(G) \subseteq I(P)$$

- But not always vice versa: there can be independencies in I(P) that are not in I(G)

Want a Sparse Graph

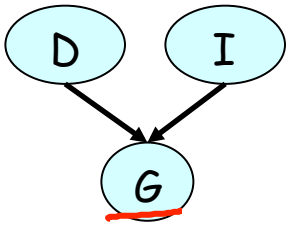
- If the graph encodes more independencies
  - it is sparser (has fewer parameters)
  - and more informative
- Want a graph that captures as much of the structure in P as possible

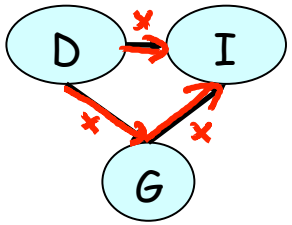
Minimal I-map

- Minimal I-map: I-map without redundant edges

$P(X \mid x^0) = P(X \mid x')$

- Minimal I-map may still not capture I(P)





D ⊥ G

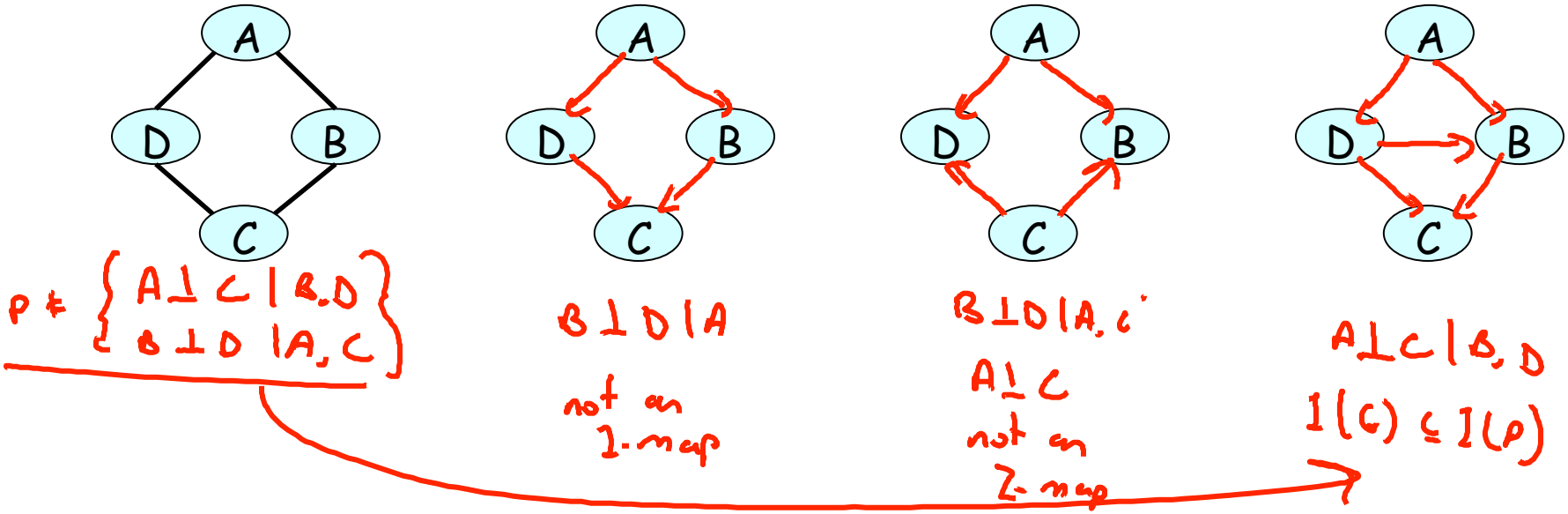
D ⊥ G | I

D ⊥ I | G

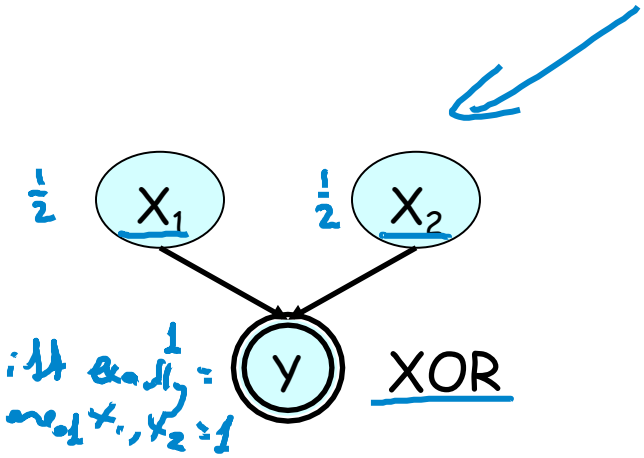
Perfect Map

- Perfect map: I(G) = I(P)
  - G perfectly captures independencies in P

Perfect Map



Another imperfect map

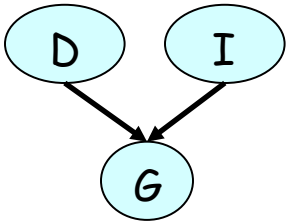


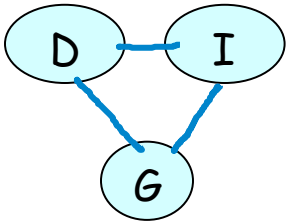
X_1	X_2	Y	Prob
0	0	0	0.25
0	1	1	0.25
1	0	1	0.25
1	1	0	0.25

$X_1 \perp X_2$ $X_1 \perp Y$ $X_2 \perp Y$

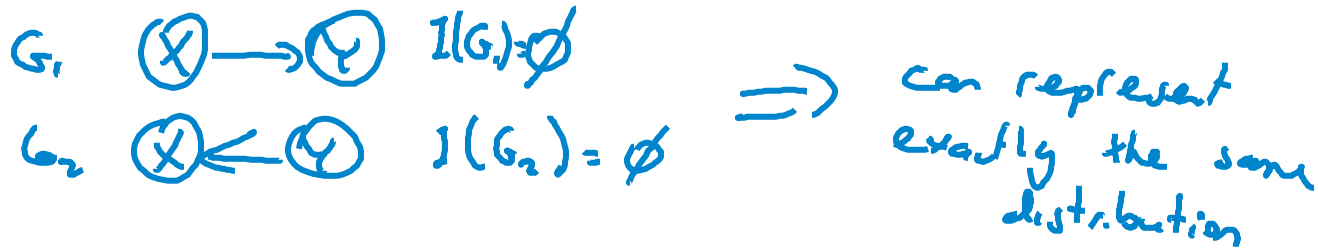
MN as a perfect map

- Perfect map: I(H) = I(P)
  - H perfectly captures independencies in P

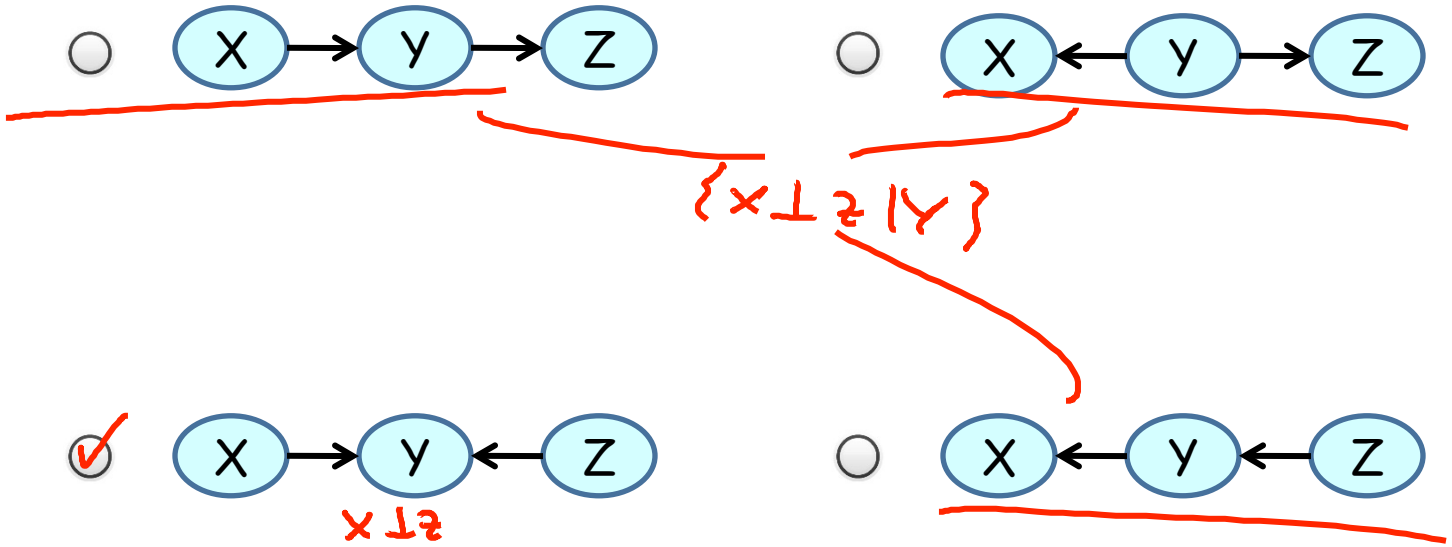




Uniqueness of Perfect Map



Which of the following graphs does not encode the same independencies as the others?



I-equivalence

Definition: Two graphs G_1 and G_2 over $X_1, \ldots, X_n$ are *I-equivalent* if $I(G_1)=I(G_2)$

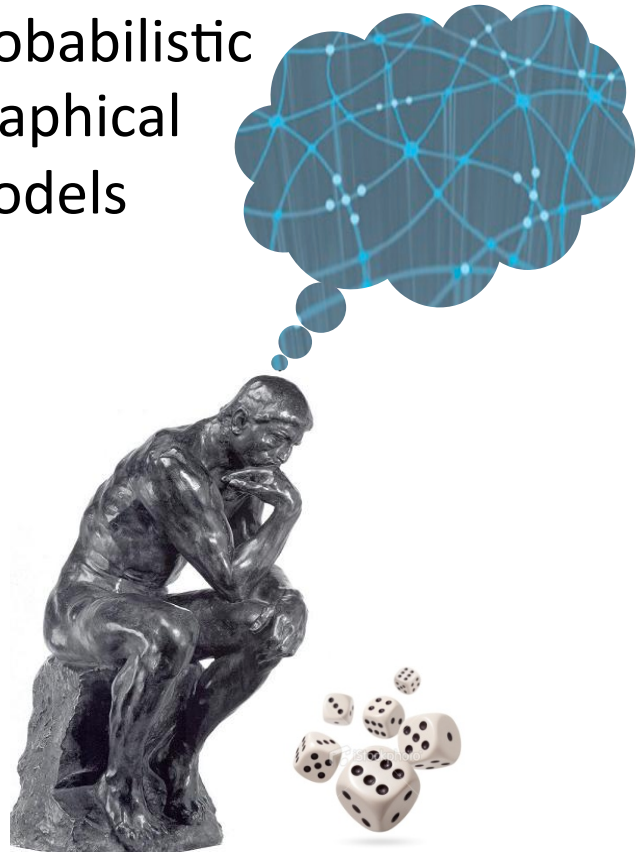
unidentifiable

Most G's have many I-equivalent variants

Summary

- Graphs that capture more of I(P) are more compact and provide more insight
- A minimal I-map may fail to capture a lot of structure even if present and representable as a PMap
- A perfect map is great, but may not exist
- Converting BNs ↔ MNs loses independencies
  - BN to MN: loses independencies in v-structures
  - MN to BN: must add triangulating edges to loops

Probabilistic Graphical Models



Representation

Local Structure

Log-Linear Models

Log-Linear Representation

$$\tilde{P} = \prod_i \phi_i(\boldsymbol{D}_i) \quad \Rightarrow \quad \tilde{P} = \exp\left(-\sum_j w_j f_j(\boldsymbol{D}_j)\right)$$

coefficients linear features

$$\tilde{P} = \prod_j \exp\left(-w_j f_j(\boldsymbol{D}_j)\right)$$

factor

- Each feature f_j has a scope D_j
- Different features can have same scope

Representing Table Factors

$\phi(X_1, X_2) = \begin{pmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{pmatrix}$

$$f_{12}^{00} = \mathbb{1}\{X_1 = 0, X_2 = 0\}$$
$$f_{12}^{01} = \mathbb{1}\{X_1 = 0, X_2 = 1\}$$
$$f_{12}^{10} = \mathbb{1}\{X_1 = 1, X_2 = 0\}$$
$$f_{12}^{11} = \mathbb{1}\{X_1 = 1, X_2 = 1\}$$

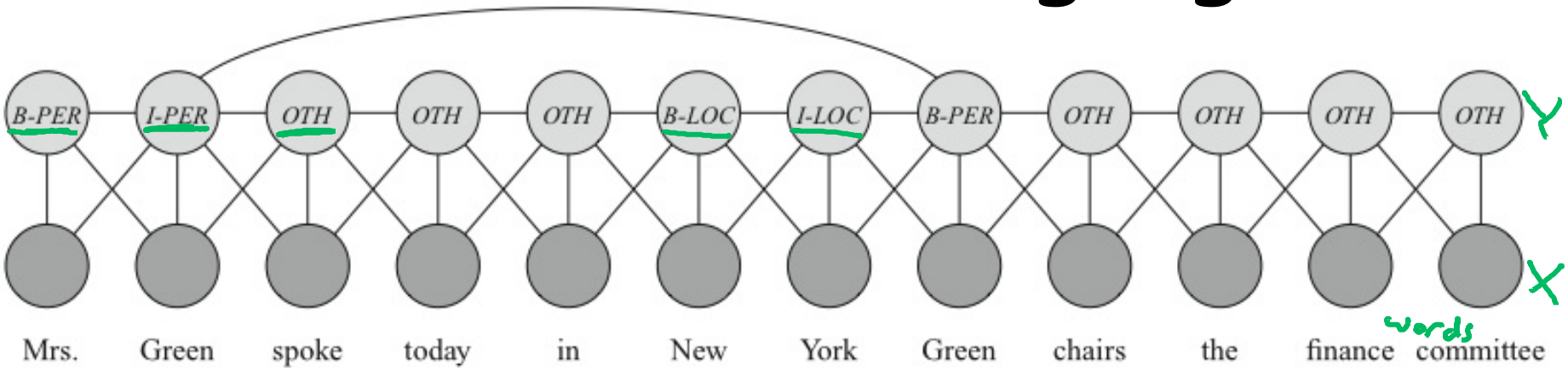
General representation

$$\phi(X_1, X_2) = \exp(-\sum_{kl} w_{kl} f_{ij}^{kl}(X_1, X_2))$$

$$w_{kl} = -\log a_{kl}$$

$\exp(-w_{00})$ when $X_1=0, X_2=0$
$\exp(-w_{0,1})$ when $X_1=0, X_2=1$

Features for Language



Features: word capitalized, word in atlas or name list, previous word is "Mrs", next word is "Times", ...

$f(Y_i, X_i) = 1\{Y_i = \text{B\_person}, X_i \text{ is capitalized}\}$

$1\{Y_i = \text{B\_loc}, X_i \text{ appears in Atlas}\}$

Ising Model

pairwise

$$E(x_1, \ldots, x_n) = -\sum_{i<j} w_{i,j} x_i x_j - \sum_i u_i x_i$$

same/different

joint spins

$x_i \in \{-1, +1\}$ $f_{i,j}(X_i, X_j) = X_i \cdot X_j$

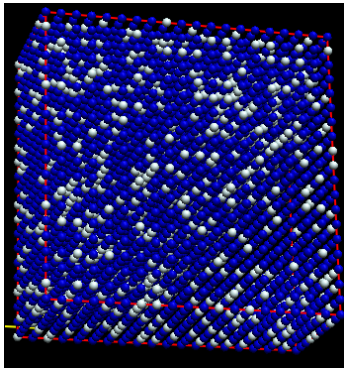
$$P(\boldsymbol{X}) \propto e^{-\frac{1}{T} E(\boldsymbol{X})}$$

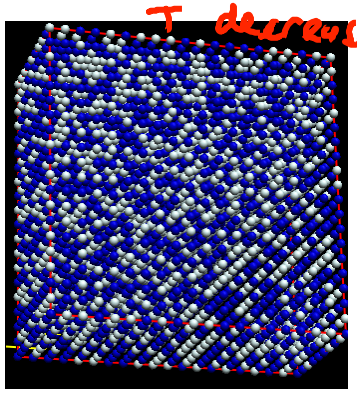
$\frac{u_i}{T} \to 0$

T grows

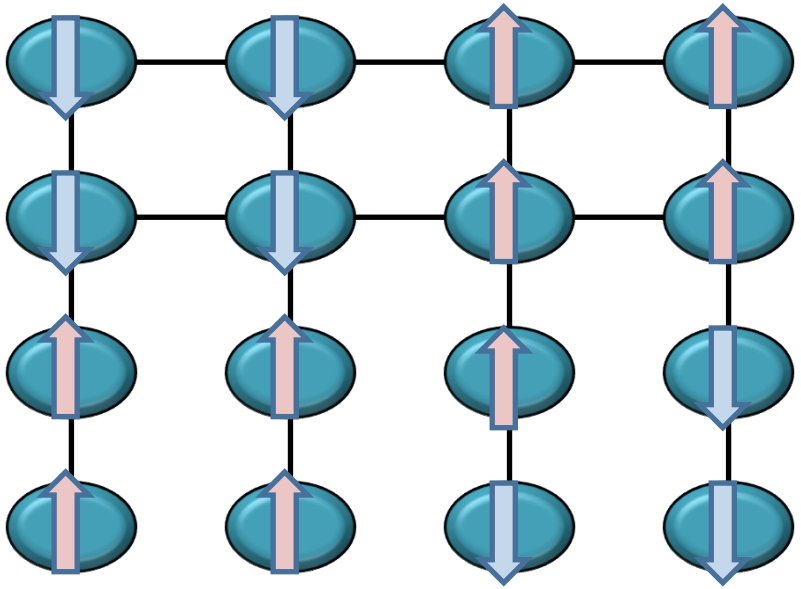
T decreases

low





high T



Daphne Koller

Metric MRFs

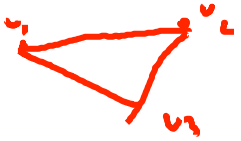
- All X_i take values in label space V

want X_i and X_j to take "similar" values

- Distance function $\mu : V \times V \rightarrow R^+$
  - Reflexivity: $\mu(v,v) = 0$ for all v
  - Symmetry: $\mu(v_1,v_2) = \mu(v_2,v_1)$ for all v_1, v_2
  - Triangle inequality: $\mu(v_1,v_2) \leq \mu(v_1,v_3) + \mu(v_3, v_2)$ for all v_1, v_2, v_3

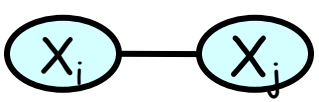
metric

semi metric



Metric MRFs

- All X_i take values in label space V



want X_i and X_j to take "similar" values

- Distance function $\mu : V \times V \rightarrow R$

$$f_{i,j}(X_i, X_j) = \mu(X_i, X_j)$$

$$\exp\left(-w_{ij} f_{ij}(X_i, X_j)\right) \qquad w_{ij} > 0$$

lower distance
higher
higher probability

w_{ij}

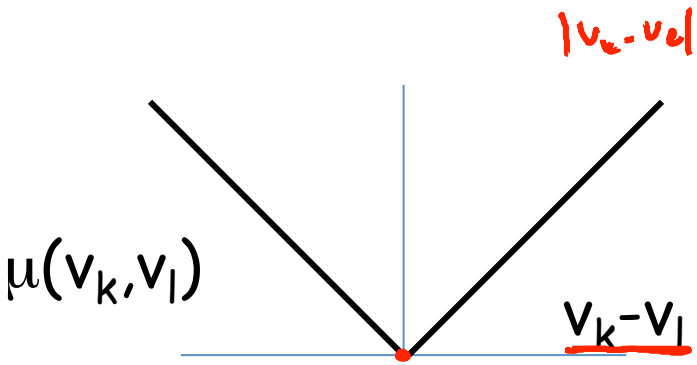
values of X_i and X_j far in μ → lower probability

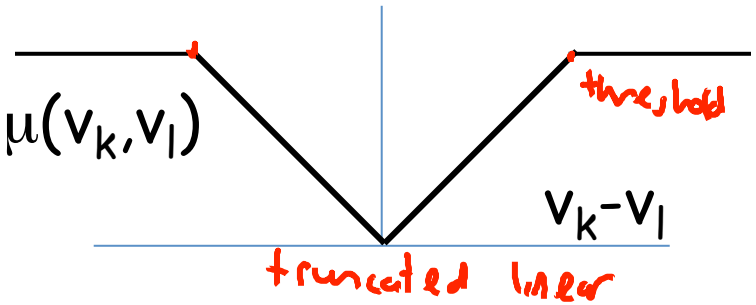
Daphne Koller

Metric MRF Examples

$$\mu(v_k,v_l) = \begin{cases} 0 & v_k = v_l \\ 1 & \text{otherwise} \end{cases}$$

$$\begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

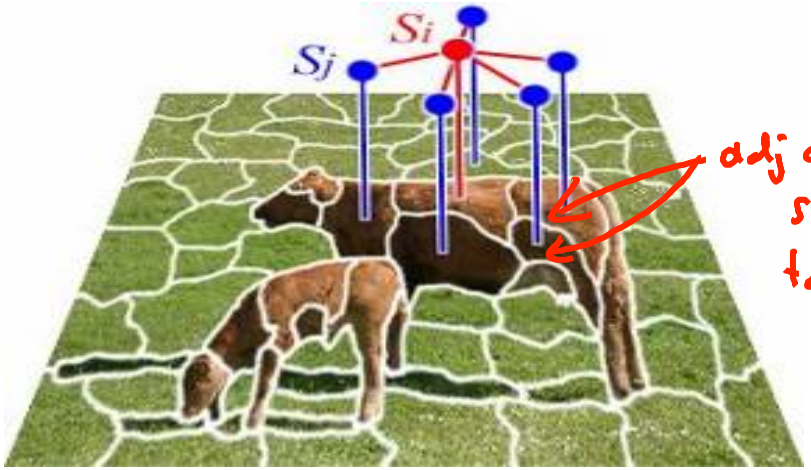




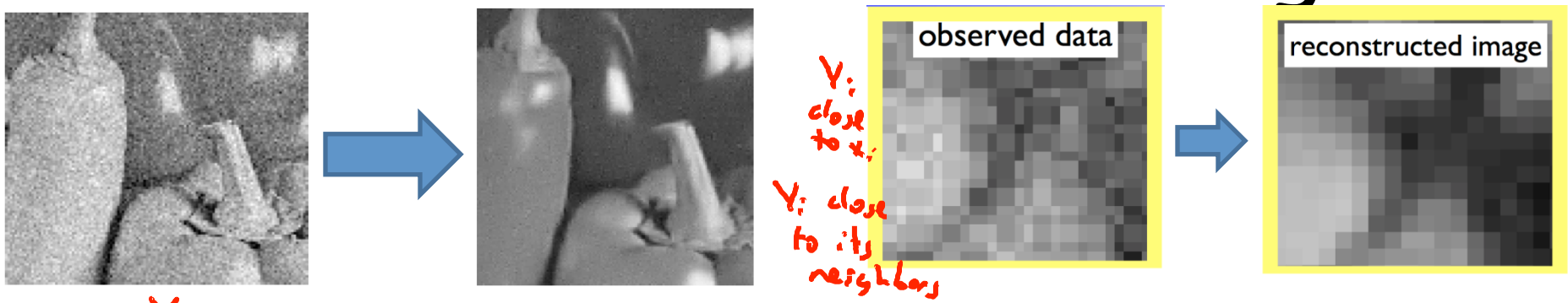
Metric MRF: Segmentation

$$\mu(v_k,v_l) = \begin{cases} 0 & v_k=v_l \\ 1 & \text{otherwise} \end{cases}$$

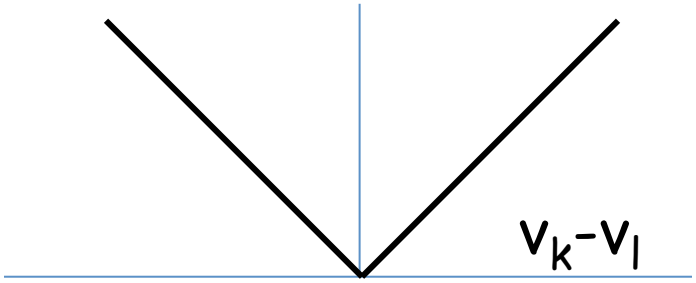
$$\begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$



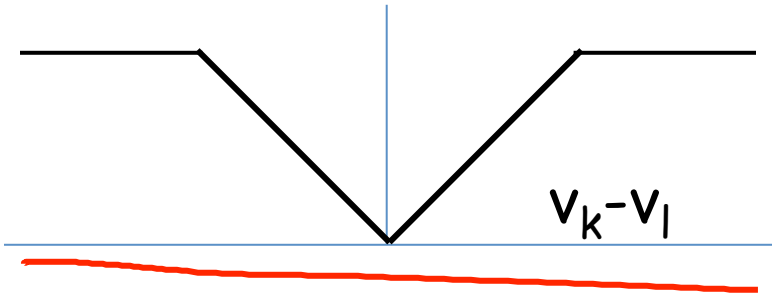
Metric MRF: Denoising



$\mu(v_k,v_l) = |v_k-v_l|$



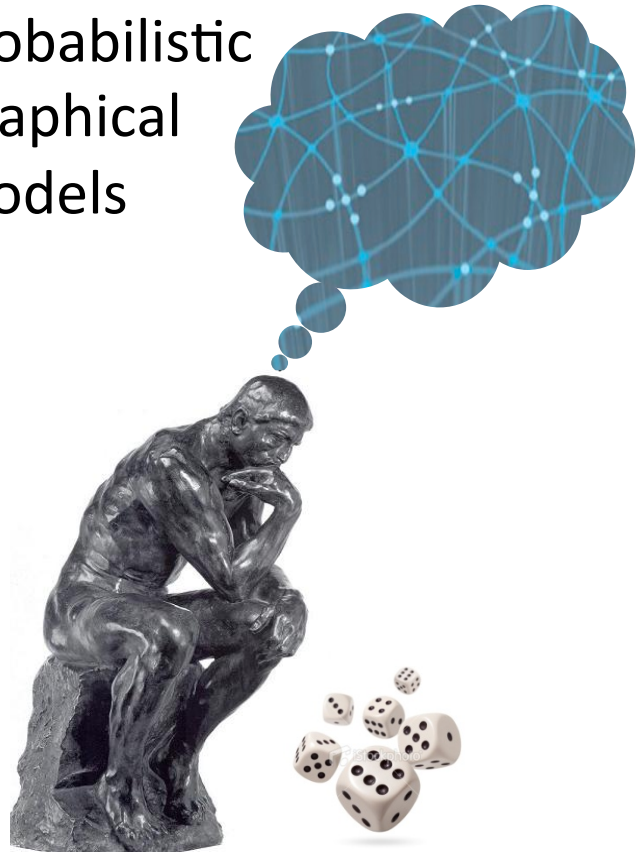
$\mu(v_k,v_l) = \max(|v_k-v_l|,d)$



Similar idea for stereo reconstruction

Daphne Koller

Probabilistic Graphical Models



Representation

Template Models

Shared Features in Log-Linear Models

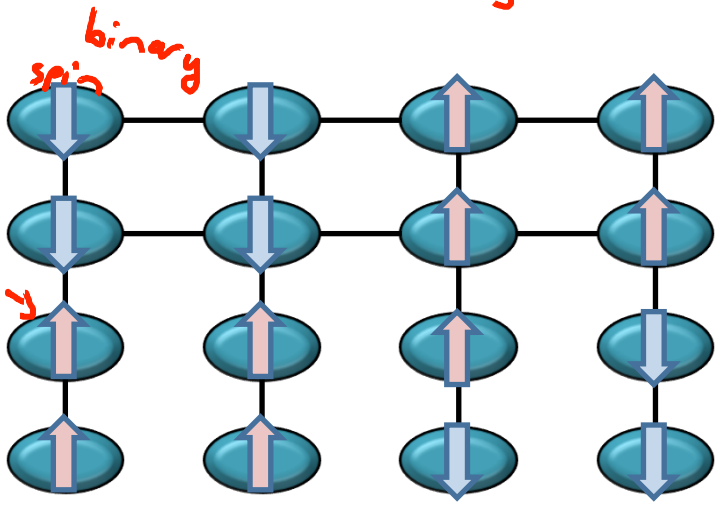
Ising Models

- In most MRFs, same feature and weight are used over many scopes

Ising Model

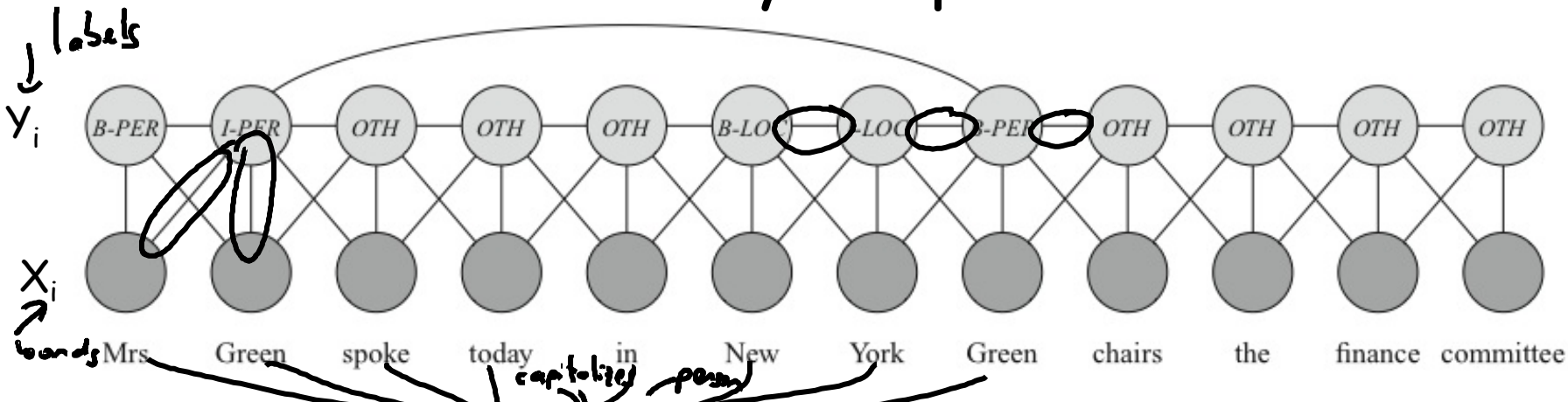
$$E(x_1, \ldots, x_n) = -\sum_{(i,j) \in \text{Edges}} w_{ij} x_i x_j - \sum_i u_i x_i$$

same weight for every adjacent pair



Natural Language Processing

- In most MRFs, same feature and weight are used over many scopes

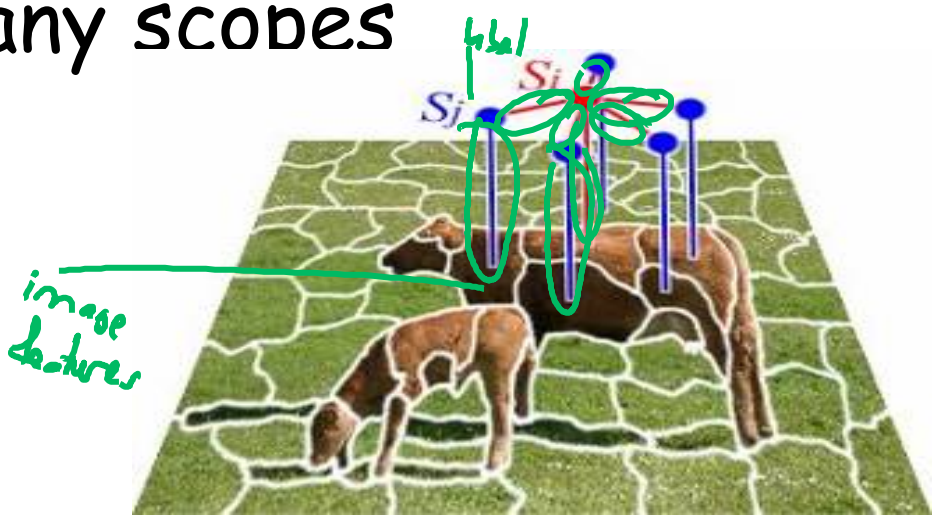


Same energy terms $w_k f_k(X_i, Y_i)$ repeat for all positions i in the sequence
Same energy terms $w_m f_m(Y_i, Y_{i+1})$ a;sp repeat for all positions i

Daphne Koller

Image Segmentation

- In most MRFs, same feature and weight are used over many scopes



Same features and weights for all superpixels in the image

Repeated Features

- Need to specify for each feature f_k a set of scopes $Scopes[f_k]$
- For each $\mathbf{D}_k \in Scopes[f_k]$ we have a term $w_k f_k(\mathbf{D}_k)$ in the energy function

$$w_k \sum_{\boldsymbol{D}_k \in \mathrm{Scopes}(f_k)} f_k(\boldsymbol{D}_k)$$

Summary

- Same feature & weight can be used for multiple subsets of variables
  - Pairs of adjacent pixels/atoms/words
  - Occurrences of same word in document
- Can provide a single template for multiple MNs
  - Different images
  - Different sentences
- Parameters and structure are reused within an MN and across different MNs
- Need to specify set of scopes for each feature