Undirected Graphical Models: Markov Random Fields

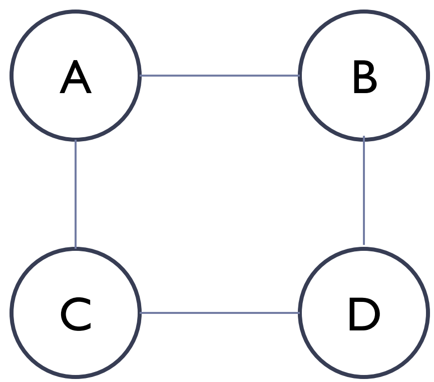
40-957 Special Topics in Artificial Intelligence: Probabilistic Graphical Models
Sharif University of Technology

Soleymani
Spring 2014

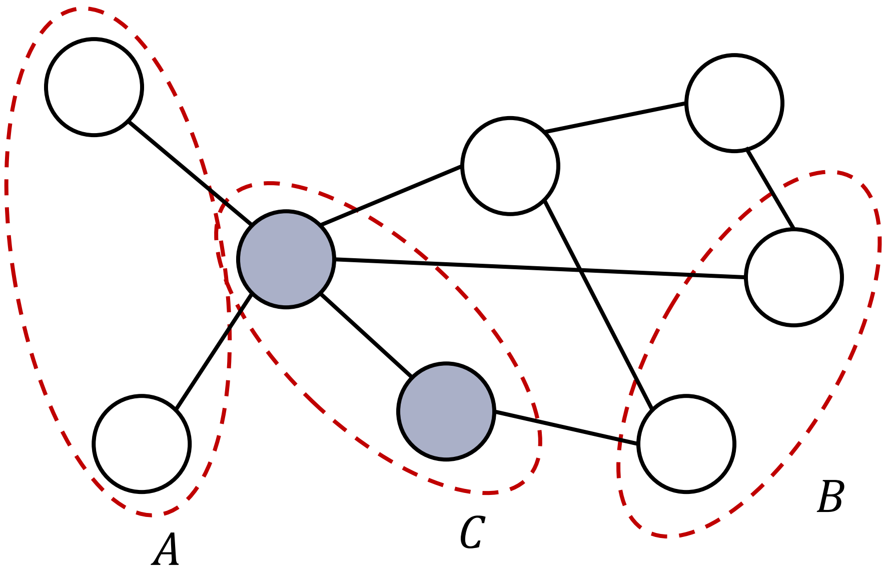
Markov Random Field

- Structure: ***undirected graph***
- Undirected edges show correlations (non-causal relationships) between variables
  - e.g., Spatial image analysis: intensity of neighboring pixels are correlated

Markov Network



Markov Random Fields (MRFs)



A path is active given C if no node in it is in C

A and B are separated given C if there is no active path between A and B given C

- Global independencies: $A \perp B | C$
  - If all paths that connect a node in A to a node in B pass through one or more nodes in set C

MRF: local independencies

- **Pairwise independencies**: $X_i \perp X_j \mid \boldsymbol{X} - \{X_i, X_j\}$

$$P\left(X_i, X_j \middle| \boldsymbol{X} - \{X_i, X_j\}\right) \\ = P\left(X_i \middle| \boldsymbol{X} - \{X_i, X_j\}\right) P\left(X_j \middle| \boldsymbol{X} - \{X_i, X_j\}\right)$$

- Markov Blanket (**local independencies**): A variable is conditionally independent of every other variables conditioned only on its neighboring nodes

$$X_i \perp \boldsymbol{X} - \{X_i\} - MB(X_i) \mid MB(X_i)$$

$$MB_H(X_i) = \{X' \in \boldsymbol{X} | (X_i, X') \in H\}$$

MRF: independencies

- Determining conditional independencies in undirected models is much easier than in directed ones

- Conditioning in undirected models can only eliminate dependencies while in directed ones can create new dependencies (v-structure)

MRF: Joint distribution

- Factor $\phi(X_1, \dots, X_k)$
  - $\phi: Val(X_1, \dots, X_k) \to \mathbb{R}$
  - Scope: $\{X_1, \dots, X_k\}$

Joint distribution parametrized by factors $\boldsymbol{\Phi} = \{\phi_1(\boldsymbol{D}_1), \dots, \phi_K(\boldsymbol{D}_K)\}$:

$$P(X_1, \dots, X_N) = \frac{1}{Z} \prod_k \phi_k(\boldsymbol{D}_k)$$

$\boldsymbol{D}_k$: the set of variables in the k-th factor

$$Z = \sum_{\boldsymbol{X}} \prod_k \phi_k(\boldsymbol{D}_k)$$

Z: normalization constant called **partition function**

MRF Factorization

- A distribution $P_{\boldsymbol{\Phi}}$ with $\boldsymbol{\Phi} = \{\phi_1(\boldsymbol{D}_1), \dots, \phi_K(\boldsymbol{D}_K)\}$ **factorizes** over an MRF H if each $\boldsymbol{D}_k$ is a **complete subgraph** of H

- If there is not a direct path between X_i and X_j then:

$$X_i \perp X_j \mid \boldsymbol{X} - \{X_i, X_j\}$$

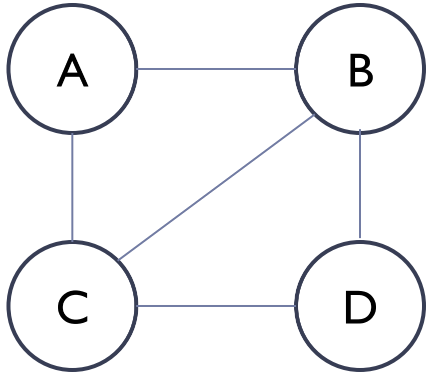
  - To hold conditional independence property, X_i and X_j that are not directly connected do not appear in the same factor in the distributions belonging to the graph

MRF Factorization: clique

- **Clique**: subsets of nodes in the graph that are fully connected (complete subgraph)
- **Maximal clique**: where no superset of the nodes in a clique are also compose a clique, the clique is maximal
- Factors are functions of the variables in the cliques
  - To reduce the number of factors we allow factors only for maximal cliques

Cliques: {A,B,C}, {B,C,D}, {A,B}, {A,C}, {B,C}, {B,D}, {C,D}, {A}, {B}, {C}, {D}

Max-cliques: {A,B,C}, {B,C,D}



MRF: Gibbs distribution

Gibbs distribution with factors $\mathbf{\Phi} = \{\phi_1(\boldsymbol{X}_{C_1}), \dots, \phi_K(\boldsymbol{X}_{C_K})\}$:

$$P_{\mathbf{\Phi}}(X_1, \dots, X_N) = \frac{1}{Z}\prod_{i=1}^{K} \phi_i(\boldsymbol{X}_{C_i})$$

$$Z = \sum_{\boldsymbol{X}} \prod_{i=1}^{K} \phi_i(\boldsymbol{X}_{C_i})$$

- $\phi_i(\boldsymbol{X}_{C_i})$: **potential function** on clique C_i
  - $\boldsymbol{X}_{C_i}$: the set of variables in the clique C_i

- **Potential functions** and **cliques** in the graph completely determine the **joint** distribution.
  - qualitative specification by potential functions

Interpretation of clique potentials

$$P(X_1, X_2, X_3) = P(X_1)P(X_2|X_1)P(X_3|X_2, X_1)$$

$$P(X_1, X_2, X_3) = P(X_1, X_2)P(X_3|X_2, X_1)$$

- All potentials cannot be marginal distributions
- All potentials cannot be conditional distributions

- A positive clique potential can be considered as general compatibility or goodness measure over values of the variables in its scope

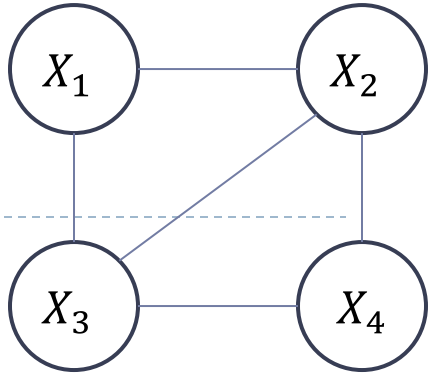
Pairwise MRF

- All of the factors on single variables or pair of variables (X_i, X_j):

$$P(\boldsymbol{X}) = \frac{1}{Z} \prod_{(X_i, X_j) \in H} \phi_{ij}(X_i, X_j) \prod_i \phi_i(X_i)$$

- Pairwise MRFs are popular (simple special case of general MRFs)
  - They consider pairwise interactions and not interactions of larger subset of variables
  - In general, do not have enough parameters to encompass the space of joint distributions

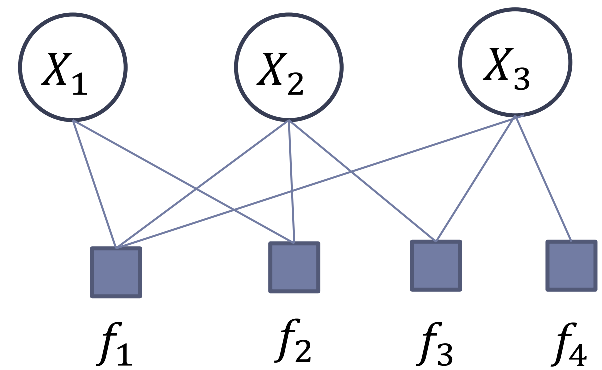
Different factorizations



- Maximal cliques:
  - $P_{\mathbf{\Phi}}(X_1, X_2, X_3, X_4) = \frac{1}{Z}\phi_{123}(X_1, X_2, X_3)\phi_{234}(X_2, X_3, X_4)$
  - $Z = \sum_{X_1, X_2, X_3, X_4} \phi_{123}(X_1, X_2, X_3)\phi_{234}(X_2, X_3, X_4)$
- Sub-cliques:
  - $P_{\mathbf{\Phi}'}(X_1, X_2, X_3, X_4)$
    $= \frac{1}{Z}\phi_{12}(X_1, X_2)\phi_{23}(X_2, X_3)\phi_{13}(X_1, X_3)\phi_{24}(X_2, X_4)\phi_{34}(X_3, X_4)$
  - $Z = \sum_{X_1, X_2, X_3, X_4} \phi_{12}(X_1, X_2)\phi_{23}(X_2, X_3)\phi_{13}(X_1, X_3)\phi_{24}(X_2, X_4)\phi_{34}(X_3, X_4)$
- Canonical representation
  - $P_{\mathbf{\Phi}'}(X_1, X_2, X_3, X_4)$
    $= \frac{1}{Z}\phi_{123}(X_1, X_2, X_3)\phi_{234}(X_2, X_3, X_4)\phi_{12}(X_1, X_2)\phi_{23}(X_2, X_3)\phi_{13}(X_1, X_3)$
    $\times\, \phi_{24}(X_2, X_4)\phi_{34}(X_3, X_4)\phi_1(X_1)\phi_2(X_2)\phi_3(X_3)\phi_4(X_4)$
  - $Z = \sum_{X_1, X_2, X_3, X_4} \phi_{123}(X_1, X_2, X_3)\phi_{234}(X_2, X_3, X_4)\phi_{12}(X_1, X_2)\phi_{23}(X_2, X_3) \times$
    $\phi_{13}(X_1, X_3)\phi_{24}(X_2, X_4)\phi_{34}(X_3, X_4)\phi_1(X_1)\phi_2(X_2)\phi_3(X_3)\phi_4(X_4)$

Factor graph

- Markov network structure doesn't fully specify the factorization of P
  - does not generally reveal all the structure in a Gibbs parameterization



- Factor graph: two kinds of nodes
  - Variable nodes
  - Factor nodes

$$P(X_1, X_2, X_3) = f_1(X_1, X_2, X_3) f_2(X_1, X_2) f_3(X_2, X_3) f_4(X_3)$$

- Factor graph is a useful structure for inference and parametrization (as we will see)

Energy function

- Constraining clique potentials to be positive could be inconvenient
  - We represent a clique potential in an unconstrained form using a real-value "energy" function

- If potential functions are strictly positive $\phi_C(\boldsymbol{X}_C) > 0$:

$$\phi_C(\boldsymbol{X}_C) = \exp\{-E_C(\boldsymbol{X}_C)\}$$

$E(\boldsymbol{X}_C)$: energy function

$$E_C(\boldsymbol{X}_C) = -\ln \phi_C(\boldsymbol{X}_C)$$

$$P(\boldsymbol{X}) = \frac{1}{Z}\exp\{-\sum_C E_C(\boldsymbol{X}_C)\}$$

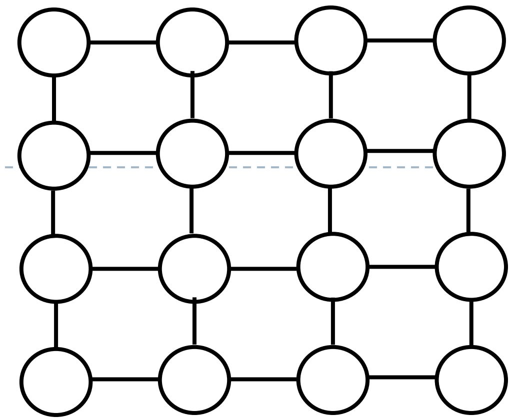
log-linear representation

Log-linear models

- Defining the energy function as a linear combination of features
- A set of m features $\{f_1(\boldsymbol{D}_1), \dots, f_m(\boldsymbol{D}_m)\}$ on complete subgraphs where $\boldsymbol{D}_i$ shows the scope of the i-th feature:
  - Scope of a feature is a complete subgraph
  - We can have different features over a sub-graph

$$P(\boldsymbol{X}) = \frac{1}{Z} \exp\left\{-\sum_{i=1}^{m} w_i f_i(\boldsymbol{D}_i)\right\}$$

Ising model



- Grid model
  - Image processing, lattice physics, etc.
  - The states of adjacent nodes are related

- Most likely joint-configurations usually correspond to a "low-energy" state

- $X_i \in \{-1,1\}$

$$P(\boldsymbol{x}) = \frac{1}{Z}\exp\left\{-\sum_i u_i x_i - \sum_{i,j} w_{ij} \underbrace{x_i x_j}_{f_{ij}(x_i, x_j) = x_i x_j}\right\}$$

Shared features in log-linear models

$$P(\boldsymbol{x}) = \frac{1}{Z}\exp\left\{-\sum_i u_i x_i - \sum_{(i,j)\in H} w_{ij}\underbrace{x_i x_j}\right\}$$

$$f_{ij}(x_i, x_j) = f(x_i, x_j) = x_i x_j$$

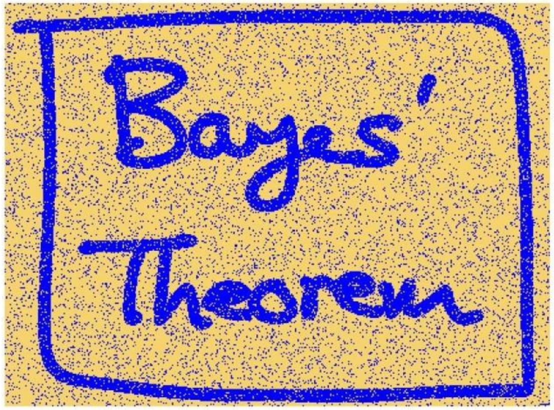
- In most practical models, same feature and weight are used over many scopes

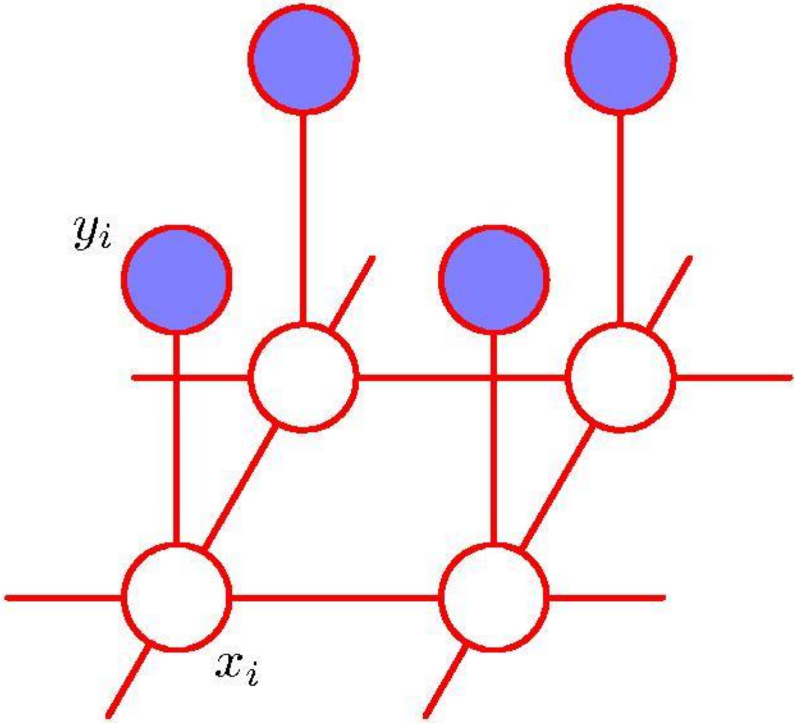
$$P(\boldsymbol{x}) = \frac{1}{Z}\exp\left\{-\sum_i u x_i - \sum_{(i,j)\in H} w\underbrace{x_i x_j}\right\}$$

$$w_{ij} = w \qquad f(x_i, x_j) = x_i x_j$$

Image denoising

- $y_i \in \{-1,1\}, i = 1, \dots, D$: array of observed noisy pixels
- $x_i \in \{-1,1\}, i = 1, \dots, D$: noise free image





[Bishop]

Image denoising

$$E(\boldsymbol{x},\boldsymbol{y}) = h\sum_i x_i - \beta \sum_{\{i,j\}\in H} x_i x_j - \eta \sum_i x_i y_i$$

$$P(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{Z}\exp\{-E(\boldsymbol{x},\boldsymbol{y})\}$$

$$\widehat{\boldsymbol{x}} = \operatorname*{argmax}_{\boldsymbol{x}} P(\boldsymbol{x}|\boldsymbol{y})$$

MPA: Most probable assignment of $\boldsymbol{x}$ variables given an evidence $\boldsymbol{y}$

Image denoising (gray-scale image)

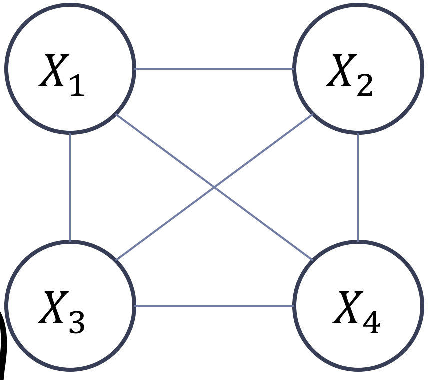
$$E(\boldsymbol{x}, \boldsymbol{y}) = -\beta \sum_{\{i,j\}\in H} \underbrace{\min(\|x_i - x_j\|_2, d)}_{f_{ij}(x_i, x_j) = f(x_i, x_j) = \min(\|x_i - x_j\|_2, d)} - \eta \sum_i \|x_i - y_j\|_2$$

$$\widehat{\boldsymbol{x}} = \operatorname*{argmax}_{\boldsymbol{x}} \frac{1}{Z} \exp\{-E(\boldsymbol{x}, \boldsymbol{y})\}$$

MPE: Most probable explanation of $\boldsymbol{x}$ variables given an evidence $\boldsymbol{y}$

Boltzmann machine

- A fully connected graph with pairwise (edge) potentials on binary-valued nodes (i.e., $X_i \in \{0,1\}$ or $X_i \in \{-1,1\}$)



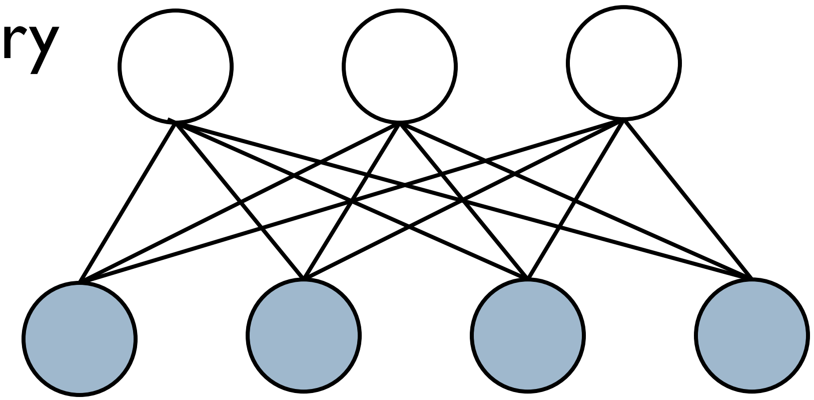
$$P(x_1, \dots, x_n) = \frac{1}{Z} \exp\left\{-\sum_{i,j>i} \phi_{ij}(x_i, x_j)\right\}$$

$$= \frac{1}{Z} \exp\left\{-\sum_{i,j>i} w_{ij}\, x_i x_j - \sum_{i} b_i\, x_i\right\}$$

$$P(x_i | x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = \sigma(-\sum_{j>i} w_{ij}\, x_j - b_i)$$

Restricted Boltzmann Machine (RBM)

- Harmonium (Smolensky –1986)
- RBM (Hinton-2002): binary
  - Efficient learning



$$P(\boldsymbol{v}, \boldsymbol{h}) = \frac{1}{Z} \exp\left\{ \sum_i a_i h_i + \sum_j b_j v_j + \sum_{i,j} w_{i,j} h_i v_j \right\}$$

Restricted Boltzmann machine

$$P(\boldsymbol{h}|\boldsymbol{v}) = \prod_i P(h_i|\boldsymbol{v})$$

$$P(\boldsymbol{v}|\boldsymbol{h}) = \prod_i P(v_i|\boldsymbol{h})$$

$$P(h_i = 1|\boldsymbol{v}) = \sigma\left(a_i + \sum_j w_{ij} v_j\right)$$

$$P\left(v_j = 1\middle|\boldsymbol{h}\right) = \sigma\left(b_i + \sum_i w_{ij} h_i\right)$$

MRF: global independencies

- Independencies encoded by H

$$I(H) = \{(\boldsymbol{X} \perp \boldsymbol{Y} | \boldsymbol{Z}) : \mathrm{sep}_H(\boldsymbol{X}, \boldsymbol{Y} | \boldsymbol{Z})\}$$

- If P satisfies $I(H)$, we say that H is an I-map (independency map) of P
  - $I(H) \subseteq I(P)$ where $I(P) = \{\boldsymbol{X}, \boldsymbol{Y} | \boldsymbol{Z} : P \vDash (\boldsymbol{X} \perp \boldsymbol{Y} | \boldsymbol{Z})\}$

Factorization & Independence

- Factorization ⇒ Independence (soundness of separation criterion)
  - **Theorem:** If P factorizes over H, and $\text{sep}_H(\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{Z})$ then P satisfies $\boldsymbol{X} \perp \boldsymbol{Y}|\boldsymbol{Z}$ (i.e., H is an I-map of P)

- Independence ⇒ Factorization
  - **Theorem** (Hammersley Clifford): For a positive distribution P, if P satisfies $I(H) = \{(\boldsymbol{X} \perp \boldsymbol{Y}|\boldsymbol{Z}) : \text{sep}_H(\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{Z})\}$ then P factorizes over H

Let **X,Y,Z** be three disjoint sets of variables:
P⊨(**X**, **Y** | **Z**) iff P(**X,Y,Z**)=ψ(**X,Z**) ψ(**Y,Z**)

Factorization & Independence

- Two equivalent views of graph structure for **positive distributions**:
  - Factorization: H allows P to be represented factorized on cliques
  - I-map: Independencies encoded by H hold also in P

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

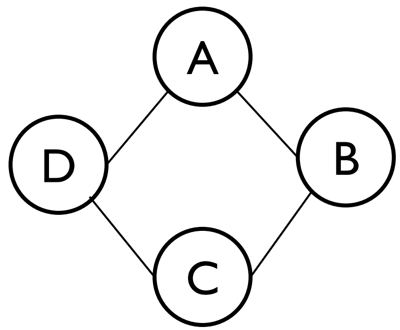
Relationship between local and global Markov properties

- If $P \models I_l(H)$ then $P \models I_p(H)$.
- If $P \models I(H)$ then $P \models I_l(H)$.

- For a **positive distribution** P, the following three statements are equivalent:
  - $P \models I_p(H)$
  - $P \models I_l(H)$
  - $P \models I(H)$

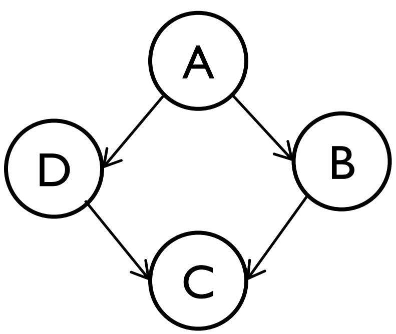
Loop of at least 4 nodes without chord has no equivalent in BNs

- Is there a BN that is a perfect map for this MN?



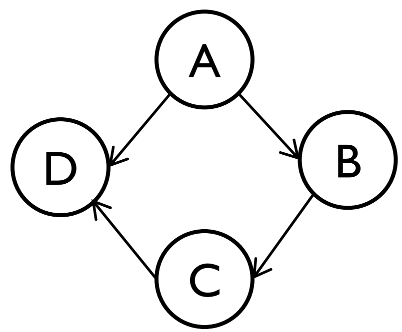
$A \perp C | \{B, D\}$

$B \perp D | \{A, C\}$



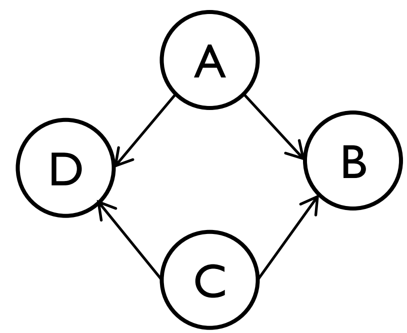
$A \perp C | \{B, D\}$

$B \perp D | \{A, C\}$



$B \perp D | \{A, C\}$

$A \perp C | \{B, D\}$

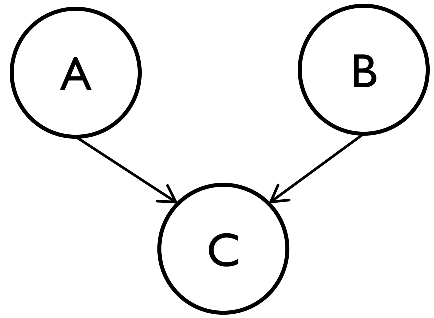


$B \perp D | \{A, C\}$

$A \perp C | \{B, D\}$

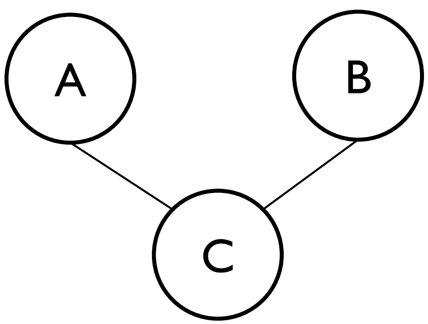
V-structure has no equivalent in MNs

- Is there an MN that is a perfect I-map of this BN?



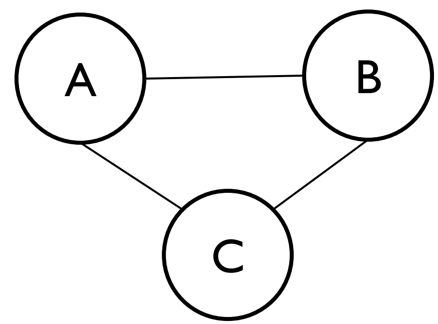
$A \perp B$

$A \perp B | C$ ✗



$A \perp B$ ✗

$A \perp B | C$

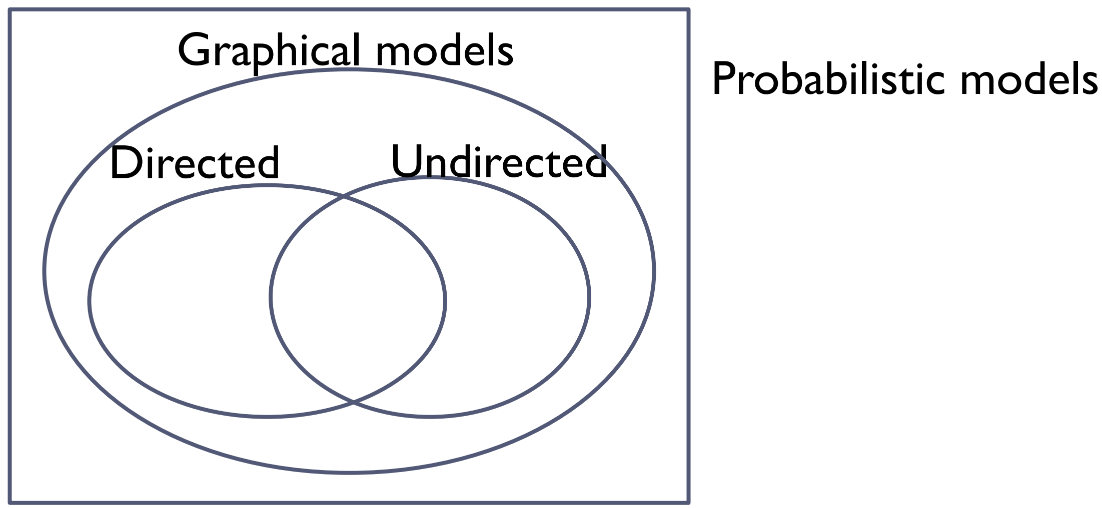


$A \perp B$ ✗

$A \perp B | C$ ✗

Perfect map of a distribution

- Not every distribution has a MN perfect map
- Not every distribution has a BN perfect map



Minimal I-map

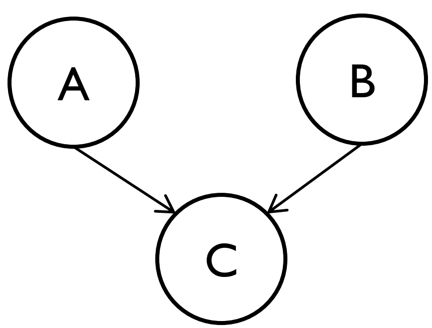
- Since we may not find an MN that is a perfect map of a BN and vice versa, we study the minimal I-map property

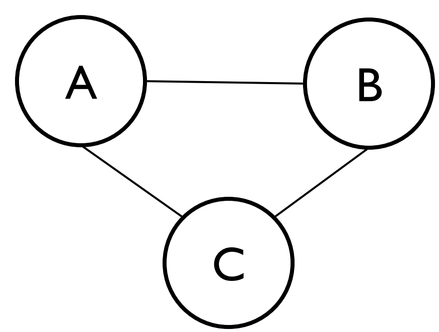
- H is a minimal I-map for G if
  - $I(H) \subseteq I(G)$
  - Removal of a single edge in H renders it is not an I-map of G

Minimal I-maps: from DAGs to MNs

- The **moral graph** $M(G)$ of a DAG G is an undirected graph that contains an undirected edge between X and Y if:
  - there is a directed edge between them in either direction
  - X and Y are parents of the same node

- Moralization turns a node and its parent into a fully connected subgraph





Minimal I-maps: from DAGs to MNs

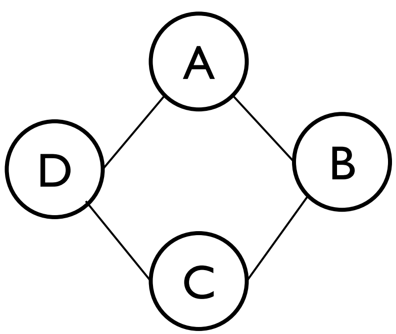
- The moral graph $M(G)$ of a DAG G is a minimal I-map for G
  - The moral graph loses some independence information
    - But all independencies in the moral graph are also satisfied in G

- If a DAG G is "moral", then its moralized graph $M(G)$ is a perfect I-map of G.

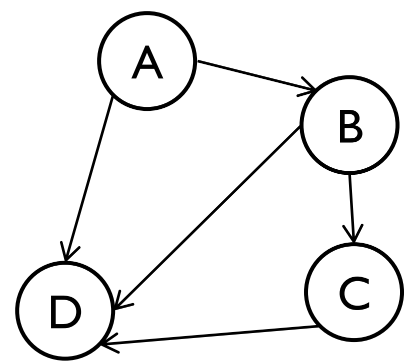
Minimal I-maps: from MNs to DAGs

- If G is a BN that is minimal I-map for an MN, then G can have no immoralities.

- $\Rightarrow$ Let G be a minimal I-map for an MN then it is **chordal**
  - Any BN that is I-map for an MN must add triangulating edges into the graph



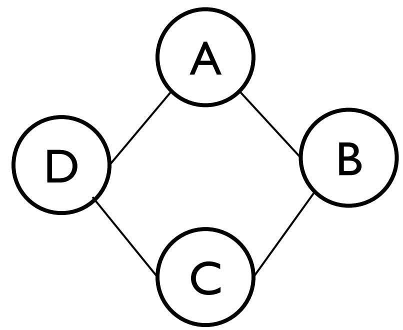
An undirected graph is chordal if any loop with more than three nodes has a chord



G is a minimal I-map of the left MN

Perfect I-map

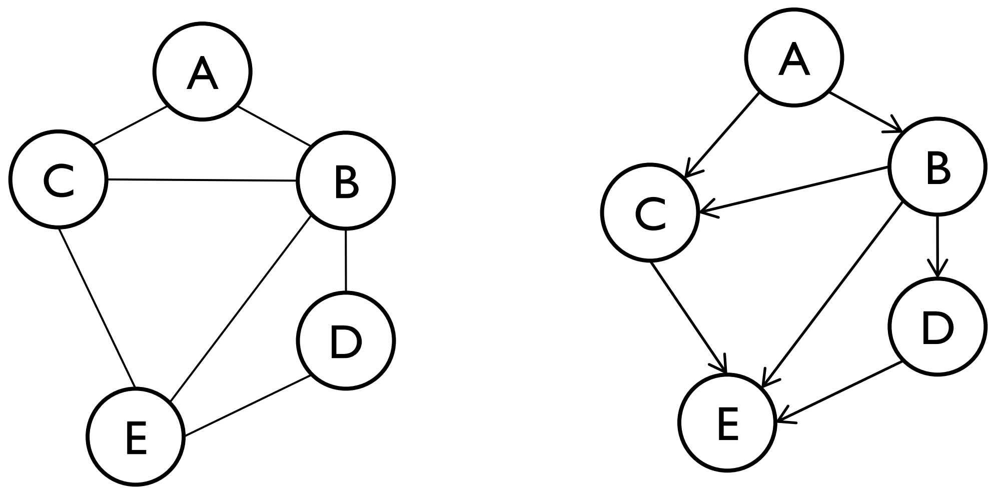
- Theorem: Let H be a non-chordal MN. Then there is no BN that is a perfect I-map for H.



- $\Rightarrow$ If the independencies in an MN can be represented via a BN then the MN graph is **chordal**

Perfect I-map

- Theorem: Let H be a chordal MN. Then there exists a DAG G that is a perfect I-map for H



- $\Rightarrow$ The independencies in a graph can be represented in both type of models **if and only if** the graph is **chordal**

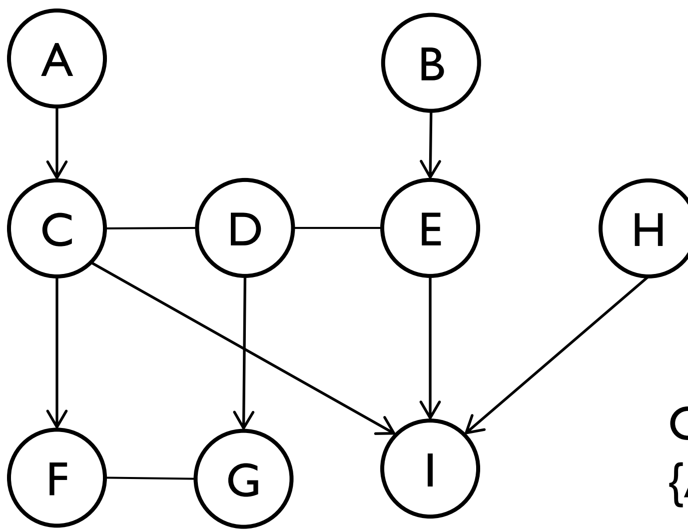
Relationship between BNs and MNs

- Directed and undirected models represent different families of independence assumptions
  - Under certain condition, they can be converted to each other
  - Chordal graphs can be represented in both BNs and MNs

- For inference, we can use a single representation for both types of these models
  - simpler design and analysis of the inference algorithm

Partially Directed Acyclic Graphs (PDAGs)

- Superset of both directed and undirected graphs
- PDAGs are also called **chain graphs**
  - Nodes can be disjointly partitioned into several chain components
  - An edge within the same chain component must be undirected
  - An edge between two nodes in different chain components must be directed



Chain components:
{A}, {B}, {C,D,E},{F,G},{H}, {I}

[Koller, PGM book]

Conditional Random Field (CRF)

- Undirected graph H with nodes $\boldsymbol{X} \cup \boldsymbol{Y}$
  - $\boldsymbol{X}$: observed variables
  - $\boldsymbol{Y}$: target variables
- Consider factors $\boldsymbol{\Phi} = \{\phi_1(\boldsymbol{D}_1), \ldots, \phi_K(\boldsymbol{D}_K)\}$ where each $\boldsymbol{D}_i \nsubseteq \boldsymbol{X}$:

$$P(\boldsymbol{Y}|\boldsymbol{X}) = \frac{1}{Z(\boldsymbol{X})} \tilde{P}(\boldsymbol{Y}, \boldsymbol{X})$$

$$\tilde{P}(\boldsymbol{Y}, \boldsymbol{X}) = \prod_{i=1}^{K} \phi_i(\boldsymbol{D}_i)$$

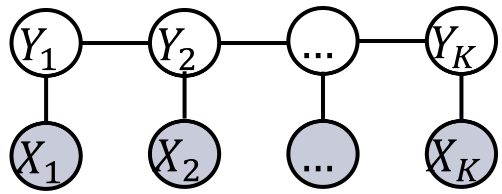
$$Z(\boldsymbol{X}) = \sum_{\boldsymbol{Y}} \tilde{P}(\boldsymbol{Y}, \boldsymbol{X})$$

- Nodes are connected by edge in H whenever they appear together in the scope of some factor

CRF: discriminative model

- Discriminative approach for labeling
- Conditional probability $P(\boldsymbol{Y}|\boldsymbol{X})$ rather than joint probability $P(\boldsymbol{Y},\boldsymbol{X})$
  - The probability of a transition between labels may depend on past and future observations
  - CRF is based on the conditional probability of label sequence given observation sequence
  - Allow arbitrary dependency between features on the observation sequence
    - As opposed to independence assumptions in generative models

Linear-chain CRF



Linear-chain CRF

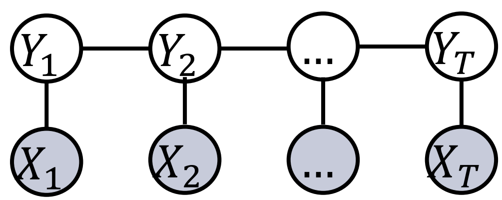
$$P(\boldsymbol{Y}|\boldsymbol{X}) = \frac{1}{Z(\boldsymbol{X})}\tilde{P}(\boldsymbol{Y},\boldsymbol{X})$$

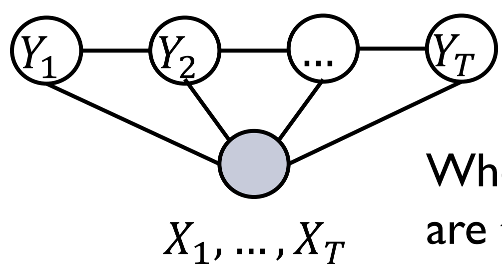
$$\tilde{P}(\boldsymbol{Y},\boldsymbol{X}) = \prod_{i=1}^{K}\phi(Y_i, Y_{i+1})\prod_{i=1}^{K}\phi(Y_i, X_i)$$

$$Z(\boldsymbol{X}) = \sum_{\boldsymbol{Y}}\tilde{P}(\boldsymbol{Y},\boldsymbol{X})$$

CRF

- CRF does not model the distribution over the observations
  - Dependencies between observed variables may be quite complex or poorly understood but we don't worry about modeling them





When labeling X_i future observations are taken into account

CRF: logistic model
Naïve Markov model

- Binary variables $\boldsymbol{Y} = \{Y\}$, $\boldsymbol{X} = \{X_1, \dots, X_m\}$

$$f_i(Y, X_i) = I(X_i = 1, Y = 1)$$

$$f_0(Y) = I(Y = 1)$$

$$\tilde{P}(Y, \boldsymbol{X}) = \exp\left\{w_0 f_0(Y) + \sum_{i=1}^{m} w_i f_i(Y, X_i)\right\}$$

$$\tilde{P}(Y = 1, \boldsymbol{X}) = \exp\left\{w_0 + \sum_{i=1}^{m} w_i X_i\right\}$$

$$\tilde{P}(Y = 0, \boldsymbol{X}) = \exp\{0\} = 1$$

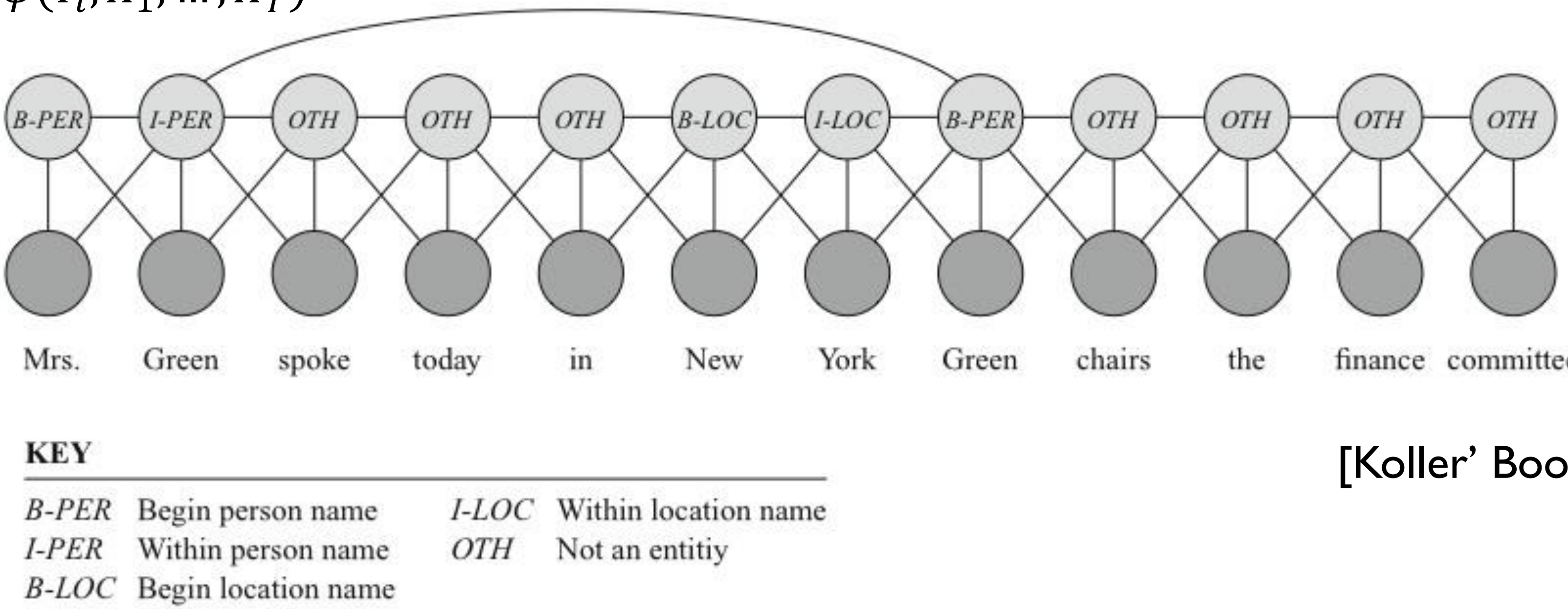
$$P(Y = 1|\boldsymbol{X}) = \sigma(w_0 + \sum_{i=1}^{m} w_i X_i)$$

Number of parameters is linear

CRF: Named Entity Recognition

$\phi(Y_i, Y_{i+1})$

$\phi(Y_i, X_1, \ldots, X_T)$



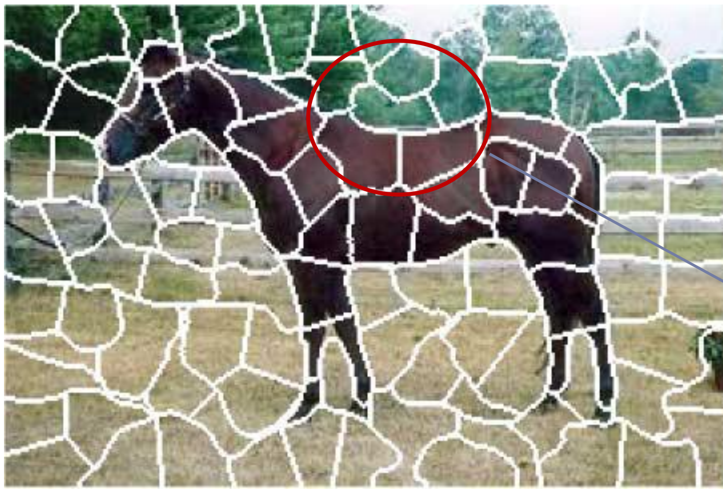
[Koller' Book]

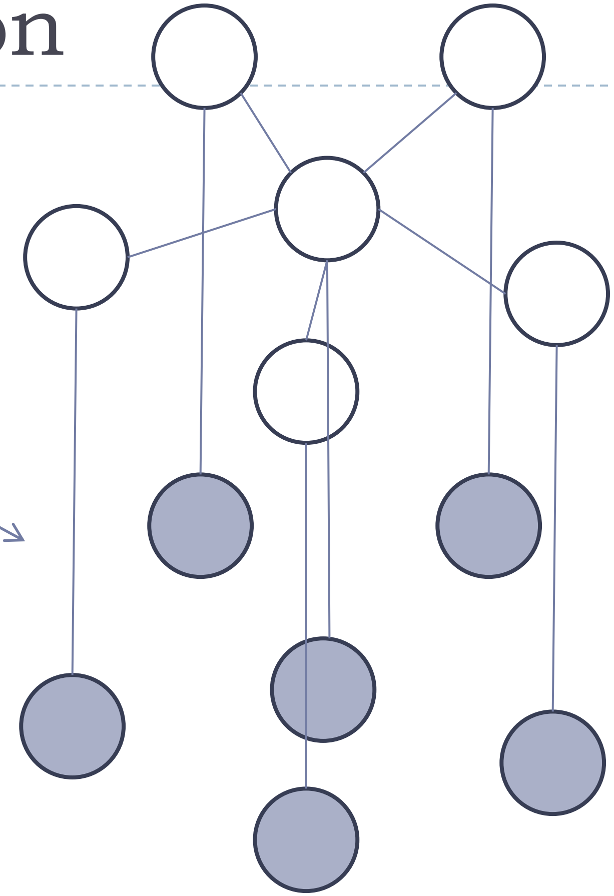
- Features: word capitalized, word in atlas of locations, previous word is "Mrs", next word is "Times", …

CRF: Image segmentation

- A node Y_i for the label of each super-pixel
  - $Val(Y_i) = \{1,2, \dots, K\}$ (i.e., grass, sky, water, …)
- An edge between Y_i and Y_j where the corresponding super-pixels share a boundary
- A node X_i for the features (e.g., color, texture, locatiob) of each super-pixel

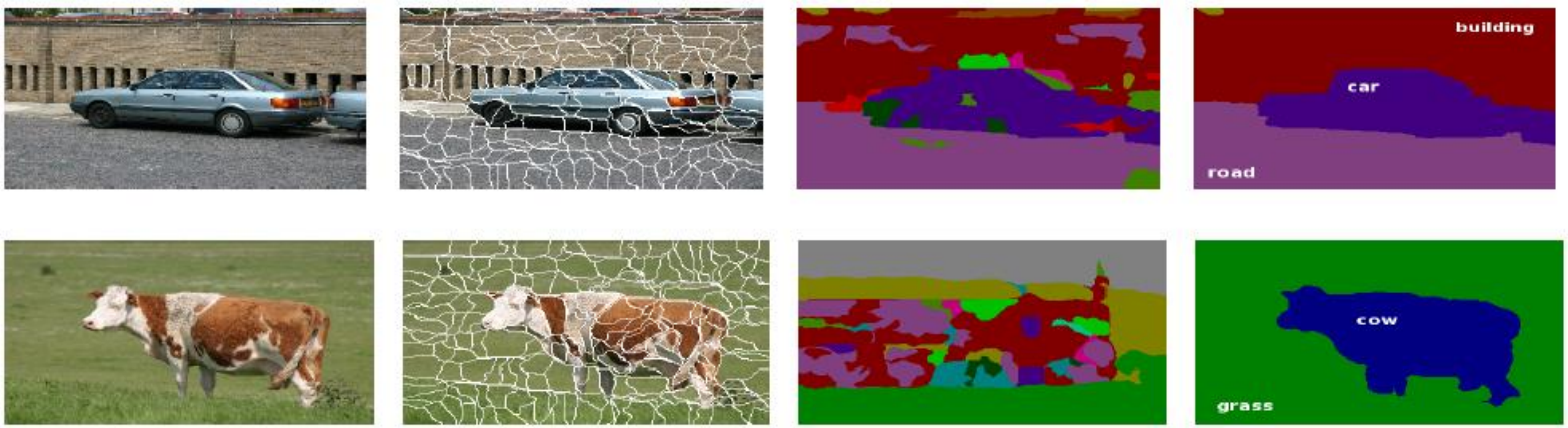
CRF: Image segmentation





- Simple: $\phi(Y_i, Y_j) = \lambda I(Y_i \neq Y_j)$
- More complex:
  - e.g., horse adjacent to vegetation than water
  - depends on the relative pixel location, e.g., water below vegetation, sky above every thing

CRF: Image segmentation



[Koller’s Book]