CS 583: Probabilistic Graphical Models

Markov Networks

Mustafa Bilgic

http://www.cs.iit.edu/~mbilgic

https://twitter.com/bilgicm

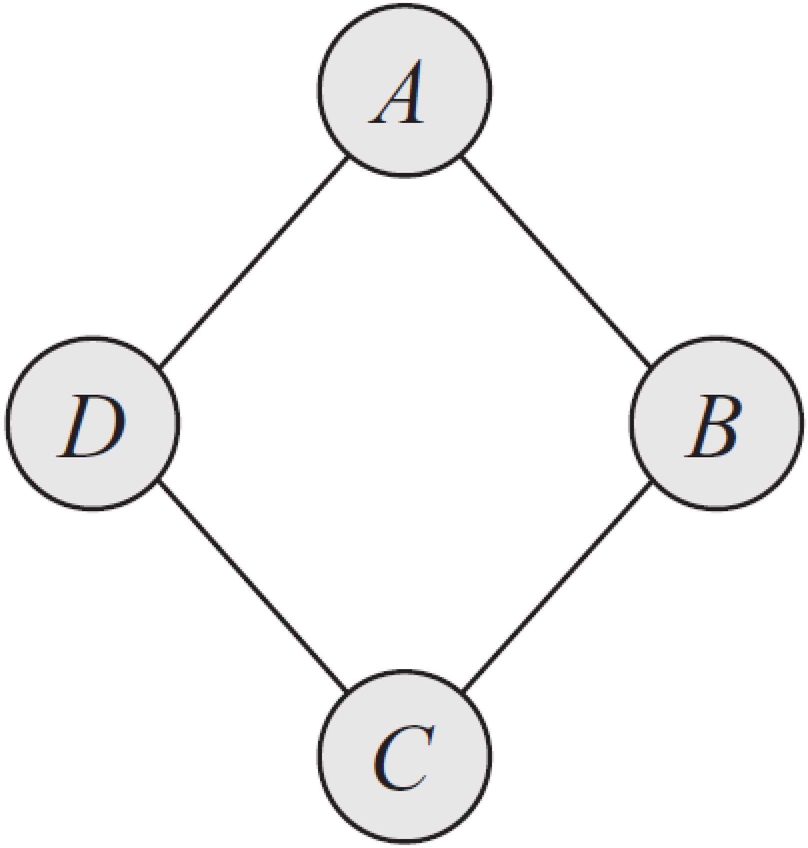
Motivation for Markov Networks

- There are distributions that cannot be represented Bayesian networks (and vice versa)
- Guaranteeing acyclicity can be hard

An example

- We'd like a graph where
  - $A \perp C \mid B, D$
  - $B \perp D \mid A, C$
- (A, B), (B, C), (C, D), and (D, A) are correlated but no causal direction exists
- Alice and Charles pair and Bob and Debbie pair do not talk to each other directly
- Alice and Bob, Bob and Charles, and Alice and Debbie pairs agree most of the time, and Charles and Debbie pair disagrees most of the time

EXAMPLE



GRAPHS

- Structure
- Parameters
- The joint distribution
- Independencies

Bayesian Networks

- Structure
  - Directed acyclic graph
- Parameters
  - Conditional probability distributions
- The joint distribution
  - $P(\mathcal{X}) = \prod P(X_i \mid Pa(X_i))$
- Independencies
  - $X_i \perp ND(X_i) \mid Pa(X_i)$
  - D-separation

Markov Networks

- Structure
  - ?
- Parameters
  - ?
- The joint distribution
  - ?
- Independencies
  - ?

Markov Networks

- Structure
  - Undirected graphs
- Parameters
  - ?
- The joint distribution
  - ?
- Independencies
  - ?

Independencies in Markov networks

1. Separation
   - $\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}$ if $\mathbf{X}$ and $\mathbf{Y}$ are separated in $\mathcal{H}$ given $\mathbf{Z}$
2. Pairwise independencies
   - $X \perp Y \mid \mathcal{X} \setminus \{X, Y\}$
3. Local independencies
   - $X \perp \mathcal{X} \setminus \mathrm{MB}(X) \mid \mathrm{MB}(X)$, where MB stands for Markov Blanket. *Markov Blanket* of a variable X in a Markov network $\mathcal{H}$ is its neighbors.

Markov Networks

- Structure
  - Undirected graphs
- Parameters
  - ?
- The joint distribution
  - ?
- Independencies
  - Separation
  - Pairwise independencies
  - Local independencies

Markov Networks

- Structure
  - Undirected graphs
- Parameters
  - Conditional Probability Distributions?
- The joint distribution
  - ?
- Independencies
  - Separation
  - Pairwise independencies
  - Local independencies

Conditioned on the neighbors?

- Consider the simple graph of $A - B$
- Can we say $P(A, B) = P(A)P(B \mid A)$
  - $P(A, B) \underset{\propto}{=} P(A \mid B)P(B \mid A)$? $P(B)P(A \mid B)$

Marginals on the (Maximal) Cliques?

- Consider the simple graph of $A – B$
- Can we say
  - $P(A, B) = P(A, B)$?
- Now consider $A – B – C$
- Can we say
  - $P(A, B, C) = P(A, B)\ P(B, C)$?
- How would you parameterize Markov Networks?

PARAMETERIZATION

- Parameterization is perhaps the least intuitive concept about MNs
- Bayesian networks
  - $P(X_i \mid Pa(X_i))$
- Markov networks
  - Cannot use probability distributions directly, but
  - MNs provide more flexibility in the parameterization

Factors

- Let **D** be a set of random variables
- **Definition:** A *factor* ϕ is a function from Val(**D**) to R.
- A factor is nonnegative if all entries are nonnegative
- The *scope* of factor, denoted as, *Scope*[ϕ], is the set of variables **D** it is associated with

An example

- Structure: A – B – C
- Factors: $\phi(A, B)$ and $\phi(B, C)$
- Remember the factors are functions from D to R.
- How can we represent the joint P(A, B, C) using factors?

Gibbs Distribution

- **Definition:** A distribution P is a *Gibbs distribution* parameterized by a set of factors $\Phi = \{\phi(\mathbf{D}_1), \ldots, \phi(\mathbf{D}_k)\}$ if it is defined as follows:

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

What is Z?

$\phi(\mathbf{D}_i)$ are factors, but what are $\mathbf{D}_i$?

Can you relate this to Bayesian Network parameterization?

MARKOV NETWORK FACTORIZATION

- We say that a distribution P with $\Phi = \{\phi(\mathbf{D}_1), \ldots, \phi(\mathbf{D}_k)\}$ *factorizes* over a Markov network $\mathcal{H}$ if each $\mathbf{D}_i$ (i=1, …, k) is a complete subgraph of $\mathcal{H}$
- The factors $\phi(\mathbf{D}_i)$ are called the *clique potentials*
- $\mathbf{D}_i$ can be maximal cliques but they do not have to be

A – B – C

A, B, C

$\frac{1}{Z} \phi(A,B) \phi(B,C)$

EXAMPLE

A	B	ϕ(A,B)
T	T	0.5
T	F	0.1
F	T	0.1
F	F	0.3

B	C	ϕ(B,C)
T	T	0.1
T	F	0.2
F	T	0.6
F	F	0.1

A	B	C	ϕ(A,B)*ϕ(B,C)	P(A,B,C)
T	T	T	0.05	0.11
T	T	F	0.10	0.22
T	F	T	0.06	0.13
T	F	F	0.01	0.02
F	T	T	0.01	0.02
F	T	F	0.02	0.04
F	F	T	0.18	0.39
F	F	F	0.03	0.07
		Z	0.46	1.00

A	B	P(A,B)
T	T	0.33
T	F	0.15
F	T	0.07
F	F	0.46

B	C	P(B,C)
T	T	0.13
T	F	0.26
F	T	0.52
F	F	0.09

Is ϕ(A, B) = P(A, B)?

What is the most likely assignment to A, B according to ϕ(A, B)? How about P(A, B)?

A	P(A)
T	0.48
F	0.52

B	P(B)
T	0.39
F	0.61

C	P(C)
T	0.65
F	0.35

Example

A	B	ϕ(A,B)
T	T	5
T	F	1
F	T	1
F	F	3

B	C	ϕ(B,C)
T	T	1
T	F	2
F	T	6
F	F	1

A	B	C	ϕ(A,B)*ϕ(B,C)	P(A,B,C)
T	T	T	5	0.11
T	T	F	10	0.22
T	F	T	6	0.13
T	F	F	1	0.02
F	T	T	1	0.02
F	T	F	2	0.04
F	F	T	18	0.39
F	F	F	3	0.07
		Z	46	1.00

A	B	P(A,B)
T	T	0.33
T	F	0.15
F	T	0.07
F	F	0.46

B	C	P(B,C)
T	T	0.13
T	F	0.26
F	T	0.52
F	F	0.09

Multiplied all the factors by 10. What changed?

A	P(A)
T	0.48
F	0.52

B	P(B)
T	0.39
F	0.61

C	P(C)
T	0.65
F	0.35

Markov Networks

- Structure
  - Undirected graphs
- Parameters
  - Factors
- The joint distribution
  - $P(\mathcal{X}) = 1/Z \prod \phi(\mathbf{D}_i)$
- Independencies
  - Separation
  - Pairwise independencies
  - Local independencies

Parameterization

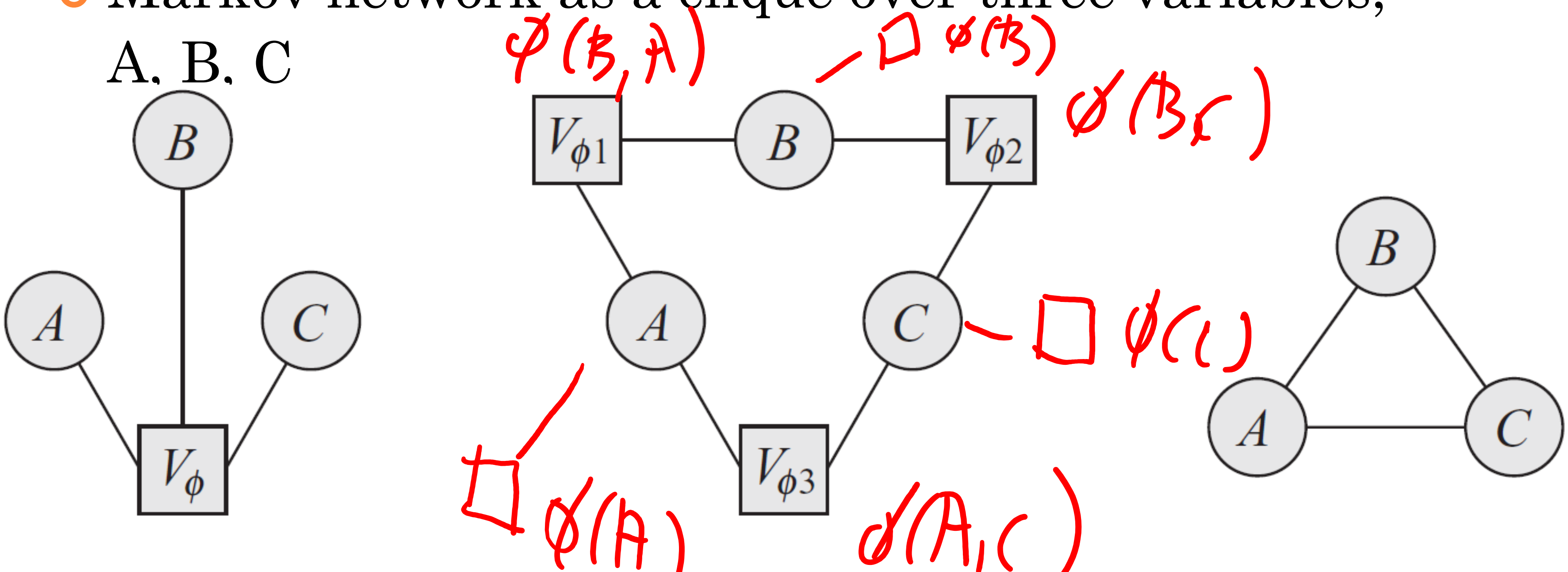
- Factors over maximal cliques
- Pairwise Markov random fields
  - Factors over nodes, and
  - Factors over connected pairs (i.e., edges)
- Pairwise Markov random fields do not introduce additional independencies, however,
  - The number of parameters is quadratic instead of exponential, but, of course,
  - The sets of distributions that can be represented over maximal cliques and pairwise interactions are not the same

FACTOR GRAPHS

- **Definition**: A *factor graph* $\mathcal{F}$ is an undirected graph containing two types of nodes
  - Random variables (ovals)
  - Factor nodes (squares).
- $\mathcal{F}$ contains edges between ovals and squares.
- $\mathcal{F}$ is parameterized by a set of factors, where each factor node (square) is associated with precisely one factor whose scope is the square's neighbor ovals.

Factor graph example

- Markov network as a clique over three variables, A, B, C



(a) (b) (c)

$\phi(A,B,C)$

How would you represent a pairwise MRF with factors over the nodes and edges?

LOG-LINEAR MODELS

$$\phi(\mathbf{D}) = e^{\left(-\varepsilon(\mathbf{D})\right)}$$

$\varepsilon(\mathbf{D}) = -\ln(\phi(\mathbf{D}))$ is often called the *energy function*.

In statistical physics, the probability of a physical state depends inversely on its energy.

Log-linear models guarantee that the factors are positive, in turn guaranteeing that the probability is positive.

Log-linear models

$$
\begin{aligned}
P(X_1, \ldots, X_n) &= \frac{1}{Z}\prod_{i=1}^{k}\phi_i(\mathbf{D}_i) \\
&= \frac{1}{Z}\prod_{i=1}^{k} e^{\left(-\varepsilon_i(\mathbf{D}_i)\right)} \\
&= \frac{1}{Z} e^{-\sum_{i=1}^{k}\varepsilon_i(\mathbf{D}_i)}
\end{aligned}
$$

Log-linear example

A	B	$\phi(A,B)$	$\varepsilon(A,B)$
T	T	5	-1.61
T	F	1	0.00
F	T	1	0.00
F	F	3	-1.10

B	C	$\phi(B,C)$	$\varepsilon(B,C)$
T	T	1	0.00
T	F	2	-0.69
F	T	6	-1.79
F	F	1	0.00

A	B	C	$\phi(A,B)*\phi(B,C)$	$\varepsilon(A,B)+\varepsilon(B,C)$	$\exp(-\Sigma\varepsilon i)$	P(A,B,C)
T	T	T	5.00	-1.61	5.00	0.11
T	T	F	10.00	-2.30	10.00	0.22
T	F	T	6.00	-1.79	6.00	0.13
T	F	F	1.00	0.00	1.00	0.02
F	T	T	1.00	0.00	1.00	0.02
F	T	F	2.00	-0.69	2.00	0.04
F	F	T	18.00	-2.89	18.00	0.39
F	F	F	3.00	-1.10	3.00	0.07
		Z	46.00		46.00	1.00

FEATURES

- **Definition**: A *feature f(***D***)*, is a function from **D** to R.
- Features provide an easy mechanism for specifying certain types of interactions more compactly.
- An important useful function is the indicator function.
  - Given a predicate, the indicator function is
    - 1 if the predicate is true, and
    - 0 otherwise.
- Example indicator functions?

LOG-LINEAR MODEL

- A distribution is a *log-linear model* over a Markov network $\mathcal{H}$ is it is associated with
  - A set of features $\mathcal{F} = \{f_1(\mathbf{D}_1), \ldots, f_k(\mathbf{D}_k)\}$, where each $\mathbf{D}$ is a complete subgraph in $\mathcal{H}$,
  - A set of weights $w_1, \ldots, w_k$

$$P(X_1, \ldots, X_n) = \frac{1}{Z} e^{\left[-\sum_{i=1}^{k} w_i f_i(\mathbf{D}_i)\right]}$$

It is possible to have several features over the same scope.

Features are especially useful for domains where variables have huge domains.

Three different parameterizations

1. Undirected graph
2. Factor graph
3. Features

- Factor graph is finer grained than the undirected graph representation and it is at least as rich
- Feature representation is finer grained than the factor graph representation and it is at least as rich
- Which representation to use?
- UGs are good for discussing independencies, factor graphs are well suited for inference, and features are well suited for learning.

ISING MODELS

Pairwise MRF

- One of the earliest types of Markov network models
- Arose in statistical physics as a model for the energy of a physical system involving a system of interacting atoms
- Each random variable X_i is binary with {+1, -1}.
- Edges: $\varepsilon(x_i, x_j) = -w_{ij}x_ix_j$
- Nodes: $\varepsilon(x_i) = -u_ix_i$
- Depending on the weights, w_{ij}, the model prefers various configurations
  - $w_{ij}>0$: x_i and x_j are preferred to have the same value
    - Ferromagnetic
  - $w_{ij}<0$: x_i and x_j are preferred to have different values
    - Antiferromagnetic
  - $w_{ij}=0$: x_i and x_j are non-interacting

$w_{ij} > 0 \quad x_i = x_j$

$w_{ij} < 0 \quad x_i \neq x_j$

METRIC MRFS

- Nodes X_1 through X_n, related by a set of edges, $\mathcal{E}$, and each X_i can take a label from $\mathcal{V} = \{v_1, \ldots, v_K\}$
- Each node has its own preferences among the possible labels
  - Node potentials
- We also want smoothness over the graph; the neighboring nodes should take similar labels.
  - Edge potentials
- Objective: MAP assignment to $\mathcal{X}$
  - So, we can drop 1/Z

Metric MRFs

$$E\left(x_1,\ldots,x_n\right)=\sum_i \varepsilon_i\left(x_i\right)+\sum_{(i,j)\in\mathcal{E}} \varepsilon_{i,j}\left(x_i,x_j\right)$$

$$\arg\min_{x_1,\ldots,x_n} E\left(x_1,\ldots,x_n\right)$$

$$\varepsilon_{i,j}\left(x_i,x_j\right) = \begin{cases} 0 & x_i = x_j \\ \lambda_{i,j} & x_i \neq x_j \end{cases}$$

$\lambda_{i,j} \geq 0$. The lowest energy, 0, is obtained when two neighboring nodes take the same value, and a higher energy when they do not.

Metric MRFs

- We may want a more general distance function between labels in the case of multiclass case
  - Maybe some labels are more similar than others
- **Definition**: A function, μ: $\mathcal{V} \times \mathcal{V} \rightarrow [0, \infty]$, is a *metric* if it satisfies
  - Reflexivity
    - $\mu(v_k, v_l)=0$, if and only if $k=l$
  - Symmetry
    - $\mu(v_k, v_l) = \mu(v_l, v_k)$
  - Triangle inequality
    - $\mu(v_k, v_l) + \mu(v_l, v_m) \geq \mu(v_k, v_m)$
- Metric MRF: $\varepsilon(v_k, v_l) = \mu(v_k, v_l)$

CRFs

- **Definition**: A conditional random field is an undirected graph $\mathcal{H}$ whose nodes correspond to $\mathbf{X} \cup \mathbf{Y}$; $\mathcal{H}$ is parameterized by a set of factors $\phi_i(\mathbf{D}_i)$, where $\mathbf{D}_i \not\subset \mathbf{X}$. The network encodes the following distribution:

$$P(\mathbf{Y} \mid \mathbf{X}) = \frac{1}{Z(\mathbf{X})} \prod_i \phi_i(\mathbf{D}_i)$$

$$Z(\mathbf{X}) = \Sigma_{\mathbf{Y}} \Pi_i \phi(\mathbf{D}_i)$$

$$P(Y|X) = \frac{P(Y,X)}{P(X)}$$

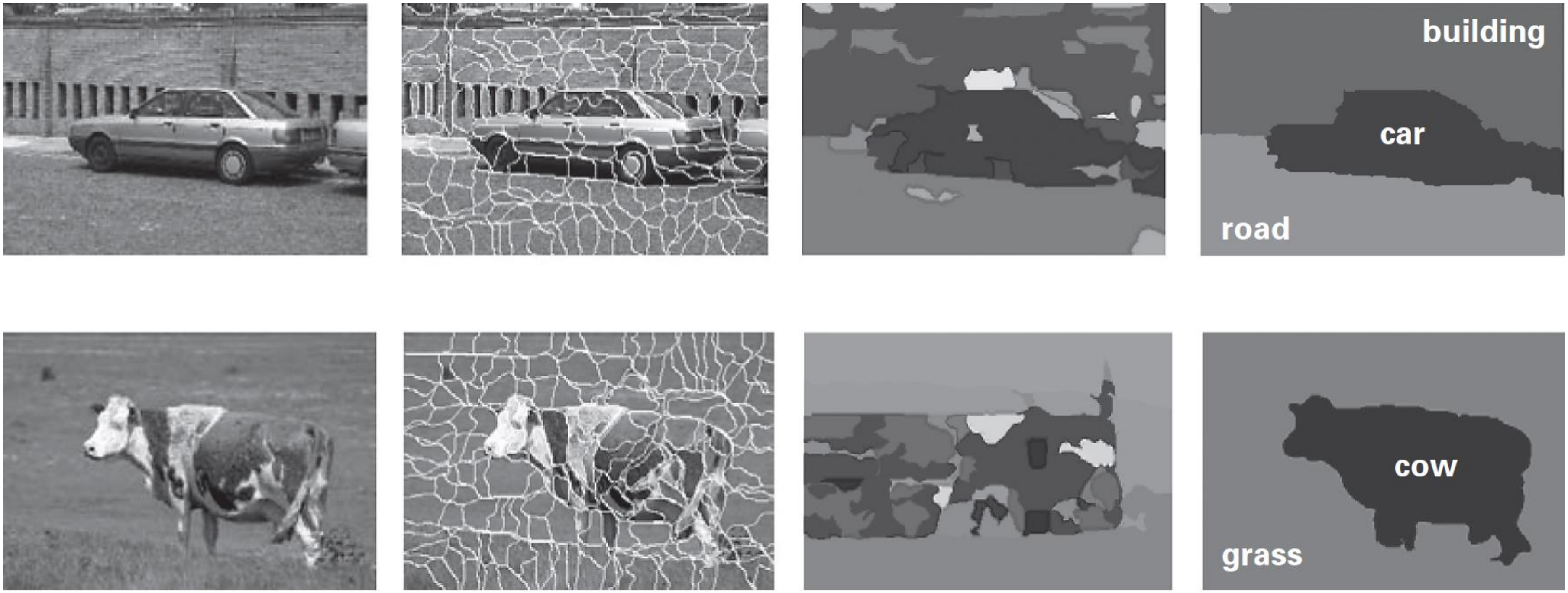
Why do we want P(Y|X) and not necessarily P(**Y**,**X**)?

Why does *Z* have **X** as an argument?

MRFs for Vision (Box 4.B)

- Tasks
  - Image segmentation, noise removal, object recognition, etc.
- Typically, pairwise MRFs are used
  - Variables are pixels and edges exist between adjacent pixels
- Image denoising
  - Restore the true value of all the pixels
  - Node potential: penalizes large deviations from the observed pixel value
  - Edge potential: prefers continuity in the predicted pixel values
    - Don't want to smooth too much to allow object boundaries

Image Segmentation Example



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

CRFs for text analysis (Box 4.E)

- **Tasks**: Part-of-speech tagging, identifying named entities, structured information extraction
- **Target**: **Y**, the labels for each word (or a phrase)
- **Input**: **X**, the text
- **Features**: Capture often domain knowledge about interactions
  - Within target variables, and
  - Between the target variables and the input
  - (No features between solely input variables)

NAMED ENTITY RECOGNITION (BOX 4.E)

- Task: Identify named entities such as people, places, organizations, etc.
- Entities span multiple words and entities might not be apparent from individual words
  - "Chicago" is a location, "Chicago Tribune" is an organization
- Given text of length T, words X_t, $1 \leq t \leq T$, define target variables Y_t.
- Y_t represents B-PERSON, I-PERSON, B-LOC., I-LOC., B-ORG., I-ORG., and OTHER.

Named entity recognition (box 4.e)

- A common structure is a linear-chain CRF
- Factors
  - $\phi_t(Y_t,\ Y_{t+1})$: Dependency between neighboring target variables
  - $\phi_t(Y_t, X_1, \ldots, X_T)$: Dependency between a target and its context
- Rather than a table, represent it as a log-linear model with features
  - Thousands of features that encode domain knowledge
- More details in the book; highly recommend to read it
- Software – many implementations out there in Java, Matlab, C++, …

FROM DISTRIBUTIONS TO GRAPHS

- **Task:** Given a P, find a Markov network structure $\mathcal{H}$ that is a minimal I-Map for P
- **Procedure 1:** Pairwise independencies
  - Add edges between X and Y, if P does not entail $X \perp Y \mid \mathcal{X} \setminus \{X, Y\}$
- **Procedure 2:** Local independencies
  - Add edges between X and all $Y \in MB_P(X)$
- **Theorems**: Let P be a positive distribution and $\mathcal{H}$ be the structure constructed through above procedures. Then $\mathcal{H}$ is a unique minimal I-Map for P.

Bayesian networks & Markov networks

- We've said that the set of distributions that can be represented using BNs and MNs are different.
- Can we go from a BN to a MN and/or vice versa?

BNs to MNs

- **Proposition:** Let $\mathcal{B}$ be a Bayesian network over $\mathcal{X}$. Then $P_{\mathcal{B}}(\mathcal{X})$ is a Gibbs distribution defined by the factors $\Phi=\{\phi(X_i)\}$ for $X_i \in \mathcal{X}$, where $\phi(X_i) = P_{\mathcal{B}}(X_i \mid \text{Parents}(X_i))$. The partition function Z is 1.

BNs to MNs

- Given a Bayesian network structure $\mathcal{G}$, find a Markov network structure $\mathcal{H}$ that is a minimal I-Map for $\mathcal{G}$.
- **Definition**: *Moralized graph*: The moral graph M[$\mathcal{G}$] of a Bayesian network structure $\mathcal{G}$ over $\mathcal{X}$ is an undirected graph over $\mathcal{X}$ that contains an undirected edge between X and Y if
  - There is a directed edge between X and Y in $\mathcal{G}$, or
  - X and Y are both parents of the same node in $\mathcal{G}$
- Moralized: Parents of a node are married by adding an edge between them

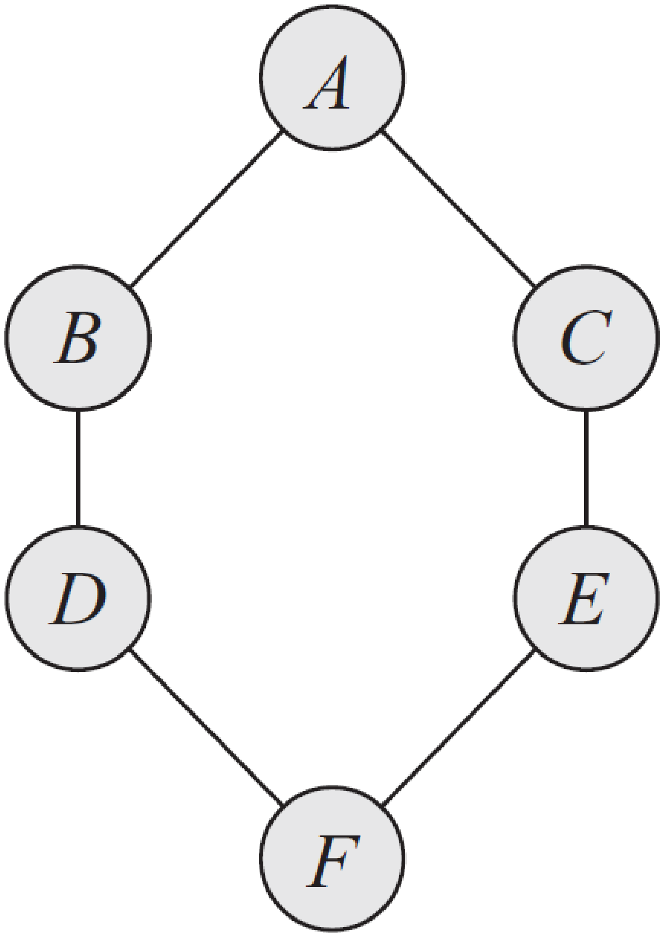
BNs to MNs

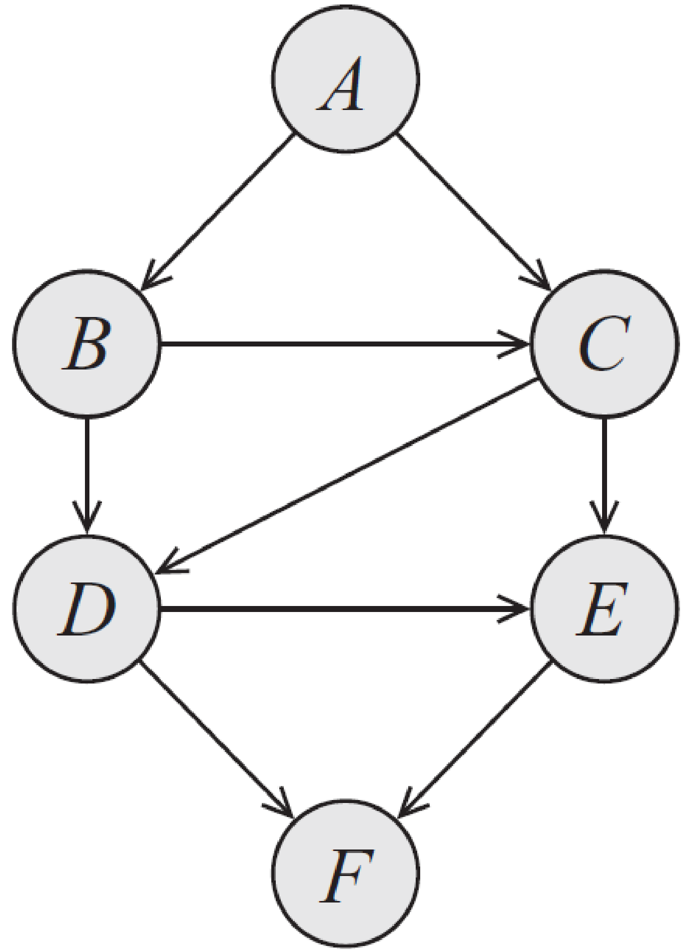
- **Proposition**: Let $\mathcal{G}$ be any Bayesian network. The moralized graph $\mathcal{M}[\mathcal{G}]$ is a minimal I-Map for $\mathcal{G}$.
- Does moralization cause loss of independencies? If so, when?
- **Proposition**: Let $\mathcal{G}$ be any moral Bayesian network. The moralized graph $\mathcal{M}[\mathcal{G}]$ is a P-Map for $\mathcal{G}$.

MNs to BNs

- Given a Markov network structure $\mathcal{H}$, find a Bayesian network structure $\mathcal{G}$ that is a minimal I-Map for $\mathcal{H}$.
- Pick an order of the variables
- Follow the procedure we discussed before.

MNs to BNs





MNs to BNs

- **Theorem:** Let $\mathcal{H}$ be a Markov network structure and let $\mathcal{G}$ be any Bayesian network structure that is a minimal I-Map for $\mathcal{H}$. Then, $\mathcal{G}$ can have no immoralities.
- **Definition**: *Chordal graph*: A graph where the longest minimal loop is a triangle. Also called *triangulated*.
- **Corollary**: Let $\mathcal{H}$ be a Markov network structure and let $\mathcal{G}$ be any Bayesian network structure that is a minimal I-Map for $\mathcal{H}$. Then, $\mathcal{G}$ is necessarily chordal.

MNs to BNs

- **Theorem**: Let $\mathcal{H}$ be a nonchordal Markov network. Then, there is no Bayesian network $\mathcal{G}$ which is a perfect map for $\mathcal{H}$.
- **Theorem**: Let $\mathcal{H}$ be a chordal Markov network. Then, there is a Bayesian network $\mathcal{G}$ which is a perfect map for $\mathcal{H}$.