Graphical Models

Lecture 4:

Undirected Graphical Models

Andrew McCallum
mccallum@cs.umass.edu

Thanks to Noah Smith and Carlos Guestrin for some slide materials.

Administrivia

- HW#1:
  - Source code was due Tuesday by 5pm.
  - Reports due today (Thursday) 5pm.

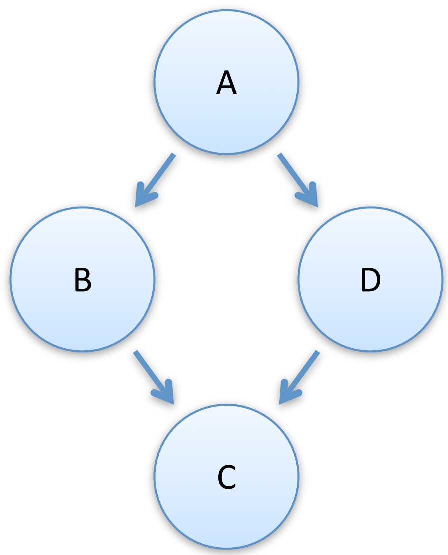
  - Heard of some successes.
  - Comments?

Motivating Example: No Bayesian Network is a P-Map

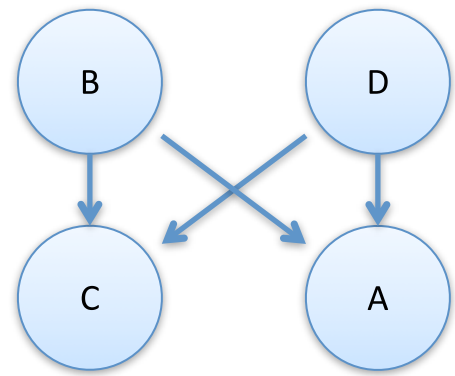
- *Misunderstanding students*

I(P):

- $A \perp C \mid B, D$
- $B \perp D \mid A, C$
- $\neg\, B \perp D$
- $\neg\, A \perp C$



Fails to capture:
$B \perp D \mid A, C$



Fails to capture:
$\neg\, B \perp D$

Want to represent preference for pairwise affinities

Undirected Graphical Models

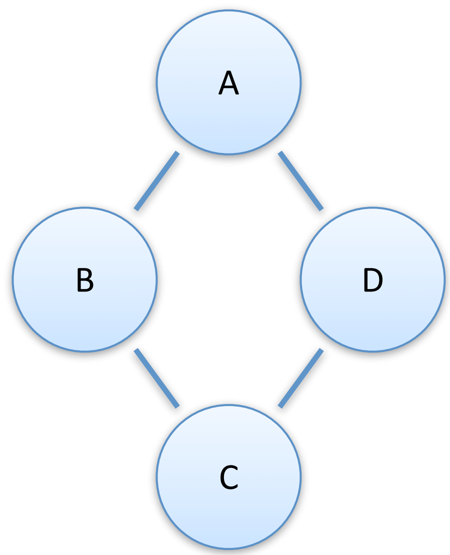
- Also known as **Markov networks** and **Markov random fields**.
- Alternative representation of graphs with probability distributions.
  - Motivation
  - Definition
  - Independence
  - Representation theorems
  - I-maps and P-maps

Motivating Example:
This Markov Network is a P-Map!

- *Misunderstanding students*

I(P):

- $A \perp C \mid B, D$
- $B \perp D \mid A, C$
- $\neg\, B \perp D$
- $\neg\, A \perp C$

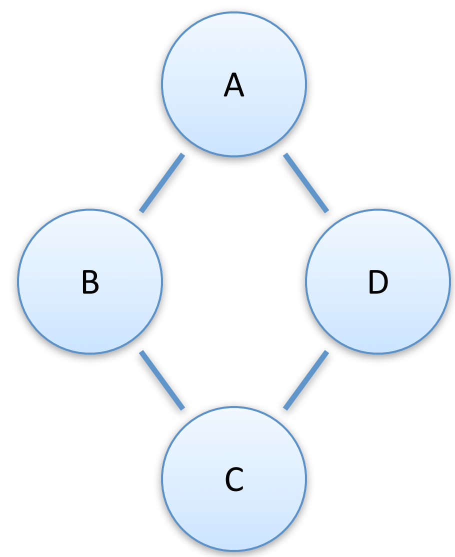


(Will explain soon why it is a P-Map.)

"affinity functions" or "compatibility functions" on edges.
Call these "factors"

Markov Networks

- Each random variable is a vertex.
- Undirected edges.
- **Factors** are associated with edges (or more generally subsets of nodes that form cliques).
  - A factor maps assignments of its nodes to nonnegative "compatibility" values.

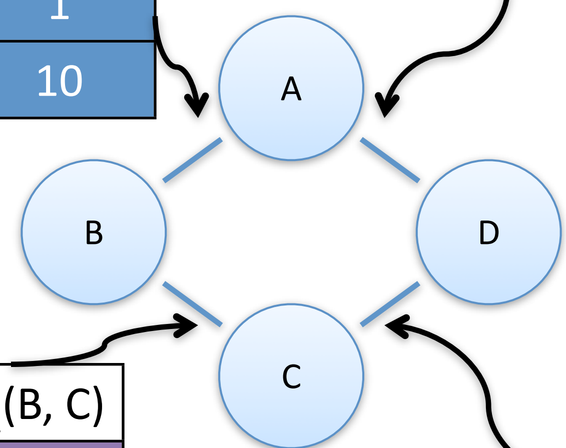


Markov Networks

- In this example, associate a factor with each edge.
  - Could also have factors for single nodes!

A	D	$\phi_4(A, D)$
0	0	100
0	1	1
1	0	1
1	1	100

A	B	$\phi_1(A, B)$
0	0	30
0	1	5
1	0	1
1	1	10



B	C	$\phi_2(B, C)$
0	0	100
0	1	1
1	0	1
1	1	100

C	D	$\phi_3(C, D)$
0	0	1
0	1	100
1	0	100
1	1	1

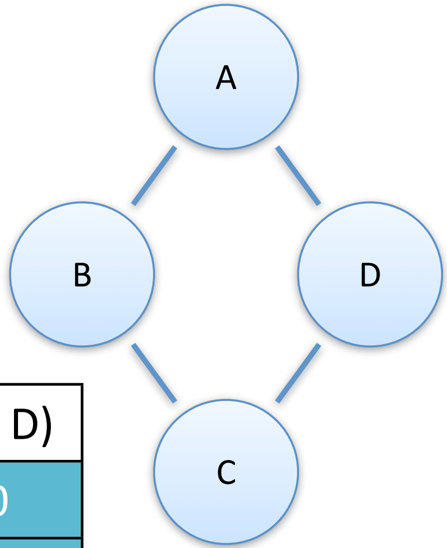
Markov Networks

- Probability distribution:

$$P(a,b,c,d) \propto \phi_1(a,b)\phi_2(b,c)\phi_3(c,d)\phi_4(a,d)$$

$$P(a,b,c,d) = \frac{\phi_1(a,b)\phi_2(b,c)\phi_3(c,d)\phi_4(a,d)}{\sum_{a',b',c',d'} \phi_1(a',b')\phi_2(b',c')\phi_3(c',d')\phi_4(a',d')}$$

$$Z = \sum_{a',b',c',d'} \phi_1(a',b')\phi_2(b',c')\phi_3(c',d')\phi_4(a',d')$$



A	B	ϕ_1(A, B)
0	0	30
0	1	5
1	0	1
1	1	10

B	C	ϕ_2(B, C)
0	0	100
0	1	1
1	0	1
1	1	100

C	D	ϕ_3(C, D)
0	0	1
0	1	100
1	0	100
1	1	1

A	D	ϕ_4(A, D)
0	0	100
0	1	1
1	0	1
1	1	100

Markov Networks

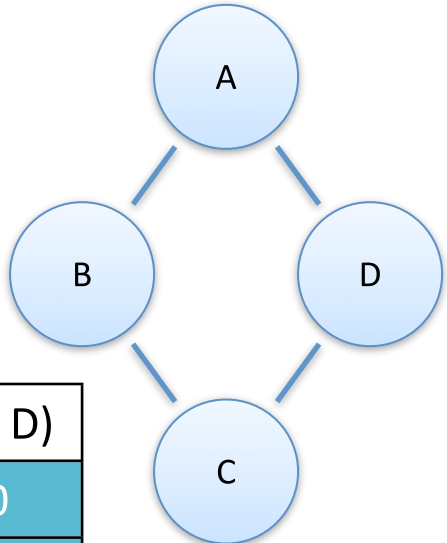
- Probability distribution:

$$P(a,b,c,d) \propto \phi_1(a,b)\phi_2(b,c)\phi_3(c,d)\phi_4(a,d)$$

$$P(a,b,c,d) = \frac{\phi_1(a,b)\phi_2(b,c)\phi_3(c,d)\phi_4(a,d)}{\sum_{a',b',c',d'} \phi_1(a',b')\phi_2(b',c')\phi_3(c',d')\phi_4(a',d')}$$

$$Z = \sum_{a',b',c',d'} \phi_1(a',b')\phi_2(b',c')\phi_3(c',d')\phi_4(a',d')$$

= 7,201,840



A	B	$\phi_1(A, B)$
0	0	30
0	1	5
1	0	1
1	1	10

B	C	$\phi_2(B, C)$
0	0	100
0	1	1
1	0	1
1	1	100

C	D	$\phi_3(C, D)$
0	0	1
0	1	100
1	0	100
1	1	1

A	D	$\phi_4(A, D)$
0	0	100
0	1	1
1	0	1
1	1	100

Markov Networks

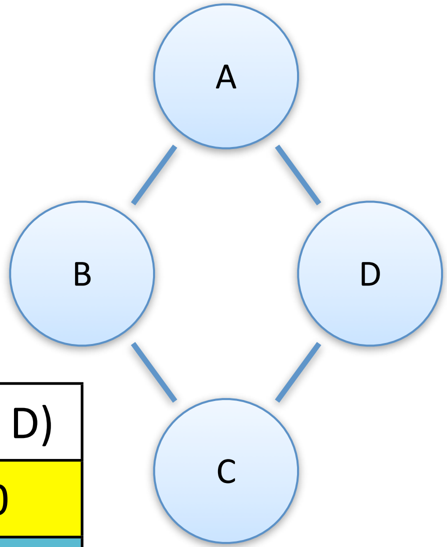
- Probability distribution:

$$P(a,b,c,d) \propto \phi_1(a,b)\phi_2(b,c)\phi_3(c,d)\phi_4(a,d)$$

$$P(a,b,c,d) = \frac{\phi_1(a,b)\phi_2(b,c)\phi_3(c,d)\phi_4(a,d)}{\sum_{a',b',c',d'} \phi_1(a',b')\phi_2(b',c')\phi_3(c',d')\phi_4(a',d')}$$

$$Z = \sum_{a',b',c',d'} \phi_1(a',b')\phi_2(b',c')\phi_3(c',d')\phi_4(a',d')$$

= 7,201,840



A	B	$\phi_1(A, B)$
0	0	30
0	1	5
1	0	1
1	1	10

B	C	$\phi_2(B, C)$
0	0	100
0	1	1
1	0	1
1	1	100

C	D	$\phi_3(C, D)$
0	0	1
0	1	100
1	0	100
1	1	1

A	D	$\phi_4(A, D)$
0	0	100
0	1	1
1	0	1
1	1	100

P(0, 1, 1, 0)
= 5,000,000 / Z
= 0.69

Markov Networks

Are factors on edges enough?
No!
Think about P with no ind.
Fully connected.
Think about # parameters.
Up to us to define alternative.

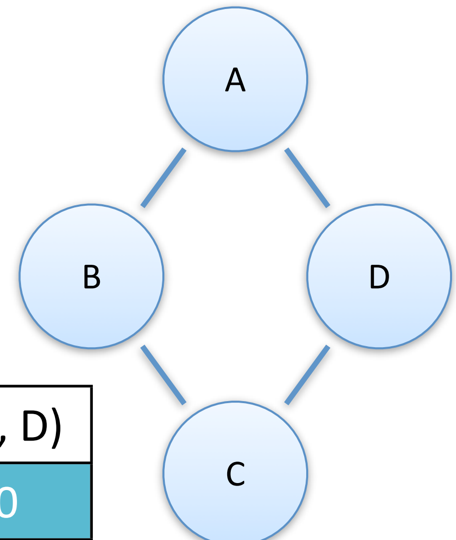
- Probability distribution:

$$P(a,b,c,d) \propto \phi_1(a,b)\phi_2(b,c)\phi_3(c,d)\phi_4(a,d)$$

$$P(a,b,c,d) = \frac{\phi_1(a,b)\phi_2(b,c)\phi_3(c,d)\phi_4(a,d)}{\sum_{a',b',c',d'} \phi_1(a',b')\phi_2(b',c')\phi_3(c',d')\phi_4(a',d')}$$

$$Z = \sum_{a',b',c',d'} \phi_1(a',b')\phi_2(b',c')\phi_3(c',d')\phi_4(a',d')$$

= 7,201,840



A	B	ϕ_1(A, B)
0	0	30
0	1	5
1	0	1
1	1	10

B	C	ϕ_2(B, C)
0	0	100
0	1	1
1	0	1
1	1	100

C	D	ϕ_3(C, D)
0	0	1
0	1	100
1	0	100
1	1	1

A	D	ϕ_4(A, D)
0	0	100
0	1	1
1	0	1
1	1	100

P(1, 1, 0, 0)
= 10 / Z
= 0.0000014

Administrivia

- CSCF mailing lists for HW submission were a disaster.
- http://nescai.cs.umass.edu/cs691/
  - username: cs691
  - password: *********

Markov Networks (General Form)

- Let $\mathbf{D}_i$ denote the set of variables (subset of **X**) in the ith clique.
- Probability distribution is a **Gibbs** distribution:

$$P(\boldsymbol{X}) = \frac{U(\boldsymbol{X})}{Z}$$

$$U(\boldsymbol{X}) = \prod_{i=1}^{m} \phi_i(\boldsymbol{D}_i)$$

$$Z = \sum_{\boldsymbol{x} \in \mathrm{Val}(\boldsymbol{X})} U(\boldsymbol{x})$$

How big is this sum?

Compare to directed!

Z = 1

Notes

- Z might be hard to calculate.
  - "Normalization constant"
  - "Partition function"
- Can get efficient calculation in some cases.
  - This is an **inference** problem; it's equivalent to marginalizing over everything.
- *Ratios* of probabilities are easy.

$$\frac{P(\boldsymbol{x})}{P(\boldsymbol{x'})} = \frac{U(\boldsymbol{x})/Z}{U(\boldsymbol{x'})/Z} = \frac{U(\boldsymbol{x})}{U(\boldsymbol{x'})}$$

Pairwise Markov Networks

- All factors associated with one node or one pair (connected by an edge).

$$P(\boldsymbol{X}) = \prod_{i} \phi_i(X_i) \prod_{(i,j)\in\mathcal{H}} \phi_{i,j}(X_i, X_j)$$

- The graph may have cliques with more than two nodes, but they do not have factors.

Markov Networks Can Always Be Made Pairwise

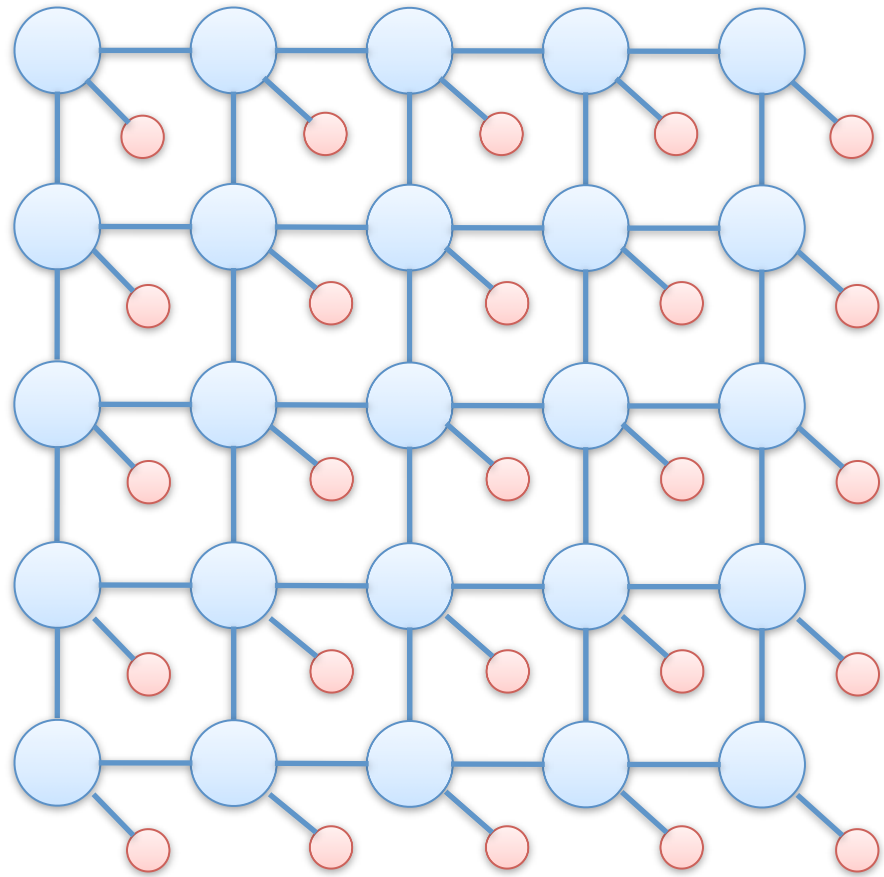
- For any factor over three or more variables, introduce a new variable.
- Val(X) has a size that is the number of values the factor can take (exponential in values of neighbors).
- Local factor structure is lost.

Show example on board

Pairwise Markov Network Example

- Classify each pixel as foreground or background.

$$
\begin{aligned}
\phi_i(X_i = \text{fg}, C_i) &= \exp \frac{-\|C_i - \mu_{\text{fg}}\|^2}{\sigma^2} \\
\phi_i(X_i = \text{bg}, C_i) &= \exp \frac{-\|C_i - \mu_{\text{bg}}\|^2}{\sigma^2} \\
\phi_{i,j}(X_i, X_j) &= \begin{cases} 10 & \text{if } X_i = X_j \\ 1 & \text{otherwise} \end{cases}
\end{aligned}
$$



Application: Image Segmentation

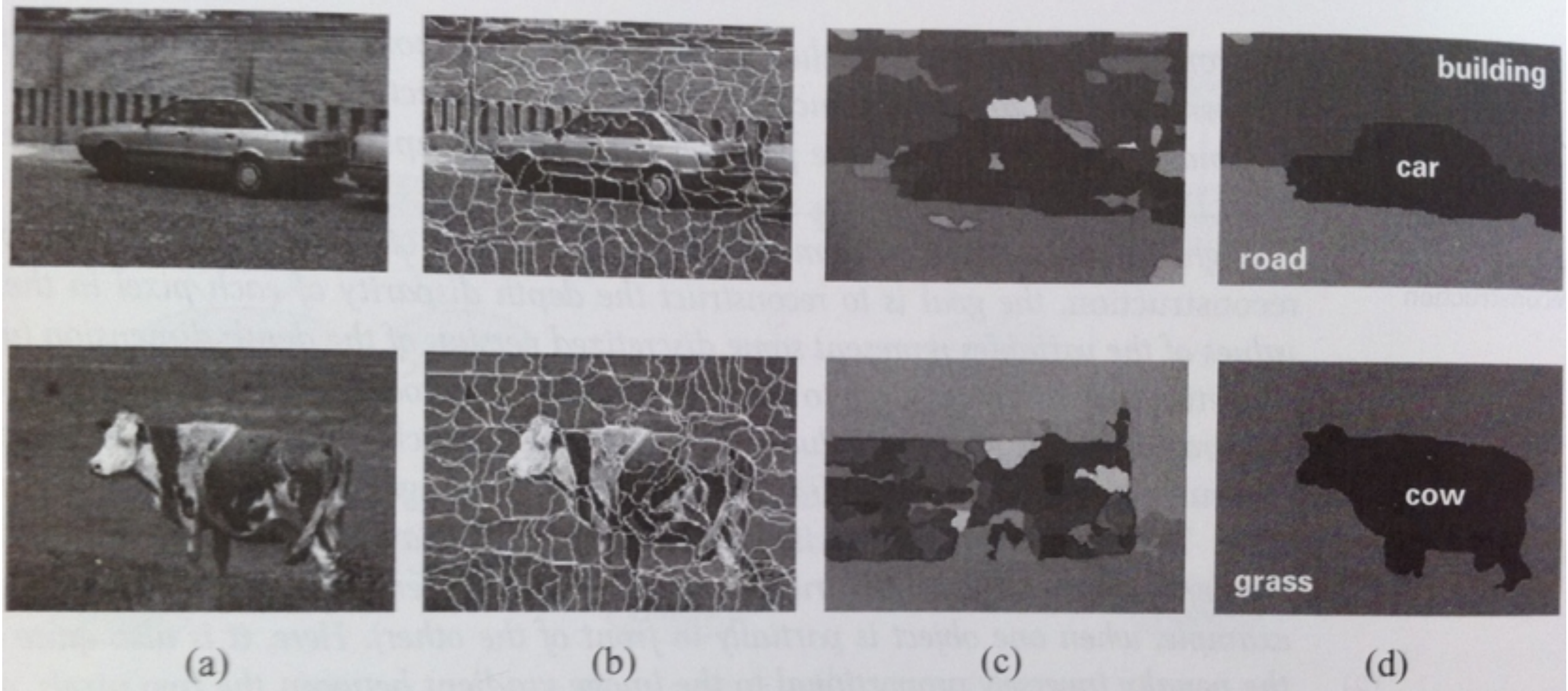
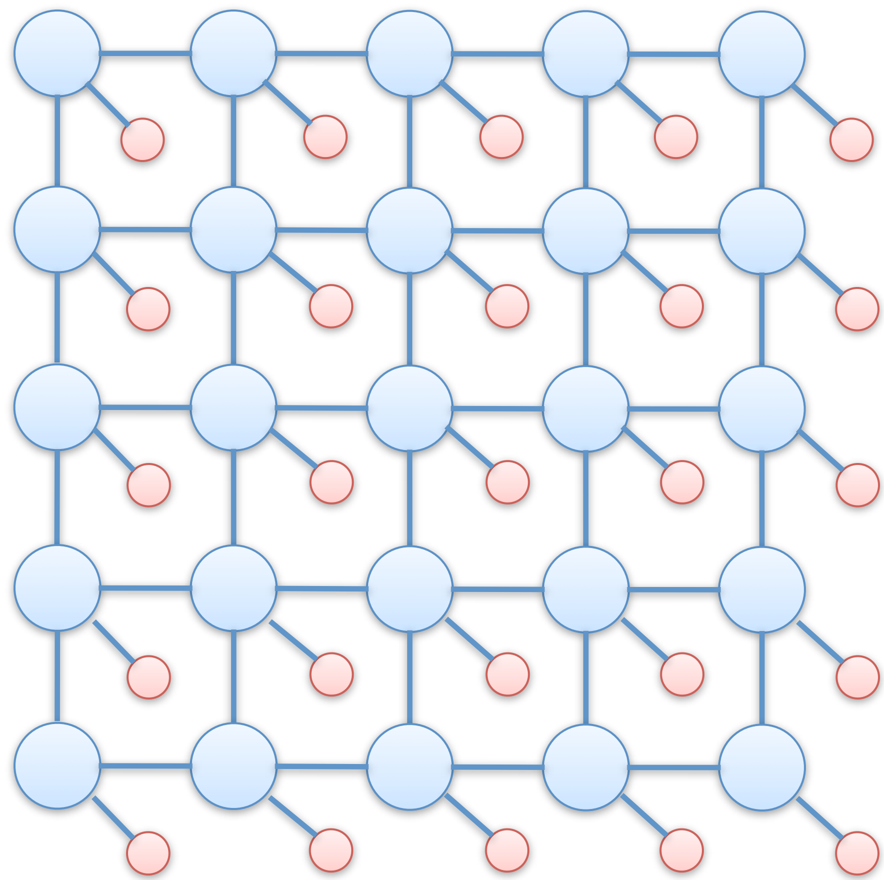


Figure 4.B.1 — Two examples of image segmentation results (a) The original image. (b) An overseg-mentation known as superpixels; each superpixel is associated with a random variable that designates its segment assignment. The use of superpixels reduces the size of the problems. (c) Result of segmentation using node potentials alone, so that each superpixel is classified independently. (d) Result of segmentation using a pairwise Markov network encoding interactions between adjacent superpixels.

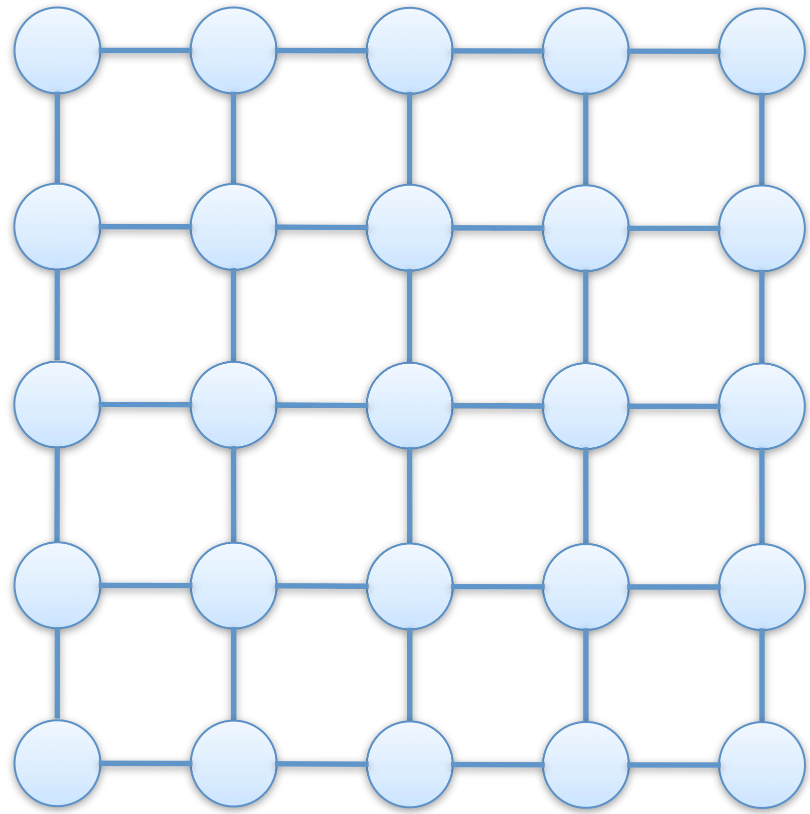
Reducing Factors

- Given some variables' values, we can **reduce** the factors to that context.
- Resulting conditional distribution is still Gibbs. (New Z.)



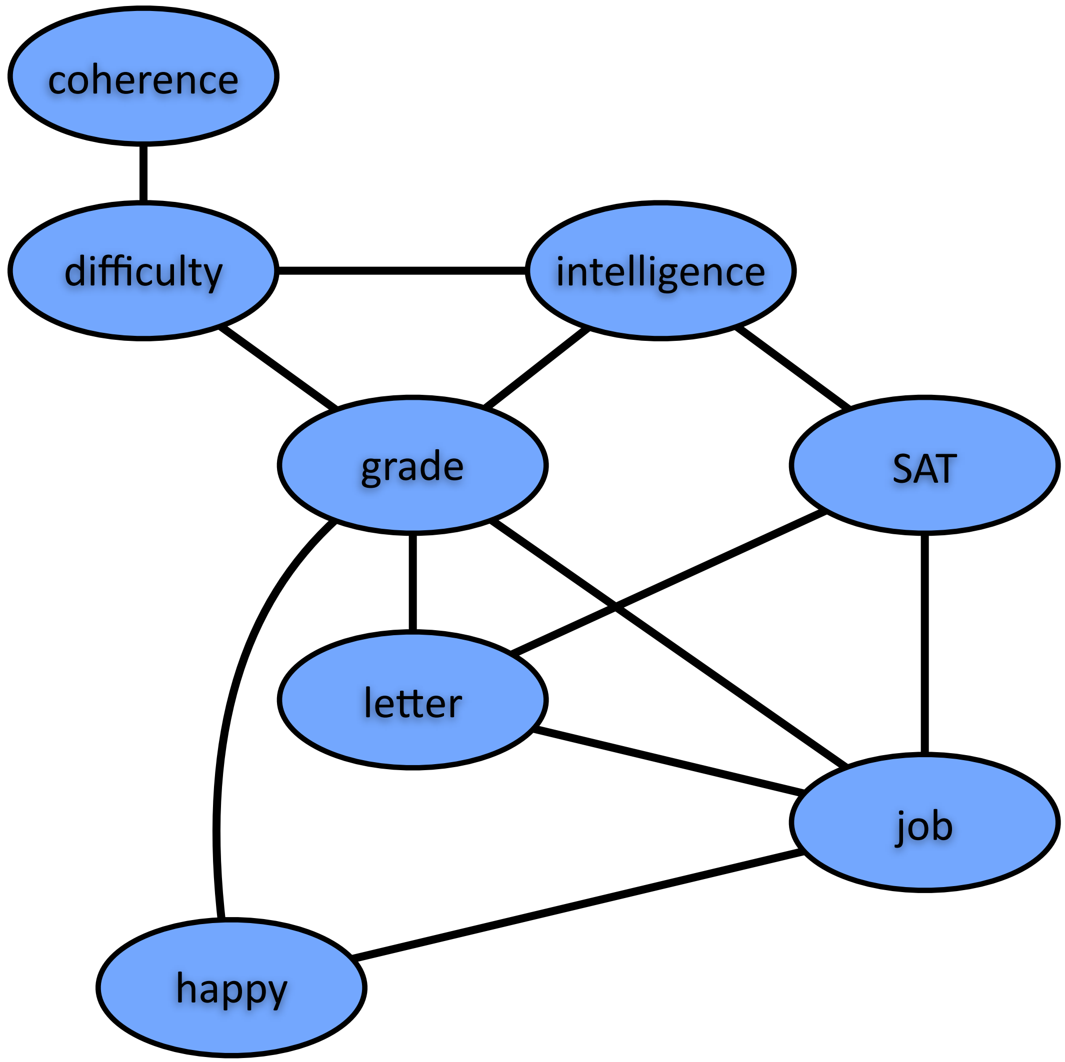
Reducing Factors

- Given some variables' values, we can **reduce** the factors to that context.
- Resulting conditional distribution is still Gibbs. (New Z.)



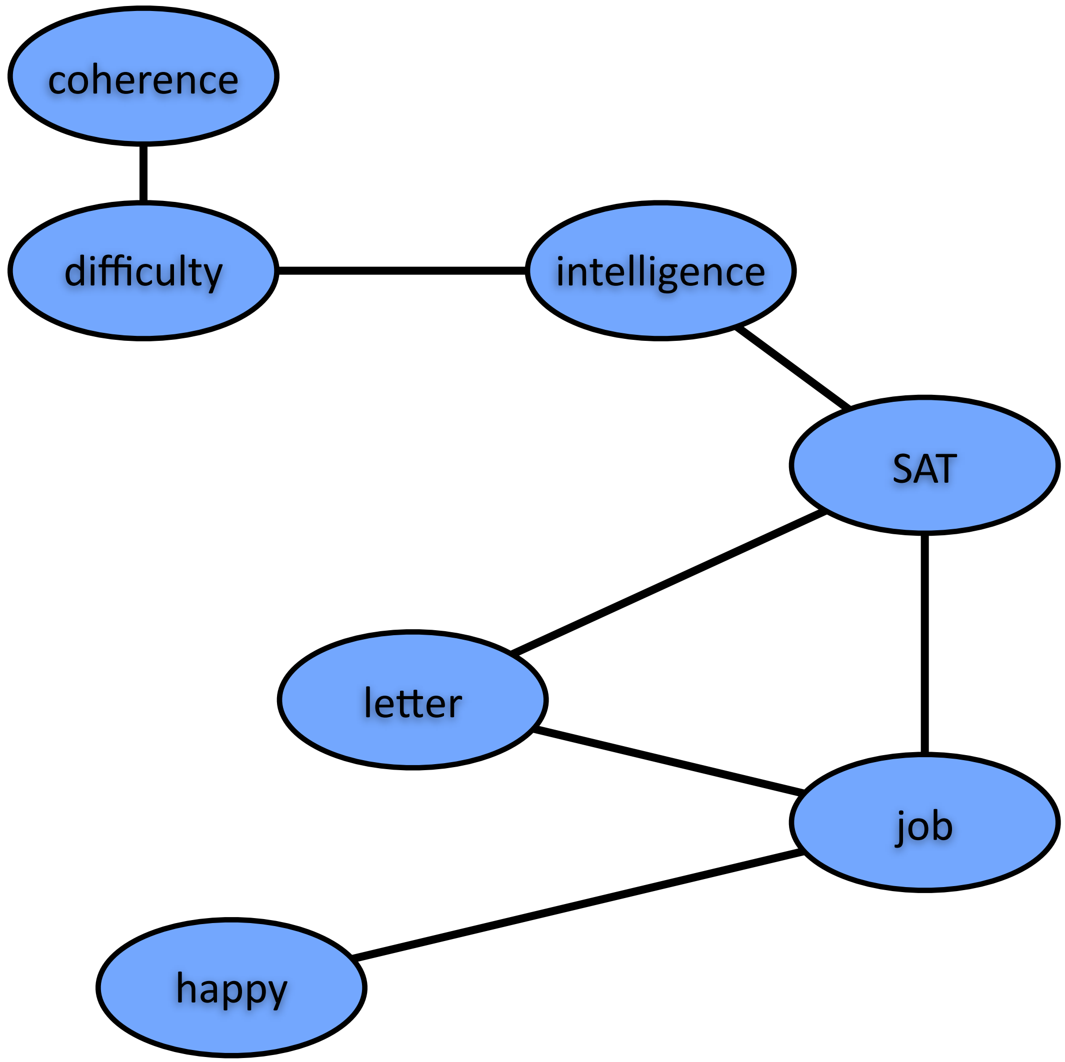
Talk about "factor product" and "factor reduction" on the board.
Relate undirected to undirected models with examples!

Reduced Markov Networks



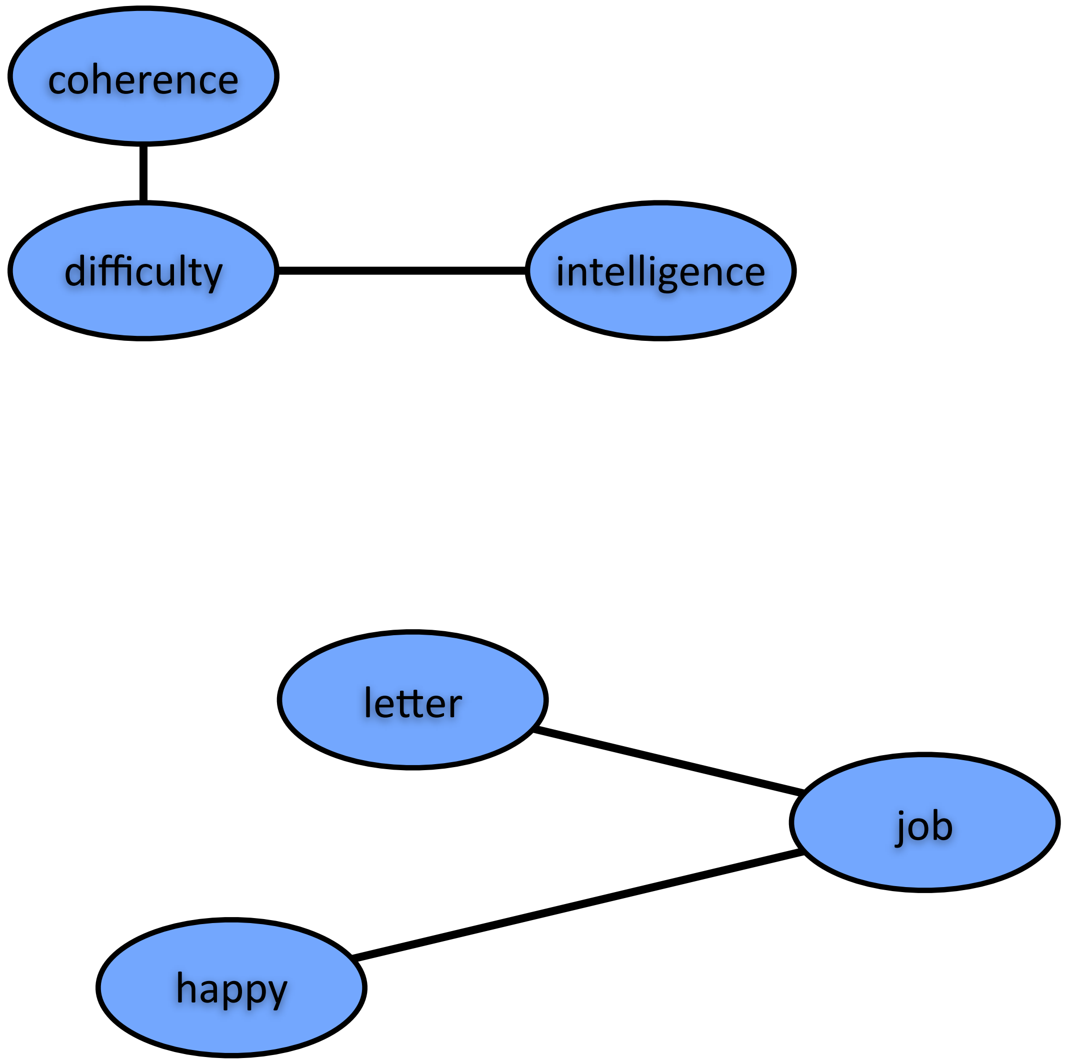
Full network

Reduced Markov Networks



Condition on "grade"

Reduced Markov Networks



Condition on
“grade” and “SAT”

Independence in Markov Networks

- Given a set of observed nodes **Z**, a path X_1-X_2-X_3-…-X_k is **active** if no nodes on the path are observed.
- Two sets of nodes **X** and **Y** in $\mathcal{H}$ are **separated** given **Z** if there is no active path between any $X_i \in \mathbf{X}$ and any $Y_i \in \mathbf{Y}$.
  - Denoted: $\text{sep}_{\mathcal{H}}(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z})$
- Global Markov assumption: $\text{sep}_{\mathcal{H}}(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z}) \Rightarrow \mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}$

Representation Theorems

- Bayesian networks …

The Bayesian network graph is an I-map for P.	⇔	$P(\boldsymbol{X}) = \prod_{i=1}^{n} P(X_i \mid \mathbf{Parents}(X_i))$

  – Independencies give you the Bayesian network.
  – Bayesian network reveals independencies.

Representation Theorems

- Bayesian networks …

The Bayesian network graph is an I-map for P. ⇔ $$P(\boldsymbol{X}) = \prod_{i=1}^{n} P(X_i \mid \mathbf{Parents}(X_i))$$

- Markov networks …

Representation Theorems

- Bayesian networks …

The Bayesian network graph is an I-map for P.

$\Leftrightarrow$

$$P(\boldsymbol{X}) = \prod_{i=1}^{n} P(X_i \mid \mathbf{Parents}(X_i))$$

- Markov networks …

The Markov network's graph is an I-map for P.

$\Rightarrow$?

$\Leftarrow$?

$$P(\boldsymbol{X}) = \frac{1}{Z} \prod_{i=1}^{m} \phi_i(\boldsymbol{D}_i)$$

Representation Theorem (I)

← Gibbs distribution satisfies the independencies associated with the graph

- Factorization into $\mathbf{D}_i$ gives a simple way to build the graph:
  - put an edge between X and Y iff $\exists \mathbf{D}_i$ such that X, Y $\in \mathbf{D}_i$.

The Markov network's graph is an I-map for P.

$$P(\boldsymbol{X}) = \frac{1}{Z} \prod_{i=1}^{m} \phi_i(\boldsymbol{D}_i)$$

Representation Theorem (I)

- Assume Gibbs.
- Consider three disjoint sets of variables, **W**, **Y**, and **Z**, such that dsep$_{\mathcal{H}}$(**W**, **Y** | **Z**).
  - For now assume these comprise all of **X**; general case is not hard.
- No edges between **W** and **Y**, so any clique is either in **W**∪**Z** or **Y**∪**Z**.

$$
\begin{aligned}
P(\boldsymbol{X}) &= \frac{1}{Z}\left(\prod_{i:\boldsymbol{D}_i \subseteq \boldsymbol{W}\cup\boldsymbol{Z}} \phi_i(\boldsymbol{D}_i)\right)\left(\prod_{i:\boldsymbol{D}_i \subseteq \boldsymbol{Y}\cup\boldsymbol{Z}} \phi_i(\boldsymbol{D}_i)\right) \\
&= \frac{1}{Z}\Phi_1(\boldsymbol{W},\boldsymbol{Z})\Phi_2(\boldsymbol{Y},\boldsymbol{Z})
\end{aligned}
$$

- It follows that **W** $\perp$ **Y** | **Z**.

Representation Theorem (II)

- Other direction?

The Markov network's graph is an I-map for P.

?

$$P(\boldsymbol{X}) = \frac{1}{Z} \prod_{i=1}^{m} \phi_i(\boldsymbol{D}_i)$$

Representation Theorem (II)

- Fails!

The Markov network's graph is an I-map for P.

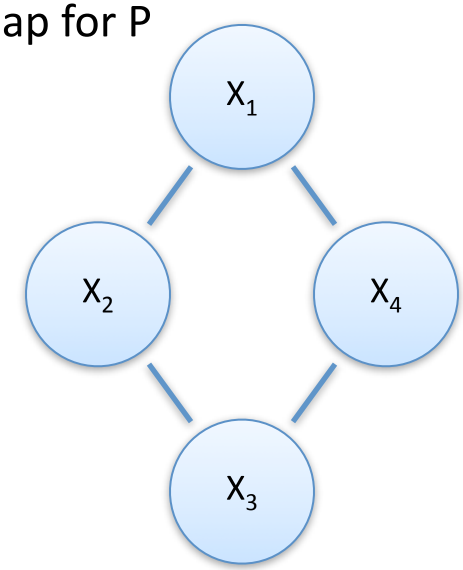
$$P(\boldsymbol{X}) = \frac{1}{Z}\prod_{i=1}^{m}\phi_i(\boldsymbol{D}_i)$$

Example

P

X_1	X_2	X_3	X_4	
0	0	0	0	0.125
0	0	0	1	0.125
0	0	1	0	0
0	0	1	1	0.125
0	1	0	0	0
0	1	0	1	0
0	1	1	0	0
0	1	1	1	0.125
1	0	0	0	0.125
1	0	0	1	0
1	0	1	0	0
1	0	1	1	0
1	1	0	0	0.125
1	1	0	1	0
1	1	1	0	0.125
1	1	1	1	0.125

I-map for P

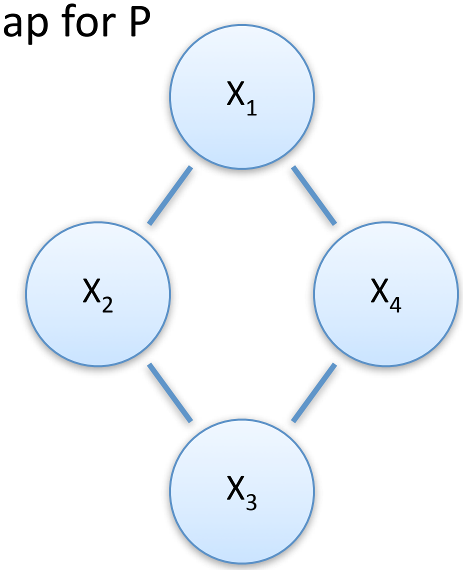


Example

P

X_1	X_2	X_3	X_4	
0	0	0	0	0.125
0	0	0	1	0.125
0	0	1	0	0
0	0	1	1	0.125
0	1	0	0	0
0	1	0	1	0
0	1	1	0	0
0	1	1	1	0.125
1	0	0	0	0.125
1	0	0	1	0
1	0	1	0	0
1	0	1	1	0
1	1	0	0	0.125
1	1	0	1	0
1	1	1	0	0.125
1	1	1	1	0.125

I-map for P



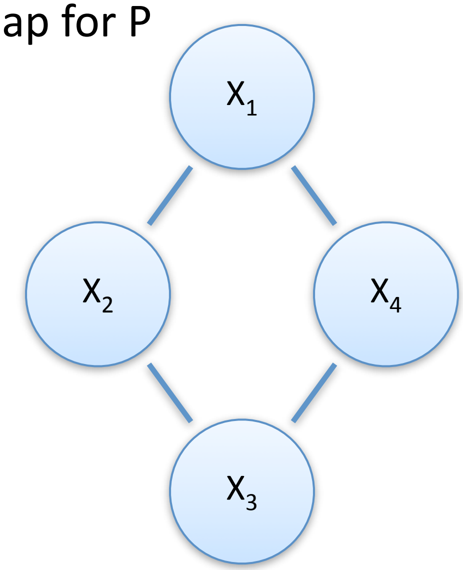
$X_1 \perp X_3 \mid X_2, X_4$ and others hold in P.

Example

P

X_1	X_2	X_3	X_4	
0	0	0	0	0.125
0	0	0	1	0.125
0	0	1	0	0
0	0	1	1	0.125
0	1	0	0	0
0	1	0	1	0
0	1	1	0	0
0	1	1	1	0.125
1	0	0	0	0.125
1	0	0	1	0
1	0	1	0	0
1	0	1	1	0
1	1	0	0	0.125
1	1	0	1	0
1	1	1	0	0.125
1	1	1	1	0.125

I-map for P



The distribution does not factorize into the graph's cliques!

Representation Theorem (II)

- Succeeds if P(**x**) > 0 for all **x**.
- Hammersley-Clifford Theorem

The Markov network's graph is an I-map for P and P is nonnegative.

⇒

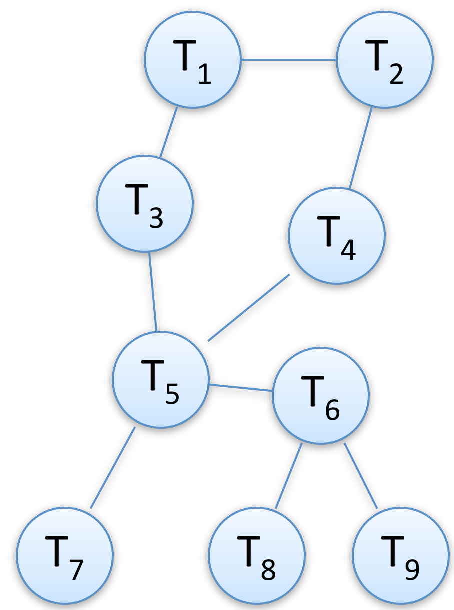
$$P(\boldsymbol{X}) = \frac{1}{Z} \prod_{i=1}^{m} \phi_i(\boldsymbol{D}_i)$$

Graphs and Independencies

	Bayesian Networks	**Markov Networks**
local independencies	local Markov assumption	?
global independencies	d-separation	separation

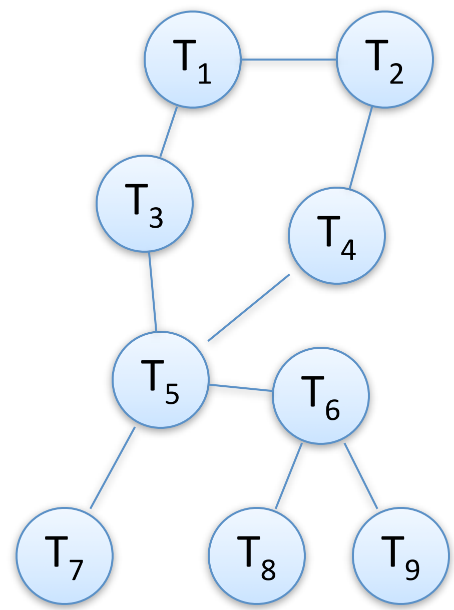
Local Independence Assumptions in Markov Networks

- **Separation** defines global independencies.



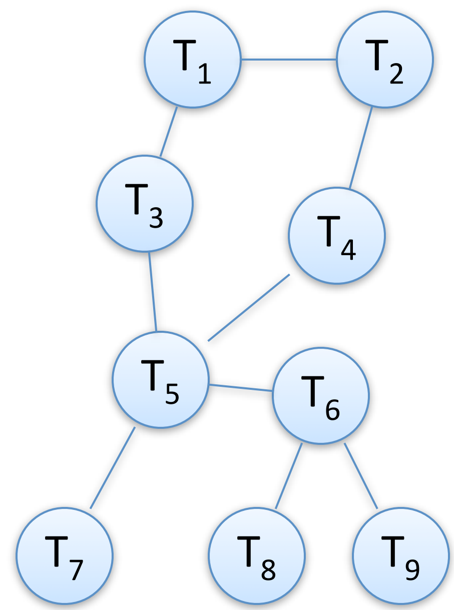
Local Independence Assumptions in Markov Networks

- **Pairwise** Markov independence: pairs of *non-adjacent* variables are independent given everything else.



Local Independence Assumptions in Markov Networks

- **Markov blanket**: each variable is independent of the rest given its *neighbors*.



Define *neighbors* on the board

Local Independence Assumptions in Markov Networks

- Separation:
  $\text{sep}_{\mathcal{H}}(\mathbf{W}, \mathbf{Y} \mid \mathbf{Z}) \Rightarrow \mathbf{W} \perp \mathbf{Y} \mid \mathbf{Z}$

- Pairwise Markov:
  $A \perp B \mid \mathbf{X} \setminus \{A, B\}$

- Markov blanket:
  $A \perp \mathbf{X} \setminus \text{Neighbors}(A) \mid \text{Neighbors}(A)$

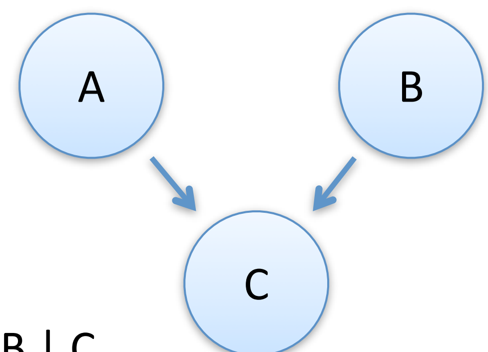
I-Maps

- Fully connected graph is an I-map (like BNs)
- *Minimal* I-maps (delete edge → not an I-map)
  - Not unique for Bayesian networks.
  - *What about Markov Networks?...*
  - Unique for Markov networks (if positive distribution)!
- Simple way to construct a minimal I-map:
  - Check each pairwise Markov assumption.
  - If it's not entailed by P, add edge.

P-Maps

- Want: independencies from the graphical model are exactly the same as in P.
- Doesn't always exist for Bayesian networks (misunderstanding students).
- *What about Markov Networks?...*
- Doesn't always exist for Markov networks.

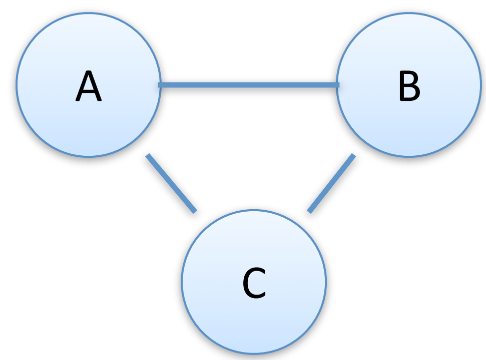
P



$A \perp B$

$\neg\ A \perp B \mid C$

Minimal I-map is not a P-map.



Bayesian Networks and Markov Networks

	Bayesian Networks	**Markov Networks**
local independencies	local Markov assumption	pairwise; Markov blanket
global independencies	d-separation	separation
relative advantages	• v-structures handled elegantly • CPDs are conditional probabilities • probability of full instantiation is easy (no partition function)	• cycles allowed • perfect maps for misunderstanding students

Markov Networks So Far

- Markov network: undirected graph, potentials over cliques (or sub-cliques), normalization via partition function
- Representation theorems
- Independence: active paths/separation; pairwise; Markov blanket
- Minimal I-maps are unique
- P-maps don't always exist

HW#2

- Due Feb 17 and Feb 22