Georgia Institute of Technology

CSIP
Center for Signal & Image Processing

ECE/CS/ISYE 8803

Probabilistic Graphical Models

Module 3 (Part A):
Undirected Graphical Models (Markov Networks)

Faramarz Fekri

Center for Signal and Information Processing

Overview

- Undirected graphical models (Markov networks)
- Parameterization of MRFs
- Independencies encoded by Markov networks
- Maximal Cliques
- Definition of MNs and various factorizations
- Example: Ising model
- I-map, and factorization of MN
- Minimal map, Perfect map, construction of MN
- More on parameterization of MRFs
- Factor graphs
- Log linear model

Read Chapter 4 of K&F

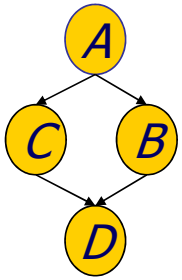
Bayesian Networks are not Enough

- Thm: not every distribution has a perfect map as DAG.

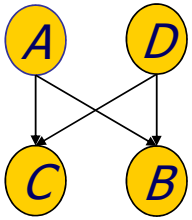
  independencies may be encoded in CPD Ind(X;Y|Z=1)
  some structures cannot be represented in a BN

Proof by counter example:

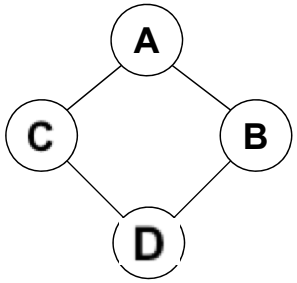
Independencies in P: Ind(A;D | B,C), and Ind(B;C | A,D)



Ind(B;C | A,D) does not hold



Ind(A,D) also holds

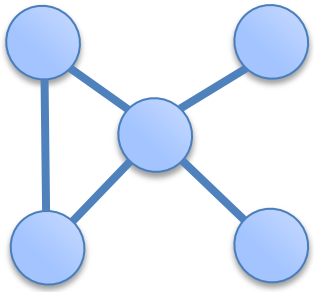


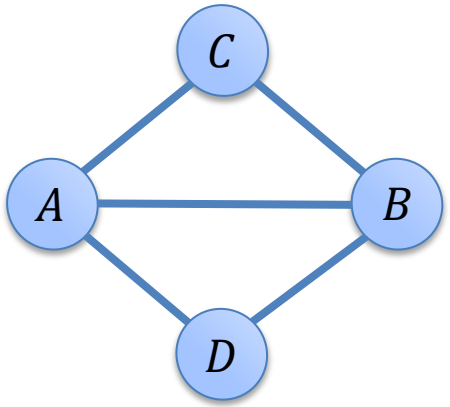
MRF

Undirected Graphical Models (Markov Networks)

Markov Random Fields (MRF)

- **Nodes** correspond to random variables
- **Edges** correspond to direct probabilistic interaction

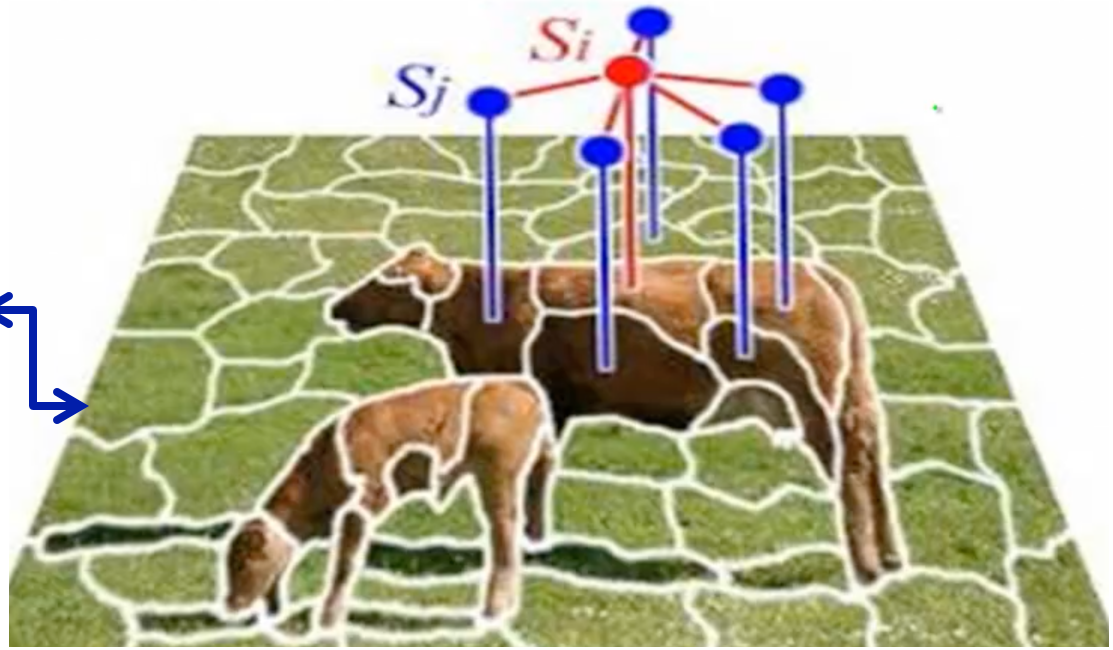




- Pairwise (non-causal) relationships
- Can write down model, and score specific configurations of the graph, but no explicit way to generate samples

Example Application: Image Segmentation

S_i labeled as one of {grass, sky, cow, water, ...}



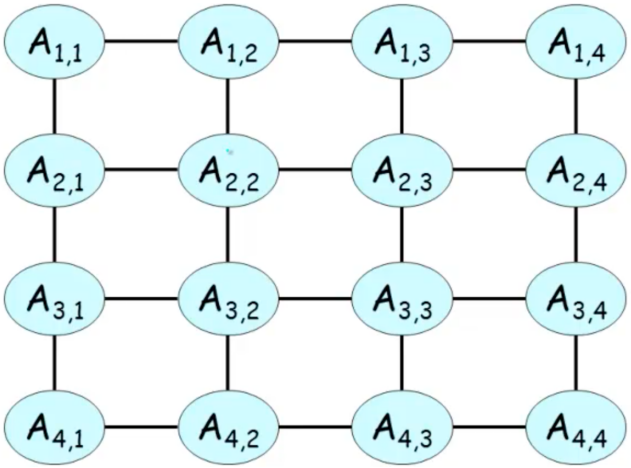
Super pixel

Markov Network is not Grid in general

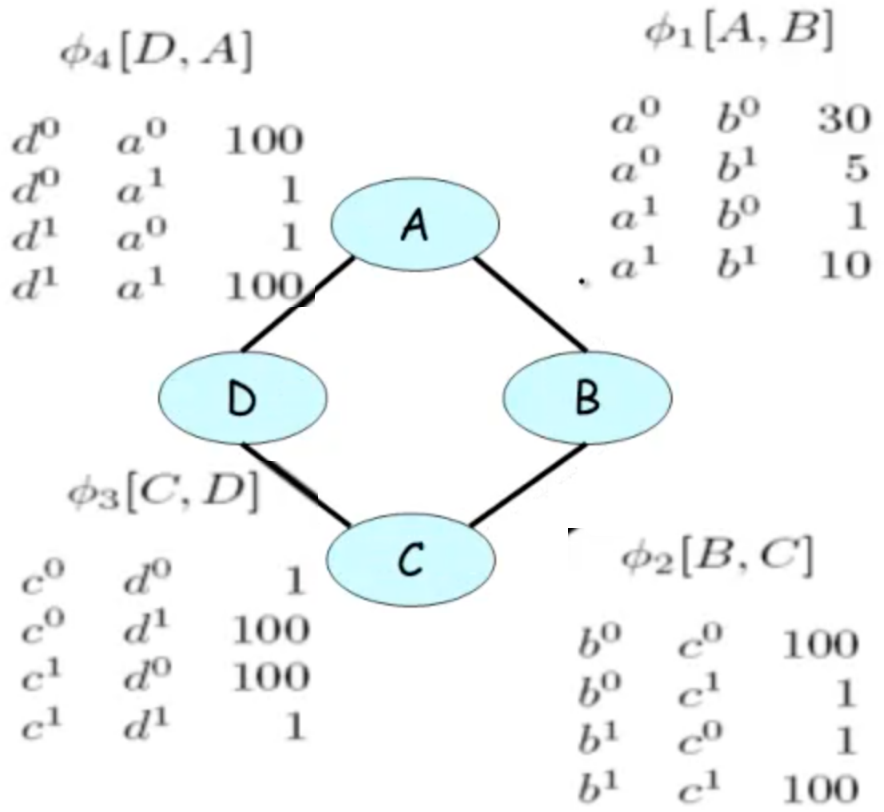
Grid Model

Grid Model

- Naturally occurs in image proc., lattice physics, etc.
- States of adjacent nodes are coupled due to pattern continuity or electro-magnetic force, etc.



How to Parameterize MRFs?



- Local factor models are attached to sets of nodes
  - Factor elements are positive
  - Do not have to sum to 1
  - Represent affinities, compatibilities

Factor Product:

$$\phi_1[A, B]\ \phi_2[B, C]\ \phi_3[C, D]\ \phi_4[D, A] \Leftrightarrow$$

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

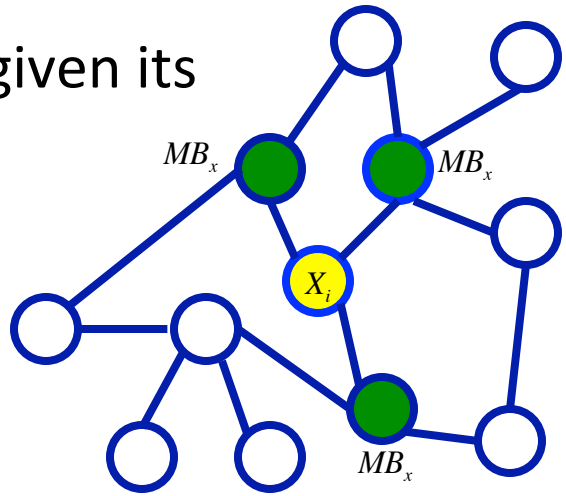
Local Markov Independencies

- For each node $X_i \in V$, there is a unique Markov blanket of X_i, denoted by MB_{X_i}, which is the set of immediate neighbors of X_i in the graph

- The local Markov independence associated with G is:

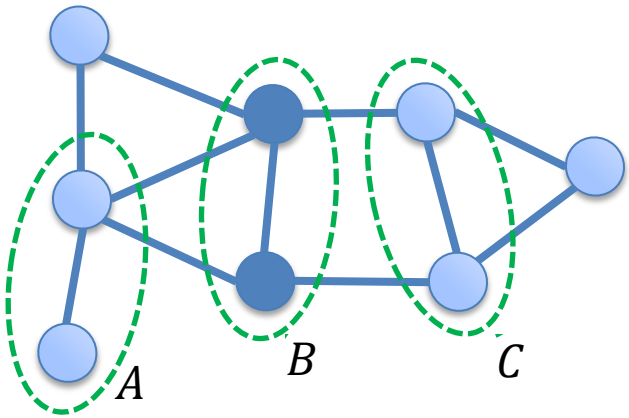
$$I_l(G) = \{X_i \perp V - \{X_i\} - MB_{X_i} \mid MB_{X_i} : \forall i\}$$

- In other words, X_i is independence of the rest given its immediate neighbors



Global Markov Independencies

- Let *H* be an undirected graph:



- *B* ***separates*** *A* and *C* if every path from a node in *A* to a node in *C* passes through a node in *B*: $\text{sep}_H(A;C|B)$
  - A path X_1-X_2-...-X_k is active if none of the X_i variables along the path are observed
- A probability distribution satisfies the ***global Markov property*** if for any disjoint *A*, *B*, *C*, such that *B* separates *A* and *C*, *A* is independent of *C* given *B:* $\text{I}(H) = \left\{A \perp C | B : \text{sep}_H(A;C|B)\right\}$

Markov Random Fields (UGM)

- Useful when edge directionality cannot be assigned

Simpler interpretation of structure

- Simpler inference
- Simpler independency structure

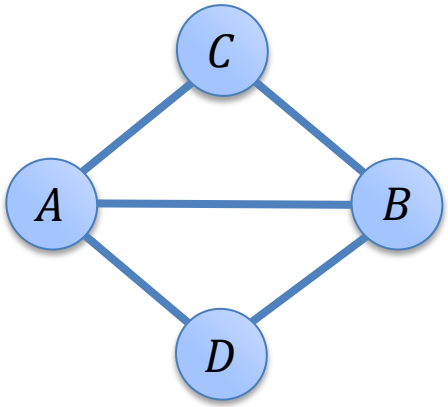
Harder to learn parameters/structures

We will also see models with combined directed and undirected edges

Maximal Cliques

- For $G = \{V, E\}$, a complete subgraph (clique) is a subgraph $G' = \{V' \subseteq V, E' \subseteq E\}$ s.t. nodes in V' are fully connected

- A (maximal) clique is a complete subgraph s.t. any superset V'', $V' \subset V''$, is not fully connected



- Example:
  - Maximal cliques = {A,B,C}, {A,B,D}
  - Sub-cliques = {A}, {B}, {A,B}, {C,D} … (all edges and singletons)

Definition of Undirected GM

- Given an undirected graph G over variables $\mathcal{X} = \{X_1, \dots, X_n\}$

- A distribution P factorizes over G if there exist
  - subset of variables $D_1 \subseteq \mathcal{X}$, …, $D_m \subseteq \mathcal{X}$ (D_i are maximal cliques in G)
  - non-negative potentials (factors/functions) $\Psi_1(D_1), \dots, \Psi_m(D_m)$
  - such that

$$P(X_1, X_2, \dots, X_n) = \frac{1}{Z} \prod_{i=1}^{m} \Psi_i(D_i)$$

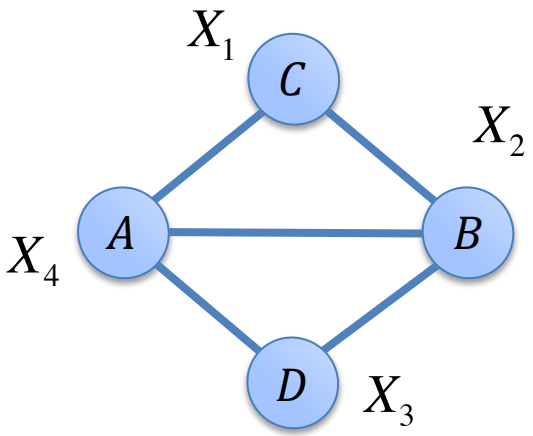
where

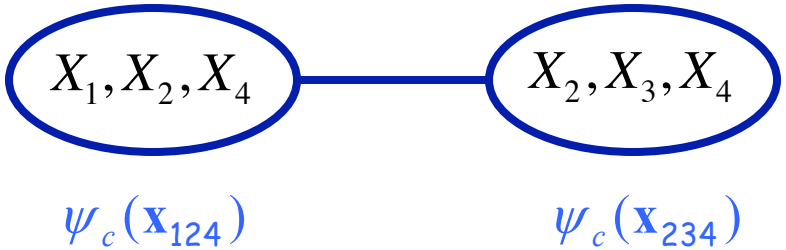
$$Z = \sum_{x_1, x_2, \dots, x_n} \prod_{i=1}^{m} \Psi_i(D_i)$$

- Z is called the partition function
- P is also called a Gibbs distribution over G

Also know as Gibbs distributions, Markov random Fields, and undirected graphical models

Example: Factorization Using Max Cliques

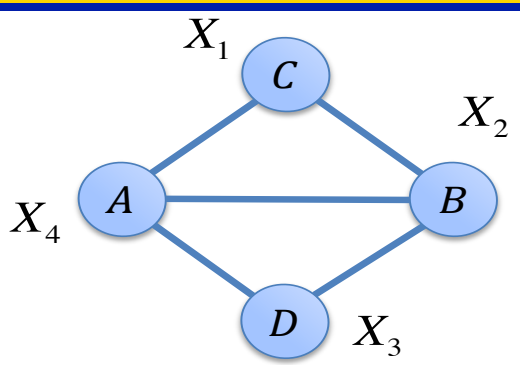


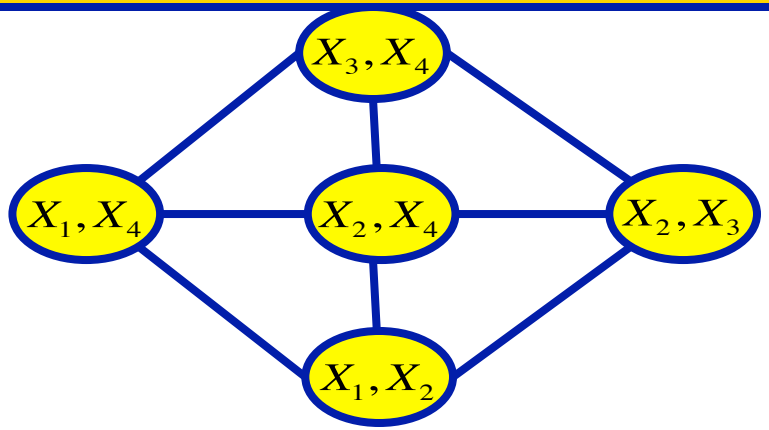


$$P'(x_1, x_2, x_3, x_4) = \frac{1}{Z}\psi_c(\mathbf{x}_{124}) \times \psi_c(\mathbf{x}_{234})$$

$$Z = \sum_{x_1, x_2, x_3, x_4} \psi_c(\mathbf{x}_{124}) \times \psi_c(\mathbf{x}_{234})$$

Example: Factorization Using Sub-cliques





$$P''(x_1, x_2, x_3, x_4) = \frac{1}{Z}\prod_{ij}\psi_{ij}(\mathbf{x}_{ij})$$

$$= \frac{1}{Z}\psi_{12}(\mathbf{x}_{12})\psi_{14}(\mathbf{x}_{14})\psi_{23}(\mathbf{x}_{23})\psi_{24}(\mathbf{x}_{24})\psi_{34}(\mathbf{x}_{34})$$

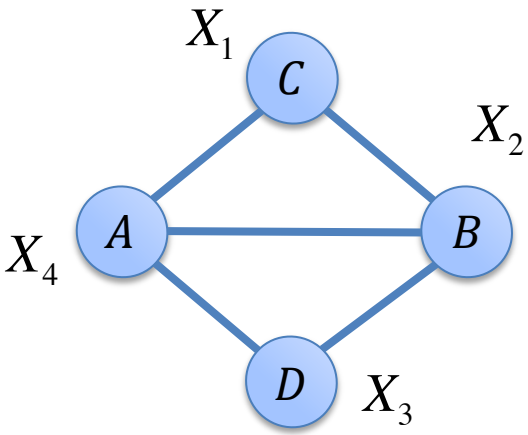
$$Z = \sum_{x_1, x_2, x_3, x_4}\prod_{ij}\psi_{ij}(\mathbf{x}_{ij})$$

- Pairwise Markov Random Field (PMRF),
  - a popular and simple special case.
- I(p') vs I(P'') ?
- D(p') vs D(P'') ?

- Node potentials $\Psi_{i(X_i)}$
- Edge potentials $\Psi_{ij}(X_i, X_j)$

$$P(X) = \prod_{i \in V}\Psi_i(X_i)\prod_{(i,j)\in E}\Psi_{ij}(X_i, X_j)$$

Example: Factorization Using Canonical Representation



$$P(x_1, x_2, x_3, x_4)$$

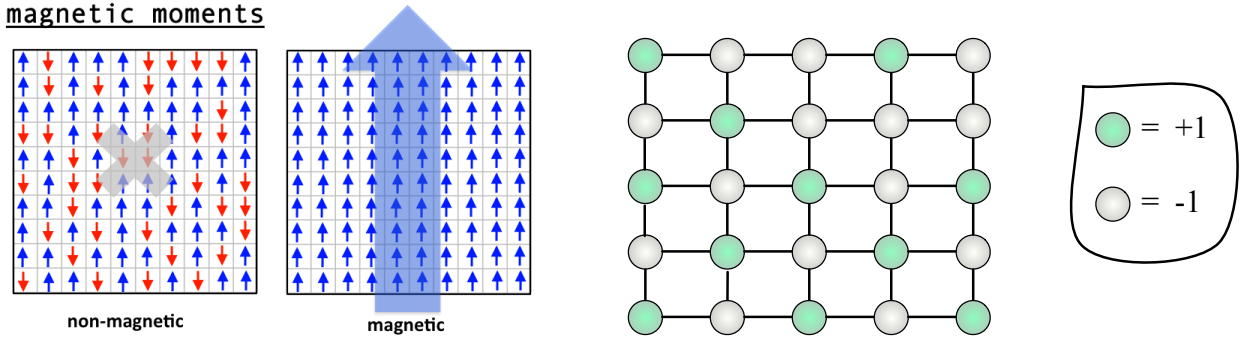
$$= \frac{1}{Z}\psi_c(\mathbf{x}_{124}) \times \psi_c(\mathbf{x}_{234})$$
$$\times \psi_{12}(\mathbf{x}_{12})\psi_{14}(\mathbf{x}_{14})\psi_{23}(\mathbf{x}_{23})\psi_{24}(\mathbf{x}_{24})\psi_{34}(\mathbf{x}_{34})$$
$$\times \psi_1(x_1)\psi_2(x_2)\psi_3(x_3)\psi_4(x_4)$$

where $Z = \sum_{x_1, x_2, x_3, x_4}$ $\psi_c(\mathbf{x}_{124}) \times \psi_c(\mathbf{x}_{234})$
$\times \psi_{12}(\mathbf{x}_{12})\psi_{14}(\mathbf{x}_{14})\psi_{23}(\mathbf{x}_{23})\psi_{24}(\mathbf{x}_{24})\psi_{34}(\mathbf{x}_{34})$
$\times \psi_1(x_1)\psi_2(x_2)\psi_3(x_3)\psi_4(x_4)$

- I(P) vs. I(p') vs. I(P'') ?
- Most general, subsume P' and P'' as special cases:
- D(p) is larger than D(p'), and D(P') is larger than D(p'')

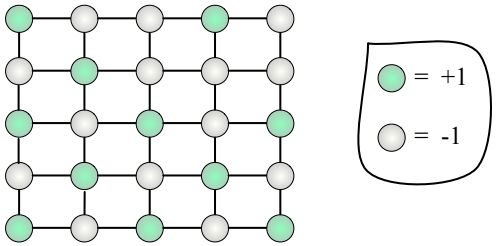
Example: Ising Model (I)

- Invented by the physicist Wilhelm Lenz (1920), who gave it as a problem to his student Ernst Ising
- Mathematical model of ferromagnetism in statistical mechanics
- The spin of an atom is biased by the spins of atoms nearby on the material:



- Each atom $X_i \in \{-1, +1\}$, whose value is the direction of the atom spin
- If a spin at position i is $+1$, what is the probability that the spin at position j is also $+1$?
- Are there phase transitions where spins go from "disorder" to "order"?

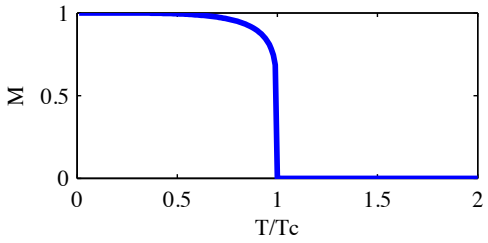
Example: Ising Model (II)



$$p(x_1, \cdots, x_n) = \frac{1}{Z} \exp\left(\sum_{i<j} w_{i,j} x_i x_j - \sum_i u_i x_i \right)$$

- When $w_{i,j} > 0$, nearby atoms encouraged to have the same spin (called **ferromagnetic**), whereas $w_{i,j} < 0$ encourages $X_i \neq X_j$
- Node potentials $\exp(-u_i x_i)$ encode the bias of the individual atoms
- Scaling the parameters makes the distribution more or less spiky

Spontaneous global behaviour



$M = |\sum_{i=1}^{N} x_i|/N$. As the temperature T decreases towards the critical temperature T_c a phase transition occurs in which a large fraction of the variables become aligned in the same state. Even though we only 'softly' encourage neighbours to be in the same state, for a low but finite T, the variables are all in the same state. Paradigm for 'emergent behaviour'.

I-Map in Markov Networks

- In BN, we defined I-map in terms of local Markov properties, and derived global independencies.
- In MN (Markov Networks), we define I-map in terms of global Markov properties.
  - Defn: An UG H is an I-map for a distribution P if $I(H) \subseteq I(P)$, i.e., P entails $I(H)$, i.e., $P \models \mathcal{I}(\mathcal{H})$.

Independencies in BN vs MN

- Bayesian network
  - Local independencies → Independence by d-separation (global)
- Markov network
  - Global independencies → Local independencies

- Can all independencies encoded by Bayesian networks be encoded by Markov networks?
  - No, immoral v-structures (explaining away)

- Can all independencies encoded by Markov networks be encoded by Bayesian networks?
  - No, counter example – $(A \perp B \mid C,D)$ and $(C \perp D \mid A,B)$

- Markov networks encode monotonic independencies
  - If $sep_H(\mathbf{X};\mathbf{Y}|\mathbf{Z})$ and $\mathbf{Z} \subseteq \mathbf{Z'}$ then $sep_H(\mathbf{X};\mathbf{Y}|\mathbf{Z'})$

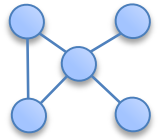
I-Map and Factorization in Markov Networks

- MN encodes global Markov assumptions $I_{(G)}$

If global conditional independence in MN are subset of conditional independence in P

$$I_{(G)} \subseteq I(P)$$

This MN is an I-map of P



obtain

P factorizes according to BN

Then the joint probability P can be written as

$$P(X_1, \dots, X_n) = \frac{1}{Z} \prod_i \Psi(D_i)$$

If the joint probability P can be written as

$$P(X_1, \dots, X_n) = \frac{1}{Z} \prod_i \Psi(D_i)$$

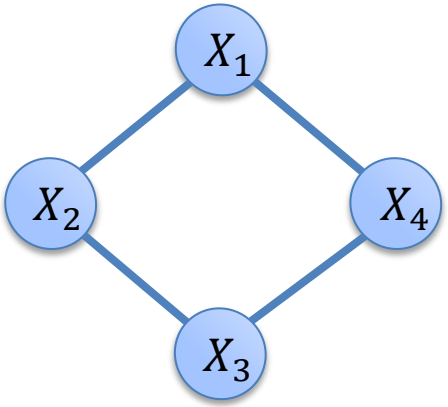
obtain

Then global conditional independence in MN are subset of conditional independence in P

$$I_{(G)} \subseteq I(P)$$

Counter Example

- $X_1, \ldots, X_4$ are binary, and only eight assignments have positive probability (each with 1/8)
  - (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)



$P(X_1, X_3 | X_2, X_4)$

X_1X_3 \ X_2X_4	00	01	10	11
00	½	½	0	0
01	0	½	0	½
10	½	0	½	0
11	0	0	½	½

$P(X_1 | X_2, X_4)$

X_1 \ X_2X_4	00	01	10	11
0	½	1	0	½
1	½	0	1	½

X_3 \ X_2X_4	00	01	10	11
0	1	½	½	0
1	0	½	½	1

$$P(X_1, X_3 | X_2, X_4) = P(X_1 | X_2, X_4) P(X_3 | X_2, X_4)$$

- But distribution does not factorize!
  - eg. $P(0,0,1,0) = 0 \neq \frac{1}{Z} \Psi_{12}(0,0)\Psi_{23}(0,1)\Psi_{34}(1,0)\Psi_{14}(0,0)$

-Song

Factorization in Markov Networks

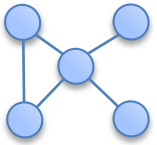
If global conditional independence in MN are subset of conditional independence in a strictly postive P

$$I(H) \subseteq I(P)$$

This MN is an I-map of P

Every P has least one MN structure H

obtain



P factorizes according to BN

Then the joint probability P can be written as

$$P(X_1, \dots, X_n) = \frac{1}{Z} \prod_i \Psi(D_i)$$

- For all x, $P(X = x) > 0$
- Known as Hammersley-Clifford Theorem

Hammersley-Clifford Theorem

- If arbitrary potentials are utilized in the following product formula for probabilities,

$$P(x_1, \ldots, x_n) = \frac{1}{Z} \prod_{c \in C} \psi_c(\mathbf{x}_c)$$

$$Z = \sum_{x_1, \ldots, x_n} \prod_{c \in C} \psi_c(\mathbf{x}_c)$$

  then the family of probability distributions obtained is exactly that set which **respects** the *qualitative specification* (the conditional independence relations)

- **Thm :** Let P be a positive distribution over $\mathbf{V}$, and *H* a Markov network graph over $\mathbf{V}$. If *H* is an I-map for P, then P is a Gibbs distribution over *H*.

-Xing

Soundness and Completeness of Global Markov Properties

- Defn: P is a Gibbs distribution over H if it can be represented as

$$P(x_1,\ldots,x_n) = \frac{1}{Z}\prod_{c\in C}\psi_c(\mathbf{x}_c)$$

- Thm (soundness): If P is a Gibbs distribution over H, then H is an I-map of P.

- Thm (completeness): If $\neg \text{sep}_H(X; Z \mid Y)$, then $X \not\perp_P Z \mid Y$ in ***some*** P that factorizes over H.

Relationship between various independencies in MN

Recall that in BN: Local and global independencies equivalent since one implies the other

In MN:

- Let I(H) be the global separation independencies
- Let $I_L(H)$ be the local (Markov blanket) independencies
- Let $I_P(H)$ be the pairwise independencies
  - Defn: The *pairwise Markov independencies* associated with UG $H = (V;E)$ are $I_p(H) = \{X \perp Y | V \setminus \{X,Y\} : \{X,Y\} \notin E\}$

  i.e., Pairs of non-adjacent variables are independent given all others.

Then:

In MN:

For any distribution P:

- $I(H) \rightarrow I_L(H)$ i.e., $I_L(H) \subseteq I(H)$ (if P entails I(H), then P also entails $I_L(H)$
- $I_L(H) \rightarrow I_P(H)$

For any positive distribution P:

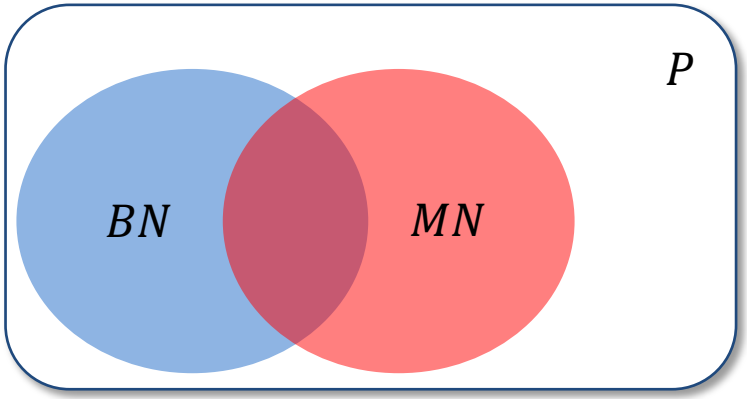
- $I(H) \leftrightarrow I_L(H) \leftrightarrow I_P(H)$

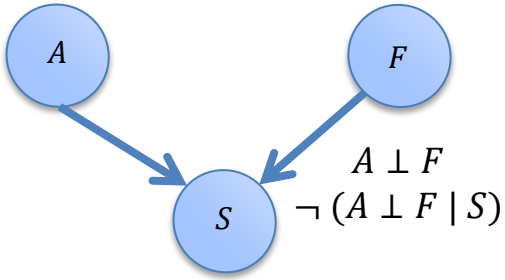
Minimal I-Map in Markov Networks

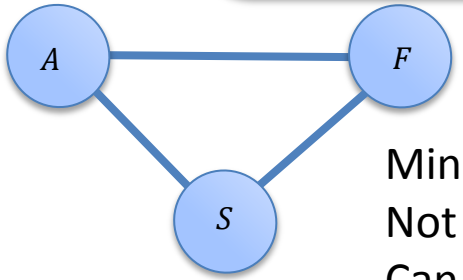
- A fully connected graph is an I-map for all distribution
- Remember minimal I-maps
  - Deleting an edge make it no longer I-map
- In a Bayesian Network, there is no unique minimal I-map
- For strictly positive distributions and Markov network, minimal I-map is unique!

Perfect Map in Markov Networks

- Perfect maps?
  - Independence in the graph are exactly the same as those in P
- For Bayesian networks, does not always exist
  - Counter example: swinging couple of variables
- How about for Markov networks?
  - Counter example: V-structure







Minimal I-map MN
Not a P-map
Can not code $A \perp F$

Construction of Markov Networks (I)

- Goal: Given a distribution, we want to construct a Markov network which is an I-map of P

Note that: If P is a positive distribution, then $I(H) \leftrightarrow I_L(H) \leftrightarrow I_P(H)$

- Thus, sufficient to construct a network that satisfies $I_P(H)$

Construction Algorithm:

- Take pairwise Markov assumption
- If P does not entail it, add edge
  - For every (X,Y) add edge if $(X \perp Y | \mathcal{U}-\{X,Y\})$ does not hold in P

Theorem: The resulting Markov network is minimal and unique I-map.

Algorithm does not work for non-positive P.

Construction of Markov Networks (II)

- Goal: Given a distribution, we want to construct a Markov network which is an I-map of P

Note that: If P is a positive distribution, then $I(H) \leftrightarrow I_L(H) \leftrightarrow I_P(H)$

- Thus, sufficient to construct a network that satisfies $I_L(H)$

Construction Algorithm:

- Connect each X to every node in the minimal set **Y** s.t.:
  $\{(X \perp \mathcal{U}-\{X\}-\mathbf{Y}|\mathbf{Y}) : X \in H\}$

Theorem: The resulting Markov network is minimal and unique I-map.

Algorithm does not work for non-positive P.

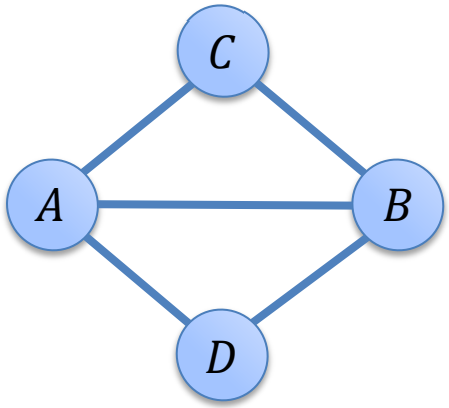
Parameterization of Markov Networks

- Markov networks have too many degrees of freedom
  - A clique over n binary variables has 2^n parameters but the joint has only 2^n-1 parameters
  - The network A—B—C has clique {A,B} and {B,C}
    - Both capture information on B which we can choose where we want to encode (in which clique)
    - We can add/subtract between the cliques
  - We can come up with infinitely many sets of factor values that lead to the same distribution

- Need: conventions for avoiding ambiguity in parameterization
  - Can be done using a canonical parameterization (see K&F 4.4.2.1)

Factor Graphs (Motivation)

- Maximal clique specification
  - $P(A,B,C,D) = \frac{1}{Z}\,\Psi_1(A,B,C)\Psi_2(A,B,D)$



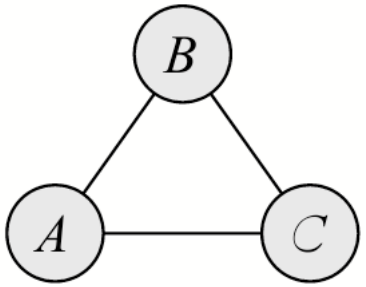
- Pairwise Markov Networks
  - $P'(A,B,C,D) = \frac{1}{Z}\Psi(AC)\Psi(BC)\Psi(AB)\Psi(AD)\Psi(BD)$

- Can not look at the graph and tell what potential is using

- Factor graph is to make this clear in graphical form

Factor Graph Definition

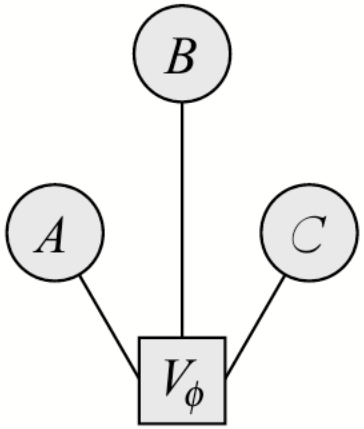
- Make factor dependency explicit
  - Useful for later inference



Markov network

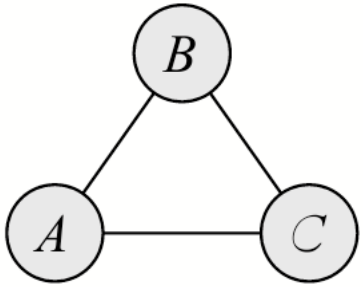
Factor Graph Definition:

- Bipartite graph:
  - Variable nodes (circle) for $X_1, \dots, X_n$
  - Factor nodes (square) for $\Psi_1, \dots, \Psi_m$
  - Edge $X_i - \Psi_j$ if $X_i \in D_j$ (Scope of $\Psi_j(D_j)$)



Factor graph for joint parameterization

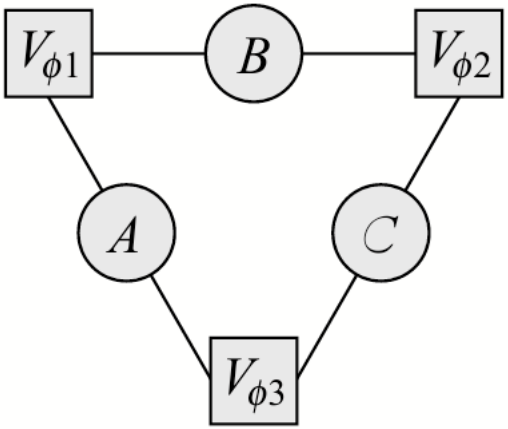
Multiple Factor Graphs for Same MRF

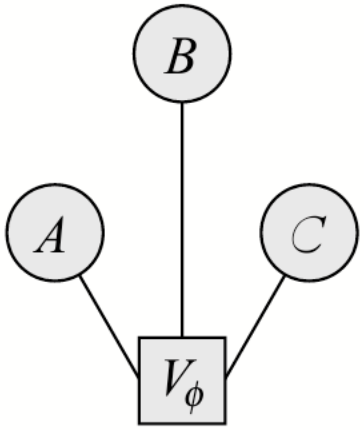


Markov network

Two Factor Graphs for the same Markov Network:

Factor graph for
pairwise parameterization

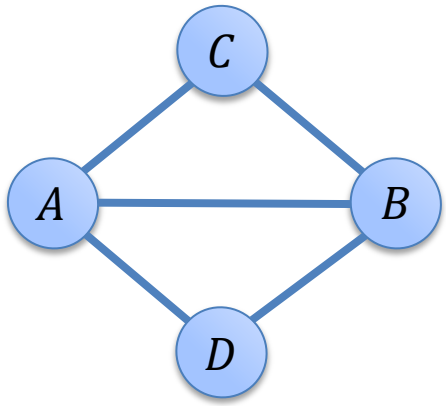




Factor graph for
joint parameterization

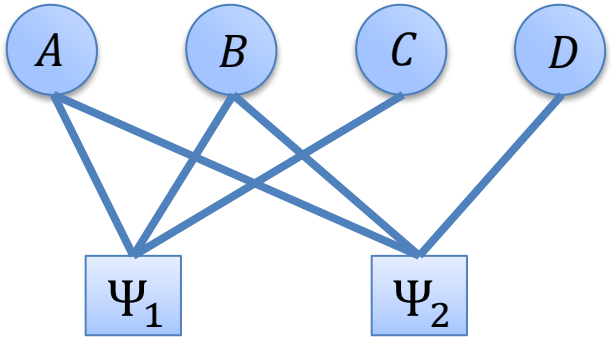
Examples of Factor Graphs

- Maximal clique specification
  - $P(A,B,C,D) = \frac{1}{Z}\,\Psi_1(A,B,C)\Psi_2(A,B,D)$

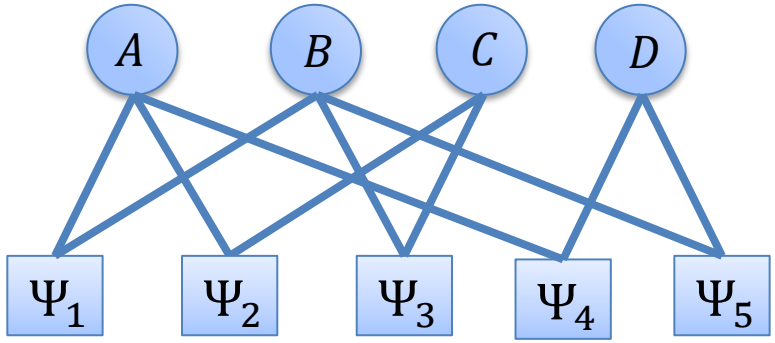


Markov network

- Pairwise Markov Networks
  - $P'(A,B,C,D) = \frac{1}{Z}\Psi(AC)\Psi(BC)\Psi(AB)\Psi(AD)\Psi(BD)$



$\frac{1}{Z}\,\Psi_1(A,B,C)\Psi_2(A,B,D)$



$\frac{1}{Z}\Psi_1(AB)\Psi_2(AC)\Psi_3(BC)\Psi_4(AD)\Psi_5(BD)$

Factor Graphs sometimes facilitate inference for MN and BN by creating factor tree.

Logarithmic Representation of MRF

A set of subsets $\mathbf{D}_1,\ldots\mathbf{D}_m$ where each $\mathbf{D}_i$ is a complete (fully connected) subgraph in H

- Standard model: $P(X_1,\ldots,X_n) = \frac{1}{Z}\prod_i \Psi(D_i)$
- Assuming strictly positive potentials:
  - $P(X_1,\ldots,X_n) = \frac{1}{Z}\prod_i \Psi(D_i)$
  - $= \frac{1}{Z}\prod_i \exp\left(\log\left(\Psi(D_i)\right)\right)$
  - $= \frac{1}{Z}\exp\left(\sum_i \log\left(\Psi(D_i)\right)\right)$
  - $= \frac{1}{Z}\exp\left(\sum_i \Phi(D_i)\right)$

Let: $\Psi(D_i) = \exp(\Phi(D_i))$

$-\Phi(D_i)$ is called energy function

Log P(X) is a linear function

- We can maintain table $\Phi(D_i)$ (can have negative entries) rather than table $\Psi(D_i)$ (strictly positive entries)

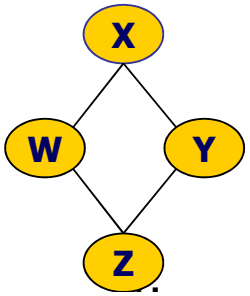
Exponential Form: Log-Linear Model of MRF

Let $\boldsymbol{D}$ *be a subset of variables. We define a* feature $f(\boldsymbol{D})$ *to be a function from* $Val(\boldsymbol{D})$ *to* $\mathbb{R}$.

Example of feature:

A feature $\phi[\mathbf{D}]$ on variables **D** is an indicator function that for some y$\in$**D**:

$$\phi[\mathbf{D}] = \begin{cases} 1 & \text{when } x = w \\ 0 & \text{otherwise} \end{cases}$$



A distribution P *is a* log-linear model *over a Markov network* $\mathcal{H}$ *if it is associated with:*

- *a set of features* $\mathcal{F} = \{f_1(\boldsymbol{D}_1), \ldots, f_k(\boldsymbol{D}_k)\}$, *where each* $\boldsymbol{D}_i$ *is a complete subgraph in* $\mathcal{H}$,
- *a set of weights* $w_1, \ldots, w_k$,

such that

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

Feature Representation (Log-Linear Model) of MRF

A feature is simply a factor without the nonnegativity requirement. One type of feature of particular interest is the *indicator feature* that takes on value 1 for some values $\boldsymbol{y} \in Val(\boldsymbol{D})$ and 0 otherwise.

- Several features can be defined on one clique

  → any factor can be represented by features, where in the most general case we define a feature and weight for each entry in the factor

- Log-linear model is more compact for many distributions especially with large domain variables

- Representation is intuitive and modular
  - Features can be modularly added between any interacting sets of variables

Summary: Markov Network Parameterizations

- Choice 1: Markov network
  - Product over potentials
  - Right representation for discussing independence queries
- Choice 2: Factor graph
  - Product over graphs
  - Useful for inference (later)
- Choice 3: Log-linear model
  - Product over feature weights
  - Useful for discussing parameterizations
  - Useful for representing context specific structures
- All parameterizations are interchangeable

Next lecture

- From BN to MN
- From MN to BN
- Restricted Boltzmann Machines (RBM)
- Partially Directed Graphs
- Conditional random fields (CRF)