ECE 6504: Advanced Topics in Machine Learning

Probabilistic Graphical Models and Large-Scale Learning

Topics

- Markov Random Fields: Representation
  - Pairwise MRFs, Gibbs distribution
  - Conditional Random Fields

Readings: KF 4.1-3; Barber 4.1-2

Dhruv Batra

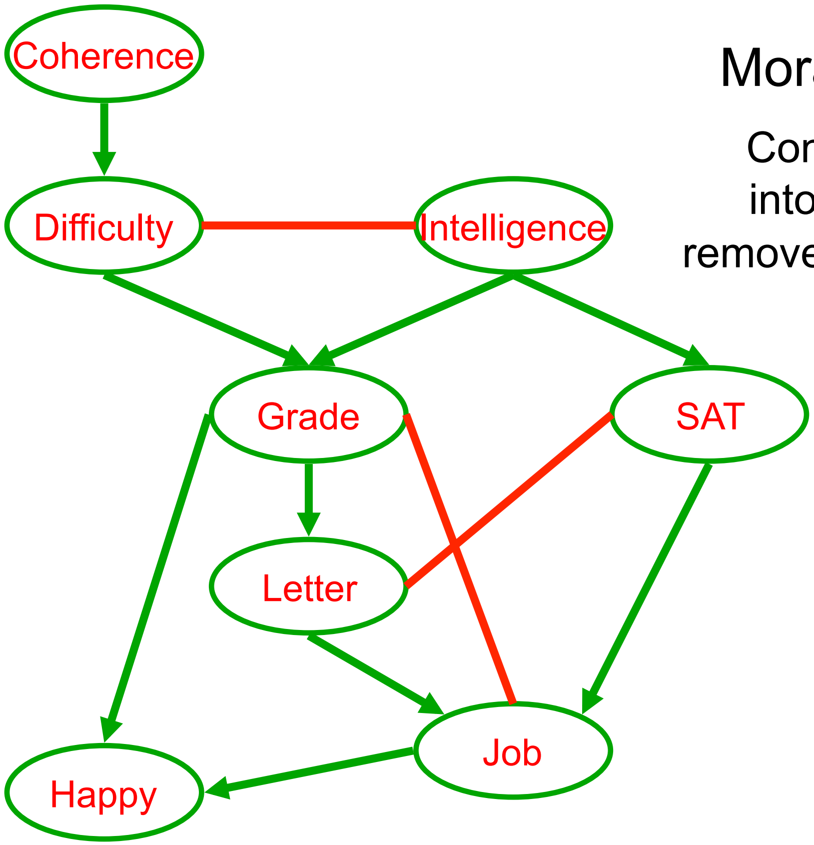
Virginia Tech

Administrativia

- Project Proposal
  - Due: ~~Mar 12~~, Mar 5, 11:59pm
  - <=2pages, NIPS format

Recap of Last Time

Moralization – "Marry" Parents

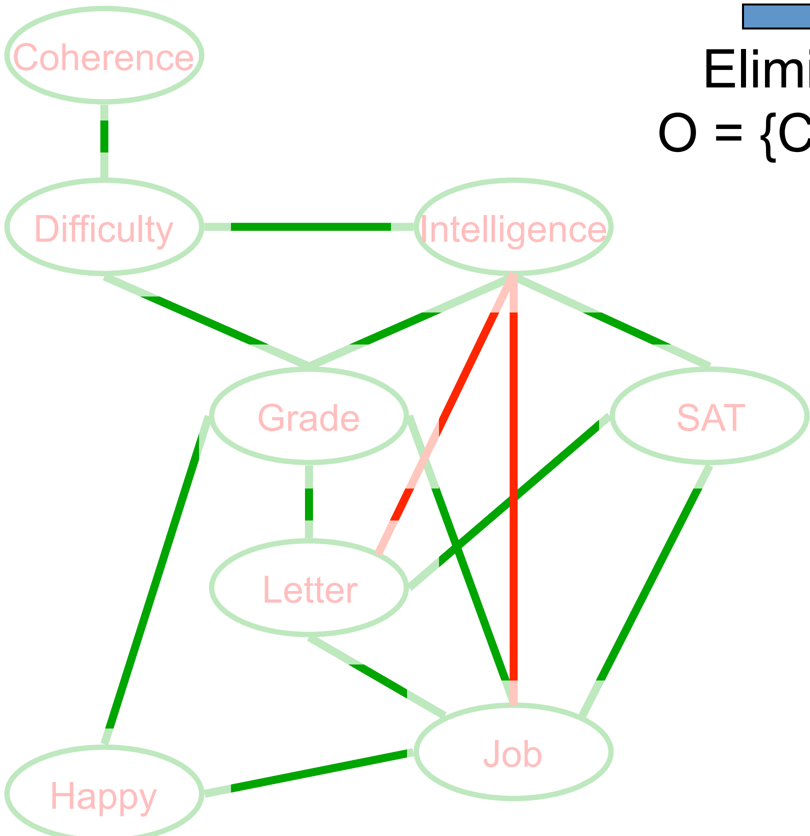


Moralize graph:

Connect parents
into a clique and
remove edge directions

Connect nodes that appear together in an initial factor

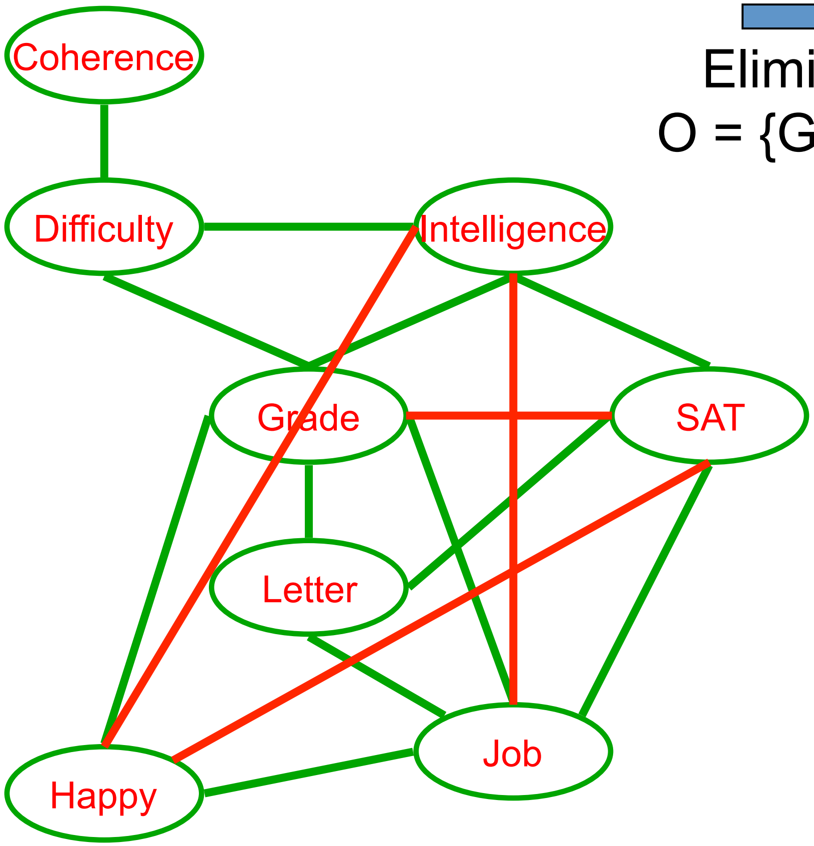
Induced graph



Elimination order:
O = {C,D,S,I,L,H,J,G}

The induced graph I_{FO} for elimination order O has an edge $X_i - X_j$ if X_i and X_j appear together in a factor generated by VE for elimination order O on factors F

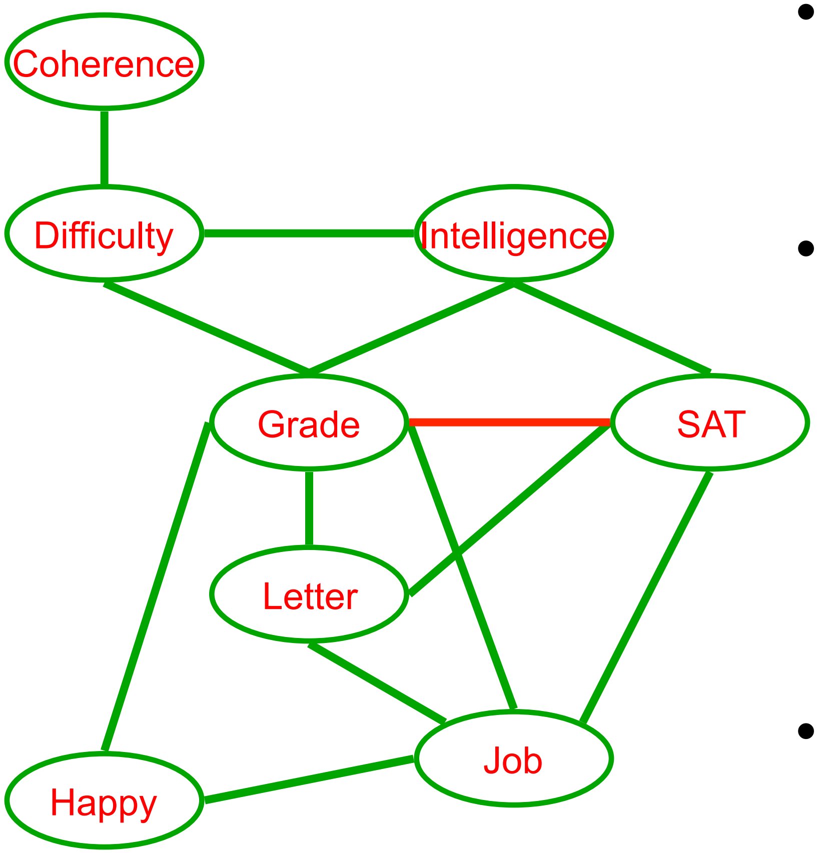
Different elimination order can lead to different induced graph



Elimination order:
O = {G,C,D,S,I,L,H,J}

Induced graph and complexity of VE

Read complexity from cliques in induced graph



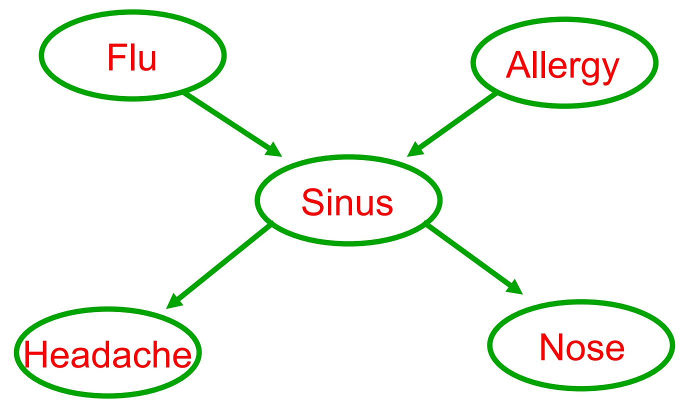
Elimination order:
O = {C,D,I,S,L,H,J,G}

- Structure of induced graph encodes complexity of VE!!!
- **Theorem**:
  - Every factor generated by VE is a clique in I_{FO}
  - Every maximal clique in I_{FO} corresponds to a factor generated by VE
- **Induced width**
  - Size of largest clique in I_{FO} minus 1
- **Treewidth**
  - induced width of best order O*

Plan for today

- Undirected Graphical Models
  - Pairwise MRFs
  - General Gibbs distribution
  - Active trails, separation
  - I-maps, P-maps
  - Conditional Random Fields

A general Bayes net



- Set of random variables
- Directed acyclic graph
  - Encodes independence assumptions
- CPTs
  - Conditional Probability Tables
- Joint distribution:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P\left(X_i \mid \mathbf{Pa}_{X_i}\right)$$

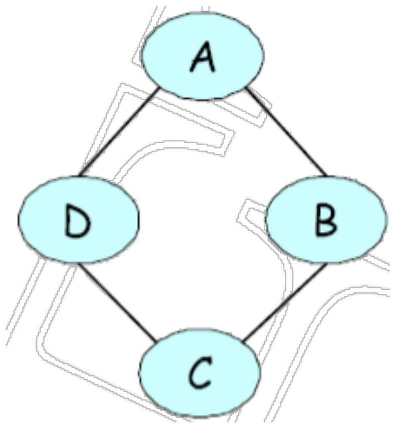
Markov Nets

- Set of random variables
- Undirected graph
  - Encodes independence assumptions
- Unnormalized Factor Tables
- Joint distribution:
  - Product of Factors

Pairwise MRFs

- Pairwise Factors
  - A function of 2 variables
    - Often unary terms are also allowed (although strictly speaking unnecessary)
  - On board

Pairwise MRF: Example



$\phi_1[A,B]$			$\phi_2[B,C]$			$\phi_3[C,D]$			$\phi_4[D,A]$		
a^0	b^0	30	b^0	c^0	100	c^0	d^0	1	d^0	a^0	100
a^0	b^1	5	b^0	c^1	1	c^0	d^1	100	d^0	a^1	1
a^1	b^0	1	b^1	c^0	1	c^1	d^0	100	d^1	a^0	1
a^1	b^1	10	b^1	c^1	100	c^1	d^1	1	d^1	a^1	100

Computing probabilities in Markov networks vs BNs

- In a BN, can compute prob. of an instantiation by multiplying CPTs

- In an Markov networks, can only compute ratio of probabilities directly

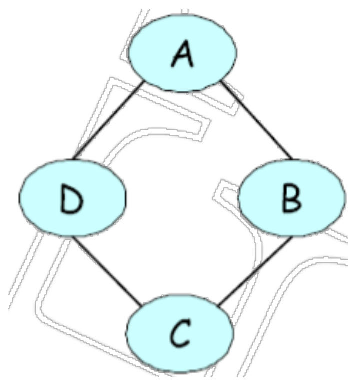
$\phi_1[A,B]$			$\phi_2[B,C]$			$\phi_3[C,D]$			$\phi_4[D,A]$		
a^0	b^0	30	b^0	c^0	100	c^0	d^0	1	d^0	a^0	100
a^0	b^1	5	b^0	c^1	1	c^0	d^1	100	d^0	a^1	1
a^1	b^0	1	b^1	c^0	1	c^1	d^0	100	d^1	a^0	1
a^1	b^1	10	b^1	c^1	100	c^1	d^1	1	d^1	a^1	100

Normalization for computing probabilities

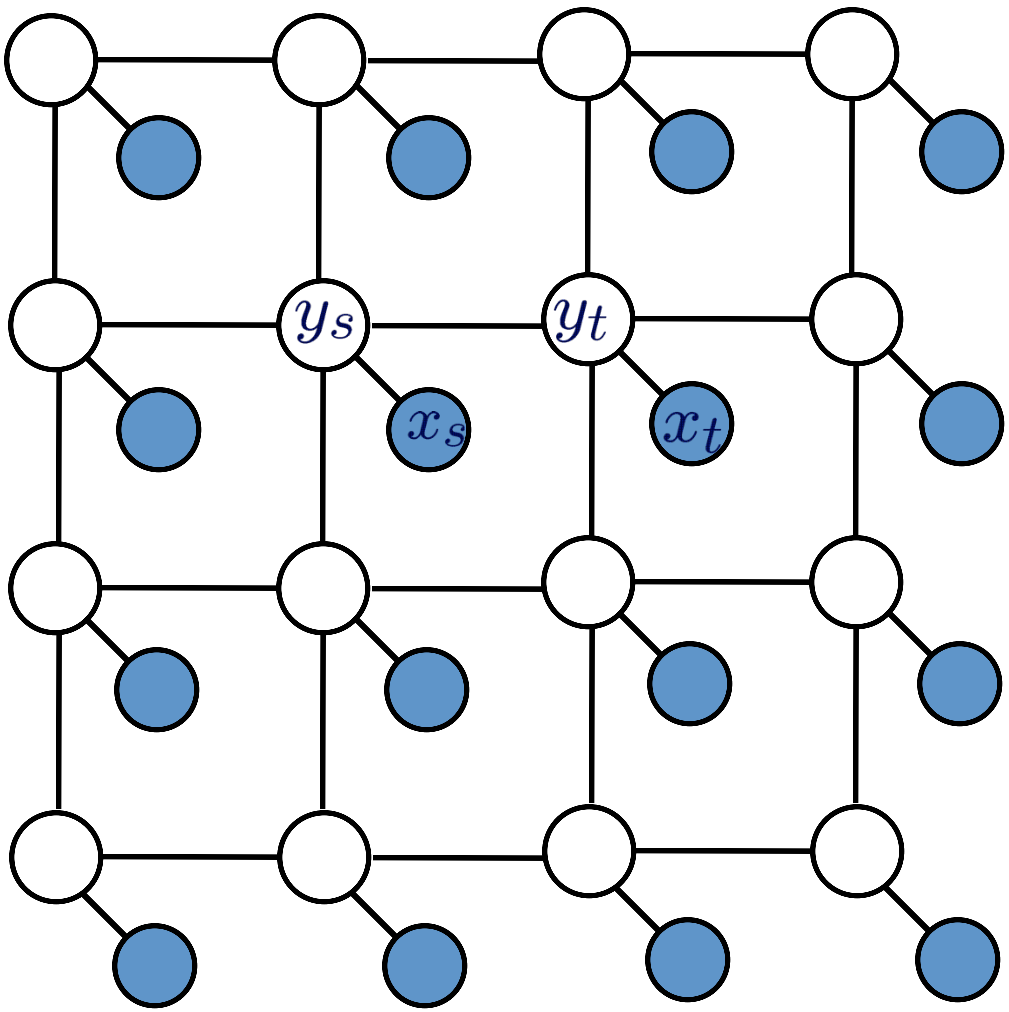
- To compute actual probabilities, must compute normalization constant (also called partition function)

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

- Computing partition function is hard! Must sum over all possible assignments



Nearest-Neighbor Grids



Low Level Vision

- Image denoising
- Stereo
- Optical flow
- Shape from shading
- Superresolution
- Segmentation

y_s → unobserved or hidden variable

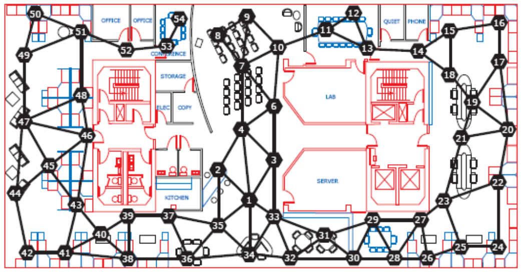
x_s → local observation

General Gibbs Distribution

- Arbitrary Factors

- “Induced” MRF Graph

Factorization in Markov networks



- Given an undirected graph *H* over variables **X**={X_1,...,X_n}

- A distribution *P* **factorizes** over *H* if there exist
  - subsets of variables $\mathbf{D}_1 \subseteq \mathbf{X}, \ldots, \mathbf{D}_m \subseteq \mathbf{X}$, such that $\mathbf{D}_i$ are *fully connected* in *H*
  - *non-negative potentials* (or factors) $\phi_1(\mathbf{D}_1), \ldots, \phi_m(\mathbf{D}_m)$
    - also known as clique potentials
  - such that

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

- Also called Markov random field *H,* or Gibbs distribution over *H*

MRFs

- Given a graph H, are factors unique?

Local Structures in BNs

- Causal Trail
  - X → Y → Z
- Evidential Trail
  - X ← Y ← Z
- Common Cause
  - X ← Y → Z
- Common Effect (v-structure)
  - X → Y ← Z

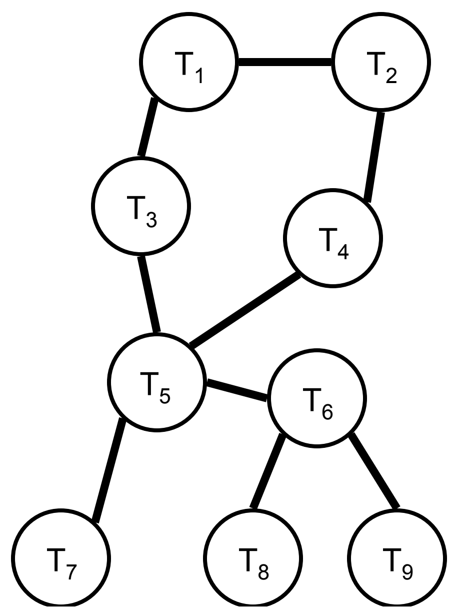
Local Structures in MNs

- On board

Active Trails and Separation

- A path $X_1 - \ldots - X_k$ is **active** when set of variables **Z** are observed
  - if none of $X_i \in \{X_1,\ldots,X_k\}$ are observed (are part of **Z**)

- Variables **X** are **separated** from **Y** given **Z** in graph
  - If no active path between any $X \in$ **X** and any $Y \in$ **Y** given **Z**



Separation

- So what if **X and Y** are separated given ***Z*?**

Factorization + d-sep ➔ Independence

- Theorem:
  - If
    - P factorizes over G
    - d-sep$_G$(**X**, **Y** | **Z**)
  - Then
    - $P \vdash (\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})$
  - Corollary:
    - $I(G) \subseteq I(P)$
    - All independence assertions read from G are correct!

The BN Representation Theorem

If G is an I-map of P → Obtain → **P factorizes to G**

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P\left(X_i \mid \mathbf{Pa}_{X_i}\right)$$

Important because:
Every P* has at least one BN structure *G

P factorizes to G

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P\left(X_i \mid \mathbf{Pa}_{X_i}\right)$$

→ Obtain → **G is an I-map of P**

Important because:
Read independencies of *P* from BN structure *G*

Markov networks representation Theorem 1

If joint probability distribution *P*:

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

Then

H is an I-map for *P*

- If
  - you can write distribution as a normalized product of factors
- Then
  - Can read independencies from graph

What about the other direction for Markov networks ?

If H is an I-map for P **Then** joint probability distribution P:

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

- Counter-example: $X_1,\ldots,X_4$ are binary, and only eight assignments have positive probability:

(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)~~

- For example, $X_1 \perp X_3 | X_2, X_4$:
  - E.g., $P(X_1=0|X_2=0, X_4=0)$

- But distribution doesn't factorize!!

Representation Theorem for Markov Networks

If joint probability distribution *P*:

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

Then →

H is an I-map for *P*

If *H* is an I-map for *P*
and
P is a positive distribution

Then →

joint probability distribution *P*:

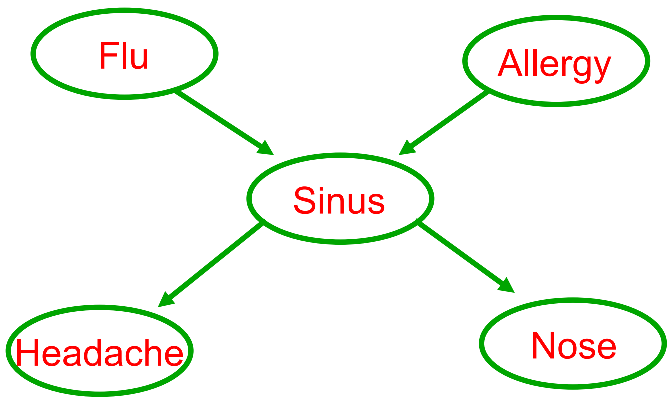
$$P(X_1, \ldots, X_n) = \frac{1}{Z} \prod_{i=1}^{m} \phi_i(\mathbf{D}_i)$$

Completeness of separation in MNs

- **Theorem: Completeness of separation**
  - For “almost all” distributions where *P* factorizes over Markov network *H*
  - we have that I(*H*) = I(*P*)
  - *“almost all” distributions*
    - except for a set of measure zero of parameterizations of the Potentials (assuming no finite set of parameterizations has positive measure)
    - Means that if **X** & **Y** are not separated given **Z**, then $P \neg (\mathbf{X} \perp \mathbf{Y} | \mathbf{Z})$

- **Analogous to BNs**

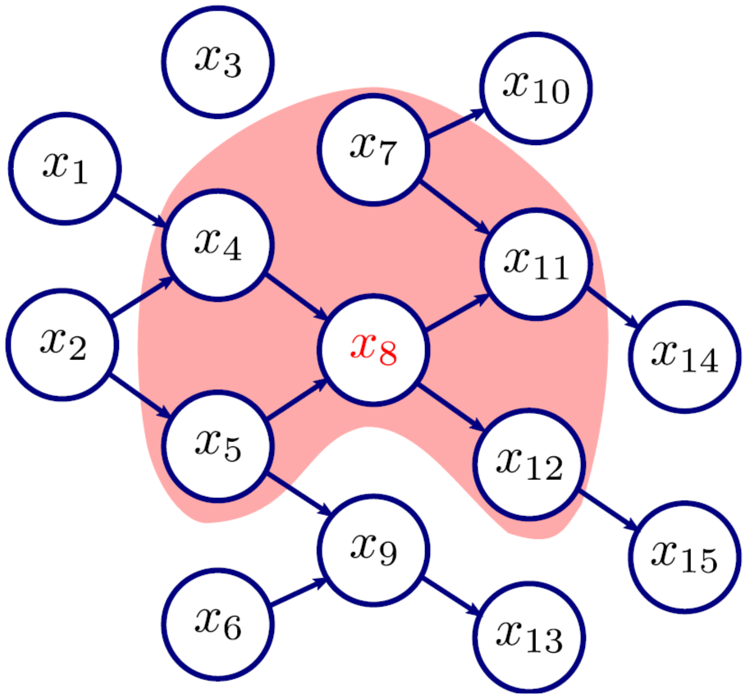
Local Markov Assumption



A variable X is independent of its non-descendants given its parents and only its parents

$(X_i \perp \mathbf{NonDescendants}_{X_i} \mid \mathbf{Pa}_{X_i})$

Markov Blanket



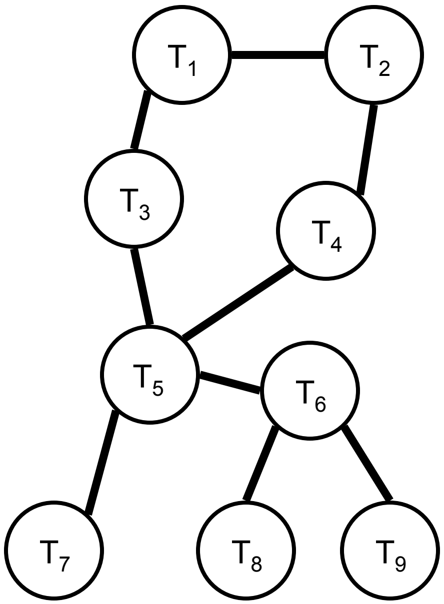
= Markov Blanket of variable x_8 – Parents, children and parents of children

Independence Assumptions in MNs

- **Separation** defines global independencies

- **Pairwise Markov Independence**:
  - Pairs of non-adjacent variables A,B are independent given all others

- **Markov Blanket**:
  - Variable A independent of rest given its neighbors



P-map

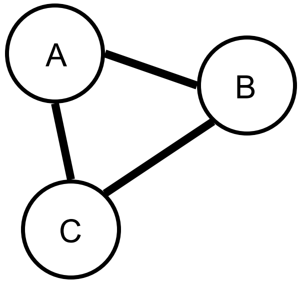
- Perfect map

- *G* is a **P-map** for *P* if
  - I(*P*) = I(*G*)

- Question: Does every distribution *P* have P-map?

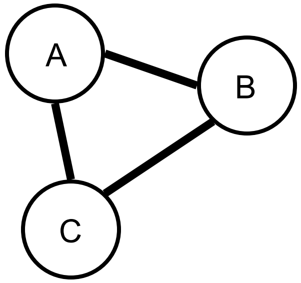
Structure in cliques

- Possible potentials for this graph:



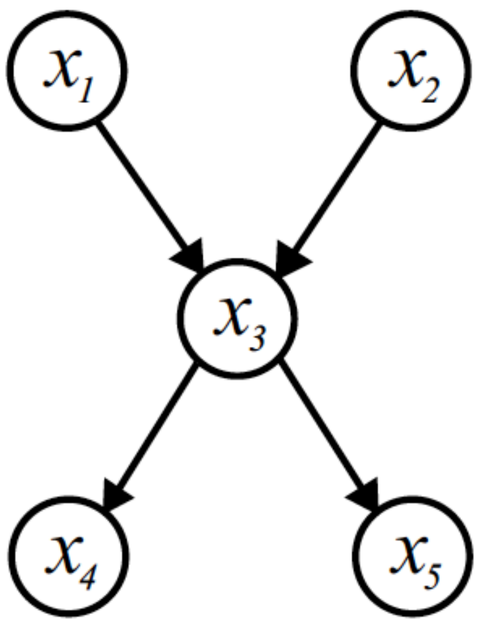
Factor graphs

- Bipartite graph:
  - variable nodes (ovals) for $X_1,\ldots,X_n$
  - factor nodes (squares) for $\phi_1,\ldots,\phi_m$
  - edge $X_i - \phi_j$ if $X_i \in Scope[\phi_j]$

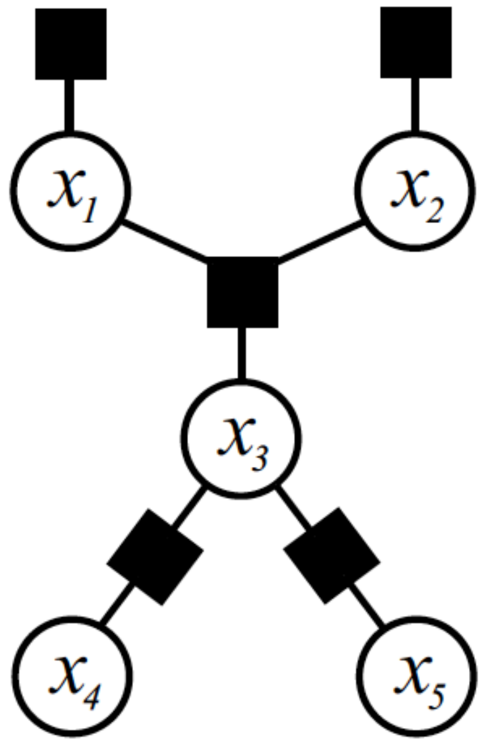


- Very useful for approximate inference
  - Make factor dependency explicit

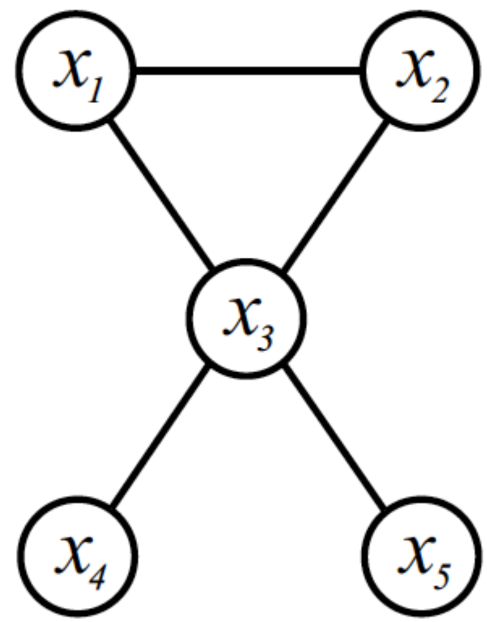
Types of Graphical Models



Directed



Factor



Undirected

Factor Graphs show Fine-grained Factorization

$$p(x) = \frac{1}{Z} \prod_{f \in \mathcal{F}} \psi_f(x_f)$$