Probabilistic Graphical Models

Undirected Graphical Models

Probabilistic Graphical Models

Directed graphical models

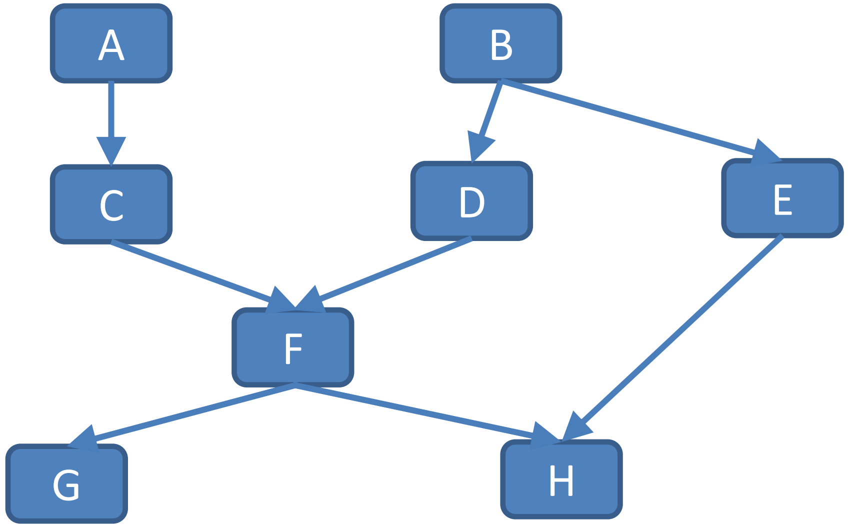
- Bayes Nets
- Conditional dependence

Undirected graphical models

- Markov random fields (MRFs)
- Factor graphs

From Markov Condition to Factorization

A Directed Acyclic Graph



A joint Probability Distribution

$$P(A, B, C, D, E, F, G, H)$$

$$P(A, \ldots, H) = \prod_{V \in \{A, \ldots, H\}} P(V \mid Pa_G(V))$$

Markov Condition:
Every variable is independent of its non-descendants given its parents (in the graph)

Summary

BN: DAG + Distribution

The distribution factorizes according to the graph based on the Markov condition: Every variable is independent from its non-descendants (in the graph) based on its parents (in the graph)

D-separation allows us to read the independencies from the graph.

sound (dsep->ind) and
complete (dcon->dep in some distribution that factorizes according to G)

If $I(G) \subseteq I(P)$ then G is an I-Map for P

Summary

BN: DAG + Distribution

The distribution factorizes according to the graph based on the Markov condition: Every variable is independent from its non-descendants (in the graph) based on its parents (in the graph)

D-separation allows us to read the independencies from the graph.

- sound (dsep->ind) and
- complete (dcon->dep in some distribution that factorizes according to G)

Summary

BN: DAG + Distribution

If $I(G) \subseteq I(P)$ then G is an I-Map for P

If G is an I-Map for P and every çthat stems from removing an edge from G is not an I-Map for P, G is minimal I-Map for P

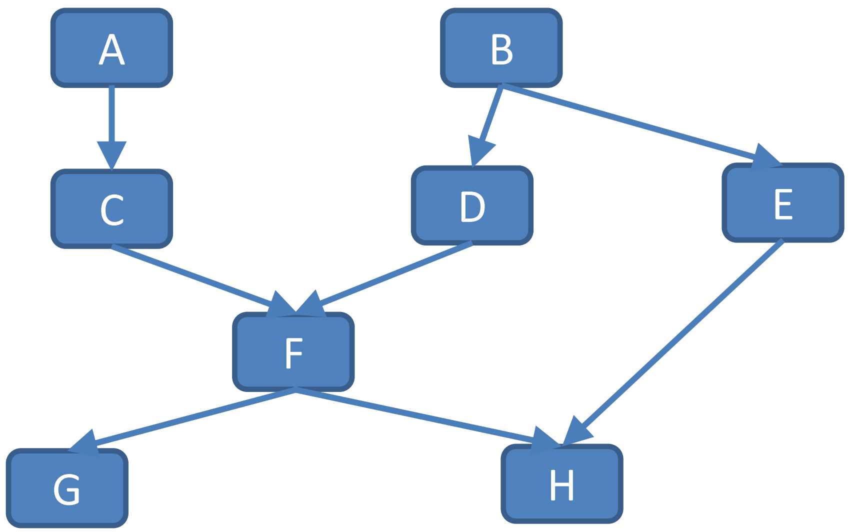
If $I(G) = I(P)$ then G is a perfect map for P

If $I(G) = I(G')$, G and G' are Markov Equivalent (I-Equivalent)

The Markov Boundary of Y is the set of Parents, Children and Spouses of G

From Markov Condition to Factorization

A Directed Acyclic Graph



A joint Probability Distribution

$$P(A, B, C, D, E, F, G, H)$$

$$P(A, \dots, H) = \prod_{V \in \{A, \dots, H\}} P(V | Pa_G(V))$$

Markov Condition:
Every variable is independent of its non-descendants given its parents (in the graph)

Example: Misconception

Four students who get together in pairs to work on homeworks.

Only the following pairs meet:

Alice and Bob;
Bob and Charles;
Charles and Debbie;
Debbie and Alice.

(Alice and Charles just can't stand each other, and Bob and Debbie had a relationship that ended badly.)
Probability of having misunderstood something in the class

Example: Misconception

Four students who get together in pairs to work on homeworks.

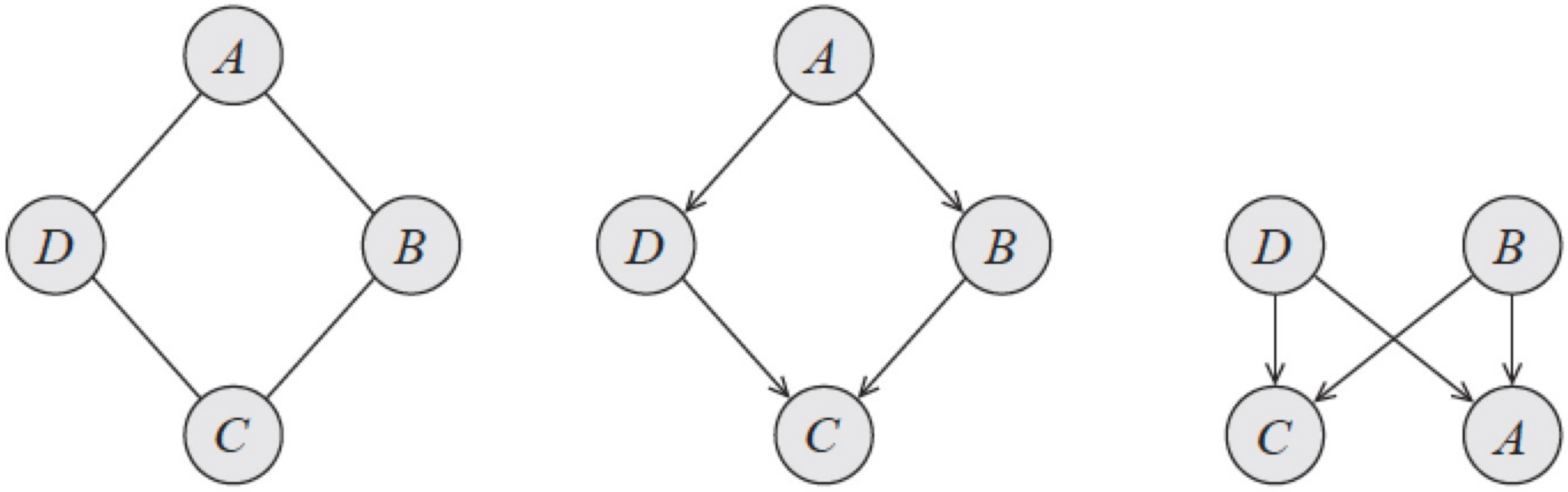
Only the following pairs meet:

Alice and Bob;
Bob and Charles;
Charles and Debbie;
Debbie and Alice.

(Alice and Charles just can't stand each other, and Bob and Debbie had a relationship that ended badly.)
Probability of having misunderstood something in the class

Ind(A, C|B,D)
Ind(B, D|A, C)

Example: Misconception

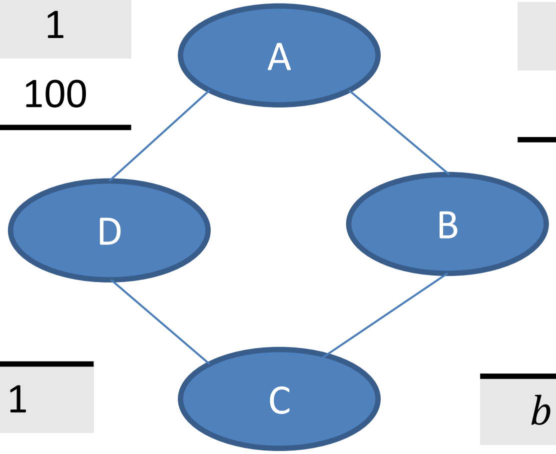


(a) (b) (c)

Pairwise Markov Networks

d^0	α^0	100
d^0	a^1	1
d^1	a^0	1
d^1	a^1	100

a^0	b^0	30
α^0	b^1	5
α^1	b^0	1
α^1	b^1	100



c^0	d^0	1
c^0	d^1	100
c^1	d^0	100
c^1	d^1	1

b^0	c^0	100
b^0	c^1	1
b^1	c^0	1
b	c^1	100

Factors

A factor $\phi(X_{1,} \ \dots, X_k)$

$$\phi: \mathrm{Val}(X_1, \dots, X_k) \rightarrow \mathbb{R}$$

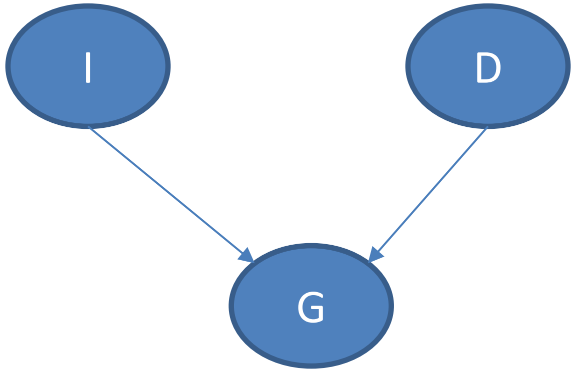
Scope $= \{X_{1,} \dots, X_k\}$

Fundamental building block for defining distributions in high-dimensional spaces

Set of basic operations for manipulating these probability distributions

Example: JPD

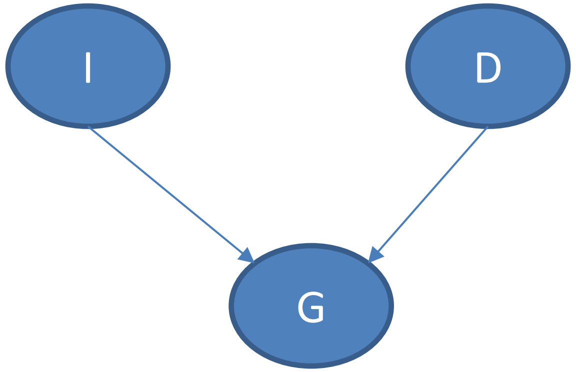
I	D	G	**Prob.**
i^0	d^0	g^1	0.126
i^0	d^0	g^2	0.168
i^0	d^0	g^3	0.126
i^0	d^1	g^1	0.009
i^0	d^1	g^2	0.045
i^0	d^1	g^3	0.126
i^1	d^0	g^1	0.252
i^1	d^0	g^2	0.0224
i^1	d^0	g^3	0.0056
i^1	d^1	g^1	0.06
i^1	d^1	g^2	0.036
i^1	d^1	g^3	0.024



$\text{Scope} = \{\text{I}, \text{D}, \text{G}\}$

Unnormalized measure

I	D	G	**Prob.**
i^0	d^0	g^1	0.126
i^0	d^0	g^2	0.168
i^0	d^0	g^3	0.126
i^0	d^1	g^1	0.009
i^0	d^1	g^2	0.045
i^0	d^1	g^3	0.126
i^1	d^0	g^1	0.252
i^1	d^0	g^2	0.0224
i^1	d^0	g^3	0.0056
i^1	d^1	g^1	0.06
i^1	d^1	g^2	0.036
i^1	d^1	g^3	0.024

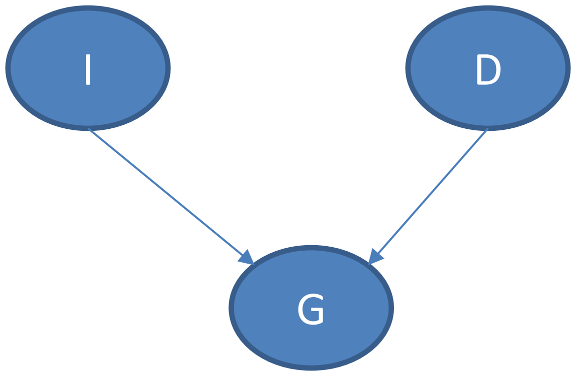


I	D	G	**Prob.**
i^0	d^0	g^1	0.126
i^0	d^1	g^1	0.009
i^1	d^0	g^1	0.252
i^1	d^1	g^1	0.06

$\text{Scope} = \{\text{I}, \text{D}\}$

Conditional Distribution

I	D	G	**Prob.**
i^0	d^0	g^1	0.126
i^0	d^0	g^2	0.168
i^0	d^0	g^3	0.126
i^0	d^1	g^1	0.009
i^0	d^1	g^2	0.045
i^0	d^1	g^3	0.126
i^1	d^0	g^1	0.252
i^1	d^0	g^2	0.0224
i^1	d^0	g^3	0.0056
i^1	d^1	g^1	0.06
i^1	d^1	g^2	0.036
i^1	d^1	g^3	0.024



	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

Scope $= \{I, D, G\}$

General Factors

D	A	$\phi(\mathrm{D}, \mathrm{A})$
d^0	α^0	100
d^0	a^1	1
d^1	a^0	1
d^1	a^1	100

Scope $= \{\mathrm{D}, \mathrm{A}\}$

Factor Product

a^1	b^1	0.5
a^1	b^2	0.8
a^2	b^1	0.1
a^2	b^2	0
a^3	b^1	0.3
a^3	b^2	0.9

b^1	c^1	0.5
b^1	c^2	0.7
b^2	c^1	0.1
b^2	c^2	0.2

a^1	b^1	c^1	$0.5 \cdot 0.5 = 0.25$
a^1	b^1	c^2	$0.5 \cdot 0.7 = 0.35$
a^1	b^2	c^1	$0.8 \cdot 0.1 = 0.08$
a^1	b^2	c^2	$0.8 \cdot 0.2 = 0.16$
a^2	b^1	c^1	$0.1 \cdot 0.5 = 0.05$
a^2	b^1	c^2	$0.1 \cdot 0.7 = 0.07$
a^2	b^2	c^1	$0 \cdot 0.1 = 0$
a^2	b^2	c^2	$0 \cdot 0.2 = 0$
a^3	b^1	c^1	$0.3 \cdot 0.5 = 0.15$
a^3	b^1	c^2	$0.3 \cdot 0.7 = 0.21$
a^3	b^2	c^1	$0.9 \cdot 0.1 = 0.09$
a^3	b^2	c^2	$0.9 \cdot 0.2 = 0.18$

Let $\boldsymbol{X}, \boldsymbol{Y}$, and $\boldsymbol{Z}$ be three disjoint sets of variables, and let $\phi_1(\boldsymbol{X}, \boldsymbol{Y})$ and $\phi_2(\boldsymbol{Y}, \boldsymbol{Z})$ be two factors. We define the factor product $\phi_1 \times \phi_2$ to be a factor ψ: $\mathrm{Val}(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) \mapsto \mathbb{R}$ as follows:

$$\psi(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) = \phi_1(\boldsymbol{X}, \boldsymbol{Y}) \cdot \phi_2(\boldsymbol{Y}, \boldsymbol{Z})$$

Factor Marginalization

a^1	b^1	c^1	$0.5 \cdot 0.5 = 0.25$
a^1	b^1	c^2	$0.5 \cdot 0.7 = 0.35$
a^1	b^2	c^1	$0.8 \cdot 0.1 = 0.08$
a^1	b^2	c^2	$0.8 \cdot 0.2 = 0.16$
a^2	b^1	c^1	$0.1 \cdot 0.5 = 0.05$
a^2	b^1	c^2	$0.1 \cdot 0.7 = 0.07$
a^2	b^2	c^1	$0 \cdot 0.1 = 0$
a^2	b^2	c^2	$0 \cdot 0.2 = 0$
a^3	b^1	c^1	$0.3 \cdot 0.5 = 0.15$
a^3	b^1	c^2	$0.3 \cdot 0.7 = 0.21$
a^3	b^2	c^1	$0.9 \cdot 0.1 = 0.09$
a^3	b^2	c^2	$0.9 \cdot 0.2 = 0.18$

$$\phi(a^1, c^1) = \sum_b \phi(a^1, c^1, b)$$

a^1	c^1	033
a^1	c^2	0.51
a^2	c^1	0.05
a^2	c^2	0.07
a^3	c^1	0.24
a^3	c^2	0.39

Let $\boldsymbol{X}$ be a set of variables, and $Y \notin \boldsymbol{X}$ a variable. Let $\phi(\boldsymbol{X}, Y)$ be a factor. We define the factor marginalization of Y in ϕ, denoted $\sum_Y \phi$, to be a factor ψ over $\boldsymbol{X}$ such that:

$$\psi(\boldsymbol{X}) = \sum_Y \phi(\boldsymbol{X}, Y)$$

This operation is also called summing out of Y in ψ.

Factor Reduction

a^1	b^1	c^1	$0.5 \cdot 0.5 = 0.25$
a^1	b^1	c^2	$0.5 \cdot 0.7 = 0.35$
a^1	b^2	c^1	$0.8 \cdot 0.1 = 0.08$
a^1	b^2	c^2	$0.8 \cdot 0.2 = 0.16$
a^2	b^1	c^1	$0.1 \cdot 0.5 = 0.05$
a^2	b^1	c^2	$0.1 \cdot 0.7 = 0.07$
a^2	b^2	c^1	$0 \cdot 0.1 = 0$
a^2	b^2	c^2	$0 \cdot 0.2 = 0$
a^3	b^1	c^1	$0.3 \cdot 0.5 = 0.15$
a^3	b^1	c^2	$0.3 \cdot 0.7 = 0.21$
a^3	b^2	c^1	$0.9 \cdot 0.1 = 0.09$
a^3	b^2	c^2	$0.9 \cdot 0.2 = 0.18$

a^1	b^1	c^1	0.25
a^1	b^2	c^1	0.08
a^2	b^1	c^1	0.05
a^2	b^2	c^1	0
a^3	b^1	c^1	0.15
a^3	b^2	c^1	0.09

Let $\phi(\boldsymbol{Y})$ be a factor, and $\boldsymbol{U} = \boldsymbol{u}$ an assignment for $\boldsymbol{U} \subseteq \boldsymbol{Y}$. We define the reduction of the factor ϕ to the context $\boldsymbol{U} = \boldsymbol{u}$, denoted $\phi[\boldsymbol{U} = \boldsymbol{u}]$ (and abbreviated $\phi[\boldsymbol{u}]$), to be a factor over scope $\boldsymbol{Y}' = \boldsymbol{Y} - \boldsymbol{U}$, such that

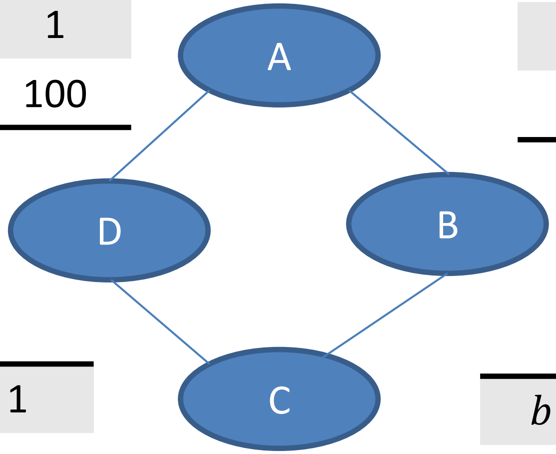
$$\phi[\boldsymbol{u}](\boldsymbol{y}') = \phi(\boldsymbol{y}', \boldsymbol{u})$$

For $\boldsymbol{U} \not\subset \boldsymbol{Y}$, we define $\phi[\boldsymbol{u}]$ to be $\phi[\boldsymbol{U}' = \boldsymbol{u}']$, where $\boldsymbol{U}' = \boldsymbol{U} \cap \boldsymbol{Y}$, and $\boldsymbol{u}' = \boldsymbol{u}\langle \boldsymbol{U}' \rangle$, where $\boldsymbol{u}\langle \boldsymbol{U}' \rangle$ denotes the assignment in $\boldsymbol{u}$ to the variables in $\boldsymbol{U}'$.

Pairwise Markov Networks

d^0	a^0	100
d^0	a^1	1
d^1	a^0	1
d^1	a^1	100

a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	100



c^0	d^0	1
c^0	d^1	100
c^1	d^0	100
c^1	d^1	1

b^0	c^0	100
b^0	c^1	1
b^1	c^0	1
b^1	c^1	100

Pairwise Markov Networks

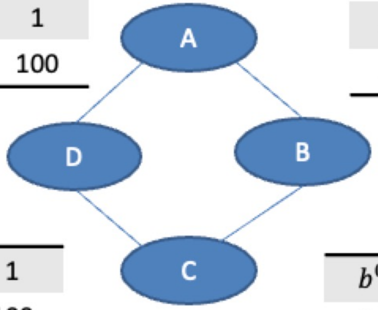
Assignment				Unnomalized
a^0	b^0	c^0	d^0	300000
a^0	b^0	c^0	d^1	300000
a^0	b^0	c^1	d^0	300000
a^0	b^0	c^1	d^1	30
a^0	b^1	c^0	d^0	500
a^0	b^1	c^0	d^1	500
a^0	b^1	c^1	d^0	5000000
a^0	b^1	c^1	d^1	500
a^1	b^0	c^0	d^0	100
a^1	b^0	c^0	d^1	1000000
a^1	b^0	c^1	d^0	100
a^1	b^0	c^1	d^1	100
a^1	b^1	c^0	d^0	10
a^1	b^1	c^0	d^1	100000
a^1	b^1	c^1	d^0	100000
a^1	b^1	c^1	d^1	100000

$$\tilde{P}(A,B,C,D) = \phi(A,B)\phi(B,C)\phi(C,D)\phi(D,A)$$

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

d^0	a^0	100
d^0	a^1	1
d^1	a^0	1
d^1	a^1	100

a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	100



c^0	d^0	1
c^0	d^1	100
c^1	d^0	100
c^1	d^1	1

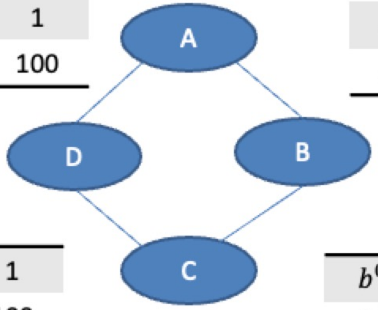
b^0	c^0	100
b^0	c^1	1
b^1	c^0	1
b^1	c^1	100

Eyeballing probabilities is hard

a^0	b^0	0.13
α^0	b^1	0.69
α^1	b^0	0.14
α^1	b^1	0.04

d^0	α^0	100
d^0	a^1	1
d^1	a^0	1
d^1	a^1	100

a^0	b^0	30
α^0	b^1	5
α^1	b^0	1
α^1	b^1	100

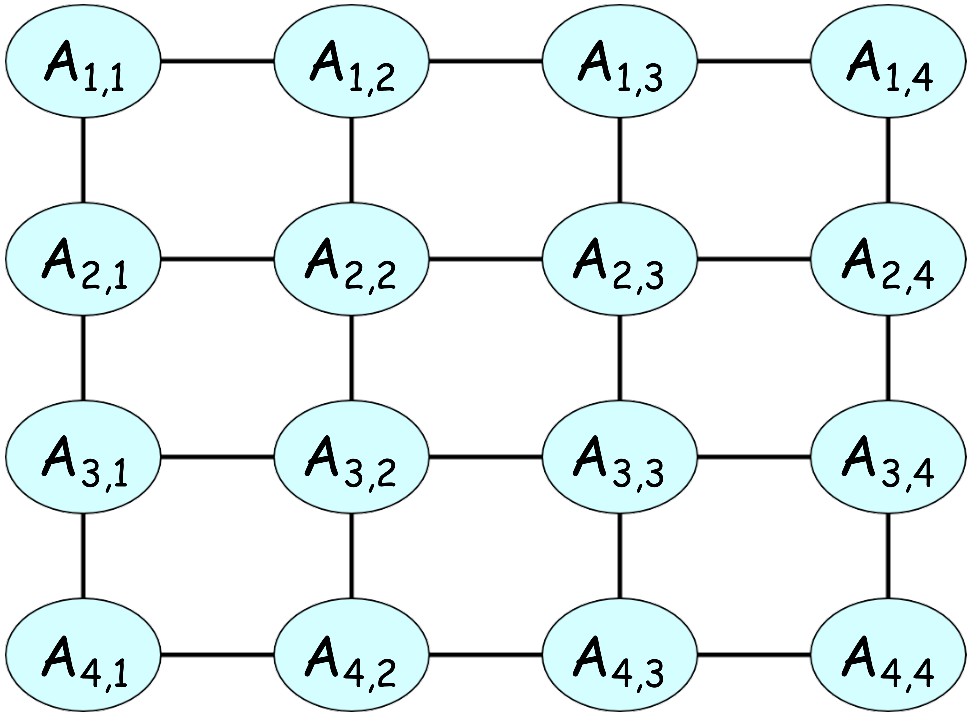


c^0	d^0	1
c^0	d^1	100
c^1	d^0	100
c^1	d^1	1

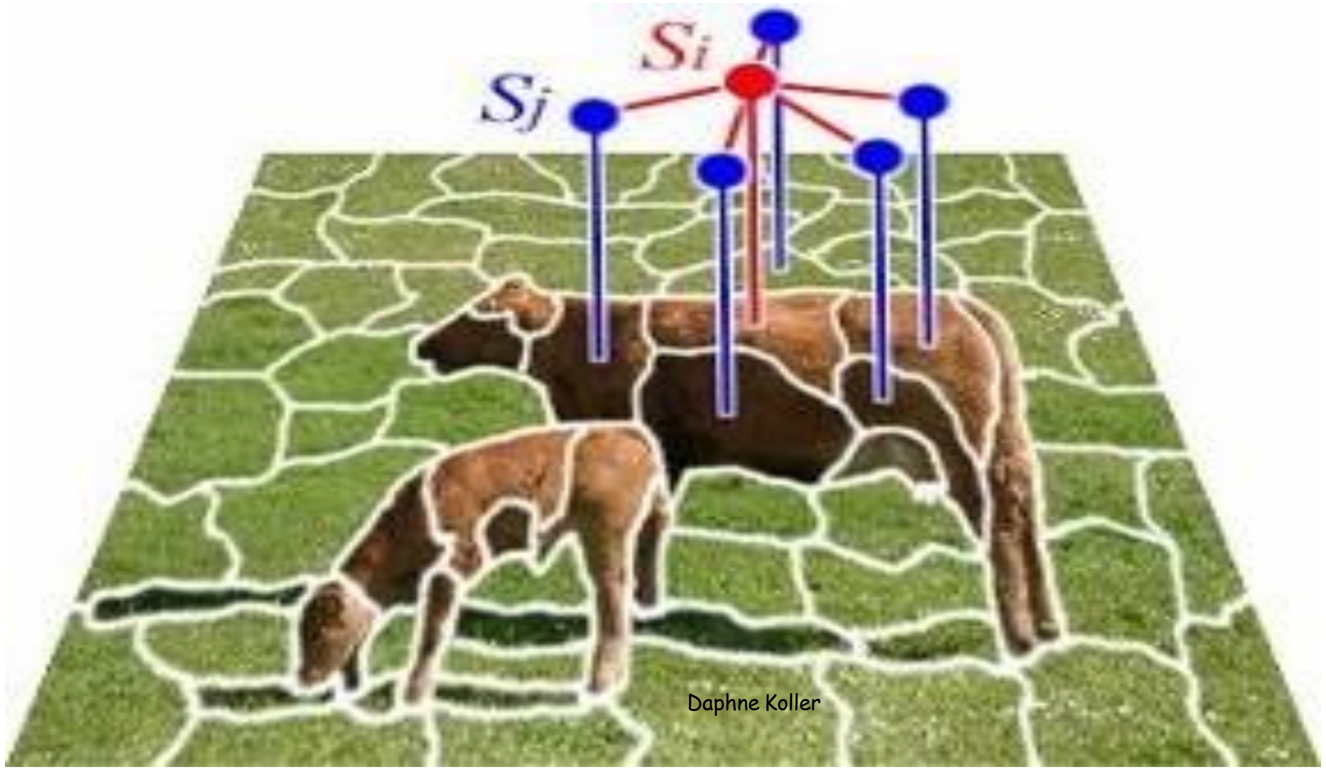
b^0	c^0	100
b^0	c^1	1
b^1	c^0	1
b^1	c^1	100

Pairwise Markov Networks

A pairwise Markov network is an undirected graph whose nodes are $X_1, \ldots, X_n$ and each edge $X_i - X_j$ is associated with a factor (potential) $\phi_{ij}\left(X_i - X_j\right)$

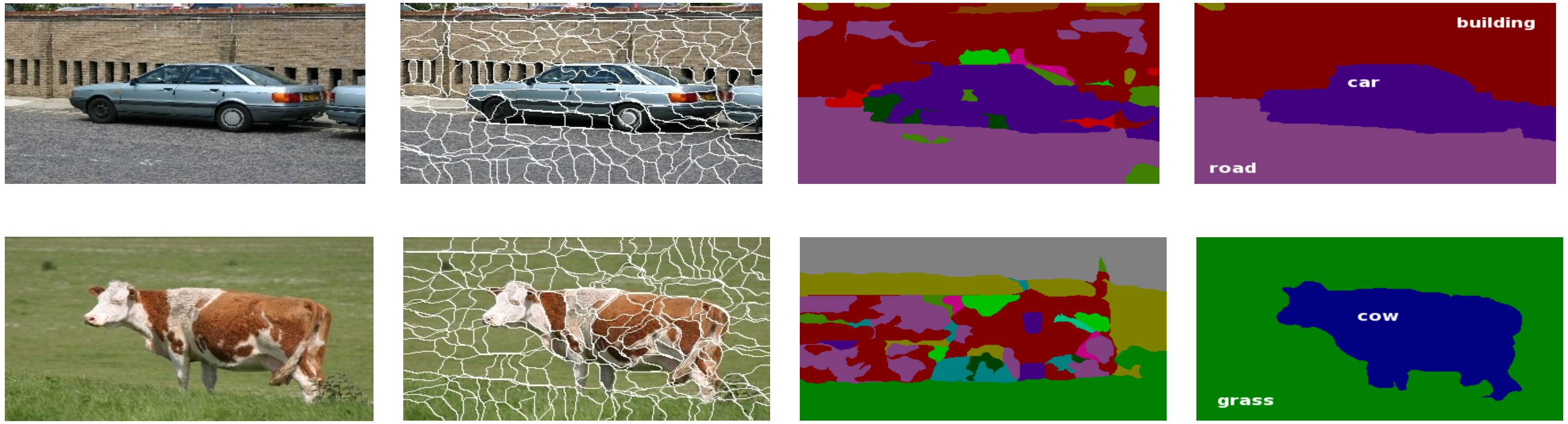


Example: Image Segmentation



Daphne Koller

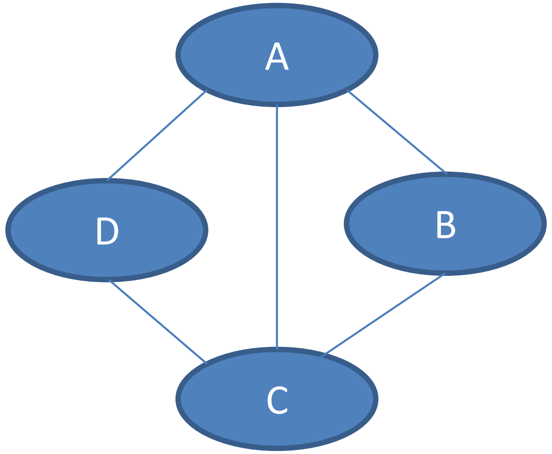
Example: Image Segmentation



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

Daphne Koller

More general Markov Networks



Consider a fully connected pairwise Markov network over $X_1, \ldots, X_n$ where each X_i has d values. How many parameters does the network have?

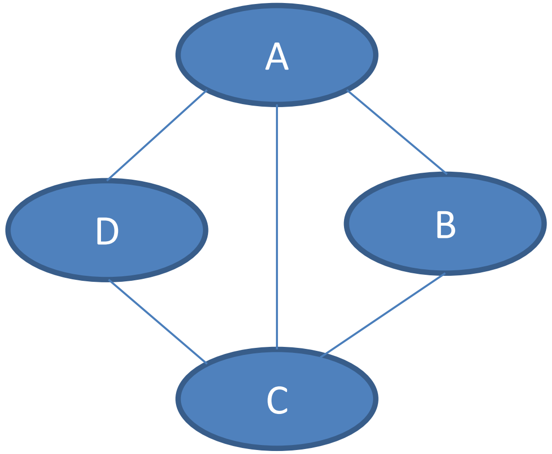
a. $O(d^n)$

b. $O(n^d)$

c. $O(n^2 d^2)$

d. $O(nd)$

More general Markov Networks

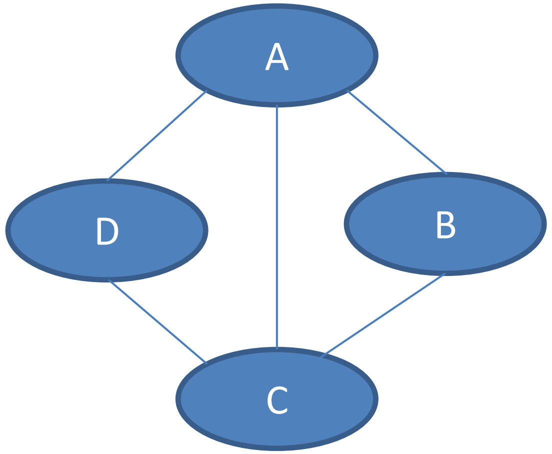


a^1	b^1	c^1	0.25
a^1	b^1	c^2	0.35
a^1	b^2	c^1	0.08
a^1	b^2	c^2	0.16
a^2	b^1	c^1	0.05
a^2	b^1	c^2	0.07
a^2	b^2	c^1	0
a^2	b^2	c^2	0
a^3	b^1	c^1	0.15
a^3	b^1	c^2	0.21
a^3	b^2	c^1	0.09
a^3	b^2	c^2	0.18

$$\Phi = \{\phi_1(\boldsymbol{D}_1), \ldots, \phi_k(\boldsymbol{D}_k)\}$$

$$\tilde{P}_\Phi(X_1, \ldots, X_n) = \prod_i \phi_i(\boldsymbol{D}_\mathrm{i})$$

More general Markov Networks



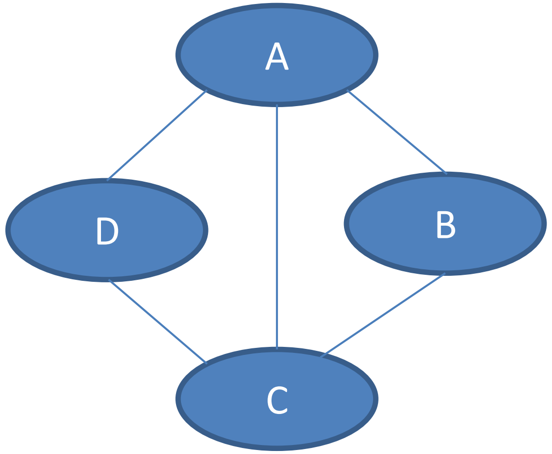
a^1	b^1	c^1	0.25
a^1	b^1	c^2	0.35
a^1	b^2	c^1	0.08
a^1	b^2	c^2	0.16
a^2	b^1	c^1	0.05
a^2	b^1	c^2	0.07
a^2	b^2	c^1	0
a^2	b^2	c^2	0
a^3	b^1	c^1	0.15
a^3	b^1	c^2	0.21
a^3	b^2	c^1	0.09
a^3	b^2	c^2	0.18

$$\Phi = \{\phi_1(\boldsymbol{D}_1), \dots, \phi_k(\boldsymbol{D}_k)\}$$

$$\tilde{P}_\Phi(X_1, \dots, X_n) = \prod_i \phi_i(\boldsymbol{D}_\mathrm{i})$$

$$Z_\Phi = \sum_{X_1 \dots, X_n} \tilde{P}_\Phi(X_1, \dots, X_n)$$

More general Markov Networks



$$\Phi = \{\phi_1(\boldsymbol{D}_1), \dots, \phi_k(\boldsymbol{D}_k)\}$$

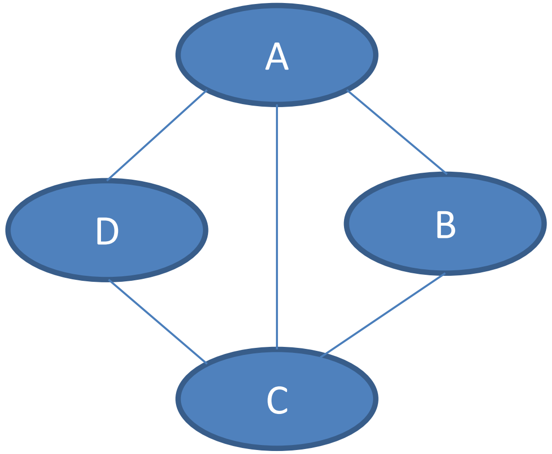
$$\tilde{P}_\Phi(X_1, \dots, X_n) = \prod_i \phi_i(\boldsymbol{D}_\mathrm{i})$$

$$Z_\Phi = \sum_{X_1 \dots, X_n} \tilde{P}_\Phi(X_1, \dots, X_n)$$

$$P_\phi(X_1, \dots, X_n) = \frac{1}{Z_\phi} \prod_i \phi_i(\boldsymbol{D}_\mathrm{i})$$

a^1	b^1	c^1	0.25
a^1	b^1	c^2	0.35
a^1	b^2	c^1	0.08
a^1	b^2	c^2	0.16
a^2	b^1	c^1	0.05
a^2	b^1	c^2	0.07
a^2	b^2	c^1	0
a^2	b^2	c^2	0
a^3	b^1	c^1	0.15
a^3	b^1	c^2	0.21
a^3	b^2	c^1	0.09
a^3	b^2	c^2	0.18

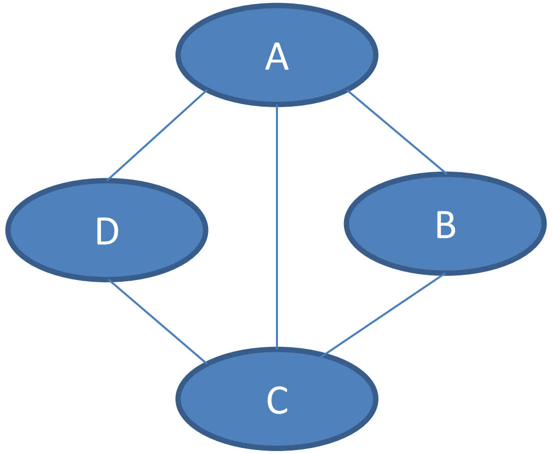
Induced Markov Network



A-B if A and B appear together in some factor.

$$\Phi = \{\phi_{A,B,C}(A,B,C),\ \phi_{A,C,D}(A,C,D)\}$$

Factorization

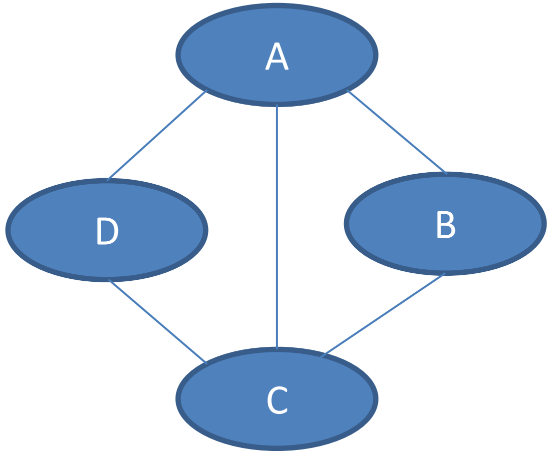


A-B if A and B appear together in some function.

We say G factorizes according to (over) P if there exists a set of factors $\Phi = \{\phi_1(\boldsymbol{D}_1), \ldots, \phi_k(\boldsymbol{D}_k)\}$ such that G is the induced graph for Φ

A graph does not imply a unique factorization

Factorization



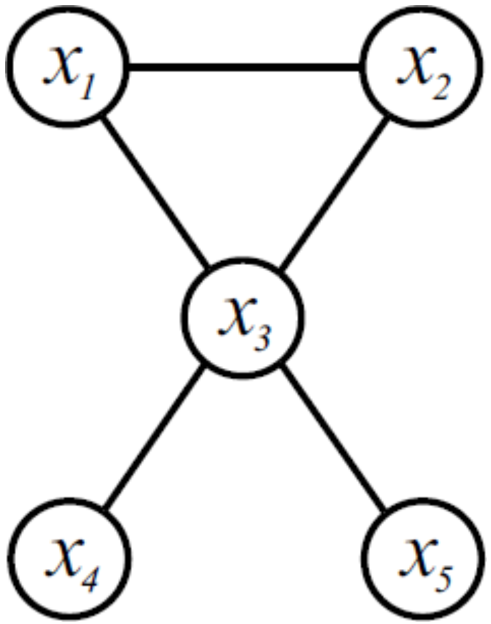
A-B if A and B appear together in some function.

*A **minimal factorization** is one where all factors are **maximal cliques** (not a strict subset of any other clique) in the MRF*

We say G factorizes according to (over) P if there exists a set of factors $\Phi = \{\phi_1(\boldsymbol{D}_1), \ldots, \phi_k(\boldsymbol{D}_k)\}$ such that G is the induced graph for Φ

A graph does not imply a unique factorization

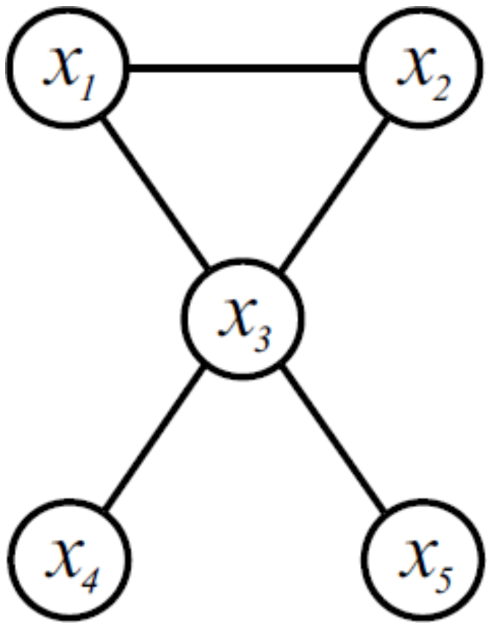
Example



Find the minimal factorization

Find a valid factorization that is not minimal

Example

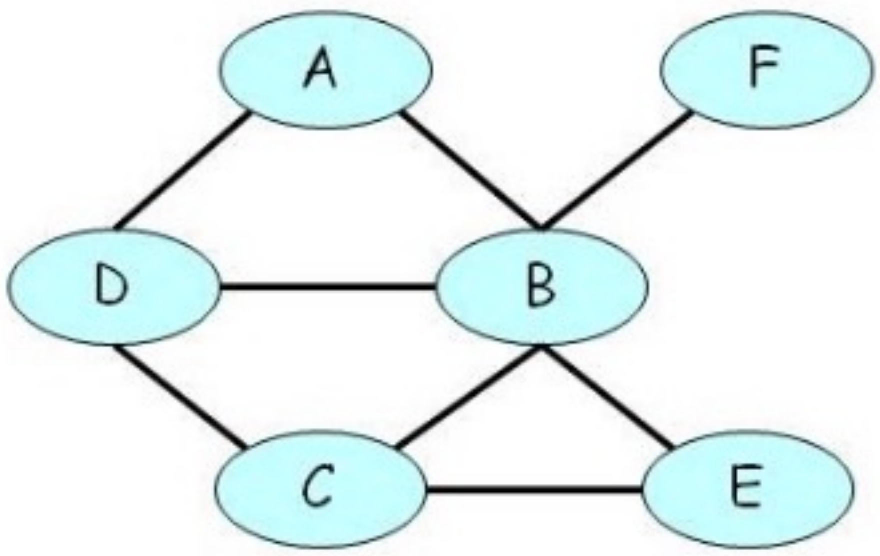


Find the minimal factorization

Find a valid factorization that is not minimal

$$\psi(x_1, x_2, x_3)\psi(x_3, x_4)\psi(x_3, x_5)$$
$$\psi(x_1, x_2)\psi(x_2, x_3)\psi(x_1, x_3)\psi(x_3, x_4)\psi(x_3, x_5)$$

Separation in Markov Networks



Definition:

X and **Y** are *separated* in H given **Z** if there is no active trail in H between **X** and **Y** given **Z**

Active trail: Undirected path

Conditioning on a node on the path blocks the path

Factorization and Independence

Factorization ⇒ Independence

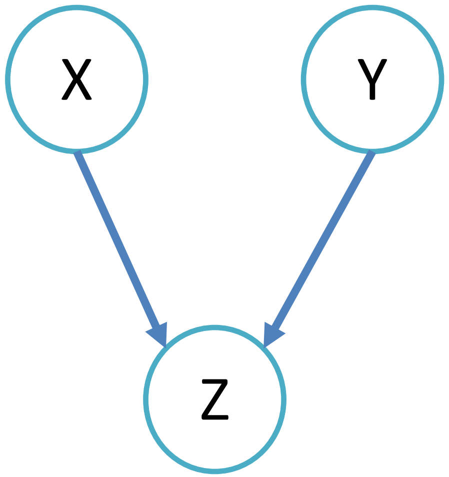
Theorem: If P factorizes over (according to) H, and sep_H(**X**, **Y** | **Z**) then Ind(**X, Y|Z**) in P

If P factorizes over H, then H is an I-map of P

Independence⇒ Factorization

Theorem (Hammersley Clifford): If H is an I-Map for P, and P is a positive distribution, then P factorizes over H

Markov Networks and DAGs

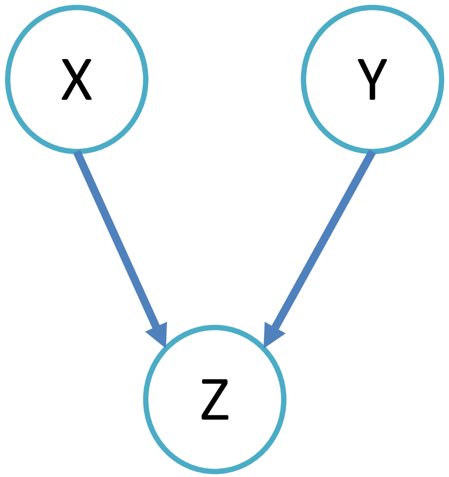


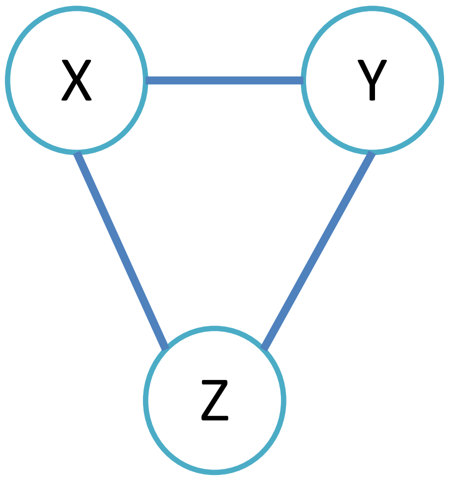
G

If G is a perfect map for P

Find an MN that is a perfect map for P

Markov Networks and DAGs



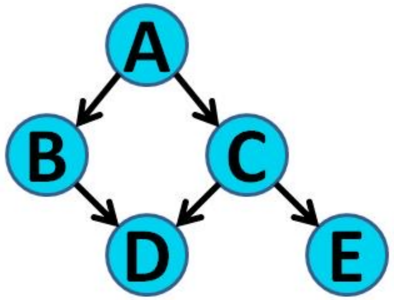


If G is a perfect map for P

Find an MN that is a perfect map for P

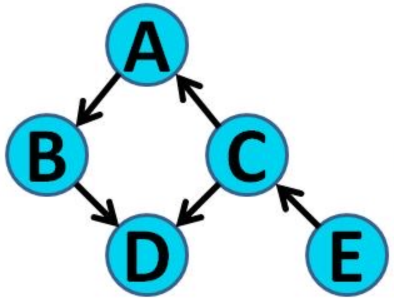
Going from BN to MN you lose some independencies

I-equivalence

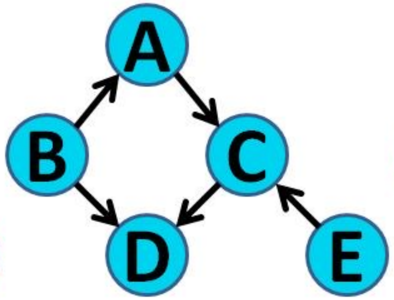


G

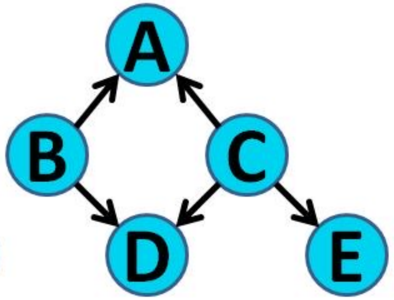
Which networks are Markov Equivalent to G?



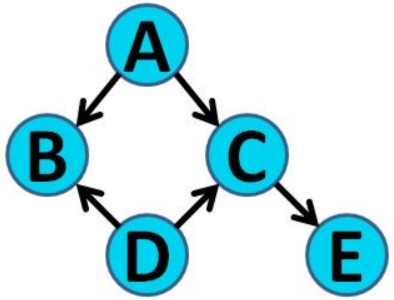
I



II

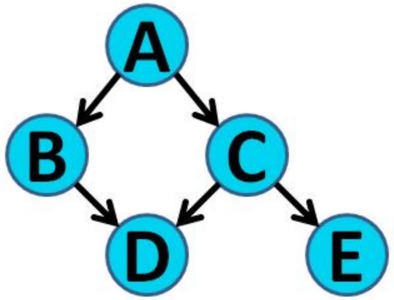


III



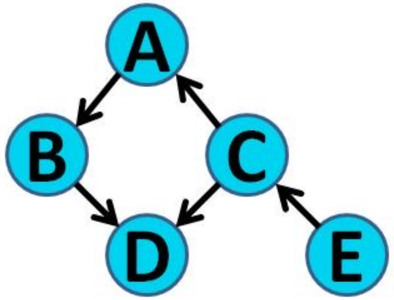
IV

Practice

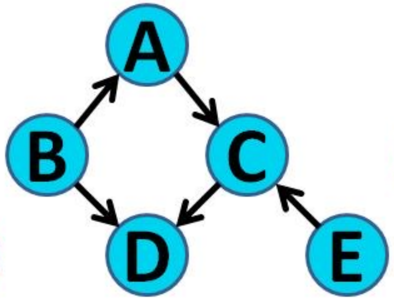


G

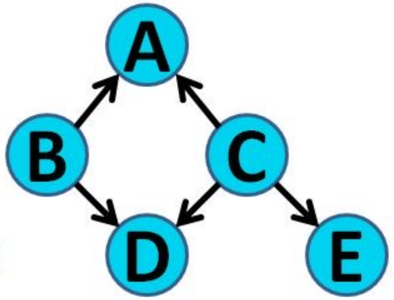
Find a MN that is an I-Map of the probability induced by G



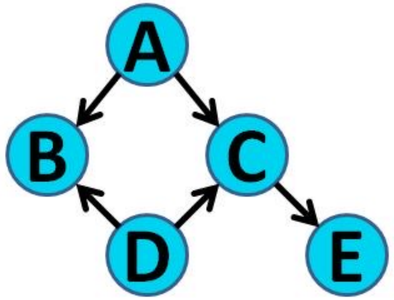
I



II

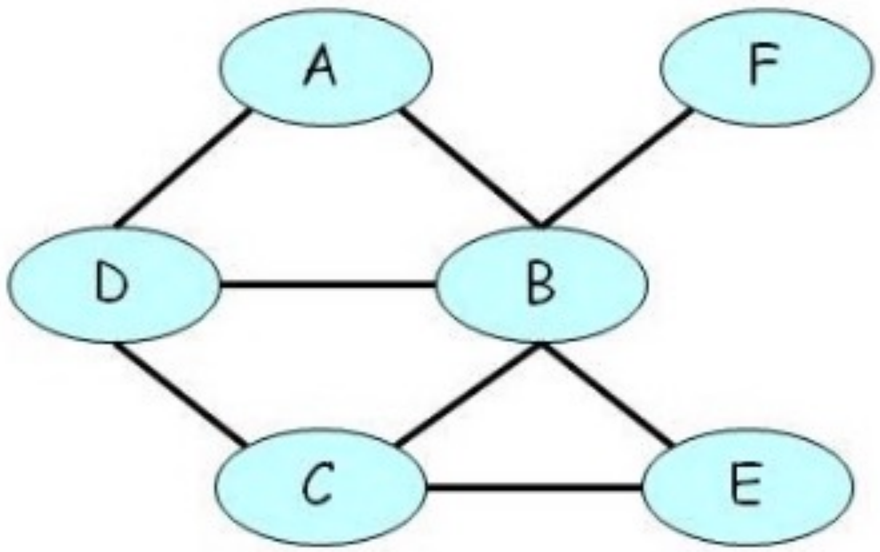


III



IV

Practice: Separations



Log-linear Representation

$$\tilde{P} = \prod_i \phi_i(\boldsymbol{D_i})$$

Original parameterization

$$\tilde{P} = \exp\left(-\sum_j w_j f_j(\boldsymbol{D_j})\right)$$

Log-linear parameterization

Features are functions (like factors) without the non-negativity assumption.
Each feature has a single weight.
Different features can have the same scope.

Log-linear Representation

$$\phi(X_1, X_2) = \begin{pmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{pmatrix}$$

$$f_{12}^{ij} = I(X_1 = i \; and \; X_2 = j)$$

One feature for each i,j value

$$\phi(X_2, X_3) = \exp\left(-\sum_{kl} w_{kl} f_{ij}^{kl}(X_1, X_2)\right)$$

$$w_{kl} = -\log(a_{kl})$$

Example: Ising Models

$$E(x_1, \dots, x_n) = -\sum_{i<j} w_{i,j} x_i x_j \; - \sum_i u_i x_i$$

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

$$f_{i,j}(X_i, X_j) = X_i \cdot X_j$$

$$P(\boldsymbol{X}) \propto e^{-\frac{1}{T} E(\boldsymbol{X})}$$

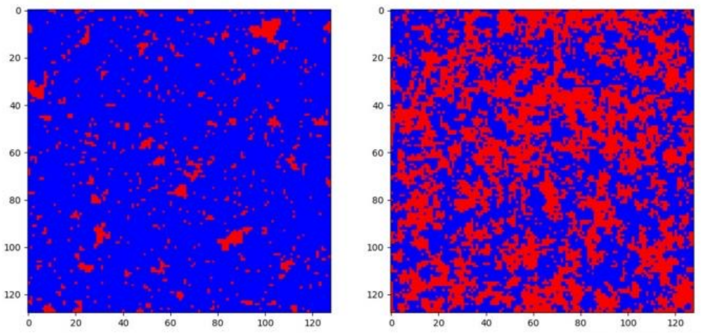
Example: Ising Models

$$E(x_1, \dots, x_n) = -\sum_{i<j} w_{i,j} x_i x_j - \sum_i u_i x_i$$

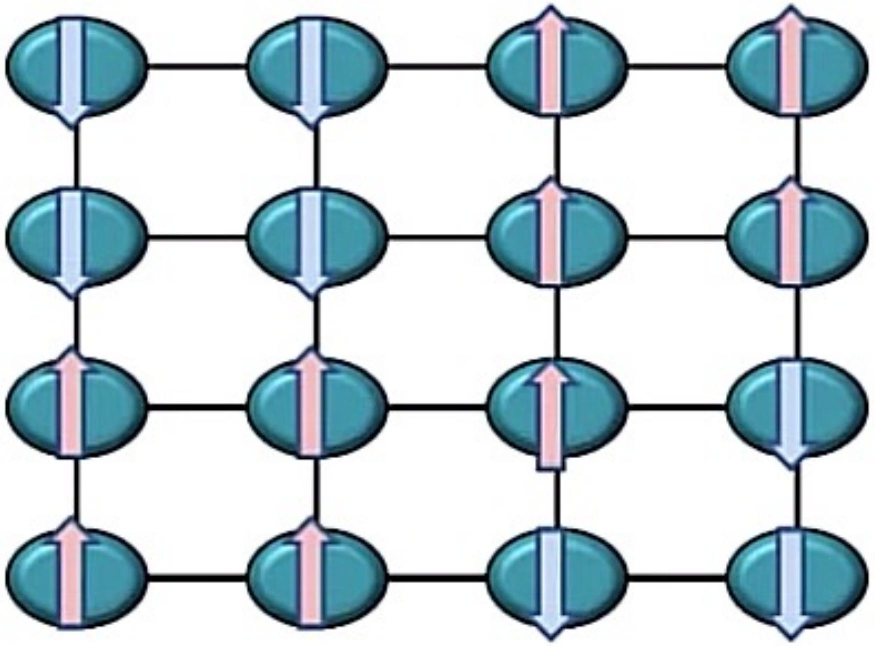
$$x_i \in \{-1, 1\}$$

$$f_{i,j}(X_i, X_j) = X_i \cdot X_j$$

$$P(\boldsymbol{X}) \propto e^{-\frac{1}{T} E(\boldsymbol{X})}$$



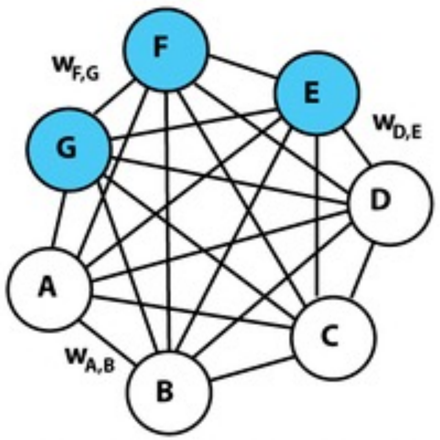
As T grows, w_{ij}'s become smaller



Example: Boltzman machine

$$E = -\sum_{i<j} w_{ij} s_i s_j + -\sum_i \theta_i s_i$$

$$s_i \in \{0, 1\}$$



- w_{ij} is the connection strength between unit j and unit i.
- s_i is the state, $s_i \in \{0,1\}$, of unit i.
- θ_i is the bias of unit i in the global energy function. ($-\theta_i$ is the activation threshold for the unit.)

Model for neural activation