Readings:
K&F: 5.1, 5.2, 5.3, 5.4, 5.5, 5.6, 5.7

Markov networks, Factor graphs, and an unified view

Start approximate inference

If we are lucky…

Graphical Models – 10708

Carlos Guestrin

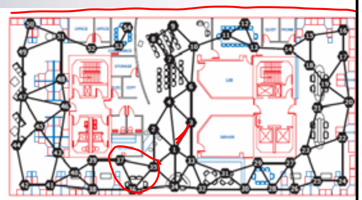
Carnegie Mellon University

October 27th, 2006

Factorization in Markov networks

- Given an undirected graph H over variables $\mathbf{X}=\{X_1,\ldots,X_n\}$
- A distribution P **factorizes** over H if $\exists$
  - subsets of variables $\mathbf{D}_1 \subseteq \mathbf{X},\ldots,\mathbf{D_m} \subseteq \mathbf{X}$, such that the $\mathbf{D}_i$ are *fully connected* in H
  - *non-negative potentials* (or factors) $\pi_1(\mathbf{D_1}),\ldots,\pi_m(\mathbf{D_m})$
    - also known as clique potentials
  - such that

$$P(X) = \frac{1}{Z} \prod_{i=1}^{m} \pi_i(D_i)$$



35, 36, 37

allowed

Stronger $\boxed{\pi(CDE)}$ K^3 values

or

Weaker $\pi(CD)$ $\pi(DE)$ $\pi(CE)$

$3K^2$ values

A B C D E

not allowed

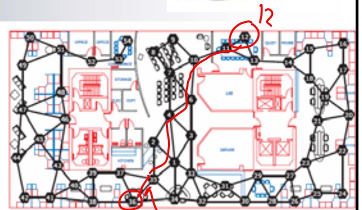
no factor AB

no factor BCD

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

Global Markov assumption in Markov networks

- 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**)



12

36

active if I don't observe 36, 1, 4...

- Variables **X** are **separated** from **Y** given **Z** in graph *H*, $sep_H(\mathbf{X};\mathbf{Y}|\mathbf{Z})$, if there is no active path between any $X \in \mathbf{X}$ and any $Y \in \mathbf{Y}$ given **Z**

- The **global Markov assumption** for a Markov network *H* is $I(H) = \{ X, Y, Z \mid sep_H(X;Y|Z) \}$

Representation Theorem for Markov Networks

If joint probability distribution *P*:

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

Then ***H*** **is an I-map for** ***P***

OK → if doing param. learning and get a zero in a potential

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

Then **joint probability distribution *P*:**

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

can get in trouble → if structural learning & P not positive

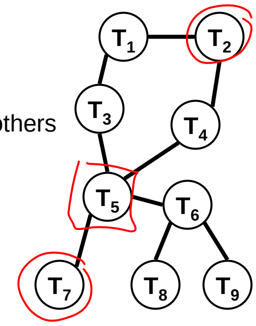
Local independence assumptions for a Markov network

- **Separation** defines global independencies $I(H)$
- **Pairwise Markov Independence**: $I_{pw}(H)$
  - Pairs of non-adjacent variables A,B are independent given all others

$A \perp B \mid \mathcal{X} - \{A,B\}$

- **Markov Blanket**: $I_{MB}(H)$
  - Variable A independent of rest given its neighbors $N(A)$

$A \perp \mathcal{X} - \{N(A), A\} \mid N(A)$



$T_7 \perp T_2 \mid T_1, T_{3-6}, T_8, T_9$

$T_7 \perp \{T_1, T_{2-6}, T_{8-9}\} \mid T_5$

Equivalence of independencies in Markov networks

- **Soundness Theorem**: For all positive distributions P, the following three statements are equivalent: $\forall x, P(x) > 0$
  - P entails the global Markov assumptions
    $P \models I(H)$
  - P entails the pairwise Markov assumptions
    $P \models I_{pw}(H)$
  - P entails the local Markov assumptions (Markov blanket)
    $P \models I_{MB}(H)$

Minimal I-maps and Markov Networks

- A fully connected graph is an I-map
- Remember minimal I-maps?
  - A "simplest" I-map $\rightarrow$ Deleting an edge makes it no longer an I-map
- In a BN, there is no unique minimal I-map
- Theorem: **In a Markov network, minimal I-map is unique!!** with positive dist.
- Many ways to find minimal I-map, e.g.,
  - Take pairwise Markov assumption: X_i & X_j not adjacent
  - If *P* doesn't entail it, add edge: $X_i \perp X_j \mid \mathcal{X} - \{X_i, X_j\}$

for X_i, X_j ask is $X_i \perp X_j \mid \mathcal{X} - \{X_i, X_j\}$
yes $\Rightarrow$ no edge
no $\Rightarrow$ edge

Problem with this alg:
$P(X_i, X_j \mid \mathcal{X} - \{X_i, X_j\}) = P(X_i \mid \mathcal{X} - \{X_i, X_j\}) \cdot P(X_j \mid \mathcal{X} - \{X_i, X_j\})$

How about a perfect map?

- Remember perfect maps?
  - independencies in the graph are exactly the same as those in *P*
- For BNs, doesn't always exist
  - counter example: Swinging Couples
- How about for Markov networks? NO!

W_1, M_1, M_2, W_2

no BN with exactly these indep.

V-structure!

F, A, S

$F \perp A$
$\neg F \perp A \mid S$

minimal I-map MN:

F — A, S

$\neg F \perp A$

Unifying properties of BNs and MNs

- BNs:
  - give you: V-structures, CPTs are conditional probabilities, can directly compute probability of full instantiation
  - but: require acyclicity, and thus no perfect map for swinging couples
- MNs:
  - give you: cycles, and perfect maps for swinging couples
  - but: don't have V-structures, cannot interpret potentials as probabilities, requires partition function
- Remember PDAGS??? (Chain Graphs)
  - skeleton + immoralities
  - provides a (somewhat) unified representation
  - see book for details

no perfect map for XOR
$X \perp Y$, $Y \perp Z$
$Z \perp X$
but $\neg X \perp Y \mid Z$

What you need to know so far about Markov networks

- Markov network representation:
  - undirected graph
  - potentials over cliques (or sub-cliques)
  - normalize to obtain probabilities
  - need partition function
- Representation Theorem for Markov networks
  - if P factorizes, then it's an I-map
  - if P is an I-map, only factorizes for positive distributions
- Independence in Markov nets:
  - active paths and separation
  - pairwise Markov and Markov blanket assumptions
  - equivalence for positive distributions
- Minimal I-maps in MNs are unique
- Perfect maps don't always exist

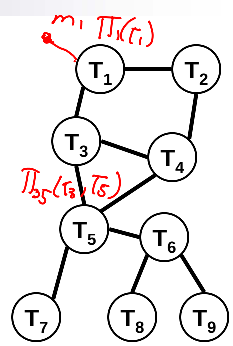
Some common Markov networks and generalizations

- Pairwise Markov networks
- A very simple application in computer vision
- Logarithmic representation
- Log-linear models
- Factor graphs

Pairwise Markov Networks

- All factors are over single variables or pairs of variables:
  - Node potentials $\pi_i(T_i)$; $\pi_i(T_i, m_i)$ ↖ observed
  - Edge potentials $\pi_{ij}(T_i, T_j)$
- Factorization:

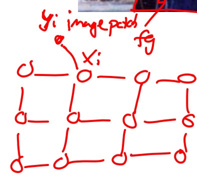
$$P(X) = \frac{1}{Z} \underbrace{\prod_i \pi_i(x_i)}_{\text{node potentials}} \prod_{\underset{\uparrow\ \text{edges}}{ij \in E}} \pi_{ij}(x_i, x_j)$$



- Note that there may be bigger cliques in the graph, but only consider pairwise potentials

A very simple vision application

- Image segmentation: separate foreground from background $X_i \in \{fg, bg\}$
- Graph structure:
  - pairwise Markov net
  - grid with one node per pixel



- Node potential:
  - "background color" v. "foreground color"

light blue brown

$\pi_i(X_i, Y_i) =$

$y_i \backslash x_i$	fg	bg
light b.	1	10
brown	20	5

- Edge potential:
  - neighbors like to be of the same class

$\pi_{ij}(X_i, X_j) =$

$x_j \backslash x_i$	fg	bg
fg	λ	1
bg	1	λ

λ low $\rightarrow$ no smoothness (focus on node pot.)

λ high $\rightarrow$ v. smooth (ignore or less focus on node pot.)

Logarithmic representation

$\log \prod_{i=1}^{m} = \sum_{i=1} \log$

- Standard model: $$P(X_1, \ldots, X_n) = \frac{1}{Z} \prod_{i=1}^{m} \pi_i(\mathbf{D}_i)$$
- Log representation of potential (assuming positive potential):
  - also called the energy function $\psi_i(D_i) = -\ln \pi_i(D_i)$

$$P(X) = \frac{1}{Z} \prod_{i=1}^{m} \pi_i(D_i) = \frac{1}{Z} \exp\left\{\log \prod_{i=1}^{m} \pi_i(D_i)\right\} = \frac{1}{Z} \exp\left\{\sum_{i=1}^{m} \log \pi_i(D_i)\right\}$$

- Log representation of Markov net: $= \frac{1}{Z} \exp\left\{-\underbrace{\sum_{i=1}^{m} \psi_i(D_i)}_{\text{energy functions}}\right\}$

Log-linear Markov network (most common representation)

- **Feature** is some function $\phi[\mathbf{D}]$ for some subset of variables $\mathbf{D}$
  - e.g., indicator function e.g. $\phi_i(D) = \begin{cases} 1 & \text{if all of } D \text{ is true} \\ 0 & \text{otherwise} \end{cases}$
- **Log-linear model** over a Markov network *H*:
  - a set of features $\phi_1[\mathbf{D}_1],\ldots,\phi_k[\mathbf{D}_k]$
    - each $\mathbf{D}_i$ is a subset of a clique in *H*
    - two ϕ's can be over the same variables
  - a set of weights $w_1,\ldots,w_k$
    - usually learned from data
  - $P(X_1,\ldots,X_n) = \frac{1}{Z}\exp\left[\sum_{i=1}^{k} w_i \phi_i(\mathbf{D}_i)\right]$

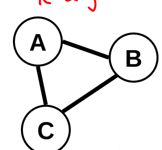
linear models: $\sum_{i=1}^{k} w_i \phi_i(D_i)$

sometimes defn as $\exp\left[-\sum_{i=1}^{K} w_i \phi_i(D_i)\right]$

Structure in cliques

- Possible potentials for this graph:

K-ary vars



how many params

$K^3-1 \longrightarrow \pi(A,B,C)$

$3(K^2-1) \longrightarrow \begin{cases} \pi(A,B) \\ \pi(B,C) \\ \pi(A,C) \end{cases}$

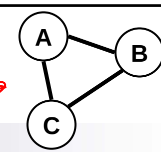
$\pi(A)$

$\pi(B)$

$\pi(C)$

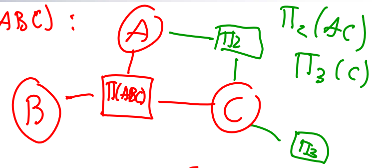
Factor graphs

[exactly equivalent to MN]

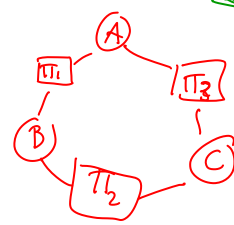


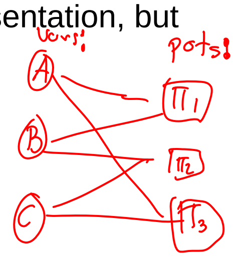
- Very useful for approximate inference
  - Make factor dependency explicit
- Bipartite graph:
  - variable nodes (ovals) for $X_1,\ldots,X_n$
  - factor nodes (squares) for $\phi_1,\ldots,\phi_m$
  - edge $X_i - \phi_j$ if $X_i \in \mathrm{Scope}[\phi_j]$
- More explicit representation, but exactly equivalent

$\pi(ABC)$:



$\pi_1(AB)$
$\pi_2(BC)$
$\pi_3(AC)$





Exact inference in MNs and Factor Graphs

- Variable elimination algorithm presented in terms of factors $\rightarrow$ exactly the same VE algorithm can be applied to MNs & Factor Graphs
- Junction tree algorithms also applied directly here:
  - triangulate MN graph as we did with moralized graph for the BN
  - each factor belongs to a clique
  - same message passing algorithms

Summary of types of Markov nets

- Pairwise Markov networks
  - very common
  - potentials over nodes and edges
- Log-linear models
  - log representation of potentials
  - linear coefficients learned from data
  - most common for learning MNs
- Factor graphs
  - explicit representation of factors
    - you know exactly what factors you have
  - very useful for approximate inference

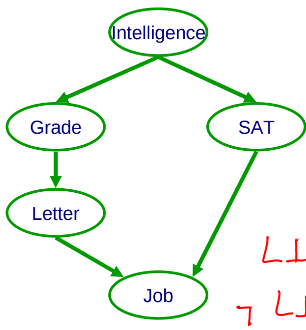
What you learned about so far

- Bayes nets
- Junction trees
- (General) Markov networks
- Pairwise Markov networks
- Factor graphs

- How do we transform between them?
- More formally:
  - I give you an graph in one representation, find an **I-map** in the other

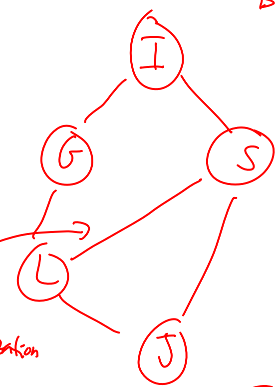
From Bayes nets to Markov nets

BN



MN

start with BN skeleton



$L \perp S \mid I$

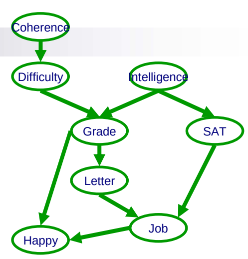
$\neg\, L \perp S \mid I, J$

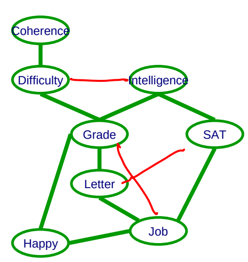
a moralization arc

minimal I-map but doesn't show that $L \perp S \mid I$

BNs → MNs: Moralization

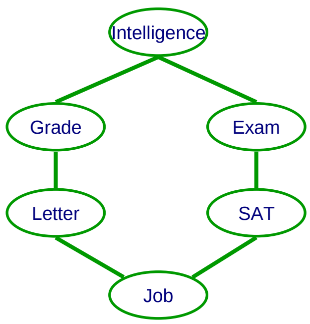
- **Theorem**: Given a BN *G* the Markov net *H* formed by moralizing *G* is the *minimal I-map* for I(*G)*
- **Intuition**:
  - in a Markov net, each factor must correspond to a subset of a clique
  - the factors in BNs are the CPTs $P(J \mid L, S)$
  - CPTs are factors over a node and its parents
  - thus node and its parents must form a clique
- **Effect**:
  - **some** independencies that could be read from the BN graph become hidden



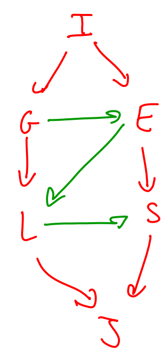


From Markov nets to Bayes nets

MN



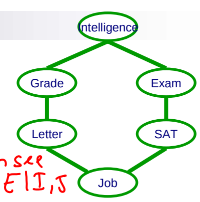
BN



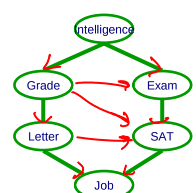
triangulation
or Chordal graph

MNs → BNs: Triangulation

- **Theorem**: Given a MN *H,* let *G* be the Bayes net that is a *minimal I-map* for I(*H)* then *G* must be **chordal**
- **Intuition**:
  - v-structures in BN introduce immoralities
  - these immoralities were not present in a Markov net
  - the triangulation eliminates immoralities
- **Effect**:
  - **many** independencies that could be read from the MN graph become hidden

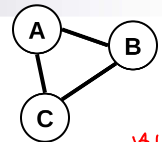


I can see
$L \perp E \mid I, J$



Can't see it
any more !!

Markov nets v. Pairwise MNs



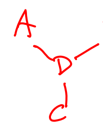
- Every Markov network can be transformed into a Pairwise Markov net
  - introduce extra "variable" for each factor over three or more variables
  - domain size of extra variable is exponential in number of vars in factor
- **Effect**:
  - any local structure in factor is lost
  - a chordal MN doesn't look chordal anymore

$\pi_1(A,B,C) =$

B \ A	t	f
tt		
tf		
ft		
ff		

Extra var D that has 8 values
↳ each value assignment of ABC

node potential $\pi(D) = \pi_1(A(D), B(D), C(D))$



edge potential

$\pi(A,D)$

D \ A	t	f
ttt	1	0
ftt	0	1

Overview of types of graphical models and transformations between them