

PROBABILISTIC GRAPHICAL MODELS  
CPSC 532C (TOPICS IN AI)  
STAT 521A (TOPICS IN MULTIVARIATE ANALYSIS)

LECTURE 5

Kevin Murphy

Monday 27 September, 2004

## ADMINISTRIVIA

---

- Homework 2 is now due on Wednesday 29th.
- Please start reading chapters 6 and 7 before Wednesday.

## TODAY'S CLASS

---

- Review of homework 1.
- Matlab vectorization.
- Review of past 4 lectures.

## TODAY'S CLASS

---

- Review of homework 1.
- Matlab vectorization.
- Review of past 4 lectures.

## CHAP 3: DAGs

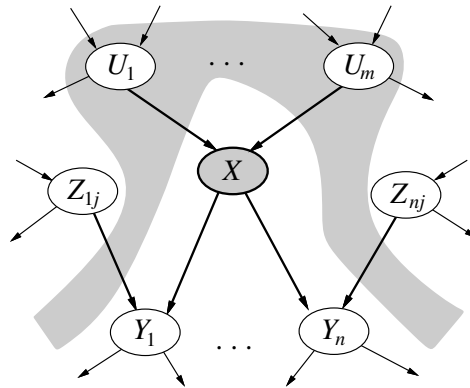
---

- Local Markov property, chain rule for Bayes nets
- Global Markov property (d-separation/ Bayes-ball)
- Minimal I-maps
- Perfect maps

## LOCAL MARKOV PROPERTY

---

- Node is conditionally independent of its non-descendants given its parents.

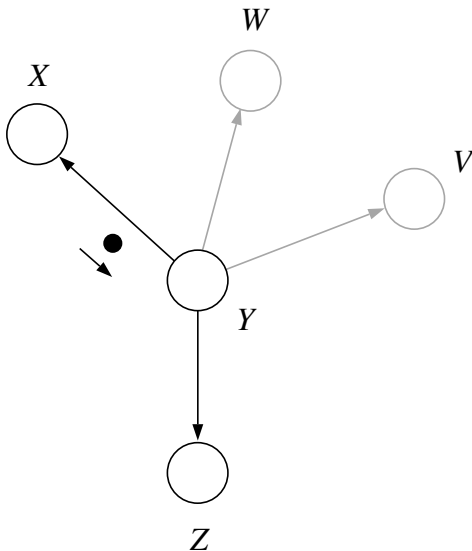


$$\begin{aligned} P(X_{1:N}) &= P(X_1)P(X_2|X_1)P(X_3|X_1, X_2) \dots \\ &= \prod_{i=1}^N P(X_i|X_{1:i-1}) \\ &= \prod_{i=1}^N P(X_i|X_{\pi_i}) \end{aligned}$$

## BAYES BALL ALGORITHM

---

- To check if  $\mathbf{x}_A \perp \mathbf{x}_B | \mathbf{x}_C$  we need to check if every variable in  $A$  is d-separated from every variable in  $B$  conditioned on all vars in  $C$ .
- In other words, given that all the nodes in  $\mathbf{x}_C$  are clamped, when we wiggle nodes  $\mathbf{x}_A$  can we change any of the node  $\mathbf{x}_B$ ?
- The *Bayes-Ball Algorithm* is a such a d-separation test.  
We shade all nodes  $\mathbf{x}_C$ , place balls at each node in  $\mathbf{x}_A$  (or  $\mathbf{x}_B$ ), let them bounce around according to some rules, and then ask if any of the balls reach any of the nodes in  $\mathbf{x}_B$  (or  $\mathbf{x}_A$ ).

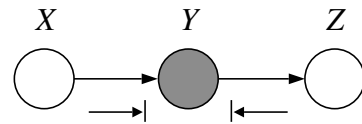


So we need to know what happens when a ball arrives at a node  $\mathbf{Y}$  on its way from  $\mathbf{X}$  to  $\mathbf{Z}$ .

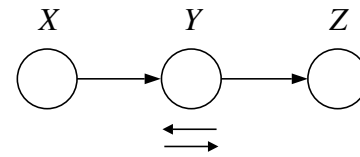
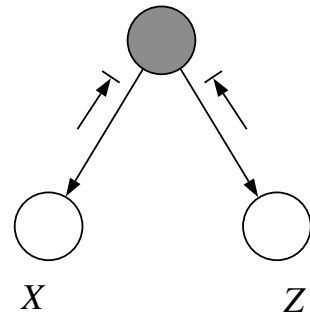
# BAYES-BALL RULES

---

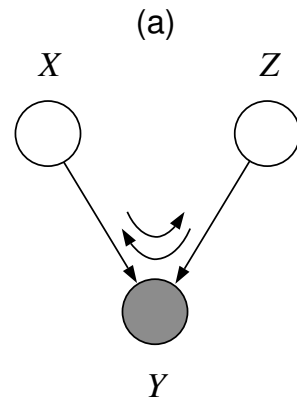
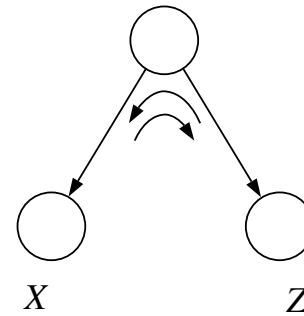
- The three cases we considered tell us rules:



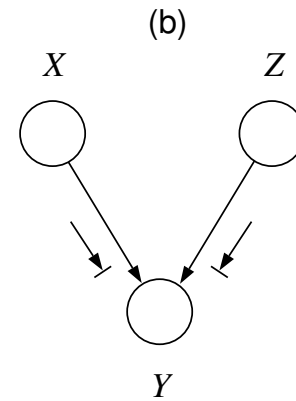
(a)  
 $Y$



(b)  
 $Y$



(a)

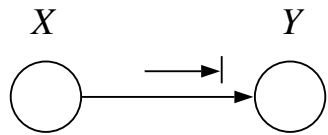


(b)

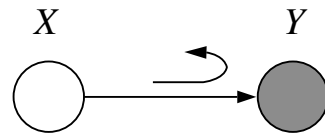
# BAYES-BALL BOUNDARY RULES

---

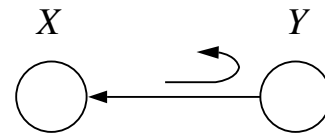
- We also need the boundary conditions:



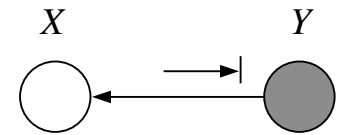
(a)



(b)



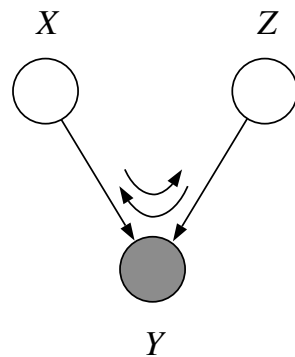
(a)



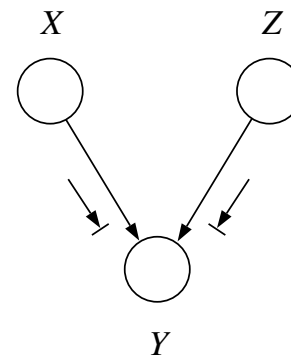
(b)

- Here's a trick for the explaining away case:

If *y* or any of its descendants is shaded, the ball passes through.



(a)



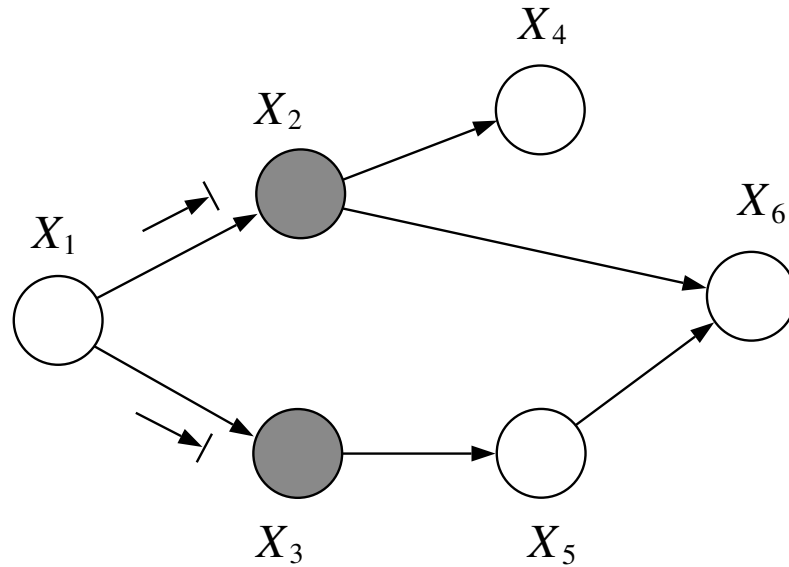
(b)

- Notice balls can travel opposite to edge directions.

# EXAMPLES OF BAYES-BALL ALGORITHM

---

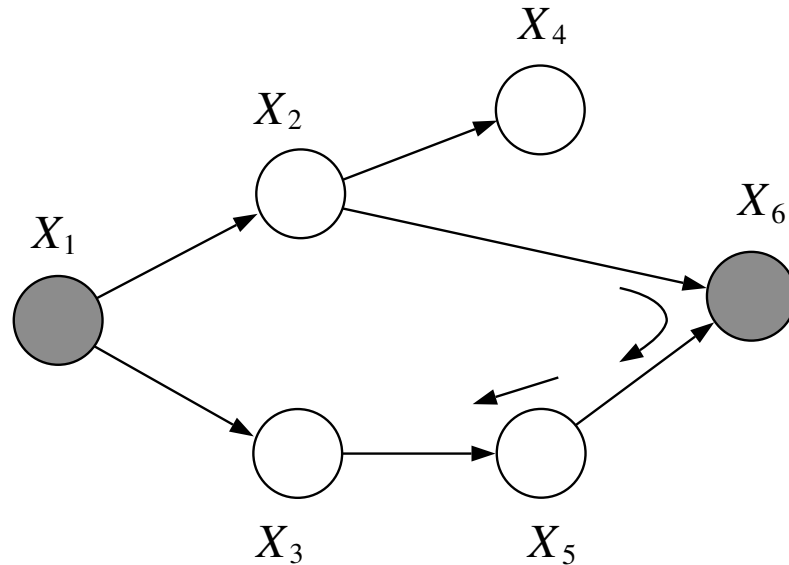
$$\mathbf{x}_1 \perp \mathbf{x}_6 \mid \{\mathbf{x}_2, \mathbf{x}_3\} \quad ?$$



# EXAMPLES OF BAYES-BALL ALGORITHM

---

$$\mathbf{x}_2 \perp \mathbf{x}_3 | \{\mathbf{x}_1, \mathbf{x}_6\} \quad ?$$



Notice: balls can travel opposite to edge directions.

## I-MAPS

---

- Defn: let  $I_l(G)$  be the set of local independence properties encoded by DAG  $G$ , namely:

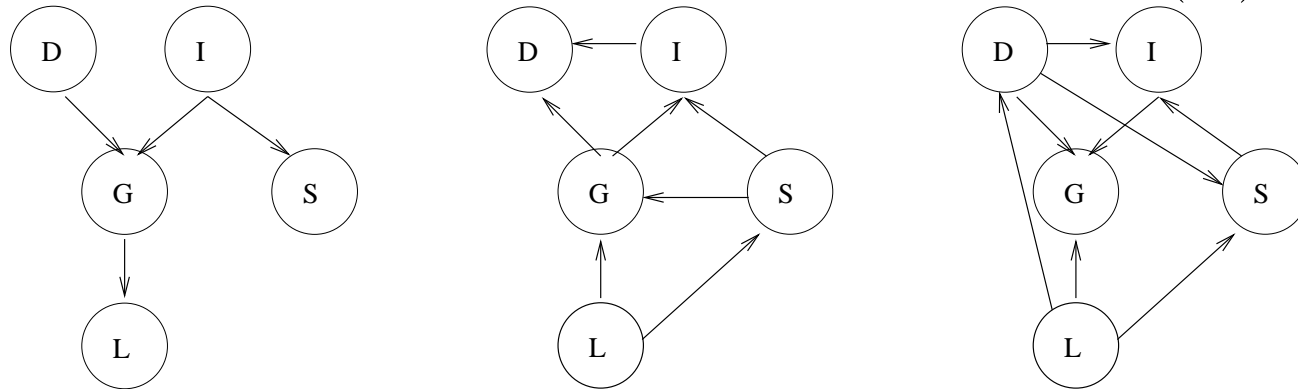
$$\{X_i \perp \text{NonDescendants}(X_i) \mid \text{Parents}(X_i)\}$$

- Defn: A DAG  $G$  is an **I-map** (independence-map) of  $P$  if  $I_l(G) \subseteq I(P)$ .
- A fully connected DAG  $G$  is an I-map for any distribution, since  $I_l(G) = \emptyset \subseteq I(P)$  for any  $P$ .
- Defn: A DAG  $G$  is a **minimal I-map** for  $P$  if it is an I-map for  $P$ , and if the removal of even a single edge from  $G$  renders it not an I-map.
- **To construct a minimal I-map**, Pick a node ordering, then let the parents of node  $X_i$  be the minimal subset  $U \subseteq \{X_1, \dots, X_{i-1}\}$   
s.t.  $X_i \perp \{X_1, \dots, X_{i-1}\} \setminus U \mid U$ .

## A DISTRIBUTION MAY HAVE SEVERAL MINIMAL I-MAPS

---

- Suppose the left DAG  $G$  perfectly captures all and only the independence properties of some distribution  $P$ , i.e.,  $I(G) = I(P)$ .

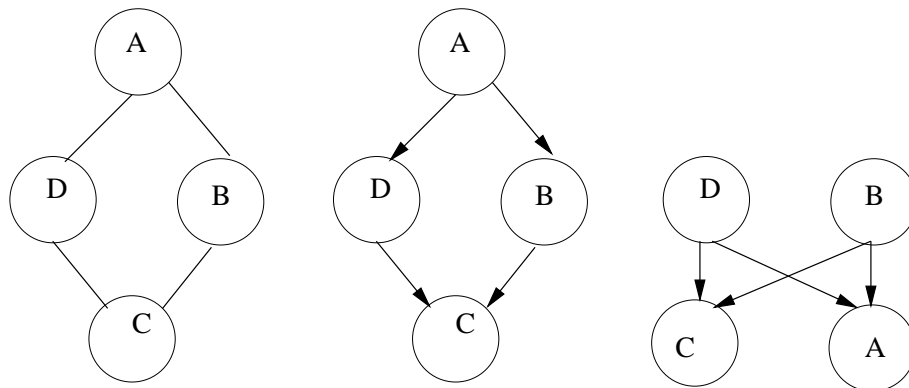


- Now consider a different node ordering:  $L, S, G, I, D$
- Now consider a different node ordering:  $L, D, S, I, G$
- All encode the same distribution!

## PERFECT MAPS

---

- Can we find a graph that captures all the independencies in an arbitrary distribution (and no more)?
- Defn: A DAG  $G$  is a **perfect map (P-map)** for a distribution  $P$  if  $I(P) = I(G)$ .
- Thm: not every distribution has a perfect map.
- Pf by counterexample. Suppose we have a model where  $A \perp C | \{B, D\}$ , and  $B \perp D | \{A, C\}$ . This cannot be represented by any Bayes net.
- e.g., BN1 wrongly says  $B \perp D | A$ , BN2 wrongly says  $B \perp D$ .



## CHAP 5: UNDIRECTED GRAPHS

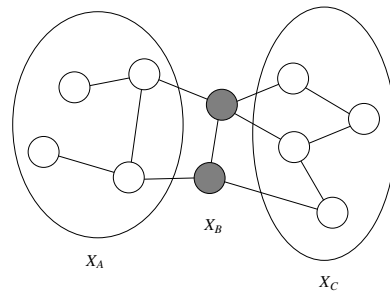
---

- Global Markov properties
- Clique potentials
- Factor graphs
- Local Markov properties
- Converting BNs  $\leftrightarrow$  MNs

# UNDIRECTED GRAPHICAL MODELS

---

- Graphs where nodes = random variables, and edges = correlation (direct dependence).
- Defn: Let  $H$  be an undirected graph. Then  $sep_H(A; C|B)$  iff all paths between  $A$  and  $C$  go through some nodes in  $B$  (simple graph separation).



- Defn: the **global Markov properties** of a UG  $H$  are

$$I(H) = \{(X \perp Y|Z) : sep_H(X; Y|Z)\}$$

- UGMs also called Markov Random Fields (MRFs) or Markov Networks.

## UNDIRECTED GRAPHICAL MODELS

---

- Defn: an undirected graphical model representing a distribution  $P(X_1, \dots, X_n)$  is an undirected graph  $H$ , and a set of positive potential functions  $\psi_c > 0$  associated with sub-cliques of  $H$ , s.t.

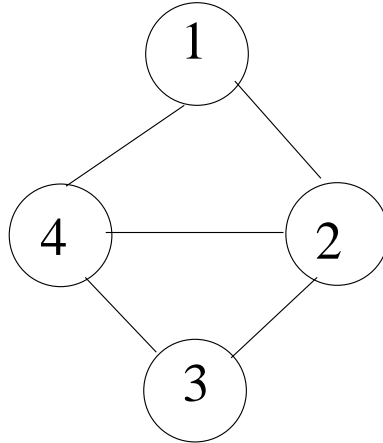
$$P(X_1, \dots, X_n) = \frac{1}{Z} \prod_{c \in C} \psi_c(x_c)$$

where  $Z$  is the partition function:

$$Z = \sum_{x_1, \dots, x_n} \prod_{c \in C} \psi_c(x_c)$$

## EXAMPLE OF UGM - MAX CLIQUES

---



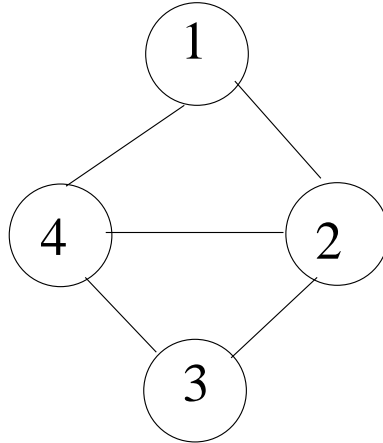
$$P(x_{1:4}) = \frac{1}{Z} \psi_{124}(x_{124}) \times \psi_{234}(x_{234})$$

$$Z = \sum_{x_1, x_2, x_3, x_4} \psi_{124}(x_{124}) \times \psi_{234}(x_{234})$$

- We can represent  $P(X_{1:4})$  as two 3D tables instead of one 4D table.

## EXAMPLE OF UGM - SUBCLIQUES

---



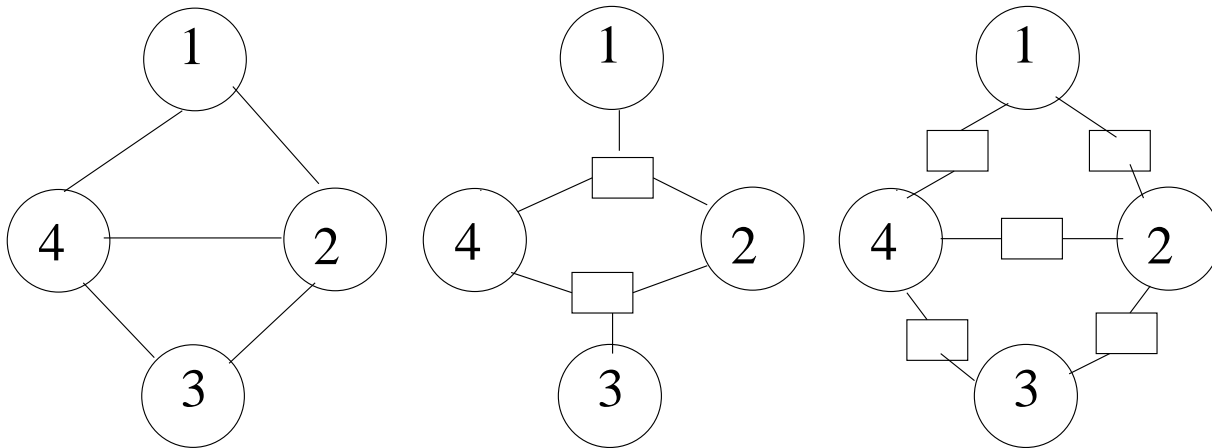
$$\begin{aligned} P(x_{1:4}) &= \frac{1}{Z} \prod_{\langle ij \rangle} \psi_{ij}(x_{ij}) \\ &= \frac{1}{Z} \psi_{12}(x_{12}) \psi_{14}(x_{14}) \psi_{23}(x_{23}) \psi_{24}(x_{24}) \psi_{34}(x_{34}) \\ Z &= \sum_{x_1, x_2, x_3, x_4} \prod_{\langle ij \rangle} \psi_{ij}(x_{ij}) \end{aligned}$$

- We can represent  $P(X_{1:4})$  as five 2D tables instead of one 4D table.

## FACTOR GRAPHS

---

- Factorized potentials can be represented graphically using a factor graph.
- Defn: a factor graph is undirected bipartite graph with two kinds of nodes. Round nodes represent variables, square nodes represent factors (potentials), and there is an edge from each variable to every factor that mentions it.



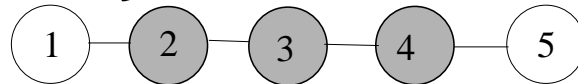
## LOCAL AND GLOBAL MARKOV PROPERTIES

---

- For directed graphs, we defined I-maps in terms of local Markov properties, and derived global independence.
- For undirected graphs, we defined I-maps in terms of global Markov properties, and will now derive local independence.
- Defn: The **pairwise markov independencies** associated with UG  $H = (V, E)$  are

$$I_p(H) = \{(X \perp Y) \mid V \setminus \{X, Y\} : \{X, Y\} \notin E\}$$

- e.g.,  $X_1 \perp X_5 \mid \{X_2, X_3, X_4\}$



## LOCAL MARKOV PROPERTIES

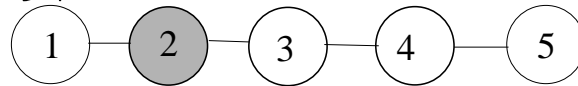
---

- Defn: The **local markov independencies** associated with UG  $H = (V, E)$  are

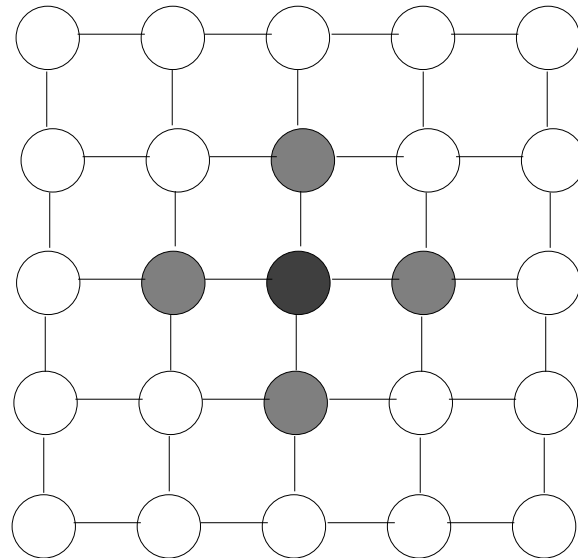
$$I_l(H) = \{(X \perp V \setminus \{X\} \setminus N_H(X) | N_H(X)) : X \in V\}$$

where  $N_H(X)$  are the neighbors

- e.g.,  $X_1 \perp \{X_3, X_4, X_5\} | X_2$



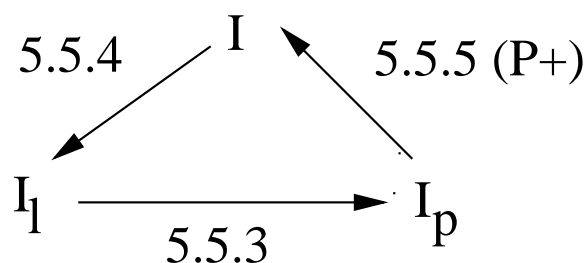
- $N_H(X)$  is also called the **Markov blanket** of  $X$ .



## RELATIONSHIP BETWEEN LOCAL AND GLOBAL MARKOV PROPERTIES

---

- Thm 5.5.3. If  $P \models I_l(H)$  then  $P \models I_p(H)$ .
- Thm 5.5.4. If  $P \models I(H)$  then  $P \models I_l(H)$ .
- Thm 5.5.5. If  $P > 0$  and  $P \models I_p(H)$ , then  $P \models I(H)$ .
- Corollary 5.5.6: If  $P > 0$ , then  $I_l = I_p = I$ .
- If  $\exists x. P(x) = 0$ , then we can construct an example (using deterministic potentials) where  $I_p \not\Rightarrow I_l$  or  $I_l \not\Rightarrow I$ .



## PERFECT MAPS

---

- Defn: A Markov network  $H$  is a **perfect map** for  $P$  if for any  $X, Y, Z$  we have that

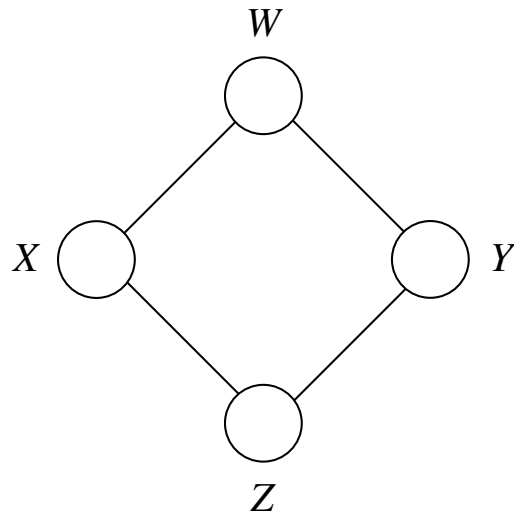
$$sep_H(X; Y|Z) \iff P \models (X \perp Y|Z)$$

- Thm: not every distribution has a perfect map.
- Pf by counterexample. No undirected network can capture all and only the independencies encoded in a v-structure  $X \rightarrow Z \leftarrow Y$ .

# EXPRESSIVE POWER

---

- Can we always convert directed  $\leftrightarrow$  undirected?
- No.

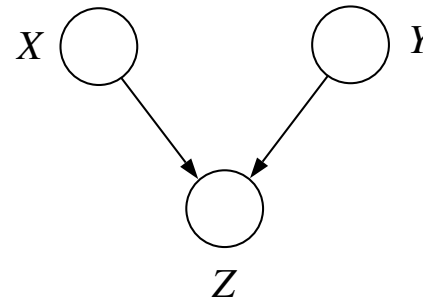


(a)

No directed model  
can represent these  
and only these  
independencies.

$$\mathbf{x} \perp \mathbf{y} \mid \{\mathbf{w}, \mathbf{z}\}$$

$$\mathbf{w} \perp \mathbf{z} \mid \{\mathbf{x}, \mathbf{y}\}$$



(b)

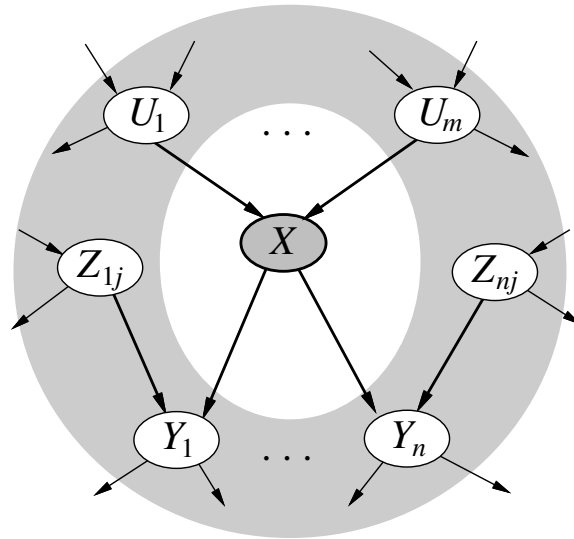
No undirected model  
can represent these  
and only these  
independencies.

$$\mathbf{x} \perp \mathbf{y}$$

## CONVERTING BAYES NETS TO MARKOV NETS

---

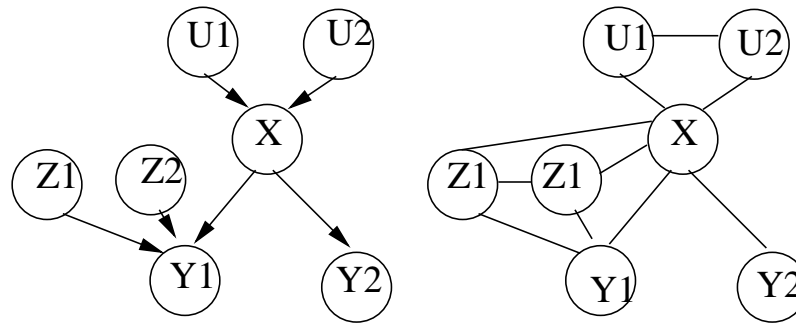
- Defn: A Markov net  $H$  is an I-map for a Bayes net  $G$  if  $I(H) \subseteq I(G)$ .
- We can construct a minimal I-map for a BN by finding the minimal Markov blanket for each node.
- We need to block all active paths coming into node  $X$ , from parents, children, and co-parents; so connect them all to  $X$ .



## MORALIZATION

---

- Defn: the moral graph  $H(G)$  of a DAG is constructed by adding undirected edges between any pair of disconnected (“unmarried”) nodes  $X, Y$  that are parents of a child  $Z$ , and then dropping all remaining arrows.
- Thm 5.7.5: The moral graph  $H(G)$  is the minimal I-map for Bayes net  $G$ .

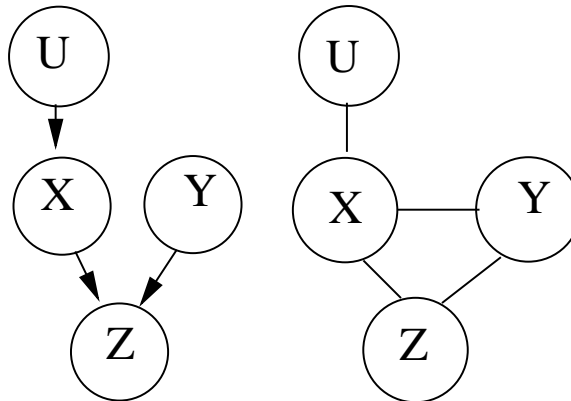


## BAYES NET TO MARKOV NET

---

- We assign each CPD to one of the clique potentials that contains it, e.g.

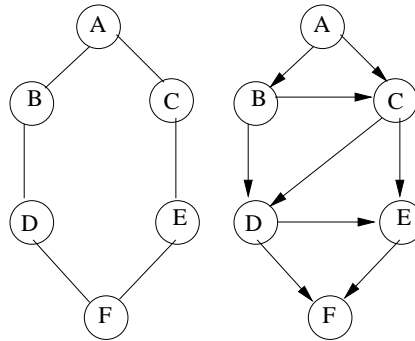
$$\begin{aligned} P(U, X, Y, Z) &= \frac{1}{Z} \psi(U, X) \times \psi(X, Y, Z) \\ &= \frac{1}{1} P(U) P(X|U) \times P(Y) P(Z|X, Y) \\ &= P(X, U) \times P(Z|X, Y) P(Y) \end{aligned}$$



## FROM MARKOV NETS TO BAYES NETS

---

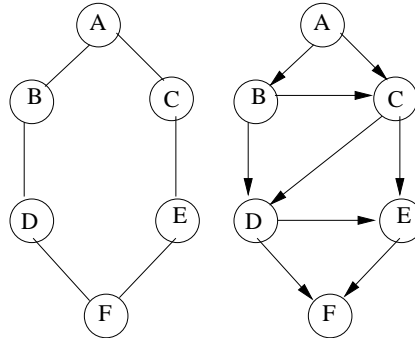
- Defn: A Bayes net  $G$  is an I-map for a Markov net  $H$  if  $I(G) \subseteq I(H)$ .
- We can construct a directed I-map by choosing a node ordering, and then picking the parents of node  $X_i$  as the subset  $U$  that renders  $X_i$  independent of its other predecessors  $X_1, \dots, X_{i-1}$ .
- e.g., when we add  $C$ , the ancestors are  $A, B$ ; since  $C \not\perp B|A$ , we need to add an edge from  $B$  to  $C$ .



- Different orderings may induce different edges.

# GRAPH TRIANGULATION

---



- The example above showed how we added extra edges to the DAG so that the largest loop was a 3-cycle (triangle).
- Defn: An undirected graph is called **chordal** or **triangulated** if every loop  $X_1 - X_2 \cdots X_k - X_1$  for  $k \geq 4$  has a chord, i.e., an edge connecting  $X_i$  and  $X_j$  for  $i, j$  non-adjacent.
- Defn: a directed graph is chordal if its underlying undirected graph is chordal.
- Thm 5.7.15: If  $G$  is a minimal I-map for Markov net  $H$ , then  $G$  is chordal.

## CHORDAL GRAPHS

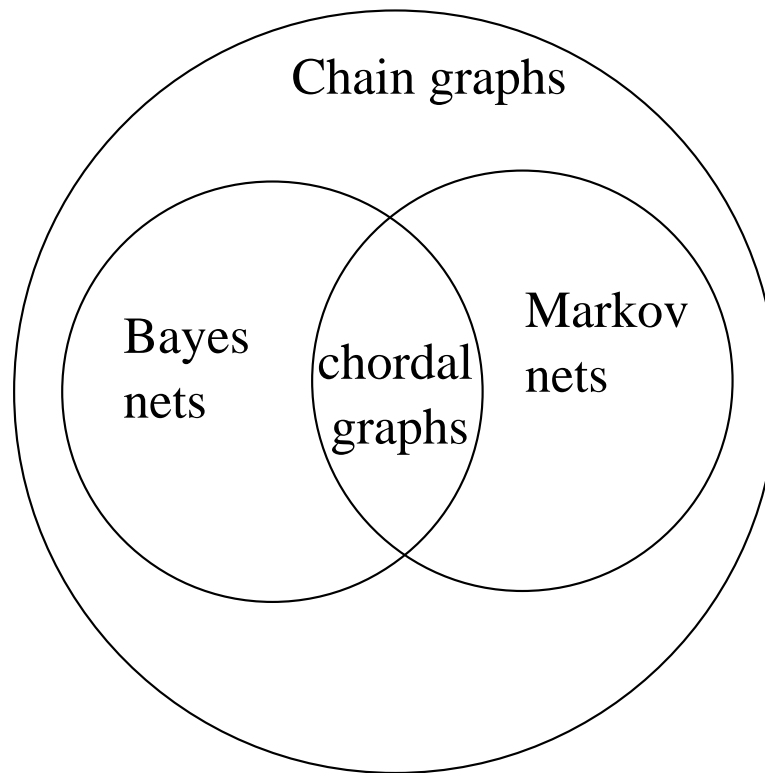
---

- Converting a Bayes net to a Markov net adds extra moralization arcs.
- Converting a Markov net to a Bayes net adds extra triangulation arcs.
- Q: When can we convert a BN to a MN or vice versa without having to add extra arcs?
- A: when the graph is chordal.
- Thm 5.7.18 (if): Let  $H$  be a chordal Markov net. Then there is a Bayes net  $G$  s.t.  $I(H) = I(G)$ .
- Thm 5.7.16 (only-if): Let  $H$  be a non-chordal Markov net. Then there is no Bayes net  $G$  s.t.  $I(H) = I(G)$ .

# CHORDAL GRAPHS

---

- Chordal graphs encode independencies that can be exactly represented by either directed or undirected graphs.
- Chain graphs combine directed and undirected graphs and represent a larger set of distributions.



## CHAP 4: LOCAL STRUCTURE IN CPDs

---

- Exponential family
- Generalized linear models
- Context-specific independence (tree-structured CPDs)
- Causal independence (noisy-or)

## EXPONENTIAL FAMILY

---

- For a random variable  $\mathbf{x}$  with no parents

$$\begin{aligned} p(\mathbf{x}|\eta) &= h(\mathbf{x}) \exp\{\eta^\top T(\mathbf{x}) - A(\eta)\} \\ &= \frac{1}{Z(\eta)} h(\mathbf{x}) \exp\{\eta^\top T(\mathbf{x})\} \end{aligned}$$

is an exponential family distribution with *natural parameter*  $\eta$ .

- Function  $T(\mathbf{x})$  is a *sufficient statistic*.
- Function  $A(\eta) = \log Z(\eta)$  is the log normalizer.
- Key idea: all you need to know about the data in order to estimate parameters is captured in the summarizing function  $T(\mathbf{x})$ .
- Examples: Bernoulli, binomial/geometric/negative-binomial, Poisson, gamma, multinomial, Gaussian, ...

## GENERALIZED LINEAR MODELS

---

- Consider the CPD for  $Y$  with parent  $X$ .
- A GLM is when  $p(\mathbf{y}|\mathbf{x})$  is exponential family with conditional mean  $\mu_i = f_i(\theta^\top \mathbf{x})$ .
- The choice of exponential family member is dictated by the *type* of  $Y$ :
  - Class labels: Bernoulli or Multinomial
  - Counts: Poisson
  - Real valued: Gaussian
- The link function  $f_i$  is usually fixed, too.

## GLM CPDS FOR $X \rightarrow Y$

$X$	$Y$	$p(Y X)$
$\mathbb{R}^n$	$\mathbb{R}^m$	Gauss( $Y; WX + \mu, \Sigma$ )
$\mathbb{R}^n$	$\{0, 1\}$	Bernoulli( $Y; p = \frac{1}{1+e^{-\theta^T x}}$ )
$\{0, 1\}^n$	$\{0, 1\}$	Bernoulli( $Y; p = \frac{1}{1+e^{-\theta^T x}}$ )
$\mathbb{R}^n$	$\{1, \dots, K\}$	Multinomial( $Y; p_i = \text{softmax}(x, \theta)$ )

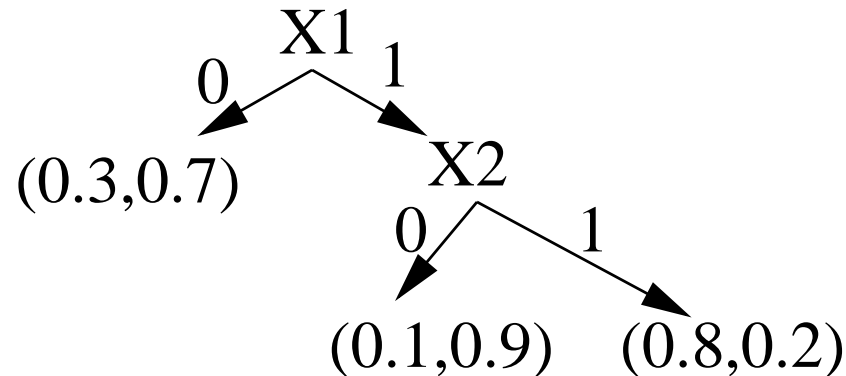
## OTHER CPDS FOR $X \rightarrow Y$

$X$	$Y$	$p(Y X)$
$\mathbb{R}^n$	$\mathbb{R}$	regression-box( $Y; X$ )
$\mathbb{R}^n$	$\{1, \dots, K\}$	classification-box( $Y; x$ )
$\{1, \dots, L\}$	$\mathbb{R}^n$	Gauss( $Y; \mu_X, \Sigma_X$ )
$\{1, \dots, L\}^n$	$\mathbb{R}$	regression-tree( $Y; X$ )
$\{1, \dots, L\}$	$\{1, \dots, K\}$	$L \times K$ CPT
$\{1, \dots, L\}^n$	$\{1, \dots, K\}$	classification-tree( $Y; X$ )
$\{0, 1\}^n$	$\{0, 1\}$	noisy-or

## CONTEXT-SPECIFIC INDEPENDENCE

---

- CSI is when some links in the graph can be removed depending on the values of certain variables.
- eg.  $P(Y|X_1, X_2)$  is represented as this decision tree:

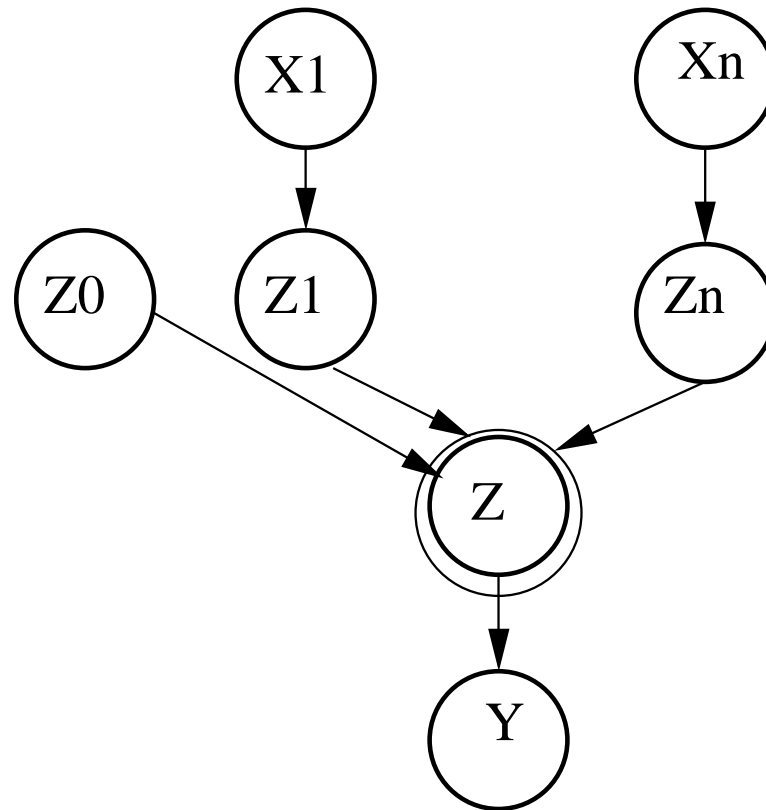


- If  $X_1 = 1$ , then the link from  $X_2 \rightarrow Y$  can be removed.
- This property arises in data association problems: let  $Z$  determine the identity of the observation; then  $P(Y|Z = i, X_{1:n}) = f(Y, X_i)$ .
- This property can be exploited in inference (condition on  $Z$  and the graph becomes sparser).

## INDEPENDENCE OF CAUSAL INFLUENCE

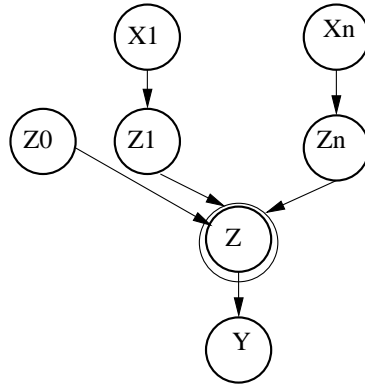
---

- A CPD  $P(Y|X_{1:n})$  exhibits ICI if it can be represented as a mini Bayes net as shown below, where  $Z$  is a deterministic function of the  $Z_i$ 's.



## NOISY-OR

---



- $Y = Z$ ,  $Z$  is deterministic OR of  $Z_i$ 's, but the link from  $X_i$  to  $Z_i$  flips 1's to 0's w.p.  $q_i$ .  $Z_0 = 1$  is always on (leak node). Hence

$$P(Y = 0 | X_{1:n}) = q_0 \prod_{i: X_i=1} q_i = q_0 \prod_i q_i^{X_i} = q_0 \sum_i e^{X_i \log q_i}$$

- Similar to sigmoid, but parameters are constrained  $q_i \in [0, 1]$ .
- Can be used to speed up inference.
- Cognitively plausible.