3. Markov property

- Markov property for MRFs
- Hammersley-Clifford theorem
- Markov property for Bayesian networks
- I-map, P-map, and chordal graphs
- (primarily based on Lauritzen book)

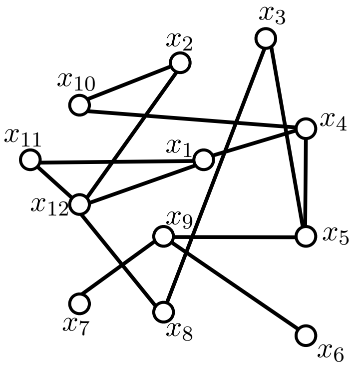
- Markov Chain

$$X\text{–}Y\text{–}Z$$

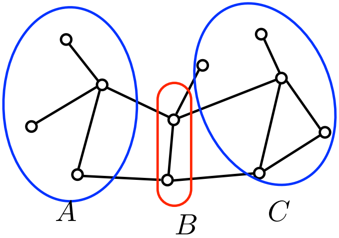
$$X \perp Z | Y$$

$$\mu(X, Y, Z) = f(X, Y) g(Y, Z)$$

Q. What independence does MRF imply?



Markov property



let $A \cup B \cup C$ be a partition of V

Definition: graph separation

- B separates A from C if any path starting in A and terminating in C has at least one node in B

Definition: global Markov property

- distribution μ over $\mathcal{X}^V$ satisfies the global Markov property on G if for any partition (A, B, C) such that B separates A from C,

$$\mu(x_A, x_C | x_B) = \mu(x_A | x_B)\, \mu(x_C | x_B)$$

Markov property for undirected graphs

- We say $\mu(\cdot)$ satisfy the **global Markov property** (G) w.r.t. a graph G if for any partition (A, B, C) such that B separates A from C,

$$\mu(x_A, x_C | x_B) = \mu(x_A | x_B)\, \mu(x_C | x_B)$$

- We say $\mu(\cdot)$ satisfy the **local Markov property** (L) w.r.t. a graph G if for any $i \in V$,

$$\mu(x_i, x_{\text{rest}} | x_{\partial i}) = \mu(x_i | x_{\partial i})\, \mu(x_{\text{rest}} | x_{\partial i})$$

- We say $\mu(\cdot)$ satisfy the **pairwise Markov property** (P) w.r.t. a graph G if for any $i, j \in V$ that are not connected by an edge

$$\mu(x_i, x_j | x_{\text{rest}}) = \mu(x_i | x_{\text{rest}})\, \mu(x_j | x_{\text{rest}})$$

obviously: (G) $\Rightarrow$ (L) $\Rightarrow$ (P)

Proof of (L)⇒(P): suppose (L) holds, then for any i and j not connected by an edge. since, for any deterministic function $h(\cdot)$,

$$X \perp Y | Z \quad \Longrightarrow \quad X \perp Y | (Z, h(Y)) \tag{1}$$

$$X \perp Y | Z \quad \Longrightarrow \quad X \perp h(Y) | Z \tag{2}$$

$$\begin{aligned} & x_i \perp x_{\text{rest}} | x_{\partial i} \\ \overset{(1)}{\Longrightarrow} \quad & x_i \perp x_{\text{rest}} | (x_{V \setminus \{i,j\}}) \\ \overset{(2)}{\Longrightarrow} \quad & x_i \perp x_j | (x_{V \setminus \{i,j\}}) \end{aligned}$$

□

proofs of (1) **and** (2)**:**

$$\begin{aligned} \mu(x, y, h(Y) = h, z) &= \mu(x, y, z) \mathbb{I}(h(y) = h) \\ &= f(x, z) g(y, z) \mathbb{I}(h(y) = h) \end{aligned}$$

this implies both $X \perp (Y, h(Y)) | Z$ and $X \perp Y | (Z, h(Y))$

$\mu(x, h, z) = \sum_y \mu(x, y, h, z) = f(x, z) \underbrace{\sum_y g(y, z) \mathbb{I}(h(y) = h)}_{\tilde{g}(h, z)}$

this implies $X \perp h(Y) | Z$

what conditions do we need for $(P) \Rightarrow (G)$ to hold?

in general $(P) \not\Rightarrow (G)$, but $(P) \Rightarrow (G)$ if the following holds for all disjoint subsets A,B,C, and $D \subseteq V$:

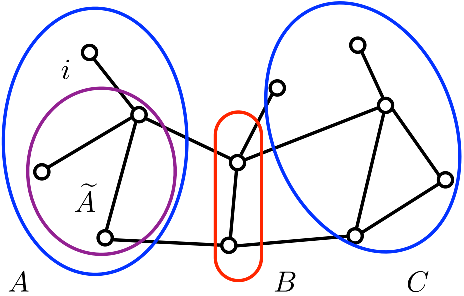
intersection lemma

if $x_A \perp x_B|(x_C, x_D)$ and $x_A \perp x_C|(x_B, x_D)$, then $x_A \perp (x_B, x_C)|x_D$ (3)

for instance, for a strictly positive $\mu(\cdot)$, i.e. if $\mu(x) > 0$ for all $x \in \mathcal{X}^V$, then the above holds [Exercise 3.1]

Proof of (P)⇒(G) when (3) **holds:** [Pearl,Paz 1987]
by induction over $s \triangleq |B|$

- initial condition: when $s = n - 2$, (P)⇔(G)
- induction step:
  assume (G) for any B with $|B| \geq s$ and prove it for $|B| = s - 1$



by induction assumption

- $x_C \perp x_{\widetilde{A}} | (x_B, x_i)$
- $x_C \perp x_i | (x_B, x_{\widetilde{A}})$

by (3),

- $x_C \perp (x_{\widetilde{A}}, x_i) | x_B$

by induction, we see that (G) holds for all sizes of B. □

Hammersley-Clifford

Theorem (Hammersley-Clifford) 1971
A strictly positive distribution $\mu(x)$ (i.e. $\mu(x) > 0$ for all x) satisfies the global Markov property (G) with respect to $G(V, E)$ if and only if it can be factorized according to G:

$$\text{(F)}: \qquad \mu(x) = \frac{1}{Z} \prod_{c \in \mathcal{C}(G)} \psi_c(x_c)$$

- i.e. any $\mu(x)$ with Markov property can be represented with MRF
- (F) $\Rightarrow$ (G) is easy [Exercise 2.1]
- (F) $\Leftarrow$ (G) requires much more effort

Proof sketch (Grimmett, Bull. London Math. Soc. 1973)
for every $S \subseteq V$, define pseudo compatibility functions

$$\widetilde{\psi}_S(x_S) \triangleq \prod_{U \subseteq S} \mu(x_U, 0_{V \setminus U})^{(-1)^{|S \setminus U|}}$$

- need strict positive $\mu(\cdot)$ for the division to make sense

example: For $S = \{i\}$ let $\mu_i^+(\,\cdot\,) \triangleq \mu(\,\cdot\,, 0_{V \setminus \{i\}})$

$$\widetilde{\psi}_i(x_i) = \frac{\mu_i^+(x_i)}{\mu_i^+(0)}$$

for $S = \{i, j\}$ (not necessarily an edge) let $\mu_{ij}^+(\,\cdot\,) \triangleq \mu(\,\cdot\,, 0_{V \setminus \{i,j\}})$

$$\widetilde{\psi}_{ij}(x_i, x_j) = \frac{\mu_{ij}^+(x_i, x_j)\,\mu_{ij}^+(0, 0)}{\mu_{ij}^+(x_i, 0)\,\mu_{ij}^+(0, x_j)}$$

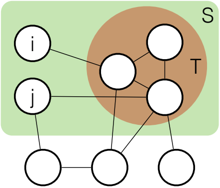
claim 1: $\widetilde{\psi}_S(x_S)$ =const. unless S is a clique
claim 2: for any $\mu(x)$, $\mu(x) = \mu(0_V) \prod_{S \subseteq V} \widetilde{\psi}_S(x_S)$

then, it follows from claims 1 and 2 that

$$\mu(x) \;=\; \frac{1}{Z}\mu(0_V) \prod_{c \in \mathcal{C}(G)} \widetilde{\psi}_c(x_c)$$

this proves that for any positive $\mu(x)$ satisfying (G) we can find a factorization as per (F) □

Proof of claim 1:



for any $i, j \in S$, expand the product as

$$\widetilde{\psi}_S(x_S) = \prod_{U \subseteq S} \mu_U^+(x_U)^{(-1)^{|S \setminus U|}}$$

$$= \prod_{U \subseteq T} \left(\frac{\mu_{U\cup\{i,j\}}^+(x_U, x_i, x_j) \;\; \mu_U^+(x_U)}{\mu_{U\cup\{i\}}^+(x_U, x_i) \;\; \mu_{U\cup\{j\}}^+(x_U, x_j)} \right)^{(-1)^{|T \setminus U|}}$$

and unless S is a clique, there exists two nodes i and j in S that are not connected by an edge, then

$$\frac{\mu_{U\cup\{i,j\}}^+(x_U, x_i, x_j)}{\mu_{U\cup\{i\}}^+(x_U, x_i)} \overset{(G)}{=} \frac{\mu(x_i|x_U, 0_{\text{rest}}) \; \mu(x_U, x_j, 0_{\text{rest}})}{\mu(x_i|x_U, 0_{\text{rest}}) \; \mu(x_U, 0_j, 0_{\text{rest}})}$$

$$= \frac{\mu(0_i|x_U, 0_{\text{rest}}) \; \mu(x_U, x_j, 0_{\text{rest}})}{\mu(0_i|x_U, 0_{\text{rest}}) \; \mu(x_U, 0_j, 0_{\text{rest}})}$$

$$= \frac{\mu(x_U, 0_i, x_j, 0_{\text{rest}})}{\mu(x_U, 0_i, 0_j, 0_{\text{rest}})} = \frac{\mu_{U\cup\{j\}}^+(x_U, x_j)}{\mu_U^+(x_U)}$$

proof of claim 2 (where did $(-1)^{|S\setminus U|}$ come from?): recall,

$$\widetilde{\psi}_S(x_S) \triangleq \prod_{U\subseteq S} \mu(x_U, 0_{V\setminus U})^{(-1)^{|S\setminus U|}}$$

Möbius inversion lemma

[see also G.C.Rota, Prob. Theor. Rel. Fields, 2 (1964) 340-368] let $f, g : \{$ subsets of $V\ \} \to \mathbb{R}$. Then the following are equivalent

$$f(S) = \sum_{U\subseteq S} g(U)\,, \quad \text{for all } S \subseteq V$$

$$g(S) = \sum_{U\subseteq S} (-1)^{|S\setminus U|} f(U)\,, \quad \text{for all } S \subseteq V$$

let, $f(S) := \log \mu(x_S, 0_{V\setminus S})$, then

$$g(S) = \sum_{U\subseteq S} (-1)^{|S\setminus U|} \log \mu(x_U, 0_{V\setminus U}) = \log \tilde{\psi}_S(x_S)$$

hence, $\exp(f(V)) = \mu(x) = \exp\Big(\sum_{U\subseteq V} \log \widetilde{\psi}_U(x_U)\Big)$

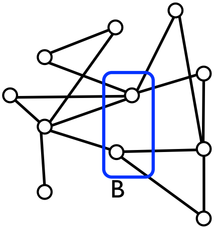
- Give an example of a distribution which is not strictly positive, that does not satisfy (G) [Exercise 3.1]

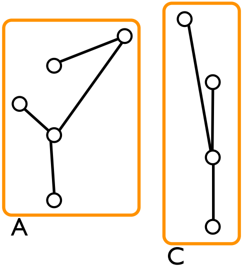
Recap

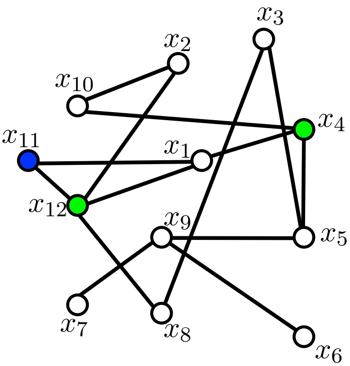
- consider a distribution $\mu(x)$ that factorizes according to an undirected graphical model on $G = (V, E)$,

$$\mu(x) \;=\; \frac{1}{Z}\prod_{c\in\mathcal{C}} \psi_c(x_c)$$

  where $\mathcal{C}$ is the set of all maximal cliques in G
- (Global Markov property) for any disjoint subsets $A, B, C \subseteq V$, $\mu(x)$ satisfy x_A–x_B–x_C whenever B separates A and C







example: color the nodes with $\{R, G, B\}$ such that no adjacent node has the same color

$$\mu(x) = \frac{1}{Z} \prod_{(i,j) \in E} \mathbb{I}(x_i \neq x_j)$$

for a node i, if we condition on the color of the neighbors of i, color of i is independent of the rest of the graph

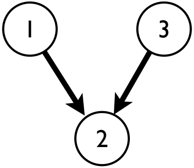
- **Hammersley-Clifford theorem**
  - (pairwise) if positive $\mu(x)$ satisfies all conditional independences implied by a graph G without any triangles, then we can find a factorization

$$\mu(x) \quad = \quad \frac{1}{Z} \prod_{(i,j)\in E} \psi_{ij}(x_i, x_j)$$

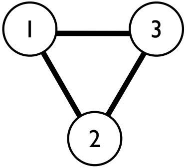
  - (general) if positive $\mu(x)$ satisfies all conditional independences implied by a graph G, then we can find a factorization

$$\mu(x) \quad = \quad \frac{1}{Z} \prod_{c\in\mathcal{C}} \psi_c(x_c)$$

- there are conditional independencies that cannot be represented by an undirected graphical model, but is represented by a Bayesian network
  Example: $\mu(x) = \mu(x_1)\mu(x_3)\mu(x_2|x_1, x_3)$



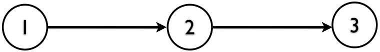
x_1 and x_3 are independent



no independence

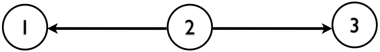
Markov property of directed graphical models

Examples



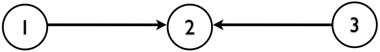
$$\mu(x) = \mu(x_1)\mu(x_2|x_1)\mu(x_3|x_2)$$

since $\mu(x_3|x_2) = \mu(x_3|x_1, x_2)$,
we have x_1–x_2–x_3



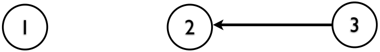
$$\mu(x) = \mu(x_2)\mu(x_1|x_2)\mu(x_3|x_2)$$

since $\mu(x_3|x_2) = \mu(x_3|x_1, x_2)$,
we have x_1–x_2–x_3



$$\mu(x) = \mu(x_1)\mu(x_3)\mu(x_2|x_1, x_3)$$

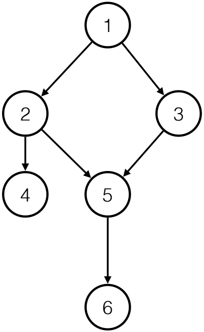
x_1 and x_3 are independent
but not x_1–x_2–x_3



$$\mu(x) = \mu(x_1)\mu(x_3)\mu(x_2|x_3)$$

x_1 is independent of (x_2, x_3)

- there are simple independencies we can immediately read from the **factorization** : $x_i \perp x_{\mathrm{pr}(i)\setminus\pi(i)} | x_{\pi(i)}$ (pr(i) is the set of predecessors in a topological ordering)



immediate

$x_2 \perp x_3 | x_1$

$x_3 \perp x_2, x_4 | x_1$

$x_4 \perp x_1, x_3, x_5, x_6 | x_2$

$x_5 \perp x_1, x_4 | x_2, x_3$

$x_6 \perp x_1, x_2, x_3, x_4 | x_5$

less immediate

$x_3 \perp x_4 | x_1$, but $x_3 \not\perp x_4 | x_1, x_6$, etc.

- however, there are more independencies that follows from the factorization
- using the idea of **d-separation**, there is a way (e.g. **Bayes ball algorithm**) to get all the independencies that are implied by all joint distributions that factorize according to a DAG G

note that a particular distribution $\mu(\cdot)$ that factorize as G might have more independence than specified by the graph (for example, complete independent distribution ($\mu(x) = \prod_i \mu(x_i)$) always can be represented by any DAG), but we are interested in the family of all μ's that factorize as per G, and **statements about the independencies satisfied by all of the μ's in that family** formally, let $\mathrm{nd}(i)$ be the set of non-descendants of node i, which are the set of nodes that are not reachable from node i via directed paths

we say $\mu(x)$ has the **local Markov property** (DL) w.r.t. a directed acyclic graph G, if for all i

$$\mu(x_i, x_{\mathrm{nd}(i)\setminus\pi(i)}|x_{\pi(i)}) = \mu(x_i|x_{\pi(i)})\mu(x_{\mathrm{nd}(i)\setminus\pi(i)}|x_{\pi(i)})$$

Definition: topological ordering is an ordering where any node comes after all of its parents (not necessarily unique)
let $\mathrm{pr}(i)$ be the set of predecessors in a topological ordering

we say $\mu(x)$ has the **ordered Markov property** (DO) w.r.t a directed acyclic graph G, if for all i and all topological ordering

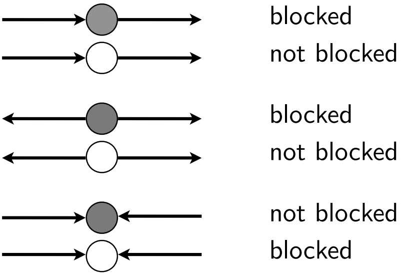
$$\mu(x_i, x_{\mathrm{pr}(i)\setminus\pi(i)}|x_{\pi(i)}) = \mu(x_i|x_{\pi(i)})\mu(x_{\mathrm{pr}(i)\setminus\pi(i)}|x_{\pi(i)})$$

(DL) $\Leftrightarrow$ (DO), since for each i there exists a topological ordering that places all $x_{\mathrm{nd}(i)}$ before x_i

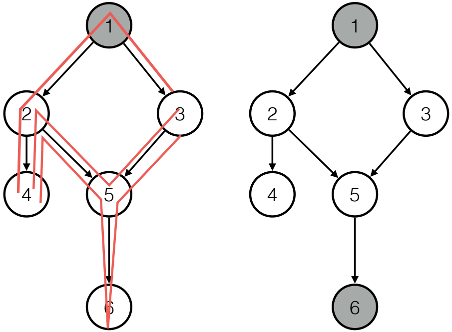
we say a distribution $\mu(x)$ has the **global Markov property** (DG) w.r.t. an acyclic directed graph G, if for all subsets A, B, C such that A and C are **d-separated** by B,

$$\mu(x_A, x_C|x_B) = \mu(x_A|x_B)\mu(x_C|x_B) \text{ or equivalently } \quad x_A \perp x_C|x_B$$

a **trail** from i to j in a directed acyclic graph is **blocked by** S if it contains a (shaded) vertex $k \in S$ such that one of the following happens



two subset A and C are **d-separated** by B if all trails from A to C are blocked by one of the above cases



$x_3 \perp x_4 | x_1$ $\qquad$ $x_3 \not\perp x_4 | x_1, x_6$

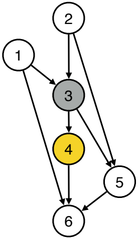
A **trail** from a node x_4 to another node x_6 in a directed graph is any path from x_4 to x_6 that ignores the direction of each of the edges in the path (you can go in the opposite direction of the edge)

- a Bayesian network on a directed acyclic graph G factorizes as

$$\text{(DF)}: \qquad \mu(x) \;=\; \prod_{i \in V} \mu_i(x_i | x_{\pi(i)})$$

(DF) $\Leftrightarrow$ (DG) $\Leftrightarrow$ (DL) $\Leftrightarrow$ (DO)

  - **proof of (DG)$\Rightarrow$(DL):** the non-descendants of a node are d-separated from the node i conditioned on $\pi(i)$ (proof by example)



by (DG) it follows that $x_4 \perp x_1, x_2 | x_3$

  - **proof of (DL)$\Rightarrow$(DF):** follows from the definition of conditional independence and Bayes rule

- **proof of (DF)⇒(DG):** requires an alternate (actually the original) definition of global Markov property

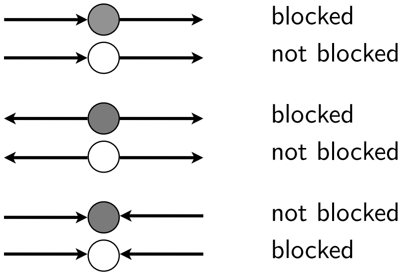
  Lemma 1: If $\mu(x)$ factorizes according to a DAG G, then for any **ancestral set** A, the marginal distribution $\mu(x_A)$ factorizes according to G_A, where G_A is the subgraph defined from G by deleting all nodes in A^C
  Proof: consider the factorization and marginalize out all not in the ancestral set.
  Lemma 2: If $\mu(x)$ factorizes according to a DAG G, then $x_A \perp x_B | x_C$ whenever A and B are separated by C in the undirected graph $\mathrm{moral}(G(\mathrm{an}(A \cup B \cup C)))$, which is the **moral graph** of the smallest ancestral set containing $A \cup B \cup C$.
  Proof: use Lemma 1.

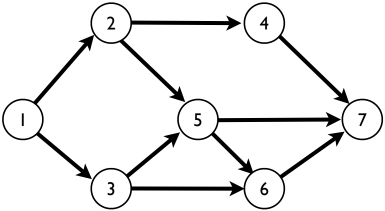
  the above definition of independence in the Lemma 2. is the original definition of global Markov property for DAGs

  it is equivalent to the condition (DG) (defined in the previous slide), but we will not prove this equivalence in the lecture □

for any distribution $\mu(x)$ that factorizes accroding to a directed acyclic graph (DAG) G, and for any disjoint subsets $A, B, C \subseteq V$, we can test whether x_A–x_B–x_C using **Bayes ball algorithm**:

1. shade all nodes in B
2. place a ball at each node in A
3. let balls propagate, duplicating itself, following the rules shown on the right (Remark: balls do not interact)
4. if no ball can reach C, then x_A–x_B–x_C

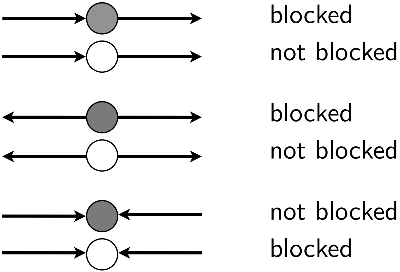


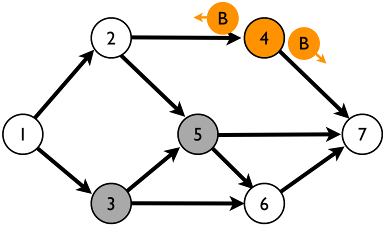


$A = \{4\}$
$B = \{3, 5\}$
$C = \{6\}$

for any distribution $\mu(x)$ that factorizes accroding to a directed acyclic graph (DAG) G, and for any disjoint subsets $A, B, C \subseteq V$, we can test whether x_A–x_B–x_C using **Bayes ball algorithm**:

1. shade all nodes in B
2. place a ball at each node in A
3. let balls propagate, duplicating itself, following the rules shown on the right (Remark: balls do not interact)
4. if no ball can reach C, then x_A–x_B–x_C

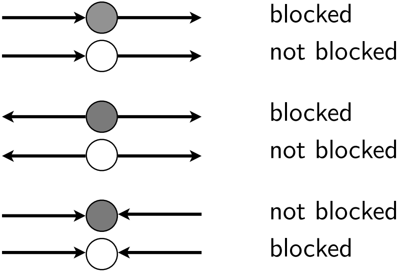


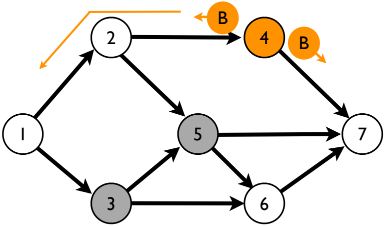


$A = \{4\}$
$B = \{3, 5\}$
$C = \{6\}$

for any distribution $\mu(x)$ that factorizes accroding to a directed acyclic graph (DAG) G, and for any disjoint subsets $A, B, C \subseteq V$, we can test whether x_A–x_B–x_C using **Bayes ball algorithm**:

1. shade all nodes in B
2. place a ball at each node in A
3. let balls propagate, duplicating itself, following the rules shown on the right (Remark: balls do not interact)
4. if no ball can reach C, then x_A–x_B–x_C

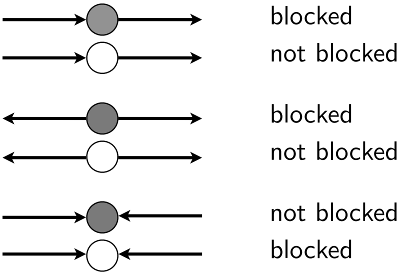


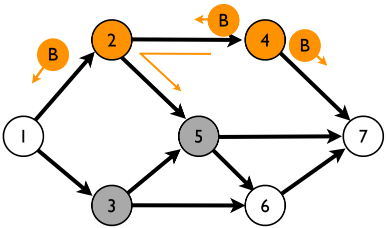


$A = \{4\}$
$B = \{3, 5\}$
$C = \{6\}$

for any distribution $\mu(x)$ that factorizes according to a directed acyclic graph (DAG) G, and for any disjoint subsets $A, B, C \subseteq V$, we can test whether x_A–x_B–x_C using **Bayes ball algorithm**:

1. shade all nodes in B
2. place a ball at each node in A
3. let balls propagate, duplicating itself, following the rules shown on the right (Remark: balls do not interact)
4. if no ball can reach C, then x_A–x_B–x_C

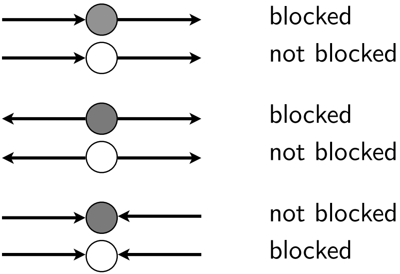


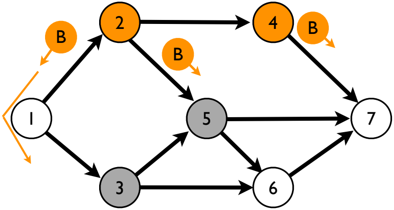


$A = \{4\}$
$B = \{3, 5\}$
$C = \{6\}$

for any distribution $\mu(x)$ that factorizes accroding to a directed acyclic graph (DAG) G, and for any disjoint subsets $A, B, C \subseteq V$, we can test whether x_A–x_B–x_C using **Bayes ball algorithm**:

1. shade all nodes in B
2. place a ball at each node in A
3. let balls propagate, duplicating itself, following the rules shown on the right (Remark: balls do not interact)
4. if no ball can reach C, then x_A–x_B–x_C

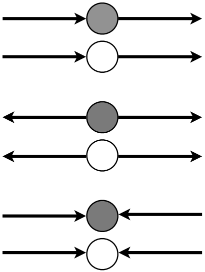




$A = \{4\}$
$B = \{3, 5\}$
$C = \{6\}$

for any distribution $\mu(x)$ that factorizes accroding to a directed acyclic graph (DAG) G, and for any disjoint subsets $A, B, C \subseteq V$, we can test whether x_A–x_B–x_C using **Bayes ball algorithm**:

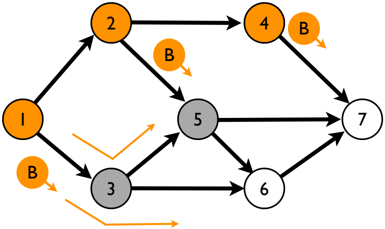
1. shade all nodes in B
2. place a ball at each node in A
3. let balls propagate, duplicating itself, following the rules shown on the right (Remark: balls do not interact)
4. if no ball can reach C, then x_A–x_B–x_C



blocked
not blocked

blocked
not blocked

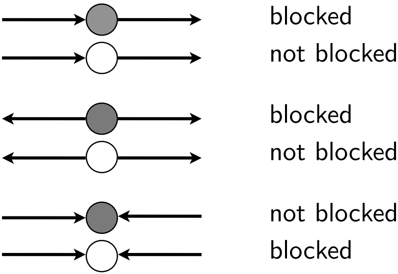
not blocked
blocked

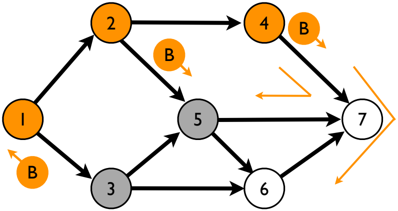


$A = \{4\}$
$B = \{3, 5\}$
$C = \{6\}$

for any distribution $\mu(x)$ that factorizes accroding to a directed acyclic graph (DAG) G, and for any disjoint subsets $A, B, C \subseteq V$, we can test whether x_A–x_B–x_C using **Bayes ball algorithm**:

1. shade all nodes in B
2. place a ball at each node in A
3. let balls propagate, duplicating itself, following the rules shown on the right (Remark: balls do not interact)
4. if no ball can reach C, then x_A–x_B–x_C

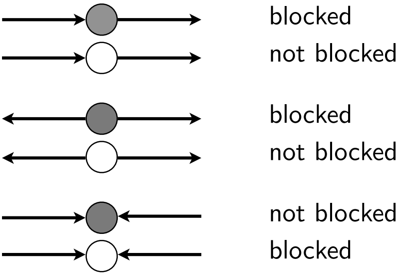


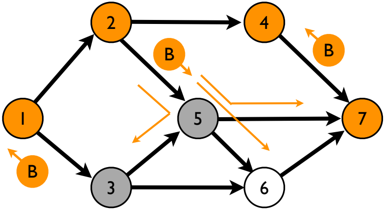


$A = \{4\}$
$B = \{3, 5\}$
$C = \{6\}$

for any distribution $\mu(x)$ that factorizes accroding to a directed acyclic graph (DAG) G, and for any disjoint subsets $A, B, C \subseteq V$, we can test whether x_A–x_B–x_C using **Bayes ball algorithm**:

1. shade all nodes in B
2. place a ball at each node in A
3. let balls propagate, duplicating itself, following the rules shown on the right (Remark: balls do not interact)
4. if no ball can reach C, then x_A–x_B–x_C

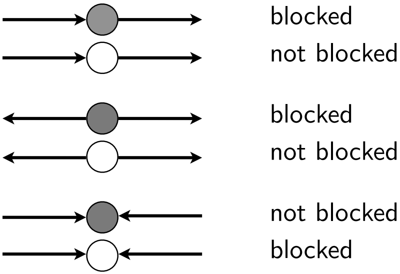


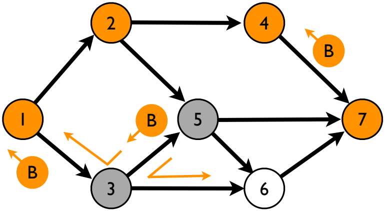


$A = \{4\}$
$B = \{3, 5\}$
$C = \{6\}$

for any distribution $\mu(x)$ that factorizes accroding to a directed acyclic graph (DAG) G, and for any disjoint subsets $A, B, C \subseteq V$, we can test whether x_A–x_B–x_C using **Bayes ball algorithm**:

1. shade all nodes in B
2. place a ball at each node in A
3. let balls propagate, duplicating itself, following the rules shown on the right (Remark: balls do not interact)
4. if no ball can reach C, then x_A–x_B–x_C

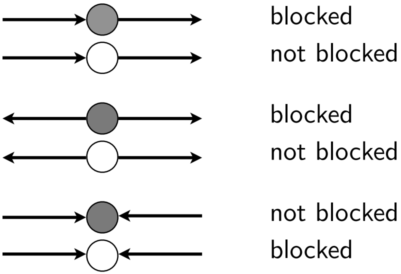




$A = \{4\}$
$B = \{3, 5\}$
$C = \{6\}$

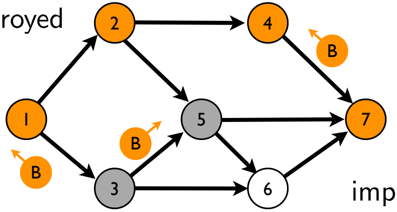
for any distribution $\mu(x)$ that factorizes accroding to a directed acyclic graph (DAG) G, and for any disjoint subsets $A, B, C \subseteq V$, we can test whether x_A–x_B–x_C using **Bayes ball algorithm**:

1. shade all nodes in B
2. place a ball at each node in A
3. let balls propagate, duplicating itself, following the rules shown on the right (Remark: balls do not interact)
4. if no ball can reach C, then x_A–x_B–x_C



blocked balls are destroyed

Q.when do we stop?



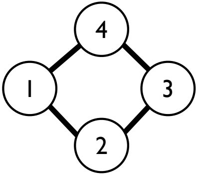
$A = \{4\}$
$B = \{3, 5\}$
$C = \{6\}$

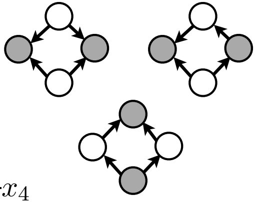
implies x_4–$\{x_3, x_5\}$–x_6

- if a distribution $\mu(x)$ factorizes according to a DAG G, i.e.

$$\mu(x) = \prod_{i \in V} \mu_i(x_i | x_{\pi(i)})$$

  then $\mu(x)$ satisfies all the conditional independencies obtainable by Bayes ball
- if a distribution $\mu(x)$ satisfies all the conditional independencies obtainable by Bayes ball on a DAG G, then we can find a factorization of $\mu(x)$ on G
- in the worst-case Bayes ball has runtime $O(|E|)$ [Shachter 1998]
- there are conditional independencies that cannot be represented by a Bayesian network, but is represented by a MRF,
  Example: $\mu(x) = \psi_{12}(x_1, x_2)\psi_{23}(x_2, x_3)\psi_{34}(x_3, x_4)\psi_{41}(x_4, x_1)$

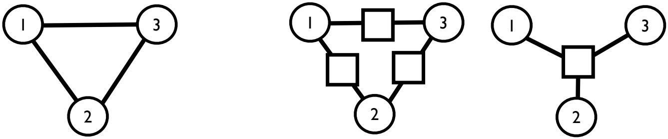




x_1–$\{x_2, x_4\}$–x_3, x_2–$\{x_1, x_3\}$–x_4

Markov property of factor graphs

Factor graphs are more 'fine grained' than undirected graphical models



$\psi(x_1, x_2, x_3)$ $\psi_{12}(x_1, x_2)\psi_{23}(x_2, x_3)\psi_{31}(x_3, x_1)$ $\psi_{123}(x_1, x_2, x_3)$

all three encodes same independencies, but different factorizations
(in particular $3|\mathcal{X}|^2$ vs. $|\mathcal{X}|^3$)

- set of independencies represented by MRF is the same as FG
- but FG can represent a larger set of factorizations
- for a factor graph $G = (V, F, E)$, for any disjoint subsets $A, B, C \subseteq V$, $\mu(x)$ satisfy x_A–x_B–x_C whenever B separates A and C

Independence maps (I-maps)

which graph (graphical model) is 'good'?

- let $\mathcal{I}(G)$ denote all conditional independencies implied by a graph G
- let $\mathcal{I}(\mu)$ denote all conditional independencies of a distribution $\mu(\cdot)$
- G is an **I-map** of μ if μ satisfy all the independencies of the graph G, i.e.

$$\mathcal{I}(G) \quad \subseteq \quad \mathcal{I}(\mu)$$

- given $\mu(x)$, can we construct a G that captures as many independencies as possible?
- G is a **minimal I-map** for $\mu(x)$ if
  - G is an I-map for $\mu(x)$, and
  - removing a single edge from G causes the graph to no longer be an I-map

Constructing minimal I-maps

Constructing minimal I-maps for Bayesian network

1. choose an (arbitrary) ordering of $x_1, x_2, \ldots, x_n$
2. consider a directed graph for Bayes rule

$$\mu(x) = \mu(x_1)\mu(x_2|x_1)\cdots\mu(x_n|x_1,\ldots,x_{n-1})$$

3. for each i, select its parents $\pi(i)$ to be the minimal subset of $\{x_1,\ldots,x_{i-1}\}$ such that

$$x_i – x_{\pi(i)} – \{x_1 \ldots, x_{i-1}\} \setminus x_{\pi(i)}$$

- minimal I-map for BN is not unique
  - there are $n!$ choices of the ordering, and a priori we cannot tell which ordering is best
  - even for the same ordering, the I-map might not be unique
    e.g. $X_1 = X_2$ and $X_3 = X_1 + Z$, then $x_3 \perp x_1|x_2$ and $x_3 \perp x_2|x_1$, but $x_3 \not\perp (x_1, x_2)$ so there are two I-maps for the ordering (1,2,3)

Constructing minimal I-maps for Markov random fields

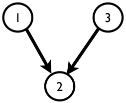
1. consider a complete undirected graph
2. for each edge (i, j), remove the edge if i and j are conditionally independent given all the rest of nodes

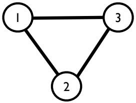
- if $\mu(x) > 0$, then the resulting graph is the unique minimal I-map of μ
- can you prove this?

  idea: when $\mu > 0$, graph obtained from the pairwise independencies is exactly the same as the graph obtained from the global independencies (from Hammersley-Clifford) and hence there is no loss in only considering the pairwise conditional independencies when constructing the graph
- when $\mu(x) = 0$ for some x, then I-map might not be unique [Homework 2.1]

Moralization

- **moralization** converts a BN D to a MRF G such that $\mathcal{I}(G) \subseteq \mathcal{I}(D)$
- resulting G is an minimal (undirected) I-map of $\mathcal{I}(D)$
  1. retain all edges in D and make them undirected
  2. connect every pair of parents with an edge





  when do we lose nothing in converting a directed graph to an undirected one? (when moralization adds no more edges than already present in D)

- there exists a G such that $\mathcal{I}(D) = \mathcal{I}(G)$ if and only if moralization of D does not add any edges

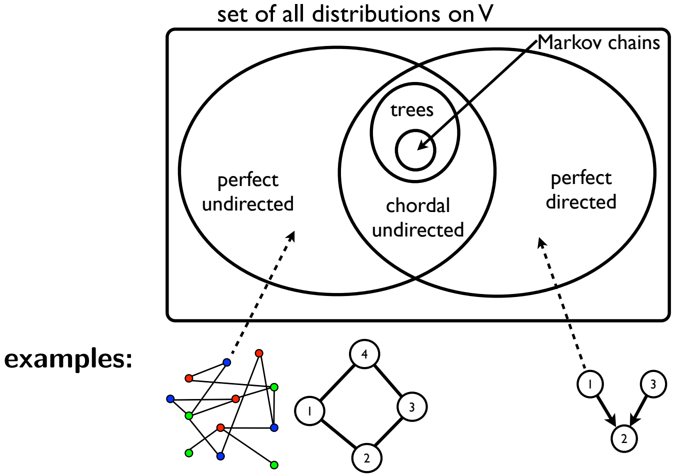
  proof of $\Rightarrow$: in the example above, the independence is not preserved, and this argument can be made general

  proof of $\Leftarrow$: we will not prove this in the lecture

Perfect maps (P-maps)

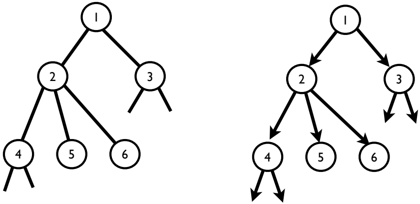
G is a **perfect map (P-map)** for $\mu(x)$ if

$$\mathcal{I}(G) \;=\; \mathcal{I}(\mu)$$



trees have efficient inference algorithms, which plays a crucial role in developing efficient algorithms for other graphs as well

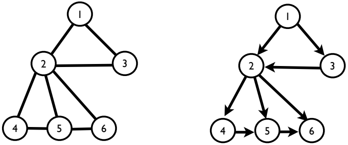
Trees



- **undirected tree** is a connected undirected graph with no cycle
- **directed tree** is a connected directed graph where each node has at most one parent
- conversion from MRF on tree to BN on tree: take any node as root and start 'directing' edges away from the root
  - non-unique
  - all conversions result in the same set of independencies
- **example**: Markov chain, hidden Markov models
- exact inference is extremely efficient on trees

Chordal graphs

- consider an undirected graph $G(V, E)$
- we say $(i_1, \ldots, i_k)$ form a **trail** in the graph $G = (V, E)$ if for every $j \in \{1, \ldots, k-1\}$, we have $(i_j, i_{j+1}) \in E$
- a **loop** is a trail $(i_1, \ldots, i_k)$ where $i_k = i_1$
- an undirected graph is **chordal** if the longest minimal loop is a triangle



- given G, if there exists a DAG D such that if $\mathcal{I}(G) = \mathcal{I}(D)$ then G is a chordal graph

  proof idea: any loop of size > 3 with no chord (a chord is an edge connecting two vertices that are not consecutive in the loop), have independency that cannot be encoded by a DAG

- the following are equivalent for MRFs
  (1) G is **chordal**
  (2) there exists an orientation of the edges in G that gives a DAG D whose **moral graph** is G
  (3) there exists a directed graphical model with conditional independencies identical to those implied by G, i.e. $\mathcal{I}(G) = \mathcal{I}(D)$

  proof of (2)⇒(3): follows from the moralization slide (we did not provide a proof)

  proof of (3)⇒(1): "proved" in the previous slide

  proof of (1)⇒(2):
  - **Lemma 3.** a chordal G is recursively simplicial.

    given a recursively simplicial graph G, we can construct a DAG D as follows:

    start with empty D_0 and fix a simplicial ordering $(1, \dots, n)$

    start from node x_1, sequentially adding one node at a time, adding ∂x_t to D_{t-1} and add edges from ∂x_t towards x_t (unless there is already an edge in the opposite direction)

    it is clear for the construction that D is acyclic

    the moral graph of D has no additional edges, since the parents of any node form a clique in the original graph

- use **triangulation** to construct such a directed graph D from an undirected chordal graph G, which we will not cover in this course

- a node in G is **simplicial** if the subgraph defined by its neighbors form a complete subgraph
- a graph G is **recursively simplicial** if it contains a simplicial node x_i and when x_i is removed from the (sub)graph, what remains is recursively simplicial
- proof of Lemma 3. is omitted here, but refer to e.g. http://www.cs.berkeley.edu/∼bartlett/courses/2009fall-cs281a/graphnotes.pdf

Roadmap

Cond. Indep. $\mu(x)$	Factorization $\mu(x)$	Graphical Model	Graph G	Cond. Indep. implied by G
x_1–$\{x_2, x_3\}$–x_4;	$\frac{1}{Z}\prod \psi_a(x_{\partial a})$	FG	Factor	Markov
x_4–$\{\}$–x_7;	$\frac{1}{Z}\prod \psi_C(x_C)$	MRF	Undirected	Markov
$\vdots$	$\prod \psi_i(x_i \vert x_{\pi(i)})$	BN	Directed	Markov

Undirected graphical models: (F) $\Leftrightarrow$ (G) $\Leftrightarrow$ (L) $\Leftrightarrow$ (P)

- For any positive $\mu(x)$ that satisfy (G), we can find a factorization (F)
- List of all positive distributions $\{\mu(x)\}$ that factorize according to (F) is equivalent as the list of all distributions that satisfy (G)

Directed graphical models: (F) $\Leftrightarrow$ (G) $\Leftrightarrow$ (L) $\Leftrightarrow$ (O)