Directed and undirected graphical models **Fall 2015**

Lecture 4 — October 21st

Lecturer: Simon Lacoste-Ju ien *Scribe: Jaime Roquero, JieYing Wu*

4.1 Notation and probability review

4.1.1 Notations

Let us recall a few notations before establishing some properties of directed graphical models. Let $X_1, X_2, \ldots, X_n$ be random variables with distribution:

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = p_X(x_1, \ldots, x_n) = p(x)$$

where x stands for $(x_1, \ldots, x_n)$. Given $A \subset \{1, \ldots, n\}$, we denote the marginal distribution of x_A by:

$$p(x_A) = \sum_{x \in A^c} p(x_A, x_{A^c}).$$

With this notation, we can write the conditional distribution as:

$$p(x_A | x_{A^c}) = \frac{p(x_A, x_{A^c})}{p(x_{A^c})}$$

We also recall the so-called 'chain rule' stating:

$$p(x_1, \ldots, x_n) = p(x_1)p(x_2|x_1)p(x_3|x_2, x_1) \ldots p(x_n|x_1, \ldots, x_{n-1})$$

4.1.2 Independence

Let A, B, and C be disjoint.
Marginal independence is defined as:

$$X_A \perp\!\!\!\perp X_B \Leftrightarrow p(x_A, x_B) = p(x_A)p(x_B) \qquad \forall x_A, x_B \tag{4.1}$$

$$X_A \perp\!\!\!\perp X_B \Leftrightarrow p(x_A|x_B) = p(x_A) \qquad \forall x_A, x_B \ s.t. \ p(x_B) > 0 \tag{4.2}$$

Conditional independence is defined as:

$$X_A \perp\!\!\!\perp X_B | X_C \Leftrightarrow p(x_A, x_B|x_C) = p(x_A|x_C)p(x_B|x_C) \qquad p(x_C) > 0 \tag{4.3}$$

$$X_A \perp\!\!\!\perp X_B | X_C \Leftrightarrow p(x_A|x_B, x_C) = p(x_A|x_C) \qquad p(x_B, x_C) > 0 \tag{4.4}$$

Three Facts About Conditional Independence

1. **Can repeat variables:** $X \perp\!\!\!\perp Y, Z|Z, W$ is the same as $X, Z \perp\!\!\!\perp Y|Z, W$. The repetition is redundant but may be convenient notation.
2. **Decomposition:** $X \perp\!\!\!\perp Y, Z|W$ implies that $X \perp\!\!\!\perp Y|W$ and $X \perp\!\!\!\perp Z|W$.
3. **Trick:** extra conditioning on both sides of the equation doesn't change anything. E.g. the following two statements are always true.

$$p(x, y) = p(x|y)p(y) \tag{4.5}$$

$$p(x, y|z) = p(x|y, z)p(y|z) \tag{4.6}$$

4.2 Directed Graphical Model

Graphical models combine probability and graph theory into an efficient data structure. We want to be able to handle probabilistic models of hundreds of variables. For example, assume we are trying to model the probability of diseases given the symptoms, as shown below.

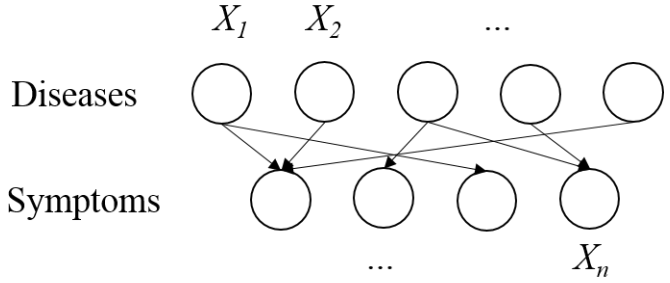


Figure 4.1. Nodes representing binary variables indicating the presence or not of a disease or a symptom.

We have n nodes, each a binary variable ($X_i \in \{0, 1\}$), indicating the presence or absence of a disease or a symptom. The number of joint probability terms would grow exponentially. For 100 diseases and symptoms, we would need a table of size 2^{100} to store all the possible states. This is clearly intractable. Instead, we will use graphical models to represent the relationships between nodes.

General issues in this class

1. Representation → DGM, UGM / parameterization → exponential family
2. Inference (computing $p(x_A|x_B)$) → sum-product algorithm
3. Statistical estimation → maximum likelihood, maximum entropy

A directed graphical model, also known as "Bayesian network", represents a *family of distributions*, denoted $\mathcal{L}(G)$, where $\mathcal{L}(G) \triangleq \{p : \exists$ legal factors, f_i, s.t. $p(x_V) = \prod_{i=1}^n f_i(x_i, x_{\pi_i})\}$, where the legal factors satisfy $f_i \geq 0$ and $\sum_{x_i} f_i(x_i, x_{\pi_i}) = 1 \ \forall i, x_{\pi_i}$.

4.2.1 First definitions and properties

Let $X_1, \ldots, X_n$ be n random variables with distribution $p(x) = p_X(x_1, \ldots, x_n)$.

Definition 4.1 *Let $G = (V, E)$ be a DAG wit $V = \{1, \ldots, n\}$. We say t at $p(x)$ factorizes in G, denoted $p(x) \in \mathcal{L}(G)$, if t ere exists some functions f_i, called factors, suc t at:*

$$\begin{aligned} \forall x, \ p(x) &= \prod_{i=1}^n f_i(x_i, x_{\pi_i}) \\ f_i \geq 0, \quad \forall i, \forall x_{\pi_i}, \ \sum_{x_i} f_i(x_i, x_{\pi_i}) &= 1 \end{aligned} \tag{4.7}$$

w ere we recall t at π_i stands for t e set of parents of t e vertex i in G.

We prove the following useful and fundamental property of directed graphical models.

Proposition 4.2 (marginalizing leaves is easy) *Suppose t at p factorizes in G, i.e. $p(x_V) = \prod_{j=1}^n f_j(x_j, x_{\pi_j})$. T en for any leaf i, we ave t at $p(x_{V\setminus\{i\}}) = \prod_{j \neq i} f_j(x_j, x_{\pi_j})$, ence $p(x_{V\setminus\{i\}})$ factorizes in $G' = (V\setminus\{i\}, E')$, t e subgrap obtained from removing t e leaf i from G.*

Proof Without loss of generality, we can assume that the leaf is labelled n. Therefore we have that $n \notin \pi_i, \forall i \leq n-1$. We have the following computation:

$$\begin{aligned} p(x_1, \ldots x_{n-1}) &= \sum_{x_n} p(x_1, \ldots, x_n) \\ &= \sum_{x_n} \left(\prod_{i=1}^{n-1} f_i(x_i|x_{\pi_i}) f_n(x_n|x_{\pi_n}) \right) \\ &= \prod_{i=1}^{n-1} f_i(x_i|x_{\pi_i}) \sum_{x_n} f_n(x_n|x_{\pi_n}) \\ &= \prod_{i=1}^{n-1} f_i(x_i|x_{\pi_i}) \end{aligned}$$

■

Remark 4.2.1 *In t is proposition we admitted t at t e new grap obtained by removing a leaf is still a DAG. Also, by induction, t is result s ows t at in t e definition of factorization we do not need to suppose t at p is a probability distribution: if a function p satisfies 4.7 t en it is a probability distribution.*

Now we try to characterize the factor functions. The following result will imply that if p factorizes in G, then we have a uniqueness of the factors.

Proposition 4.3 *If $p(x) \in \mathcal{L}(G)$ t en, for all $i \in \{1, \dots, n\}$, $f_i(x_i, x_{\pi_i}) = p(x_i|x_{\pi_i})$.*

Proof Let $|V| = n$, and i be some fixed $1 \leq i \leq n$. Without loss of generality, we suppose that $\{1, \dots, n\}$ is a topological sorting of G, such that the set $\{i+1, \dots, n\}$ corresponds to the labels of the descendants of i. We use repeatedly the previous proposition in order to pluck out (marginalize out) leaves until getting to i. Indeed, the factors f_j remain the same after plucking out the leaf n. We thus have that $p(x_{1:i}) = \prod_{j \leq i} f_j(x_j|x_{\pi_j})$.

Now, let $A = \{1, \dots, i-1\} \backslash \pi_i$. We have $1 : i$ is the disjoint union of A, π_i and $\{i\}$.

$$p(x_i, x_{\pi_i}) = \sum_{x_A} p(x_i, x_{\pi_i}, x_A) = \sum_{x_A} f_i(x_i, x_{\pi_i}) \prod_{j \leq i-1} f_j(x_j, x_{\pi_j}) = f_i(x_i, x_{\pi_i}) \sum_{x_A} \prod_{j \leq i-1} f_j(x_j, x_{\pi_j}) \tag{4.8}$$

We can factorize out $f_i(x_i, x_{\pi_i})$ because it is constant with respect to the dummy variable x_A. We can therefore compute the conditional probability:

$$p(x_i|x_{\pi_i}) = \frac{p(x_i, x_{\pi_i})}{\sum_{x_i'} p(x_i', x_{\pi_i})} = \frac{f_i(x_i, x_{\pi_i}) \sum_{x_A} \prod_{j \leq i-1} f_j(x_j, x_{\pi_j})}{\left(\sum_{x_i'} f_i(x_i', x_{\pi_i})\right) \sum_{x_A} \prod_{j \leq i-1} f_j(x_j, x_{\pi_j})} = f_i(x_i, x_{\pi_i}). \tag{4.9}$$

We justify the computations because we have that x_i' doesn't appear in $\sum_{x_A} \prod_{j \leq i-1} f_j(x_j, x_{\pi_j})$ and that $\sum_{x_i'} f_i(x_i', x_{\pi_i}) = 1$.

■

Hence we can give an equivalent definition for a DAG to the notion of factorization:

Definition 4.4 *(Equivalent definition) T e probability distribution $p(x)$ factorizes in G, denoted $p(x) \in \mathcal{L}(G)$, iff:*

$$\forall x,\ p(x) = \prod_{i=1}^{n} p(x_i|x_{\pi_i}) \tag{4.10}$$

Example 4.2.1

- *(Trivial Grap s) Assume $E = \emptyset$,* i.e. *t ere is no edges. T en we ave $p(x) = \prod_{i=1}^n p(x_i)$, implying t e random variables $X_1, \dots, X_n$ are independent. Hence variables are independent if t ey factorize in t e empty grap .*
- *(Complete Grap s) Assume now we ave a complete grap (t us wit $n(n-1)/2$ edges as we need acyclic for it to be a DAG), we ave: $p(x) = \prod_{i=1}^n p(x_i|x_1, \dots, x_{i-1})$, t e so-called 'c ain rule' w ic is always true. Every random process factorizes in t e complete grap .*

4.2.2 Graphs with three nodes

We give an insight of the different possible behaviors of a graph by thoroughly enumerating the possibilities for a 3-node graph.

- The two first options are the empty graph, leading to independence, and the complete graph that gives no further information than the chain rule.
- (Markov chain) A Markov chain is a certain type of DAG showed in Fig.(4.2). In this configuration we show that we have:

$$p(x, y, z) \in \mathcal{L}(G) \Rightarrow X \perp\!\!\!\perp Y \mid Z \tag{4.11}$$

Indeed we have:

$$p(y|z, x) = \frac{p(x, y, z)}{p(x, z)} = \frac{p(x, y, z)}{\sum_{y'} p(y', x, z)} = \frac{p(x)p(z|x)p(y|z)}{\sum_{y'} p(x)p(z|x)p(y'|z)} = p(y|z)$$

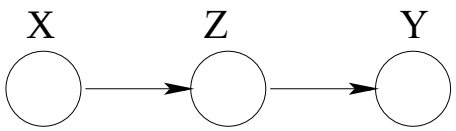


Figure 4.2. Markov Chain

- (Latent cause) It is the type of DAG given in Fig.(4.3). We show that:

$$p(x, y, z) \in \mathcal{L}(G) \Rightarrow X \perp\!\!\!\perp Y \mid Z \tag{4.12}$$

Indeed:

$$p(x, y|z) = \frac{p(x, y, z)}{p(z)} = \frac{p(z)p(y|z)p(x|z)}{p(z)} = p(x|z)p(y|z)$$

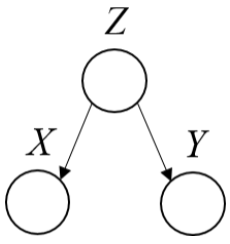


Figure 4.3. Latent cause

- (Explaining away) Represented in Fig.(4.4), we can show for this type of graph:

$$p(x) \in \mathcal{L}(G) \Rightarrow X \perp\!\!\!\perp Y \tag{4.13}$$

It basically stems from:

$$p(x,y) = \sum_z p(x,y,z) = p(x)p(y)\sum_z p(z|x,y) = p(x)p(y)$$

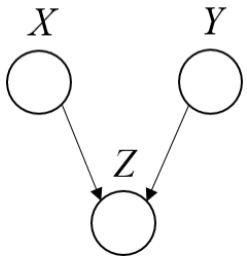


Figure 4.4. Explaining away, or V-structure

Remark 4.2.2 *T e use of 'cause' is not advised since observational statistics provide wit correlations and no causality notion. Note also t at in t e 'explaining away' grap , in general* $X \perp\!\!\!\perp Y|Z$ *is not true. Lastly, it is important to remember t at not every relations ip can be expressed in terms of grap ical models. As a counter-example take t e XOR function w ere* $Z = X \oplus Y$*, and* X, Y *are independent coin flips. T e t ree random variables are pairwise independent, but not mutually independent.*

4.2.3 Inclusion, reversal and marginalization properties

Inclusion property. Here is a quite intuitive proposition about included graphs and their factorization.

Proposition 4.5 *If* $G = (V, E)$ *and* $G' = (V, E')$ *t en:*

$$E \subset E' \Leftrightarrow \mathcal{L}(G) \subset \mathcal{L}(G') \tag{4.14}$$

Proof We have $p(x) = \prod_{i=1}^n p(x_i, x_{\pi_i(G)})$. As $E \subset E'$ it is obvious that $\pi_i(G) \subset \pi_i(G')$. Therefore, going back to the definition of graphical models through potential $f_i(x_i, x_{\pi_i})$ we get the result. ■

Reversal property. We also have some reversal properties. Let us first define the notion of V-structure.

Definition 4.6 *We say t ere is a V-structure (figure 4.4) in* $i \in V$ *if* $|\pi_i| \geq 2$*, i.e. i as two or more parents.*

Proposition 4.7 *(Markov equivalence) If $G = (V, E)$ is a DAG and if for $(i, j) \in E, |\pi_i| = 0$ and $|\pi_j| \leq 1$, t en (i, j) may be reversed,* i.e. *if $p(x)$ factorizes in G t en it factorizes in $G' = (V, E')$ wit $E' = (E - \{(i, j)\}) \cup \{(j, i)\}$.*

In terms of 3-nodes graph, this property ensures us that the Markov chain and latent cause are equivalent. On the other hand the V-structure lead to a different class of graph compared to the two others.

Definition 4.8 *An edge (i, j) is said to be covered if $\pi_j = \{i\} \cup \pi_i$.*

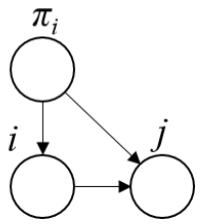


Figure 4.5. Edge (i, j) is covered

By reversing (i, j) we might not get a DAG as it might break the acyclic property. We have the following result:

Proposition 4.9 *Let $G = (V, E)$ be a DAG and $(i, j) \in E$ a covered edge. Let $G' = (V, E')$ wit $E' = (E - \{(i, j)\}) \cup \{(j, i)\}$, t en if G' is a DAG, $\mathcal{L}(G) = \mathcal{L}(G')$.*

Marginalization. The underlying question is whether the marginalization of a distribution that factorizes in a directed graphical model also does. This is true for the marginalization with respect to leaf nodes but is not true generally.

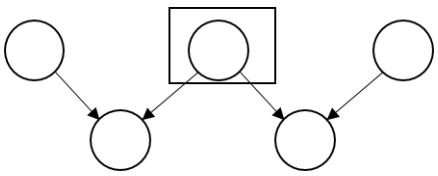


Figure 4.6. Marginalizing the boxed node would not result in family of distributions that can be exactly represented by a directed graphical model

Conditional independence. We finish this section by giving a result that explains that if $p(x)$ factorizes in G then every single random variable is independent from the set of its non-descendants given its parents.

Definition 4.10 *We define t e set of non-descendents of i by $nd(i) \triangleq \{j : \text{no pat from } i \text{ to } j\}$.*

Proposition 4.11 *If G is a DAG, t en:*

$$p(x) \in \mathcal{L}(G) \Leftrightarrow X_i \perp\!\!\!\perp X_{nd(i)} | X_{\pi_i} \tag{4.15}$$

Proof First, we consider the $\Rightarrow$ direction. Let i be fixed. The key point is that $\exists$ a topological sort such that $nd(i)$ are just before i - i.e. $(nd(i), i, V \backslash (\{i\} \cup nd(i)))$.
First we consider $\Rightarrow$ $x_i \perp\!\!\!\perp x_{nd(i)} | x_{\pi_i}$. Let $A \triangleq nd(i) \backslash \pi_i$. We want to marginalize out $x_{V \backslash (\{i\} \backslash nd(i))}$ (plucking leaves). By re-using the same argument as in the proof of Proposition 4.3, we have that $p(x_{1:i}) = p(x_i, x_{\pi_i}, x_A) = \prod_{j \leq i} p(x_j | x_{\pi_j})$. Thus:

$$p(x_i, x_{\pi_i}, x_A) = p(x_i | x_{\pi_i}) \prod_{j \in nd(i)} p(x_j | x_{\pi_j})$$

$$p(x_i | x_{nd(i)}) = \frac{p(x_i, x_{\pi_i}, x_A)}{p(x_{\pi_i}, x_A)}$$

$$p(x_i | x_{nd(i)}) = \frac{p(x_i | x_{\pi_i}) \prod_{j \in nd(i)} p(x_j | x_{\pi_j})}{\sum_{x_i'} [p(x_i' | x_{\pi_i})] \prod_{j \in nd(i)} p(x_j | x_{\pi_j})}$$

But since $\sum_{x_i'} [p(x_i' | x_{\pi_i})] = 1$ and the product terms cancel out, we get

$$p(x_i | x_{nd(i)}) = p(x_i | x_{\pi_i}),$$

and thus $X_i \perp\!\!\!\perp X_{\mathrm{nd}(i)} | X_{\pi_i}$, as we wanted to show.

Now we consider the other direction. Let $1 : n$ be a topological sort. Then $\{1, \cdots i-1\} \subseteq nd(i)$. (By contradiction, suppose $j \in \{1 \cdots i-1\}$ and $j \notin nd(i)$, then $\exists$ path from i to j, which contradicts the topological sort property as there would be an edge from i to an element of $\{1, \ldots, i-1\}$.)
By the chain rule, we have (always true):

$$p(x_V) = \prod_{i=1}^{n} p(x_i | x_{1:i-1})$$

By the conditional independence assumptions:

$$p(x_V) = \prod_{i=1}^{n} p(x_i | x_{\pi_i})$$

and thus $p(x_V) \in \mathcal{L}(G)$ as we wanted to prove. ■

4.2.4 d-separation

We want to answer queries such as, given A, B and C, three subsets, is $X_A \perp\!\!\!\perp X_B | X_c$ true? To answer those issues we need the d-separation notion, or directed separation. Indeed it is easy to see that the notion of separation is not enough in a directed graph and needs to be generalized.

Definition 4.12 *Let $a, b \in V$, a c ain from a to b is a sequence of nodes, say $(v_1, \ldots, v_n)$ suc t at $v_1 = a$ and $v_n = b$ and $\forall j, (v_j, v_{j+1}) \in E$ or $(v_{j+1}, v_j) \in E$.*

We can notice that a chain is hence a path in the symmetrized graph, *i.e.* in the graph where if the relation $\rightarrow$ is true then $\leftrightarrow$ is true as well. Assume C is a set that is observed. We want to define a notion of being 'blocked' by this set C in order to answer the underlying question above.

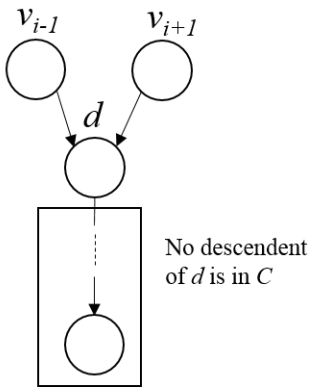


Figure 4.7. D-separation

Definition 4.13

1. *A c ain from a et b is blocked at d if:*
   - *eit er $d \in C$ and (v_{i-1}, d, v_{i+1}) is not a V-structure;*
   - *or $d \notin C$ and (v_{i-1}, d, v_{i+1}) is a V-structure and no descendants of d is in C.*
2. *A c ain from a to b is blocked if and only if it is blocked at any nodes.*
3. *A and B are said to be d-separated by C if and only if all c ains t at go from $a \in A$ to $b \in B$ are blocked.*

Example 4.2.2

- ***Markov chain:*** *If you try to prove t at any set of t e future is independent to t e past given t e present wit Markov t eory, it mig t be difficult but t e d-separation notion gives t e results directly.*

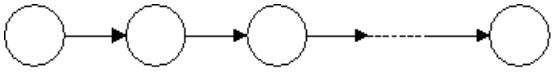


Figure 4.8. Markov chain

- ***Hidden Markov Model:*** *Often used because we only ave a noisy observation of t e random process.*

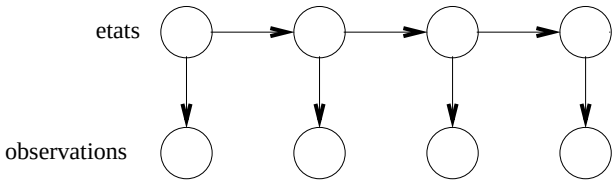


Figure 4.9. Hidden Markov Model

4.2.5 Bayes ball algorithm

This is an intuitive "reacheability" algorithm to determine conditional independence in a DAG. It is a systematic algorithm to check whether two nodes are d-separated. Suppose we want to determine if X is conditionally independent from Z given Y. Place a ball on each of the nodes in X and let them bounce around according to some rules (described below) and see if any reaches Z. $X \perp\!\!\!\perp Z|Y$ is true if none reached Z, but not otherwise.

The rules are as follows for the three canonical graph structures. Note that the balls are allowed to travel in either direction along the edges of the graph.

1. **Markov chain:** Balls pass through when we do not observe Y, but are blocked otherwise.

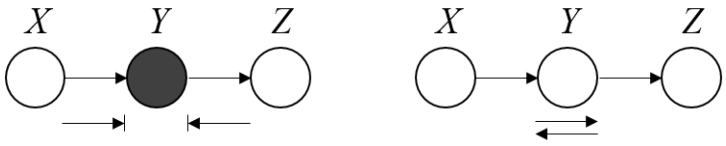


Figure 4.10. Markov chain rule: When Y is observed, balls are blocked (left). When Y is not observed, balls pass through (right)

2. **Two children:** Balls pass through when we do not observe Y, but are blocked otherwise.

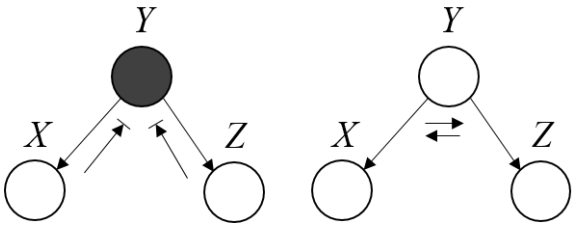


Figure 4.11. Rule when X and Z are Y's children: When Y is observed, balls are blocked (left). When Y is not observed, balls pass through (right)

3. **V-structure:** Balls pass through when we observe Y, but are blocked otherwise.

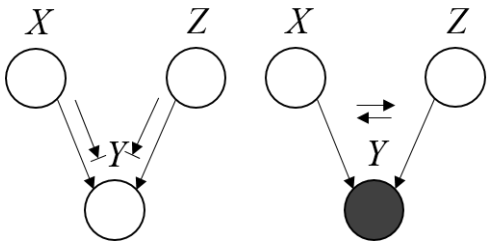


Figure 4.12. V-structure rule: When Y is not observed, balls are blocked (left). When Y is observed, balls pass through (right)

4.3 Undirected graphical models

4.3.1 Definition

Definition 4.14 *Let $G = (V, E)$ be a* **undirected graph**. *We denote by $\mathcal{C}$ a set of cliques of G i.e.a set of sets of fully connected vertices. We say t at a probability distribution p factorizes in G and denote $p \in \mathcal{L}(G)$ if $p(x)$ is of t e form:*

$$p(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C) \text{ wit } \psi_C \geq 0, Z = \sum_x \prod_{C \in \mathcal{C}} \psi_C(x_C).$$

The functions ψ_C are not probability distributions like in the directed graphical models. They are called potentials.

Remark 4.3.1 *Wit t e normalization by Z of t is expression, we see t at t e function ψ_C are defined up to a multiplicative constant.*

Remark 4.3.2 *We may restrict $\mathcal{C}$ to $\mathcal{C}_{max}$, t e set of maximal cliques.*

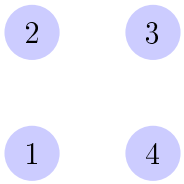
Remark 4.3.3 *T is definition can be extended to any function: f is said to factorize in $G \iff f(x) = \prod_{C \in \mathcal{C}} \psi_C(x_C)$.*

4.3.2 Trivial graphs

Empty graphs We consider $G = (V, E)$ with $E = \emptyset$. For $p \in \mathcal{L}(G)$, we get:

$$p(x) = \prod_{i=1}^{n} \psi_i(x_i) \text{ as } \mathcal{C} = \{\{i\} \in V\}$$

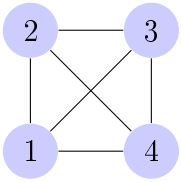
This gives us that $X_1, ..., X_n$ are mutually independent.



Complete graphs We consider $G = (V, E)$ with $\forall i, j \in V, (i, j) \in E$. For $p \in \mathcal{L}(G)$, we get:

$$p(x) = \frac{1}{Z}\psi_V(x_V) \text{ as } \mathcal{C} \text{ is reduced to a single set V}$$

This gives no further information upon the n-sample $X_1, ..., X_n$.



4.3.3 Separation and conditional dependence

Proposition 4.15 *Let $G = (V, E)$ and $G' = (V, E')$ be two undirected grap s.*

$$E \subseteq E' \Rightarrow \mathcal{L}(G) \subseteq \mathcal{L}(G')$$

Definition 4.16 *We say t at p satisfies t e* **Global Markov property** *w.r.t. G if and only if for all $A, B, S \subset V$ disjoint subsets: A and B are separated by $S \Rightarrow X_A \perp\!\!\!\perp X_B | X_S$.*

Proposition 4.17 *If $p \in \mathcal{L}(G)$ t en, p satisfies t e Global Markov property w.r.t. G.*

Proof We suppose without loss of generality that A, B, and S are disjoint sets such that $A \cup B \cup S = V$, as we could otherwise replace A and B by :

$$A' = A \cup \{a \in V / a \text{ and } A \text{ are not separated by } S\}$$
$$B' = V \setminus \{S \cup A'\}$$

A' and B' are separated by S and we have the disjoint union $A' \cup B' \cup S = V$. If we can show that $X_{A^0} \perp\!\!\!\perp X_{B^0} | X_S$, then by the decomposition property, we also have that $X_A \perp\!\!\!\perp X_B | X_S$ for any subset A of A' and B of B', giving the required general case.

We consider $C \in \mathcal{C}$. It is not possible to have both $C \cap A \neq \emptyset$ and $C \cap B \neq \emptyset$ as A and B are separated by S and C is a clique. Thus $C \subset A \cup S$ or $C \subset B \cup S$ (or both if $C \subset S$). Let $\mathcal{D}$ be the set of cliques C such that $C \subset A \cup S$ and $\mathcal{D}'$ the set of all other cliques. We have:

$$p(x) = \frac{1}{Z} \prod_{\substack{C \in \mathcal{C} \\ C \not\subset A \cup S}} \psi_C(x_C) \prod_{C \in \mathcal{D}^0} \psi_C(x_C) = f(x_{A \cup S}) g(x_{B \cup S}).$$

Thus:

$$p(x_A, x_S) = \frac{1}{Z} f(x_A, x_S) \sum_{x_B} g(x_B, x_S) \implies p(x_A | x_S) = \frac{f(x_A, x_S)}{\sum_{x'_A} f(x'_A, x_S)}.$$

Similarly: $p(x_B | x_S) = \frac{g(x_B, x_S)}{\sum_{x'_B} g(x'_A, x_S)}$. Hence:

$$p(x_A, x_S) p(x_B | x_S) = \frac{\frac{1}{Z} f(x_A, x_S) g(x_B, x_S)}{\frac{1}{Z} \sum_{x'_A} f(x'_A, x_S) \sum_{x'_B} g(x'_A, x_S)} = \frac{p(x_A, x_B, x_S)}{p(x_S)} = p(x_A, x_B | x_S).$$

i.e. $X_A \perp\!\!\!\perp X_B | X_S$. ■

Theorem 4.18 *(Hammersley - Clifford) If* $\forall x,\ p(x) > 0$ *then* $p \in \mathcal{L}(G) \iff p$ *satisfies the global Markov property.*

4.3.4 Marginalization

As for directed graphical models, we also have a marginalization notion in undirected graphs. It is slightly different. If $p(x)$ factorizes in G, then $p(x_1, \ldots, x_{n-1})$ factorizes in the graph where the node n is removed and all neighbors are connected.

Proposition 4.19 *Let* $G = (V, E)$ *be an undirected graph. Let* $G' = (V', E')$ *be the graph where* n *is removed and its neighbors are connected, i.e.* $V' = V \setminus \{n\}$*, and* E' *is obtained from the set* E *by first connecting together all the neighbours of* n *and then removing* n*. If* $p \in \mathcal{L}(G)$ *then* $p(x_1, ..., x_{n-1}) \in \mathcal{L}(G')$*. Hence undirected graphical models are closed under marginalization as the construction above is true for any vertex.*

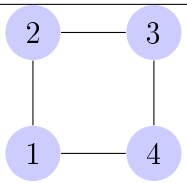
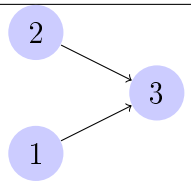
We now introduce the notion of Markov blanket

Definition 4.20 *For* $i \in V$*, the* **Markov blanket** *of a graph* G *is the smallest set of nodes that makes* X_i *independent to the rest of the graph.*

Remark 4.3.4 *The Markov blanket in an undirected graph for* $i \in V$ *is the set of its neighbors. For a directed graph, it is the union of all parents, all children and parents of children.*

4.3.5 Relation between directed and undirected graphical models

Since now we have seen that many notions developed for directed graph naturally extended to undirected graphs. The raising question is thus to know whether we can find a theory including both directed and undirected graphs, in particular, is there a way—for instance by symmetrizing the directed graph as we have done repeatedly—to find a general equivalence between those two notions. The answer is no, as we will discuss—though it might work in some special cases described above.

	Directed graphical model	Undirected graphical model
Factorization	$p(x) = \prod_{i=1}^{n} p(x_i \mid x_{\pi_i})$	$p(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C)$
Set independence	d-separation $[x_i \perp\!\!\!\perp x_{nd(i)} \mid x_{\pi_i}]$ (and many more)	separation $[X_A \perp\!\!\!\perp X_B \mid X_S]$
Marginalization	not closed in general, only when marginalizing leaf nodes	closed
Difference	grid 	v-structure 

Let G be DAG. Can we find G' undirected such that $\mathcal{L}(G) = \mathcal{L}(G')$? $\mathcal{L}(G) \subset \mathcal{L}(G')$?

Definition 4.21 *Let $G = (V, E)$ be a* DAG. *The* **symmetrized graph** *of G is $\tilde{G} = (V, \tilde{E})$, with $\tilde{E} = \{(u, v), (v, u) / (u, v) \in E\}$, ie. an edge going the opposite direction is added for every edge in E.*

Definition 4.22 *Let $G = (V, E)$ be a* DAG. *The* **moralized graph** *$\bar{G}$ of G is the symmetrized graph $\tilde{G}$, where we add edge such that for all $v \in V$, π_v is a clique.*

We admit the following proposition:

Proposition 4.23 *Let G be a* DAG *without any V-structure, then $\bar{G} = \tilde{G}$ and $\mathcal{L}(G) = \mathcal{L}(\tilde{G}) = \mathcal{L}(\bar{G})$.*

In case there is a V-structure in the graph, we can only conclude:

Proposition 4.24 *Let G be a* DAG, *then $\mathcal{L}(G) \subset \mathcal{L}(\bar{G})$.*

$\bar{G}$ is minimal for the number of edges in the set H of undirected graphs such that $\mathcal{L}(G) \subset \mathcal{L}(H)$.

Not all conditional independence structure for random variables can be factorized in a graphical model (directed or undirected).