Chapter 4

Markov Properties

In Chapter ? we highlighted the central role played by conditional independence in probabilistic modelling. We also introduced the concept of a directed graph and related the *absence* of links in the graph to statements of conditional independence of a corresponding probability distribution. We shall refer to these conditional independence statements as Markov properties, since they can be viewed as a generalization of the conditional independence property of a Markov chain discussed in Chapter ?. In this chapter we undertake a more comprehensive and systematic study of conditional independence and its graphical representation.

In Section 4.1 we introduce and prove a number of useful results relating to conditional independence. Equipped with this foundation we then explore the conditional independence properties of directed graphs in Section 4.2. In particular we introduce the concept of *d-separation* which allows conditional independences to be read directly from the graph. Since all of the conditional independence properties implied by the graph can be found using d-separation, it captures the *global* Markov properties of the graph. We then demonstrate the equivalence of d-separation to the factorization property for directed graphs introduced in Chapater ?.

Next we introduce the concept of *undirected* graphs in Section 4.3. These have a different semantics from that of the directed acyclic graphs discussed so far in the book. As we shall see in Chapter ?, undirected graphs play a crucial role in solving inference and learning problems efficiently, even for models whose definition is based originally on a directed graph. We discuss the factorization property for undirected graphs and relate this to the graph's global Markov properties.

In Section 4.4 we return to directed graphs for a further exploration of their conditional independence properties. We use the machinery of undirected graphs developed in Section 4.3 to develop an alternative formulation of the global Markov properties of directed graphs. Also we introduce *local* and *pairwise* Markov properties and investigate their relation to global Markov properties and to factorization.

In Section 4.5 we complete our discussion of the Markov properties of undirected graphs by introducing definitions of pairwise and local Markov, and relating these to global Markov properties and to factorization.

Finally, in Section 4.6 we outline some extensions of the graphical formalism, and discuss its limitations for representing conditional independencies.

Concepts from graph theory will be introduced in this chapter as they are needed. However, the key concepts are also summarized in Appendix A.

4.1 Conditional Independence

In discussing conditional independence properties, it is useful to follow Dawid's notation in which $A \perp\!\!\!\perp B \mid C$ denotes that A is independent of B given C. Here A, B and C represent variables, or more generally groups of variables, with some joint distribution $P(A, B, C)$. The conditional independence property $A \perp\!\!\!\perp B \mid C$ is simply a statement that the conditional distribution $P(A|B, C)$ satisfies

$$P(A|B, C) = P(A|C) \tag{4.1}$$

so that, once the value of C has been fixed, subsequently learning the value of B tells us nothing further about the distribution of A. It should be emphasized that the property (4.1) must hold for every possible instantiation of C.

If we multiply both sides of (4.1) by $P(B|C)$ we obtain an equivalent statement for the distribution $P(A, B|C)$ in the form

$$P(A, B|C) = P(A|C)P(B|C) \tag{4.2}$$

so that, once C is known, the joint conditional distribution of A and B factorizes into the product of the marginal conditional distributions.

A special case of conditional independence arises when there are no conditioning variables so that if, for example, A is marginally independent of B, which we write as $A \perp\!\!\!\perp B \mid \emptyset$, then $P(A|B) = P(A)$, or equivalently $P(A,B) = P(A)P(B)$.

While some conditional independence relations hold universally, others apply only to a restricted class of distributions. An important such class consists of those distributions which are strictly positive, in other words where every possible instantiation of the variables has a non-zero probability. Distributions which have zeros, and hence are not strictly positive, can arise if there are logical relations between the variables. For instance if $A = B$ then $P(A \neq B) = 0$.

Using the sum and product rules of probability we can easily prove the following four properties of conditional independence.

Theorem 4.1 (Symmetry) *If A is conditionally independent of B given C, then B is conditionally independent of A given C, so that*

$$A \perp\!\!\!\perp B \mid C \quad \Leftrightarrow \quad B \perp\!\!\!\perp A \mid C. \tag{4.3}$$

Proof: If $A \perp\!\!\!\perp B \mid C$ then $P(A|B,C) = P(A|C)$. Hence

$$P(B|A,C) = \frac{P(A,B|C)}{P(A|C)} = \frac{P(A|B,C)P(B|C)}{P(A|C)} = P(B|C) \tag{4.4}$$

and so $B \perp\!\!\!\perp A \mid C$. □

Theorem 4.2 (Decomposition)

$$A \perp\!\!\!\perp (B \cup D) \mid C \quad \Rightarrow \quad A \perp\!\!\!\perp B \mid C \quad \text{and} \quad A \perp\!\!\!\perp D \mid C. \tag{4.5}$$

Proof: Using $A \perp\!\!\!\perp (B \cup D) \mid C$ we have $P(A|B,C,D) = P(A|C)$, and hence

$$\begin{aligned} P(A|B,C,D) &= \sum_D P(A,D|B,C) \\ &= \sum_D P(A|B,C,D)P(D|B,C) \\ &= P(A|C)\sum_D P(D|B,C) \\ &= P(A|C) \end{aligned} \tag{4.6}$$

□

and hence $A \perp\!\!\!\perp B \mid C$. Similarly we can show $A \perp\!\!\!\perp D \mid C$.

Theorem 4.3 (Weak Union)

$$A \perp\!\!\!\perp (B \cup D) \mid C \quad \Rightarrow \quad A \perp\!\!\!\perp B \mid C \cup D \quad \text{and} \quad A \perp\!\!\!\perp D \mid C \cup B. \tag{4.7}$$

Proof: From $A \perp\!\!\!\perp (B \cup D) \mid C$ we have $P(A|B,C,D) = P(A|C)$ and from the decomposition property Theorem 4.2 we have $P(A|C,D) = P(A|C)$. Thus $P(A|B,C,D) = P(A|C,D)$ and hence $A \perp\!\!\!\perp B \mid C \cup D$. An analogous argument is used to show $A \perp\!\!\!\perp D \mid C \cup B$. □

Theorem 4.4 (Contraction)

$$A \perp\!\!\!\perp B \mid (C \cup D) \quad \text{and} \quad A \perp\!\!\!\perp D \mid C \quad \Rightarrow \quad A \perp\!\!\!\perp (B \cup D) \mid C. \tag{4.8}$$

Proof: From $A \perp\!\!\!\perp B \mid (C \cup D)$ we have $P(A|B,C,D) = P(A|C,D)$. Similarly from $A \perp\!\!\!\perp D \mid C$ we have $P(A|C,D) = P(A|C)$. Combining these we obtain $P(A|B,C,D) = P(A|C)$, and hence $A \perp\!\!\!\perp (B \cup D) \mid C$. □

The following property does not hold universally. It does hold, however, for distributions which are strictly positive.

Theorem 4.5 (Intersection) *For strictly positive distributions*

$$A \perp\!\!\!\perp B \mid (C \cup D) \quad \text{and} \quad A \perp\!\!\!\perp C \mid (B \cup D) \quad \Rightarrow \quad A \perp\!\!\!\perp (B \cup C) \mid D. \tag{4.9}$$

Proof: Using $A \perp\!\!\!\perp B \mid (C \cup D)$ we have

$$P(A, B, C|D) = P(A|C, D)P(B, C|D). \tag{4.10}$$

Similarly using $A \perp\!\!\!\perp C \mid (B \cup D)$ we have

$$P(A, B, C|D) = P(A|B, D)P(B, C|D). \tag{4.11}$$

Equating these two we obtain

$$P(A|C, D) = P(A|B, D) \tag{4.12}$$

where we have assumed $P(B, C|D) \neq 0$. From this it follows that the marginal distribution of A is given by

$$\begin{aligned} P(A|D) &= \sum_B P(A|B, D)P(B|D) \\ &= \sum_B P(A|C, D)P(B|D) \\ &= P(A|C, D). \end{aligned} \tag{4.13}$$

Using (4.10) we then have

$$P(A, B, C|D) = P(A|C, D)P(B, C|D) = P(A|D)P(B, C|D) \tag{4.14}$$

and hence $A \perp\!\!\!\perp (B \cup C) \mid D$. $\square$

4.2 Directed Graphs

We begin our study of Markov properties by reviewing the factorization property for directed graphs, discussed already in Chapter ?. Next we introduce the concept of d-separation, which allows the conditional independence properties to be read directly from the graph. We then show the equivalence of d-separation and factorization, and also show the equivalence to the Bayes' Ball procedure of Chapter ?.

4.2.1 Directed Factorization

We have already observed in Chapter ? that, given some ordering of the variables, a general joint distribution can be factorized into a product of conditional distributions, one for each variable. For example, given M variables $X_1, X_2, \ldots, X_M$ we have

$$P(X_1, X_2, X_3, \ldots, X_M) = P(X_1)P(X_2|X_1) \\ P(X_3|X_1, X_2) \cdots P(X_M|X_1, X_2, \ldots, X_{M-1}). \tag{4.15}$$

We can represent this factorization graphically, as shown for the case of $M = 4$ in Figure 4.1, in which each node has all lower-numbered nodes as its parent set.

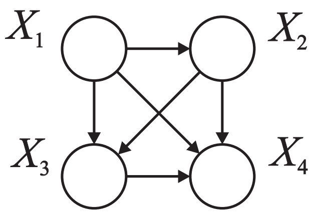


Figure 4.1: A fully connected directed graph comprising four nodes.

While this holds true for arbitrary distributions, we can define a restricted family of distributions by introducting a factorization with respect to a given directed acyclic graph in the form

$$P(S) = \prod_{i \in S} P(S_i|\mathrm{pa}(S_i)) \tag{4.16}$$

where $\mathrm{pa}(S_i)$ denotes the set of parents of S_i in the graph. A directed graph is acyclic if (and only if) the nodes can be numbered such that for every node all the parents of that node have a lower number than the node itself. For example, given the graph in Figure 4.2 we have the following factorization

$$\begin{aligned} P(X_1, X_2, X_3, X_4, X_5) &= P(X_1)P(X_2)P(X_3|X_1, X_2) \\ &\qquad P(X_4|X_2)P(X_5|X_3, X_4). \end{aligned} \tag{4.17}$$

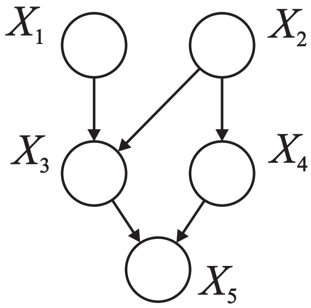


Figure 4.2: A directed graph comprising five nodes.

Comparing (4.15) with (4.17) we see that the latter has variables missing from some the conditioning sets of the conditional distributions, corresponding to missing links in the directed graph of Figure 4.2. This implies that any distribution which factorizes according to (4.17) will exhibit some conditional independence properties. A central goal in this section is to uncover these conditional independencies and to relate them quantitatively to the factorization property.

In discussing Markov properties it may be helpful to regard the directed acyclic graph as a filter. A graph with M nodes defines a particular factorization of the joint distribution according to (4.16). Any given distribution over M variables will only pass through the filter if it can be expressed in terms of the corresponding factorization. Thus the graph defines a family of distributions, namely the set of all distributions over M variables which can be expressed in the form (4.16). We denote this factorization property with respect to a directed graph by $\mathcal{DF}$ (for 'directed factorization').

4.2.2 d-separation

In order to uncover the conditional independencies implied by a given graph we consider some simple 3-node graphs, of the kind already discussed in Chapter ?. First we define a *path* from a node a to a node b as a sequence of nodes starting with a and ending with b such that successive nodes are connected by a link. Note that a path may involve traversing directed links in either direction. We will use the notion of a path to introduce the concept of nodes which are *head-to-head*, *head-to-tail* or *tail-to-tail* with respect to a particular path through the graph. From this we obtain the framework of *d-speparation* which allows all of the conditional independencies implied by the graph to be obtained from the graph itself.

We begin by considering the directed graph shown in Figure 4.3 which involves 3 nodes A, B, and C in which we have conditioned on C. As we

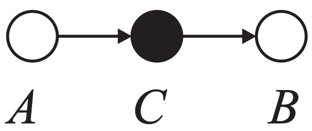


Figure 4.3: A directed graph of three nodes for which $A \perp\!\!\!\perp B \mid C$.

have already seen in Chapter ? that this graph satisfies the independence property $A \perp\!\!\!\perp B \mid C$. Converseley, in general A and C are not independent if C is not observed, so that $A \not\perp\!\!\!\perp B \mid \emptyset$. The node C is said to be *head-to-tail* with respect to the path A–C–B since the arrow on one of the links points towards node C while the arrow on the other link points away from C. Conditioning on the node C is said to *block* the path from A to B.

Similarly, consider the graph in Figure 4.4. As we discussed in Chap-

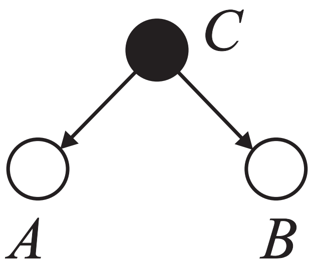


Figure 4.4: Another directed graph for which $A \perp\!\!\!\perp B \mid C$.

ter ?, the graph satisfies the conditional independence property $A \perp\!\!\!\perp B \mid C$. Similarly, as before, $A \not\perp\!\!\!\perp B \mid \emptyset$. The node C is called a *tail-to-tail* node with respect to the path A–C–B since both arrows point away from C. Again, we have the notion that conditioning on C has blocked the path from A to B, and rendered A and B conditionally independent.

Finally we consider the graph of Figure 4.5. In this case we have quite

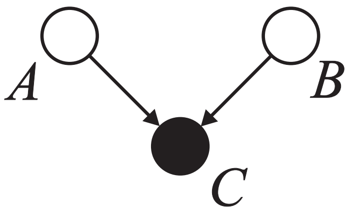


Figure 4.5: A directed graph which does not imply $A \perp\!\!\!\perp B \mid C$.

different conditional independence properties from the two previous examples since $A \perp\!\!\!\perp B \mid \emptyset$ whereas $A \not\perp\!\!\!\perp B \mid C$. Node C is called *head-to-head* with respect to the path A–C–B since both arrows point towards the node. We see that a head-to-head node has the opposite behaviour from a head-to-tail or tail-to-tail node. The path is blocked if C is unobserved, so that A and B are marginally independent. However, the node becomes unblocked if we condition on C.

It should be emphasised that, in more complex graphs, a particular node can, for example, be tail-to-tail with respect to one path through the node and head-to-tail or head-to-head with respect to a different path.

Given the above observations we might suspect that more general conditional independent statements can be read directly from a graph by considering paths through the graph and oberving whether the paths are blocked or not. This leads to the concept of *d-separation*.

We wish to ascertain whether a particular conditional independence statement $A \perp\!\!\!\perp B \mid C$ is implied by a given directed acyclic graph, where A, B and C are non-intersecting sets of nodes. A path is said to be *blocked* if it includes a node such that either

(a) the arrows on the path do not meet head-to-head at the node, and the node is in the conditioning set, or

(b) the arrows do meet head-to-head, and neither the node, nor any of its descendents, is in the conditioning set.

Given a set of conditioning nodes C, if every path from any node in a set A to any node in a set B is blocked, then A is said to be d-separated from B by

C. As we will see, this implies that the distribution will satisfy $A \perp\!\!\!\perp B \mid C$. The concept of d-separation is illustrated in Figure 4.6.

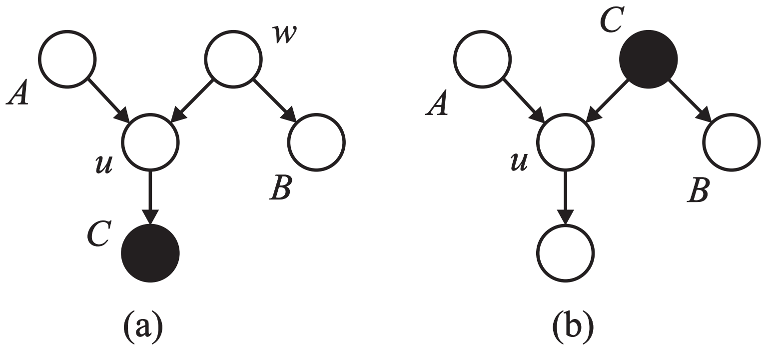


Figure 4.6: Illustration of d-separation. In graph (a) the path from A to B is not blocked by node w since it is not a head-to-head node and is not observed, nor is it blocked by node u since, although the latter is a head-to-head node, it has a descendent C which is in the conditioning set. Thus the conditional independence statement $A \perp\!\!\!\perp B \mid C$ does not follow from this graph. In graph (b) the path from A to B is blocked by node C since this is a tail-to-tail node which is observed, and is also blocked by node u since u is a head-to-head node and neither it nor its descendents are in the conditioning set, and so $A \perp\!\!\!\perp B \mid C$ will be satisfied by any distribution which factorizes according to this graph.

We can regard the d-separation property as a filter, so that, for a particular directed acyclic graph we can test whether any given joint distribution satisfies all of the conditional independence properties implied by the graph through d-separation. Clearly if the graph is fully connected (each node has all lower-numbered nodes in its parent set) then all possible joint distributions will pass the filter. Conversely if there are no links at all between nodes of the graph then only distributions which factorize completely into the product of the marginal distributions over each variable will pass the filter.

The three example graphs considered so far clearly have the property that their conditional independence properties can be determined by the d-separation criterion. We now show that this is a general result for arbitrary directed acyclic graphs, in other words if a distribution factorizes according to a particular graph, then any conditional independence property which follows from the graph will be satisfied by the distribution.

Theorem 4.6 *($\mathcal{DF} \Rightarrow$ d-separation). If a joint distribution factorizes according to a directed acyclic graph, and if A, B and C are disjoint subsets of nodes such that C d-separates A from B in the graph, then the distribution satisfies $A \perp\!\!\!\perp B \mid C$.*

Proof: We prove this result by induction on the number N of nodes in the graph. First we observe that the result holds trivially for a graph with one node. Next we assume that it holds for all DAGs with $N-1$ nodes and then prove the result for DAGs with N nodes. Let the nodes in a DAG $\mathcal{D}$ of size N be given a topological ordering and let the highest numbered node be ω (hence ω has no children). We denote by $\mathcal{D}'$ the DAG $\mathcal{D}$ with ω removed, and observe that the distribution over the reduced set of variables also factorizes with respect to $\mathcal{D}'$. Now consider three disjoint subsets of nodes A, B and C such that C d-separates A from B. The following three possibilities are exhaustive:

(a) The final vertex $\omega \notin A \cup B \cup C$. Since all paths from A to B are, by assumption, blocked it must be true that C d-separates A from B in $\mathcal{D}'$, and hence $A \perp\!\!\!\perp B \mid C$.

(b) The vertex ω is a member of A (or equivalently B). Denote $A' = A \setminus \omega$, and note that A' will be d-separated by C from B in $\mathcal{D}'$. Since there can be no direct links from A to B it follows that none of the parents of ω are in B. Let P be the set of parents of ω which are not in C. Then P is also d-separated from B by C in $\mathcal{D}'$ since any path from P to B either goes through ω (and hence is blocked since ω must be a head-to-head node for such a path and ω is not part of the conditioning set C) or the path can be extended by one step to become a path from ω to B (and such a path must be blocked by the assumption of d-separation in $\mathcal{D}$ but it is not blocked at the parent of ω in P since this is an unobserved node and is not head-to-head and so the path must be blocked elsewhere). Thus both A' and P are d-separated from B by C in $\mathcal{D}'$ and hence $A' \cup P$ is d-separated from B by C in $\mathcal{D}'$ and so $A' \cup P \perp\!\!\!\perp B \mid C$ in $\mathcal{D}'$. Since the addition of the blocking node ω cannot create new unblocked paths it follows that $A' \cup P \perp\!\!\!\perp B \mid C$ in $\mathcal{D}$. We also have $\omega \perp\!\!\!\perp B \mid A' \cup C \cup P$ since the conditioning set includes all of the parents of ω. Using the result (4.4) we then have $A \cup P \perp\!\!\!\perp B \mid C$ and hence $A \perp\!\!\!\perp B \mid C$.

(c) Finally, we consider $\omega \in C$. Note that no path can be blocked at ω, and hence A and B must be d-separated by $C' = C \setminus \omega$. Also, ω must be d-separated from A or B (or both) by C' otherwise there would be an unblocked path from A to B via ω. Suppose this holds for B so that $A \cup \omega$ is d-separated from B by C'. Using the result from case **(b)** above we then have $A \cup \omega \perp\!\!\!\perp B \mid C'$. Using (4.3) we then have $A \perp\!\!\!\perp B \mid C' \cup \omega$ and hence $A \perp\!\!\!\perp B \mid C$.

□

The converse of this theorem also holds, namely that if a joint distribution satisfies all of the conditional independence properties which can be read from a graph by d-separation then the distribution must also factorize according to that graph using (4.16). We give a formal proof of this result in Section 4.4.

Of course, not all conditional independencies present in the distribution can necessarily be determined from the graph by d-separation since there may be aditional independencies arising from the specific numerical values associated with the conditional distributions. However, it is possible to construct an example model which is such that any conditional independencies which do not correspond to d-separation are not present in the distribution, showing that in general d-separation will find all of the conditional independences which can be determined directly from the DAG.

[here we return to the Bayes' Ball algorithm introduced in Chapter ?, and demonstrate its equivalence to d-separation]

In some probability distributions there may be deterministic relationships between the variables. In this case there is probability one that the relationship occurs, and combinations of variable values which do not respect the relationship have probability zero. When the conditional distribution of a node conditioned on its parents depends deterministically on one of its parents, the corresponding edge in the directed acyclic graph is sometimes denoted by a double arrow '$\Longrightarrow$' connecting the nodes.

Since Theorem 4.6 does not require positivity of the joint distribution, we can use d-separation to read off conditional independence properties even when deterministic relations are present. However, we can extact additional independencies using D-separation, which is simply an extension of d-separation in which, if a variable is observed, then all other variables which are deterministically related to that variable are also considered to be observed. An example is shown in Figure 4.7

4.3 Undirected Graphs

So far in this book we have discussed models based on directed acyclic graphs. However, to begin our discussion of Markov properties it is con-

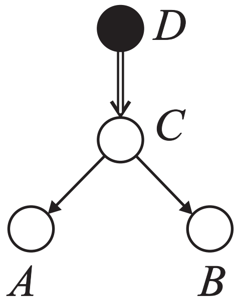


Figure 4.7: An illustration of D-separation showing a directed acyclic graph over four variables in which C depends deterministically on D. The conditional independence statement $A \perp\!\!\!\perp B \mid D$ does not follow from the standard d-separation criterion. In the D-separation criterion, however, we also infer that C is an observed node and hence that $A \perp\!\!\!\perp B \mid D$.

venient to consider undirected graphs. Not only do they have simpler semantics than directed graphs, but they will prove useful in formulating and understanding the Markov properties of directed graphs.

We first introduce the concept of graph separation on an undirected graph as follows. Given three disjoint subsets A, B, and C of nodes on the graph, we say that C separates A and B if every *path* on the graph from any node in A to any node in B passes through at least one node in C. This is illustrated in Figure 4.8.

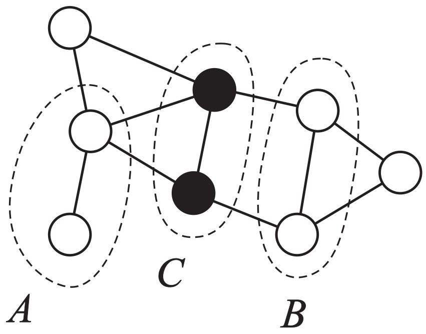


Figure 4.8: An illustration of the concept of graph separation in undirected graphs. The set C separates A from B since every path from a node in A to a node in B must pass through a node in C.

As for directed graphs we consider the global Markov properties for undirected graphs, and then relate these to a factorization property of the joint distribution.

4.3.1 Global Markov

First we introduce the *global Markov* property for undirected graphs, which we denote by $\mathcal{G}$. A probability distribution is said to be global Markov with respect to a graph if, for any disjoint subsets of nodes A, B, C such that C separates A and B on the graph, the distribution satisfies $A \perp\!\!\!\perp B \mid C$.

Again, we can think of the graph as a kind of filter. Consider the set of all possible distributions over the variables corresponding to the nodes of the graph. We could test each distribution in the set to see if it exhibits the conditional independencies implied by the graph. Those probability distributions which pass the test form a family of distributions, every member of which is global Markov with respect to the given graph.

Note that any distribution which factorizes with respect to all of the variables will always be global Markov for any graph since any possible conditional independence statement will always be statisfied. Conversely, if we consider a graph which is fully connected then any distribution will be global Markov for this graph since the graph implies no conditional independence statements.

4.3.2 Factorization

We saw in Chapter ? how a probability distribution could be constructed from a product of conditional distributions defined with respect to a directed graph. Here we introduce the corresponding factorization property for undirected graphs, which we denote by $\mathcal{F}$. First we define a set of nodes to be *complete* if there is a link from each node to every other node in the set. A probability distribution is said to factorize with respect to a given undirected graph if it can be expressed as the product of functions over the complete sets of nodes of the graph

$$P(S) = \prod_{a \text{ complete}} \psi_a(S_a) \tag{4.18}$$

where the functions $\psi_a(S_a)$ are known as potentials.

An alternative, equivalent formulation of undirected factorization can be obtained by introducing the notion of a *clique*, which is a complete set of nodes which is also maximal, so that inclusion of any other node in the set would render it incomplete. The concept of a clique is illustrated in Figure 4.9.

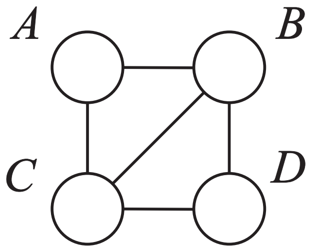


Figure 4.9: Illustration of the concept of a clique. In this graph there are two cliques comprising the sets $\{A, B, C\}$ and $\{B, C, D\}$.

Suppose we define the potential function for each clique in the graph to be initially unity. Then we take each of the potential functions $\psi_a(S_a)$ in (4.18) and multiply it into the corresponding clique potential. We then obtain a representation of the form

$$P(S) = \prod_C \psi_C(S_C) \tag{4.19}$$

where the product is over all cliques in the graph. Again, we can think of the graph as a filter in which only joint distributions which can be expressed in the form (4.19) will be accepted.

We next show that, for any graph and any distribution, the factorization property implies the global Markov property.

Theorem 4.7 *($\mathcal{F} \Rightarrow \mathcal{G}$) For any undirected graph, any distribution which satisfies the factorization property with respect to the graph will also respect the global Markov properties of the graph.*

Proof: Consider three disjoint subsets of nodes A, B and C such that C separates A from B in the graph. Let V denote the set of all nodes in the graph, and let $\widetilde{A}$ denote the set of all nodes in A together with all nodes in $V \setminus S$ which are connnected to A. By the separation assumption $\widetilde{A}$ will not contain any nodes from B. We then let $\widetilde{B} = (V \setminus S) \cup \widetilde{A}$, so that $\widetilde{A}$, $\widetilde{B}$ and S form disjoint sets such that $V = \widetilde{A} \cup \widetilde{B} \cup S$, and S separates $\widetilde{A}$ from $\widetilde{B}$. It follows that any clique is composed either of nodes from $S \cup \widetilde{A}$ or of nodes from $S \cup \widetilde{B}$, and so by the factorization property

$$P(V) \equiv P(\widetilde{A}, \widetilde{B}, S) = f(\widetilde{A}, S) g(\widetilde{B}, S). \tag{4.20}$$

We therefore have

$$\begin{aligned} P(\widetilde{A}, \widetilde{B}|S) &= \frac{P(\widetilde{A}, \widetilde{B}, S)}{\sum_A \sum_B P(\widetilde{A}, \widetilde{B}, S)} \\ &= \frac{f(\widetilde{A}, S)}{\sum_A f(\widetilde{A}, S)} \frac{g(\widetilde{B}, S)}{\sum_B g(\widetilde{B}, S)} \\ &= P(\widetilde{A}|S) P(\widetilde{B}|S) \end{aligned} \tag{4.21}$$

and hence $\widetilde{A} \perp\!\!\!\perp \widetilde{B} \mid S$. By noting that A is a subset of $\widetilde{A}$, and B is a subset of $\widetilde{B}$, and then applying Theorem 4.2 twice we obtain $A \perp\!\!\!\perp B \mid S$ as required. $\square$

The converse of Theorem 4.7 does not hold for all distributions. However, the Hammersley-Clifford theorem states that, for strictly positive probability distributions, the global Markov property is equivalent to the factorization property. The proof of this theorem is postponed to Section 4.5.3 since it makes use of additional Markov properties discussed in Section 4.5.

In the next section we exploit the formalism of undirected graphs to gain further insights into the properties of directed graphs.

4.4 Directed Graphs Revisited

We now return to a discussion of directed graphs, and make use of the techniques of undirected graphs discussed in the previous section to develop an alternative formulation of the global Markov properties for directed graphs which in some respects is simpler than the d-separation criterion discussed so far. This will also motivate the graphical concept of *moralization* which will prove useful in discussing inference algorithms in Chapter ?. We will also complete our discussion of the Markov properties of directed graphs by introducing the concepts of local and pairwise Markov, and relate these to the global Markov properties and to factorization.

4.4.1 Directed Global Markov

We have seen that the global Markov properties of an undirected graph can be obtained through simple graph separation, whereas in the case of directed graphs we have to employ the significantly more complex d-separation criterion. It is natural then to ask whether we can use the machinery of undirected graphs to obtain an alternative formulation of the global Markov properties of directed graphs.

As we shall see, simply dropping the arrows on the links of a directed graph, and then applying undirected graph separation does not yield the required conditional independencies. However, we will show that from the original directed graph we can extract an appropriate undirected graph which does exhibit the required independence properties. The specific undirected graph which is needed will depend upon the particular nodes which are included in the conditioning set.

To begin with consider the directed graph shown in Figure 4.10 involving 3 nodes A, B, and C in which we have conditioned on C. We have al-

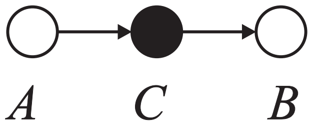


Figure 4.10: A directed graph of three nodes for which $A \perp\!\!\!\perp B \mid C$.

ready seen in Section 4.2.2 that this graph satisfies the independence property $A \perp\!\!\!\perp B \mid C$. Note that this property could have been read off from an undirected graph obtained simply by dropping the arrows on the links, as shown in Figure 4.11.

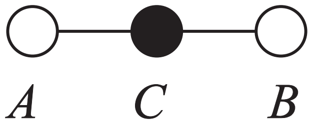


Figure 4.11: The undirected graph obtained from Figure 4.3 by dropping the arrows on the links. This graph also satisfies $A \perp\!\!\!\perp B \mid C$.

Similarly, consider the graph in Figure 4.12. Again, the conditional independence property $A \perp\!\!\!\perp B \mid C$ can be read off by graph separation from the corresponding undirected graph, again given by Figure 4.12.

However, the situation is different for the 'head-to-head' node shown in Figure 4.13. The undirected graph in this case is again given by Fig-

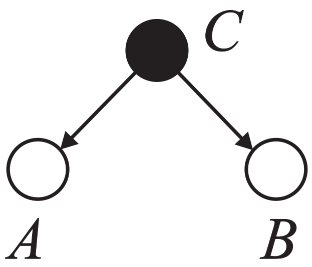


Figure 4.12: Another directed graph for which $A \perp\!\!\!\perp B \mid C$.

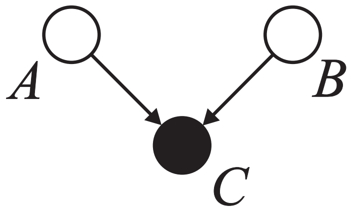


Figure 4.13: A directed graph for which $A \not\perp\!\!\!\perp B \mid C$.

ure 4.12, but this now does not give the correct conditional independence result since the undirected graph predicts $A \perp\!\!\!\perp B \mid C$, whereas this result does not follow from the original directed graph. We can avoid the introduction of this spurious conditional independence property by first introducing an extra link between the two parents of the conditioning node C and then dropping the arrows on the links to give the undirected graph of Figure 4.14. This separation properties of this graph now no longer imply

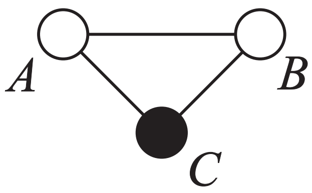


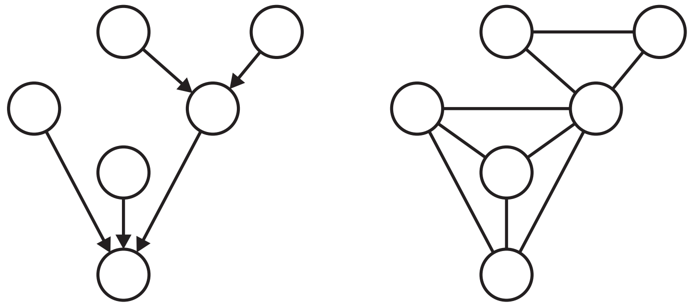
Figure 4.14: The undirected graph obtained by moralizing the directed graph of Figure 4.5.

an incorrect conditional independence statement.

The procedure of adding links to connect nodes with a common descendent is called *moralization*[1] and plays an important role in our discussion of

[1]This term originally arose from the idea of 'marrying the parents' of the node. We

graphical models. To construct the moral graph from an arbitrary directed acyclic graph we first add additional links between all pairs of nodes having a common child and then drop the arrows on the links. An example of a DAG and its moral graph is shown in Figure 4.15.



(a) (b)

Figure 4.15: An example of a directed acyclic graph (a), together with the corresponding moral graph (b).

We now show that moralization is a sufficient condition to ensure that any conditional independence statement inferred by graph separation on the moral graph holds true also in the factorized distribution.

Theorem 4.8 *If a probability distribution factorizes according to a directed acyclic graph then it respects the global Markov property of the corresponding moral graph.*

Proof: Each factor in the factorization (4.16) is a conditional distribution for a node conditioned on its parents. In the moral graph, this subgraph is complete, since all pairs of parents have been connected by a link. Thus we can take the undirected moral graph, and starting with all of the clique potentials set to unity, multiply each factor in (4.16) into the corresponding clique potential, yielding a factorized representation for the joint distribution of the form (4.19). Hence the directed factorization property $\mathcal{DF}$ implies undirected factorization $\mathcal{F}$ on the corresponding moral graph. We then invoke Theorem 4.7 which shows that factorization on the moral graph implies that the distribution is global Markov with respect to that graph. $\square$

continue to use the term moralization throughout this book since it is now in widespread use.

We now ask whether the converse is true, that is whether every conditional independence statement implied by the directed graph holds on the moral graph. The answer is clearly that it does not, as illustrated in Figure 4.16. The problem in Figure 4.16 arises from the node W. Since this

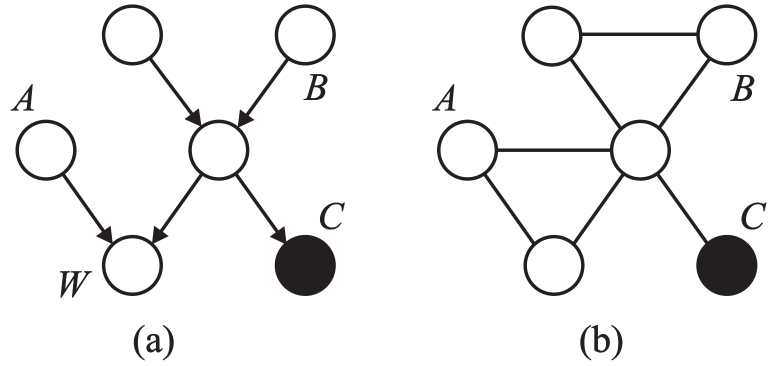


Figure 4.16: Example of a directed graph (a) which exhibits a conditional independence statement $A \perp\!\!\!\perp B \mid C$, which is not represented in the corresponding moral graph (b).

node is not part of the conditioning set it should really be removed from the graph before moralization to avoid the introduction of a spurious link.

We can see how to address this problem more generally by noting that the conditional independence statement $A \perp\!\!\!\perp B \mid C$ is a property of the marginal distribution $P(A, B, C)$ in which the remaining variables have been summed out. Any variables which are not part of $A \cup B \cup C$ or their ancestors can trivially be removed from the distribution to leave a marginal distribution over the remaining variables simply by removing the corresponding conditional distributions from the product in (4.15).

This motivates the consideration of *ancestral sets*. A node A is said to be an *ancestor* of a node B if there is a directed link from A to B. Similarly B is then said to be a *descendent* of A. We say that a sub-set of nodes within a directed acyclic graph is an ancestral set if, for every node in the set all ancestors of that node are also in the set. Using this concept we can now introduce a new definition for the global Markov property of directed graphs, which provides an alternative to d-separation.

The global markov property for directed graphs, denoted by $\mathcal{DG}$, says that $A \perp\!\!\!\perp B \mid C$ whenever C separates A from B in the moral graph of the smallest ancestral DAG containing A, B and C. This is illustrated in Figure 4.17, using the example of Figure 4.16.

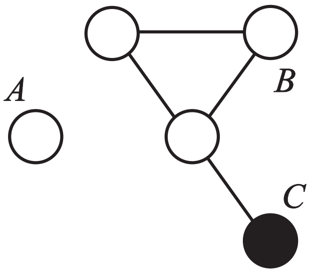


Figure 4.17: The moral graph of the smallest ancestral set from Figure 4.16(a) containing nodes A, B and C. In this case the conditional independence property $A \perp\!\!\!\perp B \mid C$ follows correctly from undirected graph separation.

We now show that any independence property which can be found by graph separation in the moral graph of the smallest ancestral set containing the relevant variables must hold also in the probability distribution.

Theorem 4.9 *($\mathcal{DF} \Rightarrow \mathcal{DG}$) If a probability distribution factorizes according to a directed acyclic graph, then $A \perp\!\!\!\perp B \mid C$ whenever A and B are separated by C in the moral graph of the smallest ancestral set containing $A \cup B \cup C$.*

Proof: If a probability distribution factorizes with respect to a particular directed graph, then the marginal distribution over a subset of nodes corresponding to an ancestral set will factorize with respect to the corresponding ancestral graph. This is easily seen by marginalizing the factored expression (4.16) over the remaining variables. Then, using Theorem 4.8 it follows that any independence obtained by graph separation in the moral graph of the smallest ancestral set containing $A \cup B \cup C$ will hold also in the marginal distribution. Hence $\mathcal{DF} \Rightarrow \mathcal{DG}$ as required. $\square$

4.4.2 Directed Local Markov

Next we describe the local Markov property for directed graphs, denoted by $\mathcal{DL}$, which is defined as follows. We define the descendents of a node α to be the set of nodes which can be reached from α by following links in the direction of the arrows. All of the remaining nodes, except α itself, constitute the non-descendents of α, denoted $\mathrm{nd}(\alpha)$. A probability distribution satisfies the directed local Markov property for a given directed graph if every variable α is conditionally independent of its non-descendents given its parents

$$\alpha \perp\!\!\!\perp \mathrm{nd}(\alpha) \mid \mathrm{pa}(\alpha). \tag{4.22}$$

We next show that the directed global Markov property implies the directed local Markov property. If we think of these properties as filters, this says that any distribution which passes the $\mathcal{DG}$ filter will also pass the $\mathcal{DL}$ filter.

Theorem 4.10 *($\mathcal{DG} \Rightarrow \mathcal{DL}$) If a probability distribution respects the directed global Markov property for a given directed acyclic graph then it also respects the directed local Markov property.*

Proof: We first note that $\alpha \cup \mathrm{nd}(\alpha)$ is an ancestral set. Next we observe that $\mathrm{pa}(\alpha)$ separates α from $\mathrm{nd}(\alpha) \setminus \mathrm{pa}(\alpha)$ within this ancestral set. The directed local Markov property then follows as a special case of the global directed Markov property. $\square$

We now complete the loop by showing that the directed local Markov property implies directed factorization.

Theorem 4.11 *($\mathcal{DL} \Rightarrow \mathcal{DF}$) For any directed graph, and any distribution, if the distribution satisfies the local Markov property with respect to the graph then it will also factorize according to the graph.*

Proof: The proof is based on induction in the number N of vertices in the graph. Suppose the result holds for an arbitrary directed acyclic graph with N nodes. Now add an additional node X_{N+1} which is a terminal node of the graph. For an arbitrary joint distribution over the $N+1$ variables we have

$$P(X_1, \dots, X_{N+1}) = P(X_1, \dots, X_N) P(X_{N+1} | X_1, \dots, X_N). \tag{4.23}$$

Using the directed local Markov property $\mathcal{DL}$ we have $P(X_{N+1}|X_1, \dots, X_N) = P(X_{N+1}|\mathrm{pa}(X_{N+1}))$. Also, by the inductive assumption $P(X_1, \dots, X_N)$ factorizes according to $\mathcal{DF}$. Hence $P(X_1, \dots, X_{N+1})$ factorizes according to $\mathcal{DF}$. $\square$

We end this section by relating d-separation back to the directed local Markov property.

Theorem 4.12 *(d-separation $\Rightarrow$ $\mathcal{DL}$) If a probability distribution satisfies the* *conditional independencies implied by d-separation over a particular directed graph, then it will also satisfy the local Markov properties implied by that graph.*

Proof: For each node α in the graph, suppose we have conditioned on the parents of α, and consider all possible paths which start at α and which end at a node in the set of non-descendents of α. Any such path must either (i) pass through a parent of α or (ii) pass through a child of α. In case (i) the parent is clearly not a head-to-head node with respect to the path, and since the parent node is in the conditioning set such a path must be blocked. In case (ii) the path must at some point pass through a head-to-head node in order to reach a non-descendent and since neither that node nor any of its descendents is in the conditioning set the pode must again block the path. Since all such paths are blocked, the d-separation criterion implies $\alpha \perp\!\!\!\perp \mathrm{nd}(\alpha) \mid \mathrm{pa}(\alpha)$. $\square$

4.4.3 Directed Pairwise Markov

Our last Markov property for directed graphs is directed pairwise Markov, denoted $\mathcal{DP}$. We say that a distribution obeys the directed pairwise Markov property in respect of a given directed acyclic graph if, for any two nodes α and β such that $\beta \in \mathrm{nd}(\alpha)$

$$\alpha \perp\!\!\!\perp \beta \mid \mathrm{nd}(\alpha) \setminus \beta. \tag{4.24}$$

This is illustrated in Figure 4.18.

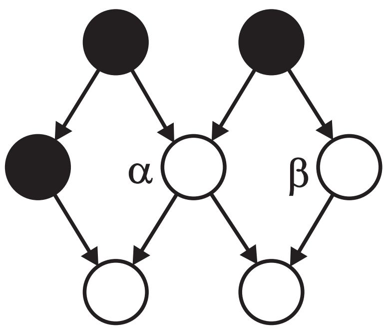


Figure 4.18: An illustration of the pairwise Markov property in which $\alpha \perp\!\!\!\perp \beta \mid \mathrm{nd}(\alpha) \setminus \beta$.

We now show that the directed local Markov property implies the directed pairwise Markov property for all graphs and for all distributions.

Theorem 4.13 *($\mathcal{DL} \Rightarrow \mathcal{DP}$) If a distribution satisfies the directed local Markov property with respect to some directed graph, then it also satisfies the directed pairwise Markov property with respect to the same graph.*

Proof: From the definition (4.22) of the directed local Markov property we have

$$\alpha \perp\!\!\!\perp (\mathrm{nd}(\alpha) \setminus \beta) \cup \beta \mid \mathrm{pa}(\alpha) \tag{4.25}$$

where we have made use of $\beta \in \mathrm{nd}(\alpha)$. From Theorem 4.3 we then have

$$\alpha \perp\!\!\!\perp \beta \mid (\mathrm{nd}(\alpha) \setminus \beta) \cup \mathrm{pa}(\alpha). \tag{4.26}$$

Finally, since $\mathrm{pa}(\alpha) \subseteq \mathrm{nd}(\alpha)$, and since $\beta \notin \mathrm{pa}(\alpha)$ we have $\alpha \perp\!\!\!\perp \beta \mid \mathrm{nd}(\alpha) \setminus \beta$. □

Note that the converse of this theorem is not in general true. However, if we restrict attention to distributions which satisfy the intersection property (4.9), for example distributions which are strictly positive, then $\mathcal{DP} \Rightarrow \mathcal{DL}$.

Theorem 4.14 *($\mathcal{DP} \Rightarrow \mathcal{DL}$) If a distribution satisfies the directed pairwise Markov property with respect to some directed graph, and respects the intersection property (4.9), then it also satisfies the directed local Markov property with respect to the same graph.*

Proof: We prove this result by induction as follows. Let the nodes in $\mathrm{nd}(\alpha)$ which are not members of $\mathrm{pa}(\alpha)$ be labelled $\beta_1, \ldots, \beta_N$, and suppose

$$\alpha \perp\!\!\!\perp \beta_1 \cup \ldots \cup \beta_m \mid \mathrm{pa}(\alpha) \cup \beta_{m+1} \cup \ldots \cup \beta_N. \tag{4.27}$$

This result clearly holds for $m = 1$ from the pairwise Markov property. Also from the pairwise Markov property we have

$$\alpha \perp\!\!\!\perp \beta_{m+1} \mid \mathrm{pa}(\alpha) \cup \beta_1 \cup \ldots \cup \beta_m \cup \beta_{m+2} \cup \ldots \cup \beta_N. \tag{4.28}$$

Now using Theorem 4.5 we have

$$\alpha \perp\!\!\!\perp \beta_1 \cup \ldots \cup \beta_{m+1} \mid \mathrm{pa}(\alpha) \cup \beta_{m+2} \cup \ldots \cup \beta_N. \tag{4.29}$$

Continuing this inductive process we arrive at (4.29) with $m+1 = N$ which is the required result. $\square$

4.4.4 Summary of Markov properties for Directed Graphs

So far in this chapter we have shown that $\mathcal{DF} \Rightarrow$ d-separation (Theorem 4.6), $\mathcal{DF} \Rightarrow \mathcal{DG}$ (Theorem 4.9), $\mathcal{DG} \Rightarrow \mathcal{DL}$ (Theorem 4.10), $\mathcal{DL} \Rightarrow \mathcal{DF}$ (Theorem 4.11), and d-separation $\Rightarrow \mathcal{DL}$ (Theorem 4.12). Hence, without restriction on the graph or the distribution, we have

$$\mathcal{DF} \Leftrightarrow \mathcal{DL} \Leftrightarrow \mathcal{DG} \Leftrightarrow \text{d-separation.} \tag{4.30}$$

Note that this confirms that the directed global Markov property defined through the moral graph of the smallest ancestral set is equivalent to the d-separation criterion. We have also shown that $\mathcal{DL} \Rightarrow \mathcal{DP}$ (Theorem 4.13) and hence $\mathcal{DF}$, $\mathcal{DL}$, $\mathcal{DG}$ and d-separation all imply $\mathcal{DP}$.

If we restrict attention to distributions satisfying the intersection property (4.9) then we have the further result $\mathcal{DP} \Rightarrow \mathcal{DL}$ (Theorem 4.14) and hence, for such distributions, all of the Markov properties as well as d-separation and the factorization property are equivalent.

4.5 Undirected Graphs Revisited

In this section we complete our discussion of the Markov properties of undirected graphs by considering pairwise and local Markov and their relation to global Markov and to factorization. This material is included mainly for technical completeness and is not required in subsequent chapters.

4.5.1 Pairwise Markov

We first define the *pairwise Markov* property for undirected graphs, which we denote by $\mathcal{P}$. A distribution is pairwise Markov with respect to a given graph if, for any two nodes α and β in the graph such that there is no direct link in the graph from α to β, then α is independent of β given the states of all of the remaining nodes, so that

$$\alpha \perp\!\!\!\perp \beta \mid S \setminus \{\alpha, \beta\} \tag{4.31}$$

where S denotes the set of all nodes in the graph.

If there is no direct link between the nodes then they are necessarily separated by the remaining nodes according to the graph separation criterion. Thus the pairwise Markov property is a special case of the global Markov property, so that $\mathcal{G} \Rightarrow \mathcal{P}$. In other words if a distribution is global Markov with respect to a particular graph then it is necessarily also pairwise Markov. However, the converse does not necessarily hold. A sufficient condition for equivalence of $\mathcal{G}$ and $\mathcal{P}$ is that the intersection property of Eq. (4.9) is satisfied, which will be the case, for example, if we restrict attention to distributions which are strictly positive. This leads to the following result.

Theorem 4.15 *($\mathcal{P} \Rightarrow \mathcal{G}$) For any undirected graph, and for distributions which satisfy the intersection property (4.9), if a distribution satisfies the pairwise Markov property with respect to the graph then it will also satisfy the global Markov property.*

Proof: Consider three disjoint subsets A, B and C of nodes, and suppose that C separates A from B on the graph. We also assume that the pairwise Markov property $\mathcal{P}$ holds and that (4.9) is valid. Our goal is to prove that $A \perp\!\!\!\perp B \mid C$. We can do this by induction on the number $|C|$ of nodes in the separating set C. Suppose $|C| = |S| - 2$ where $|S|$ is the total number of nodes. Then A and B each contain one node and the pairwise Markov property implies directly that $A \perp\!\!\!\perp B \mid C$. We then suppose that $\mathcal{G}$ holds for all values of $|C|$ greater than some value n and then show that it also holds for $|C| = n$. There are now two possibilities to consider. In the first, suppose that $A \cup B \cup C = S$. This implies that either A or B, or both, has more than one node. Suppose A has more than one node and that $\alpha \in A$, as illustrated in Figure 4.19. Since C separates A from B it follows that $C \cup \alpha$

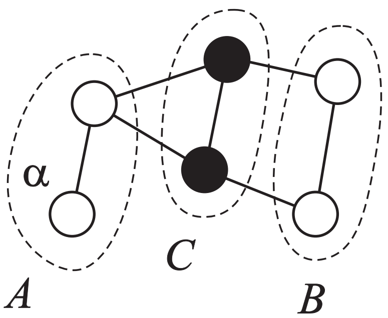


Figure 4.19: Example of a graph corresponding to case (i) in Theorem 4.15.

must separate $A \setminus \alpha$ from B. Using the inductive assumption that separation implies independence for $|C| > n$ we have $A \setminus \alpha \perp\!\!\!\perp B \mid C$. Similarly, $C \cup A \setminus \alpha$ separates α from B and so $\alpha \perp\!\!\!\perp B \mid C \cup A \setminus \alpha$. Using the intersection property (4.9) we then have $A \perp\!\!\!\perp B \mid C$ as required. The second possibility is that $A \cup B \cup C \subset S$ so that some nodes lie outside the sets A, B and C, as illustrated in Figure 4.20. We then choose $\alpha \in S \setminus (A \cup B \cup C)$. Then $S \cup \alpha$

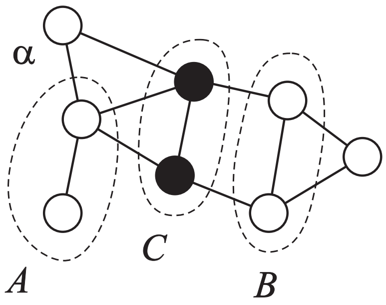


Figure 4.20: Example of a graph corresponding to case (ii) in Theorem 4.15.

separates A from B in the graph and so $A \perp\!\!\!\perp B \mid C \cup \alpha$, since adding a node to the conditioning set cannot revoke a separation property. There are now two sub-possibilities to consider. We can have $A \cup C$ separates α from B (as illustrated in Figure 4.20) in which case $\alpha \perp\!\!\!\perp B \mid A \cup C$ from which, using (4.9) we obtain $A \perp\!\!\!\perp B \mid C$ as required. Alternative we can have that $B \cup C$ separates α from A and hence $\alpha \perp\!\!\!\perp A \mid B \cup C$ from which, using (4.9), we again obtain $A \perp\!\!\!\perp B \mid C$. $\square$

4.5.2 Local Markov

Next we consider the *local Markov property*, denoted by $\mathcal{L}$, which says that the conditional distribution of a variable α given the neighbours of α in the graph is independent of the remaining nodes. We can express this more formally by introducing the concept of the *boundary* of α, denoted $\mathrm{bd}(\alpha)$, which comprises all of the nodes which have a direct link to α. Similarly we define the *closure* of α, denoted $\mathrm{cl}(\alpha)$, to be the union of α and its boundary, $\mathrm{cl}(\alpha) \equiv \alpha \cup \mathrm{bd}(\alpha)$. The local Markov property for undirected graphs can then be expressed as

$$\alpha \perp\!\!\!\perp S \setminus \mathrm{cl}(\alpha) \mid \mathrm{bd}(\alpha). \tag{4.32}$$

This is illustrated in Figure 4.21.

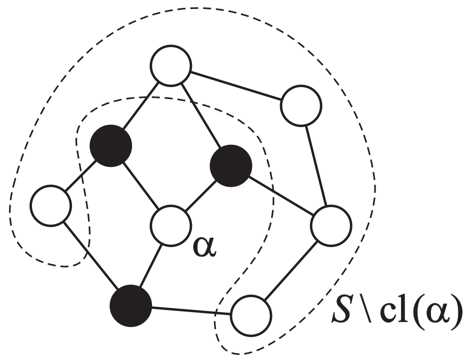


Figure 4.21: Illustration of the local Markov property (4.32) in which the black nodes represent the boundary of α.

Clearly the local Markov property is a special case of the global Markov property, so $\mathcal{G} \Rightarrow \mathcal{L}$. We now prove that local Markov implies pairwise Markov.

Theorem 4.16 *($\mathcal{L} \Rightarrow \mathcal{P}$) For any undirected graph and any distribution, if the distribution satisfies the local Markov property with respect to a graph then it will also satisfy the pairwise Markov property for that graph.*

Proof: Consider two nodes α and β, let $\mathrm{bd}(\alpha)$ denote the boundary of α and let B denote the remaining nodes including β, as illustrated in Figure 4.22. Then the local Markov property tells us that

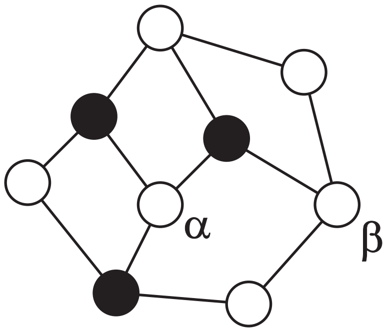


Figure 4.22: Example of an undirected graph in which the black nodes denote the boundary of a particular node α. The remaining white nodes constitute the set B which includes a specific node β.

$$\alpha \perp\!\!\!\perp B \mid \mathrm{bd}(\alpha). \tag{4.33}$$

Since $\beta \in B$ we have $B \equiv \beta \cup (B \setminus \beta)$. We can then apply the weak union property (4.7) to obtain

$$\alpha \perp\!\!\!\perp \beta \mid \mathrm{bd}(\alpha) \cup (B \setminus \beta) \tag{4.34}$$

and since the union of $\mathrm{bd}(\alpha)$ and $(B \setminus \beta)$ represents the complete set S of nodes in the graph, less α and β we have

$$\alpha \perp\!\!\!\perp \beta \mid S \setminus (\alpha \cup \beta) \tag{4.35}$$

which is the pairwise Markov property. $\square$

4.5.3 Hammersely-Clifford Theorem

We now return to the Hammersley-Clifford theorem, and use the results derived earlier in this section to provide a formal proof. First we prove the Möbius inversion lemma.

Lemma 4.1 *Consider a finite set V of elements $a \in V$, and let Ψ and Φ be functions defined over all possible subsets of V. Then the statement*

$$\Psi(a) = \sum_{b:b \subseteq a} \Phi(a) \tag{4.36}$$

is equivalent to the statement

$$\Phi(a) = \sum_{b:b \subseteq a} (-1)^{|a \setminus b|} \Psi(b) \tag{4.37}$$

where $|a|$ denotes the cardinality of the subset a.

Proof: We show that (4.37) implies (4.36), with the proof of the converse proceeding analogously. Consider

$$\begin{aligned}\sum_{b:b\subseteq a}\Phi(b) &= \sum_{b:b\subseteq a}\sum_{c:c\subseteq b}(-1)^{|b\setminus c|}\Psi(c)\\ &= \sum_{c:c\subseteq a}\left\{\sum_{b:c\subseteq b\subseteq a}(-1)^{|b\setminus c|}\right\}\\ &= \sum_{c:c\subseteq a}\left\{\sum_{h:h\subseteq a\setminus c}(-1)^{|h|}\right\}.\end{aligned}$$

The final sum on the right hand side is zero unless $\subseteq a\setminus c=\emptyset$ (i.e. $c=a$) since for any finite, non-empty set the number of subsets of even cardinality is the same as the number of subsets of odd cardinality (as is easily verified by induction). □

We now use Lemma 4.1 to prove the Hammerseley-Clifford theorem.

Theorem 4.17 *(***Hammersely-Clifford:** $\mathcal{P}\Rightarrow\mathcal{F}$*) For distributions which satisfy the intersection property (4.9), and for arbitrary undirected graphs, any distribution which satisfies the pairwise Markov property for a particular graph will factorize according to that graph.*

Proof: Consider the joint distribution $P(S)$ where $S\equiv\{S_\alpha\}$ and the S_α take values in some space χ. Now choose a particular (arbitrary) value $S^*\in\chi$. For each possible subset $a\subseteq S$ we define the function

$$H_a(S)=\ln P(\widehat{S}a) \tag{4.38}$$

where $\widehat{S}_a$ has components $\widehat{S}^{a\alpha}=S_\alpha$ for $\alpha\in a$ and $\widehat{S}_{a\alpha}=S^*$ for $\alpha\notin a$. Thus $H_a(S)$ depends on S only through S_a. Now define the following set of functions for all $a\subseteq S$

$$\phi_a(S)=\sum_{b:b\subseteq a}(-1)^{|a\setminus b|}H_b(S). \tag{4.39}$$

Again we see that $\phi_a(S)$ depends on S only through S_a. Using Lemma 4.36 we obtain

$$H_S(S) = \sum_{a:a \subseteq S} \phi_a(S). \tag{4.40}$$

From the definition of $H_A(S)$ we also have $H_S(S) = \ln P(S)$. Defining $\psi_a(S_a) = \exp \phi_a(S_a)$ and taking the exponential of both sides of (4.40) we obtain

$$P(S) = \prod_{a:a \subseteq S} \phi_a(S) \tag{4.41}$$

which has the required form of a product over potential functions. The final step is to show that $\phi_a(S)$ vanishes unless the subset a is complete. To do this we make use of the assumed pairwise Markov property. Let $\alpha, \beta \in a$ be two nodes with no direct link between them, and let $c = a \setminus \{\alpha, \beta\}$. If we let H_a denote $H_a(S)$ then

$$\phi_a(S) = \sum_{b:b \subseteq c} (-1)^{|c \setminus b|} \left\{ H_b - H_{b \cup \alpha} - H_{b \cup \beta} + H_{b \cup \{\alpha, \beta\}} \right\}. \tag{4.42}$$

If we define $d = S \setminus \{\alpha, \beta\}$, then by the pairwise Markov property $\alpha \perp\!\!\!\perp \beta \mid d$ and hence

$$
\begin{aligned}
H_{b\cup\{\alpha,\beta\}} - H_{b\cup\alpha} &= \ln \frac{P(S_b, S_\alpha, S_\beta, S^*_{d\setminus b})}{P(S_b, S_\alpha, S^*_\beta, S^*_{d\setminus b})} \\
&= \ln \frac{P(S_\alpha|S_b, S^*_{d\setminus b})P(S_\beta, S_b, S^*_{d\setminus b})}{P(S_\alpha|S_b, S^*_{d\setminus b})P(S^*_\beta, S_b, S^*_{d\setminus b})} \\
&= \ln \frac{P(S^*_\alpha|S_b, S^*_{d\setminus b})P(S_\beta, S_b, S^*_{d\setminus b})}{P(S^*_\alpha|S_b, S^*_{d\setminus b})P(S^*_\beta, S_b, S^*_{d\setminus b})} \\
&= \ln \frac{P(S_b, S^*_\alpha, S_\beta, S^*_{d\setminus b})}{P(S_b, S_\alpha, S^*_\beta, S^*_{d\setminus b})} \\
&= H_{b\cup\beta} - H_b.
\end{aligned}
$$

Hence the sum of the terms in the brackets on the right hand side of (4.42) vanishes whenever we can find two nodes α and β having no direct connection between them. Thus $\phi_a(S)$ vanishes unless a is a complete set. □

4.5.4 Summary of Markov Properties for Undirected Graphs

In this chapter we have proved the following results: $\mathcal{F} \Rightarrow \mathcal{G}$ (Theorem 4.7) and $\mathcal{L} \Rightarrow \mathcal{P}$ (Theorem 4.16). Since the local and pairwise Markov properties are special cases of the global Markov property, we also trivally have $\mathcal{G} \Rightarrow \mathcal{L}$ and $\mathcal{G} \Rightarrow \mathcal{P}$. Thus for all distributions

$$\mathcal{F} \Rightarrow \mathcal{G} \Rightarrow \mathcal{P} \Leftrightarrow \mathcal{L}. \tag{4.43}$$

For distributions which satisfy the intersection property (4.9), for example strictly positive distributions, we have also shown that $\mathcal{P} \Rightarrow \mathcal{G}$ (Theorem 4.15) and $\mathcal{G} \Rightarrow \mathcal{F}$ (Theorem 4.17), and so for such distributions the three Markov properties, as well as the factorization property, are all equivalent.

In Chapter ? we introduce the idea of triangulation of an undirected graph. Essentially this involves the addition of extra links such that every cycle of four or more nodes has a chord. It can be shown that the global Markov and the factorization properties are equivalent for all distributions (without restriction) if, and only if, the graph is triangulated.

4.6 Representational Limitations

We have considered three alternative ways in which to specify the conditional independence properties of a joint distribution: (i) write down an explicit list of conditional independence statements, (ii) specify an undirected graph, (iii) specify a directed graph. Here we consider the relationship between the corresponding families of distributions.

Earlier we considered a specific (directed or undirected) graph as being a filter, so that the set of all possible distributions over the given variables could be reduced to a sub-set which respect the conditional independencies implied by the graph. A graph is said to be a *D-map* (for 'dependency map') of a distribution if every conditional independence statement satisfied by the distribution is reflected in the graph. Thus a completely disconnected graph (no links) will be a trivial D-map for any distribution.

Alternatively, we can consider a specific distribution and ask which graphs have the appropriate conditional independence properties. If every conditional independence statement implied by a graph is satisfied by a specific distribution, then the graph is said to be an *I-map* (for 'independence map') of that distribution. Clearly a fully connnected graph will be a trivial I-map for any distribution.

If it is the case that every conditional independence property of the distribution is reflected in graph separation, and vice versa, then the graph is said to be a *perfect map*. A perfect map is therefore both an I-map and a D-map for the distribution.

Consider the set of distributions such that for each distribution there exists an directed graph which is a perfect map. This set is is distinct from the set of distributions such that for each distribution there exists an undirected graph which is a perfect map. In addition there are distributions for which neither directed nor undirected graphs offer a perfect map. This is illustrated as a Venn diagram in Figure 4.23.

Figure 4.24 shows an example of a directed graph which is a perfect map for a distribution satisfying the conditional independence properties $A \perp\!\!\!\perp B \mid \emptyset$ and $A \not\perp\!\!\!\perp B \mid C$. There is no corresponding undirected graph over the same three variables which is a perfect map.

Conversely, consider the undirected graph over four variables show in

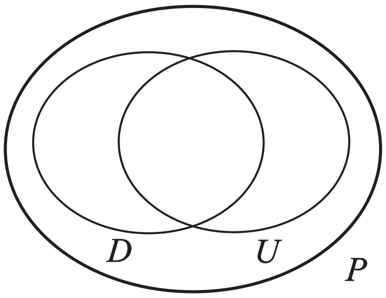


Figure 4.23: Venn diagram illustrating the set of all distributions P over a given set of variables, together the set of distributions D which can be represented as a perfect map using a directed graph, and the set U which can be represented as a perfect map using an undirected graph.

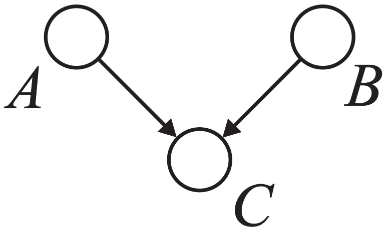


Figure 4.24: A directed graph whose conditional independence properties cannot be expressed using a directed graph over the same three variables.

Figure 4.25. This graph exhibits the properties $A \not\perp\!\!\!\perp B \mid \emptyset$, $C \perp\!\!\!\perp D \mid A \cup B$ and $A \perp\!\!\!\perp B \mid C \cup D$. There is no directed graph over four variables which implies the same set of conditional independence properties.

The graphical framework can be extended to graphs which include both directed and undirected edges and which contain the directed and undirected graphs considered so far as special cases. Such graphs are called *chain graphs*, and although they represent a broader class than either directed or undirected alone, there remain distributions for which even a chain graph cannot provide a perfect map. Chain graphs are not discussed further in this book.

4.6.1 Markov Equivalent Graphs

For distributions whose conditional independence properties can be represented graphically, we might ask whether the graph is unique. Here we

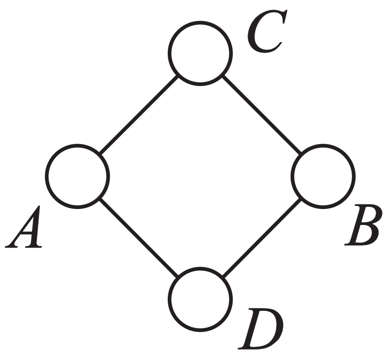


Figure 4.25: An undirected graph whose conditional independence properties cannot be expressed in terms of a directed graph over the same variables.

consider directed graphs, and discuss whether two different graphs can be Markov equivalent, that is whether they can imply the same list of conditional independence statments. [to be completed]

4.7 Historical Remarks and Bibliography

History of the Hammersley-Clifford theorem. Acknowledgement to Lauritzen's book for proofs. Castello's book. I-maps, D-maps and d-separation introduced by Pearl. The use of D-separation to determine global Markov properties for directed graphs with deterministic connections is discussed in Geiger and Pearl (1990). Geiger, Verma, and Pearl (1990) constructed an example model which is such that any conditional independencies which do not correspond to d-separation on a DAG are not present in the distribution, showing that in general d-separation will find all of the conditional independences which can be determined directly from the DAG.

Exercises

4.1 ($\star$) Consider three binary variables $A, B, C \in \{0, 1\}$ having the joint distribution given in Table 4.1. Show by direct evaluation that this distribution has the property that A and B are marginally dependent, so that $P(A, B) \neq P(A)P(B)$, but that they become independent when conditioned on C, so that $P(A, B|C) = P(A|C)P(B|C)$ for both $C = 0$ and $C = 1$.

A	B	C	$P(A,B,C)$
0	0	0	?
0	0	1	?
0	1	0	?
0	1	1	?
1	0	0	?
1	0	1	?
1	1	0	?
1	1	1	?

Table 4.1: The joint distribution over three binary variables.

4.2 (⋆) Using the result of the Exercise 4.1 show that the joint distribution specified in Table 4.1 can be expressed in the form $P(A, B, C) = P(A)P(C|A)P(B|C)$. Draw the corresponding directed graph.

4.3 (⋆) Show that there are $2^{M(M-1)/2}$ distinct undirected graphs over M variables. Draw the 8 possibilities for the case of $M = 3$.

4.4 (⋆) Consider the simple graph shown in Figure 4.26, together with the probability distribution over three binary variables X, Y and Z such that $P(X = 0) = P(X = 1) = 0.5$ and $X = Y = Z$. Show that for this graph and distribution the pairwise Markov property is statisfied, but not the local Markov property. Note that this distribution is not strictly positive since, for instance, $P(X = 0, Y = 1, Z = 1) = 0$.

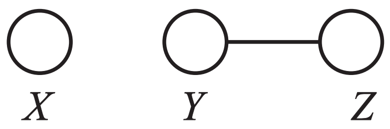


Figure 4.26: A graph which provides a counter example to the claim $\mathcal{P} \Rightarrow \mathcal{L}$.