2 Directed Graphical Models

In general, a probabilistic graphical model describes a *family* of joint probability distributions over a set of random variables. One example of a family of joint distributions are all pairs of random variables x and y that are independent, i.e., joint distributions of the factored form $p_{\mathsf{x},\mathsf{y}}(x, y) = p_{\mathsf{x}}(x)\, p_{\mathsf{y}}(y)$.

We now introduce an important class of graphical models referred to as directed graphical models, or also "belief networks." Directed graphical models, have traditionally been popular for modeling and exploring causal relationships among phenomena, but as we will see their utility is not limited to such applications. In addition to defining these models, our treatment will develop an understanding of the family of distributions that are described by a given directed graphical model.

A directed graphical model $\mathcal{G}$ consists of a collection of nodes[1] $\mathcal{V} = \{1, 2, \ldots, N\}$ (representing random variables) and a collection of directed edges (arrows) $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$. Moreover, as specific notation, $(i, j) \in \mathcal{E}$ means that there is a directed edge from node i to node j.

Directed graphs define families of distributions that factor by functions of nodes and their parents. Formally, we associate with each node i a random variable $\mathsf{x}_i \in \mathcal{X}$ and a non-negative-valued function $f_i(x_i, x_{\pi_i}) : \mathcal{X}^{|\pi_i|+1} \to \mathbb{R}_+$ such that

$$\sum_{x_i \in \mathcal{X}} f_i(x_i, x_{\pi_i}) = 1 \qquad \text{and} \qquad \prod_{i=1}^{N} f_i(x_i, x_{\pi_i}) = p_{\mathsf{x}_1,\ldots,\mathsf{x}_N}(x_1, \ldots, x_N), \tag{1}$$

where π_i denotes the set of parents of node i. i.e., $\pi_i = \{j \in \mathcal{V} \mid (j, i) \in \mathcal{E}\}$

If the graph has a directed cycle, such as Fig. 1 depicts, then there is no consistent way to assign the $f_i(\cdot)$ to meet the constraints. However, we restrict our attention to directed *acyclic* graphs (DAGs). For such graphs, which have no directed cycles, we have the following result:

Claim 1. *The unique solution to the constraints* (1) *is*

$$f_i(x_i, x_{\pi_i}) = p_{\mathsf{x}_i|\mathsf{x}_{\pi_i}}(x_i|x_{\pi_i}), \tag{2}$$

i.e., $f_i(\cdot, \cdot)$ represents the conditional probability distribution for x_i conditioned on the parent variables x_{π_i}.

[1] Without loss of generality, we are free to label the nodes in whatever way is convenient.

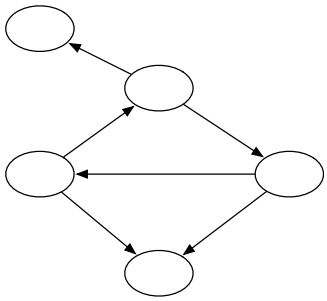


Figure 1: Example of a directed graph with a cycle, for which there is no consistent way to define a corresponding family of distributions.

Proof. To verify this, assume without loss of generality that we have chosen a topological ordering of nodes, i.e., the nodes are indexed such that the parents π_i of a node i appears before the node in the ordering. Then (1) implies that

$$p_{\mathsf{x}_1,\ldots,\mathsf{x}_{N-1}}(x_1,\ldots,x_{N-1}) = \sum_{x_N}\prod_{i=1}^{N} f_i(x_i,x_{\pi_i}) = \prod_{i=1}^{N-1} f_i(x_i,x_{\pi_i}), \tag{3}$$

where to obtain the last equality we have used that because of our choice of ordering of the variables, only the term $f_N(\cdot)$ involves the variable x_N. Hence, combining (1) and (3) we conclude

$$\begin{aligned} f_N(x_N;x_{\pi_N}) &= \frac{\prod_{i=1}^{N} f_i(x_i,x_{\pi_i})}{\prod_{i=1}^{N-1} f_i(x_i,x_{\pi_i})} \\ &= \frac{p_{\mathsf{x}_1,\ldots,\mathsf{x}_N}(x_1,\ldots,x_N)}{p_{\mathsf{x}_1,\ldots,\mathsf{x}_{N-1}}(x_1,\ldots,x_{N-1})} \\ &= p_{\mathsf{x}_N|\mathsf{x}_{N-1},\ldots,\mathsf{x}_1}(x_N|x_{N-1},\ldots,x_1) \\ &= p_{\mathsf{x}_N|\mathsf{x}_{\pi_N}}(x_N|x_{\pi_N}), \end{aligned}$$

where to obtain the last equality we have used that the lefthand side of the first equality only depends on x_{π_N}.

At this point, we can remove x_N from consideration and repeat the same procedure[2] to deduce that

$$\begin{aligned} f_{N-1}(x_{N-1};x_{\pi_{N-1}}) &= p_{\mathsf{x}_{N-1}|\mathsf{x}_{\pi_{N-1}}}(x_{N-1}|x_{\pi_{N-1}}) \\ &\vdots \\ f_2(x_2;x_{\pi_2}) &= p_{\mathsf{x}_2|\mathsf{x}_{\pi_2}}(x_2|x_{\pi_2}) \\ f_1(x_1) &= p_{\mathsf{x}_1}(x_1). \end{aligned}$$

[2]The usual way to do this formally is using induction.

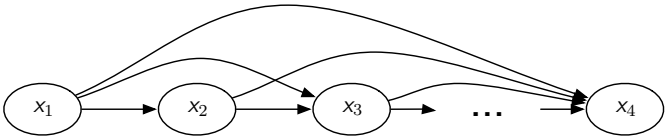


Figure 2: A fully connected DAG is universal, in that it can represent any distribution.

□

Now, by the chain rule, any joint distribution of N random variables $(x_1, \ldots, x_N)$ can in general be written as, for example,

$$p_{x_1,\ldots,x_N}(x_1, \ldots, x_N) = p_{x_1}(x_1)\, p_{x_2|x_1}(x_2|x_1) \cdots p_{x_N|x_1,\ldots,x_{N-1}}(x_N|x_1, \ldots, x_{N-1}). \quad (4)$$

Hence, by associating each of these terms with one of the functions f_i, we observe that any distribution can be described by the graph structure shown in Fig. 2. In this sense, DAGs are *universal*.

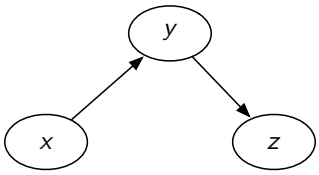
Note that the graph in Fig. 2 has a large number of edges when N is large. By contrast, of particular interest to us are *sparse* graphs, i.e. graphs where the number of edges is much smaller than the number of pairs of random variables. Such graphs can lead to efficient inference.

In general, the graph structure plays a key role of determining the size of the representation. For instance, we saw above that a fully connected DAG can represent an arbitrary distribution, and we saw in an earlier installment of the notes that the joint probability table for such a distribution requires $|\mathcal{X}|^N$ entries. More generally, the number of parameters required to represent the factorization is of order $\sum_i |\mathcal{X}|^{|\pi_i|+1} \sim |\mathcal{X}|^{\max_i |\pi_i|}$ for large N, which is dramatically smaller when $\max_i |\pi_i| \ll N$. Similarly, the graph structure affects the complexity of inference: while inference in a fully connected graph always requires $|\mathcal{X}|^N$ time, inference in a sparse graph is often (but not always) much more efficient.

2.1 Examples

There is a close relationship between how a distribution factorizes as discussed above, and conditional independence properties of such distributions, which we will now proceed to develop. However, let us first analyze some simple but insightful examples to develop our intuition before proceeding to a more general theory.

Example 1. First, consider the following scenario involving a person: x denotes whether or not the person works in Cambridge, y denotes whether the person is a leading economist or not, and z denotes whether or not the person will win a Nobel prize in the field. The relationship between these variables can be well-modeled by the graphical model



which represents the factorization

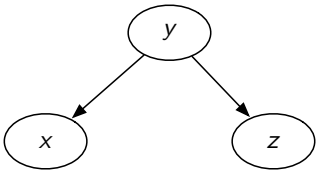
$$p_{x,y,z}(x, y, z) = p_{z|y}(z|y)\, p_{y|x}(y|x)\, p_x(x).$$

By matching these terms against those corresponding to the expansion of the distribution in terms of the chain rule, i.e., $p_{x,y,z}(x, y, z) = p_{z|y,x}(z|y, x)\, p_{y|x}(y|x)\, p_x(x)$, we see that

$$p_{x,y,z}(x, y, z) = p_{z|y}(z|y)\, p_{y|x}(y|x)\, p_x(x),$$

whence $p_{z|y}(z|y) = p_{z|y,x}(z|y, x)$, i.e., z and x are conditionally independent given y, which we express using the notation $x \perp\!\!\!\perp z \mid y$. We equivalently refer to this as *Markov* structure. Specifically, we say that x, y, z form a *Markov chain*.

Example 2. Next, consider the following scenario involving a person walking outdoors: y denotes whether or not it is raining outside, x denotes whether or not they are carrying an umbrella, and z denotes whether or not there are puddles on the ground. The relationship between these variables can be well-modeled by the graphical model



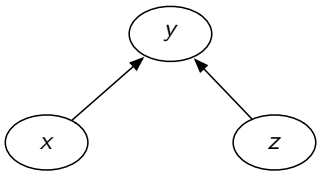
which expresses the factorization

$$p_{x,y,z}(x, y, z) = p_{z|y}(z|y)\, p_{x|y}(x|y)\, p_y(y).$$

This is sometimes referred to as a *common-cause* model. Interestingly, when we match terms in the chain rule expansion as we did in the preceding example, we find that this graph also expresses the constraint $x \perp\!\!\!\perp z \mid y$, as in that example. In other words, the models in these two examples represent exactly the same family of distributions.

One significant implication of this observation is that in general it is not possible to unambiguously infer all causal relationships from the joint distribution.

Example 3. Consider the following example of a scenario due to Pearl: y denotes whether or not your car will start, x denotes whether or not it has gas in the tank, and z denotes whether or not someone has removed the spark plugs. The relationship between these variables can be well-modeled by the following graphical model



which expresses the factorization

$$p_{x,y,z}(x, y, z) = p_x(x)\, p_{y|x,z}(y|x, z)\, p_z(z).$$

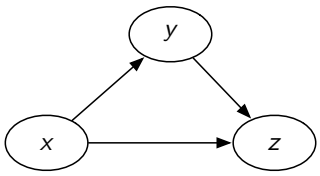
By matching terms, we find that $x \perp\!\!\!\perp z$. However, it's no longer generally true that $x \perp\!\!\!\perp z \mid y$.This implies that the direction of the edges matter in order to read off the conditional independence relations from the graph.

This model is sometimes referred to as a *common-effect* model, and, moreover, the relationship between nodes in this graph is referred to as a *V-structure* or *immorality*.

The phenomenon captured by this example is known as *explaining away* or *Berkson's paradox*. In particular, suppose we've observed an event which may result from one of two causes. If we then observe one of the causes, this makes the other one less likely—i.e., it explains it away.

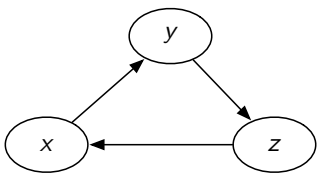
An even simpler specific instance of a scenario that can be expressed by this graphical model is as follows: x and z are binary $\mathcal{X} = \mathcal{Z} = \{0, 1\}$ random variables representing the outcome of independent fair coin flips, and y denotes their exclusive-OR, i.e., $y = x \oplus z$.

Example 4. Next, consider the graphical model



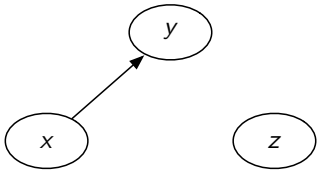
This is a fully connected graph, so as we saw before, it can represent *any* distribution over x, y, and z, and thus expresses no independence structure. This example emphasizes that one must remove edges in order to introduce independence structure.

Example 5. The following graph structure is not a valid DAG because it contains a cycle:



It cannot describe a joint distribution.

Example 6. Now consider a scenario in which x represents the age of a person, y represents how fast they can run 100m, and z represents the current price of oil. The relationship between these variables is well-modeled by the graphical model



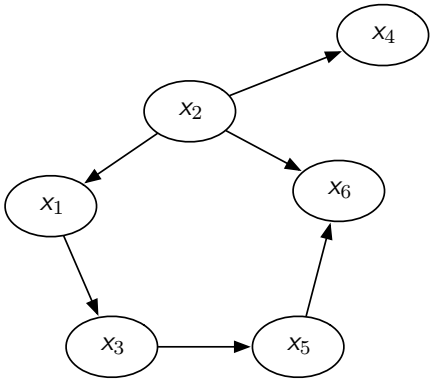
which can be viewed as the graph obtained by removing an edge from that in Example 3. The factorization represented by this graph is easily shown to be

$$p_{x,y,z}(x, y, z) = p_x(x)\, p_{y|x}(y|x)\, p_z(z),$$

which is a more constrained factorization than that in Example 3. In general, reducing edges increases the number of independencies and eliminates some of the distributions the graph could otherwise represent.

This graph has the following four independencies: $x \perp\!\!\!\perp z, y \perp\!\!\!\perp z, x \perp\!\!\!\perp z \mid y$ and $y \perp\!\!\!\perp z \mid x$.

Example 7. Here is a bigger example with more conditional independencies:



As before, we can identify many of them using the factorization properties. In particular, we can read off some conditional independencies by the following general procedure implied by our earlier analysis in these notes:

1. As used earlier, choose a topological reordering of the nodes, i.e., a reordering where any node i comes after all of its parents.

2. Let ν_i be the set of nodes that are not parents of i, i.e. $\pi_i \cap \nu_i = \varnothing$, but they appear in the topological ordering before i.

3. Then the graph implies the conditional independence $\mathsf{x}_i \perp\!\!\!\perp \mathsf{x}_{\nu_i} \mid \mathsf{x}_{\pi_i}$.

Using this approach we can conclude, for example, that this graphical model expresses, among other conditional independencies,

$$\mathsf{x}_5 \perp\!\!\!\perp \{\mathsf{x}_1, \mathsf{x}_2, \mathsf{x}_4\} \mid \mathsf{x}_3. \tag{5}$$

In particular, to get a topological ordering of the nodes, we need to simply swap the labels of nodes 1 and 2 (and the corresponding random variable indices); the remaining nodes can retain their labels. Then the factor $p_{\mathsf{x}_5|\mathsf{x}_3}$ arising out of the edge structure of node 5 implies (5).

Note that there may be many topological orderings for a graph. With the above procedure, different conditional independences can be found via different topological orderings. The resulting conditional independences can be read off from the DAG directly in the following way: Let $nd(i)$ denote the set of non-descendant nodes of node i i.e., $nd(i) = \{j \mid \nexists$ directed path from i to $j\}$.

Definition 1 (Directed local Markov Property)**.** *A distribution p over $\mathcal{X}^{|\mathcal{V}|}$ satisfies the directed local Markov property with respect to a DAG $\mathcal{G}$ if $\mathsf{x}_i \perp\!\!\!\perp \mathsf{x}_{nd(i)\setminus\pi_i} \mid \mathsf{x}_{\pi_i}$ for any $i \in \mathcal{V}$.*

By going over all possible topological orderings of the nodes and comparing the factorization obtained by the chain rule to the factorization obtained from the directed graphical model as in the procedure described above, it follows that directed local Makrov property with respect to a graph $\mathcal{G}$ holds for any distribution p that factors acording to $\mathcal{G}$.

Next, we discuss a more general procedure for testing conditional independence that does not depend on any particular topological ordering, and which can reveal additional conditional independences that are not captured by the directed local Markov property (namely conditional independences where the conditioning set is not the parent set).

2.2 Graph Separation and Conditional Independence

One of the advantages of a good graphical representation for a family of distributions is that it possible to use the tools of graph theory to answer questions about these distributions. From this perspective, the following development is an illustration of this connection in action.

Our goal in this section is to understand the relationship between how subsets of nodes are "separated" by others, and the conditional independence properties of the family of distributions it represents. From Examples 1 and 2 above, we might be

tempted to conclude that two variables are dependent if and only if they're connected by a path that isn't "blocked" by an observed node. However, this criterion fails for Example 3, where x and z are dependent only when the node between them *is* observed. As a result, we require a more elaborate notion of graph separation in which different rules apply to different scenarios, depending on the orientation of the edges involved. This notion of separation provides the correct generalization of Examples 1, 2, and 3, and it turns out that these three examples essentially capture all possible cases.

Directed-separation, or *d-separation* as it is most commonly known, describes the kind of graph separation that characterizes conditional independences in directed graphical models.

To that end, we start with the definition of a *path*. A path connecting a pair of nodes, say a and b, is an ordered collection of nodes $\{p_1, \ldots, p_k\}$ for $k \geq 2$, with $p_1 = a, p_k = b$ such that there is a directed edge between p_ℓ and $p_{\ell+1}$ (in either direction) for $1 \leq \ell < k$. For $k \geq 3$, a node say p_ℓ, on the path is said to have arrows meeting *head-to-head*, if the direction of edges along path are $p_{\ell-1} \to p_\ell$ and $p_{\ell+1} \to p_\ell$ (see the 3rd diagram in Figure 3 for an illustration).

Next, we define the notion of a *blocking* path. To that end, let $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ be disjoint subsets of the set of nodes $\mathcal{V}$. We say that a path between nodes $a \in \mathcal{A}$ and $b \in \mathcal{B}$ is *blocked* with respect to nodes in $\mathcal{C}$ if there exists a node c on the path that satisfies either of the following:

(C1) the node c is in $\mathcal{C}$ and the arrows on the path do not meet *head-to-head* for this node (see Figure 3, diagrams 1 and 2);

(C2) neither node c, nor any of it's descendents, is in $\mathcal{C}$; and the arrows at node c meet *head-to-head* (see Figure 3, diagram 3).

Definition 2 (d-separation)**.** *The set of nodes $\mathcal{A}$ is* d-separated *from the set of nodes $\mathcal{B}$ with respect to $\mathcal{C}$ if every path between nodes $a \in \mathcal{A}$ and $b \in \mathcal{B}$ is* blocked.

We now define a Markov property based on d-separation and then show that any distribution that factorizes according to a DAG $\mathcal{G}$ satisfies the conditional independence relations corresponding to d-separation.

Definition 3 (Directed global Markov property)**.** *A distribution p over $\mathcal{X}^{|\mathcal{V}|}$ satisfies the directed global Markov property with respect to a DAG $\mathcal{G}$ if $x_\mathcal{A} \perp\!\!\!\perp x_\mathcal{B} \mid x_\mathcal{C}$ for any $x_A, x_\mathcal{B}, x_\mathcal{C} \subset \mathcal{V}$ such that $\mathcal{A}$ and $\mathcal{B}$ are d-separated by $\mathcal{C}$.*

While it is easy to see that the global Markov property implies the local Markov property, the following result shows that the opposite is true as well and that the Markov properties are equivalent to factorization.

Theorem 1. *Let $\mathcal{G}$ be a DAG. The following are equivalent*

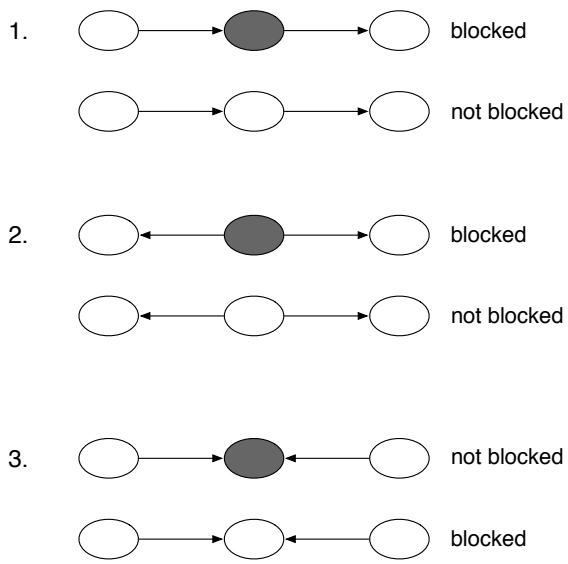


Figure 3: Rules for blocking and non-blocking in the definition of d-separation. When a nodes is in the set $\mathcal{C}$ it is shaded. In rule 3 the "not blocked" case includes the situation with the middle node being in $\mathcal{C}$ or an *ancestor* of a node in $\mathcal{C}$.

(a) p factorizes according to $\mathcal{G}$;

(b) p satisfies the directed global Markov property with respect to $\mathcal{G}$;

(c) p satisfies the directed local Markov property with respect to $\mathcal{G}$.

A useful intuition for the implication of (a)⇒ (b) in Theorem 1 comes from our three-node examples depicted in Fig. 3. In particular, we can think of paths through a middle node being "blocked" or not depending on both the direction of the edges incident to it, and whether that node is in the conditioning set or not.

More specifically, we saw that when we have the edge structure of Example 1, the variables associated with the end nodes are independent (corresponding to a blocked path) if conditioned on that associated with the middle node, and not otherwise. Moreover, the same holds for the common-cause structure of Example 2. However, the opposite holds for the common-effect (V-structure) of Example 3: the variables associated with the end nodes are independent without any conditioning on the remaining variable, and are otherwise dependent.

Checking for d-separation is straightforward to carry out on graphs of modest size, by inspection. For larger graphs, an efficient algorithm can be designed based on checking d-separation by inspecting all paths in breadth-first search manner. In the literature, this algorithm is known as the *Bayes ball* algorithm.

We now prove Theorem 1. We divide the proof into three parts: (1) if p factorizes according to $\mathcal{G}$, then d-separation implies conditional independence i.e., $(a) \Rightarrow (b)$,

(2) if p satisfies the directed global Markov property with respect to $\mathcal{G}$, then it also satisfies the directed local Markov property with respect to $\mathcal{G}$ i.e., $(b) \Rightarrow (c)$, and (3) if p satisfies the directed local Markov property with respect to $\mathcal{G}$, then it factors according to $\mathcal{G}$ i.e., $(c) \Rightarrow (a)$. The proof of (1) is somewhat more involved while the others are pretty straightforward.

Proof. We start with (1), showing that d-separation implies conditional independence. We use induction on the number of nodes N. For $N = 1$, there is nothing to show. Suppose the statement holds for any directed acyclic graph (DAG) on $N - 1$ nodes. Now consider a DAG $\mathcal{D}$ on N nodes. Consider a topological ordering of all the nodes of $\mathcal{D}$ and let ω be the node with the largest number, i.e. it has no descendent or child. Without loss of generality relabel the nodes so that $\omega = N$. Let $\mathcal{D}'$ be obtained by removing ω from $\mathcal{D}$. It is also a DAG and has $N - 1$ nodes. Crucially, the set of nodes $1, \ldots, N-1$ have a distribution that factorizes according to the DAG $\mathcal{D}'$: due to the factorization implied by $\mathcal{D}$, we can write out the joint distribution $p_{x_1,\ldots,x_{N-1}}$ as the product $\prod_{i=1}^{N-1} p_{x_i|x_{\pi_i}}(x_i|x_{\pi_i})$, and this is exactly the factorization required by $\mathcal{D}'$. Note that this is a consequence of the choice of ω as the last node in the topological ordering. There are three possibilities stated below and for each of them we prove the desired statement using the induction hypothesis:

(a) $\omega \notin \mathcal{A} \cup \mathcal{B} \cup \mathcal{C}$. Consider a pair of nodes $a \in \mathcal{A}$ and $b \in \mathcal{B}$ and any path between them in $\mathcal{D}'$. This path can be viewed as a path in $\mathcal{D}$, and we have by assumption that it is blocked in $\mathcal{D}$ by $\mathcal{C}$. If this blocking is via rule C1 (above Defn 2), then some node $c \in \mathcal{C}$, blocks the path in $\mathcal{D}$ and continues to do so in $\mathcal{D}'$. If the blocking is via rule C2, i.e., there is some node on the path where the arrows meet head-to-head and neither it nor any of its descendants $\mathcal{D}$ are in $\mathcal{C}$, we again see that this continues to hold in $\mathcal{D}'$. Thus the path is blocked in $\mathcal{D}'$ and by the induction hypothesis applied to $\mathcal{D}'$, we obtain that $\mathsf{x}_{\mathcal{A}} \perp\!\!\!\perp \mathsf{x}_{\mathcal{B}} \mid \mathsf{x}_{\mathcal{C}}$.

(b) $\omega \in \mathcal{A}$ (the same argument applies to $\omega \in \mathcal{B}$). Let $\mathcal{A}' = \mathcal{A} \backslash \omega$. Note that $\omega \notin \mathcal{A}' \cup \mathcal{B} \cup \mathcal{C}$, so the previous paragraph shows that $\mathcal{A}'$ is d-separated from $\mathcal{B}$ with respect to $\mathcal{C}$ in $\mathcal{D}'$ and therefore $\mathsf{x}_{\mathcal{A}'} \perp\!\!\!\perp \mathsf{x}_{\mathcal{B}} \mid \mathsf{x}_{\mathcal{C}}$.

Observe that no parent of ω is in set $\mathcal{B}$: Otherwise, we have an edge between ω and a node in $\mathcal{B}$, which violates the assumption of d-separation of $\mathcal{A}$ and $\mathcal{B}$ given $\mathcal{C}$. Let $P = \pi_\omega \setminus \mathcal{C}$ be the set of parent nodes of ω in $\mathcal{D}$ that are not in $\mathcal{C}$. We will show that (i) $\mathsf{x}_{\mathcal{A}' \cup P} \perp\!\!\!\perp \mathsf{x}_{\mathcal{B}} \mid \mathsf{x}_{\mathcal{C}}$, and (ii) $\mathsf{x}_{\omega} \perp\!\!\!\perp \mathsf{x}_{\mathcal{B}} \mid \mathsf{x}_{\mathcal{A}' \cup \mathcal{C} \cup P}$. From these and the definition of conditional independence one can obtain $\mathsf{x}_{\mathcal{A}' \cup P \cup \omega} \perp\!\!\!\perp \mathsf{x}_{\mathcal{B}} \mid \mathsf{x}_{\mathcal{C}}$ (be sure to check this for yourself!).

To show (i), it is sufficient to show that $\mathsf{x}_P \perp\!\!\!\perp \mathsf{x}_{\mathcal{B}} \mid \mathsf{x}_{\mathcal{C}}$ and then to combine this with $\mathsf{x}_{\mathcal{A}'} \perp\!\!\!\perp \mathsf{x}_{\mathcal{B}} \mid \mathsf{x}_{\mathcal{C} \cup P}$. We start by showing the latter conditional independence, which is similar to $\mathsf{x}_{\mathcal{A}'} \perp\!\!\!\perp \mathsf{x}_{\mathcal{B}} \mid \mathsf{x}_{\mathcal{C}}$ shown in case (a) above, and but conditioning on $\mathsf{x}_{\mathcal{C} \cup P}$ instead of $\mathsf{x}_{\mathcal{C}}$. Consider a path from $a \in \mathcal{A}'$ to $b \in \mathcal{B}$ in $\mathcal{D}'$. The same

argument as in (a) shows that if it is blocked via rule C1 by $\mathcal{C}$ in $\mathcal{D}$, then it remains so in $\mathcal{D}'$ and also when adding in P to the conditioning set, and the desired statement follows by the inductive hypothesis on $\mathcal{D}'$ as before. The remaining case to consider is that the blocking of the path under consideration is (only) via rule C2, in which case adding nodes P to the conditioning set can plausibly unblock the path between a and b. Suppose for the sake of contradiction that this occurs. This can happen only if a head-to-head node in the path has a node in P as its descendant, and let d be the node that is closest to b among such nodes. Since ω is a child of p in $\mathcal{D}$, we see that there is a path between $b \in B$ and $\omega \in A$ in $\mathcal{D}$ that is unblocked by $\mathcal{C}$, which contradicts our initial assumption. It follows that $\mathsf{x}_{\mathcal{A}'} \perp\!\!\!\perp \mathsf{x}_{\mathcal{B}} \mid \mathsf{x}_{\mathcal{C}\cup P}$.

To show $\mathsf{x}_P \perp\!\!\!\perp \mathsf{x}_{\mathcal{B}} \mid \mathsf{x}_{\mathcal{C}}$, we will argue that P is is d-separated from $\mathcal{B}$ by $\mathcal{C}$ in $\mathcal{D}'$ and use the induction hypothesis for $\mathcal{D}'$. Suppose for the sake of contradiction that this were not the case, i.e., there is a path in $\mathcal{D}'$ from a node $p \in P$ to a node $b \in \mathcal{B}$ that it is not blocked with respect to $\mathcal{C}$. Now extend this path by adding ω as the first node (there is an edge between ω and p, oriented towards ω). Node p has become an internal node on the path. It is not head-to-head since the edge to ω is oriented towards ω, and it does not belong to $\mathcal{C}$ because we are in the case that $\omega \in \mathcal{A}$. We have constructed an unblocked path in $\mathcal{D}$ from $\omega \in \mathcal{A}$ to $b \in \mathcal{B}$ with respect to $\mathcal{C}$, which violates our assumption of d-separation. Therefore, P is d-separated from $\mathcal{B}$ with respect to $\mathcal{C}$ in $\mathcal{D}'$, which proves $\mathsf{x}_P \perp\!\!\!\perp \mathsf{x}_{\mathcal{B}} \mid \mathsf{x}_{\mathcal{C}}$ using the induction hypothesis.

Now we move on to (ii) $\mathsf{x}_\omega \perp\!\!\!\perp \mathsf{x}_{\mathcal{B}} \mid \mathsf{x}_{\mathcal{A}'\cup\mathcal{C}\cup P}$. For this, again note that all parents of ω are a subset of $\mathcal{A}\cup\mathcal{C}\cup P$ and $\mathcal{B}$ are ancestors disjoints from these. Therefore, by the Markov property of the DAG we obtain (ii).

(c) Finally, consider the case $\omega \in \mathcal{C}$. Let $\mathcal{C}' = \mathcal{C}\backslash\omega$. Then, we claim that $\mathcal{A}$ and $\mathcal{B}$ must be d-separated by $\mathcal{C}'$ in $\mathcal{D}'$. Consider an arbitrary path between nodes $a \in \mathcal{A}$ and $b \in \mathcal{B}$, and by assumption it is blocked by $\mathcal{C}$ in $\mathcal{D}$. If the blockage is via rule C1, then it cannot be ω that is responsible, since ω has no descendants and both of the relevant scenarios require the blocking node to have an outgoing edge. If the blockage is via rule C2, then removing ω does not change presence of a head-to-head node with no ancestors in $\mathcal{C}'$ as $\mathcal{C}' \subset \mathcal{C}$. Thus, $\mathcal{A}$ and $\mathcal{B}$ are d-separated by $\mathcal{C}'$ in $\mathcal{D}'$, and by the induction hypothesis we have that $\mathsf{x}_{\mathcal{A}} \perp\!\!\!\perp \mathsf{x}_{\mathcal{B}} \mid \mathsf{x}_{\mathcal{C}'}$.

Now ω must be d-separated from at least one of $\mathcal{A}$ or $\mathcal{B}$ with respect to $\mathcal{C}'$. Suppose not. That is, ω has unblocked path to a node $a \in \mathcal{A}$ *and* has an unblocked path to a node $b \in \mathcal{B}$ with respect to $\mathcal{C}'$. Then, concatenation of these two paths with ω yields an unblocked path (since the two paths meet head-to-head at ω and hence rule (a) does not apply) between a and b with respect to $\mathcal{C}$ in $\mathcal{D}$. This contradicts our initial assumption and hence ω is d-separated either from $\mathcal{A}$ or $\mathcal{B}$ with respect to $\mathcal{C}'$. Without loss of generality, let

ω be d-separated from $\mathcal{B}$ with respect to $\mathcal{C}'$. That is, $\mathcal{A} \cup \omega$ is d-separated from $\mathcal{B}$ with respect to $\mathcal{C}'$ in $\mathcal{D}$. We want to conclude that $x_{\mathcal{A}\cup\omega} \perp\!\!\!\perp x_{\mathcal{B}} \mid x_{\mathcal{C}'}$.

To that end, we argue similarly to case (b). Consider parents P of ω that are not in $\mathcal{C}'$. As before, we can argue that P is d-separated from $\mathcal{B}$ with respect to $\mathcal{C}'$ in $\mathcal{D}'$. Therefore, $x_{\mathcal{A}\cup P} \perp\!\!\!\perp x_{\mathcal{B}} | x_{\mathcal{C}'}$ using the induction hypothesis. Further, as argued earlier, it must be that $P \cap \mathcal{B} = \varnothing$. Therefore, using the property of DAG, we have that $x_\omega \perp\!\!\!\perp x_{\mathcal{B}} | x_{\mathcal{A}\cup P\cup\mathcal{C}'}$ since $\mathcal{B}$ are ancestors of ω excluding it's parents and $\mathcal{A}\cup\mathcal{P}\cup\mathcal{C}'$ contains all parents of ω. Putting everything together, we have $x_{\mathcal{A}\cup\omega} \perp\!\!\!\perp x_{\mathcal{B}} \mid x_{\mathcal{C}'}$. From this, it can be checked from definition of conditional independence that $x_{\mathcal{A}} \perp\!\!\!\perp x_{\mathcal{B}} \mid x_{\mathcal{C}'\cup\omega}$. That is, $x_{\mathcal{A}} \perp\!\!\!\perp x_{\mathcal{B}} \mid x_{\mathcal{C}}$.

We now move on to part (2), showing that the directed global Markov property implies the directed local Markov property. For any $i \in \mathcal{V}$, note that $\{i\}$ is d-separated from the set $nd(i) \setminus \pi_i$ with respect to π_i. Hence by the global Markov property, it holds that $x_i \perp\!\!\!\perp x_{nd(i)\setminus\pi_i} \mid x_{\pi_i}$, which implies the local Markov property.

We now move on to part (3), showing that the directed local Markov property implies factorization according to $\mathcal{G}$. Without loss of generality assume a topological ordering of the nodes. Then using the chain rule, we can write th joint distribution as

$$p_{x_1,\ldots,x_N}(x_1,\ldots,x_N) = \prod_{i=1}^{N} p_{x_i|x_1,\ldots,x_{i-1}}(x_i|x_1,\ldots,x_{i-1}) \tag{6}$$

$$= \prod_{i=1}^{N} p_{x_i|x_{\pi_i}}(x_i|x_{\pi_i}) \tag{7}$$

where the last equality follows from the directed local Markov property. Thus, we have that p factorizes according to the DAG $\mathcal{G}$, which concludes the proof. □

We have seen that while the local Markov property implies the global Markov property, it does not describe all conditional independence relations that are implied by the graphical model. A follow-up question is whether there are conditional independence relations that are implied by the graphical model and go beyond the global Markov property. The following result shows that this is not the case, i.e., the global Markov property cannot be improved.

For this, let $\mathcal{I}(\mathcal{G})$ denote the set of conditional independence relations described by d-separation, and let $\mathcal{I}(p)$ denote the set of conditional independence relations that hold in the distribution p.

Lemma 1. *For any DAG $\mathcal{G}$ there exists a distribution p that factorizes according to $\mathcal{G}$ such that $\mathcal{I}(\mathcal{G}) = \mathcal{I}(p)$.*

Note that the inclusion $\mathcal{I}(\mathcal{G}) \subset \mathcal{I}(p)$ is the global Markov property, which we already showed in the previous theorem is implied by factorization. The reverse

inclusion shows that the global Markov property cannot be improved. For a proof, see Problem Set 1.

Note however, that a related statement does not hold, i.e., there exist distributions p such that there exists no DAG $\mathcal{G}$ such that $\mathcal{I}(\mathcal{G}) = \mathcal{I}(p)$. For a proof, see again Problem Set 1. This is an important observation from the point of view of learning a DAG from data. Given (data from) a distribution p, this motivates finding a minimal DAG that satisfies $\mathcal{I}(\mathcal{G}) \subset \mathcal{I}(p)$. We will return to this in Lecture 4.

We have seen in the examples above that different DAGs can imply the same conditional independence relations. The following result characterizes for what DAGs this is the case.

Lemma 2. *$\mathcal{I}(\mathcal{G}_1) = \mathcal{I}(G_2)$ if and only if the DAGs G_1 and G_2 have the same skeleton (i.e., underlying undirected graph) and the same immoralities (i.e., induced subgraphs of the form $i \to j \leftarrow k$).*

For a proof of this result, see Problem Set 1. Hence $\mathcal{I}(\cdot)$ describes an equivalence relation on the set of DAGs, known as *Markov equivalence*: Two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are in the same *Markov equivalence class* if $\mathcal{I}(\mathcal{G}_1) = \mathcal{I}(G_2)$. Returning to the problem of learning a directed graph from data, this result shows that if all one is given are the conditional independence relations, then the underlying DAG cannot be uniquely identified. In fact, only the Markov equivalence class of the underlying DAG can be identified. We will discuss the problem of causal structure discovery, i.e., the problem of learning the underlying directed graphical model from data on the nodes in Lectures 21 and 22.

2.3 Characterization of DAG's

The following two characterizations are equivalent descriptions of a family of probability distributions:

1. Factorization into a product of conditional probability tables according to the DAG structure

2. Complete list of conditional independencies obtainable by d-separation.

Another way of stating this is that the following two lists are equivalent:

1. List all distributions that factorize according to the graph structure.

2. List all possible distributions, and list all the conditional independencies obtainable by testing for d-separation. Discard the distributions that do not satisfy all the conditional independencies.

2.4 Some Useful Terminology

The following terminology will be useful as our development proceeds.

Definition 4. *A* forest *is a graph where each node has at most one parent.*

Definition 5. *A* connected graph *is one in which there is a path between every pair of nodes.*

Definition 6. *A* tree *is a connected forest.*

Definition 7. *A* polytree *is a singly-connected graph; specifically, there is at most one path from any node to any other node.*

Note that trees are a special case of polytrees. Trees and polytrees will both play an important role in inference.