Appendix C

The language of directed acyclic graphical models

Beginning with Chapter 8, this book makes considerable use of the formalism of DIRECTED ACYCLIC GRAPHICAL MODELS, or BAYESIAN NETWORKS (Bayes nets). In a few pages we cannot to do justice to the diversity of work within this formalism, but this appendix introduces the formalism and a few critical accompanying concepts.

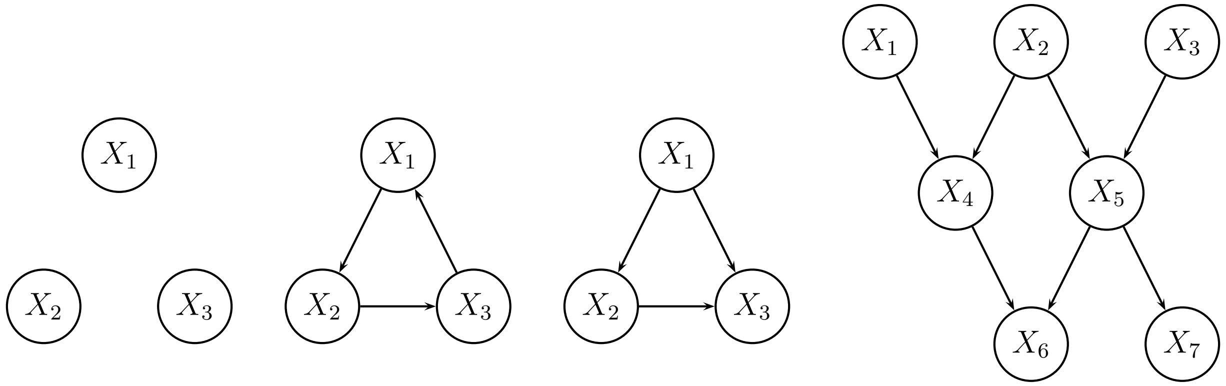
In the general case, graphical models are a set of formalisms for compactly expressing different types of conditional independence relationships between an ENSEMBLE of random variables. A graphical model on an ensemble $X_1, \dots, X_n$ is literally a graph with one node for each random variable X_i, and in which each node may or may not be connected to each other node. The class of *directed* graphical models is those graphical models in which all the inter-node connections have a direction, indicated visually by an arrowhead. The class of directed *acyclic* graphical models, or DAGs (or Bayes nets), is those directed graphical models with no cycles—that is, one can never start at a node X_i and, by traverse edges in the direction of the arrows, get back to X_i. DAGs are the only type of graphical model that you'll see in this book. Figure C.1 shows examples of several different types of graphical models.

C.1 Directed graphical models and their interpretation

The structure of a given DAG encodes what conditional independencies hold among the variables in the ensemble $X_1, \dots, X_n$. First a bit of nomenclature. The PARENTS of a node X_i are the nodes that are pointing directly to it—in Figure C.1d, for example, the parents of X_5 are X_3 and X_4. The ANCESTORS of a node are all the nodes that can be reached from the node by traveling "upstream" on edges in the direction opposite to the arrows—in Figure C.1d, for example, all other nodes are ancestors of X_5, but X_4 has no ancestors.

The set of connections between nodes in a DAG has a formal semantic interpretation whose simplest statement is as follows:

> Any node X_i is conditionally independent of its non-descendents given its parents.



(a) A graphical model with no dependencies

(b) A directed cyclic graphical model

(c) A directed acyclic graphical model (DAG)

(d) A more complex DAG

Figure C.1: Different classes of graphical models

In Figure C.1d, for instance, we have that:

$$X_6 \perp \{X_1, X_2, X_3, X_7\} \mid \{X_4, X_5\}$$

An important proviso is that—recalling from Chapter 2—conditional independencies can disappear with the accrual of new knowledge. In particular, two nodes are *not* conditionally independent of one another given a common descendent. So in Figure C.1d, for example, X_4 has many conditional independencies given only its parents:

$$X_4 \perp \{X_3, X_5, X_7\} \mid \{X_1, X_2\}$$

but two of them go away when its child X_6 is also given:

$$\begin{aligned} X_4 \not\perp X_3 &\mid \{X_1, X_2, X_6\} \\ X_4 \not\perp X_5 &\mid \{X_1, X_2, X_6\} \\ X_4 \perp X_7 &\mid \{X_1, X_2, X_6\} \end{aligned}$$

A more complete statement of conditional independence in DAGs is given in Section C.2.

This statement of conditional independence simplifies the factorization of the joint probability distribution into smaller components. For example, we could simply use the chain rule (Section 2.4) to write the joint probability distribution for Figure C.1d as follows:

$$P(X_{1\ldots7}) = P(X_7|X_{1\ldots6})P(X_6|X_{1\ldots5})P(X_5|X_{1\ldots4})P(X_4|X_{1\ldots3})P(X_3|X_1, X_2)P(X_2|X_1)P(X_1)$$

but we can use the following conditional independencies, which can be read off the connectivity in the graph, to simplify this:

$$\begin{aligned}
P(X_7|X_{1\ldots6}) &= P(X_7|X_5) \qquad \text{(C.1)}\\
P(X_6|X_{1\ldots5}) &= P(X_5|X_4, X_5)\\
P(X_5|X_{1\ldots4}) &= P(X_5|X_3)\\
P(X_4|X_{1\ldots3}) &= P(X_4|X_1, X_2)\\
P(X_3|X_1, X_2) &= P(X_3)\\
P(X_2|X_1) &= P(X_1)
\end{aligned}$$

giving us the following expression for the joint probability distribution

$$P(X_{1\ldots7}) = P(X_7|X_5)P(X_6|X_5, X_4)P(X_5|X_3)P(X_4|X_1, X_2)P(X_3)P(X_2)P(X_1)$$

which is much simpler. These minimal conditional probability distributions seen in (C.1) are the components whose form needs to be specified in order to give a complete probabilistic model of a given domain.

[say something about proper indexing of variables?]

When conducting statistical inference in DAGs, it is often the case that we observe the more "downstream" variables and need to infer some of the more "upstream" variables. The catch is that the conditional probability distributions in the DAG are specified in terms of downstream variables given upstream variables. Conducting inference upstream, then, requires Bayesian inference (the reason that DAGs are often called "Bayes nets"). As an example, in Figure C.1d suppose that we observe (or choose via prior knowledge) all variables except X_4. To draw inferences about X_4, we'd use Bayes rule, targeting the downstream variable X_6 for Bayesian inversion:

$$P(X_4|X_1, X_2, X_3, X_5, X_6, X_7) = \frac{P(X_6|X_{1\ldots5}, X_7)P(X_4|X_{1\ldots3}, X_5, X_7)}{P(X_6|X_{1\ldots3}, X_5, X_7)}$$

We can now apply the conditional independencies in the graph to simplify all the numerator of the right-hand side:

$$= \frac{P(X_6|X_4, X_5)P(X_4|X_1, X_2)}{P(X_6|X_{1\ldots3}, X_5, X_7)}$$

If we wanted to compute the denominator of Equation C.2, we'd need to do it by marginalizing over all possible values x_4 that can be taken by X_4:

$$= \frac{P(X_6|X_4, X_5)P(X_4|X_1, X_2)}{\sum_{x_4} P(X_6|X_{1\ldots5}, X_7)P(X_4|X_{1\ldots3}, X_5, X_7)}$$

Applying the conditional independencies of the graph to the explicit marginalization reveals that X_3 and X_7 can be ignored:

$$= \frac{P(X_6|X_4, X_5)P(X_4|X_1, X_2)}{\sum_{x_4} P(X_6|X_4, X_5)P(X_4|X_1, X_2)}$$

If we now drop the explicit marginalization, we obtain the simplest characterization of Bayesian inference on X_4 available for this graph:

$$P(X_4|X_1, X_2, X_3, X_5, X_6, X_7) = \frac{P(X_6|X_4, X_5)P(X_4|X_1, X_2)}{P(X_6|X_1, X_2, X_5)} \tag{C.2}$$

C.2 Conditional independence in DAGS: d-separation†

We have already briefly described the intuitive picture for when conditional independence holds in a DAG: given its parents, a node is conditionally independent of all of its non-descendents. However, we also saw that such conditional independencies can be broken when more information is conditioned on. In this section, we give the comprehensive criterion by which conditional independence can be assessed in any DAG. This criterion is known as D-SEPARATION (Pearl, 1988, Section 3.3).

Consider two disjoint subsets A and B of nodes in a DAG. A PATH between A and B is simply a sequence of edges that, when taken together, connects some node in A with some node in B (note that this definition doesn't require that the arrows along the path all point in the same direction). Any node on a given path is said to have CONVERGING ARROWS if two edges on the path connect to it and point to it. A node on the path is said to have NON-CONVERGING ARROWS if two edges on the path connect to it, but at least one does not point to it. (Note that the starting and ending nodes on the path are each connected to by only one edge on the path, so are not said to have either converging or non-converging arrows.)

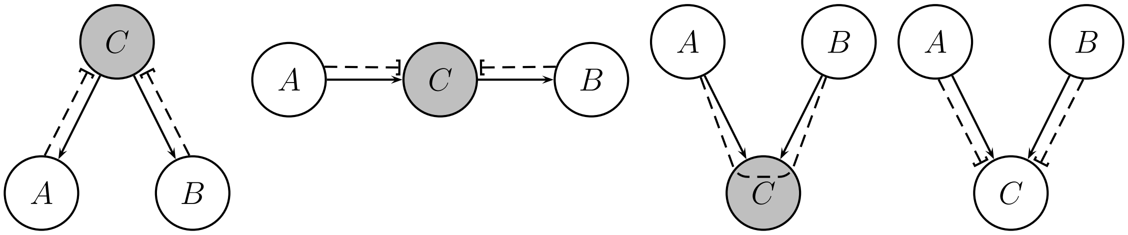
Now consider a third subset C of nodes in the DAG, disjoint from both A and B. C is said to d-separate A and B if for every path between A and B, one of the following two properties holds:

1. there is some node on the path with converging arrows which *is not* in C; or

2. there is some node on the path whose arrows do not converge and which *is* in C.

If C d-separates A and B, then A and B must be conditionally independent given C. If C does not d-separate A and B, then A and B are not in general conditionally independent.

Figure C.2 illustrates the canonical cases of d-separation and of failure of d-separation. In Figures C.2a, we have d-separation: C is on the path between A and B, and it does not have converging arrows. Therefore if C is known, then A and B become conditionally independent:



(a) Common-cause d-separation (b) Intervening d-separation (c) Explaining away: no d-separation (d) D-separation in the absence of knowledge of C

Figure C.2: Examples of d-separation and of failure of d-separation. C blocks the path between A and B in Figures C.2a, C.2b, and C.2d, but not in Figure C.2c.

$A \perp B \mid C$.[1] This configuration is sometimes called "common cause" d-separation: if A and B are the outcomes of two tosses of a possibly unfair coin, then knowing the coin's weighting (C) renders the tosses independent.

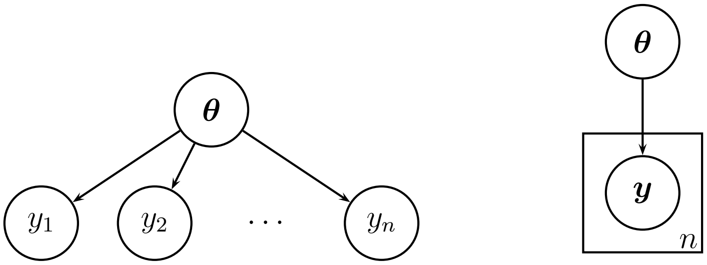
The same holds of Figure C.2b: C is on the path between A and B, and doesn't have converging arrows, so $A \perp B \mid C$. This configuration is often known as "indirect cause": if I know my mother's genome (C), then the respective contents of my genome (B) and my mother's mother's genome (A) become conditionally independent.

In Figures C.2c and C.2d, on the other hand, C is on the path between A and B but it has converging arrows. Therefore C does not d-separate A and B, so $A \not\perp B \mid C$ (Figure C.2c. This configuration is often known as "common effect": a signal (C) indicating whether the tosses of two fair coins (A and B) came up on the same side renders the two tosses conditionally dependent. However, *not* having seen this signal leaves the two tosses independent. In the language of graphical models, d-separation, and conditional independence, we have $A \perp B \mid \emptyset$ (Figure C.2d).

C.3 Plate notation

Since graphical models for structured datasets can get quite complex when the full set of variables, including observations, latent classes, and model parameters, is written out explicitly, it is common to use "plate" notation to succinctly express repetitive structure in the model. The semantics of "plate" notation are simply that any part of a graphical model on a plate with subscript n should be interpreted as being repeated n times, with all the dependencies between nodes external to the plate and nodes internal to the plate preserved and no dependencies between elements on different replicates of the plate. Figure C.3 gives

[1] Technically, since d-separation is a property holding among *sets* of nodes, we should write $\{A\} \perp \{B\} \mid \{C\}$; but for simplicity we drop the braces as a slight abuse of notation when a set consists of exactly one node.



(a) "Flat" model with no hierarchical structure (b)

Figure C.3: Equivalent directed graphical models in plate and no-plate notation

two examples of equivalent models in plate notation and "unfolded" into a plate-free format. Note in particular that in Figure C.3b in the "unfolded" version the variables XXX_i and $YYY_{i'}$ are not connected for $i \neq i'$ [**TODO!**]. Further examples of equivalent non-plate and non-plate models can be found early in Chapter 8.

C.4 Further reading

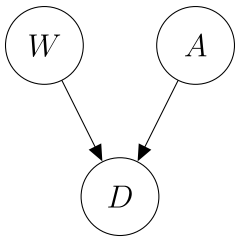
Directed graphical models are an area of considerable research activity. For further reading, some key sources are Pearl (1988, 2000); Jordan (1998); Russell and Norvig (2003, Chapter 14); Bishop (2006, Chapter 8).

Exercise C.1: Conditional independencies in Bayes nets

In each case, state the conditions (what sets of nodes must and/or must not be known) under which the specified node sets will be conditionally independent from one another. If the node sets are always independent or can never be independent, say so.

Example:

W	is the word intended to be spoken a hard word?
A	was the speaker's attention distracted?
D	was a disfluency uttered?

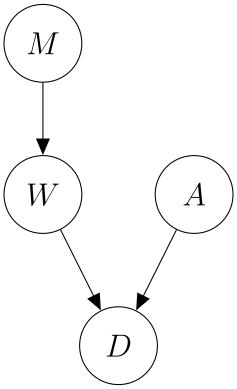


- $\{W\}$ and $\{A\}$ are conditionally independent if and only if D is unknown.
- $\{W\}$ and $\{D\}$ are never conditionally independent.

Examples to solve:

1. A variant of the disfluency model we saw earlier:

M	intended meaning to be conveyed
W	is the word intended to be spoken a hard word?
A	was the speaker's attention distracted?
D	was a disfluency uttered?



(a) $\{W\}$ and $\{A\}$

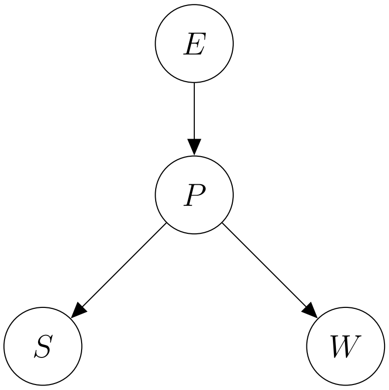
(b) $\{M\}$ and $\{D\}$

(c) $\{M\}$ and $\{A\}$

(d) $\{D\}$ and $\{A\}$

2. The relationship between a child's linguistic environment, his/her true linguistic abilities/proficiency, and measures of his/her proficiency in separate spoken and written tests

E	a child's linguistic environment
P	the child's linguistic proficiency (number of words known, etc.)
S	the child's performance on a spoken language proficiency test
W	the child's performance on a written language proficiency test



(a) $\{S\}$ and $\{W\}$

(b) $\{E\}$ and $\{P\}$

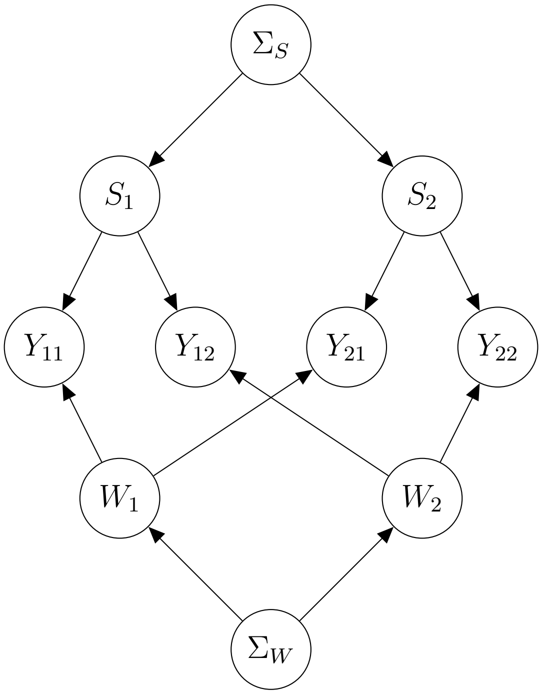
(c) $\{E\}$ and $\{S\}$

(d) $\{E, P\}$ and $\{S\}$

(e) $\{E,P\}$ and $\{S,W\}$

3. Speakers' familiarities (quantified, say, on a scale of 1 to 10) with different words

S_i	the i-th speaker's general vocabulary size
W_j	the j-th word's general difficulty/rarity
Σ_S	the variability in vocabulary sizes across speakers
Σ_W	the variability in difficulties/rarities across words
Y_{ij}	the i-th speaker's familiarity with the j-th word



- $\{\Sigma_S\}$ and $\{\Sigma_W\}$
- $\{Y_{11}\}$ and $\{Y_{22}\}$
- $\{Y_{11}\}$ and $\{Y_{12}\}$
- $\{Y_{11}\}$ and $\{S_2\}$
- $\{W_1\}$ and $\{S_1\}$, supposing that you know Y_{21}
- $\{W_1\}$ and $\{S_1\}$, supposing that you know Y_{22}