2

Graphical Models and Their Applications

2.1 Connecting Models to Data

In Chapter 1, we treated probabilities, graphs, and structural equations as isolated mathematical objects with little to connect them. But the three are, in fact, closely linked. In this chapter, we show that the concept of independence, which in the language of probability is defined by algebraic equalities, can be expressed visually using directed acyclic graphs (DAGs). Further, this graphical representation will allow us to capture the probabilistic information that is embedded in a structural equation model.

The net result is that a researcher who has scientific knowledge in the form of a structural equation model is able to predict patterns of independencies in the data, based solely on the structure of the model's graph, without relying on any quantitative information carried by the equations or by the distributions of the errors. Conversely, it means that observing patterns of independencies in the data enables us to say something about whether a hypothesized model is correct. Ultimately, as we will see in Chapter 3, the structure of the graph, when combined with data, will enable us to predict quantitatively the results of interventions without actually performing them.

2.2 Chains and Forks

We have so far referred to causal models as representations of the "causal story" underlying data. Another way to think of this is that causal models represent the *mechanism* by which data were generated. Causal models are a sort of blueprint of the relevant part of the universe, and we can use them to simulate data from this universe. Given a truly complete causal model for, say, math test scores in high school juniors, and given a complete list of values for every exogenous variable in that model, we could theoretically generate a data point (i.e., a test score) for any individual. Of course, this would necessitate specifying all factors that may have an effect on a student's test score, an unrealistic task. In most cases, we will not have such precise knowledge about a model. We might instead have a probability distribution characterizing the exogenous

Causal Inference in Statistics: A Primer, First Edition. Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell.

Companion Website: www.wiley.com/go/Pearl/Causality

variables, which would allow us to generate a distribution of test scores approximating that of the entire student population and relevant subgroups of students.

Suppose, however, that we do not have even a probabilistically specified causal model, but only a graphical structure of the model. We know which variables are caused by which other variables, but we don't know the strength or nature of the relationships. Even with such limited information, we can discern a great deal about the data set generated by the model. From an unspecified graphical causal model—that is, one in which we know which variables are functions of which others, but not the specific nature of the functions that connect them—we can learn which variables in the data set are independent of each other and which are independent of each other conditional on other variables. These independencies will be true of every data set generated by a causal model with that graphical structure, regardless of the specific functions attached to the SCM.

Consider, for instance, the following three hypothetical SCMs, all of which share the same graphical model. The first SCM represents the causal relationships among a high school's funding in dollars (X), its average SAT score (Y), and its college acceptance rate (Z) for a given year. The second SCM represents the causal relationships among the state of a light switch (X), the state of an associated electrical circuit (Y), and the state of a light bulb (Z). The third SCM concerns the participants in a foot race. It represents causal relationships among the hours that participants work at their jobs each week (X), the hours the participants put into training each week (Y), and the completion time, in minutes, the participants achieve in the race (Z). In all three models, the exogenous variables (U_X, U_Y, U_Z) stand in for any unknown or random effects that may alter the relationship between the endogenous variables. Specifically, in SCMs 2.2.1 and 2.2.3, U_Y and U_Z are additive factors that account for variations among individuals. In SCM 2.2.2, U_Y and U_Z take the value 1 if there is some unobserved abnormality, and 0 if there is none.

SCM 2.2.1 (School Funding, SAT Scores, and College Acceptance)

$$V = \{X, Y, Z\}, U = \{U_X, U_Y, U_Z\}, F = \{f_X, f_Y, f_Z\}$$

$$f_X : X = U_X$$

$$f_Y : Y = \frac{x}{3} + U_Y$$

$$f_Z : Z = \frac{y}{16} + U_Z$$

SCM 2.2.2 (Switch, Circuit, and Light Bulb)

$$V = \{X, Y, Z\}, U = \{U_X, U_Y, U_Z\}, F = \{f_X, f_Y, f_Z\}$$

$$f_X : X = U_X$$

$$f_Y : Y = \begin{cases} \text{Closed IF } (X = \text{Up AND } U_Y = 0) \text{ OR } (X = \text{Down AND } U_Y = 1) \\ \text{Open otherwise} \end{cases}$$

$$f_Z : Z = \begin{cases} \text{On IF } (Y = \text{Closed AND } U_Z = 0) \text{ OR } (Y = \text{Open AND } U_Z = 1) \\ \text{Off otherwise} \end{cases}$$

SCM 2.2.3 (Work Hours, Training, and Race Time)

$$V = \{X, Y, Z\}, U = \{U_X, U_Y, U_Z\}, F = \{f_X, f_Y, f_Z\}$$

$$f_X : X = U_X$$

$$f_Y : Y = 84 - x + U_Y$$

$$f_Z : Z = \frac{100}{y} + U_Z$$

SCMs 2.2.1–2.2.3 share the graphical model shown in Figure 2.1.

SCMs 2.2.1 and 2.2.3 deal with continuous variables; SCM 2.2.2 deals with categorical variables. The relationships between the variables in 2.2.1 are all positive (i.e., the higher the value of the parent variable, the higher the value of the child variable); the correlations between the variables in 2.2.3 are all negative (i.e., the higher the value of the parent variable, the lower the value of the child variable); the correlations between the variables in 2.2.2 are not linear at all, but logical. No two of the SCMs share any functions in common. But because they share a common graphical structure, the data sets generated by all three SCMs must share certain independencies—and we can predict those independencies simply by examining the graphical model in Figure 2.1. The independencies shared by data sets generated by these three SCMs, and the dependencies that are likely shared by all such SCMs, are these:

1. ***Z and Y are likely dependent***
   For some $z, y, P(Z = z|Y = y) \neq P(Z = z)$
2. ***Y and X are likely dependent***
   For some $y, x, P(Y = y|X = x) \neq P(Y = y)$
3. ***Z and X are likely dependent***
   For some $z, x, P(Z = z|X = x) \neq P(Z = z)$
4. ***Z and X are independent, conditional on Y***
   For all $x, y, z, P(Z = z|X = x, Y = y) = P(Z = z|Y = y)$

To understand why these independencies and dependencies hold, let's examine the graphical model. First, we will verify that any two variables with an edge between them are likely dependent. Remember that an arrow from one variable to another indicates that the first variable causes the second—that is, the value of the first variable is part of the function that determines the value of the second. Therefore, the second variable *depends* on the first for its value; there

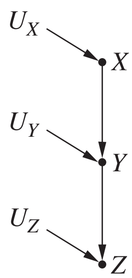

Figure 2.1 The graphical model of SCMs 2.2.1–2.2.3

is some case in which changing the value of the first variable changes the value of the second. That makes it likely that when we examine those variables in the data set, the probability that one variable takes a given value will change, given that we know the value of the other variable. So in a typical causal model, regardless of the specific functions, two variables connected by an edge are dependent. By this reasoning, we can see that in SCMs 2.2.1–2.2.3, Z and Y are likely dependent, and Y and X are likely dependent.[1]

From these two facts, we can conclude that Z and X are *likely* dependent. If Z depends on Y for its value, and Y depends on X for its value, then Z likely depends on X for its value. There are pathological cases in which this is not true. Consider, for example, the following SCM, which also has the graph in Figure 2.1.

SCM 2.2.4 (Pathological Case of Intransitive Dependence)

$$V = \{X, Y, Z\}, U = \{U_X, U_Y, U_Z\}, F = \{f_X, f_Y, f_Z\}$$

$$f_X : X = U_X$$

$$f_Y : Y = \begin{cases} a & \text{IF } X = 1 \text{ AND } U_Y = 1 \\ b & \text{IF } X = 2 \text{ AND } U_Y = 1 \\ c & \text{IF } U_Y = 2 \end{cases}$$

$$f_Z : Z = \begin{cases} i & \text{IF } Y = c \text{ OR } U_Z = 1 \\ j & \text{IF } U_Z = 2 \end{cases}$$

In this case, no matter what value U_Y and U_Z take, X will have no effect on the value that Z takes; changes in X account for variation in Y between a and b, but Y doesn't affect Z unless it takes the value c. Therefore, X and Z vary independently in this model. We will call cases such as these *intransitive cases*.

However, intransitive cases form only a small number of the cases we will encounter. In most cases, the values of X and Z vary together just as X and Y do, and Y and Z. Therefore, they are likely dependent in the data set.

Now, let's consider point 4: Z and X are independent conditional on Y. Remember that when we condition on Y, we filter the data into groups based on the value of Y. So we compare all the cases where $Y = a$, all the cases where $Y = b$, and so on. Let's assume that we're looking at the cases where $Y = a$. We want to know whether, *in these cases only*, the value of Z is independent of the value of X. Previously, we determined that X and Z are likely dependent, because when the value of X changes, the value of Y likely changes, and when the value of Y changes, the value of Z is likely to change. Now, however, examining *only the cases where* $Y = a$, when we select cases with different values of X, the value of U_Y changes so as to keep Y at $Y = a$, but since Z depends only on Y and U_Z, not on U_Y, the value of Z remains unaltered. So selecting a different value of X doesn't change the value of Z. So, in the case where $Y = a$, X is independent of Z. This is of course true no matter which specific value of Y we condition on. So X is independent of Z, conditional on Y.

This configuration of variables—three nodes and two edges, with one edge directed into and one edge directed out of the middle variable—is called a *chain*. Analogous reasoning to the above tells us that in any graphical model, given any two variables X and Y, if the only path between X and Y is composed entirely of chains, then X and Y are independent conditional

[1] This occurs for example when X and U_Y are fair coins and $Y = 1$ if and only $X = U_Y$. In this case $P(Y = 1 | X = 1) = P(Y = 1 | X = 0) = P(Y = 1) = 1/2$. Such pathological cases require precise numerical probabilities to achieve independence ($P(X = 1) = P(U_X) = 1/2$); they are rare, and can be ignored for all practical purposes.

on any intermediate variable on that path. This independence relation holds regardless of the functions that connect the variables. This gives us a rule:

Rule 1 (Conditional Independence in Chains) *Two variables, X and Y, are conditionally independent given Z, if there is only one unidirectional path between X and Y and Z is any set of variables that intercepts that path.*

An important note: Rule 1 only holds when we assume that the error terms U_X, U_Y, and U_Z are independent of each other. If, for instance, U_X were a cause of U_Y, then conditioning on Y would not necessarily make X and Z independent—because variations in X could still be associated with variations in Y, through their error terms.

Now, consider the graphical model in Figure 2.2. This structure might represent, for example, the causal mechanism that connects a day's temperature in a city in degrees Fahrenheit (X), the number of sales at a local ice cream shop on that day (Y), and the number of violent crimes in the city on that day (Z). Possible functional relationships between these variables are given in SCM 2.2.5. Or the structure might represent, as in SCM 2.2.6, the causal mechanism that connects the state (up or down) of a switch (X), the state (on or off) of one light bulb (Y), and the state (on or off) of a second light bulb (Z). The exogenous variables U_X, U_Y, and U_Z represent other, possibly random, factors that influence the operation of these devices.

SCM 2.2.5 (Temperature, Ice Cream Sales, and Crime)

$$V = \{X, Y, Z\}, U = \{U_X, U_Y, U_Z\}, F = \{f_X, f_Y, f_Z\}$$

$$f_X : X = U_X$$

$$f_Y : Y = 4x + U_Y$$

$$f_Z : Z = \frac{x}{10} + U_Z$$

SCM 2.2.6 (Switch and Two Light Bulbs)

$$V = \{X, Y, Z\}, U = \{U_X, U_Y, U_Z\}, F = \{f_X, f_Y, f_Z\}$$

$$f_X : X = U_X$$

$$f_Y : Y = \begin{cases} \text{On IF } (X = \text{Up AND } U_Y = 0) \text{ OR } (X = \text{Down AND } U_Y = 1) \\ \text{Off otherwise} \end{cases}$$

$$f_Z : Z = \begin{cases} \text{On IF } (X = \text{Up AND } U_Z = 0) \text{ OR } (X = \text{Down AND } U_Z = 1) \\ \text{Off otherwise} \end{cases}$$

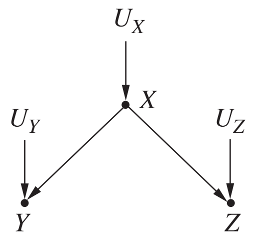

Figure 2.2 The graphical model of SCMs 2.2.5 and 2.2.6

If we assume that the error terms U_X, U_Y, and U_Z are independent, then by examining the graphical model in Figure 2.2, we can determine that SCMs 2.2.5 and 2.2.6 share the following dependencies and independencies:

1. ***X and Y are likely dependent.***
   For some $x, y, P(X = x|Y = y) \neq P(X = x)$
2. ***X and Z are likely dependent.***
   For some $x, z, P(X = x|Z = z) \neq P(X = x)$
3. ***Z and Y are likely dependent.***
   For some $z, y, P(Z = z|Y = y) \neq P(Z = z)$
4. ***Y and Z are independent, conditional on X.***
   For all $x, y, z, P(Y = y|Z = z, X = x) = P(Y = y|X = x)$

Points 1 and 2 follow, once again, from the fact that Y and Z are both directly connected to X by an arrow, so when the value of X changes, the values of both Y and Z likely change. This tells us something further, however: If Y changes when X changes, and Z changes when X changes, then it is likely (though not certain) that Y changes together with Z, and vice versa. Therefore, since a change in the value of Y gives us information about an associated change in the value of Z, Y and Z are likely dependent variables.

Why, then, are Y and Z independent conditional on X? Well, what happens when we condition on X? We filter the data based on the value of X. So now, we're only comparing cases where the value of X is constant. Since X does not change, the values of Y and Z do not change in accordance with it—they change only in response to U_Y and U_Z, which we have assumed to be independent. Therefore, any additional changes in the values of Y and Z must be independent of each other.

This configuration of variables—three nodes, with two arrows emanating from the middle variable—is called a *fork*. The middle variable in a fork is the *common cause* of the other two variables, and of any of their descendants. If two variables share a common cause, and if that common cause is part of the only path between them, then analogous reasoning to the above tells us that these dependencies and conditional independencies are true of those variables. Therefore, we come by another rule:

Rule 2 (Conditional Independence in Forks) *If a variable X is a common cause of variables Y and Z, and there is only one path between Y and Z, then Y and Z are independent conditional on X.*

2.3 Colliders

So far we have looked at two simple configurations of edges and nodes that can occur on a path between two variables: chains and forks. There is a third such configuration that we speak of separately, because it carries with it unique considerations and challenges. The third configuration contains a *collider* node, and it occurs when one node receives edges from two other nodes. The simplest graphical causal model containing a collider is illustrated in Figure 2.3, representing a common effect, Z, of two causes X and Y.

As is the case with every graphical causal model, all SCMs that have Figure 2.3 as their graph share a set of dependencies and independencies that we can determine from the graphical

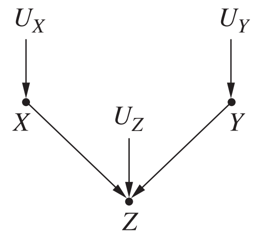

Figure 2.3 A simple collider

model alone. In the case of the model in Figure 2.3, assuming independence of U_X, U_Y, and U_Z, these independencies are as follows:

1. ***X and Z are likely dependent.***
   For some $x, z, P(X = x|Z = z) \neq P(X = x)$
2. ***Y and Z are likely dependent.***
   For some $y, z, P(Y = y|Z = z) \neq P(Y = y)$
3. ***X and Y are independent.***
   For all $x, y, P(X = x|Y = y) = P(X = x)$
4. ***X and Y are likely dependent conditional on Z.***
   For some $x, y, z, P(X = x|Y = y, Z = z) \neq P(X = x|Z = z)$

The truth of the first two points was established in Section 2.2. Point 3 is self-evident; neither X nor Y is a descendant or an ancestor of the other, nor do they depend for their value on the same variable. They respond only to U_X and U_Y, which are assumed independent, so there is no causal mechanism by which variations in the value of X should be associated with variations in the value of Y. This independence also reflects our understanding of how causation operates in time; events that are independent in the present do not become dependent merely because they may have common effects in the future.

Why, then, does point 4 hold? Why would two independent variables suddenly become dependent when we condition on their common effect? To answer this question, we return again to the definition of conditioning as filtering by the value of the conditioning variable. When we condition on Z, we limit our comparisons to cases in which Z takes the same value. But remember that Z depends, for its value, on X and Y. So, when comparing cases where Z takes some value, any change in value of X must be compensated for by a change in the value of Y—otherwise, the value of Z would change as well.

The reasoning behind this attribute of colliders—that conditioning on a collision node produces a dependence between the node's parents—can be difficult to grasp at first. In the most basic situation where $Z = X + Y$, and X and Y are independent variables, we have the following logic: If I tell you that $X = 3$, you learn nothing about the potential value of Y, because the two numbers are independent. On the other hand, if I start by telling you that $Z = 10$, then telling you that $X = 3$ immediately tells you that Y must be 7. Thus, X and Y are dependent, *given that* $Z = 10$.

This phenomenon can be further clarified through a real-life example. For instance, suppose a certain college gives scholarships to two types of students: those with unusual musical talents and those with extraordinary grade point averages. Ordinarily, musical talent and scholastic achievement are independent traits, so, in the population at large, finding a person with musical talent tells us nothing about that person's grades. However, discovering that a person is on a scholarship changes things; knowing that the person lacks musical talent then tells us immediately that he is likely to have high grade point average. Thus, two variables that are marginally independent become dependent upon learning the value of a third variable (scholarship) that is a common effect of the first two.

Let's examine a numerical example. Consider a simultaneous (independent) toss of two fair coins and a bell that rings whenever at least one of the coins lands on heads. Let the outcomes of the two coins be denoted X and Y, respectively, and let Z stand for the state of the bell, with $Z = 1$ representing ringing, and $Z = 0$ representing silence. This mechanism can be represented as a collider as in Figure 2.3, in which the outcomes of the two coins are the parent nodes, and the state of the bell is the collision node.

If we know that Coin 1 landed on heads, it tells us nothing about the outcome of Coin 2, due to their independence. But suppose that we hear the bell ring and then we learn that Coin 1 landed on tails. We now know that Coin 2 must have landed on heads. Similarly, if we assume that we've heard the bell ring, the probability that Coin 1 landed on heads changes if we learn that Coin 2 also landed on heads. This particular change in probability is somewhat subtler than the first case.

To see the latter calculation, consider the initial probabilities as shown in Table 2.1.

We see that

$$P(X = \text{“}Heads\text{”}|Y = \text{“}Heads\text{”}) = P(X = \text{“}Tails\text{”}|Y = \text{“}Tails\text{”}) = \frac{1}{2}$$

That is, X and Y are independent. Now, let's condition on $Z = 1$ and $Z = 0$ (the bell ringing and not ringing). The resulting data subsets are shown in Table 2.2.

By calculating the probabilities in these tables, we obtain

$$P(X = \text{“}Heads\text{”}|Z = 1) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}$$

If we further filter the $Z = 1$ subtable to examine only those cases where $Y =$ *“Heads”*, we get

$$P(X = \text{“}Heads\text{”}|Y = \text{“}Heads\text{”}, Z = 1) = \frac{1}{2}$$

We see that, given $Z = 1$, the probability of $X =$ *“Heads”* changes from $\frac{2}{3}$ to $\frac{1}{2}$ upon learning that $Y =$ *“Heads.”* So, clearly, X and Y are dependent given $Z = 1$. A more pronounced dependence occurs, of course, when the bell does not ring ($Z = 0$), because then we know that both coins must have landed on tails.

Table 2.1 Probability distribution for two flips of a fair coin, with X representing flip one, Y representing flip two, and Z representing a bell that rings if either flip results in heads

X	Y	Z	$P(X, Y, Z)$
Heads	Heads	1	0.25
Heads	Tails	1	0.25
Tails	Heads	1	0.25
Tails	Tails	0	0.25

Table 2.2 Conditional probability distributions for the distribution in Table 2.1. (Top: Distribution conditional on $Z = 1$. Bottom: Distribution conditional on $Z = 0$)

X	Y	$P(X, Y\|Z = 1)$
Heads	Heads	0.333
Heads	Tails	0.333
Tails	Heads	0.333
Tails	Tails	0
X	Y	$Pr(X, Y\|Z = 0)$
Heads	Heads	0
Heads	Tails	0
Tails	Heads	0
Tails	Tails	1

Another example of colliders in action—one that may serve to further illuminate the difficulty that such configurations can present to statisticians—is the Monty Hall Problem, which we first encountered in Section 1.3. At its heart, the Monty Hall Problem reflects the presence of a collider. Your initial choice of door is one parent node; the door behind which the car is placed is the other parent node; and the door Monty opens to reveal a goat is the collision node, causally affected by both the other two variables. The causation here is clear: If you choose Door *A*, and if Door *A* has a goat behind it, Monty is forced to open whichever of the remaining doors that has a goat behind it.

Your initial choice and the location of the car are independent; that's why you initially have a $\frac{1}{3}$ chance of choosing the door with the car behind it. However, as with the two independent coins, conditional on Monty's choice of door, your initial choice and the placement of the prizes are dependent. Though the car may only be behind Door *B* in $\frac{1}{3}$ of cases, it will be behind Door *B* in $\frac{2}{3}$ of cases in which you choose Door *A* and Monty opened Door *C*.

Just as conditioning on a collider makes previously independent variables dependent, so too does conditioning on any descendant of a collider. To see why this is true, let's return to our example of two independent coins and a bell. Suppose we do not hear the bell directly, but instead rely on a witness who is somewhat unreliable; whenever the bell *does not ring*, there is 50% chance that our witness will falsely report that it did. Letting W stand for the witness's report, the causal structure is shown in Figure 2.4, and the probabilities for all combinations of X, Y, and W are shown in Table 2.3.

The reader can easily verify that, based on this table, we have

$$P(X = \text{“}Heads\text{”}|Y = \text{“}Heads\text{”}) = P(X = \text{“}Heads\text{”}) = \frac{1}{2}$$

and

$$P(X = \text{“}Heads\text{”}|W = 1) = (0.25 + 0.25) \div (0.25 + 0.25 + 0.25 + 0.125) = \frac{0.5}{0.875}$$

and

$$P(X = \text{“}Heads\text{”}|Y = \text{“}Heads\text{”}, W = 1) = 0.25 \div (0.25 + 0.25) = 0.5 < \frac{0.5}{0.875}$$

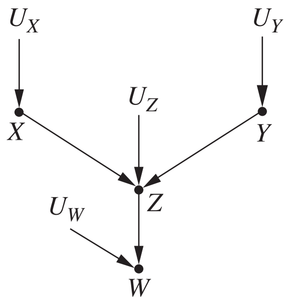

Figure 2.4 A simple collider, Z, with one child, W, representing the scenario from Table 2.3, with X representing one coin flip, Y representing the second coin flip, Z representing a bell that rings if either X or Y is heads, and W representing an unreliable witness who reports on whether or not the bell has rung

Table 2.3 Probability distribution for two flips of a fair coin and a bell that rings if either flip results in heads, with X representing flip one, Y representing flip two, and W representing a witness who, with variable reliability, reports whether or not the bell has rung

X	Y	W	$P(X, Y, W)$
Heads	Heads	1	0.25
Heads	Tails	1	0.25
Tails	Heads	1	0.25
Tails	Tails	1	0.125
Tails	Tails	0	0.125

Thus, X and Y are independent before reading the witness report, but become dependent thereafter.

These considerations lead us to a third rule, in addition to the two we established in Section 2.2.

Rule 3 (Conditional Independence in Colliders) *If a variable Z is the collision node between two variables X and Y, and there is only one path between X and Y, then X and Y are unconditionally independent but are dependent conditional on Z and any descendants of Z.*

Rule 3 is extremely important to the study of causality. In the coming chapters, we will see that it allows us to test whether a causal model could have generated a data set, to discover models from data, and to fully resolve Simpson's Paradox by determining which variables to measure and how to estimate causal effects under confounding.

Remark Inquisitive students may wonder why it is that dependencies associated with conditioning on a collider are so surprising to most people—as in, for example, the Monty Hall example. The reason is that humans tend to associate dependence with causation. Accordingly, they assume (wrongly) that statistical dependence between two variables can only exist if there is a causal mechanism that generates such dependence; that is, either one of the variables causes the other or a third variable causes both. In the case of a collider, they are surprised to find a

$$X \rightarrow R \rightarrow S \rightarrow T \leftarrow U \leftarrow V \rightarrow Y$$

Figure 2.5 A directed graph for demonstrating conditional independence (error terms are not shown explicitly)

$$\begin{tikzcd} X \arrow[r] & R \arrow[r] & S \arrow[r] & T \arrow[d] & U \arrow[l] & V \arrow[l] \arrow[r] & Y \\ & & & P & & & \end{tikzcd}$$

Figure 2.6 A directed graph in which P is a descendant of a collider

dependence that is created in a third way, thus violating the assumption of "no correlation without causation."

Study questions

Study question 2.3.1

(a) List all pairs of variables in Figure 2.5 that are independent conditional on the set $Z = \{R, V\}$.
(b) For each pair of nonadjacent variables in Figure 2.5, give a set of variables that, when conditioned on, renders that pair independent.
(c) List all pairs of variables in Figure 2.6 that are independent conditional on the set $Z = \{R, P\}$.
(d) For each pair of nonadjacent variables in Figure 2.6, give a set of variables that, when conditioned on, renders that pair independent.
(e) Suppose we generate data by the model described in Figure 2.5, and we fit them with the linear equation $Y = a + bX + cZ$. Which of the variables in the model may be chosen for Z so as to guarantee that the slope b would be equal to zero? [*Hint: Recall, a non zero slope implies that Y and X are dependent given Z.*]
(f) Continuing question (e), but now in reference to Figure 2.6, suppose we fit the data with the equation:

$$Y = a + bX + cR + dS + eT + fP$$

which of the coefficients would be zero?

2.4 *d*-separation

Causal models are generally not as simple as the cases we have examined so far. Specifically, it is rare for a graphical model to consist of a single path between variables. In most graphical models, pairs of variables will have multiple possible paths connecting them, and each path will traverse a variety of chains, forks, and colliders. The question remains whether there is a criterion or process that can be applied to a graphical causal model of *any* complexity in order to predict dependencies that are shared by all data sets generated by that graph.

There is, indeed, such a process: *d-separation*, which is built upon the rules established in the previous section. d-separation (the d stands for "directional") allows us to determine, for any pair of nodes, whether the nodes are *d-connected*, meaning there exists a connecting path between them, or *d-separated*, meaning there exists no such path. When we say that a pair of nodes are d-separated, we mean that the variables they represent are definitely independent; when we say that a pair of nodes are *d-connected*, we mean that they are possibly, or most likely, dependent.[2]

Two nodes X and Y are d-separated if every path between them (should any exist) is *blocked*. If even one path between X and Y is unblocked, X and Y are d-connected. The paths between variables can be thought of as pipes, and dependence as the water that flows through them; if even one pipe is unblocked, some water can pass from one place to another, and if a single path is clear, the variables at either end will be dependent. However, a pipe need only be blocked in one place to stop the flow of water through it, and similarly, it takes only one node to block the passage of dependence in an entire path.

There are certain kinds of nodes that can block a path, depending on whether we are performing unconditional or conditional d-separation. If we are not conditioning on any variable, then only colliders can block a path. The reasoning for this is fairly straightforward: as we saw in Section 2.3, unconditional dependence can't pass through a collider. So if every path between two nodes X and Y has a collider in it, then X and Y cannot be unconditionally dependent; they must be marginally independent.

If, however, we are conditioning on a set of nodes Z, then the following kinds of nodes can block a path:

- A collider that is not conditioned on (i.e., not in Z), and that has no descendants in Z.
- A chain or fork whose middle node is in Z.

The reasoning behind these points goes back to what we learned in Sections 2.2 and 2.3. A collider does not allow dependence to flow between its parents, thus blocking the path. But Rule 3 tells us that when we condition on a collider or its descendants, the parent nodes may become dependent. So a collider whose collision node is not in the conditioning set Z would block dependence from passing through a path, but one whose collision node, or its descendants, *is* in the conditioning set would not. Conversely, dependence can pass through noncolliders—chains and forks—but Rules 1 and 2 tell us that when we condition on them, the variables on either end of those paths become independent (when we consider one path at a time). So any noncollision node in the conditioning set would block dependence, whereas one that is not in the conditioning set would allow dependence through.

We are now prepared to give a general definition of d-separation:

Definition 2.4.1 (*d*-separation) *A path p is blocked by a set of nodes Z if and only if*

1. *p contains a chain of nodes $A \rightarrow B \rightarrow C$ or a fork $A \leftarrow B \rightarrow C$ such that the middle node B is in Z (i.e., B is conditioned on), or*
2. *p contains a collider $A \rightarrow B \leftarrow C$ such that the collision node B is not in Z, and no descendant of B is in Z.*

[2] The d-connected variables will be dependent for almost all sets of functions assigned to arrows in the graph, the exception being the sorts of intransitive cases discussed in Section 2.2.

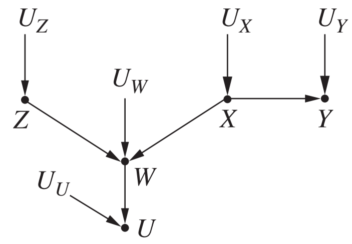

Figure 2.7 A graphical model containing a collider with child and a fork

If Z blocks every path between two nodes X and Y, then X and Y are d-separated, conditional on Z, and thus are independent conditional on Z.

Armed with the tool of d-separation, we can now look at some more complex graphical models and determine which variables in them are independent and dependent, both marginally and conditional on other variables. Let's take, for example, the graphical model in Figure 2.7. This graph might be associated with any number of causal models. The variables might be discrete, continuous, or a mixture of the two; the relationships between them might be linear, exponential, or any of an infinite number of other relations. No matter the model, however, d-separation will always provide the same set of independencies in the data the model generates.

In particular, let's look at the relationship between Z and Y. Using an empty conditioning set, they are d-separated, which tells us that Z and Y are unconditionally independent. Why? Because there is no unblocked path between them. There is only one path between Z and Y, and that path is blocked by a collider ($Z \rightarrow W \leftarrow X$).

But suppose we condition on W. d-separation tells us that Z and Y are d-connected, conditional on W. The reason is that our conditioning set is now $\{W\}$, and since the only path between Z and Y contains a fork (X) that is not in that set, and the only collider (W) on the path is in that set, that path is not blocked. (Remember that conditioning on colliders "unblocks" them.) The same is true if we condition on U, because U is a descendant of a collider along the path between Z and Y.

On the other hand, if we condition on the set $\{W, X\}$, Z and Y remain independent. This time, the path between Z and Y is blocked by the first criterion, rather than the second: There is now a noncollider node (X) on the path that is in the conditioning set. Though W has been unblocked by conditioning, one blocked node is sufficient to block the entire path. Since the only path between Z and Y is blocked by this conditioning set, Z and Y are d-separated conditional on $\{W, X\}$.

Now, consider what happens when we add another path between Z and Y, as in Figure 2.8. Z and Y are now unconditionally dependent. Why? Because there is a path between them ($Z \leftarrow T \rightarrow Y$) that contains no colliders. If we condition on T, however, that path is blocked, and Z and Y become independent again. Conditioning on $\{T, W\}$, on the other hand, makes them d-connected again (conditioning on T blocks the path $Z \leftarrow T \rightarrow Y$, but conditioning on W unblocks the path $Z \rightarrow W \leftarrow X \rightarrow Y$). And if we add X to the conditioning set, making it $\{T, W, X\}$, Z, and Y become independent yet again! In this graph, Z and Y are d-connected (and therefore likely dependent) conditional

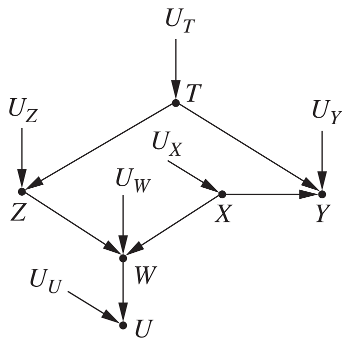

Figure 2.8 The model from Figure 2.7 with an additional forked path between Z and Y

on $W, U, \{W, U\}, \{W, T\}, \{U, T\}, \{W, U, T\}, \{W, X\}, \{U, X\}$, and $\{W, U, X\}$. They are d-separated (and therefore independent) conditional on $T, \{X, T\}, \{W, X, T\}, \{U, X, T\}$, and $\{W, U, X, T\}$. Note that T is in every conditioning set that d-separates Z and Y; that's because T is the only node in a path that unconditionally d-connects Z and Y, so unless it is conditioned on, Z and Y will always be d-connected.

Study questions

Study question 2.4.1

Figure 2.9 below represents a causal graph from which the error terms have been deleted. Assume that all those errors are mutually independent.

(a) For each pair of nonadjacent nodes in this graph, find a set of variables that d-separates that pair. What does this list tell us about independencies in the data?
(b) Repeat question (a) assuming that only variables in the set $\{Z_3, W, X, Z_1\}$ can be measured.
(c) For each pair of nonadjacent nodes in the graph, determine whether they are independent conditional on all other variables.
(d) For every variable V in the graph, find a minimal set of nodes that renders V independent of all other variables in the graph.
(e) Suppose we wish to estimate the value of Y from measurements taken on all other variables in the model. Find the smallest set of variables that would yield as good an estimate of Y as when we measured all variables.
(f) Repeat question (e) assuming that we wish to estimate the value of Z_2.
(g) Suppose we wish to predict the value of Z_2 from measurements of Z_3. Would the quality of our prediction improve if we add measurement of W? Explain.

2.5 Model Testing and Causal Search

The preceding sections demonstrate that causal models have *testable implications* in the data sets they generate. For instance, if we have a graph G that we believe might have generated

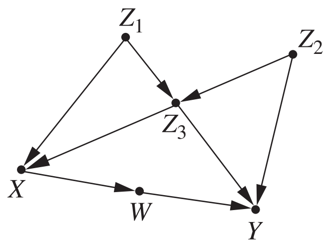

Figure 2.9 A causal graph used in study question 2.4.1, all U terms (not shown) are assumed independent

a data set S, d-separation will tell us which variables in G must be independent conditional on which other variables. Conditional independence is something we can test for using a data set. Suppose we list the d-separation conditions in G, and note that variables A and B must be independent conditional on C. Then, suppose we estimate the probabilities based on S, and discover that the data suggests that A and B are *not* independent conditional on C. We can then reject G as a possible causal model for S.

We can demonstrate it on the causal model of Figure 2.9. Among the many conditional independencies advertised by the model, we find that W and Z_1 are independent given X, because X d-separates W from Z_1. Now suppose we regress W on X and Z_1. Namely, we find the line

$$w = r_X x + r_1 z_1$$

that best fits our data. If it turns out that r_1 is not equal to zero, we know that W depends on Z_1 given X and, consequently, that the model is wrong. [Recall, conditional correlation implies conditional dependence.] Not only do we know that the model is wrong, but we also know where it is wrong; the true model must have a path between W and Z_1 that is not d-separated by X. Finally, this is a theoretical result that holds for all acyclic models with independent errors (Verma and Pearl 1990), and we also know that if every d-separation condition in the model matches a conditional independence in the data, then no further test can refute the model. This means that, for any data set whatsoever, one can always find a set of functions F for the model and an assignment of probabilities to the U terms, so as to generate the data precisely.

There are other methods for testing the fitness of a model. The standard way of evaluating fitness involves a statistical hypothesis test over the entire model, that is, we evaluate how likely it is for the observed samples to have been generated by the hypothesized model, as opposed to sheer chance. However, since the model is not fully specified, we need to first estimate its parameters before evaluating that likelihood. This can be done (approximately) when we assume a linear and Gaussian model (i.e., all functions in the model are linear and all error terms are normally distributed), because, under such assumptions, the joint distribution (also Gaussian) can be expressed succinctly in terms of the model's parameters, and we can then evaluate the likelihood that the observed samples have been generated by the fully parameterized model (Bollen 1989).

There are, however, a number of issues with this procedure. First, if any parameter cannot be estimated, then the joint distribution cannot be estimated, and the model cannot be tested.

As we shall see in Section 3.8.3, this can occur when some of the error terms are correlated or, equivalently, when some of the variables are unobserved. Second, this procedure tests models globally. If we discover that the model is not a good fit to the data, there is no way for us to determine why that is—which edges should be removed or added to improve the fit. Third, when we test a model globally, the number of variables involved may be large, and if there is measurement noise and/or sampling variation associated with each variable, the test will not be reliable.

d-separation presents several advantages over this global testing method. First, it is nonparametric, meaning that it doesn't rely on the specific functions that connect variables; instead, it uses only the graph of the model in question. Second, it tests models locally, rather than globally. This allows us to identify specific areas, where our hypothesized model is flawed, and to repair them, rather than starting from scratch on a whole new model. It also means that if, for whatever reason, we can't identify the coefficient in one area of the model, we can still get some incomplete information about the rest of the model. (As opposed to the first method, in which if we could not estimate one coefficient, we could not test any part of the model.)

If we had a computer, we could test and reject many possible models in this way, eventually whittling down the set of possible models to only a few whose testable implications do not contradict the dependencies present in the data set. It is a set of models, rather than a single model, because some graphs have indistinguishable implications. A set of graphs with indistinguishable implications is called an *equivalence class*. Two graphs G_1 and G_2 are in the same equivalence class if they share a common skeleton—that is, the same edges, regardless of the direction of those edges—and if they share common *v-structures*, that is, colliders whose parents are not adjacent. Any two graphs that satisfy this criterion have identical sets of *d*-separation conditions and, therefore, identical sets of testable implications (Verma and Pearl 1990).

The importance of this result is that it allows us to search a data set for the causal models that could have generated it. Thus, not only can we start with a causal model and generate a data set—but we can also start with a data set, and reason back to a causal model. This is enormously useful, since the object of most data-driven research is exactly to find a model that explains the data.

There are other methods of causal search—including some that rely on the kind of global model testing with which we began the section—but a full investigation of them is beyond the scope of this book. Those interested in learning more about search should refer to (Pearl 2000; Pearl and Verma 1991; Rebane and Pearl 1987; Spirtes and Glymour 1991; Spirtes et al. 1993).

Study questions

Study question 2.5.1

(a) Which of the arrows in Figure 2.9 can be reversed without being detected by any statistical test? [*Hint: Use the criterion for equivalence class.*]
(b) List all graphs that are observationally equivalent to the one in Figure 2.9.
(c) List the arrows in Figure 2.9 whose directionality can be determined from nonexperimental data.

(d) Write down a regression equation for Y such that, if a certain coefficient in that equation is nonzero, the model of Figure 2.9 is wrong.
(e) Repeat question (d) for variable Z_3.
(f) Repeat question (e) assuming the X is not measured.
(g) How many regression equations of the type described in (d) and (e) are needed to ensure that the model is fully tested, namely, that if it passes all these tests, it cannot be refuted by additional tests of these kind. [*Hint: Ensure that you test every vanishing partial reg-ression coefficient that is implied by the product decomposition (1.29).*]

Bibliographical Notes for Chapter 2

The distinction between chains and forks in causal models was made by Simon (1953) and Reichenbach (1956) while the treatment of colliders (or common effect) can be traced back to the English economist Pigou (1911) (see Stigler 1999, pp. 36–41). In epidemiology, colliders came to be associated with "Selection bias" or "Berkson paradox" (Berkson 1946) while in artificial intelligence it came to be known as the "explaining away effect" (Kim and Pearl 1983). The rule of *d*-separation for determining conditional independence by graphs (Definition 2.4.1) was introduced in Pearl (1986) and formally proved in Verma and Pearl (1988) using the theory of graphoids (Pearl and Paz 1987). Gentle introductions to *d*-separation are available in Hayduk et al. (2003), Glymour and Greenland (2008), and Pearl (2000, pp. 335–337). Algorithms and software for detecting *d*-separation, as well as finding minimal separating sets are described in Tian et al. (1998), Kyono (2010), and Textor et al. (2011). The advantages of local over global model testing, are discussed in Pearl (2000, pp. 144–145) and further elaborated in Chen and Pearl (2014). Recent applications of *d*-separation include extrapolation across populations (Pearl and Bareinboim 2014), recovering from sampling selection bias (Bareinboim et al. 2014), and handling missing data (Mohan et al. 2013).