# Chapter 3

# Causal Graphs

In his 2009 book titled *Causality: Models, Reasoning, and Inference*, Judea Pearl lays out a powerful and extensive graphical theory of causality.[1] Pearl's work provides a language and a framework for thinking about causality that differs from the potential outcome model presented in Chapter 2. Beyond the alternative terminology and notation, Pearl (2009, section 7.3) shows that the fundamental concepts underlying the potential outcome perspective and his causal graph perspective are equivalent, primarily because they both encode counterfactual causal states to define causality. Yet, each framework has value in elucidating different features of causal analysis, and we will explain these differences in this and subsequent chapters, aiming to convince the reader that these are complementary perspectives on the same fundamental issues.

Even though we have shown in the last chapter that the potential outcome model is simple and has great conceptual value, Pearl has shown that graphs nonetheless provide a direct and powerful way of thinking about full causal systems and the strategies that can be used to estimate the effects within them. Some of the advantage of the causal graph framework is precisely that it permits suppression of what could be a dizzying amount of notation to reference all patterns of potential outcomes for a system of causal relationships. In this sense, Pearl's perspective is a reaffirmation of the utility of graphical models in general, and its appeal to us is similar to the appeal of traditional path diagrams in an earlier era of social science research. Indeed, to readers familiar with path models, the directed graphs that we will present in this chapter will look familiar. There are, however, important and subtle differences between traditional path diagrams and Pearl's usage of directed graphs, which we will explain.

[1] Our references throughout will be to the 2009 second edition of Pearl's original 2000 book. Pearl recognizes that his influential work was developed in dialogue with many others (see Pearl 2009:104–5). In focusing heavily on his development of causal graphs in this book, we regrettably do not give enough credit to the active group of researchers who have participated in developing graphical representations of causality. We recommend that interested readers turn to this broader literature to learn the interconnections between the many complementary perspectives that exist, starting with pieces such as Berzuini, Dawid, and Bernardinelli (2012); Dawid (2002); Dechter, Geffner, and Halpern (2010); Elwert (2013); Glymour, Scheines, and Spirtes (2001); Koller and Friedman (2009); Lauritzen (1996); and Robins and Richardson (2010). For readers from philosophy who wish to know why we do not use neuron diagrams at all, we take the same basic position as Hitchcock (2007).

For our purposes in this book, Pearl's work is important for three different reasons. First, directed graphs encode causal relationships that are completely nonparametric and fully interactive, and as a result when considering feasible analysis strategies it is usually unnecessary to specify the nature of the functional dependence of an outcome $Y$ on the variables that cause it. A graph that includes the causal effects $X \rightarrow Y$ and $W \rightarrow Y$ simply implies that $X$ and $W$ both cause $Y$, without specifying whether their effects are linear, quadratic, interactive, or any other highly nonlinear function in the values of both $X$ and $W$. This generality allows for a model of causality without side assumptions about functional form, such as assumptions of linear additivity. Second, directed graphs show clearly the critical importance of what Pearl labels *collider variables*, which are endogenous variables that must be treated with caution in many research scenarios. Finally, Pearl uses directed graphs to develop transparent and clear justifications for the three basic methods for estimating causal effects that we will feature in this book: conditioning on variables to eliminate noncausal associations by blocking all relevant back-door paths from the causal variable, conditioning on variables that allow for estimation by a mechanism, and using an instrumental variable that is an exogenous shock to the causal variable in order to consistently estimate its effect.

In this chapter, we provide the foundations of the directed graph approach to causal analysis and discuss the relationships between directed graphs and the potential outcome model. In the course of this presentation, we provide a brief introduction to conditioning techniques, but we will hold off on the full presentation of conditioning, as well as Pearl's back-door criterion for conditioning strategies, until Chapter 4.

## 3.1 Identification

To set the stage for our introduction to directed graph representations of causal relationships, it is helpful to define the concept of identification. In Chapter 2, we defined causal effects as contrasts between well-defined potential outcomes and then proceeded to consider some of the conditions under which consistent and unbiased estimators of average causal effects are available. A complementary approach, and the one which motivates the usage of the directed graphs that we will present in this chapter, is to perform an *identification analysis*. Here, the challenges to inference that arise from the finite nature of the available sample are held aside while the analyst considers whether a causal effect could be computed if data on the full population were instead available.

In his 1995 book *Identification Problems in the Social Sciences*, Manski writes,

> it is useful to separate the inferential problem into statistical and identification components. Studies of identification seek to characterize the conclusions that could be drawn if one could use the sampling process to obtain an unlimited number of observations. (Manski 1995:4)

He continues,

> Empirical research must, of course, contend with statistical issues as well as with identification problems. Nevertheless, the two types of inferential difficulties are sufficiently distinct for it to be fruitful to study them

> separately. The study of identification logically comes first. Negative identification findings imply that statistical inference is fruitless: it makes no sense to try to use a sample of finite size to infer something that could not be learned even if a sample of infinite size were available. Positive identification findings imply that one should go on to study the feasibility of statistical inference. (Manski 1995:5)

The two most crucial ingredients for an identification analysis are these:

1. The set of assumptions about causal relationships that the analyst is willing to assert based on theory and past research, including assumptions about relationships between variables that have not been observed but that are related to both the cause and the outcome of interest.

2. The pattern of information that one can assume would be contained in the joint distribution of the variables in the observed dataset *if* all members of the population had been included in the sample that generated the dataset.

As we will begin to explain in this chapter, causal graphs can represent these ingredients effectively and efficiently and are therefore valuable tools for conducting identification analyses.[2] We will use the concept of identification frequently in this book, and we will expand upon this brief introduction as we introduce additional strategies for analysis.

## 3.2 Basic Elements of Causal Graphs

### 3.2.1 Nodes, Edges, Paths, and Cycles

The primary goal when drawing a causal system as a directed graph is to represent explicitly all causes of the outcome of interest, based on past empirical research and assumptions grounded in theory. As we discussed in Section 1.5, each node of a graph represents a random variable and is labeled by a letter, such as $A$, $B$, or $C$. Nodes that are represented by a solid circle $\bullet$ are observed random variables, whereas nodes that are represented by a hollow circle $\circ$ are unobserved random variables.

Causal effects are represented by directed edges $\rightarrow$ (i.e., single-headed arrows), such that an edge from one node to another signifies that the variable at the origin of the

[2] Nonetheless, an identification analysis can be conducted, and typically is within the econometric tradition, without utilizing directed graphs. Consider our discussion of the naive estimator in Chapter 2. The equalities across the expected values of potential outcomes that we stated as Assumptions 1 and 2 in Equations (2.15) and (2.16) are *identification assumptions*. Maintenance of these particular assumptions would allow an analyst to assert that the naive estimator in Equation (2.9) is consistent and unbiased for the true average treatment effect in Equation (2.3), as explained by the decompositions offered there. As such, the average treatment effect is "identified" or is "identifiable" when these assumptions can be maintained, even though an estimate from a finite sample may, because of sampling error, depart substantially from the true average treatment effect in the population. The value of causal graphs, as we will show in this chapter and the next, is that they allow for an efficient representation of full systems of causal relationships, which can be helpful for determining whether identification assumptions are reasonable.

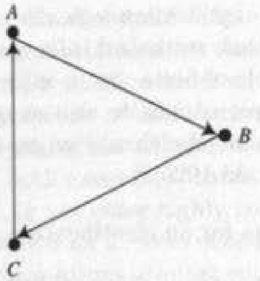


**Figure 3.1** A directed graph that includes a cycle.

directed edge causes the variable at the terminus.[3] These "directed" edges are what give graphs composed of nodes and single-headed arrows the general label of "directed graphs."

A *path* is any sequence of edges pointing in any direction that connects one variable to another. A *directed path* is a path in which all edges point in the same direction. A variable is a *descendant* of another variable if it can be reached by a directed path. All kinship terms are then duly appropriated. Most importantly, for directed paths of length one, as in $A \rightarrow B$, the variable $A$ is the *parent* while the variable $B$ is the *child*.

In this book, we will consider only a subset of directed graphs known as *directed acyclic graphs* or DAGs. For these graphs, no directed paths emanating from a causal variable also terminate at the same causal variable. In other words, no variable can be its own descendant. Figure 3.1 presents a graph that includes a directed path that forms a cycle, and as a result it is not a DAG (even though it is a directed graph because it includes only directed edges). Unlike some graphical models from the past, the prohibition of cycles in DAGs rules out representations of simultaneous causation and feedback loops.[4] All of our statements about graphs from this point onward assume that only acyclic graphs are under consideration.

Under some circumstances it is useful to use a curved and dashed bidirected edge (as in Figures 1.1–1.3) as a shorthand device to indicate that two variables are mutually dependent on one or more unobserved common causes. In this shorthand, the two graphs presented in Figures 3.2(a) and (b) are equivalent. Such shorthand can be helpful in suppressing a complex set of background causal relationships that are irrelevant

[3] In Pearl's framework, each random variable is assumed to have an implicit probability distribution net of the causal effects represented by the directed edges that point to it. This position is equivalent to assuming that background causes of each variable exist that are independent of the causes explicitly represented in the graph by directed edges. We will discuss this assumption in more detail in Section 3.3.2, where we introduce the structural equations that correspond to directed graphs.

[4] As shown in White and Chalak (2009), a broader framework that accommodates cycles is possible (a position acknowledged by Pearl and colleagues for some time, and which has its origins in the interest in reconciling recursive and nonrecursive models since the 1960s). However, the additional details of the broader setup can be daunting, and we recommend that interested readers first carefully consider how much their research questions really do require the full specification of feedback loops that generate cycles. In our experience, most such purportedly necessary loops result from a misplaced unwillingness to consider more tractable empirical research questions that can be confined to shorter spans of analytic time. We do not mean to imply, however, that theory should not attend to such feedback loops, only that most empirical projects can benefit from recursion pragmatism.

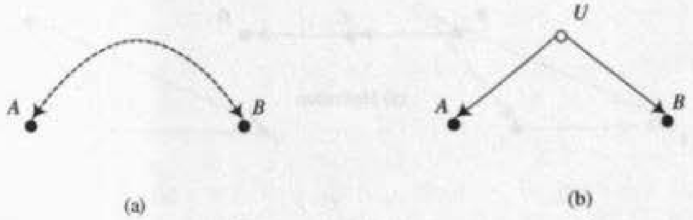


**Figure 3.2** Two representations of the joint dependence of $A$ and $B$ on unobserved common causes.

to the empirical analysis at hand. Nonetheless, these bidirected edges should not be interpreted in any way other than as we have just stated. They are not indicators of mere associations or correlations between the variables that they connect, and they do not signify that either of the two variables has a direct cause on the other one. Rather, they represent an unspecified set of unobserved common causes of the two variables that they connect.

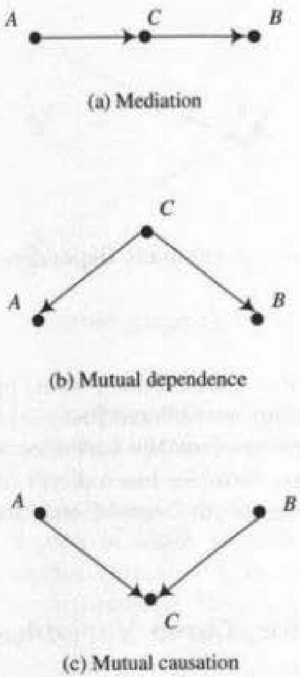
### 3.2.2 Causal Graphs for Three Variables

Figure 3.3 presents the three basic patterns of causal relationships that would be observed for any three variables that are connected to each other by only two directed edges: a chain of mediation, a fork of mutual dependence, and an inverted fork of mutual causation. Pearl's analysis of the first two types of relationship is conventional. For the graph in panel (a), $A$ affects $B$ through $A$'s causal effect on $C$ and $C$'s causal effect on $B$. This type of a causal chain renders the variables $A$ and $B$ unconditionally associated. For the graph in panel (b), $A$ and $B$ are both caused by $C$. Here, $A$ and $B$ are also unconditionally associated, but now it is because they mutually depend on $C$.[5]

For the third graph in panel (c), $A$ and $B$ are again connected by a path through $C$. But now $A$ and $B$ are both causes of $C$. Pearl labels $C$ a *collider variable*. Formally, a variable is a collider along a particular path if it has two arrows pointing directly at it. Figuratively, the causal effects of $A$ and $B$ "collide" with each other at $C$. Collider variables are common in social science applications: Any endogenous variable that has two or more causes is a collider along some path.

A path that is connected by a collider variable does not generate an unconditional association between the variables that cause the collider variable. For the mutual causation graph in panel (c) of Figure 3.3, the path between $A$ and $B$ through $C$ does not generate an unconditional association between $A$ and $B$. As a result, if nothing is known about the value that $C$ takes on, then knowing the value that $A$ takes on yields no information about the value that $B$ takes on. Pearl's language is quite helpful here.

[5] The unconditional associations between $A$ and $B$ for both graphs mean that knowing the value that $A$ takes on gives one some information on the likely value that $B$ takes on. This unconditional association between $A$ and $B$, however, is completely indirect, as neither $A$ nor $B$ has a direct causal effect on each other.



(a) Mediation

(b) Mutual dependence

(c) Mutual causation

**Figure 3.3** Basic patterns of causal relationships for three variables.

The path $A \rightarrow C \leftarrow B$ does not generate an association between $A$ and $B$ because the collider variable $C$ "blocks" the possible causal effects of $A$ and $B$ on each other.

Even though collider variables do not generate unconditional associations between the variables that determine them, we will show in the next chapter that incautious handling of colliders can create conditional dependence that can sabotage a causal analysis. The importance of considering collider variables is a key insight of Pearl's framework, and it is closely related to the familiar concerns of selecting on the dependent variable and conditioning on an endogenous variable (see Elwert and Winship 2014). Before turning to these issues in detail in Chapter 4, we first need to continue our presentation of the basic features of the directed graph approach to causal analysis.

### 3.2.3 A First Look at Confounding and Conditioning

The most common concern of a researcher seeking to estimate a causal effect with observational data is that a causal variable $D$ and an outcome variable $Y$ are determined, in part, by a third variable, $C$. This common but simple scenario is represented by the two graphs in Figure 3.4, where for panel (a) the variable $C$ is observed and for panel (b) it is not.

For both graphs in Figure 3.4, the total association between $D$ and $Y$ is composed of two pieces: (1) the genuine causal effect of $D$ on $Y$, represented by $D \rightarrow Y$, and (2) the common dependence of $D$ and $Y$ on $C$, represented by both $C \rightarrow D$ and $C \rightarrow Y$.

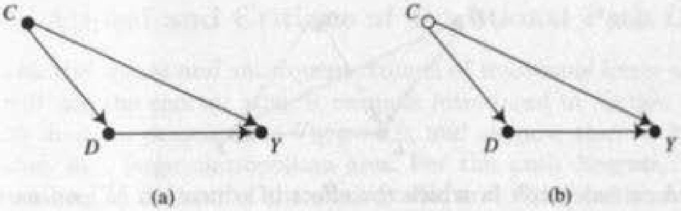


**Figure 3.4** Two graphs in which the causal effect of $D$ on $Y$ is confounded by $C$.

In such cases, it is often said that the causal effect of $D$ on $Y$ is "confounded" by $C$ or, even more simply, that $C$ is a "confounder." Regardless of the label given to $C$, the causal effects $C \to D$ and $C \to Y$ render the total association between $D$ and $Y$ unequal to the causal effect $D \to Y$.

The frequency of basic confounding has established subgroup analysis as perhaps the most common modeling strategy to prosecute causal questions in social science research. Whether referred to as subclassification, stratification, tabular decomposition, or simply adjustment for a third variable, the data are analyzed after "conditioning" on membership in groups defined by values of the confounder variable. Usage of the word "conditioning" is a reference to the " | " operator, already used extensively in Chapter 2, to define conditional expectations (see, in particular, Section 2.7). From a graphical perspective, this modeling strategy is analogous to disconnecting the conditioning variable from all other variables that it points to in the original graph, rewriting the graph without the edges from the conditioning variable for each value for the conditioning variable, analyzing the data that apply to each of these graphs separately, and then combining the results across graphs to form an overall estimate.

For Figure 3.4(a), but not for Figure 3.4(b), consistent and unbiased estimators of the causal effect of $D$ on $Y$ are available (with a large enough dataset generated by a suitably random sample) because the analyst can condition on the observed variable $C$ and eliminate the portion of the association between $D$ and $Y$ that is generated by their common dependence on $C$. We will explain this claim more formally and more generally in Chapter 4, where we introduce Pearl's back-door criterion for the identification of a causal effect.

For now, consider the essential operational data analysis routine. For Figure 3.4(a), the effect of $D$ on $Y$ can be estimated by conditioning on $C$ in two steps: (1) calculate the association between $D$ and $Y$ for each subgroup with $C$ equal to $c$ and then (2) average these $c$-specific associations over the distribution of the values $c$ that the variable $C$ takes on in the sample (which we assume is, again, a large random sample from the population of interest). The resulting weighted average is a consistent and unbiased estimator of the causal effect of $D$ on $Y$ in Pearl's framework, which would be labeled the average treatment effect (ATE) in the potential outcome model introduced in Chapter 2. For Figure 3.4(b), no such data analysis routine is feasible because the analyst has no observed variable $C$ with which to begin.

To make this example more concrete, and to begin to build connections to the examples utilized for Chapter 2, consider the graph in Figure 3.5, where we revisit

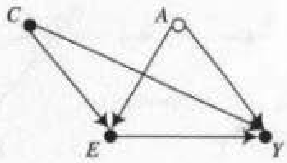


**Figure 3.5** A causal graph in which the effect of education ($E$) on earnings ($Y$) is confounded by observed variables ($C$) and by unobserved ability ($A$).

the research question on ability bias in estimates of the earning returns attributable to completed years of education. Here, education, $E$, is a cause of earnings, $Y$, but this causal effect is confounded by two sources: (1) observed variables, $C$, such as demographic characteristics and family background, and (2) an unobserved variable for ability, $A$.[6] For Figure 3.5, we have moved one step beyond the analysis of the naive estimator in Chapter 2 (see Tables 2.3 and 2.5), because now we are considering the more typical scenario in which the analyst includes in the directed graph some observed variables in $C$ that past research has established almost certainly have causal effects on both years of completed education and subsequent labor market earnings. The ability bias literature asserts that, even if we were to condition on a rich parameterization of all values of the variables in $C$, we still would not be able to eliminate the confounding generated by $A$.

Notice also that we have not asserted that the education variable, $E$, in Figure 3.5 is a two-valued cause, as was the case for the discussion of Table 2.3. Unless specified otherwise, graphs of this type do not place restrictions on the numbers of values taken on by any of their variables. This particular graph, for example, would apply to either the two-valued education variable discussed in Section 2.7.3 or the four-valued education variable introduced in Section 2.9.

## 3.3 Graphs and Structural Equations

Having introduced the basic elements of directed graphs, the next step is to introduce the structural equations that lie beneath them. As noted in the introduction to this chapter, directed graphs encode causal relationships that are completely nonparametric and fully interactive. This generality allows for a model of causality without assumptions about functional form, which is a major advantage over traditional path diagrams. An appreciation for this advantage requires first understanding the constraints imposed by the parametric assumptions that were common in the equations associated with these path diagrams and why many researchers had good reason to object to them.

[6] In the labor economics literature, it would generally be assumed that both $C$ and $A$ depend on a common unobserved cause $U$ that generates an unconditional association between $C$ and $A$. We leave out this dependence in this diagram for simplicity. Its inclusion would not change the fact that the effect of $D$ on $Y$ is not identified.

### 3.3.1 The Appeal and Critique of Traditional Path Diagrams

To explain both the appeal and subsequent critique of traditional linear additive path models, we will use the charter schools example introduced in Section 1.3.2.[7] Consider the path diagram presented in Figure 3.6, and suppose that we have data on all sixth graders in a large metropolitan area. For this path diagram, $Y$ is a standardized test taken at the end of the sixth grade, and $D$ indicates whether or not a student attended a charter school for the past year. The variable $P$ represents an omnibus parental background measure that captures differences in economic standing and other basic dimensions of resources that predict both charter school attendance and school performance. The variable $N$ is neighborhood of residence, and we assume that there are meaningful differences in the extent to which neighborhood environments are conducive to engagement with schooling. Thus, $D$ is the cause of primary interest, $P$ represents individual determinants of $D$ that also have direct causes on the outcome $Y$, and $N$ is a measure of the social context in which the effect of $D$ on $Y$ occurs.[8]

The path diagram presented in Figure 3.6 is associated with two implicit structural equations for the two endogenous variables:

$$D = a_D + b_P P + e_D, \tag{3.1}$$

$$Y = a_Y + b_D D + b_P P + b_N N + e_Y. \tag{3.2}$$

These structural equations are linear and additive in the variables $P$, $D$, and $N$, and each equation has terms, $e_D$ and $e_Y$, that are represented in the path diagram as all other determinants of $D$ and $Y$ other than $P$, $D$, and $N$.[9] The structure of the path diagram and its equations imply that the proper empirical regression specification for $Y$ is the same as Equation (3.2). Under the implicit assumption that $e_Y$ is uncorrelated with $D$, $P$, and $N$, the path-model interpretation of least squares estimates of the coefficients $b_D$, $b_P$, and $b_N$ is that they are consistent and unbiased estimates of the genuine causal effects of $P$, $D$, and $N$ on $Y$.[10]

How would such a path diagram have been presented and then discussed in a typical research methods class in the 1970s (assuming that the charter school research question was under discussion)? Following an introduction to graphical representations of causal relationships via path diagrams, at least one student would invariably ask the instructor:

[7] This section draws on material previously published in Morgan and Winship (2012).

[8] Most path models assumed the existence of unexplained correlations between all "predetermined" or "exogenous" variables, which are $P$ and $N$ for Figure 3.6. In many path diagrams, curved double-headed arrows would be drawn to represent such correlations. To avoid confusion with our usage of bidirected edges throughout this book, we have not added such a double-headed arrow to Figure 3.6.

[9] The standard approach in the early days of path modeling in the social sciences would have been to assume that $e_D$ is uncorrelated with $P$ and that $e_Y$ is uncorrelated with $P$, $D$, and $N$. More complete approaches, as exemplified by Duncan (1975), would not necessarily have maintained such assumptions.

[10] The estimated effects are presumed to be constant across individuals, as specified in Equation (3.2). However, most analysts regarded the estimates as simple average effects across individuals. We will return to this issue in detail when we discuss how regression models do not in general deliver simple average effect estimates (see Chapter 6) and when we then introduce explicit heterogeneity into causal graphs (see Chapter 8).

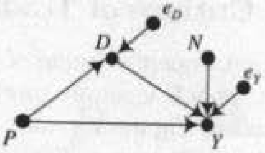


**Figure 3.6** A traditional linear additive path diagram for the effects of parental background ($P$), charter schools ($D$), and neighborhoods ($N$) on test scores ($Y$).

> Can the effect of $D$ on $Y$ vary across $P$? That seems reasonable, since it would seem that the effect of a charter school would depend on family background. Parents with college degrees probably help their kids get more out of school. Actually, now that I think about it, since $N$ captures neighborhood characteristics, don't we think that there are better schools in some neighborhoods? In fact, charter schools are more likely to be established in areas with troubled neighborhood-based schools. And neighborhoods with weaker schools also tend to have stronger deviant subcultures with gangs and such. So the effect of charter schooling probably also depends on the neighborhood in which one lives. How do we represent such variation in effects in the path model?[11]

In response, an instructor would typically explain that one can think of such effects as supplemental arrows from a variable into the middle of another arrow in the path diagram, such that the variable itself modifies the arrow under discussion. Yet, since these sorts of arrows are not formally justified in traditional path diagrams, the instructor would almost surely have then recommended a shift toward a more complex regression specification, such as

$$Y = a_Y + b_C C + b_P P + b_{C\times P}(C \times P) + b_N N + b_{C\times N}(C \times N) + e_Y. \tag{3.3}$$

In this case, the path diagram ceases to represent an underlying set of structural causal relationships and is instead best interpreted as only a simplified reflection of a more specific regression model. After all, the interaction between the effects of $D$ and $P$ on $Y$ (as well as the interaction between the effects of $D$ and $N$ on $Y$) can be estimated with little trouble. One need only calculate effects of interest, for example, by plugging in values for $\hat{b}_C C + \hat{b}_P P + \hat{b}_{C\times P}(C \times P)$, after producing standard regression output from estimation of Equation (3.3). The differences then produced can be imbued with causal interpretations based on the same justification as for Equation (3.2), assuming that no other variables that are common causes of $P$ and $Y$, $D$ and $Y$, or $N$ and $Y$ have been mistakenly omitted from Equation (3.3).

We see two related outcomes of the rise and then demise of traditional linear path models conceived and estimated in this fashion in the social sciences. First, when it became clear that there was no agreed upon way to represent within path diagrams

[11] Were this exchange occurring in the substance of the day, a path model from the status attainment tradition would be the focus of the exchange. The outcome variable $Y$ would be career success, and $D$ would be education. All of the same interactions noted for the charter school case would then apply in this case, although based on different narratives of causation.

the interactions that could be specified in regression equations to capture variability and context, path diagrams came to seem much less useful.[12] Researchers interested in such variability and context may have continued to draw path diagrams on yellow pads in their offices, but rarely did their drawings turn up in published articles.[13] Estimation and reporting became a word and number affair, often with too much of each.

Second, and far more troubling, many scholars apparently chose to retain a linear additive orientation, even while no longer using path diagrams. For some, empirical research could be fruitfully advanced by ignoring the genuine interactive nonlinearity of the real world, in pursuit of a first-approximation, linear pragmatism. This stance might have been an acceptable form of pragmatism if the approximation spirit had carried over to model interpretation. Too frequently it did not, and many causal assertions can be found in the literature based on linear additive models that are overly reductionist.

Overall, the traditional path-modeling literature, and then the more general "age of regression" that we described in Section 1.2.2, opened up quantitative research to the claims of critics that too many practitioners had fallen prey to the belief that linear regression modeling reveals strong causal laws in which variability and context play minor roles. A particularly cogent presentation of this criticism is Abbott's oft-cited "Transcending General Linear Reality" (Abbott 1988). Although its straw-man style is irksome to methodologists who knew of these problems all along, and who urged better practice, it was a reasonable critique of much practice at the time.[14]

### 3.3.2 The Shift to Nonparametric Structural Equations

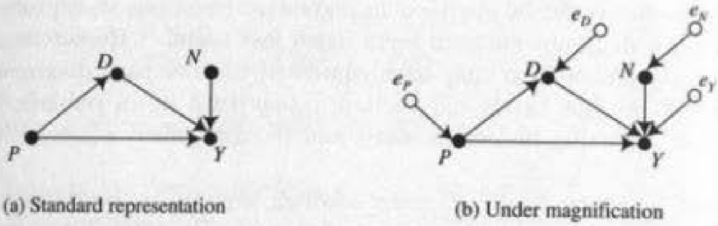
For the same substantive example depicted in the path diagram in Figure 3.6, consider now its representation by a directed graph in the new causal graph tradition. Two variants are offered in Figure 3.7. The standard representation that is depicted in panel (a) is then shown again under "magnification" in panel (b).[15] For the latter, each variable is seen, under close examination, to have its own structural "error" or "disturbance" term: $e_P$, $e_D$, $e_N$, and $e_Y$. Analogous terms are implicitly present in all directed graphs, but they are typically suppressed in their standard representation, as in panel (a), for visual simplicity.

[12] We do not mean to imply that methodologists have not proposed solutions. Bollen (1995) offers the elegant proposal of including functions of other variables as separate entities in diagrams and using sawtooth arrows, $\rightsquigarrow$, in order to represent functional assignment relations. For example, an interactive effect of $C$ and $P$ on $Y$ can be represented altogether by three paths $C \rightarrow Y$, $P \rightarrow Y$, and $(C \times P) \rightarrow Y$. The entity $(C \times P)$ is not a new source of variation with an exogenous component but rather is a deterministic function defined in $C$ and $P$. This functional dependence is signified in the diagram by including both $C \rightsquigarrow (C \times P)$ and $P \rightsquigarrow (C \times P)$.

[13] Freese and Kevern (2013:27) label such causal doodling as "arrow salad."

[14] For similar critiques in sociology in the same time period, see also Lieberson (1985) and Ragin (1987). Abbott (2001), Lieberson and Lynn (2002), and Ragin (2008) offer updates on these earlier critiques. Leamer (1983) is the analog in economics, although written in a completely different style and with alternative suggested remedies.

[15] See Pearl (2009:339) for usage of the word "magnification" to reveal the disturbance/error terms that are implicit in all directed graphs.



(a) Standard representation

(b) Under magnification

**Figure 3.7** Equivalent directed graph representations of the effects of parental background ($P$), charter schools ($D$), and neighborhoods ($N$) on test scores ($Y$).

What are these terms, $e_P$, $e_D$, $e_N$, and $e_Y$? Pearl (2009:27) states that such terms "represent errors (or 'disturbances') due to omitted factors" and are always assumed to be independent of each other and of all other variables in the graph. For Figure 3.7(b), the terms $e_P$, $e_D$, $e_N$, and $e_Y$ represent all causes of $P$, $D$, $N$, and $Y$, respectively, that can be regarded as "idiosyncratic" causes of each variable. They are assumed to be independent of each other and of $P$, $D$, $N$, and $Y$. As such, they can be suppressed in the standard representation of a directed graph, as in Figure 3.7(a).[16]

In general, all directed graphs in the new causal graph tradition must be drawn in sufficient detail so that any such disturbance terms can be pushed into the background. Pearl cautions,

> The disturbance terms represent independent background factors that the investigator chooses not to include in the analysis. If any of these factors is judged to be influencing two or more variables (thus violating the independence assumption), then that factor must enter the analysis as an unmeasured (or latent) variable and be represented in the causal graph as a hollow node. (Pearl 2009:68)

In other words, one must be able to assume that these structural error terms are mutually independent of each other and of all of the other variables in the graph, such that no pair is mutually dependent on a common third variable that has been mistakenly omitted from the graph. If this assumption is dubious, then one must re-draw the graph including hollow nodes for any mistakenly omitted unobserved common causes. These unobserved variables must then be given directed edges that point to the variables that they cause. The new error terms for the re-drawn graph can then be redefined so that they can be pushed into the background (but still rendered visible under magnification).

[16] There is considerable debate over the ontological status of these idiosyncratic causes. Their existence implies to some scholars that causality is fundamentally probabilistic. By our reading, Pearl would maintain, for example, that the variables embedded in $e_P$ are simply implicit structural causes of $P$. Under this interpretation, causality can still be considered a structural, deterministic relation. Freedman (2010, esp. chapter 15) discusses some of the drawbacks for statistical inference of assuming determinism of this form. Although convincing to some degree, his critique does not alter the utility of these sorts of causal graphs for clarifying when causal effects are identified, which is Pearl's primary contribution.

In fact, for the graphs in Figure 3.7, the existing literature on the charter school effect suggests that these graphs are incomplete. Most importantly, it is almost certainly the case that the terms $e_D$ and $e_Y$ in Figure 3.7(b) are mutually dependent on a common cause that has been mistakenly omitted from the graph, analogous to the unobserved ability confounder, $A$, in Figure 3.5. We will therefore extend this graph in Chapter 8, bringing it more into line with assumptions that analysts have been willing to maintain in existing empirical research. For now, we proceed as if the graphs in Figure 3.7 represent the true causal model, accepted hypothetically (but counterfactually!) by a clear consensus of researchers who have studied the charter school effect.

Mindful of this specific definition of the disturbance/error term in directed graphs, and supposing for now that the two causal graphs in Figure 3.7 represent a valid and accepted causal model for the charter school effect, the corresponding structural equations can now be introduced. Unlike path diagrams in traditional form, where the structural equations are linear and additive and correspond only to the the endogenous variables in the diagram (e.g., $D$ and $Y$ in Figure 3.6), for the new directed graph tradition the structural equations are written for all variables and in unrestricted form (i.e., possibly nonlinear and interactive, as explained below). The two graphs in Figure 3.7 have the same set of structural equations:

$$P = f_P(e_P), \tag{3.4}$$

$$D = f_D(P, e_D), \tag{3.5}$$

$$N = f_N(e_N), \tag{3.6}$$

$$Y = f_Y(P, D, N, e_Y). \tag{3.7}$$

Reading from left to right in the graphs and top to bottom in these equations, $P$ is generated as an unspecified function, $f_P(.)$, with $e_P$ as its sole argument. The next three equations then represent analogous unrestricted functions with inputs that represent all causes of the variables on the left-hand sides of the equations. The last is the most elaborate, in that $Y$ is a function of the three observed variables $P$, $D$, and $N$, as well as $e_Y$, all of which transmit their effects on $Y$ via the function $f_Y(.)$.

When we say that $f_P(.)$, $f_D(.)$, $f_N(.)$, or $f_Y(.)$ are unrestricted and have no assumed functional form, we mean that a value is produced for the outcome variable of each equation for every combination of the values in the corresponding function on the right-hand sides of these equations. For example, $Y$ takes on a distinct value for each combination of values, $P = p$, $D = d$, and $N = n$ (typically then with the assumption that values of $e_Y$ are drawn at random from some common distribution that is presumed to have finite moments; see Freedman 2010, chapter 15 for discussion of alternative approaches).[17]

An implication of this flexibility deserves particular emphasis, and preexisting knowledge of traditional path models and their implicit linear additive structural

[17] Restrictions are not typically placed on how the drawn value of the disturbance/error term is then transmitted by $f(.)$ to the outcome. However, because these terms are independent (by definition) of all other inputs in $f(.)$, their realization in the outcome variable does not typically require any assumption, at least when identification is the focus.

equations can hinder full recognition of its importance. Consider the nonparametric structural equation for $Y$ in Equation (3.7). *All* interactions between the effects of $P$, $D$, and $N$ on $Y$ are implicitly permitted by the lack of restrictions placed on $f_Y(.)$. Importantly, this property of the structural equations means that the causal graphs in Figure 3.7 are consistent with all such interactions because the directed edges only signify inclusion in the functions such as $f_Y(.)$. No new arrows, nor any notation of any kind, are needed to represent interactions for more specific parameterizations where, for example, the effect of $D$ on $Y$ varies with the level of $P$ or $N$.

As a result, even though it may feel natural to want to "see" a specific arrow present in the causal graph to represent an interaction effect that corresponds to a cross-product term in a regression equation, one must learn to suppress such a desire. The key point, in considering an analysis that utilizes causal graphs, is to drop regression models from one's mind when thinking about identification issues. Instead, if one must use a data analytic machine to conceptualize how to perform an appropriate empirical analysis of the puzzle under consideration, one should default to simple conditioning, as introduced in the last section.

As we will explain in far greater detail in the next part of the book, if the graphs in Figure 3.7 were the true causal system that generates the effects of charter schools on test scores, the effect of $D$ on $Y$ would be confounded by the observed parental background variables in $P$. In addition, the effect of $D$ on $Y$ may vary across the contexts defined by neighborhoods. With a dataset of sufficient size, the analyst can estimate the effect of $D$ on $Y$ for every combination of the values in $P$ and $N$, adopting a conditioning strategy for identification and estimation. These within-$P$-and-within-$N$ differences are not confounded and represent average causal effects for students within strata defined by $P$ and $N$ (again, under the assumption that the directed graph is a complete representation of the true causal system). To obtain average causal effects for larger groups of individuals, such as those living within a particular type of neighborhood defined by $N$, the analyst can combine the strata-specific estimates by forming a weighted average where the weights are proportional to the sample sizes of all strata that correspond to the type of neighborhood chosen.

In this way, causal analysis guided by the specification of directed graphs is an inherently flexible enterprise. Losing sight of the lack of functional form assumed for causal analysis with unrestricted structural equations may still lead one to fail to transcend "general linear reality." If so, the fault lies with the analyst, not the graph or the possibly highly nonlinear structural equations that it represents.

## 3.4 Causal Graphs and the Potential Outcome Model

What are the connections between the directed graph approach to causal analysis and the potential outcome model introduced in Chapter 2? In short, they are intimately related but very distinct frameworks for considering the same issues, each of which offers unique insight in particular situations. We will return to this basic point repeatedly throughout this book. In this section, we begin to explain their complementary value and then offer a brief examination of the most important formal connection between the two frameworks: how each encodes (potentially counterfactual) causal states.

### 3.4.1 Complementary Value

We already have shown through our presentation in Chapter 2 that the potential outcome model can be understood and utilized without reference to causal graphs. Individual-level potential outcomes allow one to think independently about the observed data as well as what data would have been observed if individuals had experienced alternative causal states. From these simple pieces, the model allows for transparent definitions of causal effects and encourages the analyst to consider individual-level heterogeneity as a first principle.

We also demonstrated in Chapter 2 how potential outcome random variables enable clear definitions of conditional average causal effects and provide ways to usefully decompose sources of inconsistency and bias in estimators – into unaccounted-for baseline differences between individuals and the differential responsiveness of individuals to the cause of interest. Many other advantages of thinking through causal analysis with potential outcomes will be demonstrated in the remaining chapters of this book.

However, we also began to show in Chapter 2 that the transparency of the potential outcome model begins to cloud over when more than two causal states are considered and when important assumptions, such as the stable unit treatment value assumption (SUTVA) (see Section 2.5), are asserted without considerable reflection. The latter is a specific instance of the complications of maintaining the full notational framework of the potential outcome model when many other causal variables of interest begin to enter the scientific problem.

In this chapter, we have shown that the basic elements of causal graphs do not include potential outcome random variables, such $Y^1$ and $Y^0$. The remaining chapters of the book will demonstrate how useful causal graphs can be for empirical research. As we will begin to show in the next chapter, graphs offer a disciplined framework for expressing causal assumptions for entire systems of causal relationships. In many cases, their economy of presentation cannot be matched by potential-outcome-based presentations of the same material, even though equivalent expressions are available. This advantage is especially apparent when the analyst must consider the many causal pathways that may be present in the real applications of observational research. Yet, as we will then also detail later, the price paid for such economy of presentation is then that individuals, and individual-level causal effects, which are so usefully revealed in the potential outcome model, can be covered over by causal graphs, even if the problem is not the causal graphs per se but the shallow interpretations that analysts can too easily attach to them.

Having stated our position on the useful complementarity of these two frameworks, we now lay out the most important point of connection between them. In the next section, we show how each framework encodes causal states in analogous fashion, thereby defining causal effects using counterfactuals.

### 3.4.2 Potential Outcomes and the $do(.)$ Operator

In Chapter 2, we introduced potential outcomes after first discussing the need to clearly define the states of the causal variables of primary interest. We then moved directly to the definition of potential outcomes defined by the instantiation of particular well-defined causal states, focusing for the most part on the canonical potential outcomes $Y^1$ and $Y^0$ for a two-valued cause. Only thereafter did we back out a definition of the corresponding observed variable $Y$, and only by way of Equation (2.2), $Y = DY^1 + (1-D)Y^0$.

When we then introduced directed graphs in this chapter, we skipped right over causal states and potential outcomes. We moved straight from the representation of observed random variables $A$, $B$, and $C$ to discussions of causal effects invoking the same observed variables utilized earlier in Chapter 2, the observed variables $D$ and $Y$. Corresponding potential outcomes, $Y^1$ and $Y^0$, were not incorporated into our presentation. To demonstrate the connections between potential outcomes and directed graphs, we need to introduce how causal states are represented in Pearl's variant of causal analysis.

Pearl introduces causal states in a different way, using the semantics of an ideal experimental intervention and what he labels the $do(.)$ operator. Recall the basic structure of the directed graph in Figure 3.4(a). For this graph, the causal effect of $D$ on $Y$ is represented by $D \to Y$, and the directed edge indicates that $D$ is an input in the function, $f_Y(.)$. The $do(.)$ operator, which we present in this section, is what provides the bridge to quantities such as the ATE, $E[\delta] = E[Y^1 - Y^0]$, defined earlier in Equation (2.3).

For Figure 3.4(a), consider the case where $D$ takes on two values, 0 and 1. For Pearl, there are two regimes by which $D$ takes on values of 0 or 1: pre-intervention and under-intervention. In the pre-intervention regime, the value that $D$ takes on for any given unit in the population is determined by the structural equation $D = f_D(C, e_D)$. In the under-intervention regime, the value that $D$ takes on is set by what Pearl sometimes calls an "ideal experiment" (e.g., Pearl 2009:358) and at other times calls an "atomic intervention" (e.g., Pearl 2009:70). Notationally, this intervention is $do(D=1)$ or $do(D=0)$.[18]

For Pearl, all causal quantities are defined by under-intervention distributions, not pre-intervention distributions (Pearl 2009, definitions 3.2.1 and 7.1.2-5). For $D$ and $Y$ in Figure 3.4(a), the two probability distributions that define causal effects are $Pr[Y|do(D=1)]$ and $Pr[Y|do(D=0)]$, not $Pr[Y|D=1]$ and $Pr[Y|D=0]$. In particular, the average causal effect is $E[Y|do(D=1)] - E[Y|do(D=0)]$, under the assumption that

[18] As we will explain in more detail in the appendix to this chapter, Pearl typically also assumes "modularity," which is the assumption that an intervention on a variable can be carried out without simultaneously altering anything else about the causal relationships encoded in the graph. The atomic intervention is assumed not to generate other interventions on other variables or to open up new causal pathways inconsistent with the structure of the pre-intervention graph. Although modularity is typically assumed, compound interventions are easily accommodated. More difficult, but not impossible, are cases where the assumed intervention generates initially unforeseen counterfactuals. In this case, the pre-intervention causal graph must be redrawn to represent all patterns of unfolding counterfactuals that may follow from prior events.

the individual-level causal effect is defined by the individual-level difference induced by the hypothetical intervention, $[y_i|do(d_i=1)]-[y_i|do(d_i=0)])$.[19]

Under this setup, observable associational quantities, based on $Pr[Y|D]$, do not necessarily equal causal quantities, based on $Pr[Y|do(D)]$. Most importantly, for graphs such as Figure 3.4(a), the associational difference $E[Y|D=1]-E[Y|D=0]$ does not equal the average causal effect defined by the atomic intervention on $D$, $E[Y|do(D=1)]-E[Y|do(D=0)]$. The confounding from $C$ generates additional dependence between $D$ and $Y$ that enters into the average associational difference, $E[Y|D=1]-E[Y|D=0]$, but not the average causal difference, $E[Y|do(D=1)]-E[Y|do(D=0)]$. The reason that $C$ does not enter the under-intervention difference is that $D$ is determined by a hypothetical ideal experiment in this regime. In contrast, in the pre-intervention regime, $D$ is determined by the structural equation $D=f_D(C,e_D)$ that has $C$ as one of its inputs.

Consider now the connection with potential outcomes. The *do*(.) operator is the exact analog to the superscripts given to potential outcomes in order to designate the underlying causal states that define them. In particular, Pearl states that the *do*(.) operator "is a mathematical device that helps us specify explicitly and formally what is held constant, and what is free to vary" (Pearl 2009:358). The semantics that accompany the *do*(.) operator – "ideal experiment" and "atomic intervention" – are Pearl's chosen way to express the idea that all units in the population could be assigned to the causal states in which they are observed (the "factual" ones) or to the causal states in which they are not observed (the "counterfactual" ones) and that causal effects are defined by differences attributable to movement between these alternative states. In this sense, $E[Y^1]-E[Y^0]$ is equivalent to $E[Y|do(D=1)]-E[Y|do(D=0)]$ for the graph in Figure 3.4(a), differing only in how the causal states are signified.[20]

Even though the *do*(.) operator is a crucial piece of Pearl's variant of the directed graph approach to causal analysis, we introduced and discussed causal graphs in this chapter without any reference to it. In fact, we asserted that a directed edge signifies a causal effect, and that representations such as $A \rightarrow B$ are equivalent to the assumption that $A$ is a cause of $B$. Yet, only here, at the end of this chapter, do we note that it is the *do*(.) operator that defines the causal effects that are signified by these directed edges.

As long as one maintains that it is the *do*(.) operator that defines causal effects, not associational contrasts such as $E[Y|D=1]-E[Y|D=0]$, the *do*(.) operator does not need to be represented in the causal graph in any explicit way. The primary purpose of the graph is to encode the full set of causal relationships that one assumes characterize a causal effect of interest, so that those causal relationships can be considered in both the pre-intervention regime and the under-intervention regime. The *do*(.) operator is

[19] Other causal quantities, based on comparisons of $Pr[Y|do(D=1)]$ and $Pr[Y|do(D=0)]$ could also be defined, analogous to choosing something other than the simple difference to define individual-level causal effects in the potential outcome model or selecting something other than the comparison of population expectations of individual-level potential outcomes. In fact, Pearl prefers to avoid selecting any particular comparison, emphasizing that all such comparisons are less general than focusing on the full under-intervention distribution, $Pr[Y|do(D)]$.

[20] When we introduced our chosen notation for the potential outcomes, we also noted that even for the potential outcome model there is a wide variety of notation adopted to signify causal states (see footnote 7 on page 43).

therefore a piece of the underlying structure of the graphical approach, such as the error terms of nonparametric structural equations, which can be brought into the foreground when necessary to explain an identification result.[21]

And, even though we have not done so, the *do*(.) operator can be represented in graphical fashion. For each causal graph of the type presented in this chapter, two types of graphs can be drawn to show the associated under-intervention regime. For "mutilated" graphs, the directed edges that point to the causal variable of interest are deleted, leaving a causal variable with no parents because it is set by the atomic intervention. For "augmented" graphs, the original pre-intervention graph is drawn with an additional "forcing" variable that represents the atomic intervention.

For readers who wish to have an introduction to these additional types of graphs, as well as a slightly more formal presentation of the *do*(.) operator (and its associated modularity condition within an overall definition of a Markovian causal graph), the appendix to this chapter offers an introduction to the primary literature where complete explanations can be found. The appendix also explains why potential outcome variables are rarely depicted in directed graphs. Readers who are uninterested in these details can skip the appendix and will be able to understand all of the material that follows.

## 3.5 Conclusions

In this chapter, we have introduced the directed graph approach to causal analysis. We first introduced the basic elements of causal graphs and then presented the canonical causal graph for a confounded causal effect. We explained the nonparametric nature of these graphs, as represented by the structural equations that assume no functional form for the effects of causes on outcomes. We concluded by noting the equivalence between causal effects defined by directed edges in a causal graph and causal effects defined by potential outcomes, demonstrating that the equivalence lies in their common invocation of what-if causal states grounded in counterfactual reasoning.

Along the way, we have noted that one goal of writing down a causal graph is to represent the set of causal relationships implied by past research and maintained theories. In all remaining parts of this book, we will enrich our discussion of this first goal, when, for example, we discuss whether simple graphs – such as Figure 3.7(a) for the charter school effect – are sufficiently rich to adequately represent the causal relationships that generate effects in real applications.

Another goal of writing down a causal graph is to assess the feasibility of alternative estimation strategies in light of the data that have been observed. Following up on this second goal, in the next part of the book we offer a full presentation of the rationale for conditioning as a causal effect estimation strategy. We then present three related methods for enacting conditioning estimators – matching, regression, and weighted regression – in three separate chapters.

[21] This is similar to the implementation of particular estimation strategies that are motivated by the potential outcome model. In this case, the researcher analyzes data using observed variables $Y$ and $D$. Potential outcomes, $Y^1$ and $Y^0$, are brought out, typically at the beginning of an analysis, to define the causal effects of interest and the assumptions that will be maintained to estimate them.

In subsequent parts of the book, we then further demonstrate how directed graphs can be used to represent heterogeneity and selection on unobserved variables in order to consider how one should analyze causal effects when conditioning on observed variables does not identify the causal effect. We will consider mechanistic and instrumental variable approaches and conclude with estimators that use both pretreatment and posttreatment observations on the outcome variable.

## 3.6 Appendix to Chapter 3: Graphs, Interventions, and Potential Outcomes

In this appendix, we first explain the intervention foundation of Pearl's variant of causal graph methodology, and we then explain why potential outcomes are not commonly depicted as outcome variables in causal graphs. This appendix is written for curious readers who wish to have an introduction to formal details before consulting the primary literature for more complete explanations.

**Graphs That Represent Atomic Interventions.** As noted in Section 3.4.2, the key linkage between the potential outcome model and the directed graph approach to causal analysis is Pearl's concept of an atomic intervention, as represented by the *do*(.) operator. Although the *do*(.) operator is not visible in the standard representation of a causal graph, additional related graphs can be offered to demonstrate the connection more explicitly.

The graph in Figure 3.8(a) is known as an "augmented" graph because it is the pre-intervention causal graph from Figure 3.4(a), augmented with a representation of the atomic intervention on $D$.[22] The augmentation takes the from of a special "forcing" variable, $F_D$, which is placed within □ to denote its special status as an assumed outside force that can produce a hypothetical atomic intervention. Accordingly, this forcing variable takes on three values in this case: $do(D=0)$, $do(D=1)$, and *idle*.

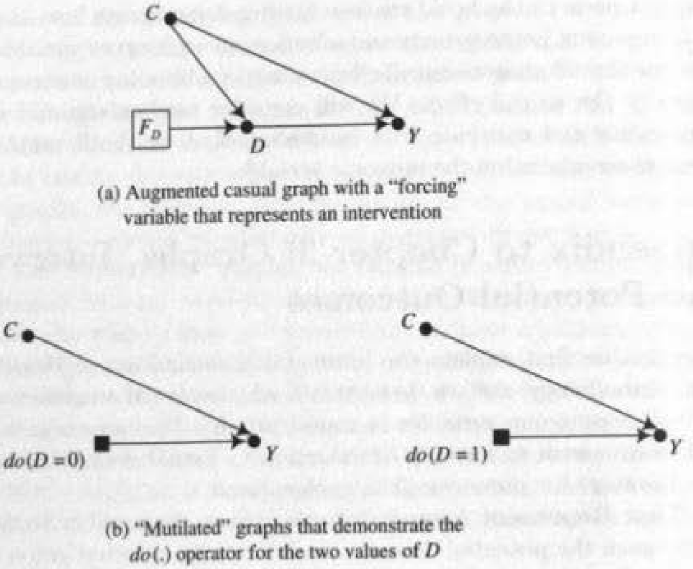
Augmented graphs have accompanying structural equations that represent both the pre-intervention and the under-intervention regimes for the setting of variables subject to the atomic intervention (see Pearl 2009, section 3.2.2). For this graph, the structural equation for Figure 3.4(a), $D=f_D(C,\varepsilon_D)$, is replaced with

$$D=Int[f_D(.),C,\varepsilon_D],$$

where the $Int[.]$ function is defined so that it reduces to 1 if $F_D=do(D=1)$, to 0 if $F_D=do(D=0)$, and to $f_D(C,\varepsilon_D)$ if $F_D=idle$. In other words, the pre-intervention structural equation $f_D(C,\varepsilon_D)$ that generates $D$ becomes the value of $F_D$ when the forcing variable is *idle* because the hypothetical atomic intervention has not been enacted.

Figure 3.8(b) then shows how to think about the graph in Figure 3.8(a) when $F_D=do(D=0)$ and $F_D=do(D=1)$. The two graphs in Figure 3.8(b) are known as the *mutilated* graphs under an atomic intervention. The mutilation refers to the removal of all edges pointing to the variable that is intervened upon. In this case, the

[22] Little would be gained by adding forcing variables for either $C$ or $Y$, since the former has no causes other than $\varepsilon_C$ and the latter causes nothing else specified in the graph. Even so, no formal rules prevent the inclusion of both $F_C$ and $F_Y$ in a fully augmented graph.



(a) Augmented casual graph with a "forcing" variable that represents an intervention

(b) "Mutilated" graphs that demonstrate the *do*(.) operator for the two values of *D*

**Figure 3.8** Two alternative representations of assumed interventions in causal graphs where the effect of $D$ on $Y$ is confounded by $C$.

observed variable for $D$, represented in Figure 3.4(a) by •, is replaced by ■'s in two separate graphs that correspond to the two values of $do(D = 0)$ and $do(D = 1)$ that are determined by the intervention. Because $D$ is no longer determined by $C$, having been set by a hypothetical atomic intervention, there is no directed edge from $C$ to $D$ in either graph in Figure 3.8(b).

Now, compare the standard representation of the canonical confounding graph in Figure 3.4(a) with its augmented variant in Figure 3.8(a). The latter shows the intervention explicitly, and the former leaves it implicit. Dawid (2010:75) claims that Pearl once regarded the inclusion of forcing variables as crucial components of causal graphs. Dawid argues that "he [Pearl], and most of those following him, have [recently] been using only the implicit version, in which the intervention variables $F_V$ are not explicitly included in the diagram, but (to comply with the Pearlian interpretation) the DAG is nevertheless to be interpreted as if they were." Dawid regards the suppression of intervention variables as a "retrograde move" that entails a "loss of transparency."[23] For simple graphs, Dawid is surely correct that forcing variables do increase transparency. For more complex diagrams, forcing variables can be a visual distraction. Accordingly, in this book, we will generally follow Pearl and not offer such representations. As we explain in the next section, however, Pearl builds very precise definitions of causal graphs, which, when kept in mind while interpreting a causal graph in its standard representation, leave little doubt about the crucial role of atomic interventions and the *do*(.) operator.

[23] Dawid prefers a more general influence diagram for causal graphs, resting on top of the decision theoretic approach he has long championed (see Dawid 2002, 2012).

**Criteria for Causal Graphs.** For the canonical confounding graph in Figure 3.4(a), only three observable variables are present: $C$, $D$, and $Y$. No reference to the $do(.)$ operator that defines causal effects is visible in the graph. Yet, we wrote in this chapter that this directed graph is also a "causal graph." The careful reader may have noticed that we have presented other directed graphs in this chapter from which we have withheld the label "causal graph." We now explain.

Pearl has specific requirements for when a directed graph can be anointed a causal graph (Pearl 2009, definitions 1.3.1-3, 2.2.2, and 7.1.1), based on whether the candidate graph and its implied joint probability distribution satisfy "Markovian" conditions (Pearl 2009, theorem 1.4.1). These criteria are difficult to convey in direct form without introducing all aspects of Pearl's approach. We offer the following simplified criteria and strongly encourage readers to consult Pearl (2009) for their more complete and original expression.[24] Accordingly, and at the risk of too much oversimplification, a directed graph can be considered a causal graph by Pearl's definitions if

1. All variables in the graph, $\{V\}$, are observed (other than those variables implicitly included in the error terms, $\{e_V\}$, that are only revealed under magnification).
2. All variables in the graph, $\{V\}$, have error terms, $\{e_V\}$, that are independent of all variables in the graph, $\{V\}$, and that are mutually independent of each other.
3. It is reasonable to assume that each variable in the graph, $V$, can be subjected to a hypothetical intervention, $do(V=v)$, that
   (a) replaces the pre-intervention probability distribution of $V$, $Pr(V)$, with a single intervention value $v$,
   (b) removes the directed edges in the graph that terminate at $V$, and
   (c) changes nothing else in the graph (even though the setting of $V$ to $v$ propagates through the probability distributions of the descendants of $V$ via the directed paths that remain in the graph).

Criteria 1 and 2 should be clear to those familiar with path models; they are stronger versions of the standard linear path model identifying assumptions that "all causal variables are uncorrelated with the error terms of all endogenous variables" and "all error terms on endogenous variables are uncorrelated with each other." Criterion 3 is typically considered to be unique to Pearl's framework, but Pearl himself makes the case that versions of criterion 3 were essential to early motivations of path-modeling techniques and were subsequently forgotten by their inheritors (see Bollen and Pearl 2013).[25]

[24] We will not, for example, discuss the basic requirement that the graph, and its associated structural equations, fully determines the joint probability distribution across all variables depicted in the graph. Using our simplified notation, this criterion is "All variables in the graph, $\{V\}$, have probability distributions, $\{Pr(V)\}$, that are generated by associated structural equations $\{f_V(.)\}$ that are functions only in (a) the variables that point directly to them in the graph (i.e, their "parents") and (b) their own error terms, $e_V$."

[25] In the broader literature on causal graphs beyond Pearl's work, criterion 3 is required only for the subset of the variables in the graph that are immediately relevant for identification of the focal causal effect (see Dawid 2002; Glymour et al. 2001; Robins and Richardson 2010).

Before addressing criterion 3, we should clarify one aspect of criteria 1 and 2. Figure 3.4(a) can be anointed a causal graph because its meets criteria 1 and 2 and because we have implicitly assumed up until now that criterion 3 is met as well. If criterion 1 is not met, but criteria 2 and 3 are, then the directed graph is "semi-Markovian" and typically thought of as if it is a causal graph (because what is observable and what is unobservable is subject to change).

Now consider criterion 3. The lead in – "It is reasonable to assume ..." – is our own writing, but it is consistent with Pearl's presentation of the same material. The key idea is that, by placing such a structure on the graph through a series of finely articulated definitions that are adopted as assumptions, causal effects can be rigorously defined. The payoff to adopting such structure is twofold. First, a directed graph that is a causal graph does not require that forcing variables be displayed to represent the atomic interventions that define all of the causal effects in the graph. Second, if a directed graph is a causal graph that satisfies criteria 1, 2, and 3, then the observed data can be manipulated to deliver estimated causal contrasts that are equal to true causal contrasts defined by the application of the *do*(.) operator separately to all variables in the graph.[26] (Weaker variants of this implication are available, and should be obvious. Even if the full causal graph does not meet these criteria, it is often possible to identify specific causal effects while other effects in the graph remain unidentified.)

We will demonstrate more fully the second implication in the next chapter, where we introduce the back-door criterion for sufficient conditioning to identify a causal effect.[27] To get a handle on this implication now, it may be helpful to consider it in reverse for the simplest estimators we have considered so far in this book. Whenever conditioning estimators, or simple naive estimators, can be used to recover all causal contrasts in a graph from observed associational contrasts, the directed graph is a causal graph. For example, for Figure 3.4(a) a conditioning estimator (described in brief in Section 3.2.3) can be used to generate a consistent estimate of the average causal effect of $D$ on $Y$, defined as $E[Y|do(D=1)] - E[Y|do(D=0)]$.[28] This result is true because the graph in Figure 3.4(a) is a causal graph: All variables are observed; the error terms that are viewable under magnification are defined to be independent of all else in the graph; and we have no reason to believe that it is unreasonable to assume that criterion 3 applies (because we have no reason to believe that intervening on $D$ would alter $C$, etc.).

For another example that also serves as a substantive bridge to the final section of this appendix, recall the graph in Figure 3.5 that includes an unobserved confounder,

[26] Written in this form, this implication holds only for an infinite sample, so as to render sampling error unable to destroy the "equal to" within it. The estimates are therefore best interpreted as consistent estimates of the true causal effects.

[27] More generally, Pearl (2009, section 3.4) presents three rules that can be applied to candidate directed graphs and their associated structural equations to assess whether all effects in the graph can be identified. He characterizes such assessment as the application of "do calculus" with the goal of reducing *do*-defined probability distributions to equivalent probability distributions based only on observable variables (or, more specifically, the joint probability distribution of all variables, as structured by the pre-intervention regime encoded in the graph).

[28] In addition, because the effects of $C$ on both $D$ and $Y$ are unconfounded, assuming criterion 2 holds, the naive estimator (or variants of it if $C$ has more than two values) can be used to consistently estimate these two effects.

$A$, for the effect of education, $E$, on earnings, $Y$. Assuming criteria 2 and 3 are met, this graph is still not fully Markovian, and hence not a causal graph, because criterion 1 is not met. Accordingly, no conditioning estimator can be undertaken with the observed data to generate values that would correspond to $E[Y|do(E=e)]$ for all values $e$ of $E$. Conditioning only on the observed confounder $C$ does not eliminate the confounding by the unobserved variable $A$. However, for Pearl, this graph would be semi-Markovian because observation of $A$ would then render it a fully Markovian causal graph, for which an effective conditioning estimator would then become available. Thus, if we can assume that criteria 2 and 3 are met, then we can state that Figure 3.5 would be a causal graph if $A$ were observed.

**The Absence of Potential Outcome Variables from Causal Graphs.** Imagine a graph that included arrows between $D$ and $Y^0$ and between $D$ and $Y^1$, suggesting purported causal effects such as $D \rightarrow Y^0$ and $D \rightarrow Y^1$. Such causal relationships are inherently nonsensical because $Y^1$ and $Y^0$ are defined in relation to each other, as they jointly determine the observed variable $Y$ in interaction with the causal states indexed by $D$.

Recall the definition offered in Equation (2.2):

$$Y = DY^1 + (1-D)Y^0. \tag{3.8}$$

One trivial way to explain why causal effects such as $D \rightarrow Y^1$ are nonsensical is to take Equation (3.8) and rewrite $Y^1$ in individual realizations as

$$y_i^1 = \frac{y_i - (1-d_i)y_i^0}{d_i}. \tag{3.9}$$

If we think of the supposed causal effect $D \rightarrow Y^1$ as being the theoretical difference in $y_i^1$ that would result from the action of switching $d_i$ from 0 to 1, we can alternatively substitute 0 and 1 into Equation (3.9) and generate the result that the supposed causal effect of $d_i$ on $y_i^1$ is the difference between "undefined" and $y_i$. Doing the same operation for $y_i^0$ delivers a supposed causal effect of $d_i$ on $y_i^0$ as the difference between $y_i$ and "undefined."

In other words, it does not make any sense to seek the causal effect of $D$ on $Y^1$ or of $D$ on $Y^0$, even though we have already offered examples where, empirically and theoretically, there may be good reason to believe that within the population of interest there may be an association between values of $D$ and $Y^1$ and/or between $D$ and $Y^0$ because of the ways individuals enter into the observed causal states (see Section 2.7). From a causal graph perspective, if such associations exist, they are produced by causal relations that connect $D$ and $Y$ but that do not travel through $D \rightarrow Y$. These are exactly the sorts of relationships that we earlier argued, with reference to Equation (2.14), generate inconsistency and bias in the the naive estimator. Individuals with $d_i = 1$ may have higher values, on average, for $y_i^0$ and/or $y_i^1$ than those with $d_i = 0$. If so, such average differences must arise because $D$ and $Y$ both mutually depend on a third variable, such as $C$ in Figure 3.4(a).[29]

[29] Pearl (2009:342) would state such dependence as "$\{Y^0, Y^1\}$ represents the sum total of all exogenous variables, latent as well as observed, which can influence $Y$ through paths that avoid $D$" (with notation changes from the original to match our example). The only point which differs from our

Can one represent the same basic insight using causal graphs? Pearl, in his own writing, has largely chosen not to do so (but see Pearl 2009, section 11.3.2).[30] The broader literature in causal graph methodology shows that it is possible to express the same ideas using directed graphs, although these are not graphs that are to be used in the same way as other graphs in this book. (A recent usage similar to ours in this section can be found in de Luna, Waernbaum, and Richardson (2011). See their citations to precursors.)

Recall Figure 3.5, which was used to bridge our first presentation of conditioning in causal graphs with our earlier presentation of the ability bias example used to introduce the potential outcome model in Chapter 2. In that graph, conditioning on $C$ will not generate a consistent and unbiased estimate of the effect of education, $E$, on earnings, $Y$, because of the presence of the unobserved ability confounder, $A$. However, in Chapter 2 we discussed such confounding by ability instead as inconsistency and bias in the naive estimator. In Section 2.7.3, we defined two types of inconsistency and bias, aided by the potential outcome definitions. Baseline bias in the naive estimator was present when $E[Y^0|D=1] \neq E[Y^0|D=0]$, and treatment effect bias was present when $E[\delta|D=1] \neq E[\delta|D=0]$ (which would be present whenever $E[Y^1|D=1] \neq E[Y^1|D=0]$, assuming no highly unlikely canceling produced by the joint distribution of $Y^0$ and $Y^1$).

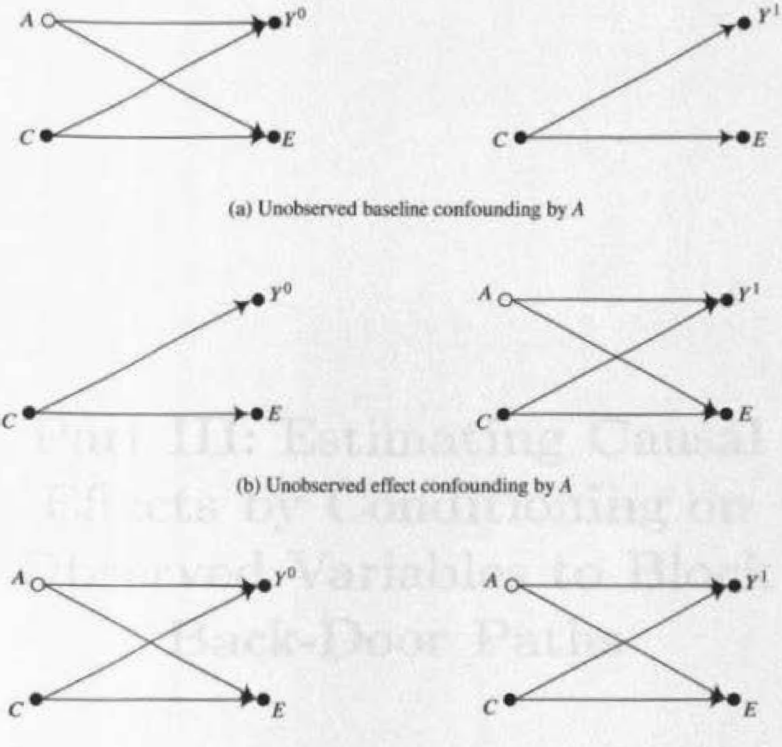
Figure 3.9 presents pairs of directed graphs that express the mutual dependence of potential outcomes and causal variables on exogenous confounders in three different scenarios. Notice first that none of the graphs in Figure 3.9 include either nonsensical causal effects, such as $E \rightarrow Y^0$ or $E \rightarrow Y^1$, or the observed outcome variable earnings, $Y$. Instead, these graphs are drawn solely to represent the mutual dependence of the potential outcomes for earnings, $Y^0$ and $Y^1$, and the focal causal variable, $E$, on observed confounders, $C$, and the unobserved ability confounder, $A$.

Recall that for the representation in Figure 3.5, it was not necessary to stipulate that $E$ took on any particular values because the graph holds under a many-valued version of $E$. For simplicity of representation, and to match our earlier discussion in Section 2.7.3, we will now consider the case where $E$ takes on only two values: 0 for those who do not obtain a bachelor's degree and 1 for those who do.

For Figure 3.9(a), confounding by $C$ generates noncausal associations between $E$ and both $Y^0$ and $Y^1$. Baseline confounding by ability, $A$, generates a noncausal association between only education and $Y^0$. This is the case, noted earlier in Section 2.7.3, where those who have college degrees would have higher earnings than those without

explanation here is that Pearl uses the $\{.,.\}$ notation to emphasize, as in ignorability, that the exogenous variables can structure a function defined in $Y^1$ and $Y^0$, such as the difference between these two. We make the same point in the graphs that follow.

[30] Although Pearl has not devoted a great deal of attention to developing graphical models that include potential outcomes, he and his colleagues have devoted considerable attention to the representation of counterfactuals in causal graphs. Shpitser and Pearl (2007) develop a powerful framework for counterfactual graphs, following on Pearl's earlier work on twinned network graphs; see Pearl (2000:213–14) and citations therein to earlier work; Shpitser (2012a, 2012b) offers particularly clear expositions of these graphs. Counterfactual graphs are beyond the scope of this book, in part because they require a more thorough understanding of Pearl's *do* calculus and full structural causal models than we have space to provide and than is necessary to understand the relevance of his framework for most forms of observational research in the social sciences.



(a) Unobserved baseline confounding by $A$

(b) Unobserved effect confounding by $A$

(c) Unobserved baseline and effect confounding by $A$

**Figure 3.9** Alternative graphs for the joint dependence of a two-valued causal variable for education ($E$) and potential outcomes for earnings ($Y^0$ and $Y^1$) on observed confounders ($C$) and on an unobserved confounder for ability ($A$).

college degrees in the counterfactual state in which they did not in fact obtain college degrees (even after adjustment for $C$). In other words, being "smarter" pays off in the labor market, even if one does not carry on to earn a bachelor's degree.

In contrast, for Figure 3.9(b), confounding by ability generates a noncausal association between education and $Y^1$ but not between education and $Y^0$. This is a form of treatment effect confounding, wherein those who do not obtain college degrees would not have earnings as high as those who do obtain college degrees in the counterfactual state in which they did in fact obtain college degrees (again, after adjustments for $C$). Here, being smarter helps one get more of an earnings payoff from obtaining a bachelor's degree, even though, contrary to Figure 3.9(a), being smarter does not lead to higher earnings in the absence of a college degree.

For Figure 3.9(c), both types of confounding by ability are present, as in the pattern presented earlier in Section 2.7.3. In this case, however, there is little advantage in using the paired graph representation instead of the single directed graph in Figure 3.5. For Figures 3.9(a), (b), and (c), the separate graphs for $Y^0$ and $Y^1$ allow for distinct patterns of confounding, and such clarity may be useful in some situations.