Appendix A

Introduction to causal graphs

Here, we only give a brief introduction to causal graphs. A thorough treatment of graphs can be found in Pearl (2009), Hernán & Robins (2014), or Spirtes, Glymour & Scheines (2001).

A ***causal graph*** consists of ***nodes*** that are linked by ***directed edges*** (arrows). The ***nodes*** represent measured or unmeasured variables like X_1, X_4, Z or Y in Figure 1(1) and the ***arrows*** represent direct causal relations between two variables. The absence of an arrow implies the absence of a direct causal relation. For instance, X_1 causally affects Z and Y, while X_4 has a causal effect on Z only. Since X_1 causally affects Z, X_1 is called a ***parent*** of Z, and Z is called a ***child*** of X_1. For the diagram to be causal all ***common causes*** (also called ***confounders***) of any two variables need to be shown in the graph. Common causes are either explicitly included as nodes with corresponding arrows from the parent node to the children (like X_1, X_2, and X_3 which are common causes of Z and Y) or as ***bidirected dashed arrows*** indicating unknown or unobserved common causes (in Figure 1(1), the bidirected dashed arrows between U_1, U_2, and U_3 indicate an association because of some unknown common causes). Since the dashed arrows are not directed, that is, the causal relation is not explicitly defined, the graph in Figure 1(1) is, strictly speaking, no longer a *causal* graph (we would obtain a causal graph if we would explicitly insert the nodes for the common causes or direct the dashed arrows). A graph that only contains directed edges (arrows) but no cycles (loops) is called a ***directed acyclic graph*** (DAG). Thus, the graph in Figure 1(1) is acyclic but not directed.

A ***path*** between two variables is an unbroken and nonintersecting sequence of arrows that may go along or against the direction of the arrows. For instance, $Z \leftarrow X_1 \rightarrow Y$ and $Z \leftarrow X_1 \leftarrow U_1 \leftarrow\!\rightarrow U_3 \rightarrow X_3 \rightarrow Y$ are two paths connecting Z and Y in Figure 1(1). A ***directed path*** is a path

that goes along the direction of arrows (e.g., $U_4 \rightarrow X_4 \rightarrow Z \rightarrow Y$). If there is a directed path between two variables, say X_4 and Y, then X_4 is said to be an ***ancestor*** or ***predecessor*** of Y, and Y is called a ***descendant*** of X_4. A variable is a ***collider*** (or inverted fork) on a path if two arrows point to or "collide" in it. For instance, Y is a collider on the path $Z \leftarrow X_1 \rightarrow Y \leftarrow X_5$. A path is said to be ***blocked*** either if one conditions on a variable on the path, which must not be a collider, or if the path contains a collider which has not been conditioned on (and no conditioning on a descendant of the collider occurred either). For instance, the path $Z \leftarrow X_1 \rightarrow Y$ can be blocked by conditioning on X_1. Conditioning on X_1 blocks the flow of information along the path $Z \leftarrow X_1 \rightarrow Y$ (conditioning corresponds to considering conditional distributions). Conditioning on a variable is frequently visualized by drawing a box around the conditioning variable: $Z \leftarrow \boxed{X_1} \rightarrow Y$. Path $Z \leftarrow X_1 \rightarrow Y \leftarrow X_5$ is naturally blocked because it contains a collider. There is only information flowing into Y, but no information flows out of Y. If one would condition on Y the naturally blocked path would be unblocked (opened) and the originally independent variables X_1 and X_5 would become dependent conditional on Y. The association resulting from conditioning on a collider is called ***collider bias*** and drawn as a dashed path (see Elwert, 2013, Elwert & Winship, in press, or Pearl, 2009, for the effect of conditioning on a collider). Thus, after conditioning on Y we obtain the correspondingly altered path: $Z \leftarrow X_1 \overset{\frown}{\rightarrow} \boxed{Y} \overset{\frown}{\leftarrow} X_5$ (note that conditioning on Y would produce several other associations between variables X and U in Figure 1(1)).

Two variables are ***d-separated*** if all paths between them are blocked. If not all paths between two variables are blocked they are said to be ***d-connected***. Consider the two variables Z and Y in Figure 1(1) which are connected by a total of five paths: one direct path $Z \rightarrow Y$, and four indirect paths $Z \leftarrow X_1 \rightarrow Y$, $Z \leftarrow X_2 \rightarrow Y$, $Z \leftarrow X_3 \rightarrow Y$, and $Z \leftarrow X_1 \leftarrow U_1 \leftarrow\!\!\rightarrow U_3 \rightarrow X_3 \rightarrow Y$. Note that all four indirect paths are open because they are not naturally blocked by a collider and

that they contain an arrow into *Z*. By conditioning on X_1, X_2, and X_3 we can block all four indirect paths but *Z* and *Y* remain d-connected because we cannot block the direct path $Z \rightarrow Y$ (no variable on the path $Z \rightarrow Y$ is available for blocking). However, since our interest is in identifying the direct causal effect of treatment *Z* on outcome *Y* we want to let the direct path $Z \rightarrow Y$ open but block all the indirect paths because they confound (bias) the treatment-effect relationship. The indirect paths from *Z* to *Y* are also called ***confounding paths***, ***biasing paths*** or ***back-door paths***. They are called back-door paths because they can be viewed as entering *Z* through the back-door (via the arrows pointing into *Z*). Since X_1, X_2, and X_3 are not descendants of *Z* and since conditioning on X_1, X_2, and X_3 blocks all the confounding back-door paths, variables X_1, X_2, and X_3 are said to satisfy the ***back-door criterion*** which implies that the causal effect of *Z* on *Y* is identifiable.

The concept of ***compatibility*** links a causal graph to the joint probability distribution of the variables in the graph. A causal graph is compatible with the variables' probability distribution if the (conditional) independencies encoded in the graph also hold for the joint probability distribution. That is, if two variables can be *d*-separated by set of variables **X**, then the two variables must be conditionally independent with respect to the conditional probability distribution (given **X**). Finally, the ***faithfulness*** assumption (also called ***stability*** assumption) requires that the (conditional) independencies implied by the joint probability distribution remain invariant to changes in the parameters of the data generating process. A faithful probability distribution guarantees that none of its embedded independence relations occurs coincidentally or deliberately by choosing parameter settings such that two variables are seemingly independent though they are *d*-connected (this would happen if two associations cancel each other out, for instance). Faithfulness is not necessarily required but it helps in avoiding "pathological" or

unnatural data generating processes which may cause theoretical problems. In practice, however, putting (deterministic) constraints on data or the data generating process might easily result in an unfaithful distribution (Mansournia, Hernán, Greenland, 2013). Matching on the propensity score is an example.

Conditioning on a variable vs. conditioning on a value of a variable (selecting a subpopulation)

For causal graphs it is important to distinguish between (i) conditioning on a *variable,* say, X_1, which we symbolize by drawing a box around X_1 and (ii) conditioning on a single *value* of a variable $X_1 = x_1$, which corresponds to restricting the data to the subpopulation with $X_1 = x_1$. While the former results in an altered joint distribution for the *overall population* (with possibly new (in)dependence relations), the latter looks at the conditional distribution of a specific *subpopulation*. Since the conditional distribution and implied independence structure might differ across subpopulations with $X_1 = x_1'$ and $X_1 = x_1^*$ (with $x_1' \neq x_1^*$), the conditional graphs will vary accordingly and need to be drawn for each subpopulation separately. The graph for a subpopulation selected on the basis of a single value of a variable, $X_1 = x_1$ (as opposed to a range of values, e.g., $X_1 = [x_1 - 1, x_1 + 1]$), does not need to show the variable X_1, neither any arrows into X_1 nor any arrows leaving X_1 because X_1 is held constant. Thus, X_1's parents do not cause any variation in $X_1 = x_1$, and $X_1 = x_1$ itself no longer causes any variation in its children. This is also reflected by the corresponding structural causal model and the conditional distribution. After conditioning on $X_1 = x_1$ in Figure 1(1), the structural equation $X_{1i} = f_i^{X_1}(U_i^{X_1})$ is no longer required because it is degenerate and so is the distribution of X_1 (a probability distribution is degenerated if a random variable only takes a single value). The structural equation

$Y_i = f_i^Y(X_1, X_2, X_3, X_5, Z_i, U_i^Y)$ reduces to $Y_i = f_i^Y(X_2, X_3, X_5, Z_i, U_i'^Y)$ where the effect of $X_1 = x_1$ is now absorbed in the error term $U_i'^Y$. After conditioning on $X_1 = x_1$, the joint distribution then refers to the remaining variables and no longer depends on X_1. However, note that any collider bias due to conditioning on a subpopulation needs to be drawn in the subpopulation graph (as dashed path). If one would not drop variable X_1 and its ingoing and outgoing arrows from the causal graph, identification results for the subpopulation would be incorrect (for instance, in the RD design, without dropping *A* and its in- and outgoing arrows from the causal graph for $A = a$, the graph would suggest that ATE is identified according to the back-door criterion though ATE is actually unidentifiable for any value of the assignment variable $A = a$). Thus, conditioning on a single value requires redrawing the graph for the corresponding subpopulation and dropping the conditioning variable. Conditioning on a variable is graphically indicated by drawing a box around the conditioning variable (Elwert, 2013; Hernán & Robins, 2014).

The same holds for conditioning on a limiting value $X_1 \to x_1$, that is, conditioning on the interval $[x_1 - \varepsilon, x_1 + \varepsilon]$ with $\varepsilon \to 0$. X_1 and all its ingoing and outgoing arrows can be removed if all the structural functions that depend on X_1 are continuous at x_1. If one of the structural functions, say, $Z_i = f_i^Z(X_1, \ldots, X_4, U_i^Z)$, is discontinuous at x_1, then X_1 and the outgoing arrow $X_1 \to Z$ cannot be removed because even in the limit X_1 causes variation in *Z*.

Finally note that *conditioning* on a specific value of a variable, $X_1 = x_1$, is different to the thought *intervention* of Pearl's *do*-operator, $do(X_1 = x_1)$, where X_1 is set to x_1 for the *entire* population (Pearl, 2009). Thus, the subgraphs or "manipulated graphs" used in the *do*-calculus are different to the conditional graphs for subpopulations.

Appendix B

Proof 1. For the compliers, we prove that the average effect of the IV on *Y* is given by τ_C:

$$\begin{aligned} &E(Y_i \mid IV_i = 1, S_i = C) - E(Y_i \mid IV_i = 0, S_i = C) \\ &\quad = E(Y_i \mid Z_i = 1, S_i = C) - E(Y_i \mid Z_i = 0, S_i = C) \\ &\quad = E(Y_i^1 \mid S_i = C) - E(Y_i^0 \mid S_i = C) = \tau_C, \end{aligned}$$

where the third line with the potential outcomes follows from the independence of *Z* and **X**.

Proof 2. We prove $\gamma = E(Z \mid IV = 1) - E(Z \mid IV = 0) = P(S = C)$. Using

$$\begin{aligned} E(Z \mid IV = 1) &= P(S = C)E(Z \mid IV = 1, S = C) + P(S = N)E(Z \mid IV = 1, S = N) \\ &\quad + P(S = A)E(Z \mid IV = 1, S = A) + P(S = D)E(Z \mid IV = 1, S = D) \\ &= P(S = C)E(1 \mid S = C) + P(S = N) \times E(0 \mid S = N) \\ &\quad + P(S = A)E(1 \mid S = A) + P(S = D)E(0 \mid S = D) \\ &= P(S = C) + P(S = A) \end{aligned}$$

and

$$\begin{aligned} E(Z \mid IV = 0) &= P(S = C)E(Z \mid IV = 0, S = C) + P(S = N)E(Z \mid IV = 0, S = N) \\ &\quad + P(S = A)E(Z \mid IV = 0, S = A) + P(S = D)E(Z \mid IV = 0, S = D) \\ &= P(S = C)E(0 \mid S = C) + P(S = N)E(0 \mid S = N) \\ &\quad + P(S = A)E(1 \mid S = A) + P(S = D)E(1 \mid S = D) \\ &= P(S = A) + P(S = D) \end{aligned}$$

we obtain $E(Z \mid IV = 1) - E(Z \mid IV = 0) = P(S = C) - P(S = D) = P(S = C)$ because $P(S = D) = 0$ due to the monotonicity assumption.

Proof 3. We prove $\text{AIVE} = E(Y \mid IV = 1) - E(Y \mid IV = 0) = \gamma\tau_C$. Using

$$\begin{aligned} E(Y \mid IV = 1) &= P(S = C)E(Y \mid IV = 1, S = C) + P(S = N)E(Y \mid IV = 1, S = N) \\ &\quad + P(S = A)E(Y \mid IV = 1, S = A) + P(S = D)E(Y \mid IV = 1, S = D) \\ &= P(S = C)E(Y^1 \mid S = C) + P(S = N)E(Y^0 \mid S = N) \\ &\quad + P(S = A)E(Y^1 \mid S = A) + P(S = D)E(Y^0 \mid S = D) \\ &= P(S = C)E(Y^1 \mid S = C) + P(S = N)E(Y^0 \mid S = N) + P(S = A)E(Y^1 \mid S = A) \end{aligned}$$

and

$$
\begin{aligned}
E(Y \mid IV = 0) &= P(S = C)E(Y \mid IV = 0, S = C) + P(S = N)E(Y \mid IV = 0, S = N) \\
&\quad + P(S = A)E(Y \mid IV = 0, S = A) + P(S = D)E(Y \mid IV = 0, S = D) \\
&= P(S = C)E(Y^0 \mid S = C) + P(S = N)E(Y^0 \mid S = N) \\
&\quad + P(S = A)E(Y^1 \mid S = A) + P(S = D)E(Y^1 \mid S = D) \\
&= P(S = C)E(Y^0 \mid S = C) + P(S = N)E(Y^0 \mid S = N) + P(S = A)E(Y^1 \mid S = A)
\end{aligned}
$$

we obtain the average effect of the IV on *Y*:

$$
\begin{aligned}
\text{AIVE} &= E(Y \mid IV = 1) - E(Y \mid IV = 0) \\
&= \Pr(S = C) \times E(Y^1 \mid S = C) - \Pr(S = C) \times E(Y^0 \mid S = C) \\
&= \Pr(S = C) \times \{E(Y^1 \mid S = C) - E(Y^0 \mid S = C)\} \\
&= \Pr(S = C) \times \tau_C = \gamma \tau_C .
\end{aligned}
$$