Identifying Challenges for Non-Experimental Causal Inference with Data from Socio-technical Systems

Rick Wash
Michigan State University
404 Wilson Rd #342
East Lansing, MI 48824

Abstract

Many research questions can be answered by analyzing the decisions that humans make while using socio-technical systems. Building on Pearl's (2009) causal graphs theory, I identify three challenges that arise when using non-experimental data from socio-technical systems to answer questions about the causal effects of human decisions. First, technical features and affordances can create an endogenous selection bias that can affect the validity of causal inference even when results are properly scoped to only be about 'the users of the system'. Second, I highlight the problem of proxy control when using only log data to make claims about humans. And third, I re-emphasize the problem of homophily bias that arises when analyzing social network data and argue that this bias can influence a wide variety of questions beyond homophily.

When people use technologies to work and to communicate, these technologies and the people who use them form *socio-technical systems*. Researchers have been looking to socio-technical systems to better understand people, to better understand technology, and to better understand the relationship between people and technology.

Much of this research can be framed as studying human decision-making as indicated by socio-technical log data. For example, we can look at people's decisions to join and use Facebook, and examine the outcomes of that choice such as contribution of content, loss of privacy, and an increase in social support. We can also use this log data to examine indicators of non-technological decisions; for example, updating your workplace on LinkedIn can indicate that you chose a new job.

One of the major goals of much research is *causal inference*: establishing the presence of and size of an effect of a given cause. Does a human decision D cause some outcome Y to increase or decrease, and if so, by how much?

When working with data from socio-technical systems, researchers rarely have the opportunity to manipulate the systems. Instead, we often end up with non-experimental data – we have to accept the data as is rather than experimentally changing features or randomly assigning conditions.

Randomized experiments are the gold standard of causal inference. But in recent years, a number of researchers have been developing statistical tools and techniques for casual inference from non-experimental data. By subclassifying, stratifying, or controlling via a regression (any of which we refer to as "conditioning") on a appropriate set of variables, it is possible to estimate the presence and magnitude of a causal effect using non-experimental data (Morgan and Winship 2014). The causal graphs theory of Pearl (2009) provides a graphical framework for identifying the necessary variables. In this paper, I pose the question that economists call *identification*: using this data, it is possible in theory – if I had perfect population data – to estimate the causal effect of a human decision? Effects are called "identified" if it is possible, and are "not identified" if there is no possible way using the data available to estimate the effect.

This paper discusses the challenges in identifying causal effects of individual decisions in socio-technical systems. The decisions being studied don't necessarily have to be *about* the system, but they are made in the context of the system and so therefore are influenced by and effected by the system. I identify two general problems that socio-technical systems create that make identifying causal effects difficult: (1) endogenous selection bias due to fixed technological affordances, and (2) proxy control of human-level variables. I also discuss how homophily bias, first identified by Shalizi and Thomas (2011), arises in social networks data.

Causal Graphs and Back-Door Paths

In the absence of a manipulation (either experimental or natural), causal inference is difficult but not impossible. For many years, the basic logic has been that if you can condition your analysis on all other causes of the outcome of interest, then whatever relationship remains must be the causal effect of interest.

For example, consider the decision of whether or not to use Facebook (which we will call D). This decision likely has a causal effect on the outcome of the amount of social support that a person receives (which we can call Y, following the literature). We are interested in estimating the size of this causal effect $D \rightarrow Y$, though we do not have an explicit manipulation of Facebook use. Traditional statistical advice says that we can estimate this causal effect by controlling for all other causes of Y (for example, by measuring them and

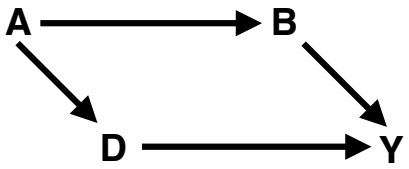

Figure 1: A simple causal graph. Letters represent (potentially unmeasured) variables. Arrows represent non-parametric causal relationships. The causal relationship $D \rightarrow Y$ is confounded by the back-door path $D \leftarrow A \rightarrow B \rightarrow Y$.

including them as predictors in a regression). However, this is rarely done because identifying all other causes is very difficult, and because this advice turns out not to be correct.

Recently, Pearl (2009) produced a new theory of causality, now known as *causal graphs*. Using this theory, we can identify a necessary and sufficient set of variables which, if we condition on them properly, we can use to accurately identify a causal effect. Pearl's theory depends on understanding the *causal graph*: A graph where nodes are variables and directed edges represent non-parametric causal relationships between variables. Using a causal graph, we can identify how associations (aka correlations) flow through causation relationships and create extra unidentified variation in outcome variables. For example, in Figure 1, the relationship between $D \rightarrow Y$ is not identified because there is an extra correlation due to D being causally related to A, A causing B, and B causing Y.

Pearl's theory goes on to identify a necessary and sufficient set of criteria for the variables that are needed to eliminate all sources of confounding other than the causal relationship of interest ($D \rightarrow Y$), which he calls the back-door criteria. Let S be a set of variables in the causal graph. Conditioning the analysis on the variables in S satisfies the back-door criteria iff:

1. All back-door paths between the causal variable and the outcome variable are blocked after conditioning on extra variables S. This happens if each back-door path
   (a) contains a chain of mediation $A \rightarrow C \rightarrow B$ and C is in S, or
   (b) contains a fork of mutual dependence $A \leftarrow C \rightarrow B$ and C is in S, or
   (c) contains an inverted fork of mutual causation $A \rightarrow C \leftarrow B$ where the middle variable C and all of C's descendents are *not* in S (C is called a "collider")
2. No variables in S are descendents of the causal variable that lie on (or descend from other variables that lie on) any of the directed paths that begin at the causal variable and reach the outcome variable.

(See the discussion in Morgan and Winship (2014) for a good discussion of the back-door criteria.)

The back-door criteria identify whether a causal effect can be estimated *in theory* from the available data. Economists state that a causal effect is *identified* if it can be estimated in theory from the available data. There are still many challenges in actually estimating the causal effect in practice, including finding matching cases, propensity score modeling, identifying the overlap between treatment and control, and ensuring balance (Morgan and Winship 2014). In this paper, I am concerned solely with whether a causal effect is identified in theory; if it isn't, then no strategy will be able to estimate the effect.

There are three important points that come out of this back-door criteria. First, to identify a causal effect, you do not need to control for *all* causes of Y; instead, you simply need to control for causes that lie on back-door paths from D to Y. Other variables do not affect the validity of the causal inference. Also, related to this, you do not necessarily need to control for the immediate cause; conditioning on any variable on the path from D to Y (such as A in Figure 1) will block the path. This can greatly reduce the number of variables that need to be measured and conditioned upon.

Second, it is possible to control for too many variables. While prior work from Heckman and others (**?**) had identified that controlling for outcomes is problematic, Pearl identifies an additional type of variable, which he calls a *collider* variable, that causes problems. Conditioning on a collider variable can actually open up paths that were previously blocked and create unwanted associations between the causal variable and the outcome variable. Elwert and Winship (2014) call this *endogenous selection bias*.

Third, in order for a causal effect to be estimated, it is important to identify all potential back-door paths and block those. This means that our causal graph must be complete enough to identify the relevant paths; only including variables we have measurements for is not enough. Both Pearl (2009) and Morgan and Winship (2014) include in their causal graph diagrams indicators for unmeasured variables for exactly this reason.

Technical Influences on Decisions: Endogenous Selection Bias

The focus of this paper is to understand how to estimate the effect on Y, some outcome variable, of a decision D made by a user. Since this user is operating in a socio-technical system, he or she must interact with the system after making the decision. The technical portion of socio-technical systems can easily be programmed to record data about user interactions with the system. Indeed, these log records are frequently already collected and often do not need additional work from the researcher to collect.

These logs record traces of behaviors that the user undertook while interacting with the system, and these traces can include evidence of the decision made (D), measurements of our outcome variable Y, and measurements of additional control variables that we may or may want to include in S, our set of conditioning variables.

However, since these logs come from a socio-technical system, the user is inevitably subject to the design and affordances of that system. There is much research that suggests that design elements of the user interface, such as the size of a text box for data entry, and whether the site supports user

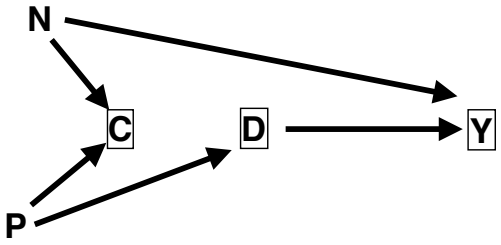

Figure 2: An example of a collider variable. Conditioning on C induces an association between P and N, and therefore creates an unblocked back-door path $D \leftarrow P \rightarrow \boxed{\text{C}} \leftarrow N \rightarrow Y$.

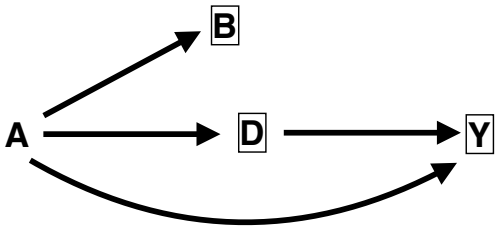

Figure 3: The problem of *proxy control*. A person's (A)ttitude can affect both their (D)ecision and the outcome (Y), which creates a back-door path $D \leftarrow A \rightarrow Y$. If we measure and control for (B)ehavior via log data, we only partially control for (A)ttitude; the residual variation can still induce unwanted associations between (D)ecision and outcome (Y). Boxed variables can be measured with log data.

profiles or not (Kraut and Resnick 2012, ch. 3), affect how people use the system. Such design elements should be included in a causal graph, since they can theoretically cause changes in both user behavior and outcome measures. Once included in a causal graph, we can then identify which of these design elements might be on a back-door path that can cause unwanted associations from decision to outcome.

In most technical systems, design elements are relatively fixed. For example, everyone who uses Facebook uses a system with the same (initial) size text box for status message input[1], and all Facebook users use a site that supports user profiles.

Most research using data from socio-technical systems ignores these fixed features and affordances of the technology. Researchers argue that they simply can put a criterion on the generalizability of the causal claim. Rather than claiming that they are estimating effects for the larger population, they scope their claims: "D causes Y as long as this set of design elements is the same" or "D causes Y for users of Facebook" or simply "D causes Y on Facebook".

However, this isn't enough. A fixed set of features and affordances effectively conditions on those features, which can create a selection effect. That is, fixed technical design elements can cause people to be selected into or out of the sample – the set of users of that system. This selection effect can act as a collider variable, and induce associations between variables that are otherwise unrelated.

For example, consider research that tries to estimate how sharing details about personal problems on Facebook (D) can lead to increased social support (Y). (See Figure 2 for a diagram of this example.) A person's decision about whether to post about a problem is certainly influenced by their privacy attitudes P. Facebook has a fixed set of privacy controls that everyone has access to. For some people, though, Facebook's available privacy controls might not be sufficient, and those people might then choose to not use Facebook. Therefore, the privacy controls create a selection effect C.

Social support is strongly influenced by the number of friends N that a person has on Facebook. The number of friends, however, also affects the selection of whether a person uses Facebook or not (C). By only including Facebook users in the study, this is effectively equivalent to conditioning on this design element (Elwert and Winship 2014). All subsequent analyses have to be considered to have conditioned on this variable. In this example, having many friends on Facebook may overcome a very private person's reluctance to join, but having fewer friends wouldn't be enough. Since C is a collider variable, it creates an association in the data between privacy attitudes P and number of friends N, and therefore opens up a back-door path $D \leftarrow P \rightarrow \boxed{\text{C}} \leftarrow N \rightarrow Y$ between the causal variable (the decision to share personal problems) D and the outcome of interest Y (social support). This back-door path means that the causal effect is not identified.

Elwert and Winship (2014) call this type of problem *endogenous selection bias*. Inclusion in the dataset implies a selection on fixed elements of the site, and this selection can act as a collider variable to create unwanted associations and back-door paths.

Proxy Control: The Problem with Log Data about Human Decisions

Many studies use purely log data to try to estimate the effects of technology use decisions. For example, we could estimate the effect of whether or not a person takes an online programming course (D) on their Github output (Y), or their use of Facebook privacy controls (D) on their use of words that indicate sensitive topics (Y).

However, in non-experimental studies, there are many back-door paths that can prevent the causal effect from being identified. Log data frequently comes with a large number of variables, and one commonly employed strategy is to control for these back-door paths by including additional variables in a regression.

Unfortunately, this strategy is very limited in practice because of the problem of *proxy control* (Elwert and Winship 2014). Consider the causal diagram in Figure 3. Almost all

[1]At least on a given platform; the input box can vary between mobile and standard interfaces.

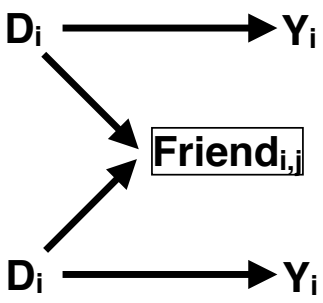

Figure 4: Homophily Bias. Individuals are indexed by i and j. By conditioning on whether two individuals are friends, as association can be induced in otherwise independent decisions D_i and D_j, creating backdoor paths in social network data. The friendship tie is a collider variable.

technology use decisions are influenced by any number of attitudes and beliefs on the part of the user: the attitude toward the efficacy of online courses, or the trust in Facebook's technology. However, we cannot directly measure attitudes with log data; instead, we measure behavioral traces that are hopefully correlated with attitudes.

In Figure 3, attitude A is unmeasured, but we measure the related behavior B via log data. Since A and B are correlated, B does provide some measure of control. However, residual variation in attitude that is not explained by B still provides a back-door path and therefore creates non-causal associations between decision D and outcome Y.

This problem of proxy control is a general problem for analysis focusing exclusively on log data. Virtually all decisions made by users are influenced by multiple attitudes, and even if we can find log data proxies for all of these attitudes, we still cannot fully remove the back-door paths due to these attitudes.

Homophily Bias In Network Data

Shalizi and Thomas (2011) identify another challenge that arises when using data from social networking systems. Much research has tried to identify the effect of homophily (being friends with people like you) separately from the effect of social contagion (becoming more like your friends). When analyzing friendship ties from social networks, you efficiently condition on whether two individuals are friends.

However, when homophily is present and people choose friends who are like them, then there is a causal relationship between the decisions of the users and the friendship tie. This makes the friendship tie a collider variable; see Figure 4 for an illustration. By conditioning on the friendship tie, this induces an association between the two otherwise independent decisions, and thus biases estimates of social contagion.

Elwert and Winship (2014) argue that this bias is best understood as an instance of endogenous selection bias. By analyzing friendship ties, you condition on a collider variable. Doing so can create back-door paths between individual outcomes that are easy to mistake for social contagion.

Summary and Conclusion

Analyzing non-experimental data from socio-technical systems to estimate causal relationships requires understanding relationships between social variables, relationships between technical variables, and relationships between social and technical variables. In this paper, I identify three challenges that frequently arise when analyzing data from socio-technical systems. These challenges can prevent a causal effect from even being identified, which means that even perfect data cannot accurately estimate the effect.

Socio-technical systems create a selection effect because they have a fixed set of features and affordances. Most prior work ignores technical features, arguing that the study only generalizes to other users of the site. However, I show that technical features can sometimes act as collider variables, inducing back-door paths between otherwise unrelated variables. Social science studies using data from socio-technical systems cannot ignore the technical features of the systems that produced or logged the data; instead they must carefully consider how those features may produce selection effects that affect the validity of their causal inferences.

Limiting analysis to log data makes it very difficult to identify the causal effects of individual's decisions. Most human decisions are influenced by attitudes, intentions, and beliefs held by the individual making the decision. Log data rarely captures those attitudes and usually is only an imperfect proxy for those attitudes. Including survey data with log data is one possible strategy to combat this issue. Surveys can capture both attitudes while logs capture behavior. By including variables about both attitudes and behaviors, researchers can fully condition the analysis to eliminate unwanted associations and block back-door paths.

Finally, care must be taken when analyzing data about relationships between people. When including friendship in an analysis, homophily can create bias in estimates by inducing an association between otherwise unrelated variables. Network ties should only be analyzed when studying decisions that have no causal effect on friendship decisions.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grants 1116544 and 1350253.

References

Elwert, F., and Winship, C. 2014. Endogenous selection bias: the problem of conditioning on a collider variable. *Annual Review of Sociology*.

Kraut, R., and Resnick, P. 2012. *Building Successful Online Communities: Evidence-Based Social Design*. MIT Press.

Morgan, S., and Winship, C. 2014. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press, 2nd edition.

Pearl, J. 2009. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition.

Shalizi, C., and Thomas, A. 2011. Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods and Research* 40:211–239.