Qualitative Inference from Causal Models

Macartan Humphreys[*] *and Alan Jacobs*[†]

Draft January 2017 (v 0.2): Comments welcome

Abstract

Process tracing is a strategy for inferring within-case causal effects from observable implications of causal processes. Bayesian nets, developed by computer scientists and used now in many disciplines, provide a natural framework for describing such processes, characterizing causal estimands, and assessing the value added of additional information for understanding different causal estimands. We describe how these tools can be used by scholars of process tracing to justify inference strategies with reference to lower level theories and to assess the probative value of new within-case information.

Contents

[*]Columbia University and WZB, mh2245@columbia.edu. Thanks to Donald Green and Steffen Huck for helpful conversations and to Fernando Martel Garcia for generous comments on an earlier draft.

[†]University of British Columbia, jacobs@politics.ubc.ca

1 Introduction

Political scientists often use process tracing to explain outcomes in single cases. Bennett and Checkel define process tracing as the "analysis of evidence on processes, sequences, and conjunctures of events within a case for the purposes of either developing or testing hypotheses about causal mechanisms that might causally explain the case" (Bennett and Checkel 2015, p 7). More broadly, we can think of the strategy as one of making inferences from within-case observations that are believed to carry information about causal relations (see also Collier (2011) and Mahoney (2010)). The approach is understood to be useful both for assessing causal effects (often by testing multiple hypotheses about the cause against one another) and for establishing the mechanism through which such effects operate in a case. While process tracing often focuses on case-level estimands—such as why or how an outcome occurred in a case—analysts sometimes use case-level findings derived from process tracing to speak to general theories or population-level claims (see, e.g., (Mahoney 2010, George and Bennett (2005))). However, process-tracing's *within-case* strategy of causal inference is generally distinguished from the *cross-case* logic used to identify causal effects in a standard statistical framework based on observed covariation between X and Y.

How does evidence derived from within a case allow analysts to draw causal inferences about the case? Van Evera (1997), Collier (2011), and Mahoney (2010) answer the question in terms of the probative value of within-case observations (see also Humphreys and Jacobs (2015)). They describe a set of tests— "hoop" tests, "smoking gun" tests or "doubly decisive" tests—that differ in probative value; that is in the inferences that can be made after observing new within-case data.[1] A smoking-gun test is a test that seeks information that is only plausibly present if a hypothesis is true (thus, generating strong evidence for the hypothesis if passed), a hoop test seeks data that should certainly be present if a proposition is true (thus generating strong evidence against the hypothesis if failed), and a doubly decisive test is both smoking-gun and hoop (for an expanded typology, see also Rohlfing (2013)).[2] Yet, conceptualizing the different ways in which probative value might operate in a sense only reframes the issues and leaves the more fundamental question unanswered: what gives within-case evidence its probative value with respect to causal relations?

In this paper, we demonstrate how casual models—that is, theory—can provide a principled underpinning for claims about the probative value of process-tracing evidence. Theory has long played a central role in accounts of the logic of process tracing. In their classic text on case study approaches, George and Bennett (2005) describe process tracing as the search for evidence of "the causal process that a theory hypothesizes or implies" (6). Similarly, P. A. Hall (2003) conceptualizes the approach as testing for the causal-process-related observable implications of a theory, Mahoney (2010) indicates that the events for which process tracers

[1] In Humphreys and Jacobs (2015) we use a fully Bayesian structure to generalize Van Evera's four test types in two ways: first, by allowing the probative values of clues to be continuous; and, second, by allowing for researcher uncertainty (and, in turn, updating) over these values. In the Bayesian formulation, use of information in K is not formally used to conduct tests in the manner suggested by the names of these strategies, but rather to update beliefs about different propositions.

[2] Note that these statements are statements about likelihood functions and do not require a specifically Bayesian mode of inference.

go looking are those posited by theory (128), and Gerring (2006) describes theory as a source of predictions that the case-study analyst tests (116).

Theory, in these accounts, is supposed to tell us where to look. What we do not yet have, however, is a systematic account of how researchers can derive within-case empirical predictions from theory and how exactly doing so provides leverage on a causal question. From what elements of a theory can scholars derive informative within-case observations? Process tracing is commonly thought of as focusing on the causal chain theorized to connect X to Y.[3] But what other elements of a theory—such as conditioning variables, pre-treatment variables, or post-outcome variables—imply potential observations with probative value on causal relations? Given a set of possible things to be observed in a case, how can theory help us distinguish more from less informative observations? Of the many possible observations suggested by a theory, how can we determine which would add probative value to the evidence already at hand? How do the evidentiary requisites for drawing a causal inference, given a theory, depend on the particular causal question of interest—on whether, for instance, we are interested in identifying the cause of an outcome, estimating an average causal effect, or identifying the pathway through which an effect is generated? In short, how exactly can we ground causal inferences from within-case evidence in background knowledge about how the world works?

This question finds an answer in methods developed in the study of Bayesian networks—a field pioneered by scholars in computer science, statistics, and philosophy—that has had limited traction to date in quantitative political science but that addresses very directly the kinds of problems that qualitative scholars routinely grapple with.[4] We begin by showing how a theory can be formalized as a causal model represented by a causal graph and a set of structural equations. Expressing theory in these terms allows us, for a wide range of causal questions, to identify a.) a set of variables (or nodes) in the model, including unobservable factors, that represent the causal query and b.) a set of observable variables that are potentially informative about the nodes in the query. Observation of data then leads to updated beliefs about queries and, simultaneously, updated beliefs about the model itself.

Graphs are already commonly employed by qualitative scholars to describe presumed causal relations between variables. Mahoney (2007), for example, uses a set of graphs to clarify the logic of arguments seeking to explain the origins of the First World War, supplementing the graphs with indicators of necessity and sufficiency that provide the kind of information generally carried by structural equations. Waldner (2015) uses causal diagrams to lay out a "completeness standard" for good process tracing. Weller and Barnes (2014) employ graphs to conceptualize the different possible pathways between causal and oucome variables among

[3]While a focus on intervening processes is probably the most common strategy in case-study practice, the methodological literature is clear that process-tracing evidence may derive from features of a process that do not intervene between X and Y. See, e.g., Bennett and Checkel (2015), Mahoney (2010), and Collier, Brady, and Seawright (2010). Nonetheless, the literature does not provide clear guidance on what kinds of non-intervening variables may be informative or when they will have probative value for causal questions.

[4]For application to quantiative analysis strategies in political science, Glynn and Quinn (2007) give a clear introduction to how these methods can be used to motivate strategies for conditioning and adjusting for causal inference; García and Wantchekon (2015) demonstrate how these methods can be used to assess claims of external validity.

which qualitative researchers may want to distinguish. Generally, in discussions of qualitative methodology, graphs are used to capture core features of theoretical accounts, but are not developed specifically to ensure a representation of the kind of independence relations implied by structural causal models (notably what is called in the literature the "Markov condition"). Moreover, efforts to tie these causal graphs to probative observations, as in Waldner (2015), are generally limited to identifying steps in a causal chain that the researcher should seek to observe.

If however we generate graphs from structural causal models—defined below—we can go much further. In particular, we can exploit well-understood properties of directed acyclic causal graphs to identify when a set of variables is uninformative about—specifically, conditionally independent of—another set of variables, given the model. In the discussion that follows, we show how researchers, having encoded their background knowledge in graphical form, can then use a set of simple graph-analytic rules to distinguish informative from uninformative data, given prior observations. As we show, the graphical relations that are critical for within-case inference are, in a certain sense, the opposite of those that are central to cross-case inference. For quantitative analysis drawing on cross-case correlations, knowledge that flows of information between variables are *blocked*, given what has been observed, provides a justification for drawing causal claims from correlations between variables; conditioning on confounders *blocks* dependencies that would otherwise bias correlational results. For qualitative analysis using case-level data, in contrast, knowledge that information flows between observable and query nodes are *open* indicates that additional within-case information would be informative about causal estimands given what is already known.

We further show how a model's structural equations—which can be non-parametric and express uncertainty about the functional form characterizing causal relations—can be used to derive the probative value of clues.

The framework thus provides guidance on within-case research design, for a given estimand of interest, conditional on the researcher's beliefs about how the world works. By the same token, the approach gives process tracers a principled way of grounding their causal inferences in their theoretical priors and thus in part addresses concerns that a focus on identification can crowd out attention to theory (Huber 2013). The case-level inferences that emerge are model-dependent, conditional on beliefs about general causal relations. Yet by logically tying data to theory, the framework also allows for *learning* about models as case-level data are observed.

In section 2 we begin with a discussion of causal relations. We give a brief overview of the counterfactual model, outlining the potential outcomes approach that is familiar already in political science, though we also highlight some implications of this model which are perhaps less familiar and sit uncomfortably with common conceptions of the logic of process tracing. We describe counterfactual relations as sets of structural equations and demonstrate how elements of these relations are visually represented by causal graphs. We then describe a broad collection of causal estimands including, but also extending beyond, traditional counterfactual causes, and show how these can be represented as statements about values of nodes on a causal graph—including, in many cases, unobservable nodes.

Section 3 describes an approach for justifying a causal model as *an implication of an underlying causal model*. We advocate conceptualizing theory in terms of underlying causal models and we describe a method for evaluating an underlying model in terms of gains in precision that the underlying model licences upon the collection of additional data—that is, by its informational content.

Section 4 then demonstrates when and how additional within-case data can be used to update beliefs on causal quantities of interest. To do so it draws on well established results in the study of probabilistic models that use "$d-$separation" to identify when additional information is informative about variables of interest conditional on whatever information is already available.

In summary, the strategy we advocate is as follows:

- Define causal estimands of interest as statements about the values of collections of "roots", Q, on a causal graph
- Identify a set of variables, K, that can be informative about Q given known data W: this is given by assessing whether K is $d-$separated from Q by W.
- Given priors over roots, evaluate the probative value of K by assessing the expected reduction in posterior variance in beliefs over the causal estimands, given K

2 Causes

We start with a description of the causal estimands of interest. We first define a causal effect and then introduce causal graphs and estimands. We give an example of a simple causal model that we return to to illustrate key ideas throughout the paper.

2.1 The counterfactual model

The counterfactual model is the dominant model of causal relations in the social sciences. The basic idea, sometimes attributed to David Hume[5] and more recently associated with Splawa-Neyman et al. (1990) and Lewis (1973)[6], conceptualizes causal relations as relations of "difference making." In the counterfactual view, X caused Y means: had X been different, Y would have been different. Importantly, the antecedent, "had X been different," imagines a *controlled* change in X, rather than a naturally arising difference in X. The counterfactual claim, then, is not that Y is different in those cases in which X is different; it is, rather, that if one could have *made* X different, Y would have been different.

In political science, the potential outcomes framework is commonly employed to describe counterfactual causal relations. Let $Y(x)$ denote the "potential" outcome (the value Y would

[5]Hume's writing contains ideas both about causality as regularity and causality as counterfactual. On the latter the key idea is "if the first object had not been, the second never had existed" (Hume and Beauchamp 2000, Section VIII).

[6]See also Lewis (1986).

take on) when $X = x$. Then, if X is a binary variable, the effect of X on Y is simply defined as $Y(1) - Y(0)$. The same type of notation can be used to describe more complex relations. For example, let $Y(x_1, x_2)$ denote the outcome when $X_1 = x_1$ and $X_2 = x_2$. Then the quantity $(Y(1,1) - Y(0,1)) - (Y(1,0) - Y(0,0))$ describes the interactive effect of two treatments: it captures how the effect of X_1 changing from 0 to 1 is different between those situations in which $X_2 = 1$ and those situations in which $X_2 = 0$.

Although the counterfactual framework is now widely employed, it contains a set of implications that might sit uncomfortably with a naive conception of how process tracing works.

First, process tracing is often thought of as a positivist enterprise, centered on careful measurement of processes connecting cause to effect. But in the counterfactural framework, a causal claim is a metaphysical statement. It involves claims not just about how the world is but how the world would be in different conditions. Thus a causal effect—including the smaller, intervening causal links between some X and some Y—can only be inferred, not directly observed. no matter how close one gets to the process or how fine grained ones data is.

Second, it is often intuitive (for both qualitative and quantitative scholars) to think of causal processes as sets of transitive relations: if we can figure out that A causes B and that B causes C, then we might think we have evidence that A causes C. Yet, in the counterfactual model, causal relations are *not* transitive. In a classic illustration, imagine a boulder that rolls down a hill, causing you to duck, and that ducking in turn saves your life. Clearly, the boulder caused the ducking and the ducking your survival, but the boulder rolling down the hill did not save your life. For discussions see N. Hall (2004) and Paul and Hall (2013).

Third, the language of "tracing" might suggest that causal relations must be continuous, connected in time and space. Yet in the counterfactural model, causes need not be temporally or spatially connected to their effects. *Potentially* intervening events that did *not* occur can have causal effects, even though they make no spatio-temporal contact with the observable events that seem to lie along the path from X to Y. The plague that put Friar John into quarantine meant that he did not deliver the letter to Romeo to inform him that Juliet was not dead, which in turn led to Romeo's death. There is a *causal* path from the plague to Romeo's death, but no *spatio-temporal* one.

Fourth, hypothesis-testing at the case level sometimes proceeds as though competing explanations amount to rival causes, where A caused B implies that C did not. But in the counterfactual model, causal relations are neither rival nor decomposable. If two out of three people vote for an outcome under majority rule, for example, then both of the two supporters caused the outcome: the outcome would not have occurred if *either* supporter's vote were different. For non-decomposability, imagine all three of three voters support an outcome, then they jointly cause the outcome; but none of their individual votes had *any* effect on the outcome.

Thus there appear to be some tensions between the counterfactual model and notions of causality common in (though by no means limited to) process tracing. These tensions largely disappear, however, once we properly specify causal models as systems of causal relations.

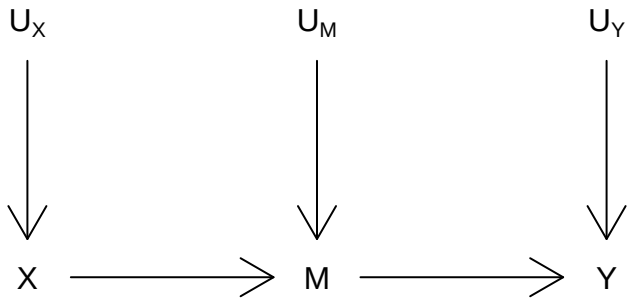


Figure 1: X, M, and Y are endogeneous variables. U_X, U_M, and U_Y are exogeneous. The arrows show relations of causal dependence between variables, from which relations of conditional independence can be deduced. Not shown on the graph are the ranges of the variables, $\mathcal{R}$, or the functional relations between them.

For this work, Directed Acyclic Graphs provide a powerful tool.

2.2 Causal Models and Directed Acyclic Graphs

In principle, highly complex causal relations can be expressed in potential outcomes notation. However, for structures involving multiple variables, it can be useful to generate a visual representation of causal relations. In this section, we show how causal models and directed acyclic graphs (DAGs) can represent substantive beliefs about how the world works. The key ideas in this section can be found in many texts (see, e.g., Halpern and Pearl (2005) and Galles and Pearl (1998)).

We consider causal models formed out of three components: variables, functions, and distributions.

The variables. Let $\mathcal{U}$ denote a set of exogenous variables. Let $\mathcal{V}$ denote a collection of variables of interest that are endogenous in the sense that they are functions of $\mathcal{U}$ and possibly other elements of $\mathcal{V}$. For example, $\mathcal{V}$ might contain a set of specified variables such as "X = Free Press" and "Y = Government removed"; $\mathcal{U}$ might then include unspecified factors that give rise to a free press and other factors that lead to government removal, other than free press. Let $\mathcal{R}$ denote a set of *ranges* for all variables in $\mathcal{U} \cup \mathcal{V}$. Thus in the binary case the range of X is $\mathcal{R}(X) = \{0, 1\}$. The triple $(\mathcal{U}, \mathcal{V}, \mathcal{R})$ is sometimes called a *signature*, $\mathcal{S}$—a listing of the variables in a model and the values that they can take on.

The functions. Let $\mathcal{F}$ denote a collection of *causal functions*, one for each variable $V_i \in \mathcal{V}$. We shall call the arguments in f_i the *parents* of variable V_i, PA_i, and by analogy we will say that V_i is a *child* of any variable in PA_i; children of V_i are descendants of PA_i and parents of PA_i are ancestors of V_i. The set of parents is required to be minimal in the sense that a variable is not included among the parents if, given the other parents, the child does not

depend on it in any state that arises with positive probability. Variables that have no parents in $\mathcal{V}$ are called *roots*.[7] We will say that $\mathcal{F}$ is a set of *ordered structural equations* if no variable is its own descendant and if no element in $\mathcal{U}$ is parent to more than one element of $\mathcal{V}$.[8]

For notational simplicity we generally write functional equations in the form $c(a, b)$ rather than $f_c(a, b)$.

The distributions. We let $P(u)$ denote a joint probability distribution over $\mathcal{U}$. A particular realization of U, u, is a *context.* A context is sufficient to determine outcomes, given a set of ordered structural equations. In what follows, we will assume that the elements of $\mathcal{U}$ are generated independently of one another. While this is not without loss of generality, it is not as constraining as it might at first appear: any graph in which two U variables are not independent can be replaced by a graph in which these U terms are listed as (possibly unobserved) nodes in $\mathcal{V}$, themselves generated by a third variable in $\mathcal{V}$ with, possibly, a parent in $\mathcal{U}$.

The U terms are sometimes described as capturing noise, or random disturbances caused by omitted forces. They can also be thought of as capturing uncertainty about functional forms. For example, suppose that in Figure 1, $U_Y \sim \text{Unif}[0, 1]$. This graph and distribution on U_Y is consistent with many possible equations for Y, including:

- $Y = X + U_Y$
- $Y = \mathbb{1}(U_Y > q)X$
- $Y = U_Y X$
- $Y = X^{U_Y}$

The first two equations capture common ways of thinking of Y as a stochastic function of X—in one case continuous, in the other binary. The third and fourth equations more obviously capture uncertainty over functional form, though also specifying certain known features (such as linearity in the third case). Thus, the use of a structural model *does not require precise knowledge of specific structural relations*, that is, of functional forms. This feature means that a model can be constructed to be very general—and can allow variables to be included as parents even if one is not sure that they matter (e.g., we could have $y_2 = a + by_1$ but allow that b might take the value 0). Some possibilities are excluded by the framework, however: for example, one cannot represent uncertainty regarding whether A causes B or B causes A.

With these elements in hand, we can define a structural causal model:[9]

Definition: A **structural causal model** *over* signature $\mathcal{S} =< \mathcal{U}, \mathcal{V}, \mathcal{R} >$ is a pair $< \mathcal{S}, \mathcal{F} >$, where $\mathcal{F}$ is a set of ordered structural equations containing a function f_i for each element $Y \in \mathcal{V}$.

[7] Thus in our usage all elements of $\mathcal{U}$ are roots, but so are variables in $\mathcal{V}$ that depend on variables in $\mathcal{U}$ only.

[8] This last condition can be achieved by shifting any parent of multiple children in $\mathcal{U}$ to $\mathcal{V}$.

[9] Note the definition here includes an assumption of acyclicity which is not found in all definitions.

Note that the definition does not include any information about $P(u)$; that is, a structural causal model describes how variables relate to each other but does not say anything about how likely any possible context or, in turn, outcome is. Thus, the model may stipulate that X causes Y, but say nothing about the distribution of X. Thus $P(y|x, u_Y)$ is defined by the model (as a degenerate distribution), but $P(x)$, $P(u_Y)$ and $P(x, y, u_Y)$ are not.

Once we introduce beliefs over $\mathcal{U}$ we have a "probabilistic causal model" which entails not just claims about how the world works under different conditions, but beliefs about what conditions we face.[10] Thus, a structural causal model might support a claim of the form "X causes Y if and only if condition C holds" whereas a probabilistic model supports a claim of the form "condition C arises with frequency π^C and so X causes Y with probability π^C." Formally:

Definition: A **probabilistic causal model** is a structural causal model coupled with a probability distribution P over $\mathcal{U}$.

The assumptions that no variable is its own descendant and that the U terms are generated independently make the model *Markovian*, and the parents of a given variable are Markovian parents. Knowing the set of Markovian parents allows one to write relatively simple factorizations of a joint probability distribution, exploiting the fact ("the Markov condition") that all nodes are *conditionally independent* of their nondescendants, conditional on their parents.[11]

To see how this Markovian property allows for simple factorization of P note that $P(X, M, Y)$ can always be written as:

$$P(X, M, Y) = P(X)P(M|X)P(Y|M, X)$$

If we believe, as above, that X causes Y only through M then we have the slightly simpler factorization:

$$P(X, M, Y) = P(X)P(M|X)P(Y|M)$$

Or, more generally:

$$P(v_1, v_2, \ldots v_n) = \prod P(v_i|pa_i)$$

The distribution P on $\mathcal{U}$ induces a joint probability distribution on $\mathcal{V}$ that captures not just information about how likely different states are to arise but also the relations of conditional independence between variables that are implied by the underlying causal process. For example, if we thought that X caused Y via M (and only via M), we would then hold that $P(Y|M) = P(Y|X, M)$: in other words if X matters for Y only via M then, conditional on M, X should not be informative about Y.

In this way, a probability distribution P over a set of variables can be consistent with some causal models but not others. This does not, however, mean that a specific causal model

[10] One could also envision "incomplete probabilistic causal models"" in which researchers claim knowledge regarding distributions over *subsets* of $\mathcal{U}$.

[11] Variables A and B are "conditionally independent" given C if $P(a|b, c) = P(a|c)$ for all values of a, b and c.

can be extracted from P. To demonstrate with a simple example for two variables, any probability distribution on (X, Y) with $P(x) \neq P(x|y)$ is consistent both with a model in which X is a parent of Y and with a model in which Y is a parent of X.

Let us now consider the graphical representation of causal models. With a causal model in hand we can represent $\mathcal{V}$ as a set of vertices (or nodes) connected by a collection of directed edges (single-headed arrows). We add a directed edge from node A to node B if and only if A is a parent of B. The resulting diagram is a *directed acyclic* graph (DAG)—the specification that $\mathcal{F}$ is a set of ordered structural equations ensures that there are no paths along directed edges that lead from any node back to itself.[12] We may or may not add nodes for elements in $\mathcal{U}$ explicitly, though in common practice, $\mathcal{U}$ is excluded from the representation, with an assumption that one element in $\mathcal{U}$ points into each element in $\mathcal{V}$.

In Figure 1 we show a simple DAG that represents a situation in which X is a parent of M, and M is a parent of Y. In this example, the three variables U_X, U_M, and U_Y are all exogenous and thus elements of $\mathcal{U}$. X, M, and Y are endogenous and members of $\mathcal{V}$. If these three variables were binary, then there would be eight possible realizations of outcomes, i.e., of $\mathcal{V}$. In the underlying model, U_X is an ancestor of X, M, and Y which are all descendants of U_X. The elements of $\mathcal{U}$ are all roots, though X is also a root as it has no parent in $\mathcal{V}$. Note that the graph contains less information than a causal model: it records which arguments enter into the structural equation for each variable but contains no other information about the form of those equations.

However, a key advantage of a DAG-representation of a causal model is that it allows for an easy reading of conditional independencies between variables in the model.

Conditional independencies can be read off the graph by checking whether paths between two variables, A and B are "active" given a set of variables, $\mathcal{C}$. Intuitively, the question is whether variables in $\mathcal{C}$ block information flows from A to B or rather allows, or possibly even creates, such flows. This can be assessed as follows. For each possible path between A and B check whether (a) there is a "chain" $X \rightarrow Y \rightarrow Z$ (going either direction) or "fork" $X \leftarrow Y \rightarrow Z$, with $Y \subset \mathcal{C}$ or (b) there is an "inverted fork" $X \rightarrow Y \leftarrow Z$ for which neither Y nor its descendants are in C. If either of these conditions holds, then the path is not active given $\mathcal{C}$. In the first case, information flows are blocked by a variable in $\mathcal{C}$. In the second case, information flows are not created by any variable in $\mathcal{C}$. If there are no active paths, then A and B are said to be "d-separated" by $\mathcal{C}$.[13]

Thus, in Figure 1, we can readily see that X and Y are conditionally independent given M: X and Y are d-separated by M. In a graph of the form $X \leftarrow Y \rightarrow Z$, we can see that X and Z are not independent but they are conditionally independent given Y; Y $d-$separates X and Z. In the graph $X \rightarrow Y \leftarrow Z$, X and Z are independent; but conditioning on Y $d-$*connects* X and Z, generating a dependency between them and so in this case X and Z

[12] More specifically, we have a "causal DAG" (Hernán and Robins 2006) since (i) the absence of an arrow between A and B means that A is not a direct cause of B (i.e., A does not enter into the functional equation for B) and (ii) any cause common to multiple variables is represented on the graph.

[13] There are multiple techniques for establishing $d-$separation. Pearl's guide "$d-$separation without tears" appears in an appendix to Pearl (2009).

are not $d-$separated by Y. In language used by Pearl and others, in this second case Y is a "collider" for X and Y, a child of two or more parents.[14]

A second advantage of causal graphs is that they provide a useful structure for thinking through the effects of interventions on outcomes. An intervention is thought of as a controlled change in some variable, X, which provides X with a value that is *not* determined by its parents. In an intervention, it is as if the function $X = f_i(pa_i)$ is replaced by the function $X = x$, with x being a constant. In the formulation used in Pearl (2009), this action is written as $do(V_i) = v_i'$, or for notational simplicity $\hat{v}_i'$ (meaning V_i is forced to take the particular value v_i') and the resulting distribution can be written:

$$P(v_1, v_2, \ldots v_n|\hat{v}_i) = \prod_{-i} P(v_j|pa_j)\mathbb{1}(v_i = v_i') \tag{1}$$

where $-i$ indicates that the product is formed over all variables V_j other than V_i, and the indicator function ensures that probability mass is only placed on vectors with $v_i = v_i'$. This new distribution has a graphical interpretation, representing the probability distribution over a graph in which all arrows into V_i are removed.

2.3 Queries

Much social science research focuses on the estimation of average causal effects. Yet many other estimands are of scholarly interest. These include case-level causal effects, causal attribution, actual causes, and causal pathways. Some causal questions involve realized values of variables only, some involve counterfactual statements, and some involve combinations of these.

Assessing many causal questions requires understanding multiple parts of a causal network. In what follows we advocate an approach in which (a) uncertainty about causal questions is represented as uncertainty about the *values of root nodes of a graph*, including unobservable roots and (b) estimands – which we term *queries* — are defined as questions about the values of collections of these nodes.[15]

[14]It is commonly understood that two variables may be independent conditional on $\mathcal{C}$ but not independent otherwise, as in the graph $A \leftarrow C \rightarrow B$. Less obviously, two variables may be unconditionally independent but *not* independent conditional on a third variable. Here, consider a situation in which variable C is a function of A and B which are each determined through independent random processes such that C acts as a collider for A and B. Conditioning on a collider (or the descendant of a collider) introduces a correlation between parents of the collider that might otherwise not exist. The reason is in fact quite simple: if an outcome is a joint function of two inputs, then if we know the outcome, information about one of the inputs can provide information about the other input. For example: If I know you have brown eyes, then learning that your mother has blue eyes makes me more confident that your father has brown eyes.

[15]With some abuse of notation we use Q generically to refer to the query itself and the the set of variables whose values determine the query. Thus a query may be written as the random variable $Q = \mathbb{1}((u_X = 1)\&(u_Y = 0))$, which takes on a value $q = 1$ if both $u_X = 1$ and $u_Y = 0$ and 0 otherwise. Assessing this query requires understanding the values of particular roots, or query nodes, $\{U_X, U_Y\}$ which we also refer to as Q.

Addressing causal questions of different kinds then involves using data on observed features of a graph to make inferences about particular unobserved or unobservable features of the graph, conditional on the graph itself. In this framework, inferences about causation amount to inferences about the *context* that a case is in: that is, whether conditions in the case (the relevant root node values) are such that a given causal effect, causal pathway, etc. would have been operating. We can translate questions about causation into questions about root nodes because, in a structural causal model, the values of all nodes in $\mathcal{U}$ is sufficient to determine the value of all nodes in $\mathcal{V}$: context determines outcomes. This further implies that, for any manipulation of an exogenous or endogenous variable, there exist one or more root nodes on the graph that suffice to determine the effect on all endogenous variables in the graph.

It is important to note a difference between this formulation and the conceptualization of causality typically employed in the potential outcomes framework. We characterize causal inference as learning about a unit *as it is*, conditional on a causal model, rather than learning about the unit as it is and as it could be. Suppose, for instance, that in a causal model a car will start if it has gas and if the key is turned.[16] Given this model, the question "Does turning the key cause the car to start?" is equivalent to the question, "Does the car have gas?" In the model-based framework, our query becomes a question about the current state of affairs—about the *context* of the case—rather than a pair of factual and counterfactual questions about outcomes with and without turning the key. Counterfactual reasoning is no less important in this framework; it has simply been displaced to the causal model, which encodes all counterfactual relations.

Case Level Counterfactual Cause. The simplest quantity of interest is the question of whether there is a case-level counterfactual causal effect. Does X cause Y in this case? The closely connected question of causal attribution (Yamamoto 2012) asks: did X cause Y in this case?

In Humphreys and Jacobs (2015) we employ the idea of response types (principal strata) to describe causal effects in a situation in which X and Y are binary (see also Frangakis and Rubin (2002)): a unit is of type a (adverse) if it has potential outcomes $Y(X) = 1 - X$, b (beneficial) if $Y(X) = X$, c (chronic) if $Y(X) = 0$ and d (destined) if $Y(X) = 1$. Written as a structural causal model, we can let Y be a function of X and of Q, a response-type variable that encodes potential outcomes. Represented as a graph, we have $X \rightarrow Y \leftarrow Q$. We let Q take on values q_{ij}, with i representing the value Y takes on if $X = 0$ and j representing the value Y takes on if $X = 1$. Thus, in a binary framework Q can take on four values: q_{00}, q_{10}, q_{01} and q_{11}. The equations for Y can be given by $Y(x, q_{ij}) = i(1 - x) + jx$. The query, "What is the case-level causal effect?", then becomes a question of learning about the root variable Q in the graph.

Note that, in this illustration, the root variable Q is not specified in substantive terms; it is a carrier for causal information. Below, however, we also provide examples in which the root variables have a stronger substantive interpretation and are not just notional stand-ins for causal types. Note, also, that there is no loss of generality in the functional form linking X and Q to Y. In the causal model framework, the structural equations, such as those linking

[16] A version of this example is in Darwiche and Pearl (1994).

X and Y conditional on another node, can be entirely non-parametric.

More generally, work in graphical models defines the causal effect of X on Y in terms of the changes in Y that arise from interventions on X. For example, using the notation for interventions given above we can describe the effect of a change in X from x' to x'' on the probability that $Y = 1$ in unit i as:

$$P(y = 1|\hat{x}_i') - P(y = 1|\hat{x}_i'') \tag{2}$$

where $P(y = 1|\hat{x}_i')$ is calculated from the marginal distribution of y given the post intervention distribution described by Equation 1 above. With y expressed as a function of x, this quantity reduces to a statement about the probability of Q taking on a given value (i.e., of the relation between Y and X taking a particular functional form).

More generically, in a graph $U_X \rightarrow X \rightarrow Y \leftarrow U_Y$, the effect of X on Y in a case will depend on the value of U_Y in that case. There are special cases in which X's effect will not depend on the value of U_Y. For instance, if U_Y operates only additively on Y (say, $Y = X + U_Y$) and Y is not bounded, then U_Y is irrelevant to X's causal effect, which will be homogeneous across cases and fixed by the model. But, in general, the causal effect of a parent on its child will depend on the value(s) of that parent's spouse(s).[17] Thus, learning about X's effect on Y within a case amounts to learning about the value of Y's other ancestors.

Note also that the distinction between the questions "would X cause Y?" and "did X cause Y?" is a difference in the nodes about which inferences are needed. The first requires information only on a response-type variable, as in the example above; the second requires information both on response type and on the value of X, in order to determine whether X *in fact* took on a value that would have produced Y, given the case's response type.

Average Causal Effects. A more general query would be to ask about average causal effects in some population. This too can be conceptualized as learning about values of root nodes. Using the same notation as above and in Humphreys and Jacobs (2015), let each unit be randomly selected from a population in which share λ_j are of type j. The population is thus characterized by a multinomial distribution with probabilities $\lambda = (\lambda_a, \lambda_b, \lambda_c, \lambda_d)$. The average causal effect is then $\lambda_b - \lambda_a$. We then include λ as a node on the causal graph pointing into Y, generating $X \rightarrow Y \leftarrow \lambda$. Now λ—the distribution of types in the population—is itself a quantity of interest and, like a response-type variable, one that cannot be directly observed. Moreover, λ can be thought of as itself drawn from a distribution, such as a Dirichlet. The hyperparameters of this underlying distribution of λ represent uncertainty over λ and hence over average causal effects. We can then use information from observable nodes (such as the value of X and Y) to learn both about the case-level causal type and about λ, and so about average causal effects for the population.

Formally, this kind of average causal effect is also calculated using Equation 2, though for a model that is not conditional on the case at hand.

[17] Nodes that share a child are spouses.

Actual Cause. Sometimes an outcome does not depend in a counterfactual sense on an antecedent condition, yet that condition may in some sense have generated or produced the outcome. Using the definition provided by (Halpern 2015), building on (Halpern and Pearl 2005) and others, we say that $X = x$ was an *actual cause* of $Y = y$ (where x and y may be collections of events) if:

1. $X = x$ and $Y = y$ both happened
2. there is some set of variables, W, such that if they are fixed at the levels that they actually took, but X is changed, Y would change
3. no strict subset of X satisfies 1 and 2

A motivating example used in much of the literature on actual causes (e.g. N. Hall 2004) imagines two characters, A and B, both great shots, simultaneously throwing stones at a bottle. A's hits first; the bottle breaks. B's would have hit had A's not hit, and would have broken the bottle, Y. Did A's throw ($A = 1$) cause the bottle to break ($Y = 1$)? Did B's?

By the usual definition of causal effects, neither A's nor B's action had a causal effect: without either throw, the bottle would still have broken. We commonly encounter similar situations in the social world. We observe, for instance, the onset of an economic crisis and the breakout of war—either of which would be sufficient to cause the government's downfall—but with the economic crisis occurring first and toppling the government before the war could do so. Yet neither economic crisis nor war made a difference to the outcome.

To return to the bottle example, while neither A's nor B's throw is a counterfactual cause, there is an important sense in which A's action obviously broke the bottle, and B's did not. This intuition is confirmed by applying the definition above. Consider first the question: Did A break the bottle? Conditions 1 and 3 are easily satisfied, since A *did* throw and the bottle *did* break (Condition 1), and "A threw" has no strict subsets (Condition 3). Condition 2 is met if A's throw made a difference, counterfactually speaking; and in determining this, we are permitted to condition on any event or set of events that actually happened (or on nothing at all). To see why Condition 2 is satisfied, we have to think of there being three steps in the process: A and B throw, A's or B's rock hits the bottle, and the bottle breaks. In actuality, B's stone did not hit the bottle. And conditioning on this actually occurring event, the bottle wouldn't have broken had A not thrown. From the perspective of counterfactual causation, it may seem odd to condition on B's stone not hitting the bottle when thinking about A not throwing the stone since throwing the stone was the very thing that prevented B from hitting the bottle. Yet Halpern argues that this is an acceptable thought experiment since it is conditioning only on facts of the case. Moreover, the same argument shows why B is not an actual cause. The reason is that B's throw is only a cause in those conditions in which A did not hit the bottle; but A *did* hit the bottle, so we are not permitted to condition on A not hitting the bottle in determining actual causation.

The striking result here is that there can be grounds to claim that X was the actual cause of Y even though, under the counterfactual definition, the effect of X on Y is 0. One immediate methodological implication follows: Since actual causes need not be causes, there are risks in research designs that seek to understand causal effects by tracing back actual causes—i.e.,

the way things actually happened.[18]

As with other causal queries, the question "Was $X = x$ the actual cause?" can be redefined as a question about which values for root nodes produce conditions under which X could have made a difference. Similarly, the question of how *common* it is for a condition to be an actual cause can be expressed as values of nodes, possibly including nodes that record parameter values for the relevant root nodes.

Notable cause An extended notion (Halpern 2016, p 81) of actual causes restricts the imagined counterfactual deviations to states that are more likely to arise (more "normal") than the factual state. We will call this notion a "notable cause." Similarly, one cause, A, is more notable than another cause, B, if a deviation in A from its realized state is more likely than a deviation in B from its realized state.

For intuition, we might wonder why a Republican was elected to the president; in looking at some minimal winning coalition of states that voted Republican we might distinguish between those that *always* vote Republican and those that are more volatile. If the coalition is minimal winning then all the states are causes of the outcome, but the volatile states are more notable causes; in a sense, only their actions were in play.

Again, whether something is a notable cause, or the likelihood in some population that a condition is a notable cause, can be expressed as a claim about the value of a set of root nodes.

Causal Paths. For a richer explanation of causal effects, researchers often seek to describe the causal path, or causal paths, through which effects propagate. Consider a DAG with $X \rightarrow M$ and $X, M \rightarrow Y$. It is possible that in a case with $X = 1$ and $Y = 1$ one might have reasonable confidence that X caused Y, but may be interested in knowing whether X caused Y *through* M. This question goes beyond assessing whether indeed $M = 1$ when $X = Y = 1$—though that might be useful information—to the question of whether in some sense $X = 1$ caused $M = 1$ and that effect in turn caused $Y = 1$.

This kind of question is taken up in work on mediation where the focus goes to understanding quantities such as the "indirect effect" of X on Y via M. Formally, this is $Y(X = 1, M = M(X = 1, U_M), U_Y) - Y(X = 1, M = M(X = 0, U_M), U_Y))$, which captures the difference to Y if M were to change in the way that it would change due to a change in X, but without an actual change in X (Pearl 2009 p 132). As stated, this is again a statement about specific nodes: U_Y and U_M. Consider, first, U_M. The structural equation for M will include X and U_M as arguments. Thus, knowing the value of M for any given value of X, conditional on a given structural equation for M, requires knowing U_M. The same logic operates for U_Y's role in determining how Y responds to a given change in M, conditional on Y's structural equation.

[18]Perhaps more surprising, it is possible that the expected causal effect is negative but that X is an actual cause in expectation. For instance, say that 10% of the time A's shot intercepted B's shot but without hitting the bottle. In that case the average causal effect of A on bottle breaking is -0.1 yet 90% of the time A is an actual cause of bottle breaking (and 10% of the time it is an actual cause of non-breaking). For related discussions see Menzies (1989).

Such a focus on causal paths does not restrict attention to questions of the form "how did X cause Y" but more generally, "what paths generated Y?" Such questions may have answers of the form "$Y = 1$ occurred because $X = 0$ led to $M = 0$, which, when $Z = 1$, gives rise to $Y = 1$ and not because $X = 1$ led to $M = 1$, which, when $Z = 0$ gives rise to $Y = 1$." Such inquiries can focus on distinct sets of conditions that give rise to an outcome ("equifinality"), as in Qualitative Comparative Analysis (QCA). While QCA analysts sometimes refer to sets of conditions as "paths", QCA does not generally involve explicit assessment of the causal steps linking conditions to outcomes. When examining paths in a causal-model framework, the analyst can address queries that involve drawing inferences about an entire chain linking X to Y or even an entire causal network. An understanding of a full causal network would, in turn, allow for any more specific estimand to be estimated.

2.4 Inference

Once queries are defined in terms of the values of roots—or *contexts*—then formation of beliefs, given data W, about estimands follows immediately from application of Bayes rule. That is, let $Q(u)$ define the value of the query in context u, the updated beliefs about the query are given by the distribution:

$$P(q|W) = \int_{u:Q(u)=q} P(u|W)du = \int_{u:Q(u)=q} \frac{P(W|u)P(u)}{\int_{u'} P(W|u')P(u')du'}du$$

This expression gathers together all the contexts that produce a given value of Q and assesses how likely these are, collectively, given the data.[19] For an abstract representation of the relations between assumptions, queries, data, and conclusions, see Figure 1 in Pearl (2012).

For illustration consider the "Two Child Problem" (Gardner 1961): *Mr Smith has two children, A and B. At least one of them is a boy. What are the chances they are both boys?* The two roots are the sexes of the two children. The query here is Q: "Are both boys?" which can be written in terms of the roots. The statement "$Q = 1$" is equivalent to the statement (A is a boy & B is a boy). Thus it takes the value $q = 1$ in just one context. Statement $q = 0$ is the statement ("A is a boy & B is a girl" or "A is a girl & B is a boy" or "A is a girl & B is a girl"). Thus $q = 0$ in three contexts. If we assume that each of the two children is equally likely to be a boy or a girl with independent probabilities, then each of the four contexts is equally likely. To be explicit about the puzzle, we will assume that the information that one child is a boy is given as a truthful answer to the question "is at least one of the children a boy?" The surprising result can then be figured out as $P(Q = 1) = \frac{1\times\frac{1}{4}}{1\times\frac{1}{4}+1\times\frac{1}{4}+1\times\frac{1}{4}+0\times\frac{1}{4}} = \frac{1}{3}$. This answer requires summing over only one context. $P(Q = 0)$ is of course the complement of this, but using the Bayes formula one can see that it can be found by summing over the posterior probability of three contexts in which the statement $Q = 0$ is true.

[19] Learning about roots from observed data is sometimes termed *abduction*; see Pearl (2009), p 206.

2.5 A running example

Consider a simple probabilistic causal model of a political process. Begin with two features of context: there may or may not be a free press (X) and a government may or may not be sensitive to public opinion (S).[20] Then, say that the government will act honestly only if there is a free press and the government is sensitive; if there is government corruption and a free press, the press will report on the corruption; the government will be removed from office if indeed it has acted corruptly and this gets reported by the press. We expand on underlying logics for this example later.

As a set of equations, this simple structural causal model may be written as follows:

$X = \mathbb{1}(u_X < \pi^X)$	Whether the press is free
$S = \mathbb{1}(u_S < \pi^S)$	Whether the government is sensitive
$C = 1 - X \times S$	Whether the government is corrupt
$R = C \times X$	Whether the press reports on corruption
$Y = C \times R$	Whether the government is removed from office

where π^S and π^X are parameters governing the probability of S and X, respectively, taking on the value of 1.

To generate a probabilistic causal model, we also need distributions on $\mathcal{U} = (U_S, U_X)$. These are given by:

$u_S \sim \text{Unif}[0,1]$	Stochastic component of government type
$u_X \sim \text{Unif}[0,1]$	Stochastic component of press freedom

Note that in this model, unlike in Figure 1, only the most "senior" specified variables, X and S, have a stochastic component (i.e., include a U term in their function); all other endogenous variables are deterministic functions of other specified variables.

Substituting through the causal processes, the functional equation for the outcome can be written as $Y = (1-S)X$. In Boolean terms, where Y stands for the occurrence of government removal, $Y = \neg S \wedge X$; and the function for the outcome "government retained" can be written $\neg Y = (S \wedge X) \vee (S \wedge \neg X) \vee (\neg S \wedge \neg X)$ or, equivalently, $\neg Y = S + \neg S \neg X$.

The corresponding causal diagram for this model is shown in Figure 2. The first graph includes the U terms explicitly, though these are often left implicit. In addition the figure shows all possible "realizations" of the graph given the four different possible combinations of the root nodes, S and X, that might arise from U_S and U_X. We illustrate the four possible histories (in which there is no remaining uncertainty, by construction of the model), built in each case by assessing outcomes for each possible combination of S and X values. The arrows indicate the changes that would arise from an intervention that altered each variable independently, given the values realized by all other variables that are not that variable's descendants.[21]

[20] Government sensitivity here can be thought of as government sophistication—does it take the actions of others into account when choosing decisions—or as a matter of preferences—does the government have a dominant strategy to engage in corruption.

[21] Though similar, these graphs are not natural beams or submodels. To construct "natural beams" (Pearl

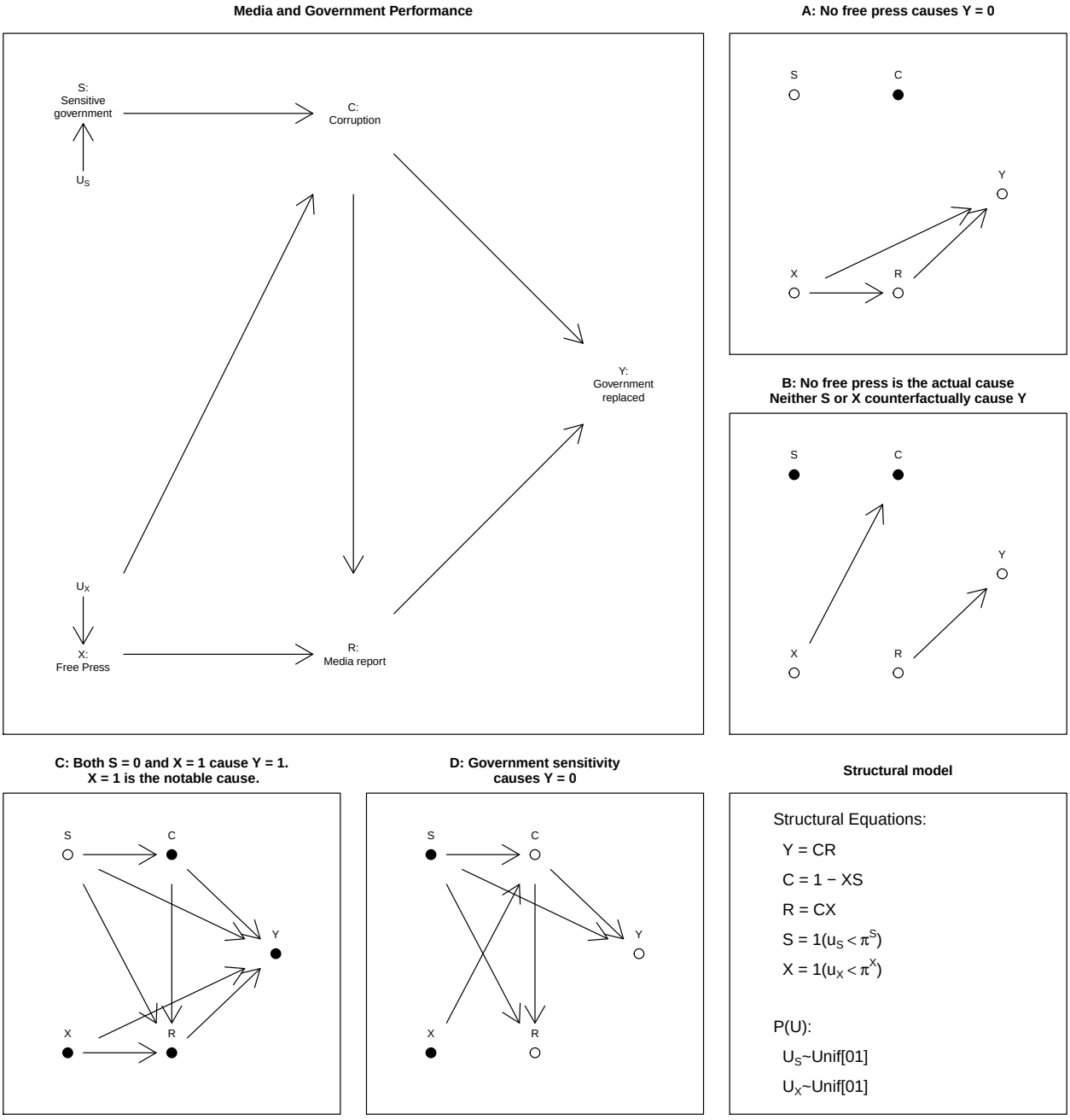


Figure 2: The main panel shows a simple causal model. S and X are stochastic, other variables determined by their parents, as shown in bottom right panel. Other panels show four possible histories that can arise depending on values taken by S and X, along with causal relations in each case. The equations for S and X are written with indicator variables, which take a value of 1 whenever the u value is less than the π value.

These graphs, together with information on π^S and π^X allow us to assess all causal claims of interest. The graphs illustrate, in other words, how causal queries can be represented as the value of the root nodes in a causal diagram

Case-level causal effect. We can read a set of case-level causal effects between two variables off of the submodels in the graph.[22] These submodels are themselves derived from application of Equation 2. We work through an example to demonstrate how this is done.

Consider the effect of R on Y given $S = 0, X = 0$. This is the arrow between R and Y in panel A. Removing the arrows pointing to R, the distribution over nodes when $R = r'$ is: $P(c, y|\hat{r} = r', s = 0, x = 0)$. We are interested in $P(y = 1|\hat{r} = 1, s = 0, x = 0) - P(y = 1|\hat{r} = 0, s = 0, x = 0)$. The second term is easy as for all cases in which $r = 0$, $y = 0$; and so $P(y = 1||\hat{r} = 0) = 0$. We focus then on $P(y = 1|\hat{r} = 1, s = 0, x = 0)$. Taking the marginal distribution, this can be written $\sum_{c=0}^{1} P(y = 1|r = 1, c)P(c|s = 0, x = 0)$. From the structural equations, we know that $P(c = 1|s = 0, x = 0) = 1$ and that $P(y = 1|r = 1, c = 1) = 1$. So the marginal distribution is $P(y = 1|\hat{r} = 1, s = 0, x = 0) = 1$; and the treatment effect of R on Y, conditional on the characteristics of this case, is then 1. This positive effect is represented with the arrow from the $R = 0$ node to the $Y = 0$ node in panel A.

To put the point differently, the subgraphs show that we can determine the effect of R and Y in a case if we know the *value of the root nodes* X and S: R has a positive causal effect on Y in all configurations of root node values (i.e., in all subgraphs) except when $X = 1$ and $S = 1$, in which case R's effect on Y is 0. If X and S are observable, then estimating the case-level effect of R on Y is simply a matter of measuring these two root nodes. If X or S or both are unobservable, then our research design would involve using information from other, observable, nodes to draw inferences about them—a strategy that we discuss below.

Average causal effects. Average causal effects are simply averages of case-level causal effects integrated over the distribution of case-level conditions. The average causal effect thus depends on how commonly the relevant case-level conditions occur. In the logic of the running example, the free press makes a difference if and only if the government is *non-sensitive*: the non-sensitive government gets exposed as corrupt if and only if there is a free press, while the sensitive government never gets replaced since it adjusts by eliminating corruption. Similarly, government quality matters only if there *is* a free press. Without a free press, corrupt governments stay on; with a free press, non-sensitive (and, thus, corrupt) governments get replaced. Put differently, the average effect of each cause in the example depends on the probability with which the other cause is absent or present, and thus is defined over repeated draws of u_S or u_X (for X and S, respectively).

These quantities can be calculated from the distributions in the same way as we calculated the case-level effects. Removing the arrows pointing to R, the distribution over nodes when

2009, 10.3), we fix a realization of root variables, U, (here, $U = (S, X)$); then for each variable, V_i we partition $pa(V_i)$ into a set of "engaged parents," S, and "disengaged parents," with the property that (a) $f_i(S(u), \overline{s}, u) = V_i(u)$ for *all* values of $\overline{s}$ and (b) $f_i(s', \overline{S}(u), u) \neq V_i(u)$ for *some* s'. Thus a natural beam would connect a parent to a child if, given the particular history, the parent mattered for the child's outcome.

[22] These four panels represent submodels in that they reflect outcomes conditional on the values of S and X, but they are not themselves DAGs because they indicate the values taken by nodes and include arrows between two nodes whenever one causes the other, directly or indirectly.

$R = r'$—but this time not fixing S and X—is $P(s, x, c, y|\hat{r} = r')$. Again the key part is $P(y = 1|\hat{r} = 1)$, which can be written $\sum_x \sum_s \sum_c P(x)P(s)P(c|x, s)P(y|c, r = 1)$. Using the structural equations, this simplifies to $\sum_x \sum_s P(x)P(s)P(c = 1|x, s) = P(x = 0)P(s = 0) + P(x = 0)P(s = 1) + P(x = 1)P(s = 0)$, or, $1 - \pi^S\pi^X$.

In the same way, we can construct the average treatment effect for each of the exogenous variables:

- $\tau_X = E_S(Y(X = 1|S) - Y(X = 0|S)) = -(1 - \pi^S)$
- $\tau_S = E_X(Y(S = 1|X) - Y(S = 0|X)) = \pi^X$

We can also arrive at these same quantities by reasoning with the submodel panels in Figure 2. Reading off the presence or absence of the arrows (which represent counterfactual causal effects), we see that R has a causal effect of 1 in panels A, B and C—that is, whenever it is not the case that $X = 1$ and $S = 1$. Thus, the average causal effect is simply $1 - \pi^S\pi^X$, or the probability of not seeing both $X = 1$ and $S = 1$. The average causal effect of R conditional on $S = 1$ is $1 - \pi^X$ (the probability of ending up in panel B, rather than D); and the average causal effect of R given $S = 0$ is 1 (since it has an effect in both panels A and C).

Given the model, data will be useful for estimating average effects only if one is uncertain about the distributions of S and X, which are a function of U_S and π^S and U_X and π^X, respectively. In this example π^S and π^X are fixed in the model and so we do not learn anything about them from data. If however π^S and π^X are represented as nodes that are themselves produced by some other distribution – such as a Beta distribution — then the question of understanding average effects is the question of making inferences about these nodes.

Actual cause. The concept of an actual cause becomes useful when outcomes are overdetermined. Suppose that there is a sensitive government ($S = 1$) and no free press ($X = 0$), as in panel B. Then the *retention* of the government is over-determined: neither government sensitivity nor the free press is a counterfactual cause. Nevertheless, we can distinguish between the causes. Conditioning on there being corruption, if there had been a free press, then the government would have been removed. This would make the lack of a free press an actual cause—that is, a counterfactual cause when the presence of corruption is fixed. The values of the root nodes, S and X, in this case tell us whether such conditioning is permitted for the determination of actual causes. Since corruption *is* present whenever $S = 1$ and $X = 0$, the values in this case, we are permitted to condition on its presence, and the free press is an actual cause of government retention. In contrast, the sensitivity of the government is not an actual cause under these same root node values: with no free press, there is no chance of reporting on corruption; there is thus no subset of actual events, which, when kept fixed, would make a change to a non-sensitive government result in the government's removal.

Notable cause. In the event that that there is a non-sensitive government ($S = 0$) and a free press ($X = 1$), as in panel C, the government gets replaced and *both* of the two causes matter for government replacement. Again however, we can distinguish between them, this time on the basis of both the values of S and X and normality, which depends on π^S and π^X. If for instance governments are frequently non-sensitive, but free presses are rare, i.e. $\pi^X < 1 - \pi^Q$, then the notable cause is the free press.

Causal Paths. Note finally that different causal paths can give rise to the same outcome, where the different paths can be distinguished based on values of root nodes S and X. For example the government may be retained because there is no free press ($X = 0$) and so no negative reporting on the government, regardless of the value of S; or because, there is a free press ($X = 1$) and a sensitive government ($S = 1$) takes account of this and does not engage in corruption.

3 Theories and DAGs

Characterizing beliefs about causal dependencies in terms of causal models and directed acyclic graphs provides a language for formalizing what is meant by a theory and its empirical content.

3.1 What is a theory?

We will say that a causal model (probabilistic or functional), M', is a *theory* of M if M can be derived from M'. In such cases we will refer to M as a *higher*-level model relative to M', and to M' as a *lower*-level model relative to M.[23]

Higher-level models can be generated from lower-level models in two ways, both of which are consistent with common understandings of what it is for a set of claims to constitute or, conversely, derive from a "theory."

1. Aggregating nodes: A higher-level model M', can be a representation of M in which multiple nodes in M' have been aggregated into a single node or in which one or more nodes have been dropped. Conversely, M, can be theorized by a lower-level model, M', in which new nodes have been added and existing nodes split.

For instance, suppose we start with M as represented in Figure 3(a). We can then offer the graph M' in panel (b) as a *theory* of M. Informally, we have added a step in the causal chain between X and Y, a familiar mode of theorization. However, to see why and when adding a node may be helpful for inference we have to formalize how the two models relate to each other.

In the model M, Y is a function of just X and a disturbance U_Y, the latter representing all things other than X than can affect Y. When we add K, X now does not directly affect Y but only does so via K. Further, in the general case, we would explicitly model X as acting on K "with error" by modeling K as a function of both X and U_K. As we emphasize further below, it is in fact only this "error" in the $X \rightarrow K$ link that makes the addition of K potentially informative as a matter of research design: if K were a deterministic function

[23]This definition differs somewhat from that given in Pearl (2009) (p207): there a theory is a (functional) causal model and a restriction over $\times_j \mathcal{R}(U_j)$, that is, over the collection of contexts envisionable. Our definition also considers probabilistic models as theories, allowing statements such as "the average effect of X on Y is 0.5."

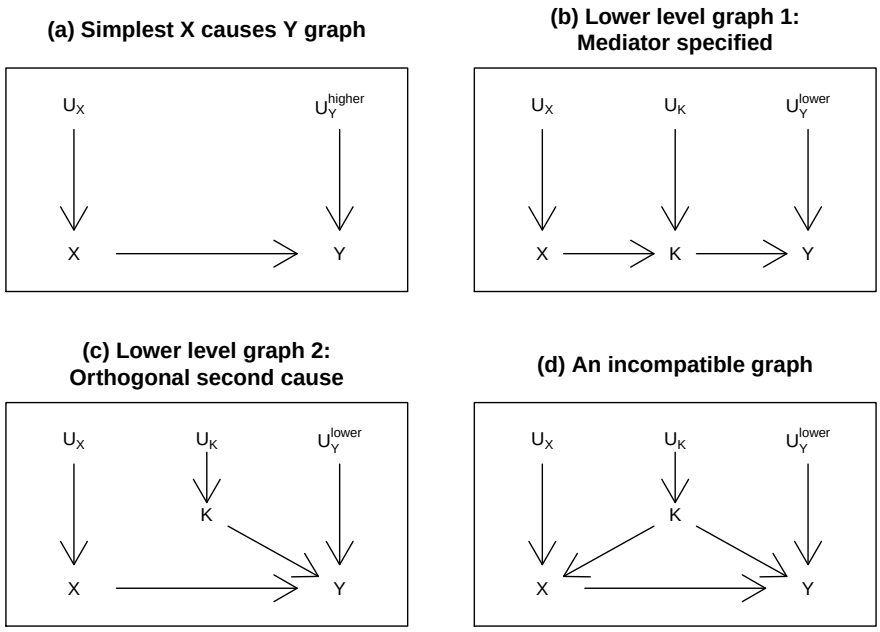


Figure 3: A model with one explanatory variable (top left), two lower level models that can imply it, and one model that does not.

of X only, then knowledge of X would provide full knowledge of K, and nothing could be learned from observing K. What U_K represents, then, is that part of the original U_Y that is a consequence of some variable other than X operating at the *first* step of the causal chain from X to Y. We are thus taking that part of Y not determined by X and splitting it in two: into a non-X input into K and a non-K (and thus also non-X) input into Y. Addition and splitting thus go hand-in-hand: the *insertion* of a mediator between X and Y generally also involves the *splitting* of Y's unspecified parent (U_Y). Importantly, we distinguish between the U_Y's at the two levels by referring to U_Y^{lower} in M' and U_Y^{higher} in M.

Consider next the model in Figure 3(c), which also implies the higher-level theory in panel (a). But the logical relationship between models (a) and (c) is somewhat different. Here the lower level theory *specifies* one or more of the conditions comprising U_Y^{higher}. As we have now extracted U^K from U_Y^{higher}, the unspecified term pointing into Y is now relabelled U_Y^{lower} because it represents a different distribution. M' is a *theory* of M in that it, in a sense, helps explain the dependencies of Y on X more fully than does M.

To turn the situation around, when we move up a level and eliminate a node, we must be careful to preserve all causal dependencies among remaining nodes. In particular, all of the eliminated node's parents become parents of all of that node's children. Thus, for instance in M', since K is a function of both X and U_K, in a higher-level model omitting K, both X and U_K become parents of K's child, Y. Recall that U_K represents the part of K not fully determined by X; thus, to retain all causal determinants of Y in the graph, U_K must (along with X) be carried forward as a parent of Y when K is removed. Rather than drawing two separate U terms going into Y, however, we simply represent the combined root as U_Y^{higher}, with the "higher" signaling the aggregation of roots.

Nodes with no parents in $\mathcal{U} \cup \mathcal{V}$ cannot be eliminated as this would entail a loss of information. The graph in Figure 3(d) illustrates the importance of this. Here K is a cause of both X and Y, in other words it is a possible confounder. A higher-level graph that does not include K still requires a U_K node pointing into both K and Y to capture the fact that there is a confounder.

In Figure 4, we show the permissible reductions of our running example (from Figure 2). We can think of these reductions as the full set of simpler claims (involving at least two nodes) than can be derived from the lower-level theory. In each subgraph, we mark eliminated nodes in grey. Those nodes that are circled must be replaced with U terms. The arrows represent the causal dependencies that must be preserved. Note, for instance, that neither S (because it has a spouse) nor X (because it has multiple children) can be simply eliminated; each must be replaced with an unspecified variable. Also, the higher-level graph with nodes missing can contain edges that do not appear in the lower-level graph: eliminating D, for instance, forces an edge running from X to Y, just as eliminating K produces a $S \rightarrow Y$ arrow. The simplest elimination is of Y itself since it does not encode any feature of dependencies (not conditional on Y itself) between other variables.

We can also read Figure 4 as telling us the set of claims for which the lower-level graph in Figure 2 can serve as a theory. For each reduction, there may be other possible lower-level graphs consistent with it.

One effect of elimination is to render seemingly deterministic relations effectively probabilistic. For example, in the lower level graph C is a deterministic function of X and S. But in higher level graphs it can depend probabilistically on one of these: in submodel 21, C depends probabilistically on X since S is now a stochastic disturbance; in 34 C depends probabilistically on S. This illustrates how unobserved or unidentified features render a model "as-if" stochastic. Conversely, models that exclude this form of uncertainty implicitly claim model-completeness.

Aggregation of nodes may also take the form of "encapsulated conditional probability distributions" (Koller and Friedman 2009) where in a system of nodes, $\{Z_i\}$ is represented by a single node, Z, that takes the parents of $\{Z_i\}$ not in $\{Z_i\}$ as parents to Z and issues the children of (Z_i) that are not in (Z_i) as children.

2. A higher level model may be formed by conditioning on values of nodes in a lower level model. Conversely, a higher-level functional model, M, can be theorized via a lower-level M' in which conditions shaping the operation of the causal effect in M, unspecified in M, are now specified.

To illustrate this approach, consider again the graphs in Figure 3. Above we described how the graph in panel (a) can be produced by aggregating U_Y^{lower} and U_K from panel (c). An alternative possibility is to simplify by conditioning: we derive a higher-level graph from M' by fixing the value of K. For instance, if $Y = XK + U_Y^{lower}$ in M', then at $K = 1$, we have the submodel M_k in which $Y = X + U_Y^{lower}$. Note that, in generating a submodel by conditioning on K, we retain the term U_Y^{lower} as we have not added causal force into Y's unspecified parent.

As we will see, thinking about models as conditionally nested within one another can be empirically useful in providing a way for analysts to more fully specify incomplete higher-level claims by reference to lower-level models within which they are implicitly embedded and thus to make explicit unspecified conditions on which the higher-level relationships depend.

Note that the mapping from theories to higher-level claims may not be one-to-one. A single theory can support multiple higher-level theories. Moreover, a single higher-level relation can be supported by multiple, possibly incompatible lower-level theories. To illustrate, consider two theories:

L_1: $X_1 \rightarrow X_2 \rightarrow Y$

L_2: $X_1 \rightarrow Y \leftarrow X_2$

These two theories record different relations of conditional independence: in L_2, X_1 and X_2 are independent, but they are not independent in L_1. Also, in L_1, X_1 is independent of Y conditional on X_2; but this is not the case in L_2. Now consider the following higher-level models:

H_1: $X_1 \rightarrow Y$

H_2: $X_2 \rightarrow Y$

H_3: $X_1 \rightarrow X_2$

Both H_1 and H_2 are consistent with both L_1 and L_2. However, H_3 can be supported only by L_1 and not by L_2. In addition, the *conditional* higher-level model $((X_1 \rightarrow Y)|X_2 = x_2)$ can be supported by model L_2 but not by model L_1.

Thus multiple (possibly *incompatible*) theories can usually be proposed to explain any given causal effect; and any given theory implies multiple (necessarily *compatible*) causal effects. This suggests that there is no generic sense in which a lower-level theory is more or less general than a higher-level theory. For example, a higher-level theory that is formed by conditioning on a node in a lower-level theory is less general in that it makes sense of fewer cases. On the other hand, a higher-level theory that is formed by aggregating nodes may be more general in that it is consistent with multiple lower-level theories that explain the relationships it contains, even if these lower-level theories are not consistent with each other.

Perhaps surprisingly, in this treatment, the theoretical support for a causal model is itself just another causal model: a set of beliefs about structural relations between variables. Thus, a theory is an object that is formally similar to an empirical claim.

The approach can even handle theoretical propositions in the form of structural causal models, as described above, that make no immediate empirical claims but still have "empirical content" in the sense of being able to inform *conditional* claims. The claim "if X then Y" says nothing about $P(Y)$, but it says a lot if $P(X)$ is known.

This approach allows for an assessment of two features sometimes considered important to assess empirical content of a theory: the level of *universality* of a theory and the degree of *precision* of a theory (Glöckner and Betsch 2011). For instance, consider a theory over X_1, X_2, A, B, Y that specified $X_1, X_2 \rightarrow Y \leftarrow A, B, g$ with functional equations:

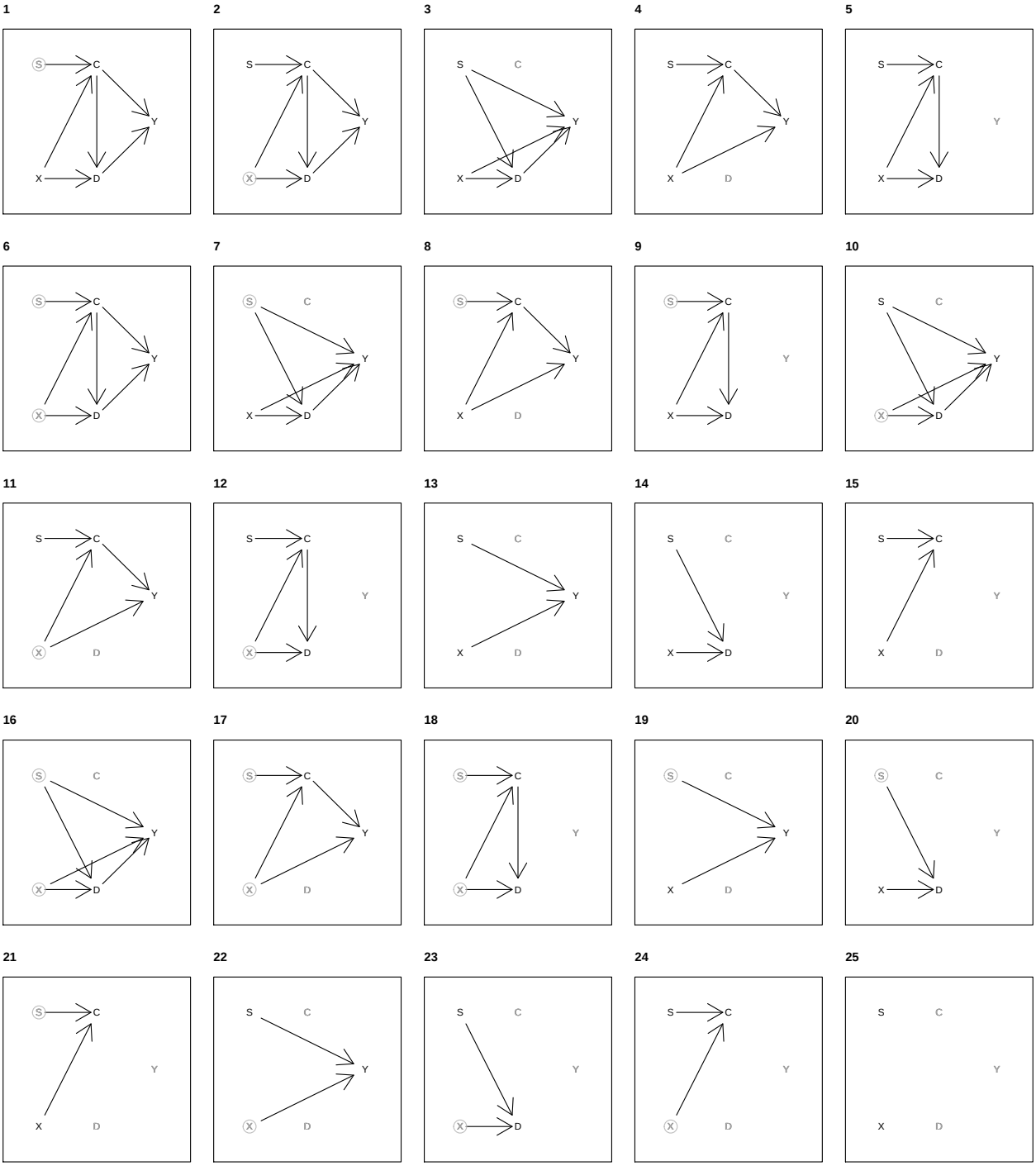


Figure 4: Higher level models derived from the model in Figure 2. Nodes that are eliminated are marked in grey; circles denote root nodes that are replaced in subgraphs by unidentified variables. (A circled node pointing into two children could equivalently be indicated as undirected edge connecting the children.) Note that C, D, and Y are deterministic functions of X and S in this example.

$$Y = \begin{cases} A + BX_1 & \text{if} \quad X_2 = 1 \\ g(X_1) & \text{if} \quad X_2 = 0 \end{cases}$$

where the domain of g, $\mathcal{R}(g)$, is the set of all functions that map from $\mathbb{R}^1$ to $\mathbb{R}^1$, and the ranges of A and B are the real number line. Say the distributions over A, B, X_1, X_2, and g are not specified. Then the theory makes a precise claim conditional on u_1, u_2, X_1, X_2, and g. But since the distribution over $\mathcal{R}(g)$ is not provided by the theory, the theory only claims knowledge of a functional form for Y for those cases in which $X_2 = 1$. Thus in this case the *universality* of the theory for the claim "Y is a linear function of X," is $P(X_2 = 1)$. This is the domain over which the theory has something to say about this proposition. Note that in this case the universality is not provided by the theory, but is rather an external proposition that depends on additional data. The *precision* of the theory depends both on the claim of interest and the distribution of root variables. For example, the precision of the theory for the causal effect of X_1 on Y when $X_2 = 1$ depends on the distribution of B: the theory is more precise about this causal effect the less uncertainty there is about the value of B. Moreover, a theory that specified that B has large variance would be making a precise claim about causal *heterogeneity*, even if it was imprecise about the causal effect. Again this feature cannot be read from the theory without access to ancillary information that the theory itself does not provide.

Functional (but not probabilistic) causal models allow for the representation of logically derived relations between nodes without implying any unconditional empirical claims; that is, all claims may be of the *if-then* variety, as is typical for example of propositions derived from game theoretic models. The process of connecting such models to the empirical claims can be thought of as the embedding of these incomplete models within larger structures.

Consider for example the claim that in normal form games, players play Nash equilibrium. This claim in itself is not a tautology; that is, it is not a result. It can be contrasted for example with the *analytic result* that when rational players play a game and players have common knowledge of the game structure and of player rationality they will only play "rationalizable" strategies. Even still, the Nash claim does provide a set of analytically derived functional equations that relate nodes that describe game forms to actions taken, and from actions to utilities. Representation as a causal graph can make explicit what conditional independencies are assumed in the move from analytic results to empirical claims. For example, are actions independent of the game form conditional on beliefs about the game form; are utilities independent of expectations conditional on actions, and so on.

We give an example of one such model below when we turn to extensive-form games for a lower-level theory that supports our running example.

3.1.1 Assessing the gains from a theory

The observation that theories vary in their precision points to a method for describing the learning that is attributable to a lower-level theory relative to a higher level theory. When a lower-level theory represents a disaggregation, the lower-level theory identifies a set of

potentially observable variables that are not listed by the the higher-level theory. This allows one to assess the gains in precision (for some collection of unobserved variables) that can arise from learning the values of additional observables in the lower-level theory.

Suppose that the contribution of a lower-level theory is to allow for inferences from new data K about some set of query variables Q, after we have already observed variables W from the higher-level model. If we use the expected squared error from the mean posterior estimate as a measure of precision for collection Q, then we have a measure of loss:

$$E_{k,q}\left(\left(\int q' P(q'|k,w)dq' - q\right)^2\right)$$

where the expectation is taken over the joint distribution of K and Q, given W. This is an expected loss—or the *Bayes risk*. The inner term $P(q'|k,w)$ is the posterior distribution on q' given observation of k and w.

Another way to think of the gains is as the expected reduction in the variance of the Bayesian posterior: how certain do you expect you will be after you make use of this new information?

In fact these two quantities are equivalent (see for example Scharf (1991)). Moreover, it is easy to see that whenever inferences are sensitive to K, the expected variance of the posterior will be lower than the variance of the prior. This can be seen from the law of total variance, written here to highlight the gains from observation of K, given what is already known from observation of W.[24]

$$Var(Q|W) = E_{K|W}(Var(Q|K,W)) + Var_{K|W}(E(Q|K,W))$$

The contribution of a theory can then be defined as the mean reduction in Bayes risk:

$$\text{Gains from theory} = 1 - \frac{E_{K|W}(Var(Q|K,W))}{Var(Q|W)}$$

This is a kind of R^2 measure (see also Gelman and Pardoe (2006)).

Other loss functions could be used, including functions that take account of the costs of collecting additional data,[25] or to the risks associated with false diagnoses.[26]

[24] A similar expression can be given for the expected posterior variance from learning K in addition to W when W is not yet known. See, for example, Proposition 3 in Geweke and Amisano (2014).

[25] Further, one might call into question the value of a theory if the gains in precision depend upon data that are practically impossible to gather.

[26] For instance, in Heckerman, Horvitz, and Nathwani (1991), an objective function is generated using expected utility gains from diagnoses generated based on new information over diagnoses based on what is believed already. In their treatment (Heckerman, Horvitz, and Nathwani 1991, Equation 6), the expected value of new information K, given existing information W is: $\sum KP(K|W)(EU(d(Q,W,K)|W,K) - EU(d(Q,W)|W,K))$ where EU is expected utility and d is the optimal inference (diagnosis) given available data. Note that the diagnosis can take account of K when it is observed, but the expected utility depends on K whether or not it is observed, as K carries information about the state of interest.

For illustration say that it is known that $X = 1, Y = 1$ and that, given this information (playing the role of W), the posterior probability that a unit is of type b (and not type d) is p. Say then that a theory specifies that K will take a value 1 with probability ϕ_j if the unit is of type j. Then what is the value added of this theory? Define Q here as the query regarding whether the unit is a b type. Then the prior variance, $Var(Q|W)$, is simply $p(1-p)^2 + (1-p)p^2 = p(1-p)$.

To calculate $E_{K|W}(Var(Q|K,W))$, note that the posterior if K is observed is $\frac{\phi_b p}{\phi_b p + \phi_d(1-p)}$. Let us call this $\hat{q}_K$, and the belief when K is not observed $\hat{q}_{\overline{K}}$. In that case the *expected error* is:

$$\text{Expected Error} = p\phi_b \left(1-\hat{q}_K\right)^2 + (1-p)\phi_d \hat{q}_K^2 + p(1-\phi_b)\left(1-\hat{q}_{\overline{K}}\right)^2 + (1-p)(1-\phi_d)\hat{q}_{\overline{K}}^2$$

where the four terms are the errors when K is seen for a b type, when K is seen for a d type, when K is not seen for a b type, and when K is not see for a d type.

Defining $\rho_K = (p\phi_b + (1-p)\phi_d)$ as the probability of observing K given the prior, we can write the posterior variance as:

$$\text{Expected Posterior Variance} = \rho_K \hat{q}_K(1-\hat{q}_K) + (1-\rho_K)\hat{q}_{\overline{K}}(1-\hat{q}_{\overline{K}})$$

With a little manipulation, both of these expressions simplify to:

$$\text{Expected Posterior Variance} = p(1-p)\left(\frac{\phi_b\phi_d}{\phi_b p + \phi_d(1-p)} + \frac{(1-\phi_b)(1-\phi_d)}{(1-\phi_b)p + (1-\phi_d)(1-p)}\right)$$

The gains are then:

$$\text{Gains} = 1 - \frac{\phi_b\phi_d}{\phi_b p + \phi_d(1-p)} - \frac{(1-\phi_b)(1-\phi_d)}{(1-\phi_b)p + (1-\phi_d)(1-p)}$$

Other natural measures of gains from theory might include the simple correlation between K and Q, or entropy-based measures (see Zhang and Srihari (2003) for many more possibilities).

For this problem the correlation is given by (see appendix):

$$\rho_{KQ} = \frac{(\phi_b + \phi_d)(1-2p)(p(1-p))^{.5}}{(p\phi_b + (1-p)\phi_d)(1-(p\phi_b + (1-p)\phi_d)))^{.5}}$$

One might also use a measure of "mutual information" from information theory:

$$I(Q,K) = \sum_q \sum_k P(q,k)\log\left(\frac{P(q,k)}{P(q)P(k)}\right)$$

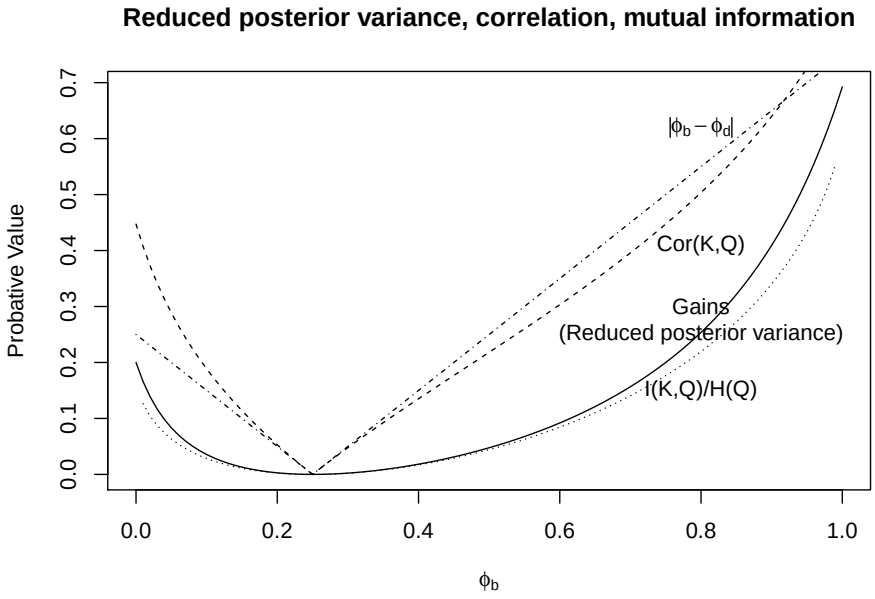


Figure 5: The solid line shows gains in precision (reduced posterior variance) for different values of ϕ_b given $\phi_d = 0.25$ and $p = .5$ for the example given in the text. Additional measures of probative value are also provided including $|\phi_b - \phi_d|$, the correlation of K and Q, and the reduction in entropy in Q due to mutual information in Q and K.

To express this mutual information as a share of variation explained, we could divide $I(Q, K)$ by the entropy of Q, $H(Q)$ where $H(Q) = -\sum_q P(q)\log(P(q))$. The resulting ratio can be interpreted as 1 minus the ratio of the entropy of Q conditional (on K) to the unconditional entropy of Q.

For this example, Figure 5 shows gains as a function of ϕ_b given a fixed value of ϕ_d. The figure also shows other possible measures of probative value, with, in this case, the reduction in entropy tracking the reduced posterior variance closely.

3.2 Illustration of simple theories of moderation and mediation

3.2.1 Mediation as Theory

We begin with a simple theory: there are two binary variables, X and Y, and X causes Y (probabilistically). This theory, such as it is, is represented in Figure 3(a) above.

Although simple, one could imagine many structural equations representing this relationship. For example if u_Y^{higher} is distributed normally and Y takes on the value 1 if $bX + u_Y^{higher}$ is above some threshold, we have a probit model. In a very general formulation we can let u_Y^{higher} be a variable that selects among four different causal types represented with the notation t_{ij}: we read the subscripts to mean that a unit of type t_{ij} has outcome i when $X = 0$ and j when $X = 1$. Then let u_Y^{higher} have a multinomial distribution over the four values of t_{ij} with event probabilities λ_{ij}^{higher}. Note also that in this graph X is independent of u_Y^{higher},

which means that it is as if X is randomly assigned; for example, let $u_X \sim \text{Unif}[0,1]$ and $X = \mathbb{1}(u_K < \pi^K)$.[27]

The functional equation for Y is then given by:

$$Y(x, t_{ij}^{higher}) = \begin{cases} i & \text{if } x = 0 \\ j & \text{if } x = 1 \end{cases}$$

Now consider a theory that specifies a mediating variable between X and Y. This theory is depicted in Figure 3(b) above.

The lower-level functional equations are formally similar though now each unit's outcome (given X) depends on two event probabilities: one that determines type with respect to the effect of X on K (t_{ij}^K), and one with respect to the effect of K on Y (t_{ij}^Y):

$$Y(K, t_{ij}^Y) = \begin{cases} i & \text{if } K = 0 \\ j & \text{if } K = 1 \end{cases}$$

$$K(X, t_{ij}^K) = \begin{cases} i & \text{if } X = 0 \\ j & \text{if } X = 1 \end{cases}$$

Thus, in the lower-level model, there are sixteen types that derive from the cross product of two independent random terms.

Critically, one can derive the higher-level types from the lower level types, and beliefs about the higher level types from beliefs about the lower level types. For example, using the nomenclature in Humphreys and Jacobs (2015):

$$\begin{aligned}
\text{adverse: } t_{10}^{high} &= t_{01}^K \& t_{10}^Y \text{ or } t_{10}^K \& t_{01}^Y \\
\text{beneficial: } t_{01}^{high} &= t_{01}^K \& t_{01}^Y \text{ or } t_{10}^K \& t_{10}^Y \\
\text{chronic: } t_{00}^{high} &= t_{00}^Y \text{ or } t_{00}^K \& t_{01}^Y \text{ or } t_{11}^K \& t_{10}^Y \\
\text{destined: } t_{11}^{high} &= t_{11}^Y \text{ or } t_{00}^K \& t_{10}^Y \text{ or } t_{11}^K \& t_{01}^Y
\end{aligned}$$

In the same way, the higher-level probabilities are implied by the lower level probabilities.

$$\begin{aligned}
\text{adverse: } \lambda_{10}^{high} &= \lambda_{01}^K \lambda_{10}^Y + \lambda_{10}^K \lambda_{01}^Y \\
\text{beneficial: } \lambda_{01}^{high} &= \lambda_{01}^K \lambda_{01}^Y + \lambda_{10}^K \lambda_{10}^Y \\
\text{chronic: } \lambda_{00}^{high} &= \lambda_{00}^Y + \lambda_{00}^K \lambda_{01}^Y + \lambda_{11}^K \lambda_{10}^Y \\
\text{destined: } \lambda_{11}^{high} &= \lambda_{11}^Y + \lambda_{00}^K \lambda_{10}^Y + \lambda_{11}^K \lambda_{01}^Y
\end{aligned}$$

[27] The types here map directly into the four types, a, b, c, d, used in Humphreys and Jacobs (2015) and into principal strata employed by Rubin and others. The literature on probabilistic models also refers to such strata as "canonical partitions" or "equivalence classes." Note that this model is not completely general as the multinomial distribution assumes that errors are iid.

Importantly, even without specifying a distribution over U_K or U_Y^{lower}, a lower-level structural model could be informative by restricting the *ranges* of U_K or U_Y^{lower}. For instance, a lower level theory that imposed a monotonicity condition (no adverse effects) might exclude t_{10}^K and t_{10}^y—that is, increasing X never reduces K, and increasing K never reduces Y.

We return to this example below and show how observation of K can yield inference on causal estimands when the theory places this kind of a structure on theory.

3.2.2 Moderator as Theory

Now consider an alternative lower-level theory. This theory is represented in Figure 3(c) above, in which K is assumed to be a second cause of Y. This graph contains the substantive assumption that K is orthogonal to X as well as the assumption that X is as-if randomly assigned.

In this graph u_K determines the value of K. For example, let $u_K \sim \text{Unif}[0,1]$ and $K = \mathbb{1}(u_K < \pi^K)$. Now u_Y^{lower} is more complex as it determines the mapping from two binary variables into $\{0,1\}$. With this structure, u_Y^{lower} selects among 16 causal types. Let t_{ij}^{gh} denote a unit that has outcome i when $X=0, K=0$, j when $X=1, K=0$, g when $X=0, K=1$, h when $X=1, K=1$. We let u_Y^{lower} in this graph denote a multinomial distribution over the sixteen values of t_{ij}^{gh} with event probabilities λ_{ij}^{gh}.

The sixteen types are illustrated in Table 1 in the appendices.

Again, the types in the higher level mapping are functions of the types in the lower-level mapping. For example, a unit has type t_{01} in the higher level model if $K=1$ and it is of type $t_{00}^{01}, t_{10}^{01}, t_{01}^{01}$, or t_{11}^{01}, or if $K=0$ and it is of type $\lambda_{01}^{00}, \lambda_{01}^{10}, \lambda_{01}^{01}$, or λ_{01}^{11}.

We write this as:

$$t_{01} = ((K=1) \wedge (t^{lower} \in \{t_{00}^{01} \cup t_{10}^{01} \cup t_{01}^{01} \cup t_{11}^{01}\}) \vee ((K=0) \wedge (t^{lower} \in \{\lambda_{01}^{00} \cup \lambda_{01}^{10} \cup \lambda_{01}^{01} \cup \lambda_{01}^{11}\})$$

In the same way, the probability of type t_{01} can be written in terms of the parameters of the lower-level graph. Importantly, the parameters of the higher-level distribution u_Y^{higher} depend on both u_K and u_Y^{lower}. Thus, unlike the mediation case above, the probative value depends on the likelihood of an *observable* event occurring. Specifically, the share of a given higher-level type is given by:

$$\lambda_{ij} = P(u_Y^{higher} = t_{ij}) = \pi^K \left(\lambda_{00}^{gh} + \lambda_{10}^{gh} + \lambda_{01}^{gh} + \lambda_{11}^{gh}\right) + (1-\pi^K)\left(\lambda_{ij}^{00} + \lambda_{ij}^{10} + \lambda_{ij}^{01} + \lambda_{ij}^{11}\right)$$

For example:

$$\lambda_{00} = P(u_Y^{higher} = t_{00}) = \pi^K \left(\lambda_{00}^{00} + \lambda_{10}^{00} + \lambda_{01}^{00} + \lambda_{11}^{00}\right) + (1-\pi^K)\left(\lambda_{00}^{00} + \lambda_{00}^{10} + \lambda_{00}^{01} + \lambda_{00}^{11}\right)$$

Conditional probabilities follow in the usual way. Consider, for instance, the case where it is known that $X = Y = 1$ and so the posterior probability of type t_{01} is simply $P(i \in t_{01}|X = Y = 1) = \frac{\lambda_{01}}{\lambda_{01}+\lambda_{11}}$. Note that π^x does not appear here as this X is orthogonal to u_Y. The probability of type t_{01}, knowing that $X = Y = 1$, can be written in terms of the parameters of the u distributions in the lower-level graph.

$$P(i \in t_{01}|X = Y = 1) = \frac{\pi^K\left(\lambda_{00}^{01} + \lambda_{10}^{01} + \lambda_{01}^{01} + \lambda_{11}^{01}\right) + \left(1 - \pi^K\right)\left(\lambda_{01}^{00} + \lambda_{01}^{10} + \lambda_{01}^{01} + \lambda_{01}^{11}\right)}{\sum_{i=0}^{1}\left(\pi^K\left(\lambda_{00}^{i1} + \lambda_{10}^{i1} + \lambda_{01}^{i1} + \lambda_{11}^{i1}\right) + \left(1 - \pi^K\right)\left(\lambda_{i1}^{00} + \lambda_{i1}^{10} + \lambda_{i1}^{01} + \lambda_{i1}^{11}\right)\right)}$$

We return below to this example and describe how the lower-level model can be used to generate inferences on relations implied by the higher level model.

3.3 Illustration of a Mapping from a Game to a DAG

Our running example supports a set of higher level models, but it can also be *implied* by a lower level models. Here we illustrate with an example in which the lower level model is a game theoretic model, together with a solution.[28]

In Figure 6 we show a game in which nature first decides on the type of the media and the politician – is it a media that values reporting on corruption or not? Is the politician one who has a dominant strategy to engage in corruption or one who is sensitive to the risks of media exposure? In the example the payoffs to all players are fully specified, though for illustration we include parameter b in the voter's payoffs which captures utility gains from sacking a politician that has had a negative story written about them *whether or not they actually engaged in corruption*. A somewhat less specific, though more easily defended, theory would not specify particular numbers as in the figure, but rather assume ranges on payoffs that have the same strategic implications.

The theory is then the game plus a solution to the game. Here for a solution the theory specifies subgame perfect equilibrium.

In the subgame perfect equilibrium of the game; marked out on the game tree (for the case $b = 0$) the sensitive politicians do not engage in corruption when there is a free press – otherwise they do; a free press writes up any acts of corruption, voters throw out the politician if indeed she is corrupt and this corruption is reported by the press.

As with any structural model, the theory says what will happen but also what *would* happen if things that should not happen happen.

To draw this equilibrium as a DAG we include nodes for every action taken, nodes for features that determine the game being played, and the utilities at the end of the game.

[28]Such representations have been discussed as multi agent influence diagrams, for example in Koller and Milch (2003) or White and Chalak (2009) on "settable systems"— an extension of the "influence diagrams" described by Dawid (2002).

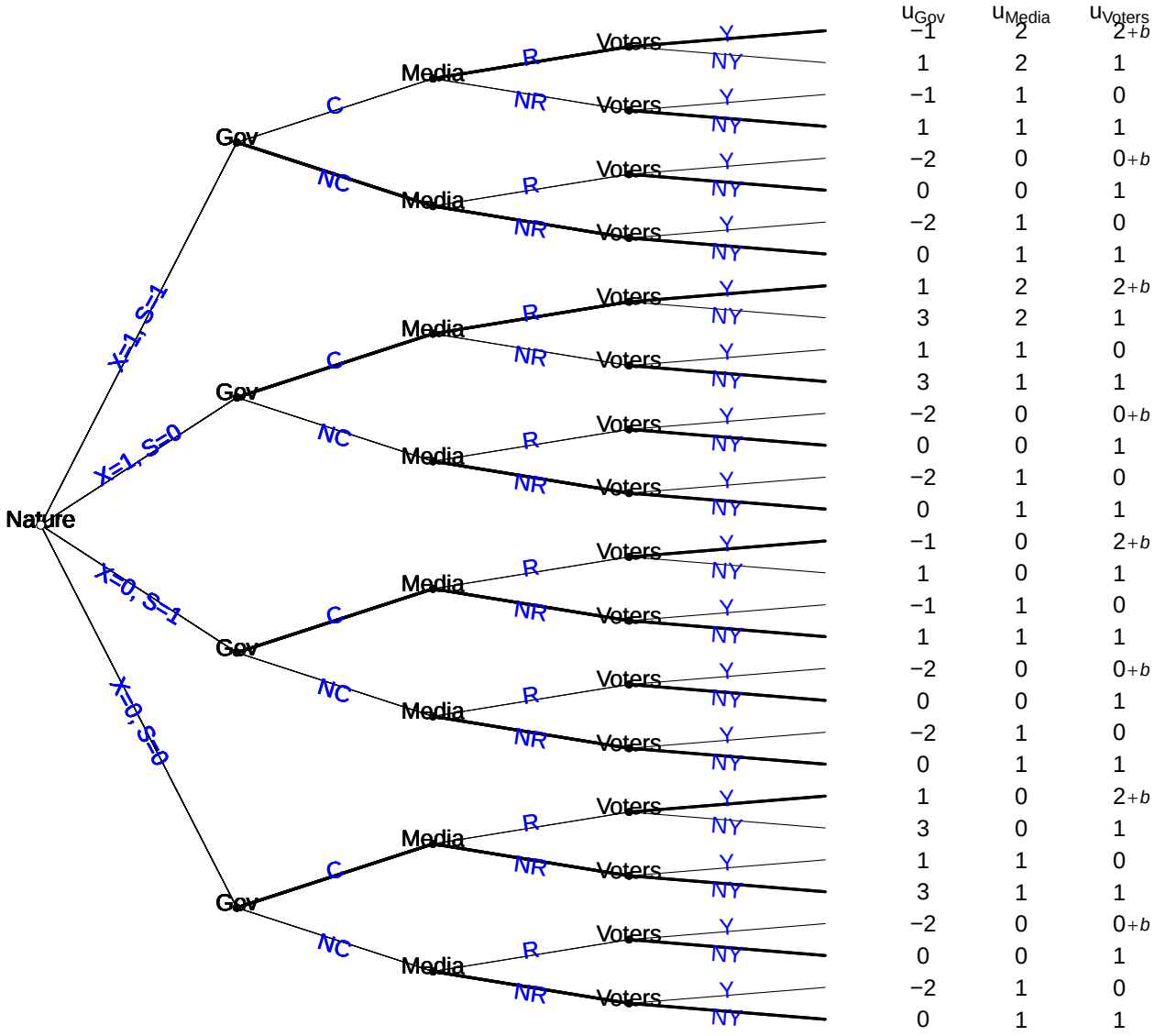


Figure 6: A Game Tree. Solid lines represent choices on the (unique) equilibrium path of the subgames starting after nature's move for the case in which $b = 0$.

If equilibrium claims are justified by claims about the beliefs of actors then these could also appear as nodes. To be clear however these are not required to represent the game or the equilibrium, though they can capture assumed logics underlying the equilibrium choice. For instance a theorist might claim that humans are wired so that whenever they are playing a "Stag Hunt" game they play "defect." The game and this solution can be represented on a DAG without reference to the beliefs of actors about the action of other players. However, if the *justification* for the equilibrium involves optimization given the beliefs of other players, a lower level DAG could represent this by having a node for the game description that points to beliefs about the actions of others, that then points to choices. In a game with dominant strategies, in contrast, there would be no arrows from these beliefs to actions.

For our running example, nodes could usefully include the politician's expectations, since the government's actions depend on expectations of the actions of others. However, given the game there is no gain from including the media's expectations of the voter's actions since in this case the media's actions do not depend on expectations of the voters actions then these expectations should be included.

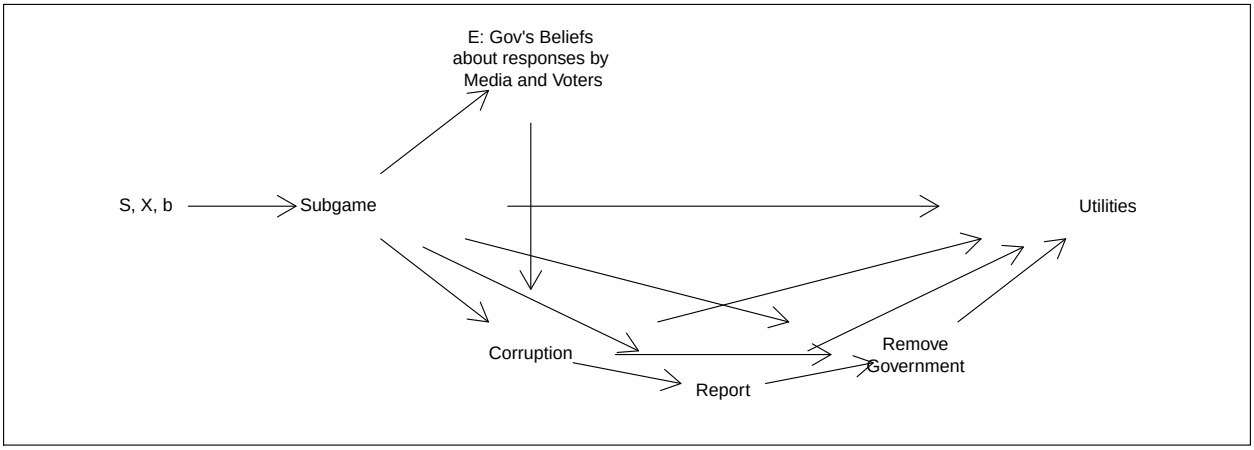
In Figure 7 we provide two examples of DAGs that illustrate lower level models that support our running example.

The upper graph gives a DAG reflecting equilibrium play in the game described in Figure 6. Note that in this game there is an arrow between C and Y even though Y does not depend on C for some values of b—this is because conditional independence requires that two variables are independent for *all* values of the conditioning set. For simplicity also we mark S and X, along with b as features that affect which subgame is being played—taking the subgames starting after Nature's move. Note that the government's expectations of responses by others matters, but the expectations of other players do not matter given this game and solution. Note that the utilities appear twice in a sense. They appear in the subgame node, as they are part of the definition of the game–though here they are the utilities that players expect at each terminal node; when they appear at the end of the DAG they are the utilities that actually arise (in theory at least).

The lower level DAG is very low and much more general, representing the theory that in three player games of complete information, players engage in backwards induction and choose the actions that they expect to maximize utility given their beliefs about the actions of others. The DAG assumes that players know what game is being played ("Game"), though this could also be included for more fundamental justification of behavioral predictions. Each action is taken as a function of the beliefs about the game, the expectations about the actions of others, and knowledge of play to date. The functional equations—not shown—are given by optimization and belief formation assuming optimization by others.

These lower level graphs can themselves provide clues for assessing relations in the higher level graphs. For instance, the lower level model might specify that the value of b in the game affects the actions of the government only through their beliefs about the behavior of voters, E. These beliefs may themselves have a stochastic component, U_E. Thus b high might be thought to reduce the effect of media on corruption. For instance if $b \in \mathbb{R}_+$, we have $C = 1 - FG(1 - \mathbb{1}(b > 1))$. If X is unobserved and one is interested in whether $S = 0$ caused

Lower DAG: Backwards induction in a game with 3 players with one move each



Still lower: Backwards induction, 3 player game with one move for each player

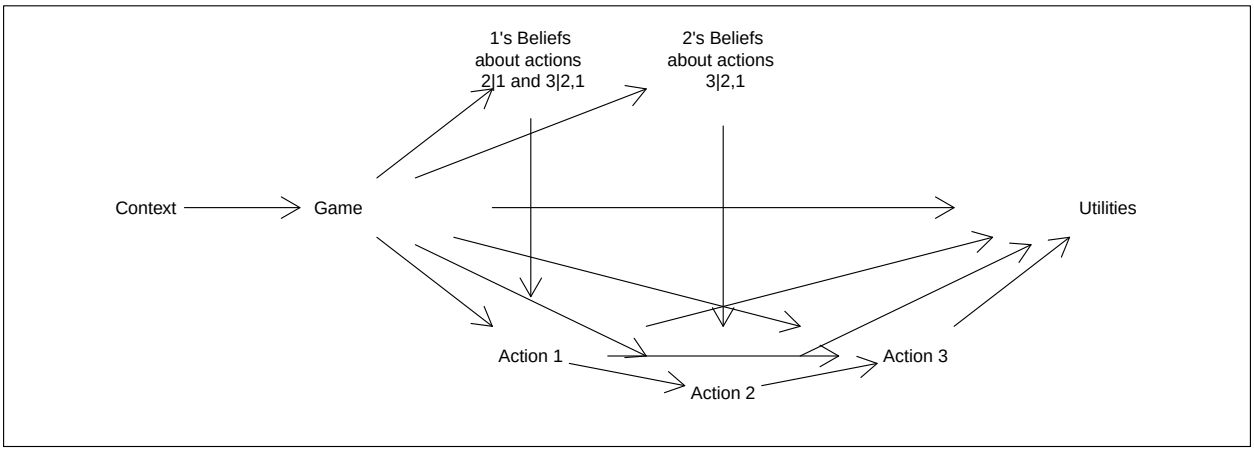


Figure 7: The upper panel shows a causal graph that describes relations between nodes suggested by analysis of the game in Figure 6 and which can imply the causal graph of Figure 2. The game itself (or beliefs about the game) appear as a node, which are in turn determined by exogneous factors. The lower panel represents a still lower level and more general theory "players use backwards induction in three step games of complete information."

corruption, knowledge of b is informative. It is a root node in the causal estimand. If $b > 1$ then $S = 0$ did not cause corruption. However if b matters only because of its effect on E then the query depends on U_E. In this case, while knowing b is informative about whether $S = 0$ caused $C = 1$, knowing E from the lower level graph is more informative.

Note that the model we have examined here involves no terms for U_C, U_R and U_Y—that is, shocks to outcomes given action. Yet clearly any of these could exist. One could imagine a version of this game with "trembling hands," such that errors are always made with some small probability, giving rise to a much richer set of predictions. These can be represented in the game tree as moves by nature between actions chosen and outcomes realized. Importantly in a strategic environment such noise could give rise to different types of conditional independence. For instance say that a Free Press only published its report on corruption with probability π^R, then with π^R high enough the sensitive government might decide it is worth engaging in corruption even if there is a free press; in this case the arrow from X to C would be removed. Interestingly in this case as the error rate rises, R becomes less likely, meaning that the effect of a S on Y becomes gradually weaker (since governments that are not sensitive become more likely to survive) and then drops to 0 as sensitive governments start acting just like nonsensitive governments.

4 DAGs and clues

We have described the general problem of process tracing as using observable data to make inferences about unobserved, or unobservable parts of a causal structure. Is it possible to say when, in general, this is possible? The literature on Bayesian nets gives a positive answer to this question.

4.1 A Condition for probative value

As argued above, causal estimands can be expressed as the values of root nodes in a causal graph. We thus define case-level causal inference as the assessment of the value of unobserved (possibly unobservable) root nodes on a causal graph, given observable data. Let Q denote the set of query variables of interest, W a set of previously observed variables, and K a set of additional variables—clues— that we can design a research project to collect information on. Our interest is thus in knowing whether K is informative about Q given W.

This question is equivalent to asking whether K and Q are conditionally independent given W. This question is answered by the structure of a DAG. The following proposition, with only the names of the variable sets altered, is from Pearl (2009) (Proposition 1.2.4):

Proposition 1: If sets Q and K are d-separated by W in a DAG, G, then Q is independent of K conditional on W in every distribution compatible with G. Conversely, if Q and K are *not* d-separated by W in DAG W, then Q and K are dependent conditional on W in at least one distribution compatible with G.

Thus, given graph G, and observed data, W, a collection of clues K is uninformative about the distribution of query nodes, Q, if K is d-separated from Q by W. And it is possibly informative if K is not d-separated from Q by W.[29]

Note, moreover, that under quite general conditions (referred to in the literature as the *faithfulness* of a probability distribution) then for *some* values of W, K will be informative about Q. That is, there will not be any conditional independencies that are *not* represented in the DAG. This does not, however, mean that we can tell from the DAG alone whether K is informative given *particular* values of W. It is possible, for example, that K and Q are not d-separated by W but that, given $W = w$, K is uninformative about Q. As a simple example, let $q = kw + (1 - w)z$: here, if $W = 1$ then learning K is informative about Q (and Z is not) ; but K is uninformative (and Z is informative) if $W = 0$.

Let us examine Proposition 1 in practice: first, in the simplest case possible, and then for more complex models. The very simplest graph is $X \rightarrow Y$, with X determined by a coin flip. Assuming that there is some heterogeneity—that is, it is unknown in any particular case whether X causes Y—the graph that explicitly includes this heterogeneity is: $X \rightarrow Y \leftarrow Q$. Here, Q determines the value of Y given X. Understanding the causal effect of X means learning Q.

Let us ask what we learn about Q from observing X, Y, and both X and Y. Note that, although there is no variable labelled K in this model, each of X and Y can play the role of informative signals of the values of Q.

In the case where we observe only X, the posterior on Q is:

$$\begin{aligned} P(Q = q|X = x) &= \frac{\sum_{j=0}^{1} p(X = x)P(Q = q)P(Y = j|X = x, Q = q)}{\sum_{q'} \sum_{j=0}^{1} p(X = x)P(Q = q')P(Y = j|X = x, Q = q')} \\ &= \frac{P(Q = q)}{\sum_{q'} P(Q = q')} \end{aligned}$$

which is simply the prior on Q. Thus, nothing is learned about Q from observing X only. This reflects the fact that in the graph, X is d-separated from Q given the empty set. We can see this visually in that there is no active path from X to Q (the path from X to Q is blocked by colliding arrow heads).

[29]This proposition is almost coextensive with the definition of a DAG. A DAG is a particular kind of dependency model ("graphoid") that is as a summary of a collection of "independency statements", (I), over distinct subsets of V (Pearl and Verma 1987), where $I(Q, D, K)$ means "we learn nothing about Q from K if we already know D". More formally:

$$I(K, D, Q) \leftrightarrow P(K, Q|D) = P(K|D)P(Q|D)$$

A Directed Acyclic Graph Dependency model is one where the set of independencies correspond exactly to the relations that satisfy d-separation (Pearl and Verma 1987, p376). Thus on DAG G, $I(K, D, Q)_G$ implies that K and Q are d-separated by D.

In the case where we observe Y only we have:

$$P(Q = q|Y = y) = \frac{\sum_{j=0}^{1} p(X = j)P(Q = q)P(Y = y|X = j, Q = q)}{\sum_{q'} \sum_{j=0}^{1} p(X = j)P(Q = q')P(Y = y|X = j, Q = q')}$$

Here terms involving Y and Q cannot be separated, so the same kind of reduction is not possible. This implies scope for learning about Q from Y. For instance, if we have $P(Q = j) = 1/4$ for type $j \in \{a, b, c, d\}$ and $P(X = j) = \frac{1}{2}$, then we have $P(Q = a|Y = 1) = P(Q = b|Y = 1) = \frac{1}{4}$, $P(Q = c|Y = 1) = 0$ and $P(Q = d|Y = 1) = 1$.

Where we observe both Y and X, we have:

$$P(Q = q|Y = y, X = x) = \frac{P(X = x)P(Q = q)P(Y = y|X = x, Q = q)}{\sum_{q'} P(X = x)P(Q = q')P(Y = y|X = x, Q = q')}$$

which does not allow separation either of Q and X or of Q and Y. Thus, there is again learning from Y and, given Y, there is *also* learning from X. Put differently, we have $P(Q|Y, X) \neq P(Q|Y)$. We can again read this result more simply in terms of d-separation on the graph: given Y, X is no longer $d-$separated from Q because Y is a collider for X and Q. That is, Y *d-connects* X and Q, rendering X informative about Q. Similarly, the informativeness of Y, having observed X, arises from the fact that X does not $d-$separate Q from Y.

More generally, we put this condition to work in Figure 8 by showing different $d-$relations on all graphs of variables X, Y, K, and Q for causal models in which (a) all variables are connected (b) X causes Y directly or indirectly (c) Q causes Y but is not caused by any other variable in the model and is thus a root variable.[30] The titles of the figures report when K is possibly informative about Q depending on whether X, Y, both or none are observed.[31]

A number of features are worth highlighting.

- **Clues at many stages.** K can be informative about Q in cases in which it is pretreatment (with respect to X—e.g. Figure 8(3)), post treatment but pre-outcome (that is, "between" X and Y—e.g. Figure 8(26)) or post-outcome (after Y—e.g. Figure 8(15)). In the case where X is a direct cause of Y, and K is a joint product of X and Y, K can be informative given X or Y, but not both (e.g. Figure 8(31)).
- **Symptoms and surrogates as clues.** The graph $X \rightarrow Y; Y \rightarrow K; Q \rightarrow K, Y$ is one in which the symptoms are clues to the cause (Figure 8(17)). Here the symptoms are informative, even conditional on knowledge of the outcome, because the same underlying features of the case that generate its response type also cause the symptoms. In the simpler graph $X \rightarrow Y; Y \rightarrow K; Q \rightarrow Y$, the symptom, K, is a function of Y but is independent of Q given Y (Figure 8(15)). Thus, here, K is uninformative once the outcome is known. However, here K can be informative, as a surrogate for Y, if the outcome itself is not known.

[30] Graphs generated with Dagitty R package (Textor, Hardt, and Knüppel (2011)).

[31] Note the "possibly" can be dropped under the assumption that the underlying probability model is "stable" and with the interpretation that K is informative about Q for some, but not necessarily all , values of W.

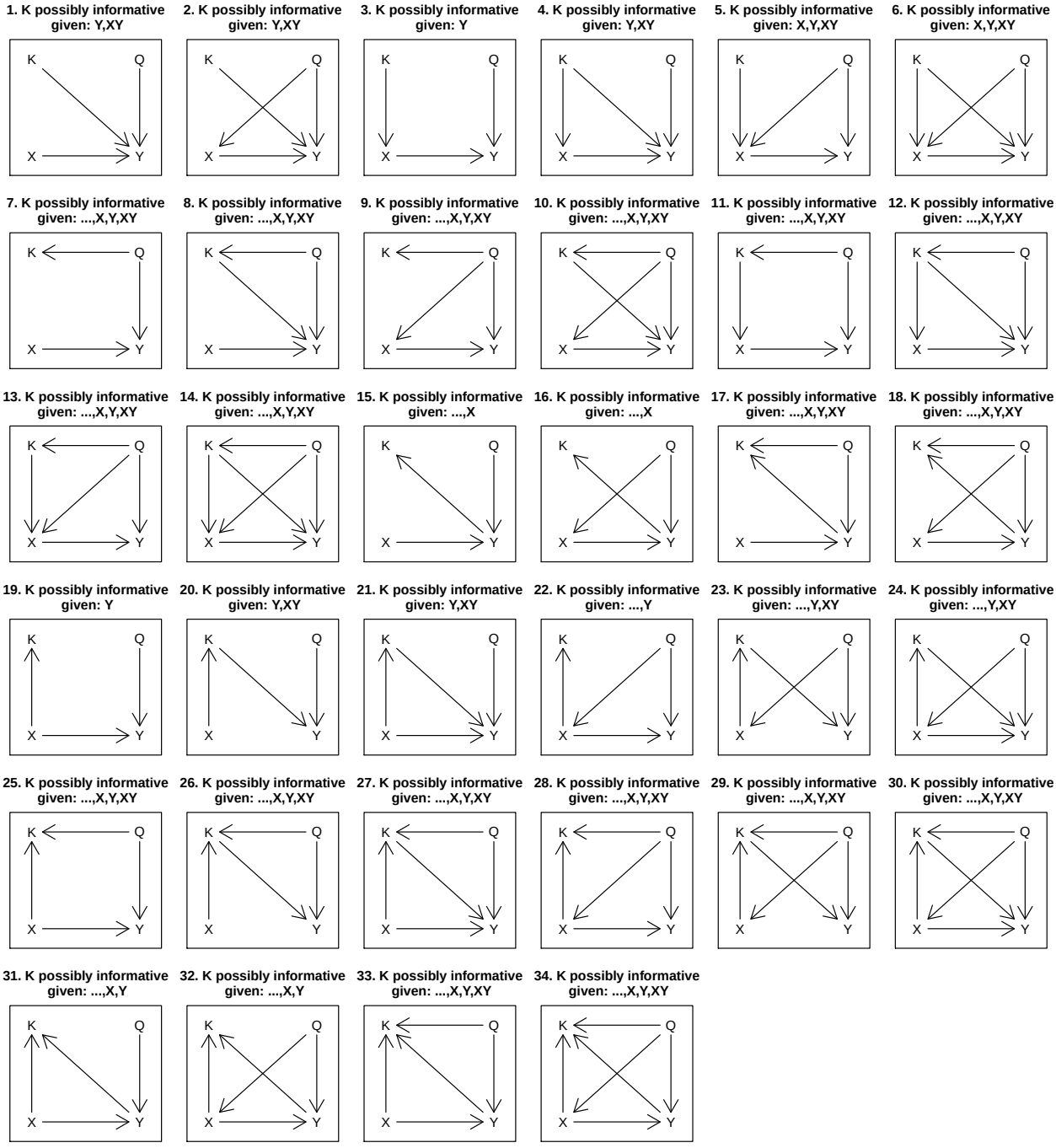


Figure 8: All connected directed acyclic graphs over X, Y, K, Q, in which Q is a root that causes Y, and X is a direct or indirect cause of Y.

- **Instruments as clues.** Consider the graph $K \rightarrow X, Q \rightarrow X, Y; X \rightarrow Y$ (Figure 8(5)). Here, K, is an *instrument* for the effect of X on Y. Notice that, if X is observed, this instrument is informative because Q causes both X and Y. In graphical terms, X is a collider for K and Q, rendering a dependency between the two. The basic logic of instrumentation does not require exactly this causal structure, however. For example, in the case $U \rightarrow X, K; Q \rightarrow X, Y; X \rightarrow Y$, variable K, though not a cause of X is a "surrogate instrument" (Hernán and Robins 2006) as it is a descendant of an unobserved instrument, U. This case can also be seen as one in which there is learning *from the assignment processes*. The graph is similar to one discussed in (Hausman and Woodward 1999) in which there is learning from a pretreatment clue because X is a collider for K and Q. As a simple example one might imagine a system in which $X = K$ if $q \in \{a, b\}$ and $X = 1 - K$ if $q \in \{c, d\}$. Then if we observe, say, $K = Y = K = 1$, we can infer that $q = b$.
- **Mediators and Front-Door Clues**. In graph $X \rightarrow K \rightarrow Y \leftarrow Q$ (Figure 8(20)), the mediating variable K is informative about Q but only when Y is observed, as Y acts as a collider for K and Q.

Note that these figures identify when a clue K is possibly informative for unobserved node Q, a parent of Y. This setup does not, however, fully indicate when clues will be informative about the causal effect of X on Y, or other causal estimands of interest. Even if a clue is uninformative for some parent of Y, it may still help establish whether X causes Y since the statement X causes Y will for some graphs be a statement about a *collection* of nodes that form the set of query variables Q. For example, in the graph $X \rightarrow K \rightarrow Y$, X causes Y if it affects K and if K affects Y. Using our earlier notation, inferring the effect of X on Y requires inferring the value of both u_K and u_Y^{lower}. A clue K that is $d-$separated from u_Y^{lower} may nevertheless be informative about X's effect on Y if it is not $d-$separated from u_K. Additionally, K may be informative about the causal pathway through which X affects Y—even if Y is not observed—again via an inference about U_K.

4.2 Probative value from lower level models of moderation and mediation

So far, we have demonstrated principles of inference from clues, given causal graphs, without imposing any structure on functional forms on the underlying structural equations. We now go deeper, placing structure on functional forms but in a very general way. To do so, we return to the two examples of moderation and mediation discussed in the last section.

4.2.1 Inference from a lower level model of mediating effects

We return to the mediation example described in Figure 3(b) above. Say now, one knows that $X = Y = 1$ (for instance) and the question is, what more can be learned from seeing K?

In this case, before observing K, the belief on t_{01}^{higher}—that is that a unit is a b type—is:

$$P(t_i^{higher} = t_{01}^{higher} | X = Y = 1) = \frac{\lambda_{01}^K \lambda_{01}^Y + \lambda_{10}^K \lambda_{10}^Y}{\lambda_{01}^K \lambda_{01}^Y + \lambda_{10}^K \lambda_{10}^Y + \lambda_{11}^Y + \lambda_{00}^K \lambda_{10}^Y + \lambda_{11}^K \lambda_{01}^Y}$$

Then, if we observe $K = 1$, the posterior is:

$$P(t_i^{high} = t_{01}^{high} | X = Y = 1, K = 1) = \frac{\lambda_{01}^K \lambda_{01}^Y}{\lambda_{01}^K \lambda_{01}^Y + \lambda_{11}^Y + \lambda_{11}^K \lambda_{01}^Y}$$

After observing $K = 0$ the posterior is:

$$P(t_i^{high} = t_{01}^{high} | X = Y = 1, K = 0) = \frac{\lambda_{10}^K \lambda_{10}^Y}{\lambda_{10}^K \lambda_{10}^Y + \lambda_{11}^Y + \lambda_{00}^K \lambda_{10}^Y}$$

Thus updating is possible. For this strategy to work however, K must vary independently of X—that is, u_K has to matter. To see why, imagine that X fully determined K. In this situation, the clue has no probative value since it takes on the same value whether there was a causal effect of X on Y or not; observing X would itself already provide full information on K. Put differently, all variation in causal effects across cases would derive from U_Y^{lower}, and once X has been observed, K would provide no further information about U_Y. Graphically, the region in which $X = 1, K = 1$ is orthogonal to the types region.

More broadly, we can ask: what is the value added of the lower-level model, relative to the higher model. Is the lower-level model *useful*?

As we have discussed, the lower-level model can be thought of as partitioning an unobservable quantity, u_Y^{higher} into a potentially observable quantity K, and an unobservable quantity, u_Y^{lower}.

We can then calculate the expected error after seeing K as:

$$\text{Expected Posterior Var} = \frac{\left(\lambda_{01}^K + \lambda_{11}^K\right) \lambda_{01}^K \lambda_{01}^Y \left(\lambda_{11}^Y + \lambda_{11}^K \lambda_{01}^Y\right)}{\left(\lambda_{01}^K \lambda_{01}^Y + \lambda_{11}^Y + \lambda_{11}^K \lambda_{01}^Y\right)^2} + \frac{\left(1 - \lambda_{01}^K - \lambda_{11}^K\right) \lambda_{10}^K \lambda_{10}^Y \left(\lambda_{11}^Y + \lambda_{00}^K \lambda_{10}^Y\right)}{\left(\lambda_{10}^K \lambda_{10}^Y + \lambda_{11}^Y + \lambda_{00}^K \lambda_{10}^Y\right)^2}$$

which we compare with the prior error: $P(t_{01}^{higher})(1 - P(t_{01}^{higher}))$:

$$\text{Prior variance} = \frac{\left(\lambda_{01}^K \lambda_{01}^Y + \lambda_{10}^K \lambda_{10}^Y\right) \left(\lambda_{11}^Y + \lambda_{00}^K \lambda_{10}^Y + \lambda_{11}^K \lambda_{01}^Y\right)}{\left(\lambda_{01}^K \lambda_{01}^Y + \lambda_{10}^K \lambda_{10}^Y + \lambda_{11}^Y + \lambda_{00}^K \lambda_{10}^Y + \lambda_{11}^K \lambda_{01}^Y\right)^2}$$

To assess the gains provided by the lower-level theory we compare the expected posterior variance with the prior variance. Whether or not these are different depends however on the values of U_K and U_Y^{lower}; the graph structure itself does not guarantee any learning.

To see why, imagine that we want to know X's effect on Y, and we know $X = 1, Y = 1$. Thus, we know that U_Y^{higher}, the higher-level type variable for Y's effect on X, takes on either

t_{01} or t_{11}: either X has had a positive effect on Y or it has no effect because Y would be 1 regardless of X's value. We now want the mediator clue, K, to help us distinguish between these possibilities. Suppose that we then observe $K = 1$. This clue observation eliminates the values of t^K_{00} and t^K_{10} for U_K and the values t^Y_{00} and t^Y_{10} for U^{lower}_Y. In other words, by eliminating these lower-level types, we have excluded the possibility that X has had no effect on Y either because K is always 0 or because Y is always zero. We have also excluded the possibility that X has a positive effect on Y via a negative effect on K followed by a negative effect of K on Y. However, we have not yet eliminated either of the original two *higher*-level types t_{01} or t_{11} defining Y's response to X: the data still allow X to have a positive effect on Y via linked positive effects ($U_K = t^K_{01}$ and $U^{lower}_Y = t^Y_{01}$) or to have no effect, either because K is always 1 ($U_K = t^K_{11}$) or because Y is always 1 ($U^{lower}_Y = t^Y_{11}$) or both. If our prior knowledge gives us no way of distinguishing among the lower-level types—if theory permits and places equal prior probabilities on all of them—then the clue does not allow us to update our beliefs about the higher-level types that are of interest.

On the other hand, if our priors about lower-level types are informative, clues can in turn be informative about higher-level types. To take an extreme example, K would be "doubly decisive" if:

1. $\lambda^K_{01}, \lambda^Y_{01} > 0$: It is possible that $X = 1$ causes $K = 1$ which in turn causes $Y = 1$,
2. $\lambda^K_{10} = 0$ or $\lambda^Y_{10} = 0$: $X = 1$ can only cause $Y = 1$ by first causing $K = 1$, and so seeing $K = 0$ would be sure evidence that X did not cause Y, and
3. $\lambda^Y_{11} = 0$ and $\lambda^K_{11} = 0$: We rule out that K would be 1 no matter what the value of X, or that Y would be 1 no matter what the value of K

On the other hand, nothing would be learned about this causal effect if:

1. $\lambda^K_{10}\lambda^Y_{10} = \lambda^K_{01}\lambda^Y_{01}$; that is, a path via $K = 1$ is as likely as a path through $K = 0$, and
2. $\lambda^K_{00}\lambda^Y_{10} = \lambda^K_{11}\lambda^Y_{01}$. That is, K is just as likely to be 0 or 1 in those situations in which X does not affect K, but K produces $Y = 1$.

These features can also be seen clearly if we write down the probability of observing $K = 1$ conditional on causal type and X, using notation from Humphreys and Jacobs (2015). Here ϕ_{jx} refers to the probability of observing a clue in a case of type j when $X = x$. We can thus derive, for the probabilities of seeing a clue in treated b (positive effect) or d (no effect, Y always 1) types:

$$\begin{aligned}
\phi_{b1} &= \frac{\lambda^K_{01}\lambda^Y_{01}}{\lambda^K_{01}\lambda^Y_{01} + \lambda^K_{10}\lambda^Y_{10}} \\
\phi_{d1} &= \frac{\lambda^Y_{11}(\lambda^K_{01} + \lambda^K_{11}) + \lambda^K_{11}\lambda^Y_{01}}{\lambda^Y_{11} + \lambda^K_{00}\lambda^Y_{10} + \lambda^K_{11}\lambda^Y_{01}}
\end{aligned}$$

These quantities allow for easy mapping between the distributions on variables in $\mathcal{U}$ and the classic process tracing tests in Van Evera (1997). Figure 9 illustrates. In the left panel, we see that as we place a lower prior probability on K's being negatively affected by X, seeking $K = 1$ increasingly takes on the quality of a hoop test for X having a positive effect on

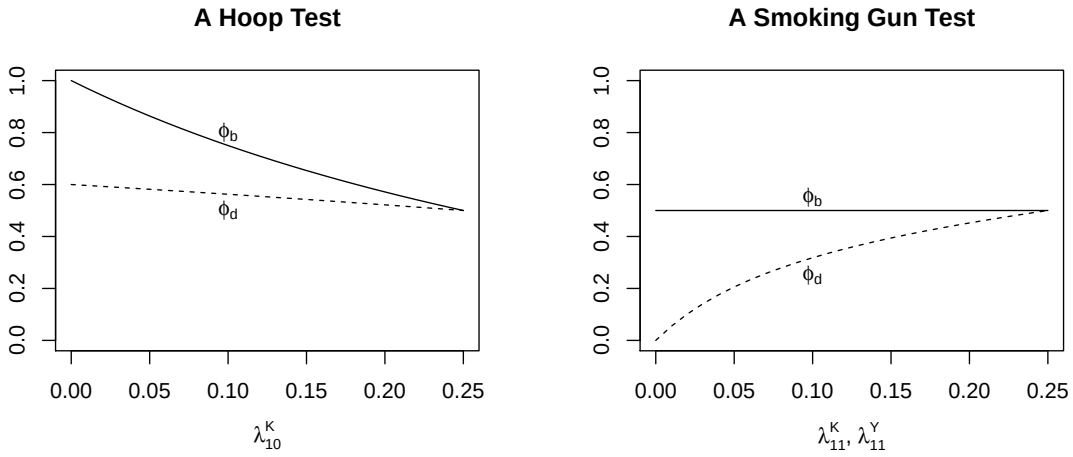


Figure 9: The probability of observing K given causal type for different beliefs on primitives. In the left figure priors on all parameters are flat except for the probability that X is a negative cause of K. In the extreme when one is certain that X does not have a negative effect on K, then K becomes a 'hoop' test for the hypothesis that a unit is of type b. The second figure considers (simultaneous) changes in λ^K_{11} and λ^Y_{11}—the probabilities that K arises regardless of X and Y regardless of K, with flat distributions on all other parameters. With λ^K_{11}, λ^Y_{11} both close to 0, K becomes a 'smoking gun' test for the proposition.

Y: to the extent that $K = 0$ cannot be caused by $X = 1$, then observing $K = 0$ becomes diminishingly consistent with X having a positive causal effect on Y via K. Likewise, as the prior probabilities of K and Y being 1 regardless of the values of X and K, respectively, diminish, seeking $K = 1$ increasingly becomes a smoking-gun test for a positive effect of X on Y, as it becomes increasingly unlikely that $K = Y = 1$ could have occurred without $X = 1$.

At a more informal level, the implication is that a lower-level theory of mediation is useful for inference *to the extent that we have more prior knowledge about smaller, intermediate links in a causal chain than about the $X \rightarrow Y$ relationship taken as a whole.* We are arguably often in this situation. To return to our running example involving government replacement, we may not know much about the frequency with which a free press makes government replacement more likely. However, we may have some prior knowledge indicating that a free press increases reports of government corruption more often than it has no effect; and that greater reports of corruption are more likely to reduce governments' survival in office than to leave their survival prospects untouched. It is precisely those differential weights that we are able to put on causal effects at the lower-level—and not for the higher-level claim of interest—that allow the observation of K to be informative.

4.2.2 Inference from a lower level model of moderating effects

Return now to the example described in Figure 3(c) above. We have shown for this example that the probability of type t_{01} (positive case-level causal effect of X on Y), knowing that $X = Y = 1$, can be written in terms of the parameters of the lower-level graph:

$$P(i \in t_{01}|X=Y=1) = \frac{\pi^K\left(\lambda_{00}^{01}+\lambda_{10}^{01}+\lambda_{01}^{01}+\lambda_{11}^{01}\right)+(1-\pi^K)\left(\lambda_{01}^{00}+\lambda_{01}^{10}+\lambda_{01}^{01}+\lambda_{01}^{11}\right)}{\sum_{i=0}^{1}\left(\pi^K\left(\lambda_{00}^{i1}+\lambda_{10}^{i1}+\lambda_{01}^{i1}+\lambda_{11}^{i1}\right)+(1-\pi^K)\left(\lambda_{i1}^{00}+\lambda_{i1}^{10}+\lambda_{i1}^{01}+\lambda_{i1}^{11}\right)\right)}$$

We can now see that the posterior probabilities after observing K can be written:

$$P(i \in t_{01}|X=Y=1,K=0) = \frac{\left(\lambda_{00}^{01}+\lambda_{10}^{01}+\lambda_{01}^{01}+\lambda_{11}^{01}\right)}{\sum_{i=0}^{1}\left(\lambda_{00}^{i1}+\lambda_{10}^{i1}+\lambda_{01}^{i1}+\lambda_{11}^{i1}\right)} \quad (3)$$

$$P(i \in t_{01}|X=Y=1,K=1) = \frac{\left(\lambda_{01}^{00}+\lambda_{01}^{10}+\lambda_{01}^{01}+\lambda_{01}^{11}\right)}{\sum_{i=0}^{1}\left(\lambda_{i1}^{00}+\lambda_{i1}^{10}+\lambda_{i1}^{01}+\lambda_{i1}^{11}\right)} \quad (4)$$

These posterior probabilities can also be derived by calculating $\phi_{t_{01}1}$ and $\phi_{t_{11}1}$ as used in Humphreys and Jacobs (2015), that is, the probability of observing clue K given causal type. By assumption, K is as-if randomly assigned: in particular (without conditioning on Y), K is orthogonal to u_Y^{lower}—that is, to Y's responsiveness to K. While the probability of observing K is thus the same for all lower-level Y types, that probability is—critically—not the same for all the types in the *higher*-level theory. The higher-level types—those of interest for determining X's effect on Y—are themselves related to the probability of K. Moreover, K is also correlated with the *lower*-level type variable, u_Y^{lower}, *conditional* on Y. In particular, given $Y=1$, we have:

$$\phi_{t_{j1}1} = \frac{\pi^K\left(\lambda_{00}^{j1}+\lambda_{10}^{j1}+\lambda_{01}^{j1}+\lambda_{11}^{j1}\right)}{\pi^K\left(\lambda_{00}^{j1}+\lambda_{10}^{j1}+\lambda_{01}^{j1}+\lambda_{11}^{j1}\right)+(1-\pi^K)\left(\lambda_{j1}^{00}+\lambda_{j1}^{10}+\lambda_{j1}^{01}+\lambda_{j1}^{11}\right)}$$

In this setup, we can think of learning about K as shaping inferences through two channels. When we observe a moderator clue, we learn (1) about the laws governing the case and (2) about the case being governed by those laws.

First, the moderator clue provides information about u_Y^{lower}—that is, how the case *would* respond in different contingencies. For example, suppose that one first believed that the type was either t_{01}^{10} or t_{01}^{01}. That is, a type for which $Y=1$ if either $K=1$ or $X=1$ *but not both*, or it is a case where X causes Y no matter what the value of K. Having observed $X=Y=1$, seeing $K=1$ can rule out the possibility that the unit is of type t_{01}^{10}. We conclude it is of type t_{01}^{01}, and so X caused Y—and *would have cuased Y even if K were 0.*

Second, observing a moderator clue identifies *which* contingencies the case is facing. The clue thus aids inference about X's effect *even if u_Y^{lower} is known.* For instance, suppose it is known that the unit is of type t_{01}^{11}. For this unit $Y=1$ if either $X=1$ of $K=1$. In this case, as X is 1, whether X caused Y or not depends on K. If $K=0$ then X indeed caused Y but not if $K=1$. Using the expressions above, the prior before observing K is $P(i \in t_{01}|X=Y=1) = \frac{\pi^K\lambda_{01}^{11}}{\pi^K\lambda_{01}^{11}+(1-\pi^K)\lambda_{01}^{11}} = \pi^K$, and the posterior after observing $K=1$ (say) is $P(i \in t_{01}|X=Y=1,K=0) = \frac{\lambda_{01}^{11}}{\lambda_{01}^{11}} = 1$.

4.3 Qualitative inferences strategies in the running example

Returning to the running example involving government replacement in Figure 2, we can identify a host of causal estimands of interest and associated strategies of inference. We have shown above how one can use the structural equations in this model to provide a set of conditional causal graphs that let one see easily what caused what at different values of the root nodes S and X. Each of these plots graphs a particular context. We can thus readily see which collection of root nodes constitutes a given query, or estimand. With larger graphs, continuous variables, and more stochastic components, it may not be feasible to graph every possible context; but the strategy for inference remains the same.

For example, suppose one can see that $X = 0$ and $Y = 0$ but does not know the causal effect of X on Y This is equivalent to saying that we know that we are in either panel A or B but we do not know which one. Defining the query in terms of root nodes, the question becomes $S \stackrel{?}{=} 1$, or $P(S = 1|X = 0, Y = 0)$; the difference between the contexts in the two panels is that $S = 0$ when, and only when, $X = 0$ causes $Y = 0$. Given the structural equation for S, $P(S|X = 0, Y = 0) = P(S|X = 0)$, and given independence of X and S, $P(S = 1|X = 0) = \pi^S$. Figuring out S fully answers the query: that is, S is doubly decisive for the proposition.

Graphically what is important is that S is informative not because it is $d-$connected with Y, but because it is $d-$connected to the query variable—here, simply, to itself. For example, suppose that C were already observed together with X and Y. C and X $d-$separate S from Y. Yet S would continue to be informative about the causal effect of X on Y. We can, again, test this by comparing panel A to panel B: the values of C, X, and Y are the same in both graphs; it is only the value of S that distinguishes between the two contexts. This highlights the importance of stating the estimands of interest in terms of root nodes.

We can also see how existing data can make clues uninformative. Say one wanted to know if X causes C in a case. As we can see from inspection of the panels, this query is equivalent to asking whether $S = 1$ (as X causes C only in those two panels, B and D, where $S = 1$. R is unconditionally informative about this query as R is not $d-$separated from S. For example, $R = 1$ implies $S = 0$. However, if C and X are already known, then R is no longer informative because C and X together d-separate R from S. We can come to the same conclusion by reasoning with the graphs: if $X = 0$ and $C = 1$, we know we are in subfigure A or B, and X causes C only in panel B. However, R is of no help to us in distinguishing between the two contexts as it takes the same value in both graphs.

More generally, we can now see in this example how different types of clues can be informative, sometimes conditional on other data. Each of these types corresponds to a set of relations highlighted in Figure 8. We can also read each of these results off of the subpanels in Figure 2.

1. **Informative spouses** Spouses—parents of the same child—can inform on one another. As we have seen in other examples, when an outcome has multiple causes, knowing the value of one of those causes helps assess the effect(s) of the other(s). For example, here, S and X are both parents of C; S is thus informative for assessing whether X causes

C. Indeed this query, written in terms of roots, is simply $P(S)$: X causes C if and only if $S = 1$. Likewise, S causes C (negatively) if and only if $X = 1$.

2. **Pre-treatment clues.** Did the absence of media reports on corruption ($R = 0$) cause government survival ($Y = 0$)? Look to the pre-treatment clue, X: $X = 0$ is a smoking gun establishing that the absence of a report produced government survival. Or, substantively, if there were a free press, then a missing report would never be a cause of survival since it would occur only in the absence of corruption, which would itself be sufficient for survival. More broadly, this example illustrates how knowledge of selection into treatment can be informative about treatment effects.

3. **Post-outcome clues.** Suppose we observe the presence of a free press ($X = 1$) and want to know if it caused a lack of corruption ($C = 0$), but cannot observe the level of corruption directly. Observing Y—which occurs after the outcome—is informative here: if $X = 1$, then X causes C (negatively) if and only if $Y = 0$. When an outcome is not observed, a consequence of that outcome can be informative about its value and, thus, about the effect of an observed suspected cause.

4. **Mediators as clues**: We see a politically sensitive government ($S = 1$) and its survival ($S = 0$). Did the government survive because of its sensitivity to public opinion? Here, the mediation clue C is helpful: a lack of corruption, $C = 0$, is evidence of S's negative effect on Y.

While the above examples focused on case level causal effects, we can also how clues are informative for different types of estimand:

1. **Average casual effects.** Analysis of a single case is informative about average causal effects if there is uncertainty about the distributions of root nodes. Recall that the values of X and S, in the model, are determined by the parameters π^X and π^S (not pictured in the graph). Recall, further, that average causal effects are functions of these parameters: in a model in which π^X is explicitly included as a root, the query "what is the average causal effect of S on Y" is $P(\pi^X)$. Simple observation of Y is informative about the value of X and, in turn, about $P(\pi^X)$. If we start with a Beta prior with parameters α, β over π^X and we observe $Y = 1$, say, then the posterior is:

$$P(\pi^X|Y=1) = \frac{P(Y=1|\pi^X)P(\pi^X)}{P(Y=1)} = \frac{P(X=1,S=1|\pi^X)P(\pi^X)}{P(X=1,S=1)} = \frac{P(X=1|\pi^X)P(\pi^X)}{P(X=1)}$$

   Note that S drops out of the expression because of the assumed independence of X and S. We are left with the problem of updating a belief about a proportion, π^X, in a population given a positive draw from the population: $(P(\pi^X|X=1))$. In the case of the Beta prior, the posterior would be $Beta(\alpha+1, \beta)$. Note also that, using the $d-$separation criterion, learning Y would be uninformative if we already knew X or if we already knew C and R, as each of these sets of nodes blocks Y from π^X, a parent of X.

2. **Actual cause, notable cause.** Consider now the query: is $X = 0$ an *actual cause* of $Y = 0$? The definition of actual causes, together with our structural model, would

require that that $X = 0$ and $Y = 0$, and given $X = 0$ the condition for X to make a difference is $C = 1$. So our query is $C \stackrel{?}{=} 1$ which in terms of roots is $P((X = 0)\&((X = 0)\text{or}(X = 1\&S = 0)))$, which is simply equal to $P(X = 0)$. This means that in this example, whenever $X = 0$, X is an actual cause of $Y = 0$. X itself is decisive about this query. The likelihood that $X = 0$ will actually cause $Y = 0$ is $1 - \pi^X$. The query whether $X = 0$ is a *notable* cause of $Y = 0$ in a case is a query about both π^X and u_X (as u_X causes X given π^X).

3. **Path estimands.** Consider now a causal path as an estimand; for example "does X cause R which in turn causes Y"; this path arises in panel A and panel C only and so in terms of roots the query is $S \stackrel{?}{=} 0$. S is doubly decisive for this proposition. For such paths note that transitivity does not necessarily follow: for example $S = 1, X = 1$ is evidence for the path "Free press prevents corruption and the absence of corruption causes the government to survive," but here the free press does not cause survival.

5 Conclusion

Qualitative methodologists in political science have been exploring how process-tracing strategies can be formalized using Bayesian logic (e.g., Bennett (2015), Fairfield and Charman (2015), Humphreys and Jacobs (2015)). This move to formalization has helped clarify how within-case information can be used to support causal claims in the study of a single case. In spelling out the logic of inference more fully, this literature also clarifies the oftentimes strong assumptions about the probative value of clues that are required for qualitative causal inference. Work on Bayesian process tracing encourages case-study researchers to justify their inferences by reference to beliefs about the likelihood of the evidence under alternative hypotheses. Yet, in doing so, this literature also raises the question of where beliefs about probative value should come from.

Fortunately, the formalization of the basic inference strategy used in process tracing opens connections to a very large body of research on probabilistic models developed in computer science, biomedics, and philosophy. Drawing on insights from this literature, the strategy for justifying inferences that we explore here is one in which the researcher makes use of background knowledge about a domain to embed process-tracing clues within a causal model of the phenomenon of interest. A formalization of the logic of process tracing inference, together with a formalization of theoretical priors, permits researchers to provide a clear account of the probative value of within-case evidence, conditional on theory. Techniques developed in the study of probabilistic models then let one assess the informational flows between variables given a structural model in order to mark out a strategy to form posteriors over estimands of interest, and in doing so update the theory.

The procedure that emerges involves three steps: representing existing knowledge as a causal graph and underlying structural model; defining causal queries in terms of root nodes on that graph; and identifying variables that are informative about these query nodes given already-observed data. This last step may also require recourse to lower-level models that provide

theoretical justification for inference strategies. While the basic properties of structural and graphical models that we exploit are well understood, they have not, to our knowledge, been used to underpin the inferential logic of process tracing.

The approach to inference that we describe here is quite general. We have focused our discussion on inferences in single cases, as is common in accounts and the practice of process tracing, yet the logic extends to multiple cases, mixed data, and population or super-population inference. For extension to multiple cases, some nodes in a causal model can be interpreted as vectors, with an entry for each case. For mixed-methods inference, some vectors can contain data sought for only subsets of cases while other nodes contain complete data: for example, the researcher might collect data on dependent and independent variables for all cases, but data on other parts of the causal network for just a subsample. For superpopulation estimates, nodes can include population and superpopulation parameters feeding into the causal network.

While we have focused on inference from single clues, the approach accommodates multiple clues in a simple way. Proposition 1 is a statement about the flow of information between sets of random variables, and so covers both a situation in which estimands are formed from statements about multiple nodes and a situation in which the within-case data involves observations of multiple nodes. Moreover, the assessment of gains from clues always conditions on existing data, which opens up strategies for assessing gains from the sequencing of clue information. Given many possible clues, for instance, the approach can tell the analyst who observes one clue whether there are gains to also observing a second clue.[32]

The formalization of process tracing strategies using structural causal models brings a number of benefits. First, the full specification of a model can allow for an unpacking of the notion of a causal type, allowing for a partitioning of uncertainty around causal responses into those parts that are explained by observable variables and those parts left unexplained by the causal model. Second, in clarifying the role of theory in generating probative value, we can conceptualize and assess the gains from theory to causal inference. Third, we can learn about more complex causal estimands than causal effects. The specification of theories that contain more complete descriptions of causal structure may shift attention towards learning about models of the world, rather than generating model-free estimates of average effects. Last, the connection to Bayesian nets helps clarify how causal forces move through theoretical structures, and in doing so provides access to relatively simple strategies for assessing when more within-case information can, or certainly does not, have probative value.

[32]As a simple illustration, if the theory takes the form $Q \rightarrow K_1 \rightarrow K_2 \rightarrow W$, observing W does not make observing K_1 or K_2 uninformative; but observing K_1 renders K_2 uninformative by the logic of d-separation. Optimal clue choice has been a long-standing concern in medical diagnostics. Heckerman, Horvitz, and Nathwani (1991) developed a tool PATHFINDER that suggested evidence based on possible inferences from a causal graph.

6 Appendix

We provide details on probative value from a moderation model. The sixteen types implied by u_Y^{lower} in the model described in section 4.2.2 are as show in Table 1 .

Type	Label	$(Y\|X=0, C=0)$	$(Y\|X=1, C=0)$	$(Y\|X=0, C=1)$	$(Y\|X=1, C=1)$
t_{00}^{00}	chronic	0	0	0	0
t_{00}^{01}	jointly-beneficial	0	0	0	1
t_{00}^{10}	2-alone-beneficial	0	0	1	0
t_{00}^{11}	2-beneficial	0	0	1	1
t_{01}^{00}	1-alone-beneficial	0	1	0	0
t_{01}^{01}	1-beneficial	0	1	0	1
t_{01}^{10}	any-alone-beneficial	0	1	1	0
t_{01}^{11}	any-beneficial	0	1	1	1
t_{10}^{00}	any-adverse	1	0	0	0
t_{10}^{01}	any-alone-adverse	1	0	0	1
t_{10}^{10}	1-adverse	1	0	1	0
t_{10}^{11}	1-alone-adverse	1	0	1	1
t_{11}^{00}	2-adverse	1	1	0	0
t_{11}^{01}	2-alone-adverse	1	1	0	1
t_{11}^{10}	jointly-adverse	1	1	1	0
t_{11}^{11}	destined	1	1	1	1

Table 1: Types given two treatments (or one treatment and one covariate)

The posterior probability on higher level types, derived directly from lower level types in the text, can also by derived using Bayes' rule and the ϕ values, as done in Humphreys and Jacobs (2015):

$$P(t_i = t_{01}|X = Y = 1, K = 1) = \frac{\phi_{t_{01}1} P(t_i = t_{01}|X = Y = 1)}{\phi_{t_{01}1} P(t_i = t_{01}|X = Y = 1) + \phi_{t_{11}1} P(t_i = t_{11}|X = Y = 1)}$$

Note that $\phi_{t_{01}1} P(t_i = t_{01}|X = Y = 1)$ can be written:

$$\phi_{t_{01}1} P(t_i = t_{01}|X = Y = 1) = \frac{\pi^K \left(\lambda_{00}^{01} + \lambda_{10}^{01} + \lambda_{01}^{01} + \lambda_{11}^{01}\right)}{\sum_{i=0}^{1} \left(\pi^K \left(\lambda_{00}^{i1} + \lambda_{10}^{i1} + \lambda_{01}^{i1} + \lambda_{11}^{i1}\right) + \left(1 - \pi^K\right) \left(\lambda_{i1}^{00} + \lambda_{i1}^{10} + \lambda_{i1}^{01} + \lambda_{i1}^{11}\right)\right)}$$

with a similar expression for $\phi_{t_{11}1} P(t_i = t_{11}|X = Y = 1)$, then entering into the Bayes equation the denominators and the π^K terms cancel out giving the expression for $P(i \in t_{01}|X = Y = 1, K = 0)$ in the text.

References

Bennett, Andrew. 2015. "Appendix." In *Process Tracing: From Metaphor to Analytic Tool*, edited by Andrew Bennett and Jeffrey Checkel. New York: Cambridge University Press.

Bennett, Andrew, and Jeffrey Checkel. 2015. "Process Tracing: From Philosophical Roots to Best Practices." In *Process Tracing: From Metaphor to Analytic Tool*, edited by Andrew Bennett and Jeffrey Checkel, 3–37. New York: Cambridge University Press.

Collier, David. 2011. "Understanding Process Tracing." *PS: Political Science & Politics* 44 (04). Cambridge Univ Press: 823–30.

Collier, David, Henry E Brady, and Jason Seawright. 2010. "Sources of Leverage in Causal Inference: Toward an Alternative View of Methodology." In *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, edited by David Collier and Henry E Brady, 161–99. Lanham MD: Rowman; Littlefield.

Darwiche, Adnan, and Judea Pearl. 1994. "Symbolic Causal Networks." In *AAAI*, 238–44.

Dawid, A Philip. 2002. "Influence Diagrams for Causal Modelling and Inference." *International Statistical Review* 70 (2). Wiley Online Library: 161–89.

Fairfield, Tasha, and Andrew Charman. 2015. "Formal Bayesian Process Tracing: Guidelines, Opportunities, and Caveats."

Frangakis, Constantine E, and Donald B Rubin. 2002. "Principal Stratification in Causal Inference." *Biometrics* 58 (1). Wiley Online Library: 21–29.

García, Fernando Martel, and Leonard Wantchekon. 2015. "A Graphical Approximation to Generalization: Definitions and Diagrams." *Journal of Globalization and Development* 6 (1): 71–86.

Gardner, Martin. 1961. *The Second Scientific American Book of Mathematical Puzzles and Diversions*. Simon; Schuster New York.

Gelman, Andrew, and Iain Pardoe. 2006. "Bayesian Measures of Explained Variance and Pooling in Multilevel (Hierarchical) Models." *Technometrics* 48 (2). Taylor & Francis: 241–51.

George, Alexander L, and Andrew Bennett. 2005. *Case Studies and Theory Development in the Social Sciences*. Cambridge, MA: MIT Press.

Gerring, John. 2006. *Case Study Research: Principles and Practices*. New York: Cambridge University Press.

Geweke, John, and Gianni Amisano. 2014. "Analysis of Variance for Bayesian Inference." *Econometric Reviews* 33 (1-4). Taylor & Francis: 270–88.

Glöckner, Andreas, and Tilmann Betsch. 2011. "The Empirical Content of Theories in Judgment and Decision Making: Shortcomings and Remedies." *Judgment and Decision Making* 6 (8). Society for Judgment & Decision Making: 711.

Glynn, Adam, and Kevin Quinn. 2007. "Non-Parametric Mechanisms and Causal Modeling."

working paper.

Hall, Ned. 2004. "Two Concepts of Causation." *Causation and Counterfactuals*. Cambridge, MA: MIT Press, 225–76.

Hall, Peter A. 2003. "Aligning Ontology and Methodology in Comparative Research." In *Comparative Historical Analysis in the Social Sciences*, edited by James Mahoney and Dietrich Rueschemeyer, 373–404. New York: Cambridge University Press; Cambridge University Press.

Halpern, Joseph Y. 2015. "A Modification of the Halpern-Pearl Definition of Causality." *ArXiv Preprint ArXiv:1505.00162*.

———. 2016. *Actual Causality*. MIT Press.

Halpern, Joseph Y, and Judea Pearl. 2005. "Causes and Explanations: A Structural-Model Approach. Part I: Causes." *The British Journal for the Philosophy of Science* 56 (4). Br Soc Philosophy Sci: 843–87.

Hausman, Daniel M, and James Woodward. 1999. "Independence, Invariance and the Causal Markov Condition." *The British Journal for the Philosophy of Science* 50 (4). Br Soc Philosophy Sci: 521–83.

Heckerman, David E, Eric J Horvitz, and Bharat N Nathwani. 1991. "Toward Normative Expert Systems: The Pathfinder Project." *Methods of Information in Medicine* 31: 90I105.

Hernán, Miguel A, and James M Robins. 2006. "Instruments for Causal Inference: An Epidemiologist's Dream?" *Epidemiology* 17 (4). LWW: 360–72.

Huber, John. 2013. "Is Theory Getting Lost in the 'Identification Revolution'?" *Washington Post Monkey Cage: Retrieved June* 15: 2013.

Hume, David, and Tom L Beauchamp. 2000. *An Enquiry Concerning Human Understanding: A Critical Edition*. Vol. 3. Oxford University Press.

Humphreys, Macartan, and Alan M Jacobs. 2015. "Mixing Methods: A Bayesian Approach." *American Political Science Review* 109 (04). Cambridge Univ Press: 653–73.

Koller, Daphne, and Nir Friedman. 2009. *Probabilistic Graphical Models: Principles and Techniques*. MIT press.

Koller, Daphne, and Brian Milch. 2003. "Multi-Agent Influence Diagrams for Representing and Solving Games." *Games and Economic Behavior* 45 (1). Elsevier: 181–221.

Lewis, David. 1973. "Counterfactuals and Comparative Possibility." In *Ifs*, 57–85. Springer.

———. 1986. "Causation." *Philosophical Papers* 2: 159–213.

Mahoney, James. 2007. "The Elaboration Model and Necessary Causes." *Explaining War and Peace: Case Studies and Necessary Condition Counterfactuals*. Routledge, 281–306.

———. 2010. "After Kkv: The New Methodology of Qualitative Research." *World Politics* 62

(01). Cambridge Univ Press: 120–47.

Menzies, Peter. 1989. "Probabilistic Causation and Causal Processes: A Critique of Lewis." *Philosophy of Science*. JSTOR, 642–63.

Paul, Laurie Ann, and Edward Jonathan Hall. 2013. *Causation: A User's Guide*. Oxford University Press.

Pearl, Judea. 2009. *Causality*. Cambridge university press.

———. 2012. "The Causal Foundations of Structural Equation Modeling." DTIC Document.

Pearl, Judea, and Thomas Verma. 1987. *The Logic of Representing Dependencies by Directed Graphs*. University of California (Los Angeles). Computer Science Department.

Rohlfing, Ingo. 2013. "Comparative Hypothesis Testing via Process Tracing." *Sociological Methods & Research* 43 (04). Sage Publications: 0049124113503142.

Scharf, Louis L. 1991. *Statistical Signal Processing*. Vol. 98. Addison-Wesley Reading, MA.

Splawa-Neyman, Jerzy, DM Dabrowska, TP Speed, and others. 1990. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." *Statistical Science* 5 (4). Institute of Mathematical Statistics: 465–72.

Textor, Johannes, Juliane Hardt, and Sven Knüppel. 2011. "DAGitty: A Graphical Tool for Analyzing Causal Diagrams." *Epidemiology* 22 (5). LWW: 745.

Van Evera, Stephen. 1997. *Guide to Methods for Students of Political Science*. Ithaca, NY: Cornell University Press.

Waldner, David. 2015. "What Makes Process Tracing Good? Causal Mechanisms, Causal Inference, and the Completeness Standard in Comparative Politics." In *Process Tracing: From Metaphor to Analytic Tool*, edited by Andrew Bennett and Jeffrey Checkel, 126–52. New York: Cambridge University Press.

Weller, Nicholas, and Jeb Barnes. 2014. *Finding Pathways: Mixed-Method Research for Studying Causal Mechanisms*. Cambridge, UK: Cambridge University Press.

White, Halbert, and Karim Chalak. 2009. "Settable Systems: An Extension of Pearl's Causal Model with Optimization, Equilibrium, and Learning." *Journal of Machine Learning Research* 10 (Aug): 1759–99.

Yamamoto, Teppei. 2012. "Understanding the Past: Statistical Analysis of Causal Attribution." *American Journal of Political Science* 56 (1). Wiley Online Library: 237–56.

Zhang, Bin, and Sargur N Srihari. 2003. "Properties of Binary Vector Dissimilarity Measures." In *Proc. Jcis Int'l Conf. Computer Vision, Pattern Recognition, and Image Processing*. Vol. 1.