University of California
Los Angeles

Generalizability in Causal Inference: Theory and Algorithms

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Computer Science

by

Elias Bareinboim

2014

© Copyright by
Elias Bareinboim
2014

Abstract of the Dissertation

Generalizability in Causal Inference: Theory and Algorithms

by

Elias Bareinboim
Doctor of Philosophy in Computer Science
University of California, Los Angeles, 2014
Professor Judea Pearl, Chair

In the empirical sciences, experiments are invariably conducted with the intent of being used elsewhere (e.g., outside the laboratory), where conditions are likely to be different. This practice is based on the premise that, owing to certain commonalities between the source and target environments, causal claims will be valid even where experiments have never been performed. Yet, despite the extensive amount of empirical work relying on this premise, practically no formal treatments have been attempted to reveal the conditions under which environments can differ and still allow, in some formal sense, generalizations to be valid.

This work develops a theoretical framework for understanding, representing, and algorithmizing the generalization problem described above and brings other types of generalization problems, of both causal and statistical character, under the same theoretical umbrella.

The generalization problems addressed in this thesis are as follows:

Problem 1. Transportability (generalizing experimental findings across settings, populations, or domains). How to reuse causal information acquired by experiments in one setting to answer causal queries in another, possibly different setting where only passive observations can be collected? This question embraces

several sub-problems treated informally in the literature under rubrics such as "external validity," "meta-analysis," "quasi-experiments," and "heterogeneity."

Problem 2. Selection Bias (generalizing statistical findings across sampling conditions (preferential exclusion of units from the sample)). How can knowledge from a sampled subpopulation be generalized to the entire population when the sampling process is not random, but determined by variables in the analysis?

Problem 3. Experimental identifiability (generalizing experimental findings across experimental conditions in the same population). How can accessible experiments be used as surrogates for other experiments that are too difficult, expensive, or unethical to be conducted in practice?

Building on the modern theory of causation, we provide algebraic, graphical, and algorithmic conditions to support the inductive step required in the corresponding task in each of these problems. This characterization delineates the formal boundary between estimable and non-estimable effects, and identifies which pieces of scientific knowledge need to be collected in each study to construct a bias-free estimate of the target query. The theory provided in this work is general, in the sense that it takes as input any arbitrary set of generalizability assumptions and decides whether this specific instance admits solution.

The problems discussed in this thesis have applications in several empirical sciences such as bioinformatics, medicine, economics, social sciences as well as in data-driven fields such as machine learning, artificial intelligence and statistics.

The dissertation of Elias Bareinboim is approved.

Eleazar Eskin

David Heckerman

Rosa Matzkin

Song-Chun Zhu

Judea Pearl, Committee Chair

University of California, Los Angeles

2014

This thesis is dedicated to the memory of my father, Julio Bareinboim.

Table of Contents

List of Figures

Acknowledgments

I am in debt to many people for being part of this long but rewarding journey.

First of all, I am grateful to my advisor, Judea Pearl, for his kindness in sharing his time, energy, and dreams with me. I feel very fortunate for having the chance to learn from him this very delicate type of craftsmanship called research. Since the beginning, Judea has been very generous with his time and ideas and treated me as a peer (as if I already had a PhD), always being critical but at the same time supportive and open minded. No idea or thought was ever dismissed a priori, but was rather considered carefully on its merits. It is difficult to forget the great discussions we have had at the dawn of day in his office about research, history, and the mysteries of science with such passion and intensity. It has also been very instructive for me, as a young researcher, to see Judea's attitude towards peer pressure, revered authorities, and "sacred cows;" he would never make a detour from his principled positions and explorational goals. I am very thankful to Judea for the freedom, encouragement, and inspiration he has given to me throughout these years.

I would like to thank the members of my doctoral committee: Eleazar Eskin, David Heckerman, Rosa Matzkin, and Song-Chun Zhu, who were always helpful, gracious and accommodating. Special thanks to David for his insightful questions and enthusiasm, and to Eleazar for his effort in making my results applicable in Bioinformatics.

I would also like to thank many professors who have shared their time, advice, and different types of support throughout these years: Valmir Barbosa, Junghoo Cho, Vitor Santos Costa, Frederick Eberhardt, Noah Goodman, Sander Greenland, Kristian Kersting, Manabu Kuroki, Adam Meyerson, James Robins, Bernhard Schölkopf, João Carlos Pereira Silva, Jennifer Wortman Vaughan, Ying Nian Wu, and Alan Yuille.

I feel fortunate and proud of being part of the Cognitive Systems Lab, and would like to thank all its previous members for creating this great research track that one can build on. I am especially thankful for the time, attention, and support given to me by Carlos Brito and Jin Tian; I feel humble for having the chance to interact with them. I would particularly like to thank Kaoru Mulvihill, who has always been very kind and saved me uncountably many times from real trouble with bureaucracy. I am also in debt to Bryant Chen, who has been such an accommodating officemate, our interactions have always been very informative and entertaining. Likewise, I thank Jacob Matthew and Andrew Forney for the nice conversations, and Madelyn Glymour for editing parts of my thesis.

There are also friends who were always interested in discussing causality,

science, or life more broadly; I am sure the PhD wouldn't have been as much fun without them: Ana Luisa Pessoa de Araujo, Amila Ariyaratne, Alberto Busetto, Gabriela Cybis, Michal Danielczyk, Paul Daniell, Doris Entner, Andre Freitas, Myung Soo Ko, James Lou, Salvatore Marcantonio, Humberto Naves, Pedro Ortega, Alan Roytman, Paul Wais, Lucas Wanner, and Kun Zhang.

Last but not least, I would like to thank my parents, Julio Bareinboim (of blessed memory) and Fortune Bareinboim; without their love and unconditional support, being here in the U.S. and pursuing my dream would not be possible. I am very thankful to my brother, Leonardo Bareinboim, who has always enthusiastically supported my endeavors and covered for me in Brazil during all these years. I am in debt to all my family, especially my cousins Pedro, Gloria, and Fremi, who have given me love and encouragement throughout my life.

It is really amazing if what we call free will is really just an illusion, as conjectured by many, since it is so obvious that we have options, that we are autonomous agents, that we can act intentionally, and that we are in control of our own fate. It is a great opportunity to live in these times and have the chance of being inspired and challenged by these questions, and perhaps one day, being able to resolve such mysteries. The groundbreaking theory developed by Judea can perhaps help us answer some of these fundamental questions about the essence of human nature.

CHAPTER 1

Introduction

Science is about generalization, and generalization requires an inductive leap from an observed reality to one that has not been seen before. Informal discussions concerning the difficulties of generalizing findings across populations have been going on for at least half a century (Cox58; CS63; Hec92; HIM05; Man07) and appear to accompany every textbook in experimental design. By and large, however, these discussions have not led to more than the obvious conclusions that researchers should be extremely cautious about unwarranted generalization, that many threats may await the unwary, and that extrapolation across differing studies requires "some understanding of the reasons for the differences" (Cox58, p. 10). Surprisingly, very few formal treatments have been attempted to reveal the conditions under which such generalizations are indeed possible.

In contrast, the assumptions needed for generalizations in the specific context of standard statistical analysis as well as in standard causal inference are extensively studied and well-understood. In the former, both the sample and the population are governed by the same probabilistic structure – the joint distribution of the observed variables; based on this and perhaps other assumptions, statistical theory provides guarantees on the inductive step of going from a sampled subpopulation to the entire population (e.g., law of large numbers, central limit theorem). In the latter, both the passive (observational) and experimental regimes are considered over the same population, share a data-generating mechanisms, and differ only in a local modification of the treatment assignment; based on certain assumptions, causal theory provides guarantees on the inductive step of going from a population in a passive pre-interventional regime to conclusions about the same population under an interventional regime.

The generalizability problems analyzed in this work are fundamentally different from those two accounts. In transportability, for example, we deal with two distinct populations that differ both in their inherent causal characteristics and the regimes under which they are studied. Interestingly enough, we will show that even when experiments cannot be conducted in the target population, and despite glaring differences between the two populations, it might still be possible to compute causal effects by borrowing experimental knowledge from the source environments.

The standard literature on this topic, falling under rubrics such as "external validity,"[1] "quasi-experiments," "meta-analysis," and "heterogeneity," consists primarily of "threats," namely, verbal narratives of what can go wrong when we try to transport results from one study to another (e.g., (SCC02, Chapter 3; HGH10)). Rarely do we find an analysis of "licensing assumptions," namely, formal and transparent conditions under which the transport of results across differing environments or populations is licensed from first principles. [2]

The reasons for this asymmetry are several. First, threats are safer to cite than assumptions. He who cites "threats" appears prudent, cautious and thoughtful, whereas he who seeks licensing assumptions is immediately accused of endorsing those assumptions, thus legitimizing unwarranted transport, or of pretending to know in advance when those assumptions hold true.

Second, assumptions are self-destructive in their honesty. The more explicit the assumption, the more criticism it invites, for it tends to trigger a richer space of alternative scenarios in which the assumption may fail. Researchers prefer therefore to declare threats in public and make assumptions in private.

More importantly, whereas threats can be communicated in plain English, supported by anecdotal pointers to familiar experiences, assumptions require a formal language within which the notion "environment" (or "population," "setting," "domain") is given precise characterization, and differences among environments can be encoded and analyzed.

The advent of causal diagrams (Pea95; GPR99; SGS00; Pea09b) provides such a language and renders the formalization of problems in causal generalizability possible. Building on the theory of non-parametric structural equations, we tackle three classes of problems related to the generalization of empirical findings. These classes were mentioned briefly in the abstract, and will be realized below, together

[1] (Man07) defines "external validity" as follows: "An experiment is said to have "external validity" if the distribution of outcomes realized by a treatment group is the same as the distribution of outcome that would be realized in an actual program." (CS63, p. 5) take a slightly broader view: "'External validity' asks the question of generalizability: To what population, settings, treatment variables, and measurement variables can this effect be generalized?" To the best of my knowledge no formal treatment of these problems was attempted by these authors.

[2] The machine learning literature, on the other hand, while seriously concerned about discrepancies between training and test environments (DM06; Sto09), has focused almost exclusively on predictive, or classification tasks as opposed to effect-learning tasks. Moreover, even in classification tasks, machine learning researchers have rarely allowed apriori causal knowledge to guide the learning process and, as a result, have not sought theoretical guarantees in the form of sufficient conditions under which discrepancies between the training and test environments can be circumvented, or necessary conditions without which bias will persist regardless of sample size. Some recent work on anticausal learning leverages knowledge about invariances of the data-generating model across domains (using representation equivalent to that discussed here), moving the literature towards more general modalities of learning (SJP12; ZSM13).

with examples and the chapters in which they are presented and analyzed.

Problem 1. Transportability (Chapters 3-4). Generalizing experimental findings across settings, populations, or domains. How can one reuse causal information acquired by experiments in one setting to answer causal queries in another, possibly different setting where only passive observations can be collected?

For instance, a researcher may perform experiments on mice and aim to generalize the conclusions to human beings. What mathematical principles support this leap of generalization? We show that one key ingredient necessary to formalize this type of questions is to identify areas of commonalities and disparities between the two species. After having a coarse description of these areas, we can formally decide what knowledge is or is not transportable across species.

Problem 2. Selection Bias (Chapter 5). Generalizing statistical findings across sampling conditions (preferential exclusion of units). How can knowledge from a sampled subpopulation be generalized to the entire population if the sampling selection is not random, but is affected by variables in the analysis?

For instance, one might have collected data in a specific hospital, and ask whether (under what conditions) this data could be generalized to the population as a whole, given that the hospital exercises a peculiar admission policy which depends on applicants' financial ability and symptoms.

Problem 3. Experimental Identifiability (Chapter 6). Generalizing experimental findings across experimental conditions in the same domain. How can some experimental knowledge be used as a surrogate for other experiments that are too difficult, expensive, or unethical to perform in practice?

For instance, one can conduct an experiment by randomizing diet, and ask whether (and under what conditions) the available data could help in establishing the causal-effects of cholesterol level on heart attack, when it is infeasible to randomize cholesterol level in practice. The problem relates to that of finding instrumental variables but, given the non-parametric setting, it is more involved.

These problems appear every time data is being collected and are pervasive in the empirical sciences (e.g., economics, bioinformatics, public health) as well as in machine learning and artificial intelligence (which are also data-oriented). Interestingly, they appear to be disparate types of generalizations and, to the extent that they have been addressed, they have evoked different tools and vocabulary in the literature; this work puts them under the same theoretical umbrella.

This thesis provides a characterization in the form of algebraic, graphical, and algorithmic conditions to support the inductive step in the corresponding task. This characterization establishes a theoretical boundary between estimable and non-estimable realities, and it goes further, for it identifies what pieces of scientific knowledge need to be collected in each study, and how to cement them

together, to achieve consistent estimates of the desired queries.

We list below the specific contributions achieved in this thesis by chapter and include the publications in which they were presented:

- Chapter 3. Transportability (PB11a; PB11b; BP12c):
  - introduces a formal language for expressing differences and commonalities between environments and reduces the problem of transportability to an exercise in symbolic calculus;
  - develops sufficient conditions for solving simple transportability problems such as when causal effects are the same in both environments;
  - constructs an intuitive algorithm for special transportability problems;
  - derives a general graphical condition for deciding transportability of causal effects (i.e., transportability is feasible if and only if a certain graph structure does not appear as an edge subgraph of the inputted diagram);
  - proves completeness of the do-calculus (Pea95) for recognizing transportability (that is, if a causal effect cannot be expressed in terms of the available data by repeated application of the three rules of the do-calculus, such an expression does not exist);
  - constructs an effective and complete algorithm for deciding experimental transportability of joint causal effects and returning a transport formula whenever those effects are transportable.
- Chapter 4. Transportability from multiple domains with limited experimental information (BP13a; BP13b; BLH13; BP14):
  - formulates the transportability problem over multiple source domains;
  - relaxes the requirement that all experiments be feasible in the same source domain;
  - derives a general graphical condition for deciding transportability of causal effects under these relaxed assumptions;
  - proves completeness of the do-calculus for recognizing general transportability (that is, shows that if a desired causal effect cannot be expressed in terms of the available (passive and experimental) distributions by repeated application of the three rules of the do-calculus, such an expression does not exist);
  - constructs a complete algorithm for deciding general transportability of causal effects and allowing for generic weighting schemes, which generalizes standard statistical procedures and leads to the construction of statistically more powerful estimators.

- Chapter 5. Selection Bias (BP12b; BTP14):
  - derives a general graphical condition for deciding whether conditional distributions are recoverable from selection bias without resorting to external information (from properly sampled data);
  - derives a sufficient graphical condition for deciding whether conditional distributions are recoverable from selection bias when external information is available (e.g., census data);
  - extends the back-door criterion to decide when sets of variables are sufficient for eliminating both selection and confounding biases;
  - derives a general graphical condition for deciding whether population and covariate-specific odds ratios can be recovered from selection bias;
  - constructs an efficient and complete algorithm for deciding recoverability from selection bias and returning an unbiased estimand for the odds ratio whenever such estimand exists;
  - shows that for measures such as the risk difference, when selection and confounding biases are simultaneously present, the former can be entirely removed with certain instrumental variables even when the latter cannot;
  - develops sufficient conditions for recovering from selection bias in common scenarios (complementing the back-door condition).
- Chapter 6. General (Experimental) Identification (BP12a):
  - derives a necessary and sufficient graphical condition for the general identification problem (that is, identification in terms of auxiliary experiments);
  - proves completeness of the do-calculus for recognizing general identifiability (that is, if a causal effect cannot be expressed in terms of the available data by repeated application of the three rules of the do-calculus, such an expression does not exist);
  - constructs an efficient and complete algorithm for deciding general identification of joint causal effects and returning the correct estimand whenever those effects are computable from the available data.

Moreover, we start reviewing the basic notation and the main results used throughout this thesis in Chapter 2, and conclude by summarizing the results and pointing to new problems in Chapter 7.

CHAPTER 2

Logical Foundations of Causal Inference

In this chapter, we review the basic semantical framework of our analysis that rests in nonparametric Structural Equations Models (SEM), which unifies and generalizes several approaches to Causal Inference (Rub74; Rob86; Daw02; SGS00).

2.1 Causal Models as Inference Engines

From a logical viewoint, causal analysis relies on causal assumptions that cannot be deduced from (nonexperimental) data. Thus, every approach to causal inference must provide a systematic way of encoding, testing and combining these assumptions with data. Accordingly, we view causal modeling as an inference engine that takes three inputs and produces three outputs.

The inputs are:

I-1. A set A of qualitative causal *assumptions* which the investigator is prepared to defend on scientific grounds, and a model M_A that encodes these assumptions mathematically. (In SEM, M_A takes the form of a diagram or a set of unspecified functions. A typical assumption is that no direct effect exists between a pair of variables, or that an omitted factor, represented by an error term, is uncorrelated with some other factors.)

I-2. A set Q of *queries* concerning causal or counterfactual relationships among variables of interest. In linear SEM, Q concerned the magnitudes of structural coefficients but, in general, Q may address causal relations directly, e.g.,

Q_1 : What is the effect of treatment X on outcome Y?

Q_2 : Is this employer guilty of gender discrimination?

In principle, each query $Q_i \in Q$ should be computable from any fully specified model M compatible with A.

I-3. A set D of experimental or non-experimental *data*, governed by a probability distribution presumably consistent with A.

The outputs are:

O-1. A set A^* of statements which are the logical implications of A, separate from the data at hand. For example, that X has no effect on Y if we hold Z constant, or that Z is an instrument relative to $\{X, Y\}$.

O-2. A set C of data-dependent *claims* concerning the magnitudes or likelihoods of the target queries in Q, each contingent on A. C may contain, for example, the estimated mean and variance of a given structural parameter, or the expected effect of a given intervention. Auxiliary to C, a causal model should also yield an estimand $Q_i(D)$ for each query in Q, or a determination that Q_i is not computable from D (in this case, we would typically say that Q is not identifiable form P).

O-3. A list T of testable statistical implications of A, and the degree $g(T_i), T_i \in T$, to which the data agrees with each of those implications. A typical implication would be a conditional independence assertion, or an equality constraint between two probabilistic expressions. Testable constraints should be read from the model M_A (see Definition 3.), and used to confirm or disconfirm the model against the data.

In this chapter, we deepen our discussion about the methodological issues; for a more comprehensive review on this topic, see (Pea09a; Pea12a).

2.2 Causal Assumptions in Nonparametric Models

Structural equation modeling (SEM) has been the main vehicle for effect analysis in economics and the behavioral and social sciences (Gol72; Dun75; Bol89). However, the bulk of SEM methodology was developed for linear analysis and, until recently, no comparable methodology has been devised to extend its capabilities to models involving dichotomous variables or nonlinear dependencies. A central requirement for any such extension is to detach the notion of "effect" from its algebraic representation as coefficient in an equation, and redefine "effect" as a general capacity to transmit *changes* among variables. Such an extension, based on simulating hypothetical interventions in the model, was proposed in (Haa43; SW60; SGS93; Pea93b; Pea00; Hec00; Hec05; Mat07) and has led to new ways of defining and estimating causal effects in nonlinear and nonparametric models (that is, models in which the functional form of the equation is unknown). These observations lead to the following definition of SEM:

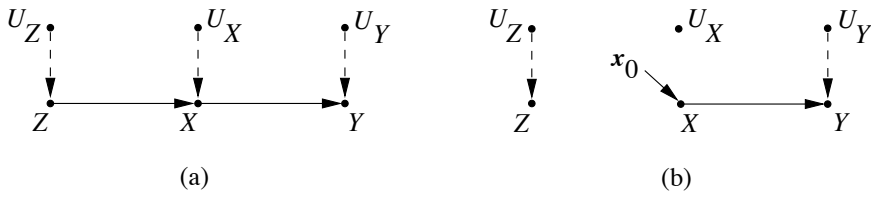


Figure 2.1: The diagrams associated with (a) the structural model of equation (2.1) and (b) the modified model of equation (2.4), representing the intervention $do(X = x_0)$.

Definition 1 (Structural Equation Model)**.** (Pea00, p. 203) *A structural causal model M is a tuple $M = \langle U, V, F, P(u) \rangle$, where:*

1. *A set U of background or exogenous variables, representing factors outside the model, which nevertheless affect relationships within the model.*
2. *A set $V = \{V_1, ..., V_n\}$ of endogenous variables, assumed to be observable. Each of these variables is functionally dependent on some subset PA_i of $U \cup V \setminus \{V_i\}$.*
3. *A set F of functions $\{f_1, ..., f_n\}$ such that each f_i determines the value of $V_i \in V$, $v_i = f_i(pa_i, u)$.*
4. *A joint probability distribution $P(u)$ over U.*

The central idea is to exploit the invariant characteristics of structural equations without committing to a specific functional form. For example, the nonparametric interpretation of the diagram of Fig. 2.1(a) corresponds to a set of three functions, each corresponding to one of the observed variables:

$$\begin{aligned} z &= f_Z(u_Z) \\ x &= f_X(z, u_X) \\ y &= f_Y(x, u_Y), \end{aligned} \tag{2.1}$$

where in this particular example, U_Z, U_X and U_Y are assumed to be jointly independent but otherwise arbitrarily distributed. Conversely, every set of structural equations like (2.1) corresponds to a unique diagram in which arrows are drawn from the arguments of f_i into the endogenous variable V_i.

Each of the functions represents a causal process (or mechanism) that determines the value of the left variable (output) from the values on the right variables (inputs), and is assumed to be invariant unless explicitly intervened on.

The absence of a variable from the right-hand side of an equation encodes the assumption that Nature ignores that variable in the process of determining the value of the output variable. For example, the absence of variable Z from the arguments of f_Y conveys the empirical claim that variations in Z will leave Y unchanged, as long as variables U_Y and X remain constant. A system of such functions are said to be *structural* if they are assumed to be autonomous, that is, each funciton is invariant to possible changes in the form of the other functions (Sim53; Koo53; Ald89).

Assuming that F has a unique solution for every variable, the distribution over the exogenous variables $P(u)$ together with F induce a distribution $P(v)$ over the endogenous variables V.[1] Interestingly, certain topological patterns in the diagram imply conditional independencies in the distribution over observables $P(v)$. To explain this relationship, we start with the simpler class of Markovian models in which the graphs are acyclic (i.e., containing no directed cycles) and all the exogenous variables are jointly independent. (Non-Markovian models, such as those involving correlated errors (resulting from unmeasured confounders), can be modelled similarly by introducing latent variables to account for the dependencies among the exogenous variables, to be discussed later on.) In this class, the relationship between distributions and graphs lies in the following theorem:

Theorem 1 (Causal Markov Condition (PV91))**.** *Any distribution generated by a Markovian model M can be factorized as:*

$$P(v_1, ..., v_n) = \prod_i P(v_i|pa_i) \tag{2.2}$$

We can now define the notion of compatibility between distributions and graphs:

Definition 2. *If a probability function P admits the factorization of (2.2) relative to a diagram G, we say that G and P are compatible, or that P is Markov relative to G.*

For example, the distribution associated with the model in Fig. 2.1(a) can be factorized as

$$P(z, y, x) = P(z)P(x|z)P(y|x) \tag{2.3}$$

since X is the (endogenous) parent of Y, Z is the parent of X, and Z has no parents. We also say that the distribution $P(z, y, x)$ in eq. (2.3) is compatible with the graph in Fig. 2.1(a).

[1] We will later define other distributions induced by the pair $F, P(u)$, for example, post-interventional distributions and counterfactuals.

This factorization implies that regardless of the idiosyncrasies of the properties of the distribution of exogenous $P(u)$ and the functions F (e.g., monotonicity, linearity, separability, continuity, differentiability), separation in G (defined next) unveils conditional independences constrains over $P(v)$.[2] The criterion for reading these constraints is known as *d-separation*, which will be instrumental in our analysis.

Definition 3 (d-separation (Pea88))**.**
A set S of nodes is said to block a path p if either

1. *p contains at least one arrow-emitting node that is in S, or*
2. *p contains at least one collision node that is outside S and has no descendant in S.*

If S blocks all *paths from set X to set Y, it is said to "d-separate X and Y," and then, it can be shown that variables X and Y are independent given S, written* $(X \perp\!\!\!\perp Y|S)$.[3]

Theorem 2 (Probabilistic Implications of d-separation (VP88; GVP90))**.** *If X and Y are d-separated by Z in a diagram G, then X is independent of Y conditional on Z in every distribution compatible with G. Conversely, if X and Y are not d-separated by Z in a diagram G, then X and Y are dependent conditional on Z in at least one distribution compatible with G.*

D-separation reflects conditional independencies that hold in any distribution $P(v)$ that is compatible with the causal assumptions A embedded in the diagram. To illustrate, the path $U_Z \to Z \to X \to Y$ in Figure 2.1(a) is blocked by $S = \{Z\}$ and by $S = \{X\}$, since each emits an arrow along that path. Consequently we can infer that the conditional independencies $(U_Z \perp\!\!\!\perp Y|Z)$ and $(U_Z \perp\!\!\!\perp Y|X)$ will be satisfied in any probability function that this model can generate, regardless of how we parametrize the arrows. Likewise, the path $U_Z \to Z \to X \leftarrow U_X$ is blocked by the null set $\emptyset$, but it is not blocked by $S = \{Y\}$ since Y is a descendant of the collision node X. Consequently, the marginal independence $(U_Z \perp\!\!\!\perp U_X)$ will hold in the distribution, but $(U_Z \perp\!\!\!\perp U_X|Y)$ may or may not hold. [4]

[2] Constraints of this nature will also appear in the interventional distributions.

[3] *See (HCS03), (Mul09), and (Pea09b), pp. 335 for a gentle introduction to d-separation.*

[4] This special handling of collision nodes (or *colliders*, e.g., $Z \to X \leftarrow U_X$) reflects a general phenomenon known as *Berkson's paradox* (Ber46), whereby observations on a common consequence of two independent causes render those causes dependent. For example, the outcomes of two independent coins are rendered dependent by the testimony that at least one of them is a tail. The colliders play a key role in the problem of selection bias as discussed in Chapter 4.

Notice that with this tool one is able to enumerate all testable implications entailed by a structural equation model, so test whether the data was generated by the hypothesized model. Furthermore, one might use these constraints to decide if two models are equivalent, or more ambitiously, one might use these constraints to identify all models that are compatible with a given dataset.

2.3 Representing Causal Effects and Counterfactuals

Another feature that SEM provides is the ability to predict the effect of interventions. [5] This is done through a mathematical operator called $do(x)$, which simulates physical interventions by deleting certain functions from the model, replacing them with a constant $X = x$, while keeping the rest of the model unchanged. For example, to emulate an intervention $do(x_0)$ that holds X constant (at $X = x_0$) in model M of Figure 2.1(a), we replace the equation for x in equation (2.1) with $x = x_0$, and obtain a new model, M_{x_0},

$$\begin{aligned} z &= f_Z(u_Z) \\ x &= x_0 \\ y &= f_Y(x, u_Y), \end{aligned} \tag{2.4}$$

the graphical description of which is shown in Figure 2.1(b).

The joint distribution associated with the modified model is denoted by $P(z, y|do(x_0))$, which describes the post-intervention distribution of variables Y and Z (also called "controlled" or "experimental" distribution), to be distinguished from the pre-interventional distribution, $P(x, y, z)$, associated with the original model of equation (2.1).[6] For example, if X represents a treatment variable, Y a response variable, and Z some covariate that affects the amount of treatment received, then the distribution $P(z, y|do(x_0))$ gives the proportion of individuals that would attain response level $Y = y$ and covariate level $Z = z$ under the hypothetical situation in which treatment $X = x_0$ is administered uniformly to the population.[7]

In general, we can formally define the post-interventional distribution by the

[5] A completely specified model can also be used to compute counterfactuals and their probabilities, but this lies beyond the scope of this work.

[6] This can be translated for the parametric setting, for instance linear, where the meaning of $\frac{\partial}{\partial x} P(y|do(X = x))$ is equivalent to the β coefficient in the respective structural equation.

[7] Equivalently, $P(z, y|do(x_0))$ can be interpreted as the joint probability of $(Z = z, Y = y)$ under a randomized experiment among units receiving treatment level $X = x_0$. Readers versed in potential-outcome notations may interpret $P(y|do(x), z)$ as the probability $P(Y_x = y|Z_x = z)$, where Y_x is the potential outcome under treatment $X = x$.

equation

$$P_M(y|do(x)) = P_{M_x}(y) \tag{2.5}$$

In words, in the framework of model M, the post-intervention distribution of outcome Y is defined as the probability that model M_x assigns to each outcome level $Y = y$. From this distribution, which is readily computed from any fully specified model M, we are able to assess treatment efficacy by comparing aspects of this distribution at different levels of x_0.[8]

From this distribution, one is able to assess the treatment efficacy by comparing aspects of this distribution at different levels of x_0. A common measure of treatment efficacy is the average difference

$$P(y|do(x_0')) - P(y|do(x_0)) \tag{2.6}$$

where x_0' and x_0 are two levels (or types) of treatment selected for comparison.

One might surmise that this definition requires that to predict the effect of interventions, one would need to literally simulate the intervention by mutilating the model as in (2.4), which requires a fully specified model M. This is not the case, and the definition encoded in equation (2.5) only represents the semantics of the interventional operator; in the sequel, we show how to compute the effect of interventions based on this semantics in only partially specified models.

2.4 Identification in partially specified models: The emergence of the Causal Calculus

A central question in causal analysis is the question of *identification of causal effects* in partially specified models: Given assumptions set A (as embodied in the model), can the controlled (post-intervention) distribution, $P(y|do(x))$, be estimated from data governed by the pre-interventional distribution $P(z, x, y)$?

The problem of identification of causal effects has received considerable attention in econometrics (Hur50; Mar50; Koo53) and social science (Dun75; Bol89), usually in linear parametric settings, were it reduces to asking whether some model parameter, β, has a unique solution in terms of parameters of P (say the population variance-covariance matrix). In the nonparametric, the notion of "has a unique solution" does not directly apply since quantities such as $Q(M) = P(y|do(x))$ have no parametric signature and are defined procedurally by simulating an intervention in a causal model M, as in equation (2.4). The following definition captures the requirement that Q be estimable from the data:

[8] Counterfactuals are defined similarly through the equation $Y_x(u) = Y_{M_x}(u)$ (see (Pea09b, Ch. 7)), but will not be needed for the discussions in this work.

Definition 4 (Identifiability)**.** *A causal query $Q(M)$ is identifiable, given a set of assumptions A, if for any two (fully specified) causal models M_1 and M_2 that satisfy A, we have*

$$P(M_1) = P(M_2) \Rightarrow Q(M_1) = Q(M_2) \tag{2.7}$$

In words, the functional details of M_1 and M_2 do not matter; what matters is that the assumptions in A (e.g., those encoded in the diagram) would constrain the variability of those details in such a way that equality of P's would entail equality of Q's. When this happens, Q depends on P only, and should therefore be expressible in terms of the parameters of P.

For Markovian systems, all queries can be expressed in terms of the distribution of the observed variable, which follows from Theorem 1:

Corollary 1 (Truncated factorization[9])**.** *For any Markovian model, the distribution generated by an intervention $do(X = x_0)$ on a set X of endogenous variables is given by the truncated factorization*

$$P(v_1, v_2, ..., v_k|do(x_0)) = \prod_{i|V_i \notin X} P(v_i|pa_i)\Big|_{X=x_0}, \tag{2.8}$$

where $P(v_i|pa_i)$ are the pre-intervention conditional probabilities.

When the system is non-Markovian, the problem of identifiability can be decided systematically using an algebraic procedure known as the do-calculus (Pea95), which is discussed next.

Do-calculus: The Algorithmization of Causal Effects[10]

Let us consider the problem of identifiability related to the century-old debate on relation between smoking (X) and lung cancer (Y). According to reports of the time, the tobacco industry has managed to prevent anti-smoking legislation to pass by arguing that the observed correlation between smoking and lung cancer could be explained by some sort of carcinogenic genotype (U) that involves innate dependency for nicotine. In other words, the claim was that smoking did not cause cancer, but the same people that smoke were doomed to contract cancer given that this was a deterministic genetic trait. Based on scientific grounds, one might

[9] A simple proof of the Causal Markov Theorem is given in (Pea00, pp. 30). This theorem was first presented in (PV91), but it is implicit in the works of (KSC84) and others. Corollary 1 was named "Manipulation Theorem" in (SGS93), and is also implicit in Robin's (1987) G-computation formula. See (Lau01).

[10] For a more thorough discussion on the do-calculus and identifiability, see (Pea00, Ch. 3)

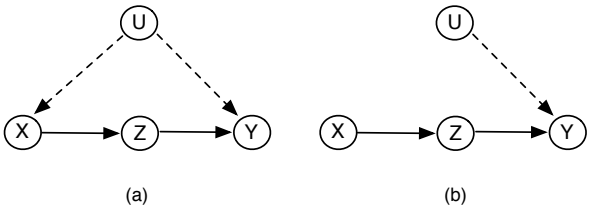


Figure 2.2: (a) Causal graph where X and Y are confounded by U, and there is a mediator Z such that all effects of X on Y pass through Z (so called frontdoor). The effect $P(y|do(x))$ is computable from observational data. (b) Interventional causal graph where X's incoming edges were cut.

further hypothesize that the effect of smoking on cancer is mediated through the deposit of tar in the lungs, as depicted in the model in Fig. 2.2(a).

Remarkably, the dependence between U_x and U_y is represented by a latent variable U, which turns the model into a Markovian model. This family of models is called Semi-Markovian and all the machinery developed for Markovian models is applicable here as well, but with one provision: the U-nodes cannot be manipulated nor condition on.

The policy question is whether the effect of smoking on cancer is computable from passive data, i.e., whether the query $Q = P(y|do(x))$ can be established without resorting to a randomized experiment that is unethical in this case.[11]

To solve the identification problem, we rely on the data-generating model G depicted in Fig. 2.2(a), and note that G factorizes accordingly to Theorem 1:

$$P(X, Y, Z, U) = P(X|U)P(Z|X)P(Y|Z, U)P(U), \tag{2.9}$$

We can express Q "wiping out" the factor $P(X|U)$ that accounts for the decision of smoke based on the truncated factorization (Corollary 1), which yields:

$$Q = \sum_{Z=z, U=u} P(Z = z|x)P(y|Z = z, U = u)P(U = u) \tag{2.10}$$

The challenge now is that not all factors necessary to compute Q are readily available from the data collected over observables. In particular, it is not obvious whether (and how) either $P(y|Z = z, U = u)$ or $P(U = u)$ is estimable from the data collected over the observables $P(X, Y, Z)$. Note that, at this point, there

[11]Note that this is a generalization question since the decision-maker needs to anticipate whether the effect Q is substantial before adopting the policy (which was never experienced before), so she/he might suggest the implementation of such policy in the real world (in this case, it could be to inhibit the commercializing of cigarettes).

are no constraints imposed over neither the dimensionality, nor the form of the unmeasured variable U, and our analysis will proceed in a non-parametric fashion. Still, it is unclear whether Q is computable from the assumptions encoded in G together with the pre-interventional distribution $P(V)$.[12] This graphical structure is known in the literature as the front-door graph.

The problem depicted above shows that a principled way to decide whether there exists a mapping between the observational (available) and the interventional distributions (target) is needed. The do-calculus consists of rules that permit the transformation of expressions involving the *do*-operator into equivalent expressions whenever certain conditions hold in G. The causal graph G licenses transformations between expressions based on the underlying assumptions.

Let X, Y, Z, and W be arbitrary disjoint sets of nodes in a causal diagram G. We denote by $G_{\overline{X}}$ the graph obtained by deleting from G all arrows pointing to nodes in X. Likewise, we denote by $G_{\underline{X}}$ the graph obtained by deleting from G all arrows emerging from nodes in X. To represent the deletion of both incoming and outgoing arrows, we use the notation $G_{\overline{X}\underline{Z}}$.

Theorem 3 (Rules of do-calculus (Pea95))**.** *Let G be a causal diagram generated by a structural equation model (Definition 1), and let $P(v)$ stand for the probability distribution induced by that model over the endogenous variables V. For any disjoint subsets of endogenous variables X, Y, Z, and W, the following rules are valid for every interventional distribution compatible with G.*

Rule 1 *(Insertion/deletion of observations):*

$$P(y|do(x), z, w) = P(y|do(x), w) \text{ if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}}} \tag{2.11}$$

Rule 2 *(Action/observation exchange):*

$$P(y|do(x), do(z), w) = P(y|do(x), z, w) \text{ if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\underline{Z}}} \tag{2.12}$$

Rule 3 *(Insertion/deletion of actions):*

$$P(y|do(x), do(z), w) = P(y|do(x), w) \text{ if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\overline{Z(W)}}}, \tag{2.13}$$

where $Z(W)$ is the set of Z-nodes that are not ancestors of any W-node in $G_{\overline{X}}$.

The first rule affirms that the d-separation criterion also holds for graphs under intervention. The second rule gives the condition for when *observing* and *intervening* are equivalent (when the confounding can be controlled, or the "criterion backdoor"). The third rule gives the conditions for when the do-operator

[12] Recall the provision for Semi-Markovian models that says that we cannot condition on U.

can be completely removed from the expression, i.e., there is no causal effect of Z on Y whatsoever. [13] [14]

We will be try to remove the do-operator from the r.h.s. of the expression since its absence represents the fact that the target relation (l.h.s.) is expressible in terms of the observable variables, hence computable from the available data (independently of the underlying functions and exogenous variables).

So, following this strategy to the smoking problem, it is clear that $(X \perp\!\!\!\perp Y)$, $(X \perp\!\!\!\perp Y)_{G_{\underline{X}}}$, and $(X \perp\!\!\!\perp Y)_{G_{\overline{X}}}$ do not hold (Figs. 2.2(a), 2.3(a), 2.2(b), respectively), so we cannot apply any of the rules of do-calculus. Still, we can simply condition on Z, which yields:

$$P(y|do(x)) = \sum_{Z=z} P(y|do(x), Z = z)P(Z = z|do(x)) \tag{2.14}$$

Considering the second term $P(Z = z|do(x))$, it is clear that only the second rule is applicable since $(Z \perp\!\!\!\perp X)_{G_{\underline{X}}}$ holds in Fig. 2.3(a) (intervening and observing are equivalent), which allows us to rewrite equation (2.14) as:

$$P(y|do(x)) = \sum_{Z=z} P(y|do(x), Z = z)P(Z = z|x) \tag{2.15}$$

Let us consider the first expression $P(y|do(x), Z = z)$. Note that to replace the do-operator with the see-operator (applying the second rule), we would need $(X \perp\!\!\!\perp Y|Z)_{G_{\underline{X}}}$, which does not hold in Fig. 2.3(a). Alternatively, we could try to fully remove the do-operator (applying the third rule), but $(X \perp\!\!\!\perp Y|Z)_{G_{\overline{X(Z)}}}$ does not hold since $G_{\overline{X(Z)}}$ does not allow us to cut the edges incoming to X (since X is ancestor of Z; the correct graph is Fig. 2.2(a)).

There is no clear way to proceed. Similarly to a derivation in infinitesimal calculus which is non-monotonic, even though our goal is to remove the do-operator, we may first add the do-operator to Z (since $(Z \perp\!\!\!\perp Y|X)_{G_{\overline{X}\underline{Z}}}$ holds in Fig. 2.3(d)), and see what happens. So, rewriting equation (2.15) yields:

$$P(y|do(x)) = \sum_{Z=z} P(y|do(x), do(Z = z))P(Z = z|x) \tag{2.16}$$

The problem became apparently harder than the original one since there are more do-operators in the expression than before (more experiments are required).

[13] Derivations illustrating the use of do-calculus can be find in (Pea09b, pp. 87).

[14] The *do*-calculus was proven to be complete to the identifiability of causal effects (SP06a; HV06b), which means that if a causal effect $P(y|do(x))$ cannot be expressed in terms of the probability of observables $P(v)$ by repeated application of these three rules, such an expression does not exist, and the effect is called non-identifiable.

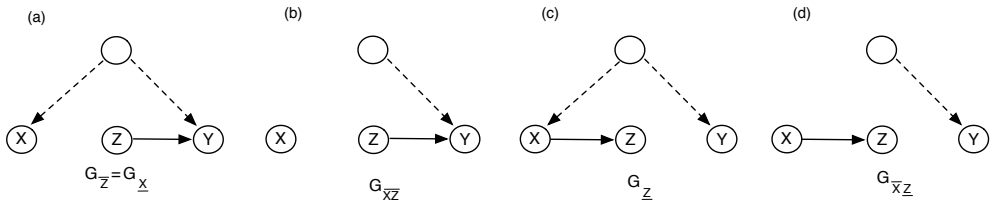


Figure 2.3: Mutilated graphs representing the conditions that license the multiple steps perfomed in do-calculus to derive $P(y|do(x))$ of Fig. 2.2(a).

However, the third rule of do-calculus can be applied in the first expression since $(X \perp\!\!\!\perp Y|Z)_{G_{\overline{X(Z)}}}$ holds in Fig. 3.3(b). So, we can express equation (2.16) as:

$$P(y|do(x)) = \sum_{Z=z} P(y|do(Z=z))P(Z=z|x) \tag{2.17}$$

To solve the problem, it suffices to remove the do-operator from the first expression. Again, there is no rule of the do-calculus that is directly applicable now, but note that X block all directed paths with edges going towards Z, so we condition on X:

$$P(y|do(x)) = \sum_{Z=z} P(Z=z|x)\Bigg(\sum_{X'=x'} P(y|do(Z=z),x')P(x'|do(Z=z))\Bigg), \tag{2.18}$$

and note that the second rule can be applied in $P(y|do(Z=z),x')$ since $(Z \perp\!\!\!\perp Y|X)_{G_{\underline{Z}}}$ holds in Fig. 3.3(c), which yields:

$$P(y|do(x)) = \sum_{Z=z} P(Z=z|x)\Bigg(\sum_{X'=x'} P(y|Z=z,x')P(x'|do(Z=z))\Bigg). \tag{2.19}$$

Note that the third rule can be applied in the last factor since Z is not an ancestor of X (or, $(Z \perp\!\!\!\perp X)_{G_{\overline{Z(W)}}}$ holds in Fig. 3.3(a)), which yields:

$$P(y|do(x)) = \sum_{Z=z} P(Z=z|x)\Bigg(\sum_{X'=x'} P(y|Z=z,x')P(x')\Bigg). \tag{2.20}$$

Finally, we note that the do-operator does not appear in the r.h.s. of equation (2.20), so even though we do not possess any quantitative knowledge about the unobservable variable U (neither its distribution nor its dimensionality), besides the fact that it influences both $\{X, Y\}$, we are still able to compute Q purely from the pre-interventional distribution $P(V)$ together with the assumption encoded in G. The final expression is somehow intuitive and appealing, and it can be seen as the causal effect of X on Z and Z on Y, but non-parametrically.

Remarkably, it follows immediately from the graphical interpretation of the second rule of the do-calculus what is known as the "adjustment formula" (also known as "conditional ignorability" (Rub74)). This represents the condition present when observing and intervening are essentially the same. This operation algebraically is equivalent to condition on a variable and take the average[15] . This provides the answer to the common question "what variables should we adjust for." The set of variables on which we need to adjust is none other but a set of nodes that satisfies a criterion called "Back-Door" and is defined as follows:

Definition 5 (Back-Door)**.** *A set of variables Z satisfies the back-door criterion relative to an ordered pair of variables (X, Y) in a causal diagram G if:*

1. *no node in Z is a descendent of X; and*
2. *Z blocks every path between X and Y that contains an arrow into X.*

And so, the following is immediate:

Theorem 4 (Back-Door Adjustment)**.** *If a set of variables Z satisfies the back-door criterion relative to (X, Y), then the causal effect of X on Y is identifiable and is given by the formula*

$$P(y|do(x)) = \sum_z P(y|x, z)P(z) \tag{2.21}$$

This criterion becomes especially useful when some variables in the graph are unobservable. The name "back-door" echoes conditions (ii) since it requires that only paths with arrow pointing to X be blocked; these paths can be viewed as entering in X through the back door.

For instance, our goal is to estimate the effect $P(y|do(Z = z))$ in Fig. 2.2. Note that X blocks all backdoor paths from Z to Y, so $P(y|do(Z = z)) = \sum_{x'} P(y|z, x')P(x')$, which is precisely the result entailed by the steps from equations (2.17) to (2.20).

In the real world, it is not uncommon that some causal effects are not identifiable from the observational data. For instance, consider the graph in Fig. 2.4(a) known as the 'bow-graph' (named due to its format) that is the smallest possible graph in which the effect of X on Y is not identifiable from observational data.

There are complementary ways to understand the problem of non-identifiability in this example. First, note that the correlation between X and Y can be expressed through $P(Y|X)$, and graphically this can be seen as the flow of influence passing through all open paths between X and Y, including the ones

[15]The average is by the prior of the variable, as can be seen next.

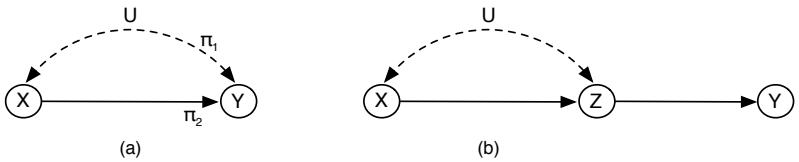


Figure 2.4: (a) Causal graph known as 'bow-graph' where X and Y are confounded by U and the effect $P(y|do(x))$ is not computable from observational data. (b) Extension of the bow-graph where Y is not confounded with X, but one of its ancestors is, so $P(y|do(x))$ is not computable from passive data.

passing through the unobservable U – i.e., $P(y|x) = \pi_1 + \pi_2$. On the other hand, the causal effect of X on Y can be seen as the influence being transmitted only through the directed edge from X to Y (assming a mutilated model) – i.e., $P(y|x) = \pi_1$ – which cannot be disentangle from the correlation coming from the passively observed data (i.e., $P(Y, X)$).

We can note that none of the rules of the do-calculus is applicable here; alternatively, we can write based on the truncated factorization (Corollary 1):

$$P(y|do(x)) = \sum_{U=u} P(y|x, u)P(u) \tag{2.22}$$

Note that it is not possible to manipulate equation (2.22) in such a way that U does not appear, given that the graph does not display any independence to be exploited (all variables are connected). This indeed precludes identifiability.

More formally, the non-computability of a certain quantity from the data means that two structural equation models (or, Nature) M_1, M_2 might be generating the same distribution over observables, $P_1(X, Y) = P_2(X, Y)$, but each one entails a different answer for the causal effect, i.e., $P_1(y|do(x)) \neq P_2(y|do(x))$.

Example. Consider two causal processes M_1 and M_2 as defined next. All variables are binary and the distribution over exogenous is $P(U) = 1/2$ in both models. In M_1, $F_1 = \{X = U,\ Y = 0\}$, and in M_2, $F_2 = \{X = U,\ Y = (U \text{ XOR } X)\}$, where XOR stands for the exclusive-or function. Note that in both models, $P(X = x, Y = 0) = 1/2$, for $X = \{0, 1\}$. However, $P_1(Y = 1|do(X = 1)) = 1/2$, while $P_2(Y = 1|do(X = 1)) = 0$, which show that $P(Y|do(X))$ is not computable from the assumptions encoded in G and the available data $P(v)$. [16]

In the general case, it is not easy to establish whether a given quantity is or is not computable from the combination of assumption in the form of the causal

[16] In this example, we just displayed one witness for the non-identifiability of the target quantity Q, but it is the case that the effect $P(y|do(x))$ will not be computable for *almost all* valid parametrizations of G.

diagram G and the observable data $P(v)$. For instance, even though X and Y are not confounded in the diagram G in Fig. 2.4(b), the causal effect $P(y|do(x))$ is not identifiable in G. However, note that $P(y|do(z))$ is identifiable in G and equals $P(y|z)$, by the second rule of do-calculus. This exemplifies that in the non-parametric settings, it might be the case that some quantities are computable and others are not given the same causal diagram. The only difference between the front-door graph (Fig. 2.2(a)) and the diagram in Fig. 2.4(b) is that the unobservable variable U confounds the $\{X, Y\}$ relationship in the former while it confounds the $\{X, Z\}$ in the latter.

Throughout this thesis, the do-calculus will show to be instrumental in other problems besides identification of causal effects. For instance, in Chapter 3, we shall see that, to establish transportability, the goal will be different; instead of eliminating do-operators, we will need to separate them from a special set of variables S that represent disparities between the mechanisms of the populations.

Furthermore, in Chapter 5, we shall see that, to establish z-identifiability, the goal will also be different; instead of eliminating do-operators altogether, we will need to replace a do-operator associated with the treatment X with a do-operator associated with the set of variables Z that represent external auxiliary experiments that are assumed to be available for use.

CHAPTER 3

Transportability Across Studies

3.1 Introduction

The generalizability of empirical findings to new environments, settings or populations, often called "external validity," is essential in most scientific explorations. This chapter treats a particular problem of generalizability, called "transportability", defined as a license to transfer causal effects learned in experimental studies to a new population, in which only observational studies can be conducted. In this chapter, we consider instances with only a pair of source and target domains and assuming that any experiment can be conducted in the source domain.

We introduce a formal representation called "selection diagrams" for expressing knowledge about differences and commonalities between populations of interest and, using this representation, we reduce questions of transportability to symbolic derivations in the do-calculus. This new representation also supports graphical and algorithmic methods for deciding whether causal effects in the target population can be inferred from experimental findings in the study population. When the answer is affirmative, the procedures identify what experimental and observational findings need be obtained from the two populations, and how they can be combined to ensure bias-free transport.

The chapter is organized as follows. In section 3.2, we motivate the question of transportability through simple examples, and illustrate how the solution depends on the causal story behind the problem. In section 3.3, we formally define the notion of transportability and reduce it to a problem of symbolic transformations in do-calculus. In section 3.4, we provide an intuitive sufficient graphical criterion for deciding transportability and estimating transported causal effects. In section 3.5, we provide a complete graphical criterion for deciding transportability and estimating transported causal effects. In section 3.6, we construct an algorithm based on the graphical criterion for deciding transportability without relying on algebraic manipulations. We conclude in section 3.7 briefly connecting transportability with other problems of generalizability that can benefit from the analysis developed throughout this chapter (e.g., surrogate endpoints).

3.2 Inference Across Populations: Motivating Examples

To motivate the formal treatment of Section 3.3, we first demonstrate some of the subtle questions that transportability entails through three simple examples, graphically depicted in Fig. 3.1.

Example 1. *We conduct a randomized trial in Los Angeles (LA) and estimate the causal effect of exposure X on outcome Y for every age group $Z = z$ as depicted in Fig. 3.1(a). We now wish to generalize the results to the population of New York City (NYC), but data alert us to the fact that the study distribution $P(x, y, z)$ in LA is significantly different from the one in NYC (call the latter $P^*(x, y, z)$). In particular, we notice that the average age in NYC is significantly higher than that in LA. How are we to estimate the causal effect of X on Y in NYC, denoted $P^*(y|do(x))$.*

Our natural inclination would be to assume that age-specific effects are invariant across cities and so, if the LA study provides us with (estimates of) age-specific causal effects $P(y|do(x), Z = z)$, the overall causal effect in NYC should be

$$P^*(y|do(x)) = \sum_z P(y|do(x), z)P^*(z) \tag{3.1}$$

This *transport formula* combines experimental results obtained in LA, $P(y|do(x), z)$, with observational aspects of NYC population, $P^*(z)$, to obtain an experimental claim $P^*(y|do(x))$ about NYC.[1]

Our first task will be to explicate the assumptions that renders this extrapolation valid. We ask, for example, what must we assume about other confounding variables beside age, both latent and observed, for Eq. (3.1) to be valid, or, would the same transport formula hold if Z was not age, but some proxy for age, say, language proficiency. More intricate yet, what if Z stood for an exposure-dependent variable, say hyper-tension level, that stands between X and Y?

Let us examine the proxy issue first.

Example 2. *Let the variable Z in Example 1 stand for subjects language proficiency, and let us assume that Z does not affect exposure (X) or outcome (Y), yet it correlates with both, being a proxy for age which is not measured in either study (see Fig. 3.1(b)). Given the observed disparity $P(z) \neq P^*(z)$, how are we*

[1] At first glance, Eq. (3.1) may be regarded as a routine application of "standardization" – a statistical extrapolation method that can be traced back to a century-old tradition in demography and political arithmetic (Wes16; Yul34; LN82; CS10). On a second thought it raises the deeper question of why we consider age-specific effects to be invariant across populations. See discussion following Example 2.

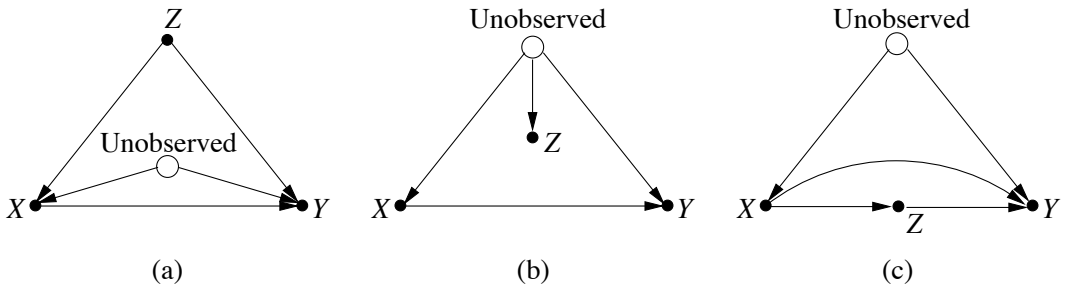


Figure 3.1: Causal diagrams depicting Examples 1–3. In (a) Z represents "age." In (b) Z represents "linguistic skills" while age (in hollow circle) is unmeasured. In (c) Z represents a biological marker situated between the treatment (X) and a disease (Y).

to estimate the causal effect $P^*(y|do(x))$ *for the target population of NYC from the z-specific causal effect* $P(y|do(x), z)$ *estimated at the study population of LA?*

The inequality $P(z) \neq P^*(z)$ in this example may reflect either age difference or differences in the way that Z correlates with age. If the two cities enjoy identical age distributions and NYC residents acquire linguistic skills at a younger age, then, since Z has no effect whatsoever on X and Y, the inequality $P(z) \neq P^*(z)$ can be ignored and, intuitively, the proper transport formula would be

$$P^*(y|do(x)) = P(y|do(x)) \tag{3.2}$$

If, on the other hand, the conditional probabilities $P(z|age)$ and $P^*(z|age)$ are the same in both cities, and the inequality $P(z) \neq P^*(z)$ reflects genuine age differences, Eq. (3.2) is no longer valid, since the age difference may be a critical factor in determining how people react to X. We see, therefore, that the choice of the proper transport formula depends on the causal context in which population differences are embedded.

This example also demonstrates why the invariance of Z-specific causal effects should not be taken for granted. While justified in Example 1, with Z = age, it fails in Example 2, in which Z was equated with "language skills." Indeed, using

Fig. 3.1(b) for guidance, the Z-specific effect of X on Y in NYC is given by:

$$\begin{aligned} P^*(y|do(x),z) &= \sum_{age} P^*(y|do(x),z,age)P^*(age|do(x),z) \\ &= \sum_{age} P^*(y|do(x),age)P^*(age|z) \\ &= \sum_{age} P(y|do(x),age)P^*(age|z) \end{aligned}$$

Thus, if the two populations differ in the relation between age and skill, i.e.,

$$P(age|z) \neq P^*(age|z)$$

the skill-specific causal effect would differ as well.

The intuition is clear. A NYC person at skill level $Z = z$ is likely to be in a totally different age group from his skill-equals in Los Angeles and, since it is age, not skill that shapes the way individuals respond to treatment, it is only reasonable that Los Angeles residents would respond differently to treatment than their NYC counterparts at the very same skill level.

The essential difference between Examples 1 and 2 is that age is normally taken to be an exogenous variable (not assigned by other factors in the model) while skills may be indicative of earlier factors (age, education, ethnicity) capable of modifying the causal effect. Therefore, conditional on skill, the effect may be different in the two populations.

Example 3. *Examine the case where Z is a X-dependent variable, say a disease bio-marker, standing on the causal pathways between X and Y as shown in Fig. 3.1(c). Assume further that the disparity $P(z) \neq P^*(z)$ is discovered in each level of X and that, again, both the average and the z-specific causal effect $P(y|do(x),z)$ are estimated in the LA experiment, for all levels of X and Z. Can we, based on information given, estimate the average (or z-specific) causal effect in the target population of NYC?*[2]

Here, Eq. (3.1) is wrong for two reasons. First, as in the case of age-proxy, it matters whether the disparity in $P(z)$ represents differences in susceptibility to X or differences in propensity to receiving X. In the latter case, Eq. (3.2) would be valid, while in the former, more information is needed. Second, the overall

[2] *This is precisely the problem that motivated the unsettled literature on "surrogate endpoint" (Pre89; FGS92; FR02; Bak06; JG09; Pea11), that is, using the effect of X on Z to predict the effect of X on Y in a population with potentially differing characteristics. An initial solution to this problem is offered in (PB11a).*

causal effect (in both LA and NYC) is no longer a simple average of the z-specific causal effects. To witness, consider an unconfounded Markov chain $X \rightarrow Z \rightarrow Y$; the z-specific causal effect $P(y|do(x), z)$ is $P(y|z)$, independent of x, while the overall causal effect is $P(y|do(x)) = P(y|x)$ which is clearly dependent on x. The latter could not be obtained by averaging over the former. The correct weighing rule is

$$P(y|do(x)) = \sum_z P(y, z|do(x)) \tag{3.3}$$

$$= \sum_z P(y|do(x), z)P(z|do(x)) \tag{3.4}$$

which reduces to (3.1) only in the special case where Z is unaffected by X, as is the case in Fig. 3.1(a). Thus, in general, both $P(y|do(x), z)$ and $P(z|do(x))$ need be measured in the experiment before we can transport results to populations with differing characteristics. In the Markov chain example, if the disparity in $P(z)$ stems only from a difference in people's susceptibility to X (say, due to preventive measures taken in one city and not the other) then the correct transport formula would be

$$P^*(y|do(x)) = \sum_z P(y|do(x), z)P^*(z|x) \tag{3.5}$$

$$= \sum_z P(y|z)P^*(z|x) \tag{3.6}$$

which is different from both (3.1) and (3.2), and hardly makes any use of experimental findings.

In case X and Y are confounded and directly connected, as in Fig. 3.1(c), it is Eq. (C.11) which provides the correct transport formula (to be proven in Section 3.4), calling for the z-specific effects to be weighted by the conditional probabilities $P^*(z|x)$, estimated at the target population.

3.3 Formalizing Transportability

3.3.1 Selection diagrams and selection variables

A few patterns emerge from the examples discussed in Section 3.2. First, transportability is a causal, not statistical notion. In other words, the conditions that license transport as well as the formulas through which results are transported depend on the causal relations between the variables in the domain, not merely on their statistics. When we asked, for instance (in Example 3), whether the change in $P(z)$ was due to differences in $P(x)$ or due to a change in the way

Z is affected by X, the answer cannot be determined by comparing $P(x)$ and $P(z|x)$ to $P^*(x)$ and $P^*(z|x)$. If X and Z are confounded (e.g., Fig. 3.4(e)), it is quite possible for the inequality $P(z|x) \neq P^*(z|x)$ to hold, reflecting differences in confounding, while the way that Z is affected by X, (i.e., $P(z|do(x))$) is the same in the two populations.

Second, licensing transportability requires knowledge of the mechanisms, or processes, through which population differences come about; different localization of these mechanisms yield different transport formulae. This can be seen most vividly in Example 2 (Fig. 3.1(b)) where we reasoned that no weighing is necessary if the disparity $P(z) \neq P^*(z)$ originates with the way language proficiency depends on age, while the age distribution itself remains the same. Yet, because age is not measured, this condition cannot be detected in the probability distribution P, and cannot be distinguished from an alternative condition,

$$P(age) \neq P^*(age) \quad \text{and} \quad P(z|age) = P^*(z|age)$$

one that may require weighting according to to Eq. (3.1). In other words, every probability distribution $P(x, y, z)$ that is compatible with the process of Fig. 3.1(b) is also compatible with that of Fig. 3.1(a) and, yet, the two processes dictate different transport formulas.

Based on these observations, it is clear that if we are to represent formally the differences between populations (similarly, between experimental settings or environments), we must resort to a representation in which the causal mechanisms are explicitly encoded and in which differences in populations are represented as local modifications of those mechanisms.

To this end, we will use causal diagrams augmented with a set, S, of "selection variables," where each member of S corresponds to a mechanism by which the two populations differ, and switching between the two populations will be represented by conditioning on different values of these S variables.

Intuitively, if $P(v|do(x))$ stands for the distribution of a set V of variables in the experimental study (with X randomized) then we designate by $P^*(v|do(x))$ the distribution of V if we were to conduct the study on population Π^* instead of Π. We now attribute the difference between the two to the action of a set S of selection variables, and write[3] [4]

$$P^*(v|do(x)) = P(v|do(x), s^*).$$

[3]Alternatively, one can represent the two populations' distributions by $P(v|do(x), s)$, and $P(v|do(x), s^*)$, respectively. The results, however, will be the same, since only the location of S enters the analysis.

[4]Pearl ((Pea95; Pea09b, p. 71)) and (Daw02), for example, use conditioning on auxiliary variables to switch between experimental and observational studies. (Daw02) further uses such variables to represent changes in parameters of probability distributions.

Remark. Similarly to the missing-links in Bayesian networks that encode the probabilistic invariance in the form of conditional independences, the S-variables that are missing encode the assumptions of invariance in structural models. For instance, the absence of a S node pointing to Y in Fig. 3.2(a) is what entails the age-specific effects to be invariant across the two populations (see more below).

The selection variables in S may represent all factors by which populations may differ or that may "threaten" the transport of conclusions between populations. For example, the age disparity $P(z) \neq P^*(z)$ discussed in Example 1 will be represented by the inequality

$$P(z) \neq P(z|s)$$

where S stands for all factors responsible for drawing subjects at age $Z = z$ to NYC rather than LA.

This graphical representation, which we will call "selection diagrams" is defined as follows:[5]

Definition 6 (Selection Diagram)**.** *Let $\langle M, M^* \rangle$ be a pair of structural causal models (Definition 1) relative to domains $\langle \Pi, \Pi^* \rangle$, sharing a causal diagram G. $\langle M, M^* \rangle$ is said to induce a selection diagram D if D is constructed as follows:*

1. *Every edge in G is also an edge in D;*
2. *D contains an extra edge $S_i \rightarrow V_i$ whenever there exists a discrepancy $f_i \neq f_i^*$ or $P(U_i) \neq P^*(U_i)$ between M and M^*.*

In summary, the S-variables locate the *mechanisms* where structural discrepancies between the two populations are suspected to take place. Alternatively, the absence of a selection node pointing to a variable represents the assumption that the mechanism responsible for assigning value to that variable is the same in the two populations. In the extreme case, we could add selection nodes to all variables, which means that we have no reason to believe that the populations share any mechanism in common, and this, of course would inhibit any exchange of information among the populations. The invariance assumptions between populations, as we will see, will open the door for the transport of some experimental findings.

For clarity, we will represent the S variables by squares, as in Fig. 3.2, which uses selection diagrams to encode the three examples discussed in Section 3.2.

[5]The assumption that there are no structural changes between domains can be relaxed starting with $D = G^*$ and adding S-nodes following the same procedure as in Def. 6, while enforcing acyclicity.

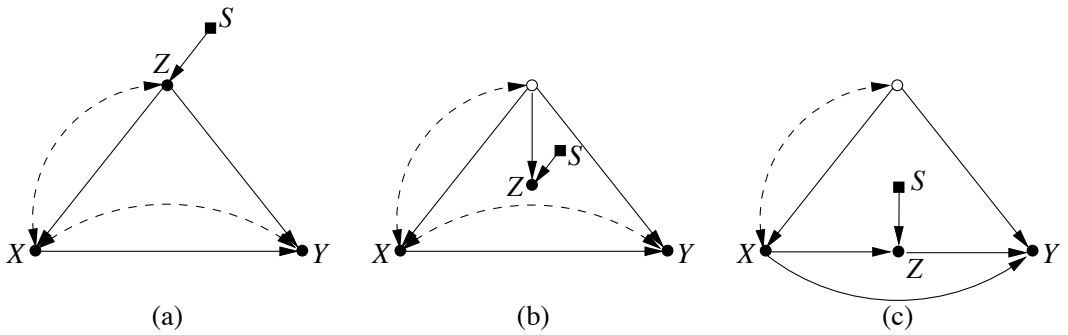


Figure 3.2: Selection diagrams depicting Examples 1–3. In (a) the two populations differ in age distributions. In (b) the populations differs in how Z depends on age (an unmeasured variable, represented by the hollow circle) and the age distributions are the same. In (c) the populations differ in how Z depends on X.

In particular, Fig. 3.2(a) and 3.2(b) represent, respectively, two different mechanisms responsible for the observed disparity $P(z) \neq P^*(z)$. The first (Fig. 3.2(a)) dictates transport formula (1) while the second (Fig. 3.2(b)) calls for direct, unadjusted transport (2). Clearly, if the age distribution in the target is different relative to that of the study population (Fig. 3.2(a)), we will represent this difference in the form of an unspecified influence that operates on the age variable Z and results in the difference between $P^*(age) = P(age|S = s^*)$ and $P(age)$.

In this chapter, we will address the issue of transportability assuming that scientific knowledge about invariance of certain mechanisms is available and encoded in the selection diagram through the S nodes. Such knowledge is, admittedly, more demanding than that which shapes the structure of each causal diagram in isolation. It is, however, a prerequisite for any scientific extrapolation, and constitutes therefore a worthy object of formal analysis.

3.3.2 Transportability: Definitions and Examples

Using selection diagrams as the basic representational language, and harnessing the concepts of intervention, *do*-calculus, and identifiability (Section 2.4), we can now give the notion of transportability a formal definition.

Definition 7 (Transportability)**.** *Let D be a selection diagram relative to domains $\langle \Pi, \Pi^* \rangle$. Let $\langle P, I \rangle$ be the pair of observational and interventional distributions of Π, and P^* be the observational distribution of Π^*. The causal relation $R(\Pi^*) = P^*(y|do(x), z)$ is said to be transportable from Π to Π^* in D if $R(\Pi^*)$ is uniquely computable from P, P^*, I in any model that induces D.*

Two interesting connections between identifiability and transportability are worth noting. First, note that all identifiable causal relations in D are also transportable, because they can be computed directly from P^* and require no experimental information from Π. Second, note that given causal diagram G, one can produce a selection diagram D such that identifiability in G is equivalent to transportability in D. First set $D = G$, and then add selection nodes pointing to all variables in D, which represents that the target domain does not share any mechanism with its counterpart – this is equivalent to the problem of identifiability because the only way to achieve transportability is to identify R from scratch in the target population.

While the problems of identifiability and transportability are related, proofs of non-transportability are more involved than those of non-identifiability for they require one to demonstrate the non-existence of two competing models compatible with D, agreeing on $\{P, P^*, I\}$, and disagreeing on $R(\Pi^*)$.

Definition 7 is declarative, and does not offer an effective method of demonstrating transportability even in simple models. Theorem 5 offers such a method using a sequence of derivations in do-calculus.

Theorem 5. *Let D be the selection diagram characterizing two populations, Π and Π^*, and S a set of selection variables in D. The relation $R = P^*(y|do(x), z)$ is transportable from Π to Π^* if the expression $P(y|do(x), z, s)$ is reducible, using the rules of do-calculus, to an expression in which S appears only as a conditioning variable in do-free terms.*

Proof. Every relation satisfying the condition of Theorem 5 can be written as an algebraic combination of two kinds of terms, those that involve S and those that do not. The formers can be written as P^*-terms and are estimable, therefore, from observations on Π^*, as required by Definition 7. All other terms, especially those involving *do*-operators, do not contain S; they are experimentally identifiable therefore in Π. □

This criterion was proven to be both sufficient and necessary for causal effects, namely $R = P(y|do(x))$ (see section 3.6).

Theorem 5, though procedural, does not specify the sequence of rules leading to the needed reduction when such a sequence exists. In the sequel (Theorem 7), we establish a more effective procedure of confirming transportability, which is guided by two recognizable subgoals.

Definition 8. *(Trivial Transportability)*
A causal relation R is said to be trivially transportable *from Π to Π^*, if $R(\Pi^*)$ is identifiable from (G^*, P^*).*

This criterion amounts to an ordinary test of identifiability of causal relations using graphs, as given by Definition 4. It permits us to estimate $R(\Pi^*)$ directly from observational studies on Π^*, un-aided by causal information from Π.

Example 4. *Let R be the causal effect $P(y|do(x))$ and let the selection diagram of Π and Π^* be given by $X \rightarrow Y \leftarrow S$, then R is trivially transportable, since $R(\Pi^*) = P^*(y|x)$.*

Another special case of transportability occurs when a causal relation has identical form in both domains – no recalibration is needed.

Definition 9. *(Direct Transportability)*
A causal relation R is said to be directly transportable *from Π to Π^*, if $R(\Pi^*) = R(\Pi)$.*

A graphical test for direct transportability of $R = P^*(y|do(x), z)$ follows from do-calculus and reads: $(S \perp\!\!\!\perp Y|X, Z)_{G_{\overline{X}}}$; in words, X blocks all paths from S to Y once we remove all arrows pointing to X and condition on Z. As a concrete example, this test is satisfied in Fig. 3.1(a), and therefore, the z-specific effects is the same in both populatons; it is directly transportable.

Remark. The notion of "external validity" as defined by (Man07) (footnote 1) corresponds to Direct Transportability, for it requires that R retains its validity without adjustment, as in Eq. (3.2). Such conditions restrict us from using information from Π^* to recalibrate R.

Example 5. *Let R be the causal effect of X on Y, and let D have a single S node pointing to X, then R is directly transportable, because causal effects are independent of the selection mechanism (see Pea09b, pp. 72–73).*

Example 6. *Let R be the z-specific causal effect of X on Y $P^*(y|do(x), z)$ where Z is a set of variables, and P and P^* differ only in the conditional probabilities $P(z|pa(Z))$ and $P^*(z|pa(Z))$ such that $(Z \perp\!\!\!\perp Y|pa(Z))$, as shown in Fig. 3.2(b). Under these conditions, R is not directly transportable. However, the $pa(Z)$-specific causal effects $P^*(y|do(x), pa(Z))$ are directly transportable, and so is $P^*(y|do(x))$. Note that, due to the confounding arcs, none of these quantities is identifiable.*

3.4 Transportability of Causal Effects - A Graphical Criterion

We now state and prove two theorems that permit us to decide algorithmically, given a selection diagram, whether a relation is transportable between two populations, and what the transport formula should be.

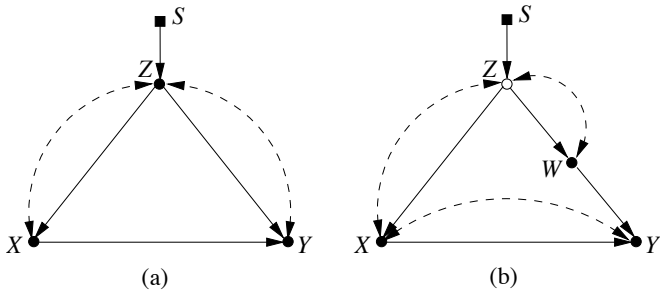


Figure 3.3: Selection diagrams illustrating S-admissibility. (a) has no S-admissible set while in (b), W is S-admissible.

Theorem 6. *Let D be the selection diagram characterizing two populations, Π and Π^*, and S the set of selection variables in D. The strata-specific causal effect $P^*(y|do(x), z)$ is transportable from Π to Π^* if Z d-separates Y from S in the X-manipulated version of D, that is, Z satisfies $(Y \perp\!\!\!\perp S|Z, X)_{D_{\overline{X}}}$.*

Proof.

$$P^*(y|do(x), z) = P(y|do(x), z, s^*)$$

From Rule-1 of *do*-calculus we have: $P(y|do(x), z, s^*) = P(y|do(x), z)$ whenever Z satisfies $(Y \perp\!\!\!\perp S|Z)$ in $D_{\overline{X}}$. This proves Theorem 6. □

Definition 10. *(S-admissibility)*
A set T of variables satisfying $(Y \perp\!\!\!\perp S|T, X)$ in $D_{\overline{X}}$ will be called S-admissible (with respect to the causal effect of X on Y).

Corollary 2. *The average causal effect $P^*(y|do(x))$ is transportable from Π to Π^* if there exists a set Z of observed pre-treatment covariates that is S-admissible. Moreover, the transport formula is given by the weighting of Eq. (3.1).*

Example 7. *The causal effect is transportable in Fig. 3.2(a), since Z is S-admissible, and in Fig. 3.2(b), where the empty set is S-admissible. It is also transportable by the same criterion in Fig. 3.3(b), where W is S-admissible, but not in Fig. 3.3(a) where no S-admissible set exists.*

Corollary 3. *Any S variable that is pointing directly into X as in Fig. 3.4(a), or that is d-connected to Y only through X can be ignored.*

This follows from the fact that the empty set is S-admissible relative to any such S variable. Conceptually, the corollary reflects the understanding that differences in propensity to receive treatment do not hinder the transportability of

treatment effects; the randomization used in the experimental study washes away such differences.

We now generalize Theorem 6 to cases involving treatment-dependent Z variables, as in Fig. 3.2(c).

Theorem 7. *The average causal effect $P^*(y|do(x))$ is transportable from Π to Π^* if either one of the following conditions holds*

1. *$P^*(y|do(x))$ is trivially transportable*
2. *There exists a set of covariates, Z (possibly affected by X) such that Z is S-admissible and for which $P^*(z|do(x))$ is transportable*
3. *There exists a set of covariates, W that satisfy $(X \perp\!\!\!\perp Y|W,S)_{D_{\overline{X(W)}}}$ and for which $P^*(w|do(x))$ is transportable.*

Proof. 1. Condition (1) entails transportability.

2. If condition (2) holds, it implies

$$P^*(y|do(x)) = P(y|do(x),s) \tag{3.7}$$
$$= \sum_z P(y|do(x),z,s)P(z|do(x),s) \tag{3.8}$$
$$= \sum_z P(y|do(x),z)P^*(z|do(x)) \tag{3.9}$$

We now note that the transportability of $P(z|do(x))$ should reduce $P^*(z|do(x))$ to a star-free expression and would render $P(y|do(x))$ transportable.

3. If condition (3) holds, it implies

$$P^*(y|do(x)) = P(y|do(x),s) \tag{3.10}$$
$$= \sum_w P(y|do(x),w,s)P(w|do(x),s) \tag{3.11}$$
$$= \sum_w P(y|w,s)P^*(w|do(x)) \tag{3.12}$$
(by Rule-3 of *do*-calculus)
$$= \sum_w P^*(y|w)P^*(w|do(x)) \tag{3.13}$$

We similarly note that the transportability of $P^*(w|do(x))$ should reduce $P(w|do(x),s)$ to a star-free expression and would render $P^*(y|do(x))$ transportable. This proves Theorem 7.

□

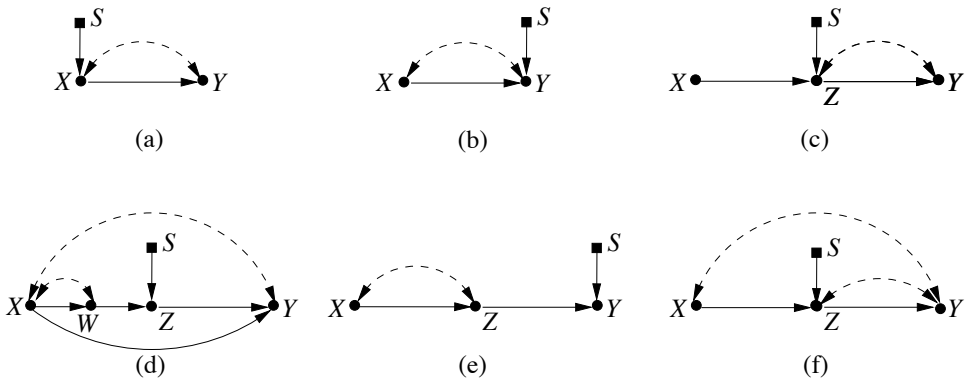


Figure 3.4: Selection diagrams illustrating transportability. The causal effect $P(y|do(x))$ is (trivially) transportable in (c) but not in (b) and (f). It is transportable in (a), (d), and (e) (see Corollary 3).

Remark. The test entailed by Theorem 7 is recursive, since the transportability of one causal effect depends on that of another. However, given that the diagram is finite and feedback-free, the sets Z and W needed in conditions 2 and 3 of Theorem 7 would become closer and closer to X, and the iterative process will terminate after a finite number of steps. This occurs because the causal effects $P^*(z|do(x))$ (likewise, $P^*(w|do(x))$) is trivially transportable and equals $P(z)$ for any Z node that is not a descendant of X. Thus, the need for reiteration applies only to those members of Z that lie on the causal pathways from X to Y.

Example 8. *Fig. 3.4(d) requires that we invoke both conditions of Theorem 7, iteratively. To satisfy condition 2 we note that Z is S-admissible, and we need to prove the transportability of $P^*(z|do(x))$. To do that, we invoke condition 3 and note that W d-separates X from Z in D. There remains to confirm the transportability of $P^*(w|do(x))$, but this is guaranteed by the fact that the empty set is S-admissible relative to W, since $(W \perp\!\!\!\perp S)$. Hence, by Theorem 6 (replacing Y with W) $P^*(w|do(x))$ is transportable, which bestows transportability on $P^*(y|do(x))$. Thus, the final transport formula (derived formally in Appendix A) is:*

$$P^*(y|do(x)) = \sum_z P(y|do(x), z) \sum_w P(w|do(x)) P^*(z|w) \tag{3.14}$$

The first two factors on the right are estimable in the experimental study, and the third through observational studies on the target population. Note that the joint effect $P^*(y, w, z|do(x))$ need not be estimated in the experiment; a decomposition that results in improved estimation power.

A similar analysis proves the transportability of the causal effect in Fig. 3.4(e)

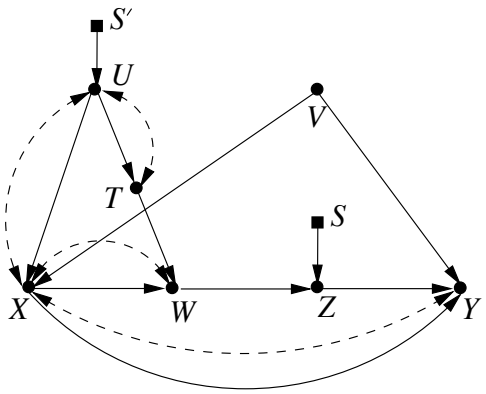


Figure 3.5: Selection diagram in which the causal effect is shown to be transportable in multiple iterations of Theorem 7 (see Appendix A).

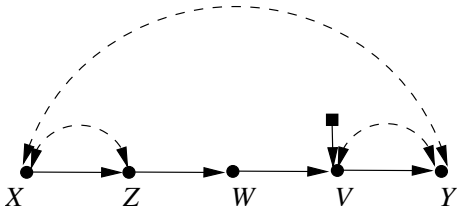


Figure 3.6: Selection diagram in which the effects $P^*(y|do(x))$ is transportable, but Theorem 7 is incapable to determine it. (See Corollary 8 in Appendix A.)

(see (PB11a)). The model of Fig. 3.4(f) however does not allow for the transportability of $P(y|do(x))$ because there is no S-admissible set in the diagram and, furthermore, condition 3 of Theorem 7 cannot be invoked.

Example 9. *To illustrate the power of Theorem 7 in discerning transportability and deriving transport formulae, Fig. 3.5 represents a more intricate selection diagram, which requires several iteration to discern transportability. The transport formula for this diagram is given by (derived formally in Appendix A):*

$$P^*(y|do(x)) = \sum_z P(y|do(x),z) \sum_w P^*(z|w) \sum_t P(w|do(x),t) P^*(t) \quad (3.15)$$

The main power of this formula is to guide investigators in deciding what measurements need be taken in both the experimental study and the target population. It asserts, for example, that variables U and V need not be measured. It likewise asserts that the W-specific causal effects need not be estimated in

the experimental study and only the conditional probabilities $P^*(z|w)$ and $P^*(t)$ need be estimated in the target population. The derivation of this formulae is given in the Appendix A.

Despite its power, Theorem 7 in not complete, namely, it is not guaranteed to approve all transportable relations or to disapprove all non-transportable ones. An example of the former is shown in Fig. in 3.6, which motivates the need of an alternative, perhaps complete (necessary and sufficient) conditions for transportability. Such conditions have been established in the next sections, where they are given in a graphical and algorithmic form. Theorem 7 provides, nevertheless, a simple and powerful method of establishing trasportability in practice.

3.5 Characterizing Transportable Relations

To characterize the class of transportable relations, we need to better understand when a relationship is non-transportable, which is the main subject of this section. The following lemma provides an auxiliary tool to prove non-transportability and is based on refuting the uniqueness property required by Definition 7.

Lemma 1. *Let X, Y be two sets of disjoint variables, in population Π and Π^*, and let D be the selection diagram. $P^*_x(y)$ is not transportable from Π to Π^* if there exist two causal models M^1 and M^2 compatible with D such that $P_1(V) = P_2(V)$, $P^*_1(V) = P^*_2(V)$, $P_1(V \setminus W|do(W)) = P_2(V \setminus W|do(W))$, for any set W, all families have positive distribution, and $P^*_1(y|do(x)) \neq P^*_2(y|do(x))$.*

Proof. Let I be the set of interventional distributions $P(V \setminus W|do(W))$, for any set W. The latter inequality rules out the existence of a function from P, P^*, I to $P^*_x(y)$. □

Lemma 1 explicitly indicates that proofs of non-transportability are more involved than those of non-identifiability, showing that to prove non-transportability one needs to construct two models agreeing on $\langle P, I, P^* \rangle$, while to prove non-identifiability is only required the models to agree on the distribution P.

The simplest non-transportable structure is an extension of the famous 'bow arc' graph named here 's-bow arc', see Fig. 3.7(a). The s-bow arc has two endogenous nodes: X, and its child Y, sharing a hidden exogenous parent U, and a S-node pointing to Y. This and similar structures that prevent transportability will be useful in our proof of completeness, which requires a demonstration that whenever a method fails to transport a causal relation, this relation is indeed non-transportable.

Theorem 8. *$P^*_x(y)$ is not transportable in the s-bow arc graph.*

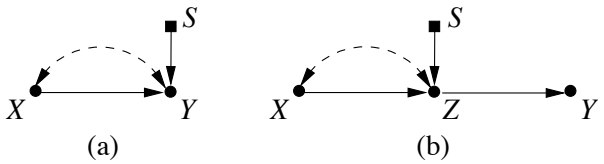


Figure 3.7: (a) Smallest selection diagram in which $P(y|do(x))$ is not transportable (s-bow graph). (b) A selection diagram in which even though there is no S-node pointing to Y, the effect of X on Y is still not-transportable due to the presence of a sC-tree (see Corollary 5).

Proof. The proof will show a counter-example to the transportability of $P^*_x(Y)$ through two models M_1 and M_2 that agree in $\langle P, P^*, I\rangle$ and disagree in $P^*_x(y)$.

Assume that all variables are binary. Let the model M_1 be defined by the following system of structural equations: $X_1 = U, Y_1 = ((X \oplus U) \oplus S), P_1(U) = 1/2$, and M_2 by the following one: $X_2 = U, Y_2 = S \vee (X \oplus U), P_2(U) = 1/2$, where $\oplus$ represents the *exclusive or* function.

Lemma 2. *The two models agree in the distributions* $\langle P, P^*, I\rangle$.

Proof. We show that the following equations must hold for M_1 and M_2:

$$\begin{cases} P_1(X|S) = P_2(X|S), \quad S = \{0,1\} \\ P_1(Y|X,S) = P_2(Y|X,S), \quad S = \{0,1\} \\ P_1(Y|do(X), S=0) = P_2(Y|do(X), S=0) \end{cases}$$

for all values of X, Y. The equality between $P_i(X|S)$ is obvious since $(S \perp\!\!\!\perp X)$ and X has the same structural form in both models. Second, let us construct the truth table for Y:

X	S	U	Y_1	Y_2
0	0	0	0	0
0	0	1	1	1
0	1	0	1	1
0	1	1	0	1
1	0	0	1	1
1	0	1	0	0
1	1	0	0	1
1	1	1	1	1

To show that the equality between $P_i(Y = 1|X, S = 0), X = \{0,1\}$ holds, we

rewrite it as follows:

$$P_i(Y=1|X,S=0) = \frac{P_i(Y=1|X,S=0,U=1)P_i(X|U=1)P_i(U=1)}{P_i(X)} + \frac{P_i(Y=1|X,S=0,U=0)P_i(X|U=0)P_i(U=0)}{P_i(X)} \quad (3.16)$$

In eq. (3.16), the expressions for $X=\{0,1\}$ are functions of the tuples $\{(X=1,S=0,U=1),(X=0,S=0,U=0)\}$, which evaluate to the same value in both models. Similarly, the expressions $P_i(Y=1|X,S=1)$ for $X=\{0,1\}$ are functions of the tuples $\{(X=1,S=1,U=1),(X=0,S=1,U=0)\}$, which also evaluate to the same value in both models.

We further assert the equality between the interventional distributions in Π, which can be written using the do-calculus as

$$\begin{aligned} P_i(Y=1|do(X),S=0) &= \sum_U P_i(Y|do(X),S=0,U)P_i(U|do(X),S=0) \\ &= P_i(Y=1|X,S=0,U=1)P_i(U=1) \\ &+ P_i(Y=1|X,S=0,U=0)P_i(U=0), \quad X=\{0,1\} \end{aligned} \quad (3.17)$$

Evaluating this expression points to the tuples $\{(X=1,S=0,U=1),(X=1,S=0,U=0)\}$ and $\{(X=0,S=0,U=1),(X=0,S=0,U=0)\}$, which map to the same value in both models. □

Lemma 3. *There exist values of X,Y such that $P_1(Y|do(X),S=1) \neq P_2(Y|do(X),S=1)$.*

Proof. Fix $X=1,Y=1$, and let us rewrite the desired quantity in Π^* as

$$\begin{aligned} P_i(Y=1|do(X=1),S=1) &= \sum_U P_i(Y|do(X=1),S=1,U)P_i(U|do(X=1),S=1) \\ &= P_i(Y=1|X=1,S=1,U=1)P_i(U=1) \\ &+ P_i(Y=1|X=1,S=1,U=0)P_i(U=0) \end{aligned} \quad (3.18)$$

Since R_i is a function of the tuples $\{(X=1,S=1,U=1),(X=1,S=1,U=0)\}$, it evaluates in M_1 to $\{1,1\}$ and in M_2 to $\{1,0\}$.

Hence, together with the uniformity of $P(U)$, it follows that $R_1=1$ and $R_2=1/2$, which finishes the proof. □

By Lemma 1, Lemmas 2 and 3 prove Theorem 8. □

The concept of confounded components (*C-components*) was introduced in (TP02) to represent clusters of variables connected through bidirected edges, and was instrumental in establishing a number of conditions for ordinary identification. If G is not a C-component itself, it can be uniquely partitioned into a set $\mathcal{C}(G)$ of C-components. We recast this concept in the context of transportability.[6]

Definition 11 (sC-component)**.** *Let G be a selection diagram such that a subset of its bidirected arcs forms a spanning tree over all vertices in G. Then G is a sC-component (selection confounded component).*

A special subset of C-components that embraces the ancestral set of Y was noted by (SP06b) to play an important role in deciding identifiability – this observation can also be applied to transportability, as formulated next.

Definition 12 (sC-tree)**.** *Let G be a selection diagram such that $\mathcal{C}(G) = \{G\}$, all observable nodes have at most one child, there is a node Y, which is a descendent of all nodes, and there is a selection node pointing to Y. Then G is called a Y-rooted sC-tree (selection confounded tree).*

The presence of this structure (and generalizations) will prove to be an obstacle to transportability of causal effects. For instance, the s-bow arc in Fig. 3.7(a) is a Y-rooted sC-tree where we know $P^*_x(y)$ is not transportable there.

In certain classes of problems, the absence of such structures will prove sufficient for transportability. One such class is explored below, and consists of models in which the set X coincides with the parents of Y.

Theorem 9. *Let G be a selection diagram. For any node Y, the effects $P^*_{Pa(Y)}(y)$ is transportable if there is no subgraph of G which forms a Y-rooted sC-tree.*

Proof. See Appendix A. □

Theorem 9 provides a tractable transportability condition for the Controlled Direct Effect (CDE) – a key concept in modern mediation analysis, which permits the decomposition of effects into their direct and indirect components (Pea01; Pea12b). CDE is defined as the effect of X on Y when all other parents of Y are held constant, and it is identifiable if and only if $P^*_{Pa(Y)}(y)$ is identifiable (Pearl, 2009, pp. 128).

[6] Departing from results given in (SGS93; GP95; PR95; Hal98; KM99), the advent of C-components complements the notion of *inducing path*, which was earlier introduced in (VP90), and opened the path for several observations culminating in the results proving completeness of the *do-calculus* for non-parametric identification of causal effects by (HV06a; SP06b).

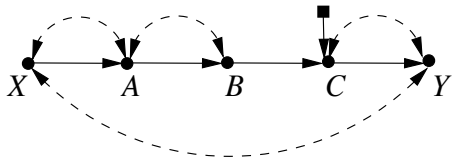


Figure 3.8: Example of a selection diagram in which $P(Y|do(X))$ is not transportable, there is no sC-tree but there is a sC-tree.

The selection diagram in Fig. 3.2(a) does not contain any Y-rooted sC-trees as subgraphs, and therefore the direct effects (causal effects of Y's parents on Y) is indeed transportable. In fact, the transportability of CDE can be determined by a more visible criterion:

Corollary 4. *Let G be a selection diagram. Then for any node Y, the direct effect $P^*_{Pa(Y)}(y)$ is transportable if there is no S node pointing to Y.*

Proof. See Appendix A. □

Generalizing to arbitrary effects, the following result provides a necessary condition for transportability whenever the whole graph is a sC-tree.

Theorem 10. *Let G be a Y-rooted sC-tree. Then the effects of any set of nodes in G on Y are not transportable.*

Proof. See Appendix A. □

The next corollary demonstrates that sC-trees are obstacles to the transportability of $P^*_x(y)$ even when they do not involve Y, i.e., transportability is not a local problem – if there exists a node W that is an ancestor of Y but not necessarily "near" it, transportability is still prohibited (see Fig. 3.7(b)). This fact anticipates that transporting causal effects for singletons is not necessarily easier than the general problem of transportability.

Corollary 5. *Let G be a selection diagram, and X and Y a set of variables. If there exists a node W that is an ancestor of some node $Y \in Y$ such that there exists a W-rooted sC-tree which contains any variables in X, then $P^*_x(y)$ is not transportable.*

Proof. See Appendix A. □

We now generalize the definition of sC-trees (and Theorem 10) in two ways: first, Y is augmented to represent a set of variables; second, S-nodes can point to any variable within the sC-component, not necessarily to root nodes. For instance, consider the graph G in Fig. 3.8. Note that there is no Y-rooted sC-tree nor W-rooted sC-tree in G (where W is an ancestor of Y), and so the previous results cannot be applied even though the effect of X on Y is not transportable in G – still, there exists a Y-rooted sC-forest in G, which will prevent the transportability of the causal effect.

Definition 13 (sC-forest)**.** *Let G be a selection diagram, where Y is the maximal root set. Then G is a Y-rooted sC-forest if G is a sC-component, all observable nodes have at most one child, and there is a selection node pointing to some vertex of G (not necessarily in Y).*

We next conveniently introduce a structure that witnesses non-transportability characterized by a pair of sC-forests. Transportability will be shown impossible whenever such structure exists as an edge subgraph of the given selection diagram.

Definition 14 (s-hedge)**.** *Let X, Y be set of variables in G. Let F, F' be R-rooted sC-forests such that $F \cap X \neq 0$, $F' \cap X = 0$, $F' \subseteq F$, $R \subset An(Y)_{G_{\overline{X}}}$. Then F and F' form a s-hedge for $P^*_x(y)$ in G.*

For instance, in Fig. 3.8, the sC-forests $F' = \{C, Y\}$, and $F = F' \cup \{X, A, B\}$ form a s-hedge to $P_x(y)$. [7] The idea here is similar to the hedge, and we can see a s-hedge as a growing sC-forest F', which doesn't intersect X, to a larger sC-forest F that do intersect X.

We state below the formal connection between s-hedges and non-transportability.

Theorem 11. *Assume there exist F, F' that form a s-hedge for $P^*_x(y)$ in Π and Π^*. Then $P^*_x(y)$ is not transportable from Π to Π^*.*

Proof. See Appendix A. □

To prove that the s-hedges characterize non-transportability in selection diagrams, we construct in the next section an algorithm which transport any causal effects that do not contain a s-hedge.

[7] Note that, by definition, at least one S-node has to appear in both F', F.

3.6 A Complete Algorithm For Transportability of Joint Effects

The algorithm proposed to solve transportability is called **sID** (see Fig. 3.9) and extends previous analysis and algorithms of identifiability given in (Pea95; KM99; TP02; SP06b; HV06a), and we choose to start with the version called **ID** (SP06b) since the hedge structure is explicitly employed, which will show to be instrumental to prove completeness. We build on two observations developed along the chapter:

1. Transportability: Causal relations can can be partitioned into trivially and directly transportable.

2. Non-transportability: The existence of a s-hedge as an edge subgraph of the inputted selection diagram can be used to prove non-transportability.

The algorithm **sID** first applies the typical c-component decomposition on top of the inputted selection diagram D (which, by definition, is also a causal diagram of Π^*), partitioning the original problem into smaller blocks (call these blocks sc-factors) until either the entire expression is transportable, or it runs into the problematic s-hedge structure.

More specifically, for each sc-factor Q, **sID** tries to directly transport Q. If it fails, **sID** tries to trivially transport Q, which is equivalent to solving an ordinary identification problem. **sID** alternates between these two types of transportability, and whenever it exhausts the possibility of applying these operations, it exits with failure with a counterexample for transportability – that is, the graph local to the faulty call witnesses the non-transportability of the causal query since it contains a s-hedge as edge subgraph.

Before showing the more formal properties of **sID**, we demonstrate how **sID** works through the transportability of $Q = P^*(y|do(x))$ in the graph in Fig. 1(c).

Since $D = An(Y)$ and $\mathcal{C}(D \setminus \{X\}) = (C_0, C_1, C_2)$, where $C_0 = D(\{Z\})$, $C_1 = D(\{W\})$, and $C_2 = D(\{V, Y\})$, we invoke line 4 and try to transport respectively $Q_0 = P^*_{x,w,v,y}(z)$, $Q_1 = P^*_{x,z,v,y}(w)$, and $Q_2 = P^*_{x,z,w}(v, y)$. Thus the original problem reduces to transporting $\sum_{z,w,v} P^*_{x,w,v,y}(z)P^*_{x,z,v,y}(w)P^*_{x,z,w}(v, y)$.

Evaluating the first expression, **sID** triggers line 2, noting that nodes that are not ancestors of Z can be ignored. This implies that $P^*_{x,w,v,y}(z) = P^*_x(z)$ with induced subgraph $G_0 = \{X \rightarrow Z, X \leftarrow U_{xz} \rightarrow Z\}$, where U_{xz} stands for the hidden variable between X and Z. **sID** goes to line 5, in which in the local call $\mathcal{C}(D \setminus \{X\}) = \{G_0\}$. Note that in the ordinary identifiability problem the procedure would fail at this point, but **sID** proceeds to line 6 testing whether

function **sID**(y, x, P^*, I, D)
INPUT: x, y value assignments, P^* observational distribution in Π^*, I set of interventional distributions in Π, D a selection diagram, S set of selection nodes.
OUTPUT: Expression for $P^*_x(y)$ in terms of P^*, I or $FAIL(F, F')$.

1 **if** $x = \emptyset$, **return** $\sum_{V \setminus Y} P^*(V)$
2 **if** $V \setminus An(Y)_D \neq \emptyset$,
return sID$(y, x \cap An(Y)_D, \sum_{V \setminus An(Y)_D} P^*, An(Y)_D)$
3 Set $W = (V \setminus X) \setminus An(Y)_{D_{\overline{X}}}$.
if $W \neq \emptyset$, **return sID**$(y, x \cup w, P^*, D)$
4 **if** $\mathcal{C}(D \setminus X) = \{C_0, C_1, ..., C_k\}$,
return $\sum_{V \setminus \{Y,X\}} \prod_i$ **sID**$(c_i, V \setminus c_i, P^*, D)$
5 **if** $\mathcal{C}(D \setminus X) = \{C_0\}$
6 **if** $(S \perp\!\!\!\perp Y \mid X)_{D_{\overline{X}}}$, **return** $P(y|do(x))$
7 **if** $\mathcal{C}(D) = \{D\}$, **FAIL**(D, C_0)
8 **if** $C_0 \in \mathcal{C}(D)$, **return** $\sum_{C_0 \setminus Y} \prod_{i|V_i \in C_0} P^*(v_i|V_D^{(i-1)})$
9 **if** $(\exists C')C_0 \subset C' \in \mathcal{C}(D)$, **return sID**$(y, x \cap C'$,
$\prod_{i|V_i \in C'} P^*(V_i|V_D^{(i-1)} \cap C', v_D^{(i-1)} \setminus C'), C')$.

Figure 3.9: Modified version of identification algorithm capable of recognizing transportable relations.

$(S \perp\!\!\!\perp Z|X)_{D_{\overline{X}}}$. The test comes true, which makes **sID** directly transport Q_0 with data from the experimental population Π, i.e., $P^*_x(z) = P_x(z)$.

Evaluating the second expression, **sID** again triggers line 2, which implies that $P^*_{x,z,v,y}(w) = P^*_{x,z}(w)$ with induced subgraph $G_1 = \{X \to Z, Z \to W, X \leftarrow U_{xz} \to Z\}$. **sID** goes to line 5, in which in the local call $\mathcal{C}(D \setminus \{X\}) = \{G_1\}$. Thus it proceeds to line 6 testing whether $(S \perp\!\!\!\perp W|X, Z)_{D_{\overline{X,Z}}}$. The test comes true again, which makes **sID** directly transport Q_1 with data from the experimental population Π, i.e., $P^*_{x,z}(w) = P_{x,z}(w)$.

Evaluating the third expression, **sID** goes to line 5 in which $\mathcal{C}(D \setminus \{X, Z, W\}) = \{G_2\}$, where $G_2 = \{V \to Y, S \to V, V \leftarrow U_{vy} \to Y\}$. It proceeds to line 6 testing whether $(S \perp\!\!\!\perp W|X, Z)_{D_{\overline{X,Z}}}$, which is false in this case. It tests the other conditions until it reaches line 9, in which $C' = G_0 \cup G_2 \cup \{X \leftarrow U_{xy} \to Y\}$. Thus it tries to transport $Q'_2 = P^*_{x,z}(v, y)$ over the induced graph C', which stands for ordinary identification, and trivially yields $\sum_v P^*(v|w)P^*(y|v, w)$. The return of these calls composed indeed coincide with the expression provided in the first section.

We prove next soundness and completeness of **sID**.

Theorem 12 (soundness)**.** *Whenever **sID** returns an expression for $P^*_x(y)$, it is correct.*

Proof. See Appendix A. □

Theorem 13. *Assume **sID** fails to transport $P^*_x(y)$ (executes line 7). Then there exists $X' \subseteq X$, $Y' \subseteq Y$, such that the graph pair D, C_0 returned by the fail condition of **sID** contain as edge subgraphs sC-forests F, F' that form a s-hedge for $P^*_{x'}(y')$.*

Proof. See Appendix A. □

Corollary 6 (completeness)**.** ***sID** is complete.*

Proof. See Appendix A. □

Corollary 7. *$P^*_x(y)$ is transportable from Π to Π^* in G if and only if there is not s-hedge for $P_{x'}(y')$ in G for any $X' \subseteq X$ and $Y' \subseteq Y$.*

Proof. See Appendix A. □

Theorem 14. *The rules of do-calculus, together with standard probability manipulations are complete for establishing transportability of all effects of the form $P^*_x(y)$.*

Proof. See Appendix A. □

3.7 Conclusions

Given judgemental assessments of how target populations may differ from those under study, the chapter offers a formal representational language for making these assessments precise and for deciding whether causal relations in the target population can be inferred from those obtained in an experimental study.

This chapter introduces a set of intuitive conditions for deciding transportability. When such inference is possible, the criteria provided by Theorems 6 and 7 yield transport formulae, namely, principled ways of calibrating the transported relations so as to properly account for differences in the populations. These transport formulae enable the investigator to select the essential measurements in both the experimental and observational studies, and thus minimize measurement costs and sample variability.

Despite the power and intuitiveness of these results, they are only sufficient to decide transportability, so we seek a complete characterization for the class of all transportable relations. Accordingly, Theorem 7 provides the first step towards this direction, introducing a necessary graphical condition for deciding transportability. In the sequel, we provide complete (necessary and sufficient) algorithmic (Corollary 5) and graphical (Corollary 6) conditions for deciding whether causal effects in the target population are estimable from both the statistical information available and the causal information transferred from the experiments. Furthermore, the procedure known as **sID** (Fig. 3.9) not only decides, but also returns a transport formula in case it exists, that is, a way of combining observational and experimental information to synthesize bias-free estimate of the desired causal relation. Finally, Theorem 10 shows that the algebraic method of the do-calculus is also complete for establishing transportability.

The inferences licensed hitherto represent the worst-case analysis, since we have assumed, in the tradition of non-parametric modeling, that every variable may potentially be an effect-modifier (or moderator.) If one is willing to assume that certain relationships are non-interactive (as in additive models), additional transport licenses may be issued, beyond those sanctioned by these results.

While the results of this chapter concern the transfer of causal information from experimental to observational studies, the method can also benefit in transporting statistical findings from one observational study to another ((PB11a)). The methodology described here is also applicable in the selection of *surrogate endpoints*, namely, variables that would allow good predictability of an outcome for both treatment and control. ((EH89)) Using the representational power of "selection diagrams", we have proposed a causally principled definition of "surrogate endpoint" and suggested how valid surrogates can be identified in a complex network of cause-effect relationships (PB11a).

Of course, our entire analysis is based on the assumption that the analyst is in possession of sufficient background knowledge to determine, at least qualitatively, where two populations may differ from one another. In practice, such knowledge may only be partially available and, as is the case in every mathematical exercise, the benefit of the analysis lies primarily in understanding what knowledge is needed for the task to succeed and how sensitive conclusions are to knowledge that we do not possess.

This chapter provides the semantics as well as a set of graphical and algorithmic conditions for deciding transportability, which constitutes the blueprint for any analysis of problems within the transportability family. In the sequel, we build on these results and relax two assumptions made throughout this chapter: we allow for many source domains (instead of only one) and we assume that only a limited set of experiments are available for use (instead of all experiments).

CHAPTER 4

Transportability from Multiple Studies with Limited Experiments

4.1 Introduction

In this chapter, we consider the problem of transferring causal knowledge collected in several heterogeneous domains to a target domain in which only passive observations and limited experimental data can be collected, which we call mz-transportability. Specifically, the mz-transportability problem concerns the transfer of causal knowledge from a heterogeneous collection of source domains $\Pi = \{\pi_1, ..., \pi_n\}$ to a target domain π^*. In each domain $\pi_i \in \Pi$, experiments over a set of variables Z_i can be performed, and causal knowledge gathered. In π^*, potentially different from π_i, only passive observations and limited experimental data over Z^* can be collected. This problem generalizes the one-dimensional version of transportability with unrestricted experiments previously studied.

Interestingly, while certain effects might not be individually transportable to the target domain from the experiments in any of the available sources, combining different pieces from the various sources may enable their estimation. Conversely, it is also possible that effects are not estimable from multiple experiments in individual domains, but they are from experiments scattered throughout domains.

The goal of this chapter is to make sense of these intricacies, and more generally, to understand under what conditions a causal effect is (non-parametrically) estimable from the available data scattered throughout the different domains. We provide graphical and algorithmic conditions for deciding mz-transportability.

The chapter is organized as follows. In section 4.2, we show relaxations of the assumptions of transportability as encountered in practical settings in the literature. First, we show examples in which a causal relation is not transportability from the individual domains but is transportable from multiple domains combined. Second, we show examples of transportability in which only limited experiments are available in the source domain. In section 4.4, we elaborate on the combination of the previous relaxations noticing that it yields non-trivial instances of transportability, which we call mz-transportability. Then, we formally define mz-transportability and reduce it to a problem of symbolic transformations

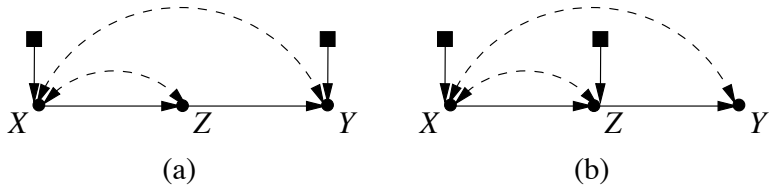


Figure 4.1: Selection diagrams illustrating impossibility of obtaining $P^*(y|do(x))$ through individual transportability from π_a and π_b to π^*, yet a more elaborated analysis yield the desired result combining different pieces from both domains.

in do-calculus. In section 4.5, we establish a necessary and sufficient condition for deciding the feasibility of mz-transportability, i.e., whether causal effects in the target domain are estimable from the information available. In section 4.5, we construct an algorithm based on the graphical criterion for deciding transportability without relying on algebraic manipulations. We show that the algorithm for computing the transport formula is in fact complete, that is, failure of the algorithm implies non-existence of a transport formula. Finally, we show that the do-calculus is complete for the mz-transportability class.

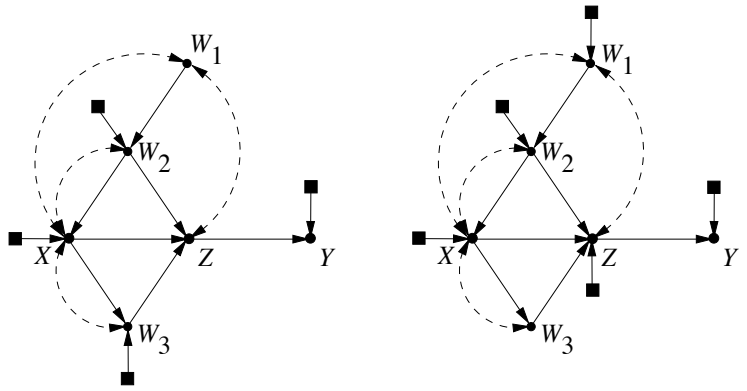
4.2 Relaxations of Transportability

4.2.1 Transportability from Multiple Domains

Consider the problem of transporting causal relations when unrestricted experiments are available in different source domains, which we call m-transportability. One might surmise that multiple pairwise transportability, as studied in the previous chapter, would be sufficient to solve this problem, but this is not the case. To witness, consider Fig. 4.1 which concerns the transference of experimental results from two sources ($\{\pi_a, \pi_b\}$) to infer the effect of X on Y in π^*, $R(\pi^*) = P^*(y|do(x))$. In these graphs, X may represent the treatment (e.g., drug), Z represents an intermediate variable (e.g., biomarker), and Y represents the outcome (e.g., survival).

If we try to directly transport $R(\pi^*)$ from each source domain separately, this is not allowed since we can rewrite $P^*(y|do(x)) = P(y|do(x), S)$, and the condition for directly transporting this relation does not hold, i.e., it is not true that $(S \perp\!\!\!\perp Y|X)_{D^i_{\overline{X}}}$ in π_i, for $i = \{a, b\}$.[5] However, we can decompose the target

[5]Indeed, the impossibility follows more specifically from the completeness of the do-calculus for ordinary transportability as shown in the previous chapter (Thm. 14).



(a) (b)

Figure 4.2: Selection diagrams illustrating a more involved analysis that yields an estimand (Eq. 4.5) for the target quantity which combines information from three domains, the two sources π_a and π_b together with the target π^*.

relation recursively as follows:

$$R(\pi^*) = \sum_z P^*(y|do(x),z)P^*(z|do(x)) \tag{4.1}$$

$$= \sum_z P^*(y|do(x),do(z))P^*(z|do(x)) \tag{4.2}$$

$$= \sum_z P^*(y|do(z))P^*(z|do(x)), \tag{4.3}$$

where Eq. (4.1) follows by conditioning on Z, Eq. (4.2) follows by rule 2 of the do-calculus since $(Z \perp\!\!\!\perp Y|X)_{D_{\overline{X}\underline{Z}}}$ holds, and Eq. (4.3) follows from rule 3 of the do-calculus since $(X \perp\!\!\!\perp Y|Z)_{D_{\overline{X,Z}}}$ holds, where D is the causal diagram of π^* (despite the S-nodes).

Finally, we can now directly transport each of these pieces individually from the source domains noticing that $P^*(y|do(z))$ is directly transportable from π_b giving that $(S \perp\!\!\!\perp Y|Z)_{D^{(b)}_{\overline{Z}}}$, and $P^*(z|do(x))$ is directly transportable from π_a given that $(S \perp\!\!\!\perp Z|X)_{D^{(a)}_{\overline{X}}}$. These yields, respectively, $P^*(y|do(z)) = P^{(b)}(y|do(z))$ and $P^*(z|do(x)) = P^{(a)}(z|do(x))$, and therefore the target relation can be written as,

$$R(\pi^*) = \sum_z P^{(a)}(z|do(x))P^{(b)}(y|do(z)), \tag{4.4}$$

which is a type of convolution, we combine over the entire support of Z the effect X on Z on π_a ($P^{(a)}(z|do(x))$) with the effect of Z on Y in π_b ($P^{(b)}(y|do(z))$).

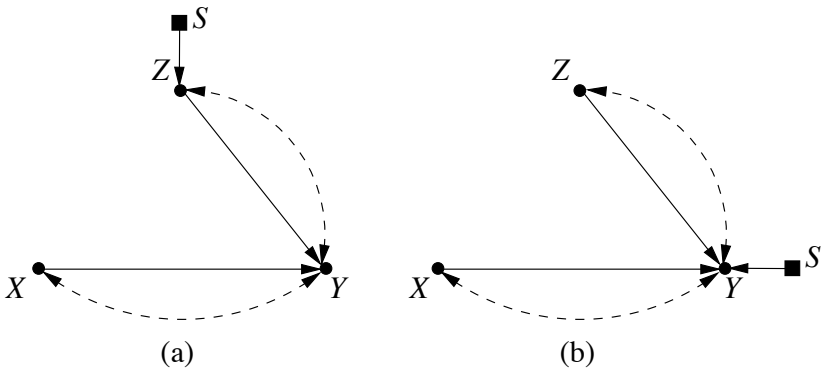


Figure 4.3: Collection of heterogeneous selection diagrams in which the target relation $P^*(y|do(x))$ is not m-transportable from both domains.

For a somewhat more involved example, consider the selection diagrams in Fig. 4.2, and the task of deciding whether there exists an unbiased estimand for the relation $R(\pi^*) = P^*(y|do(x))$. It is not difficult to show that $R(\pi^*)$ is not (individually) transportable from the domains π_a and π_b (Thm. 14), however, it turns out that this relation is transportable from the domains when treated in conjunction (i..e., m-transportable). A less trivial analysis is required in this case though, which yields the following transport formula for $R(\pi^*)$:

$$\sum_{w_1,w_2,w_3,z} P^*(y|z)P^{(a)}_{x,w_2,w_3}(w_1,z)P^*(w_2|w_1)P^{(b)}_{x,w_1,w_2}(w_3) \tag{4.5}$$

In this case we have a witness showing that $R(\pi^*)$ is transportable from the combination of the two sources together with the target domain, but the question arises how to perform a systematic decomposition guided by a guarantee that when it fails, there is no alternative way to decompose the target relation $R(\pi^*)$ in order to transport it from the available data in the source domains.

Consider again the s-bow arc (Fig. 3.7(a)), which is the smallest graph where $R(\pi^*) = P^*(y|do(x))$ is not transportable (Thm. 8). This structure can be trivially extended to the m-transportability case assuming that two domains have selection diagrams identical to the s-bow arc. It is obvious that $R(\pi^*)$ cannot be obtained from the available data; note that there is no possible alternative decomposition for R, and R is neither trivially nor directly transportable from any of the domains. The reduction of m-transportability to ordinary transportability can be justified for any causal relation and collection of domains where (1) the selection diagrams coincide and (2) the target quantity is not pairwise-transportable, which implies that the target relation is not m-transportable.

This, however, does not exhaust the possible cases of impossibility of m-transportability. Consider Fig. 4.3 in which the source domains do not share

selection diagrams and the target quantity is $R(\pi^*) = P^*(y|do(x))$. If an oracle claims that $R(\pi^*)$ is not m-transportable, it is still not trivial to show that this claim is true. Formally, we need to display two models M_1, M_2 such that the following relations hold:

$$\begin{cases} P^{(i)}_{M_1}(X,Z,Y) = P^{(i)}_{M_2}(X,Z,Y), \\ P^{(i)}_{M_1}(X,Y|do(Z)) = P^{(i)}_{M_2}(X,Y|do(Z)), \\ P^{(i)}_{M_1}(X,Z|do(Y)) = P^{(i)}_{M_2}(X,Z|do(Y)), \\ P^{(i)}_{M_1}(Y|do(X),do(Z)) = P^{(i)}_{M_2}(Y|do(X),do(Z)), \\ P^{*}_{M_1}(X,Z,Y) = P^{*}_{M_2}(X,Z,Y), \end{cases} \tag{4.6}$$

for $i = \{a, b\}$ and all values of X, Y, Z, and also,

$$P^*_{M_1}(Y|do(X)) \neq P^*_{M_2}(Y|do(X)), \tag{4.7}$$

for some value of X and Y. This is certainly a more involved task than proving lack of ordinary transportability since there are more equalities and inequalities constraints to maintain. We show how to construct such a certificate in Appendix B, which will be useful for the completeness proof of the most general case.

4.2.2 Transportability with Limited Experiments

In real world applications, it may happen that certain controlled experiments cannot be conducted in the source environment (for financial, ethical, or technical reasons), so only a limited amount of experimental information can be gathered. A natural question arises whether the investigator in possession of a limited set of experiments would still be able to estimate the desired effects at the target. For simplicity, in this section, we consider the case when only one source domain is available. We call this variant the z-transportability problem.

To illustrate z-transportability, consider Fig. 4.4(a) and assume we wish, again, to estimate $P^*(y|do(x))$ but, now, X cannot be randomized in the source. Instead, variable Z can be randomized, and we ask whether we can still estimate $P^*(y|do(x))$ despite this constraint and despite the fact that the two populations differ in the prior probabilities of Z (as indicated by the S-node). [1]

Fortunately, in this case, the problem has a positive solution as can be seen from the following derivation. First apply Rule 3 of the *do*-calculus to add $do(z)$ to the expression,

$$P^*(y|do(x)) = P^*(y|do(x),do(z)) \text{ since } (Y \perp\!\!\!\perp Z|X)_{G_{\overline{XZ}}}$$

[1] A typical example is whether we can estimate the effect of cholesterol (X) on heart failure (Y) by experiments on diet (Z) given that cholesterol levels cannot be randomized.

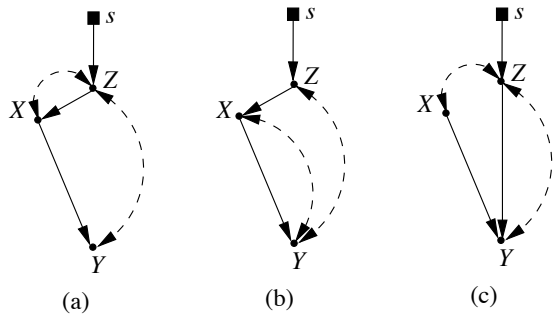


Figure 4.4: Selection diagrams illustrating transportability with limited experiments of the causal effect $R = P^*(y|\hat{x})$. R can be transported with experiments on Z in model (a), but not in (b) and (c).

Then apply Rule 2 to exchange $do(x)$ with x:

$$P^*(y|do(x), do(z)) = P^*(y|x, do(z)) \text{ since } (Y \perp\!\!\!\perp X|Z)_{G_{\underline{X}\overline{Z}}}$$

This last expression can be rewritten as,

$$P^*(y|x, do(z)) = P(y|x, do(z), s) = \frac{P(y, x|do(z))}{P(x|do(z))}, \tag{4.8}$$

where the first equality follows from the definition of selection diagram and the second using the separation of S from $\{X, Y\}$ after intervening on Z. Therefore, performing an experiment on Z in Π suffices to estimate the causal effect of X on Y in Π^* (without resorting to experimentation on X.)

There are subtle features of z-transportability that are worth illustrating. Whereas the graph in Fig. 4.4(a) permits the effect to be z-transported (with experiments over Z), the graph in Fig. 4.4(b) does not. One is tempted to explain this difference by noting that in the mutilated graph from which the edges incoming to Z are cut (to simulate intervention), the causal effect of X on Y is identifiable in Fig. 4.4(a) but not in (b). The fact that this is not the case is shown in the graph in Fig. 4.4(c). The resulting mutilated graph in this case entails both the identifiability and transportability of $P^*(y|do(x))$, but this effect is neither identifiable, nor transportable, nor z-transportable (see Appendix B).

In a more involved manner, one might surmise that the solution for the z-identification problem (Chapter 6) could yield the solution for z-transportability – z-identification asks for expressing the causal relation $R = P(y|do(x))$ in terms of experiments on Z (in a fixed domain Π) – however, this turns out to not be the case as well. To witness, consider the diagram G in Fig. 4.5(a), and note that even though R is z-identifiable in Π, it is not the case that R is z-transportable. (The reason is that the S-node pointing to W, which is in a

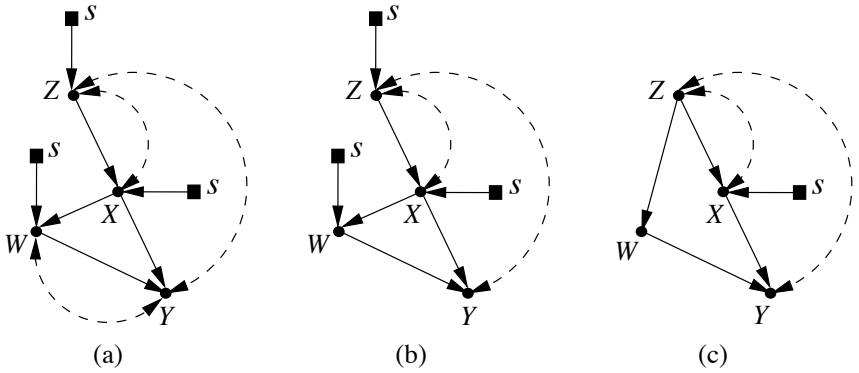


Figure 4.5: Selection diagrams illustrating the non-trivial relationship among the problems of z-identifiability, transportability, and restricted transportability.

confounded relationship with Y, disallows transportability, but is unrelated to its identification counterpart.)

Furthermore, consider the same task in regard to Fig. 4.5(b), a simple analysis for z-identification in the source would yield expression similar to the one in Fig. 4.4(a),

$$P(y|do(x)) = \frac{P(y, x|do(z))}{P(x|do(z))}, \tag{4.9}$$

but in this case, the availability of the ratio in eq. (4.9) is not sufficient for estimating the target quantity $R = P^*(y|do(x))$ in Π^*. Interestingly enough, the quantity R is z-transportable through the transport formula

$$P^*(y|do(x)) = \sum_w P(y|x, w, do(z))P^*(w|x, z), \tag{4.10}$$

which combines experimental results over Z obtained in the source Π, $P(y|x, w, do(z))$, with observational aspects of the target domain, $P^*(w|x, z)$, to obtain an experimental claim $P^*(y|do(x))$ about the target (see Appendix B).

One might further surmise that the s-hedge structure (Def. 14), which characterizes the set of transportable relations, could lead to a characterization for z-transportability as well, but this is not the case. To witness, note that there is no s-hedge in Fig. 4.4(b) and 4.5(c), so the effect $R = P^*(y|do(x))$ is transportable in both scenarios, but R is not z-transportable in these cases. Specifically, there is no s-hedge in Fig. 4.4(b) because, even though there are s^*-trees $F' = \{Y\}$, $F = \{X, Z\} \cup F'$, there is not a selection node pointing F'. Clearly, if a quantity R is not transportable, R is also not z-transportable. The converse is obviously not true, when R is transportable, it is the case that R might be either z-transportable (e.g., Fig. 4.5(b)) or not (e.g., Fig. 4.4(b) and 4.5(c)).

In spite of these observations, it is clear that z-transportability reduces neither to ordinary transportability nor to z-identifiability, which leaves open the question of how to algorithmically characterize transportability with limited experiments.

4.3 Formalizing mz-Transportability

In practice, it is common that different experiments can be conducted in various different domains, which can be seen as the combination of the relaxations previously discussed and constitutes the so-called mz-transportability problem.

One might surmise that multiple pairwise z-transportability would be sufficient to solve mz-transportability, but this is not the case. To witness, consider Fig. 4.6(a,b) which concerns the transport of experimental results from two sources ($\{\pi_a, \pi_b\}$) to infer the effect of X on Y in π^*, $R = P^*(y|do(x))$. In these diagrams, X may represent the treatment (e.g., cholesterol level), Z_1 represents a pre-treatment variable (e.g., diet), Z_2 represents an intermediate variable (e.g., biomarker), and Y represents the outcome (e.g., heart failure). We assume that experimental studies randomizing $\{Z_1, Z_2\}$ can be conducted in both domains.

A simple analysis can show that R cannot be z-transported from either source alone (even with both experiments), but it turns out that combining the experiments from both sources allows one to determine the effect in the target. In order to show how this is possible, we can conveniently decompose R as follows:

$$\begin{aligned} P^*(y|do(x)) &= P^*(y|do(x), do(Z_1)) && (4.11)\\ &= \sum_{z_2} P^*(y|do(x), do(Z_1), z_2) P^*(z_2|do(x), do(Z_1)) && (4.12)\\ &= \sum_{z_2} P^*(y|do(x), do(Z_1), do(z_2)) P^*(z_2|do(x), do(Z_1)), && (4.13) \end{aligned}$$

where Eq. (4.11) follows by rule 3 of the do-calculus since $(Z_1 \perp\!\!\!\perp Y|X)_{D_{\overline{X,Z_1}}}$ holds, we condition on Z_2 in Eq. (4.12), and Eq. (4.13) follows by rule 2 of the do-calculus since $(Z_2 \perp\!\!\!\perp Y|X, Z_1)_{D_{\overline{X,Z_1},\underline{Z_2}}}$, where D is the diagram in π^* (despite the location of the S-nodes). Now, we rewrite the first term of Eq. (4.13):

$$\begin{aligned} P^*(y|do(x), do(Z_1), do(z_2)) &= P(y|do(x), do(Z_1), do(z_2), S_a, S_b) && (4.14)\\ &= P(y|do(x), do(Z_1), do(z_2), S_b) && (4.15)\\ &= P(y|do(z_2), S_b) && (4.16)\\ &= P^{(a)}(y|do(z_2)), && (4.17) \end{aligned}$$

where Eq. (4.14) follows from the definition of selection diagram, Eq. (4.15) follows from rule 1 of the do-calculus since $(S_a \perp\!\!\!\perp Y|Z_1, Z_2, X)_{D^{(a)}_{\overline{Z_1,Z_2,X}}}$, Eq. (4.16)

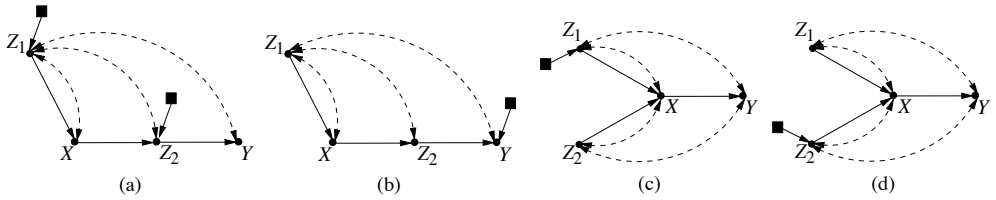


Figure 4.6: (a,b) Selection diagrams illustrating the impossibility of estimating R through individual transportability from π_a and π_b when experiments over $\{Z_1, Z_2\}$ are available. If experiments over $\{Z_2\}$ is available in π_a and over $\{Z_1\}$ in π_b, R is transportable. (c,d) Selection diagrams illustrating the opposite phenomenon – transportability through multiple domains is not feasible, but if experiments over $Z = \{Z_1, Z_2\}$ is available in one domain, transportability is feasible.

follows from rule 3 of the do-calculus since $(Z_1, X \perp\!\!\!\perp Y | Z_2)_{D^{(a)}_{\overline{Z_1,Z_2,X}}}$, and Eq. (4.17) follows from the definition of selection diagrams.

Now, we can rewrite the second term of Eq. (4.13) as follows:

$$\begin{aligned} P^*(z_2|do(x), do(Z_1)) &= P(z_2|do(x), do(Z_1), S_a, S_b) && (4.18)\\ &= P(z_2|do(x), do(Z_1), S_a) && (4.19)\\ &= P(z_2|x, do(Z_1), S_a) && (4.20)\\ &= P^{(b)}(z_2|x, do(Z_1)), && (4.21) \end{aligned}$$

where Eq. (4.18) follows from the definition of selection diagram, Eq. (4.19) follows from rule 1 of the do-calculus since $(S_b \perp\!\!\!\perp Z_2 | Z_1, X)_{D^{(b)}_{\overline{Z_1,X}}}$, Eq. (4.20) follows from rule 2 of the do-calculus since $(X \perp\!\!\!\perp Z_2 | Z_1)_{D^{(b)}_{\overline{Z_1}\underline{X}}}$, and Eq. (4.21) follows from the definition of selection diagrams. Finally, we can substitute back Eqs. (4.17) and (4.21) in Eq. (4.13),which yields the following transport formula

$$R = \sum_{z_2} P^{(a)}(y|do(z_2)) P^{(b)}(z_2|x, do(Z_1)) \quad (4.22)$$

This transport formula is a mixture of the experimental result over $\{Z_1\}$ from π_b, $P^{(b)}(z_2|x, do(Z_1))$, with the result of the experiment over $\{Z_2\}$ in π_a, $P^{(a)}(y|do(z_2))$, and constitute a consistent estimand of the target relation in π^*. Conversely, it is the case that if the domains in which experiments were conducted were reversed – i.e., $\{Z_1\}$ is randomized in π_a and $\{Z_2\}$ in π_b) – it will not be possible to transport R by any method, the target relation is simply not computable given the assumptions encoded in the diagrams (formally shown in the next section). [2]

[2] Despite the fact that the directed paths from Z_2 to Y were *not* blocked by X, randomizing

Further consider Fig. 4.6(c,d) which illustrates the opposite phenomenon. In this case, if experiments over $\{Z_2\}$ are available in domain π_a and over $\{Z_1\}$ in π_b, R is not mz-transportable. However, if $\{Z_1, Z_2\}$ are available in the same domain, say π_a, R is mz-transportable. To witness,

$$\begin{aligned} P^*(y|do(x)) &= P^*(y|do(x), do(z_1), do(z_2)) && (4.23)\\ &= P(y|do(x), do(z_1), do(z_2), S_a) && (4.24)\\ &= P(y|do(x), do(z_1), do(z_2)) && (4.25)\\ &= P^{(a)}(y|do(x), do(z_1), do(z_2)) && (4.26) \end{aligned}$$

where Eq. (4.23) follows from rule 3 of the do-calculus since $(Y \perp\!\!\!\perp Z_1, Z_2|X)_{D^{(a)}_{\overline{X,Z_1,Z_2}}}$, Eq. (4.24) follows from the definition of selection diagram, Eq. (4.25) follows from rule 1 of the do-calculus since $(S_a \perp\!\!\!\perp Y|X, Z_1, Z_2)_{D^{(a)}_{\overline{X,Z_1,Z_2}}}$, and Eq. (4.26) follows from the definition of selection diagrams. [3]

These results illustrate some of the subtle issues mz-transportability entails, which cannot be immediately cast in terms of other variants of transportability. In order to pursue a more systematic treatment of this problem, and using a collection of selection diagrams as basic representational language, harnessing the concepts of interventions, do-calculus, identifiability, and transportability, we can formally define the mz-transportability problem as follows.

Definition 15 (mz-Transportability)**.** *Let $\mathcal{D} = \{D^{(1)}, ..., D^{(n)}\}$ be a collection of selection diagrams relative to source domains $\Pi = \{\pi_1, ..., \pi_n\}$, and target domain π^*, respectively, and Z_i (and Z^*) be the variables in which experiments can be conducted in domain π_i (and π^*). Let $\langle P^i, I^i_z\rangle$ be the pair of observational and interventional distributions of π_i, where $I^i_z = \bigcup_{Z' \subseteq Z_i} P^i(v|do(z'))$, and in an analogous manner, $\langle P^*, I^*_z\rangle$ be the observational and interventional distributions of π^*. The causal effect $R = P^*_x(y)$ is said to be mz-transportable from Π to π^* in $\mathcal{D}$ if $P^*_x(y)$ is uniquely computable from $\bigcup_{i=1,...,n}\langle P^i, I^i_z\rangle \cup \langle P^*, I^*_z\rangle$ in any model that induces the diagrams $\mathcal{D}$.*

While this definition might appear convoluted, it is nothing more than a formalization of the statement "R needs to be uniquely computable from the information set IS alone." Naturally, when IS has many components (multiple observational and interventional distributions), it becomes lengthy.

There are interesting connections between mz-transportability and the other variants of transportability discussed so far. The mz-transportability problem

Z_2 was instrumental to yield transportability in this case, which suggests how different this type of transportability is from the z-identifiability problem.

[3] Despite the fact that the transport formula relies on experiments over Z_1 and Z_2, the formula is independent of their specific values.

can be reduced to m-transportability whenever $Z_i = V$, for all i, and $Z^* = \emptyset$. Further, the mz-transportability problem can be reduced to z-transportability whenever $n = 1$ and $Z^* = \emptyset$. Also, the mz-transportability problem can be reduced to transportability whenever $n = 1$, $Z_i = V$, and $Z^* = \emptyset$. [4]

The requirement of computability from $\langle P^*, I_z^* \rangle$ and $\langle P^i, I_z^i \rangle$ from all sources given in definition of mz-transportability has a syntactic image in the do-calculus, which is captured by the following sufficient condition:

Theorem 15. *Let $\mathcal{D} = \{D^{(1)}, ..., D^{(n)}\}$ be a collection of selection diagrams relative to source domains $\Pi = \{\pi_1, ..., \pi_n\}$, and target domain π^*, respectively, and S_i represents the collection of S-variables in the selection diagram $D^{(i)}$. Let $\{\langle P^i, I_z^i \rangle\}$ and $\langle P^*, I_z^* \rangle$ be respectively the pairs of observational and interventional distributions in the sources Π and target π^*. The effect $R = P^*(y|do(x))$ is mz-transportable from Π to π^* in $\mathcal{D}$ if the expression $P(y|do(x), S_1, ..., S_n)$ is reducible, using the rules of the do-calculus, to an expression in which (1) do-operators that apply to subsets of I_z^i have no S_i-variables or (2) do-operators apply only to subsets of I_z^*.*

Proof. See Appendix B. □

This result provides a powerful way to syntactically establish mz-transportability, but it is not obvious whether a sequence of applications of the rules of the do-calculus that achieves the reduction required by the theorem exists. If such sequence does not exist, it is not immediately clear whether this would entail the non-existence of a transport formula, so the infeasibility of estimating the target relation (Sec. 4.4). On the other hand, even if such sequence does exist, it is not obvious how to efficiently obtain the eventual transport formula (Sec. 4.5).

4.4 Characterizing mz-Transportable Relations

The goal of this section is to demonstrate whether the non-existence of the reduction sequence in do-calculus as required by Theorem 15 entails the impossibility of transport. We will collect different examples of non-transportability and try to make sense whether there is a pattern in such cases and how to generalize them.

For instance, consider Fig. 4.7(a,b) and the goal of transporting the effect $R = P^*(y|do(x))$ when experiments over $\{X\}$ are available in π_a and over $\{Z\}$ are available in π_b. It is not difficult to see that there is no reducing sequence in this

[4] The requirement that the number of domains is equal to one ($n = 1$) can be seen as if all other selection diagrams besides the one under consideration have selection nodes pointing to all variables, which means that all other source domains are unrelated to the target domain.

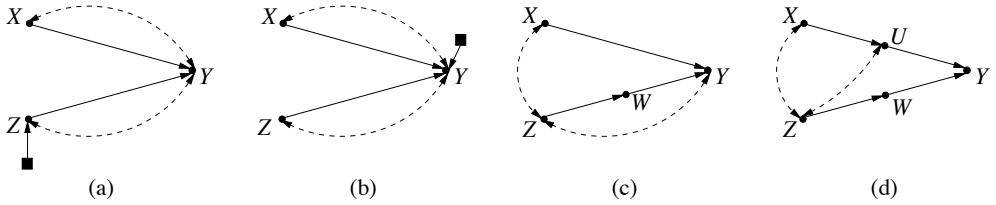


Figure 4.7: (a,b) Diagrams which is not possible to transport $P^*(y|do(x))$ with experiments over $\{X\}$ in π_a and $\{Z\}$ in π_b. (c,d) Example of diagrams in which some paths need to be extended for satisfying the definition of mz^*-shedge.

case, which turns out to imply that the target relation is indeed not transportable. This means that there exist two models that are equally compatible with the data (i.e., both could generate the same dataset) while each model entails a different answer for the effect R (violating the uniqueness requirement of Def. 15). [5] To demonstrate this fact explicitly, we show the existence of two models M_1 and M_2 such that the following equalities and inequality between distributions hold,

$$\begin{cases} P^{(a)}_{M_1}(X,Z,Y) = P^{(a)}_{M_2}(X,Z,Y), \\ P^{(b)}_{M_1}(X,Z,Y) = P^{(b)}_{M_2}(X,Z,Y), \\ P^{(a)}_{M_1}(Z,Y|do(X)) = P^{(a)}_{M_2}(Z,Y|do(X)), \\ P^{(b)}_{M_1}(X,Y|do(Z)) = P^{(b)}_{M_2}(X,Y|do(Z)), \\ P^{*}_{M_1}(X,Z,Y) = P^{*}_{M_2}(X,Z,Y), \end{cases} \tag{4.27}$$

for all values of X, Z, and Y, and

$$P^{*}_{M_1}(Y|do(X)) \neq P^{*}_{M_2}(Y|do(X)), \tag{4.28}$$

for some value of X and Y.

Let V be the set of observable variables and U be the set of unobservable variables in $\mathcal{D}$. Let us assume that all variables are binary. Let $U_1, U_2 \in U$ be the common causes of X and Y and Z and Y, respectively; let $U_3, U_4 \in U$ be the random disturbances exclusive to Z and Y, respectively, and $U_5, U_6 \in U$ be extra random disturbances exclusive to Y. Let S_a and S_b index the model in the following way: the tuples $\langle S_a = 1, S_b = 0\rangle$, $\langle S_a = 0, S_b = 1\rangle$, $\langle S_a = 0, S_b = 0\rangle$ represent domains π_a, π_b, and π^*, respectively. Define the two models as follows:

$$M_1 = \begin{cases} X = U_1 \\ Z = U_2 \oplus (U_3 \wedge S_a) \\ Y = ((X \oplus Z \oplus U_1 \oplus U_2 \oplus (U_4 \wedge S_b)) \wedge U_5) + (\neg U_5 \wedge U_6), \end{cases}$$

[5] This is usually an indication that the current state of scientific knowledge about the phenomenon under consideration (encoded in the form of a selection diagram) does not constraint the observed distributions in such a way that an answer is entailed independently of the details of the functions and probability distribution over the exogenous variables.

and

$$M_2 = \begin{cases} X = U_1 \\ Z = U_2 \oplus (U_3 \wedge S_a) \\ Y = ((Z \oplus U_2 \oplus (U_4 \wedge S_b)) \wedge U_5) \oplus (\neg U_5 \wedge U_6) \end{cases}$$

where $\oplus$ represents the *exclusive or* function. Both models agree in respect to $P(U)$, which is defined as $P(U_i) = 1/2$, $i = 1, ..., 6$. It is not difficult to evaluate these models and note that the constraints given in Eqs. (4.29) and (4.30) are indeed satisfied (including positivity), the result follows. [6]

Consider again the example given in Fig. 4.6(a,b). While it was almost immediate to obtain a do-calculus sequence in which the target relation $R = P^*(y|do(x))$ is transportable with experiments over $\{Z_2\}$ in π_a and over $\{Z_1\}$ in π_b (Eq. (4.22)), it is not the case that such a sequence exists when we exchange the experiments and assume that interventional data is available over $\{Z_1\}$ in π_a and over $\{Z_2\}$ in π_b. Similar to the previous case, to show this fact formally we display two models M_1, M_2 such that the following relations hold (using Def. 15):

$$\begin{cases} P^{(a)}_{M_1}(Z_1, X, Z_2, Y) = P^{(a)}_{M_2}(Z_1, X, Z_2, Y), \\ P^{(b)}_{M_1}(Z_1, X, Z_2, Y) = P^{(b)}_{M_2}(Z_1, X, Z_2, Y), \\ P^{(a)}_{M_1}(X, Z_2, Y|do(Z_1)) = P^{(a)}_{M_2}(X, Z_2, Y|do(Z_1)), \\ P^{(b)}_{M_1}(Z_1, X, Y|do(Z_2)) = P^{(b)}_{M_2}(Z_1, X, Y|do(Z_2)), \\ P^{*}_{M_1}(Z_1, X, Z_2, Y) = P^{*}_{M_2}(Z_1, X, Z_2, Y), \end{cases} \tag{4.29}$$

for all values of Z_1, X, Z_2, and Y, and also,

$$P^*_{M_1}(Y|do(X)) \neq P^*_{M_2}(Y|do(X)), \tag{4.30}$$

for some value of X and Y.

Let V be the set of observable variables and U be the set of unobservable variables in $\mathcal{D}$. Let us assume that all variables in $U \cup V$ are binary. Let $U_1, U_2 \in U$ be the common causes of Z_1 and X and Z_2, respectively; let $U_3, U_4, U_5 \in U$ be a random disturbance exclusive to Z_1, Z_2, and Y, respectively, and $U_6 \in U$ be an extra random disturbance exclusive to Z_2, and $U_7, U_8 \in U$ to Y. Let S_a and S_b index the model in the following way: the tuples $\langle S_a = 1, S_b = 0\rangle$, $\langle S_a = 0, S_b = 1\rangle$, $\langle S_a = 0, S_b = 0\rangle$ represent domains π_a, π_b, and π^*, respectively. Define the two models as follows:

$$M_1 = \begin{cases} Z_1 = U_1 \oplus U_2 \oplus \big(U_3 \wedge S_a\big) \\ X = Z_1 \oplus U_1 \\ Z_2 = (X \oplus U_2 \oplus (U_4 \wedge S_a)) \vee \overline{U_6} \\ Y = (Z_2 \wedge U_5) \oplus (\overline{U_5} \wedge U_7) \oplus (S_b \wedge U_8) \end{cases}$$

[6] See Appendix B for a more refined argument on how to evaluate these models.

and

$$M_2 = \begin{cases} Z_1 = U_1 \oplus U_2 \oplus \big(U_3 \wedge S_a\big) \\ X = U_1 \\ Z_2 = \big(U_4 \wedge S_a \wedge U_6\big) \oplus \overline{U_6} \\ Y = (Z_2 \wedge U_5) \oplus (\overline{U_5} \wedge U_7) \oplus (S_b \wedge U_8) \end{cases}$$

Define in both models $P(U_i) = 1/2$, $i = 1, ..., 8$. It is not difficult to evaluate these models and note that the constraints given in Eqs. (4.29) and (4.30) are satisfied (including positivity), which demonstrates that R is not transportable.

After accumulating positive and negative examples of mz-transportability, we note that one syntactic subtask of mz-transportability is to determine whether certain effects are identifiable in some source domains where interventional data is available. There are two fundamental results developed for identifiability that will be relevant for mz-transportability as well. For completeness, we will repeat some definitions given in the previous chapter. First, we should consider confounded components (or *c-components*), which were defined in (Tia02) and stand for a cluster of variables connected through bidirected edges (which are not separable through the observables in the system). One key result is that each causal graph (and subgraphs) induces an unique C-component decomposition ((Tia02, Lemma 11)). This decomposition was indeed instrumental for a series of conditions for ordinary identification (TP02) and the inability to recursively decompose a certain graph was later used to prove completeness.

Definition 16 (C-component)**.** *Let G be a causal diagram such that a subset of its bidirected arcs forms a spanning tree over all vertices in G. Then G is a C-component (confounded component).*

Subsequently, (SP06c) proposed an extension of *C-components* called *C-forests*, essentially enforcing that each C-component has to be a spanning forest and closed under ancestral relations (Tia02).

Definition 17 (C-forest)**.** *Let G be a causal diagram where Y is the maximal root set. Then G is a Y-rooted C-forest if G is a C-component and all observable nodes have at most one child.*

For concreteness, consider Fig. 4.6(a) and note that there exists a C-forest over nodes $\{Z_1, X, Z_2\}$ and rooted in $\{Z_2\}$. There exists another C-forest over nodes $\{Z_1, X, Z_2, Y\}$ rooted in $\{Y\}$. It is also the case that $\{Z_2\}$ and $\{Y\}$ are themselves trivial C-forests. When we have a pair of C-forests as $\{Z_1, X, Z_2\}$ and $\{Z_2\}$ or $\{Z_1, X, Z_2, Y\}$ and $\{Y\}$ – i.e., the root set does not intersect the treatment variables; these structures are called *hedges* and identifiability was shown to be infeasible whenever a hedge exists (SP06c). Clearly, despite the

existence of hedges in Fig. 4.6(a,b), the effects of interest were shown to be mz-transportable. This example is an indication that hedges alone do not capture the structure needed for characterizing mz-transportability – i.e., a graph might be a hedge (or have a hedge as an edge sub–graph) but the target quantity might still be mz-transportable.

Based on these observations, we propose the following definition that may lead to the boundaries of the class of mz-transportable relations:

Definition 18 (mz^*-shedge)**.** *Let $\mathcal{D} = (D^{(1)}, \ldots, D^{(n)})$ be a collection of selection diagrams relative to source domains $\Pi = (\pi_1, \ldots, \pi_n)$ and target domain π^*, respectively, S_i represents the collection of S-variables in the selection diagram $D^{(i)}$, and let $D^{(*)}$ be the causal diagram of π^*. Let $\{\langle P^i, I^i_z\rangle\}$ be the collection of pairs of observational and interventional distributions of $\{\pi_i\}$, where $I^i_z = \bigcup_{Z' \subseteq Z_i} P^i(v|do(z'))$, and in an analogous manner, $\langle P^*, I^*_z\rangle$ be the observational and interventional distributions of π^*, for Z_i the set of experimental variables in π_i. Consider a pair of R-rooted C-forests $\mathcal{F} = \langle F, F'\rangle$ such that $F' \subset F$, $F' \cap X = \emptyset$, $F \cap X \neq \emptyset$, and $R \subseteq An(Y)_{G_{\overline{X}}}$ (called hedge). We say that the induced collection of pairs of R-rooted C-forests over each diagram, $\langle \mathcal{F}^{(*)}, \mathcal{F}^{(1)}, ..., \mathcal{F}^{(n)}\rangle$, is an mz-shedge for $P^*_x(y)$ relative to experiments $(I^*_z, I^1_z, ..., I^n_z)$ if they are all hedges and one of the following conditions hold for each domain π_i, $i = \{*, 1, ..., n\}$:*

1. *There exists at least one variable of S_i pointing to the induced diagram $F'^{(i)}$, or*
2. *$(F^{(i)} \setminus F'^{(i)}) \cap Z_i$ is an empty set, or*
3. *The collection of pairs of C-forests induced over diagrams, $\langle \mathcal{F}^{(*)}, \mathcal{F}^{(1)}, \ldots, F^{(i)} \setminus Z^*_i, \ldots, \mathcal{F}^{(n)}\rangle$, is also an mz-shedge relative to $(I^*_z, I^1_z, ..., I^i_{z \setminus z^*_i}, ..., I^n_z)$, where $Z^*_i = (F^{(i)} \setminus F'^{(i)}) \cap Z_i$.*

We call mz^*-shedge the mz-shedge in which there exist one directed path from $R \setminus (R \cap De(X)_F)$ to $(R \cap De(X)_F)$ not passing through X (see appendix B).

The definition of mz^*-shedge might appear involved, but it is nothing more than the articulation of the computability requirement of Def. 15 (and implicitly the syntactic goal of Thm. 15) in a more explicit graphical fashion. Specifically, for a certain factor Q^*_i needed for the computation of the effect $Q^* = P^*(y|do(x))$, in at least one domain, (i) it should be enforced that the S-nodes are separable from the inducing root set of the component in which Q^*_i belongs, and further, (ii) the experiments available in this domain are sufficient for solving Q^*_i. For instance, assuming we want to compute $Q^* = P^*(y|do(x))$ in Fig. 4.6(a, b), Q^* can be decomposed into factors $Q^*_1 = P^*_{z_1,x}(z_2)$ and $Q^*_2 = P^*_{z_1,x,z_2}(y)$. It is the case

that for Q_1^*, (i) holds true in π_b and (ii) the experiments available over Z_1 are enough to guarantee the computability of this factor (similar analysis applies to Q_2^*) – i.e., there is no mz^*-shedge and Q^* is computable from the available data.

Def. 18 also asks for the explicit existence of a path from the nodes in the root set $R \backslash (R \cap De(X)_F)$ to $(R \cap De(X)_F)$, a simple example can help to illustrate this requirement. Consider Fig. 4.7(c) and the goal of computing $Q = P^*(y|do(x))$ without extra experimental information. There exists a hedge for Q induced over $\{X, Z, Y\}$ without the node W (note that $\{W\}$ is a c-component itself) and the induced graph $G_{\{X,Z,Y\}}$ indeed leads to a counter-example for the computability of $P^*(z, y|do(x))$. Using this subgraph alone, however, it would not be possible to construct a counter-example for the marginal effect $P^*(y|do(x))$. Despite the fact that $P^*(z, y|do(x))$ is not computable from $P^*(x, z, y)$, the quantity $P^*(y|do(x))$ is identifiable in $G_{\{X,Z,Y\}}$, and so any structural model compatible with this subgraph will generate the same value under the marginalization over Z from $P^*(z, y|do(x))$. Also, it might happen that the root set R must be augmented (Fig. 4.7(d)), so we prefer to add this requirement explicitly to the definition. (There are more involved scenarios that we prefer to omit for the sake of presentation.) After adding the directed path from Z to Y that passes through W, we can construct the following counter-example for Q:

$$M_1 = \begin{cases} X = U_1 \\ Z = U_1 \oplus U_2 \\ W = ((Z \oplus U_3) \vee B) \oplus (B \wedge (1 \oplus Z)) \\ Y = ((X \oplus W \oplus U_2) \wedge A) \oplus \ (A \vee (1 \oplus X \oplus W \oplus U2)), \end{cases}$$

and

$$M_2 = \begin{cases} X = U_1 \\ Z = U_2 \\ W = ((Z \oplus U_3) \vee B) \oplus (B \wedge (1 \oplus Z)) \\ Y = ((W \oplus U_2) \wedge A) \oplus \ (A \vee (1 \oplus W \oplus U2)), \end{cases}$$

with $P(U_i) = 1/2, \forall i$, $P(A) = P(B) = 1/2$. It is not immediate to show that the two models produce the desired property, see Appendix B for a formal proof.

Given that the definition of mz^*-shedge is justified and well-understood, we can now state the connection between hedges and mz^*-shedges more directly:

Theorem 16. *If there is a hedge for $P_x^*(y)$ in G and no experimental data is available (i.e., $I_z^* = \{\}$), there exists an mz^*-shedge for $P_x^*(y)$ in G.*

Proof. See Appendix B. □

Whenever one domain is considered and no experimental data is available, this result states that an mz^*-shedge can always be constructed from a hedge, which

implies that we can operate with mz^*-shedges from now on (the converse holds for $Z = \{\}$). Finally, we can concentrate on the most general case of mz^*-shedges with experimental data in multiple domains as stated in the sequel:

Theorem 17. *Let $\mathcal{D} = \{D^{(1)}, ..., D^{(n)}\}$ be a collection of selection diagrams relative to source domains $\Pi = \{\pi_1, ..., \pi_n\}$, and target domain π^*, respectively, and $\{I_z^i\}$, for $i = \{*, 1, ..., n\}$ defined appropriately. If there is an mz^*-shedge for the effect $R = P_x^*(y)$ relative to experiments $(I_z^*, I_z^1, ..., I_z^n)$ in $\mathcal{D}$, R is not mz-transportable from Π to π^* in $\mathcal{D}$.*

Proof. See Appendix B. □

This is a powerful result that states that the existence of a mz^*-shedge precludes mz-transportability. For concreteness, consider the selection diagrams $\mathcal{D} = (D^{(a)}, D^{(b)})$ relative to domains π_a and π_b in Fig. 4.7(a,b). Our goal is to mz-transport $Q = P^*(y|do(x))$ with experiments over $\{X\}$ in π_a and $\{Z\}$ in π_b. It is the case that there exists an mz^*-shedge relative to the given experiments. To witness, note that $F' = \{Y, Z\}$ and $F = F' \cup \{X\}$, and also that there exists a selection variable S pointing to F' in both domains – the first condition of Def. 18 is satisfied. This is a trivial graph with 3 variables that can be solved by inspection, but it is somewhat involved to efficiently evaluate the conditions of the definition in more intricate structures, which motivates the search for a procedure for recognizing mz^*-shedges that can be coupled with the previous theorem.

4.5 A Complete Algorithm For mz-Transportability of Joint Effects

In this section, we generalize the algorithm given in section 3.6 to obtain a mechanic procedure in which a collection of selection diagrams and experimental data is inputted, and the procedure returns a transport formula whenever one exists. The new algorithm is called $\mathbf{TR}^{mz}$ (see Fig. 4.8) and is based on previous results in transportability as well as identifiability (TP02; SP06c).

The *main idea* of the algorithm is to leverage the c-component factorization (Tia02) and recursively decompose the target relation into manageable pieces (line 4), so as to try to solve each of them separately. Whenever this standard evaluation fails in the target domain π^* (line 6), $\mathbf{TR}^{mz}$ tries to use the experimental information available from the target and source domains (line 10), which essentially implements the declarative condition delineated in Theorem 15. Our ultimate goal is to understand what happens when the algorithm returns an expression and also when it exits with failure.

PROCEDURE $\mathbf{TR}^{mz}$ $(y, x, \mathcal{P}, \mathcal{I}, \mathcal{S}, \mathcal{W}, D)$
INPUT: x, y: value assignments; $\mathcal{P}$: local distribution relative to domain $\mathcal{S}$ ($\mathcal{S} = 0$ indexes π^*) and active experiments $\mathcal{I}$; $\mathcal{W}$: weighting scheme; D: backbone of selection diagram; S_i: selection nodes in π_i ($S_0 = \emptyset$ relative to π^*); [The following set and distributions are globally defined: $Z_i, P^*, P^{(i)}_{Z_i}$.]
OUTPUT: $P^*_x(y)$ in terms of $P^*, P^*_Z, P^{(i)}_{Z_i}$ or $FAIL(D, C_0)$.

1 **if** $x = \emptyset$, **return** $\sum_{V \setminus Y} \mathcal{P}$.
2 **if** $V \setminus An(Y)_D \neq \emptyset$, **return** $TR^{mz}(y, x \cap An(Y)_D, \sum_{V \setminus An(Y)_D} \mathcal{P}, \mathcal{I}, \mathcal{S}, \mathcal{W}, D_{An(Y)})$.
3 **set** $W = (V \setminus X) \setminus An(Y)_{D_{\overline{X}}}$.
if $W \neq \emptyset$, **return** $TR^{mz}(y, x \cup w, \mathcal{P}, \mathcal{I}, \mathcal{S}, \mathcal{W}, D)$.
4 **if** $\mathcal{C}(D \setminus X) = \{C_0, C_1, ..., C_k\}$, **return** $\sum_{V \setminus \{Y,X\}} \prod_i TR^{mz}(c_i, v \setminus c_i, \mathcal{P}, \mathcal{I}, \mathcal{S}, \mathcal{W}, D)$.
5 **if** $\mathcal{C}(D \setminus X) = \{C_0\}$,
6 **if** $\mathcal{C}(D) \neq \{D\}$,
7 **if** $C_0 \in \mathcal{C}(D)$, **return** $\prod_{i|V_i \in C_0} \sum_{V \setminus V_D^{(i)}} \mathcal{P} / \sum_{V \setminus V_D^{(i-1)}} \mathcal{P}$.
8 **if** $(\exists C')C_0 \subset C' \in \mathcal{C}(D)$,
for $\{i|V_i \in C'\}$, **set** $\kappa_i = \kappa_i \cup v_D^{(i-1)} \setminus C'$.
return $TR^{mz}(y, x \cap C', \prod_{i|V_i \in C'} \mathcal{P}(V_i|V_D^{(i-1)} \cap C', \kappa_i), \mathcal{I}, \mathcal{S}, \mathcal{W}, C')$.
9 **else**,
10 **if** $\mathcal{I} = \emptyset$, **for** $i = 0, ..., |\mathcal{D}|$,
if $\left((S_i \perp\!\!\!\perp Y \mid X)_{D^{(i)}_{\overline{X}}} \wedge (Z_i \cap X \neq \emptyset)\right)$, $E_i = TR^{mz}(y, x \setminus z_i, \mathcal{P}, Z_i \cap X, i, \mathcal{W}, D \setminus \{Z_i \cap X\})$.
11 **if** $|E| > 0$, **return** $\sum_{i=1}^{|E|} w_i^{(j)} E_i$.
12 **else**, **FAIL**(D, C_0).

Figure 4.8: Algorithm capable of recognizing mz-transportable relations.

Before showing the more formal properties of the algorithm, we demonstrate how $\mathbf{TR}^{mz}$ works through the transportability of $Q = P^*(y|do(x))$ in Fig. 4.9(a,b) with $Z^* = \{Z_1\}$, $Z_a = \{Z_2\}$, and $Z_b = \{Z_1\}$.

Since $(V \setminus X) \setminus An(Y)_{D_{\overline{X}}} = \{Z_2\}$, $\mathbf{TR}^{mz}$ invokes line 3 with $\{Z_2\} \cup \{X\}$ as interventional set. The new call triggers line 4 and $\mathcal{C}(D \setminus \{X, Z_2\}) = \{C_0, C_1, C_2, C_3\}$, where $C_0 = D_{Z_1}$, $C_1 = D_{Z_3}$, $C_2 = D_U$, and $C_3 = D_{W,Y}$, we invoke line 4 and try to mz-transport individually $Q_0 = P^*_{x,z_2,z_3,u,w,y}(z_1)$, $Q_1 = P^*_{x,z_1,z_2,u,w,y}(z_3)$, $Q_2 = P^*_{x,z_1,z_2,z_3,w,y}(u)$, and $Q_3 = P^*_{x,z_1,z_2,z_3,u}(w, y)$. Thus the original problem reduces to try to evaluate the equivalent expression $\sum_{z_1,z_3,u,w} P^*_{x,z_2,z_3,u,w,y}(z_1) P^*_{x,z_1,z_2,u,w,y}(z_3)$ $P^*_{x,z_1,z_2,z_3,w,y}(u) P^*_{x,z_1,z_2,z_3,u}(w, y)$.

First, $\mathbf{TR}^{mz}$ evaluates the expression Q_0 and triggers line 2, noting that all nodes can be ignored since they are not ancestors of $\{Z_1\}$, which implies after line 1 that $P^*_{x,z_2,z_3,u,w,y}(z_1) = P^*(z_1)$.

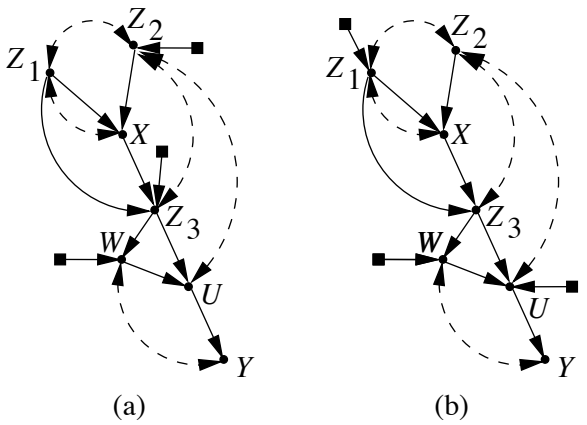


Figure 4.9: Selection diagrams illustrating how the procedure $\mathbf{TR}^{mz}$ works when transporting $R = P^*(y|do(x))$ with experiments $Z_a = \{Z_2\}$, $Z_b = Z^* = \{Z_1\}$.

Second, $\mathbf{TR}^{mz}$ evaluates the expression Q_1 triggering line 2, which implies that $P^*_{x,z_1,z_2,u,w,y}(z_3) = P^*_{x,z_1,z_2}(z_3)$ with induced subgraph $D_1 = D_{X,Z_1,Z_2,Z_3}$. TR^{mz} goes to line 5, in which in the local call $\mathcal{C}(D \setminus \{X, Z_1, Z_2\}) = \{D_{Z_3}\}$. Thus it proceeds to line 6 testing whether $\mathcal{C}(D \setminus \{X, Z_1, Z_2\})$ is different from D_1, which is false. In this call, ordinary identifiability would fail, but TR^{mz} proceeds to line 9. The goal of this line is to test whether some experiment can help for computing Q_1. In this case, π_a fails immediately the test in line 10, but π_b and π^* succeed, which means experiments in these domains may eventually help; the new call is $P^{(i)}_{x,z_2}(z_3)_{D\setminus Z_1}$, for $i = \{b, *\}$ with induced graph $D'_1 = D_{X,Z_2,Z_3}$. Finally, TR^{mz} triggers line 8 since X is not part of Z_3's components in D'_1 (or, $Z_3 \in C' = \{Z_2 \dashleftarrow\dashrightarrow Z_3\}$), so line 2 is triggered since Z_2 is no longer an ancestor of Z_3 in D'_1, and then line 1 is triggered since the interventional set is empty in this local call, so $P^*_{x,z_1,z_2}(z_3) = \sum_{Z'_2} P^{(i)}_{z_1}(z_3|x, Z'_2)P^{(i)}_{z_1}(Z'_2)$, for $i = \{b, *\}$.

Third, evaluating the expression Q_2, TR^{mz} goes to line 2, which implies that $P^*_{x,z_1,z_2,z_3,w,y}(u) = P^*_{x,z_1,z_2,z_3,w}(u)$ with induced subgraph $D_2 = D_{X,Z_1,Z_2,Z_3,W,U}$. $\mathbf{TR}^{mz}$ goes to line 5, and in this local call $\mathcal{C}(D \setminus \{X, Z_1, Z_2, Z_3, W\}) = \{D_U\}$, and the test in 6 succeed, since there are more components in D. So, it triggers line 8 since W is not part of U's component in D_2. The algorithm makes $P^*_{x,z_1,z_2,z_3,w}(u) = P^*_{x,z_1,z_2,z_3}(u)_{D_2|W}$ (and update the working distribution); note that in this call, ordinary identifiability would fail since the nodes are in the same C-component and the test in line 6 fails. But $\mathbf{TR}^{mz}$ proceeds to line 9 trying to find experiments that can help in Q_2's computation. In this case, π_b cannot help but π_a and π^* perhaps can, noting that new calls are launched for computing $P^{(a)}_{x,z_1,z_3}(u)_{D_2\setminus Z_2|W}$ relative to π_a, and $P^*_{x,z_2,z_3}(u)_{D_2\setminus Z_1|W}$ relative to π^* with the corresponding data structures set. In π_a, the algorithm triggers line 7, which

yields $P^{(a)}_{x,z_1,z_3}(u)_{D_2\setminus Z_2|W} = P^{(a)}_{z_2}(u|w,z_3,x,z_1)$, and a bit more involved analysis for π_b yields (after simplification) $P^*_{x,z_2,z_3}(u)_{D_2\setminus Z_1|W} = \big(\sum_{Z'_2} P^*_{z_1}(u|w,z_3,x,Z'_2)$ $P^*_{z_1}(z_3|x,Z'_2)P^*_{z_1}(Z'_2)\big)/\big(\sum_{Z''_2} P^*_{z_1}(z_3|x,Z''_2)P^*_{z_1}(Z''_2)\big)$.

Fourth, $\mathbf{TR}^{mz}$ finally evaluates the expression Q_3 and triggers line 5, $\mathcal{C}(D\setminus\{X,Z_1,Z_2,Z_3,U\}) = D_{W,Y}$. In turn, both tests at lines 6 and 7 succeed, which makes the procedure to return $P^*_{x,z_1,z_2,z_3,u}(w,y) = P^*(w|z_3,x,z_1,z_2)P^*(y|w,x,z_1,z_2,z_3,u)$.

The composition of the return of these calls generates the following expression:

$$
\begin{aligned}
P^*_x(y) \;=\; & \sum_{z_1,z_3,w,u} P^*(z_1)\left(w_1^{(1)}\sum_{Z'_2} P^*_{z_1}(z_3|x,Z'_2)P^*_{z_1}(Z'_2) + w_2^{(1)}\sum_{Z'_2} P^{(b)}_{z_1}(z_3|x,Z'_2)P^{(b)}_{z_1}(Z'_2)\right) \\
& \left(w_1^{(2)}\Big(\sum_{Z'_2} P^*_{z_1}(u|w,z_3,x,Z'_2)P^*_{z_1}(z_3|x,Z'_2)P^*_{z_1}(Z'_2)\Big)/\Big(\sum_{Z''_2} P^*_{z_1}(z_3|x,Z''_2)P^*_{z_1}(Z''_2)\Big)\right. \\
& \left. +\; w_2^{(2)}P^{(a)}_{z_2}(u|w,z_3,x,z_1)\right) P^*(w|x,z_1,z_2,z_3)\; P^*(y|x,z_1,z_2,z_3,w,u) \qquad (4.31)
\end{aligned}
$$

where $w_i^{(k)}$ represents the weight for each factor in estimand k ($i = 1,...,n_k$), and n_k is the number of feasible estimands of k.

Remarkably, the derived transport formula depicts a powerful way to estimate $P^*(y|do(x))$ in the target domain, and depending on the weighting scheme a different estimand will be entailed. For instance, one might use an analogous to *inverse-variance weighting*, which sets the weights for the normalized inverse of their variances (*i.e.*, $w_i^{(k)} = \sigma_i^{-2}/\sum_{j=1}^{n_k}\sigma_j^{-2}$, where σ_j^2 is the variance of the jth component of estimand k). Our strategy resembles the approach taken in meta-analysis (HO85), albeit the latter usually disregards the intricacies of the relationships between variables, so producing a statistically less powerful estimand. Our method leverages this non-trivial and highly structured relationships, as exemplified in Eq. (4.31), which allows one to obtain an estimand with less variance and statistically more powerful than standard ones.

Finally, we can formally consider the soundness of the algorithm.

Theorem 18 (soundness)**.** *Whenever* $\boldsymbol{TR}^{mz}$ *returns an expression for* $P^*_x(y)$*, it is correct.*

Proof. See Appendix B. □

Theorem 19. *Assume* $\boldsymbol{TR}^{mz}$ *fails to transport the effect* $P^*_x(y)$ *(exits with failure executing line 12). Then there exists* $X' \subseteq X$*,* $Y' \subseteq Y$*, such that the graph pair* D, C_0 *returned by the fail condition of* TR^{mz} *contain as edge subgraphs sC-forests F, F' that spans a mz*-shedge for* $P^*_{x'}(y')$*.*

Proof. Let D be the subgraph local to the call in which $\mathbf{TR}^{mz}$ failed, and R be the root set of D. It is possible to remove some directed arrows from D while preserving R as root, which result in a R-rooted c-forest F. Since by construction $F' = F \cap C_0$ is closed under descendents and only directed arrows were removed, both F, F' are c-forests. Also by construction $R \subset An(Y)_{G_{\overline{X}}}$ together with the fact that X and Y from the recursive call are clearly subsets of the original input. Before failure, $\mathbf{TR}^{mz}$ evaluated false consecutively at lines 6, 10, and 11, and it is not difficult to see that an S-node points to F' or the respective experiments were not able to break the local hedge (lines 10 and 11). It remains to be shows that this mz-shedge can be stretched to generate a mz^*-shedge, but now the same construction given in Thm. 16 can be applied (see also Appendix B). □

In the sequel, we state the completeness of the algorithm and graphical condition discussed in this section.

Corollary 8 (completeness)**.** ***TR***^{mz} *is complete.*

Proof. See Appendix B. □

Corollary 9 (mz^*-shedge characterization)**.** $P^*_x(y)$ *is* mz*-transportable from* Π *to* π^* *in* $\mathcal{D}$ *if and only if there is not* mz^**-shedge for* $P_{x'}(y')$ *in* $\mathcal{D}$ *for any* $X' \subseteq X$ *and* $Y' \subseteq Y$*.*

Proof. See Appendix B. □

Furthermore, we show below that the do-calculus is complete for establishing mz-transportability, which means that failure in the exhaustive application of its rules implies the non-existence of a mapping from the available data to the target relation (i.e., there is no mz-transport formula), independently of the method used to obtain such mapping.

Corollary 10 (do-calculus characterization)**.** *The rules of do-calculus together with standard probability manipulations are complete for establishing* mz*-transportability of all causal effects.*

Proof. See Appendix B. □

4.6 Conclusions

In this chapter, we introduced the most general variant of transportability known to date in which experiments can be conducted over limited sets of variables in

the various domains under study (source and target), and the goal is to (non-parametrically) infer whether a certain effect can be estimated in the target using the information gathered through randomized experiments in all the domains. Using the language of selection diagrams, the work developed in this chapter puts in mathematical language, solves, and generalizes the problems of external validity, meta-analysis, and fusion of causal knowledge that have been only semi-formally discussed in the literature for at least half a century.

We provided a complete characterization for deciding transportability in the form of graphical, algorithmic, and algebraic conditions. Specifically, we derived a general graphical condition for deciding transportability of causal effects, which means that transportability is feasible if and only if a certain graph structure does not appear as an edge subgraph of the inputted collection of selection diagrams.

Furthermore, we constructed a procedure for deciding transportability, that is, generate a formula for fusing the available observational and experimental data to synthesize an estimate of the desired causal effects. We showed that this procedure is complete, which means that the set of transportable instances identified by the algorithm cannot be broadened without strengthening the assumptions. Our algorithm also allows for generic weighting schemes, which generalizes standard statistical procedures and leads to the construction of statistically more powerful estimands. We further showed that the do-calculus is complete for this class of problems, which means that finding a proof strategy in this language suffices to solve the problem.

While practical applications of these results are predicated on the availability of problem-specific selection diagrams, the general understanding of why some problems permit information transfer and others do not have scientific merit on its own. It informs investigators what kind of disparities between environments would make transportability theoretically impossible, and what disparities can be circumvented by clever information fusion strategies. Even though the construction of a selection diagram might be a demanding task, the completeness result makes such construction unavoidable if one seeks theoretical guarantees for a given method of information transfer. Fortunately, the knowledge necessary to construct a diagram is not much different than that required for ordinary causal diagrams as used, for example, to establish internal validity (i.e., identifiability).

The non-parametric characterization established in this chapter gives rise to a new set of research questions. While our analysis aimed at achieving consistent transport under asymptotic conditions, when no transport formula exists, approximation techniques must be resorted to, for example, replacing the requirement of non-parametric analysis with assumptions about linearity or monotonicity of certain relationships in the domains. The nonparametric characterization provided in this chapter should serve as a guideline for such approximation schemes.

CHAPTER 5

Controlling Selection Bias in Causal and Statistical Inference

5.1 Introduction

Selection bias is caused by preferential exclusion of units from the samples and represents a major obstacle to valid causal and statistical inferences; it cannot be removed by randomized experiments and can rarely be detected in either experimental or observational studies. This chapter provides methods to help to understand, formalize, solve or alleviate this problem in a broad range of data-intensive applications that appears in many disciplines.

The chapter is organized as follows. In section 5.2, we discuss the distinction between selection and confounding biases, which are the most common biases encountered in the literature. We then put our work in perspective and elaborate on the different types of assumptions behind the current methods aiming to solve the selection problem, namely, qualitative assumptions about the selection mechanism, parametric assumptions regarding the data-generating model, and quantitative assumptions about the selection process. In section 5.3, we provide complete graphical and algorithmic conditions for recovering conditional probabilities from selection biased data. In section 5.4, we relax the previous assumptions and provide graphical conditions for recoverability when unbiased data is also available for use (e.g., census data). In section 5.5, we provide a graphical condition that generalizes the backdoor criterion and serves to recover causal effects when the data is collected under preferential selection. In section 5.6, we relax the assumptions and introduce conditions for recoverability when the target quantity is the odd ratio. We provide a complete graphical and algorithmic criteria for recoverability of the population and conditions odds ratio. We relate the odds ratio with other common measures such as risk differences and risk ratios. In section 5.7, we consider the challenging task of recovering from selection when confounding bias is simultaneously present. Given that the bounds obtained from the previous analysis is biased given the selection process, we provide a graphical criterion for recoverability under general conditions.

5.2 The Structure of the Selection Problem

Selection bias is induced by preferential selection of units for data analysis, usually governed by unknown factors including treatment, outcome, and their consequences, and represents a major obstacle to valid causal and statistical inferences. It cannot be removed by randomized experiments and can rarely be detected in either experimental or observational studies.[1] For instance, in a typical study of the effect of training program on earnings, subjects achieving higher incomes tend to report their earnings more frequently than those who earn less. The data-gathering process in this case will reflect this distortion in the sample proportions and, since the sample is no longer a faithful representation of the population, biased estimates will be produced regardless of how many samples were collected.

This preferential selection challenges the validity of inferences in several tasks in AI (Coo95; Elk01; Zad04; CMR08) and Statistics (Whi78; LR86; Jew91; KC06) as well as in the empirical sciences (e.g., Genetics (PDS12; MW12), Economics (Hec79; Ang97), and Epidemiology (Rob01; GG08)).

To illuminate the nature of preferential selection, consider the data-generating model in Fig. 5.1(a) in which X represents an action, Y represents an outcome, and S represents a binary indicator of entry into the data pool ($S = 1$ means that the unit is in the sample, $S = 0$ otherwise). If our goal is to compute the population-level conditional distribution $P(y|x)$, and the samples available are collected under selection, only $P(y, x|S = 1)$ is accessible for use. [2] Given that in principle these two distributions are just loosely connected, the natural question to ask is under what conditions $P(y|x)$ can be recovered from data coming from $P(y, x|S = 1)$. In this specific example, both action and outcome affect the entry in the data pool, which will be shown not to be recoverable (see Corollary 11) – i.e., there is no method capable of unbiasedly estimating the population-level distribution using data gathered under this selection process.

The bias arising from selection differs fundamentally from the one due to *confounding*, though both constitute threats to the validity of causal inferences. The former bias is due to treatment or outcome (or ancestors) affecting the inclusion of the subject in the sample (Fig. 5.1(a)), while the latter is the result of treatment X and outcome Y being affected by a common omitted variables U (Fig. 5.1(b)). In both cases, we have unblocked extraneous "flow" of information between treatment and outcome, which appear under the rubric of "spurious correlation," since it is not what we seek to estimate.

[1] Remarkably, there are special situations in which selection bias can be detected even from observations, as in the form of a non-chordal undirected component (Zha08).

[2] In a typical AI task such as classification, we could have X being a collection of features and Y the class to be predicted, and $P(y|x)$ would be the classifier that needs to be trained.

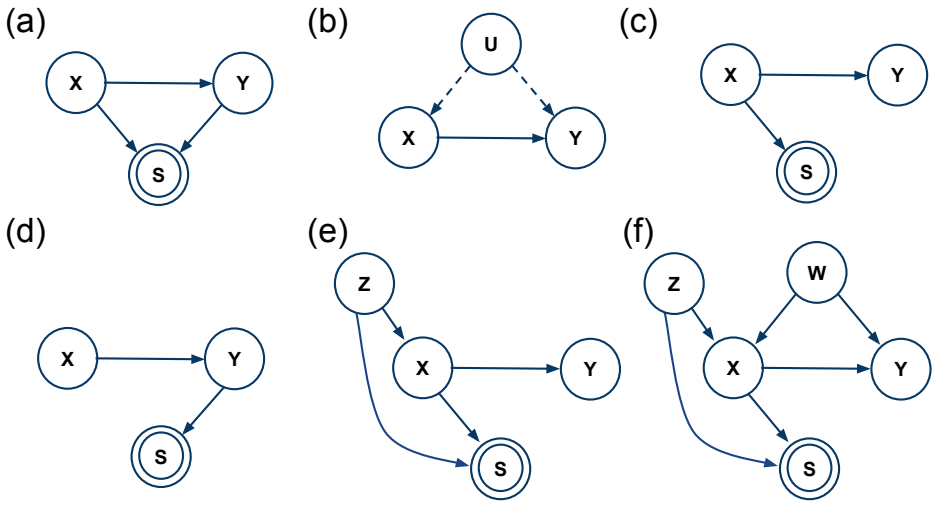


Figure 5.1: (a,b) Simplest examples of selection and confounding bias, respectively. (c,d) Treatment-dependent and outcome-dependent studies under selection, $Q = P(y|x)$ is recoverable in (c) but not in (d). (e,f) Treatment-dependent study where selection is also affected by driver of treatment Z (e.g., age); Q is recoverable in (e) but not in (f).

It is instructive to understand selection graphically, as in Fig. 5.1(a). The preferential selection that is encoded through conditioning on S creates spurious association between X and Y through two mechanisms. First, given that S is a collider, conditioning on it induces spurious association between its parents, X and Y (Pea88). Second, S is also a descendant of a "virtual collider" Y, whose parents are X and the error term U_Y (also called "hidden variable") which is always present, though often not shown in the diagram.[3]

5.2.1 Related work and Our contributions

There are three sets of assumptions that are enlightening to acknowledge if we want to understand the procedures available in the literature for treating selection bias – qualitative assumptions about the selection mechanism, parametric assumptions regarding the data-generating model, and quantitative assumptions about the selection process.

In the data-generating model in Fig. 5.1(c), the selection of units to the sample is treatment-dependent, which means that it is caused by X, but not Y. This case has been studied in the literature and $Q = P(y|x)$ is known to be

[3]See (Pea00, pp. 339-341) and (Pea13) for further explanations of this bias mechanism.

non-parametrically recoverable from selection (GP11). Alternatively, in the data-generating model in Fig. 1(d), the selection is caused by Y (outcome-dependent), and Q is not recoverable from selection (formally shown later on), but is the odds ratio (Cor51; Whi78; Gen92; DKK10). As mentioned earlier, Q is also not recoverable in the graph in Fig. 5.1(a). By and large, the literature is concerned with treatment-dependent or outcome-dependent selection, but selection might be caused by multiple reasons and embedded in more intricate realities. For instance, a driver of the treatment Z (e.g., age, sex, socio-economic status) may also be causing selection, see Fig. 5.1(e,f). As it turns out, Q is recoverable in Fig 5.1(e) but not in (f), so different qualitative assumptions need to be modelled explicitly since each topology entails a different answer for recoverability.

The second assumption is related to the parametric form used by recoverability procedures. For instance, one variation of the selection problem was studied in Econometrics, and led to the celebrated method developed by James Heckman (Hec79). His two-step procedure removes the bias by leveraging the assumptions of linearity and normality of the data-generating model. A graph-based parametric analysis of selection bias is given in (Pea13).

The final assumption is about the probability of being selected into the sample. In many settings in Machine learning and Statistics (Elk01; Zad04; SE07; Sto09; Hei09; CMR08), it is assumed that this probability, $P(S = 1|Pa_s)$, can be modelled explicitly, which often is an unattainable requirement for the practitioner (e.g., it might be infeasible to assess the differential rates of how salaries are reported).

Our treatment differs fundamentally from the current literature regarding these assumptions. First, we do not constrain the type of data-generating model as outcome- or treatment-dependent, but we take arbitrary models (including these two) as input, in which a node S indicates selection for sampling. Second, we do not make parametric assumptions (e.g. linearity, normality, monotonicity) but operate non-parametrically based on causal graphical models (Pea00), which is more robust, less prone to model misspecifications. Third, we do not rely on having the selection's probability $P(S = 1|Pa_s)$, which is not always available in practice. Our work hinges on exploiting the qualitative knowledge encoded in the data-generating model for yielding recoverability. This knowledge is admittedly a demanding requirement for the scientist, but we now understand formally its necessity for *any* approach to recoverability – any procedure aiming for recoverability, implicitly or explicitly, relies on this knowledge. [4]

The analysis of selection bias requires a formal language within which the notion of data-generating model is given precise characterization, and the qual-

[4] A trivial instance of this necessity is Fig. 5.1(c,d) where the odds ratio is recoverable, yet $P(y|x)$ is recoverable in 5.1(c) but not in (d).

itative assumptions regarding how the variables affect selection can be encoded explicitly. The advent of causal diagrams (Pea95; SGS00; Pea00; KF09) provides such a language and renders the formalization of the selection problem possible.

Using this language, we address the issue of recovering from selection bias under different assumptions encountered in the literature:

1. **Selection without external data (Sec. 5.3):** The dataset is collected under selection bias, $P(v|S = 1)$; under which conditions is $P(y|x)$ recoverable?

2. **Selection with external data (Sec. 5.4):** The dataset is collected under selection bias, $P(v|S = 1)$, but there are unbiased samples from $P(t)$, for $T \subseteq V$; under which conditions is $P(y|x)$ recoverable?

3. **Selection in causal inferences (Sec. 5.5):** The dataset is collected under selection bias, $P(v|S = 1)$, but there are unbiased samples from $P(t)$, for $T \subseteq V$; under which conditions is the interventional distribution $P(y|do(x))$ estimable?

4. **Selection of the odds ratio (Sec. 5.6):** The dataset is collected under selection bias, $P(v|S = 1)$; under which conditions is $OR(X, Y|C)$ recoverable, for some set C?

5. **Selection with confounding (Sec. 5.7):** The dataset is collected under selection bias, $P(v|S = 1)$; under which conditions is the bounds for the interventional distribution $P(y|do(x))$ recoverable from selection?

5.3 Recoverability without External Data

To address the selection problem formally, we need to model the selection mechanism explicitly, so we add a variable S to represent this mechanism, and assume that $S = 1$ represents presence in the sample, and zero otherwise. In this augmented representation, independence of a certain variable from S encodes the assumption that entry to the data pool is not affected by this variable (possibly conditioned in a third set). A similar representation was used in (Coo95; LR08; GRB09; DKK10). We denote the set of all variables by V except for the selection mechanism S.

We next introduce the formal notion of recoverability for conditional distributions when the data is collected under selection bias.

Definition 19 (s-Recoverability)**.** *Given a causal graph G_s augmented with a node S encoding the selection mechanism, the distribution $Q = P(y \mid x)$ is said*

to be s-recoverable from selection biased data in G_s if the assumptions embedded in the causal model renders Q expressible in terms of the distribution under selection bias $P(v \mid S = 1)$. Formally, for every two probability distributions P_1 and P_2 compatible with G_s, $P_1(v \mid S = 1) = P_2(v \mid S = 1) > 0$ implies $P_1(y \mid x) = P_2(y \mid x)$.

Consider the graph G_s in Fig. 5.1(c) and assume that our goal is to establish s-recoverability of $Q = P(y|x)$. Note that by d-separation (Pea88), X separates Y from S, (or $(Y \perp\!\!\!\perp S|X)$), so we can write $P(y|x) = P(y|x, S = 1)$. This is a very special situation since these two distributions can be arbitrarily distant from each other, but in this specific case G_s constrains Q in such a way that despite the fact that data was collected under selection and our goal is to answer a query about the overall population, there is no need to resort to additional data external to the biased study.

Now we want to establish whether Q is s-recoverable in the graph G_s in Fig. 5.1(d). In this case, S is not d-separated from Y if we condition on X, so $(S \perp\!\!\!\perp Y|X)$ does not hold in at least one distribution compatible with G_s, and the identity $P(y|x) = P(y|x, S = 1)$ is not true in general. One may wonder if there is another way to s-recover Q in G_s, but this is not the case as formally shown next. That is, the assumptions encoded in G_s imply a universal impossibility; no matter how many samples of $P(x, y|S = 1)$ are accumulated or how sophisticated the estimation technique is, the estimator of $P(y|x)$ will never converge to its true value.

Lemma 4. *$P(y|x)$ is not s-recoverable in Fig. 5.1(d).*

Proof. We construct two causal models such that P_1 is compatible with the graph G_s in Fig. 5.1(d) and P_2 with the subgraph $G_2 = G_s \setminus \{Y \rightarrow S\}$. We will set the parameters of P_1 through its factors and then computing the parameters of P_2 by enforcing $P_2(V \mid S = 1) = P_1(V \mid S = 1)$. Since $P_2(V|S = 1) = P_2(V)$, we will be enforcing $P_1(V|S = 1) = P_2(V)$. Recoverability should hold for any parametrization, so we assume that all variables are binary. Given a Markovian causal model (Pea00), P_1 can be parametrized through its factors in the decomposition over observables, $P_1(X), P_1(Y|X), P_1(S = 1|Y)$, for all X, Y.

We can write the conditional distribution in the second causal model as follows:

$$P_2(y|x) = P_1(y|x, S = 1) = \frac{P_1(y, x, S = 1)}{P_1(x, S = 1)} \tag{5.1}$$

$$= \frac{P_1(S = 1|y)P_1(y|x)}{P_1(S = 1|y)P_1(y|x) + P_1(S = 1|\overline{y})P_1(\overline{y}|x)}, \tag{5.2}$$

where the first equality, by construction, should be enforced, and the second and third by probability axioms. The other parameters of P_2 are free and can be chosen to match P_1.

Finally, set the distribution of every family in P_1 but selection variable equal to $1/2$, and set the distribution $P_1(S = 1|y) = \alpha, P_1(S = 1|\overline{y}) = \beta$, for $0 < \alpha, \beta < 1$ and $\alpha \neq \beta$. This parametrization reduces eq. (2) to $P_2(y|x) = \alpha/(\alpha + \beta)$ and $P_1(y|x) = 1/2$, the result follows. □

Corollary 11. *$P(y|x)$ is not s-recoverable in Fig. 5.1(a).*

The corollary follows immediately noting that lack of s-recoverability with a subgraph (Fig. 5.1(d)) precludes s-recoverability with the graph itself since the extra edge can be inactive in a compatible parametrization (Pea88) (the converse is obviously not true). Lemma 4 is significant because Fig. 5.1(d) can represent a study design that is typically used in empirical fields known as case-control studies. The result is also theoretically instructive since Fig. 5.1(d) represents the smallest graph structure that is not s-recoverable, and its proof will set the tone for more general and arbitrary structures that we will be interested in (see Theorem 20).

Furthermore, consider the graph in Fig. 5.1(e) in which the independence $(S \perp\!\!\!\perp Y|X)$ holds, so we can also recover Q from selection $(P(y|x, S = 1) = P(y|x))$. However, $(S \perp\!\!\!\perp Y|X)$ does not hold in Fig. 5.1(f) – there is an open path passing through X's ancestor W (i.e. $S \leftarrow Z \rightarrow X \leftarrow W \rightarrow Y$) – and the natural question that arises is whether Q is recoverable in this case. It does not look obvious whether the absence of an independence precludes s-recoverability since there are other possible operators in probability theory that could be used leading to the s-recoverability of Q. To illustrate this point, note that it is not the case in causal inference that the inapplicability of the backdoor criterion (Pea00, Ch. 3), which is also an independence constraint, implies the impossibility of recovering certain effects.

Remarkably, the next result states that the lack of this independence indeed precludes s-recoverability, i.e., the probe of one separation test in the graph is sufficient to evaluate whether a distribution is or is not s-recoverable.

Theorem 20. *The distribution $P(y|x)$ is s-recoverable from G_s if and only if $(S \perp\!\!\!\perp Y|X)$.*

Proof. See Appendix C. □

In words, Theorem 20 provides a powerful test for s-recoverability without external data, which means that when it disavows s-recoverability, there exists no

procedure that would be capable of recovering the distribution from selection bias (without adding assumptions). Its sufficiency part is immediate, but the proof of necessity is somewhat involved since we need to show that for *all* graphical structures in which the given d-separation test fails, each of these structures does not allow for s-recoverability (i.e., a counter-example can always be produced showing agreement on $P(v|S=1)$ and disagreement on $P(y|x)$).

The next corollary provides a test for s-recoverability of broader joint distributions (including Y alone):

Corollary 12. *Let $Z = An(S) \setminus An(Y)$ including S, and $A = Pa(Z) \cap (An(Y) \setminus \{Y\})$. $P(Y, An(Y) \setminus (A \setminus \{Y\})|A)$ is s-recoverable if and only if Y is not an ancestor of S.*

This result can be embedded as a step reduction in an algorithm to s-recover a collection of distributions in the form of the corollary. We show such algorithm in Appendix C. [5]. The main idea is to traverse the graph in a certain order s-recovering all joint distributions with the form given in the corollary (updating S along the way). If the algorithm exits with failure, it means that the distributions of its predecessors are not s-recoverable.

5.4 Recoverability with External Data

A natural question that arises is whether additional measurements in the population level over certain variables can help recovering a given distribution. For example, $P(age)$ can be estimated from census data which is not under selection bias.

To illustrate how this problem may arise in practice, consider Fig. 5.2 and assume that our goal is to s-recover $Q = P(y|x)$. It follows immediately from Theorem 20 that Q cannot be s-recovered without additional assumptions. Note, however, that the parents of the selection node $Pa_s = \{W_1, W_2\}$ separates S from all other nodes in the graph, which indicates that it would be sufficient for recoverability to measure $T = \{W_1, W_2\} \cup \{X\}$ from external sources. To witness, note that after conditioning Q on W_1 and W_2, we obtain:

$$\begin{aligned} P(y|x) &= \sum_{w_1,w_2} P(y|x,w_1,w_2)P(w_1,w_2|x) \\ &= \sum_{w_1,w_2} P(y|x,w_1,w_2,S=1)P(w_1,w_2|x), \end{aligned} \tag{5.3}$$

[5] This listing is useful when one needs to examine properties of the collection of distributions, analogously to the list of all backdoor admissible sets by (TL11).

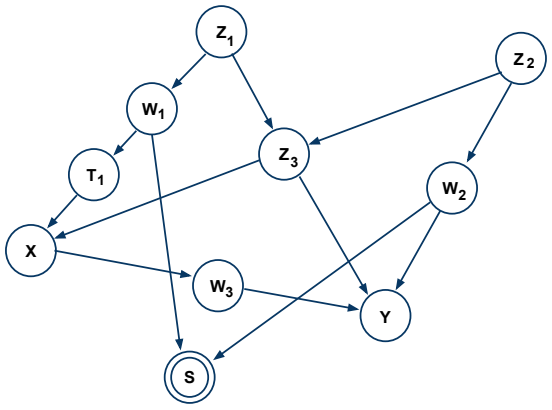


Figure 5.2: Causal model in which $Q = P(y|x)$ is not recoverable without external data (Thm. 20), but it is recoverable if measurements on the set $Pa_s = \{W_1, W_2\}$ are taken (Thm. 21). Alternatively, even if not all parents of S are measured, any set including $\{W_2, Z_3\}$ would yield recoverability of Q.

where the last equality follows from $(Y \perp\!\!\!\perp S \mid X, W_1, W_2)$. That is, Q can be s-recovered and is a combination of two different types of data; the first factor comes from biased data under selection, and the second factor is available from external data collected over the whole population.

Our goal is to understand the interplay between measurements taken over two types of variables, $M, T \subseteq V$, where M are variables collected under selection bias, $P(M|S = 1)$, and T are variables collected in the population-level, $P(T)$. In other words, we want to understand when (and how) can this new piece of evidence $P(T)$ together with the data under selection ($P(M|S = 1)$) help in extending the treatment of the previous section for recovering the true underlying distribution $Q = P(y|x)$.[6]

Formally, we need to redefine s-recoverability for accommodating the availability of data from external sources.

Definition 20 (s-Recoverability)**.** *Given a causal graph G_s augmented with a node S, the distribution $Q = P(y \mid x)$ is said to be s-recoverable from selection bias in G_s with external information over $T \subseteq V$ and selection biased data over $M \subseteq V$ (for short, s-recoverable) if the assumptions embedded in the causal model render Q expressible in terms of $P(m \mid S = 1)$ and $P(t)$, both positive. Formally, for every two probability distributions P_1 and P_2 compatible with G_s, if they agree*

[6]This problem subsumes the one given in the previous section since when $T = \emptyset$, the two problems coincide. We separate them since they come in different shades in the literature and also just after solving the version without external data we can aim to solve its more general version; we discuss more about this later on.

on the available distributions, $P_1(m \mid S = 1) = P_2(m \mid S = 1) > 0$, $P_1(t) = P_2(t) > 0$, *they must agree on the query distribution,* $P_1(y \mid x) = P_2(y \mid x)$.

The observation leading to eq. (5.3) provides a simple condition for s-recoverability when we can choose the variables to be collected. Let Pa_s be the parent set of S. If measurements on the set $T = Pa_s \cup \{X\}$ can be taken without selection, we can write $P(y|x) = \sum_{pa_s} P(y|x, pa_s, S = 1)P(pa_s|x)$, since S is separated from all nodes in the graph given its parent set. This implies s-recoverability where we have a mixture in which the first factor is obtainable from the biased data and the second from external sources.

This solution is predicated on the assumption that Pa_s can be measured in the overall population, which can be a strong requirement, and begs a generalization to when part of Pa_s is not measured. For instance, what if in Fig. 5.2 W_1 cannot be measured? Would other measurements over a different set of variables also entail s-recoverability?

This can be expressed as a requirement that subsets of T and M can be found satisfying the following criterion:

Theorem 21. *If there is a set* C *that is measured in the biased study with* $\{X, Y\}$ *and in the population level with* X *such that* $(Y \perp\!\!\!\perp S|\{C, X\})$, *then* $P(y|x)$ *is* s*-recoverable as*

$$P(y|x) = \sum_c P(y|x, c, S = 1)P(c|x). \tag{5.4}$$

Proof. See Appendix C. □

In the example in Fig. 5.2, it is trivial to confirm that any (pre-treatment) set C containing W_2 and Z_3 would satisfy the conditions of the theorem. In particular, $\{W_2, Z_3\}$ is such a set, and it allows us to s-recover Q without measuring W_1 ($W_1 \in Pa_s$) through eq. (C.11). Note, however, that the set $C = \{W_2, Z_1, Z_2\}$ is not sufficient for s-recoverability. It fails to satisfy the separability condition of the theorem since conditioning on $\{X, W_2, Z_1, Z_2\}$ leaves an unblocked path between S and Y (i.e., $S \leftarrow W_1 \rightarrow T_1 \rightarrow X \leftarrow Z_3 \rightarrow Y$).

It can be computationally difficult to find a set satisfying the conditions of the theorem since this could imply a search over a potentially exponential number of subsets. Remarkably, the next result shows that the existence of such a set can be determined by a single d-separation test.

Theorem 22. *There exists some set* $C \subseteq T \cap M$ *such that* $Y \perp\!\!\!\perp S|\{C, X\}$ *if and only if the set* $(C' \cup X)$ *d-separates* S *from* Y *where* $C' = [(T \cap M) \cap An(Y \cup S \cup X)] \setminus (Y \cup S \cup X)$.

Proof. See Appendix C. □

In practice, we can restrict ourselves to minimal separators, that is, looking only for minimal set $C \subseteq T \cap M$ such that $(Y \perp\!\!\!\perp S | \{C, X\})$. The algorithm for finding minimal separators has been given in (AC96; TPP98).

Despite the computational advantages given by Thm. 22, Thm. 21 still requires the existence of a separator C measured in both the biased study (M) and in the overall population (T), and it is natural to ask whether this condition can be relaxed. Assume that all we have is a separator $C \subseteq M$, but C (or some of its elements) is not measured in T, and therefore $P(c|x)$ in eq. (C.11) still needs to be s-recovered. We could s-recover $P(c|x)$ in the spirit of Thm. 21 as

$$P(c|x) = \sum_{c_1} P(c|x, c_1, S = 1) P(c_1|x), \tag{5.5}$$

if there exists a set $C_1 \subseteq M \cap T$ such that $(S \perp\!\!\!\perp C | X, C_1)$. Now if this fails in that we can only find a separator $C_1 \subseteq M$ not measured in T, we can then attempt to recover $P(c_1|x)$ in the spirit of Thm. 21 by looking for another separator C_2, and so on. At this point, it appears that Thm. 21 can be extended.

We further extend this idea by considering other possible probabilistic manipulations and embed them in a recursive procedure. For $W, Z \subseteq M$, consider the problem of recovering $P(w|z)$ from $P(t)$ and $P(m|S = 1)$, and define procedure $RC(w, z)$ as follows:

1. If $W \cup Z \subseteq T$, then $P(w|z)$ is s-recoverable.
2. If $(S \perp\!\!\!\perp W | Z)$, then $P(w \mid z)$ is s-recoverable as $P(w \mid z) = P(w \mid z, S = 1)$.
3. For minimal $C \subseteq M$ such that $(S \perp\!\!\!\perp W | (Z \cup C))$, $P(w|z) = \sum_c P(w|z, c, S = 1) P(c|z)$. If $C \cup Z \subseteq T$, then $P(w|z)$ is s-recoverable. Otherwise, call $RC(c, z)$.
4. For some $W' \subset W$, $P(w|z) = P(w'|w \setminus w', z) P(w \setminus w'|z)$. Call $RC(w', \{w \setminus w'\} \cup z)$ and $RC(w \setminus w', z))$.
5. Exit with FAIL (to s-recover $P(w|z)$) if for a singleton W, none of the above operations are applicable.

Now, we define recoverability based on this procedure:

Definition 21. *We say that $P(w|z)$ is C-recoverable if and only if it is recovered by the procedure $RC(w, z)$.*

Remarkably, the manipulations considered in $RC()$ are not actually more powerful than Thm. 21, as shown next.

Theorem 23. *For $X \subseteq T$, $Y \notin T$, $Q = P(y|x)$ is C-recoverable if and only if it is recoverable by Theorem 21, that is, if and only if there exists a set $C \subseteq T \cap M$ such that $(Y \perp\!\!\!\perp S|\{C, X\})$ (where C could be empty). If s-recoverable, $P(y|x)$ is given by $P(y|x) = \sum_c P(y|x, c, S = 1)P(c|x)$.*

Proof. See Appendix C. □

This result suggests that the constraint between measurement sets cannot be relaxed through ordinary decomposition and Thm. 21 captures the bulk of s-recoverable relations. Importantly, this does not constitute a proof of necessity of Thm. 21.

Now we turn our attention to some special cases that appear in practice. Note that, so far, we assumed X being measured in the overall population, but in some scenarios Y's prevalence might be available instead. So, assume $Y \in T$ but some variables in X are not measured in the population-level. Let $X^0 = X \cap T$ and $X^m = X \setminus X^0$, we have

$$P(y|x) = \frac{P(x^m|y, x^0)p(y|x^0)}{\sum_y P(x^m|y, x^0)p(y|x^0)} \tag{5.6}$$

Therefore, $P(y|x)$ is recoverable if $P(x^m|y, x^0)$ is recoverable. We could use the previous results to recover $P(x^m|y, x^0)$. In particular, Theorems 21 and 22 lead to:

Corollary 13. *$P(y|x)$ is recoverable if there exists a set $C \subseteq T \cap M$ (C could be empty) such that $(X^m \perp\!\!\!\perp S|\{C \cup Y \cup X^0\})$. If recoverable, $P(y|x)$ is given by Eq. (5.6) where*

$$P(x^m|y, x^0) = \sum_c P(x^m|y, x^0, c, S = 1)P(c|y, x^0) \tag{5.7}$$

Corollary 14. *$P(y|x)$ is recoverable via Corollary 13 if and only if the set $(C' \cup Y \cup X^0)$ d-separates S from X^m where $C' = [(T \cap M) \cap An(Y \cup S \cup X)] \setminus (Y \cup S \cup X)$.*

For example, in Fig. 5.2, assuming $M = \{X, Y, W_1, W_3, Z_3\}$ and $T = \{Y, W_1, W_3, Z_3\}$, we have $S \perp\!\!\!\perp X|\{Y, W_1, W_3, Z_3\}$, therefore we can s-recover

$$P(x|y) = \sum_{w_1, w_3, z_3} P(x|y, w_1, w_3, z_3, S = 1)P(w_1, w_3, z_3|y), \tag{5.8}$$

as well as $P(y|x)$ by substituting back eq. (5.8) in eq. (5.6).

Furthermore, it is worth examining when no data is gathered over X or Y in the population level. In this case, $P(y|x)$ may be recoverable through $P(x, y)$, as shown in the sequel.

Corollary 15. *$P(y|x)$ is recoverable if there exists a set $C \subseteq T \cap M$ such that $(\{Y\} \cup X \perp\!\!\!\perp S|C)$. If recoverable, $P(y, x)$ is given by $P(y, x) = \sum_c P(y, x|c, S = 1)P(c)$.*

For instance, $P(x, y)$ is s-recoverable in Fig. 5.2 if $T \cap M$ contains $\{W_2, T_1, Z_3\}$ or $\{W_2, T_1, Z_1\}$ (without $\{X, Y\}$) .

5.5 Recoverability of Causal Effects

We now turn our attention to the problem of estimating causal effects from selection biased data.

Our goal is to recover the effect of X on Y, $P(y|do(x))$ given the structure of G_s. Consider the graph G_s in Fig. 5.3(a), in which X and Y are not confounded, hence, $P(y|do(x)) = P(y|x)$ and, based on Theorem 20, we conclude that $P(y|do(x))$ is not recoverable in G_s. Fig. 5.3(b) and (c), on the other hand, contains covariates W_1 and W_2 that may satisfy conditions similar to those in Theorem 20 that would render $P(y|do(x))$ recoverable. These conditions, however, need to be strengthened significantly, to account for possible confounding between X and Y which, even in the absence of selection bias, might require adjustment for admissible covariates, namely, covariates that satisfy the backdoor condition (Pea93a). For example, $\{W_2\}$ satisfies the backdoor condition in both Fig. 5.3(b) and (c), while $\{W_1\}$ satisfies this condition in (b) but not in (c).

Definition 22 below extends the backdoor condition to selection bias problems by identifying a set of covariates Z that accomplishes two functions. Conditions (i) and (ii) assure us that Z is backdoor admissible (PP10) [7], while conditions (iii) and (iv) act to separate S from Y, so as to permit recoverability from selection bias.

Definition 22 (Selection-backdoor criterion)**.** *Let a set Z of variables be partitioned into $Z^+ \cup Z^-$ such that Z^+ contains all non-descendants of X and Z^- the descendants of X. Z is said to satisfy the selection backdoor criterion (s-backdoor, for short) relative to an ordered pairs of variables (X, Y) and an ordered pair of sets (M, T) in a graph G_s if Z^+ and Z^- satisfy the following conditions:*

[7] These two conditions extend the usual backdoor criterion (Pea93a) to allow descendants of X to be part of Z.

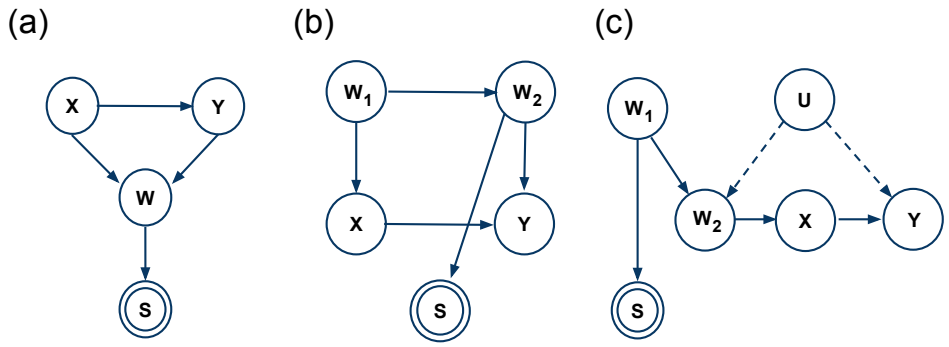


Figure 5.3: (a) Causal diagram in which $(S \perp\!\!\!\perp Y|\{X, W\})$ but $P(y|do(x))$ is not s-backdoor admissible. (b) $P(y|do(x))$ is s-recoverable through $T = \{W_2\}$ but not $\{W_1\}$. (c) $\{W_2\}$ does not satisfy the s-backdoor criterion but $P(y|do(x))$ is still recoverable.

(i) Z^+ *blocks all back door paths from* X *to* Y*;*

(ii) X *and* Z^+ *block all paths between* Z^- *and* Y*, namely,* $(Z^- \perp\!\!\!\perp Y|X, Z^+)$*;*

(iii) X *and* Z *block all paths between* S *and* Y*, namely,* $(Y \perp\!\!\!\perp S|X, Z)$*;*

(iv) $Z \cup \{X, Y\} \subseteq M$*, and* $Z \subseteq T$*.*

Consider Fig. 5.3(a) where $Z^- = \{W\}, Z^+ = \{\}$ and Z^- is *not* separated from Y given $\{X\} \cup Z^+$ in G_s, which means that condition (ii) of the s-backdoor is violated. So, despite the fact that the relationship between X and Y is unconfounded and $(Y \perp\!\!\!\perp S|\{W, X\})$, it is improper to adjust for $\{W\}$ when computing the target effect.

For the admissible cases, we are ready to state a sufficient condition that guarantees proper identifiability and recoverability of causal effects under selection bias:

Theorem 24 (Selection-backdoor adjustment)**.** *If a set* Z *satisfies the* s*-backdoor criterion relative to the pairs* (X, Y) *and* (M, T) *(as given in def. 20), then the effect of* X *on* Y *is identifiable and* s*-recoverable and is given by the formula*

$$P(y|do(x)) = \sum_z P(y|x, z, S = 1)P(z) \tag{5.9}$$

Proof. See Appendix C. □

Interestingly, X does not need to be measured in the overall population when the s-backdoor adjustment is applicable, which contrasts with the expression given in Theorem 21 where both X and Z (equivalently C) are needed.

Consider Fig. 5.3(b) and assume our goal is to establish $Q = P(y|do(x))$ when external data over $\{W_2\}$ is available in both studies. Then, $Z = \{W_2\}$ is s-backdoor admissible and the s-backdoor adjustment is applicable in this case. However, if $T = \{W_1\}$, $Z = \{W_1\}$ is backdoor admissible, but it is *not* s-backdoor admissible since condition (iii) is violated (i.e., $(S \perp\!\!\!\perp Y|\{W_1, X\})$ does not hold in G_s). This is interesting since the two sets $\{W_1\}$ and $\{W_2\}$ are c-equivalent (PP10), having the same potential for bias reduction in the general population. To understand why c-equivalence is not sufficient for s-recoverability, note that despite the equivalence for adjustment, $\sum_{w_1} P(y|x, w_1)P(w_1) = \sum_{w_2} P(y|x, w_2)P(w_2)$, the r.h.s. is obtainable from the data, while the l.h.s. is not.

Now we want to recover $Q = P(y|do(x))$ in Fig. 5.3(c) (U is a latent variable) with $T = \{W_2\}$. Condition (iii) of the s-backdoor fails since $(S \perp\!\!\!\perp Y|\{X, W_2\})$ does not hold. Alternatively, if we discard W_2 and consider the null set for adjustment ($Z = \{\}$), condition (i) fails since there is an open backdoor path from X to Y ($X \leftarrow W_2 \leftarrow U \rightarrow Y$). Despite the inapplicability of the s-backdoor, $P(y|do(x))$ is still s-recoverable since, using do-calculus, we can show that $Q = P(y|do(x), S = 1)$, which reduces to $\sum_{w_2} P(y|x, w_2, S = 1)P(w_2|S = 1)$, both factors s-recoverable without the need for external information.

The reliance on the do-calculus in recovering causal effects is expected since even when selection bias is absent, there exist identifiability results beyond the backdoor. Still, this criterion, which is generalized by the s-backdoor criterion, is arguably the most used method for identifiability of causal effects currently available in the literature.

5.6 Recoverability of the Odds Ratio

In this section, we consider the measure of association known as the odds ratio (OR) and show that exploring its functional form allows the recoverability of more quantities than in the previous scenarios.

Consider the chain structure in Fig. 5.4(a) which might represent the qualitative assumptions of a study of the effect of a training program (X) on earnings after 5 years of completion (Y) without confounding between treatment and outcome. Assume that subjects achieving higher income tend to report their status more frequently than those with lower income. Given that all available data is obtained under selection bias, is the unbiased odds ratio recoverable?

The chain structure is the simplest structure exhibiting selection bias and known to not be recoverable for conditional distributions by Thm. 20. The intuition gained from analyzing the odds ratio in this example will serve as a basis for treating more complicated structures.

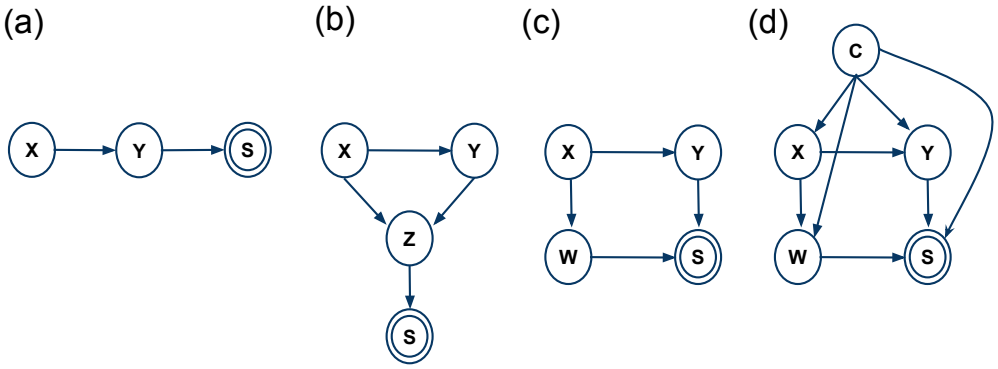


Figure 5.4: (a) Chain graph where X represents treatment, Y is the outcome, and S an indicator variable for the selection mechanism. (b) Scenario where there exists a blocking set from $\{X, Y\}$ to S yet the OR is not G-recoverable. (c) Example where the selection is outcome dependent and $P(y|x)$ is not recoverable, but is the OR. (d) Example where the C-specific OR is G-recoverable.

We define next some key concepts used along the chapter and state some results that will support answering about the recoverability of the odds ratio.

Definition 23 (Odds ratio)**.** *Consider two variables X and Y and a set Z, the conditional odds ratio $OR(Y, X \mid Z = z)$ is given by the ratio: $\big(Pr(y \mid z, x')/Pr(y' \mid z, x')\big)/\big(Pr(y \mid z, x)/Pr(y' \mid z, x)\big)$.*

$OR(Y, X \mid Z)$ measures the strength of association between X and Y conditioned on Z and it is symmetric, i.e., $OR(Y, X \mid Z) = OR(X, Y \mid Z)$.

Definition 24 (G-Recoverability)**.** *Given a graph G, $OR(X, Y \mid Z)$ is said to be G-recoverable from s-biased data if the assumptions embedded in G renders it expressible in terms of the observable distribution $P(V_{xy} \mid S = 1)$ where $V_{xy} = V \setminus \{S\}$. Formally, for every two probability distributions $P_1(.)$ and $P_2(.)$ compatible with G, $P_1(v_{xy} =| S = 1) = P_2(v_{xy} \mid S = 1)$ implies $OR_1(X, Y \mid Z) = OR_2(X, Y \mid Z)$.*

Definition 25 (Collapsibility)**.** *Consider two variables X and Y and disjoint sets Z and W. We say that the odds ratio $OR(X, Y \mid Z, W)$ is collapsible over W if $OR(X, Y \mid Z = z, W = w) = OR(X, Y \mid Z = z, W = w') = OR(X, Y \mid Z = z)$, for all $w \neq w'$.*

Definition 25 and the following Lemma are stated in (DKK10) and are based on long tradition in Epidemiology starting with (Cor51) and followed by (Whi78;

Gen92).[8]

Lemma 5. *For any two sets, Z and W, the conditional odds ratio $OR(Y, X \mid Z, W)$ is collapsible over W (that is, $OR(Y, X \mid Z, W) = OR(Y, X \mid Z)$), if either $(X \perp\!\!\!\perp W \mid \{Y, Z\})$ or $(Y \perp\!\!\!\perp W \mid \{X, Z\})$.*

The following Corollary provides a graphical test for G-recoverability (Def. 24) based on Lemma 5:

Corollary 16. *Given a graph G in which node S represents selection, the $OR(X, Y \mid Z)$ is G-recoverable from s-biased data if Z is such that $(X \perp\!\!\!\perp S \mid \{Y, Z\})_G$ or $(Y \perp\!\!\!\perp S \mid \{X, Z\})_G$.*

Corollary 16 conveys the main distinction between the types of recoverability allowed in the OR case versus the other measures previously discussed. Note that while Theorem 20 requires the independence of S from Y (given X), Corollary 16 allows recoverability through the independence of S from X (given Y), which follows directly from the symmetric form of the OR. More concretely, consider the graph in Fig. 5.4(a) and note that X is d-separated from S by Y, which implies that the odds ratio is recoverable in this case – this independence encodes the assumption that entry to the data pool is determined by the outcome Y only, not by X – while $P(y|x)$ is not recoverable from selection in this case.

There is another important subtlety here. One might surmise that selection bias of $OR(X, Y)$ can be removed if the condition of Corollary 16 holds, i.e., there exists a separating set Z such that $(X \perp\!\!\!\perp S \mid \{Y, Z\})_G$ or $(Y \perp\!\!\!\perp S \mid \{X, Z\})_G$, but this is not the case. Consider Fig. 5.4(b) where the set Z d-separates $\{X, Y\}$ from S and therefore permits us to remove S by writing $OR(X, Y \mid Z, S = 1)$ as $OR(X, Y \mid Z)$, yet the unconditional OR is not G-recoverable because we cannot re-apply the condition of Corollary 16 to eliminate Z from $OR(X, Y \mid Z)$. Moreover, the resulting quantity, $OR(X, Y \mid Z)$, though estimable for every level $Z = z$, does not represent a meaningful relation for decision making or interpretation, because it does not stand for a causal effect in a stable subset of individuals (see discussion about the causal OR at the end of this section). Since Z is X-dependent in G, the class of units for which $Z = z$ under $do(X = 1)$ is not the same as the class of units for which $Z = z$ under $do(X = 0)$. The conditional odds ratio $OR(X, Y \mid Z)$ would be meaningful only if Z is restricted to pre-treatment covariates, which are X-invariant, hence stable.

We next introduce a criterion, followed by a procedure to decide whether it is legitimate to replace Z with a set C of pre-treatment covariates, for which

[8] Cornfield's result and some of its graphical ramifications were brought to our attention by Sander Greenland. See also (GP11).

$OR(Y, X \mid C)$ is a meaningful c-specific causal effect. Typical examples of c-specific effects would be $C = \{age, sex\}$ or, when average behavior is desired, $C = \{\}$.

Definition 26 (OR-admissibility)**.** *A set $Z = \{Z_1, ..., Z_n\}$ is OR-admissible relative to an ordered triplet (X, Y, C) whenever an ordering $(Z_1, ..., Z_n)$ exists such that for each Z_k, either $(X \perp\!\!\!\perp Z_k \mid C, Y, Z_1, ..., Z_{k-1})$ or $(Y \perp\!\!\!\perp Z_k \mid C, X, Z_1, ..., Z_{k-1})$.*

Corollary 17 ((DKK10))**.** *OR-admissibility of Z implies $OR(Y, X \mid C, Z) = OR(Y, X \mid C)$.*

This Corollary follows by successive application of Lemma 5 to the elements $Z_1, ..., Z_n$ of Z.

Theorem 25 (OR G-recoverability)**.** *Let graph G contain the arrow $X \to Y$ and a set C of measured X-independent covariates. The c-specific odds ratio $OR(Y, X \mid C)$ is G-recoverable from s-biased data if and only if there exists an additional set Z of measured variables such that the following conditions hold in G:*

1. *$(X \perp\!\!\!\perp S \mid \{Y, Z, C\})_G$ or $(Y \perp\!\!\!\perp S \mid \{X, Z, C\})_G$.*
2. *Z is OR-admissible relative to (X, Y, C).*

Moreover, $OR(Y, X \mid C) = OR(Y, X \mid C, Z, S = 1)$. [9]

Proof. See Appendix C. □

Note that unlike the control of confounding, which requires averaging over the adjusted covariates, a single instantiation of the variables in Z is all that is needed for removing selection bias.

Now consider the problem found in the medical literature reported in (HF78; HHR04; GRB09) and depicted in Fig. 5.4(c) in which the effect of Oestrogen (X) on Endometrial Cancer (Y) was noticed to be overestimated in the data studied. One of the symptoms of the use of Oestrogen is vaginal bleeding (W), and the hypothesis was that women noticing bleeding are more likely to visit their doctors,

[9] *This Theorem builds on and extends the results in (DKK10) which are summarized by Definition 26 and Corollary 17. First, it supplements the sufficient condition with its necessary counterpart. This is made possible by defining G-recoverability in terms of identifiability (Def. 24). Second, Theorem 25 explicitly avoid meaningless ORs (i.e., $OR(X, Y \mid Z)$, where Z is X-dependent). Finally, the proof of the sufficiency part prepares the ground for a procedure for finding an admissible sequence if such exists, to be shown next.*

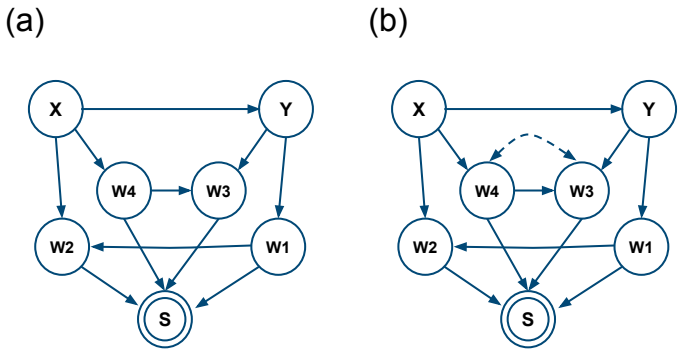


Figure 5.5: Scenario where OR is G-recoverable and $Z = \{W_1, W_2, W_4\}$ (a), and it is not G-recoverable in (b).

causing women using Oestrogen to be overrepresented in the study. This problem can be easily solved applying Theorem 25 with $Z = \{W\}$ – we can immediately verify that Z is OR-admissible relative to $(X, Y, \{\})$ (i.e., $(W \perp\!\!\!\perp Y \mid X)$), and $(X \perp\!\!\!\perp S \mid \{Y, W\})$ holds. Thus, we can write $OR(Y, X) = OR(Y, X \mid W) = OR(X, Y \mid W) = OR(X, Y \mid W, S = 1)$, which shows a mapping from the target (unbiased) quantity (without any S) to the s-biased data (conditioned on $S = 1$, which was measured). (In the sequel we will drop G finding no need to distinguish conditional independencies from d-separation statements.) [10]

Theorem 25 defines the boundary that distinguishes the class of graphs that permit G-recoverability of OR from those that do not. To show the power of Theorem 25, let us consider the more intricate scenario of Fig. 5.5(a), in which $Z = \{W_1, W_2, W_4\}$ satisfies the conditions of Theorem 25. This can be seen through the following sequence of reductions verified by the graph: $(X \perp\!\!\!\perp S \mid \{Y, W_1, W_2, W_4\}) \rightarrow (Y \perp\!\!\!\perp W_2 \mid \{X, W_1, W_4\}) \rightarrow (X \perp\!\!\!\perp W_1 \mid \{Y, W_4\}) \rightarrow (Y \perp\!\!\!\perp W_4 \mid X)$. The final result is

$$OR(Y, X) = OR(Y, X \mid W_1, W_2, W_4, S = 1)$$

where the term on the left is our target quantity and the one on the right is estimable from the s-biased data. Fig. 5.5(b) shows an example where OR is not G-recoverable, because we must start with $Z = \{W_1, W_2, W_3, W_4\}$ or $Z = \{W_1, W_3, W_4\}$ to separate S from X or Y, respectively – these two sets are not OR-admissible since each set contains the variable W_3 which cannot be separated from X or Y by any set.

[10] Furthermore, the graph symmetric to Fig. 5.4(c) where the positions of X and Y are interchanged yields the same result. Similarly, another common variant of Fig. 5.4(c), with the edge $X \rightarrow W$ reversed, is solvable as well.

Theorem 25 relies on OR-admissibility, for which Definition 26 gives a declarative, non-procedural criterion. Taken literally, it requires that we first find a proper Z and then, out of the $n!$ orderings of the elements in Z, find one that will satisfy the d-separation tests specified in Definition 26. We will now supplement Theorem 25 with a simple graphical condition, followed by an effective procedure for finding such a sequence if one exists.

Theorem 26. *Let graph G contain the arrow $X \to Y$, a necessary condition for G to permit the G-recoverability of $OR(Y, X \mid C)$ for a given set C of pre-treatment covariates is that S and every ancestor A_i of S that is also a descendant of X have a separating set T_i that either d-separates A_i from X given Y, or d-separates A_i from Y given X.* [11]

Proof. See Appendix C. □

Theorem 27. *Let G be a DAG containing the arrow $X \to Y$ and two sets of variables, measured V and unmeasured U. A necessary and sufficient condition for G to permit the G-recoverability of $OR(Y, X \mid C)$ for a given set C of pre-treatment variables is when the sink-procedure below terminates. Moreover, $OR(Y, X \mid C) = OR(Y, X \mid C, Z, T, S = 1)$, where $Z = \big(An(S) \setminus An(Y)\big) \cap V$ and T is given by the sink-procedure.*

Procedure (Sink reduction)

1. Set $T = \{\}$, and consider Z as previously defined. Remove $V \setminus An(Y \cup S)$ from G, and name the new graph G^*. Consider an ordering compatible with G^* such that $Z_i < Z_j$ whenever Z_i is non-descendant of Z_j.

2. Test if sink Z_i of G^* satisfies the following condition: $(Z_i \perp\!\!\!\perp X \mid C, T, Y, Z_1, ..., Z_{i-1})$ or $(Z_i \perp\!\!\!\perp Y \mid C, T, X, Z_1, ..., Z_{i-1})$. If so, go to step 4. Otherwise, continue.

3. Test if there exists a minimal set T_i of non-descendants of X that, if added to T would render step 2 successful, if none exists, exit with failure.[5] Else, add T_i to T and continue with step 4.

4. Remove Z_i from G^* and Z, and repeat step 2 recursively until Z is empty. If so, go to step 5.

[11] *A polynomial time algorithm for finding a minimal separating set in DAGs is given in (TPP98). The restricted minimal separation version of that algorithm finds a minimal separator in a DAG with latent variables (equivalently, semi-Markovian models). A fast test for the non-separability of X and A_i is the existence of an inducing path between the two variables (VP90). For example, the path $X \to W_4 \to W_3$ in Fig. 5.5(b).*

5. Test if $(T \perp\!\!\!\perp Y \mid C, X)$, if so, the sequence $(Z_1, Z_2, ...Z_m)$ with T constitutes a witness for the OR-admissibility of Z relative to (X, Y, C), for a set C of X-independent variables. Otherwise, exit with failure.

Proof. See Appendix C. □

The algorithm exploits the graph structure to construct a mapping from the observed s-biased data and the desired target OR. Since the OR is symmetric, it is not necessary to separate S from X and Y simultaneously, but only from one of them (given the other.) For simplicity, denote the expression "X given Y or Y given X" by the symbol Φ_{xy}. A separating set from S to Φ_{xy} is first sought in step 2, starting with all observable ancestors of S that are non-ancestors of Y. If the test succeeds and this set is a separator, the algorithm iterates trying to separate Φ_{xy} from the deepest node in the remaining set. In case of failure, the algorithm attempts (step 3) to achieve separability using pre-treatment covariates T_i. In case no separability can be found using these added covariates, the algorithm fails. Otherwise, at the end, the algorithm further requires that all T_i added along these iterations be separable from Y (step 5).

To illustrate, running the procedure on the graph of Fig. 5.5(b) with $C = \{\}$, the graph remaining after the removal of S has two sink nodes, W_2 and W_3. Removing W_2 leaves two other sinks, W_3, and W_1. Removing W_1 leaves W_3 as the only remaining sink node which fails the test of Step 3. Since no non-descendant of X exists that yields separability, we must exit with failure. On the other hand, if we are able to measure U, the hidden variable responsible for the double arrow arc between W_3 and W_4, we would add this node to T, W_3 will pass the test, followed by W_4, and we will end up with U as the only non-descendant of X remaining in T. In step 5 we remove U from T, yielding $OR(X, Y) = OR(X, Y \mid W, U, S = 1)$.

Thus far, we assumed that the treatment X is unconfounded, therefore the OR is identical to the causal OR defined as $COR(X, Y) = \frac{P(y|do(x))P(y'|do(x'))}{P(y|do(x'))P(y'|do(x))}$. In the presence of confounding, it is not enough to recover OR in s-biased data, we need to go further and assure that the recovered $OR(X, Y \mid C)$ is such that C satisfies the back-door criterion (2nd rule of do-calculus, observing and intervening are equivalent), in which case $OR(X, Y \mid C)$ will represent the c-specific causal OR. For example, in Fig. 5.4(d) the $COR(X, Y \mid C)$ will be G-recoverable because once we condition on C all conditional independencies will be identical to those of Fig. 5.4(c), and $P(Y \mid do(X), C) = P(Y \mid X, C)$.

Note, however that although we can recover the c-specific causal OR, we cannot recover the population $COR(X, Y)$. For such measure to be recoverable we need to add assumptions which will make it possible to infer averageable measures of causal effects such as RD and RR, to be handle next.

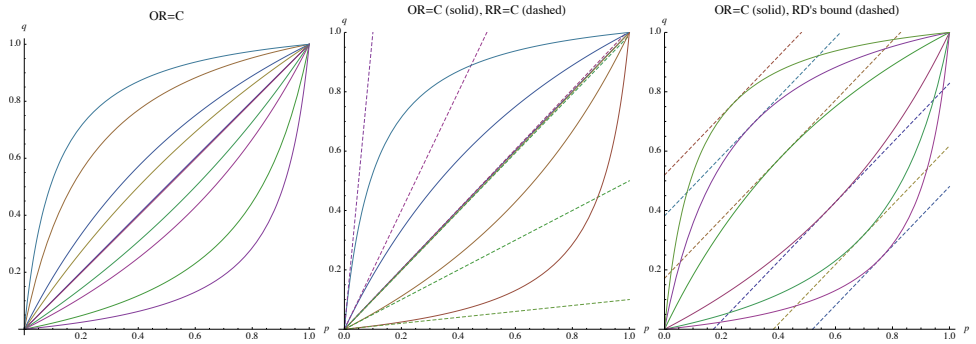


Figure 5.6: (a) Constant odds ratio curves for $c = \{1.00, 1.01, 1.50, 2.00, 5.00, 10.00\}$ and their inverses; Superimposed constant odds ratio with constant risk ratio (b) and constant risk difference curves (c).

5.6.1 The Relationship Between OR and Other Effects

Consider again the chain structure in Fig. 5.4(a) and define the causal effect as $COR(X, Y)$. The fact that X and Y are not confounded permits us to estimate the causal effect $COR(X, Y)$ by the odd ratio $OR(X, Y)$ which, by the results in the previous section, will remain invariant to conditioning on $S = 1$. However, if we define the causal effect as $ACE = Pr(y \mid do(x)) - Pr(y \mid do(x'))$ (also known as the causal risk difference), a bias will be introduced upon conditioning.

The invariance of OR can be represented in the following intuitive and pictorial way. We characterize the conditional distribution $P(Y \mid X)$ by two independent parameters $p = P(y \mid x)$ and $q = P(y \mid x')$, which define a point (p, q) in the unit square. The condition $OR(X, Y) = c$ describes a curve in the $(p, q)-$plane. For $c = 1$, the curve is the unit slope line. For $c > 1$, this curve separate points with $OR(.) > c$ from those with $OR(.) < c$ in the region below the unit slope line (symmetrically for the inverses ($c < 1$) in the region above $q = p$). See Fig. 5.6.

Now, by conditioning on $S = 1$, we obtain a new conditional probability, also characterized by two independent parameters $p_s = P(y \mid x, S = 1), q_s = P(y \mid x', S = 1)$. The fact that $OR(Y, X \mid S = 1) = OR(Y, X)$ means that conditioning on $S = 1$ must shift the initial (p, q) point along a constant OR curve, not anywhere else. We show these universal curves of constant OR for $c = \{1.00, 1.01, 1.50, 2.00, 5.00, 10.00\}$ and their respective inverses in Fig. 5.6(a). Fig. 5.6(b) shows curves for constant risk ratio (RR: $\frac{p}{q} = c$), which are variable slope lines going through the origin, and bounded by the slope $\frac{1}{c}$. Similarly, Fig. 5.6(c) shows curves for constant risk difference.

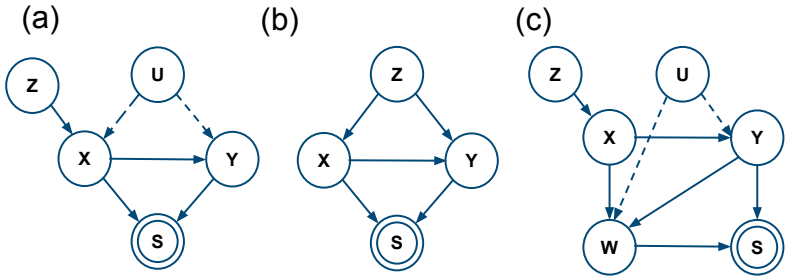


Figure 5.7: Different scenarios in which Theorem 28 can be applied. (a) Typical study with randomization and non-compliance (IV as incentive-mechanism) where selection and confounding are both present. (b) Selection bias in the back-door case. (c) More complex study with an intermediary variable W between treatment and selection. In this case, Y directly cause W and there is a common cause between them (extension of Fig. 5.4(c), see Corollary 20.)

We see that even though RR does not remain constant (upon conditioning), the constancy of OR constrains the behavior of the RR. This follows by noting (after some algebra) that $RR = c + (1 - c)p$, i.e., RR has intercept c and slope $1 - c$. For instance, if OR is constant and $c = 1$, we have unit slope line for OR, but RR does not move and is equal to one. For constant OR and $\frac{1}{2} < c < 1$, the slope is positive but less than $\frac{1}{2}$, and the intercept is greater than $c = \frac{1}{2}$, which implies that RR lies inside the interval $[c, 1]$. Similar bounds can be obtained for other values of c.

5.7 Recoverability with Instrumental Variabless

In this section, we consider the problem of recoverability when confounding and selection biases are simultaneously present, see Fig. 5.7(a) for an example.

Our goal is to infer the most accurate bounds for the causal effect of X on Y, knowing that there is no unbiased estimate for this quantity even when selection bias is not present (i.e., $S \perp\!\!\!\perp V$). This scenario is usually presented under the rubric of "randomization with non-compliance", and it is pervasive in the Economics literature, we defer to (Pea09b, Ch. 8) for a more comprehensive discussion of the relevance of this setup, we focus here on the technical aspects of the problem.

Generally, the bounding analysis assumes no selection bias, and the natural question that arises is whether selection bias can be treated and under what conditions bounds free from selection can be recovered.

We show next that this problem can be solved assuming the existence of two instrumental variables Z_1 and Z_2. [12] Noteworthy, the set of assumptions used in our analysis are commonplace in daily Econometrics practice, and its convoluted appearance is diluted when one observes them more vividly through the causal graph depicted in Fig. 5.7(a). In a nutshell, they are the same assumptions of randomization with non-compliance together with selection bias (such that treatment and outcome affect entry in the data pool).

Theorem 28. *The joint distribution of $P(X, Y, Z)$ is recoverable from s-biased data whenever the following conditions hold: (i) the S node is affected by the set Z only through $\{X, Y\}$; (ii) the set Z is d-connected to $\{X, Y\}$ (and combinations); (iii) the dimensionality of Z matches the dimensionality of $\{X, Y\}$; (iv) the marginal probability of Z is known. In other words, the distribution $P(X, Y, Z)$ is recoverable from s-biased data whenever $(S \perp\!\!\!\perp Z \mid X, Y)$, $(Z \not\perp\!\!\!\perp \{X, Y\})$, $(Z \not\perp\!\!\!\perp X \mid Y)$,$(Z \not\perp\!\!\!\perp Y \mid X)$, the dimensionality of Z and $X \cup Y$ matches, and the marginal distribution of $P(Z)$ is given.*

Proof. See Appendix C. □

Corollary 18. *The bounds for $P(y \mid do(x))$ in the scenario of randomization with non-compliance (Fig.5.7(a)) are recoverable from s-biased data whenever the conditions of the Theorem 28 hold.*

Proof. It follows directly from Theorem 28 together with the bounds in (BP97). □

Corollary 19. *The causal effect $P(y \mid do(x))$ in the back-door scenario (Fig. 5.7(b)) is recoverable from s-biased data whenever the conditions of the Theorem 28 hold.*

Proof. It follows directly from Theorem 28. □

Corollary 20. *The causal effect of Oestrogen (X) on Endometrial Cancer (Y) as studied in (HF78; HHR04) (Fig. 5.7(c)) is recoverable from s-biased data whenever there is an IV set Z pointing to X, and the conditions of the Theorem 28 hold. Moreover, the same holds without relying on Z whenever the following conditions hold: (i) X has the same dimensionality of $\{W, Y\}$; (ii) the marginal distribution of $P(X)$ is available.*

Proof. See Appendix C. □

[12] Call $Z = Z_1 \cup Z_2$, or consider one IV with the same number of levels. Let us name both cases by instrumental variable set.

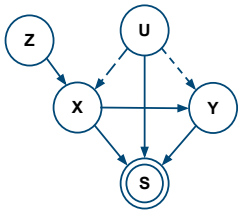


Figure 5.8: Scenario in which selection and confounding biases are present, entangled, and thus not recoverable.

Some observations on the method

Methods that handle selection bias try to model the distribution of S, which is unobservable and not always easy to estimate; we take a different approach and avoid doing this explicit manipulation of the selection mechanism by exploiting the topology of the causal graph and the underlying data-generating process. We are not aware of other approaches trying to do so.

The main idea is to exploit the conditional independence of the IV set Z and the selection mechanism S given the distribution of the treatment and outcome – interestingly, the latter is what we seek to estimate. The method hinges on two properties about the induced system, that it is linearizable and full rank – both facts were not obvious nor expected a priori.

It is worth to make some additional remarks that follow the proof of Theorem 28. First note that the proposed method relies on a sample size approaching infinity, which is difficult to obtain in practice. As a possible improvement, the problem could be cast as an optimization problem. The formulation goes as follows. We associate error terms $\epsilon_{z_1 z_2, xy}$ to each $\gamma_{z_1 z_2, xy}$ term, and proceed the analysis minimizing the (square) mean error subject to constraints. The constraints emerge naturally from the induced system of equations together with the additional constraints of positivity and integrality. Our original goal was to show feasibility of removing selection bias (identifiability) but not the estimation per se, still, this should be an interesting exercise to pursue. Further investigation is needed to check the applicability of this suggestion.

We envision our method being used as a first step in a pre-processing stage, before the application of any bounding (BP97) or estimation procedure. The method returns the same values of $P(X, Y, Z)$ whenever the collected data is not under selection bias, which means that its usage will not hurt and should be considered as a "good practice."

Finally, note that there are scenarios not solvable by our method or in which our assumptions are not applicable. For instance, we show in Fig. 5.8 one of this

kind, in which selection and confounding biases are entangled in such way that it does not seem possible to detach one from another. We conjecture that this case is not solvable in general without further assumptions. Notice that even if we remove the edge $U \to X$, the example is still hard to resolve.

5.8 Conclusions

In this chapter, we provided conditions for recoverability from selection bias in statistical and causal inferences applicable for arbitrary structures in non-parametric settings. Theorem 20 provided a complete characterization of recoverability when no external information is available. Theorem 21 provided a sufficient condition for recoverability based on external information; it is optimized by Theorem 22 and strengthened by Theorem 23. Verifying these conditions takes polynomial time and could be used to decide what measurements are needed for recoverability. Theorem 24 further gave a graphical condition for recovering causal effects, which generalized the back-door adjustment.

We further relaxed our requirements and considered recoverability of the odds ratio (OR). Theorem 25 provided a complete graphical condition under which the population OR and a covariate-specific causal OR can be recovered from selection. We then devised an effective procedure for testing this condition (Theorems 26 and 27). These results, although motivated by causal considerations, are applicable to classification tasks as well, since the process of eliminating selection bias is separated from that of controlling for confounding bias. We presented universal curves that show the behavior of OR as the distribution $P(y|x)$ changes, and how the risk ratio (RR) and risk difference (RD) are related to OR.

Finally, we considered the problem of recovering from selection when confounding bias is also present (Fig. 5.7(a)); we showed the former can be entirely removed with the help of instrumental variables (Theorem 28). This result is surprising because bias removal in the presence of confounding is generally expected to be a more challenging task than only under selection. We finally showed how this result is applicable to scenarios where other structural assumptions hold, for instance, when an instrument is not available but a certain back-door admissible set can be identified (Corollary 19).

Given that selection bias is a common problem across many disciplines, the methods developed in this chapter should help to understand, formalize, and alleviate this problem in a broad range of data-intensive applications.

CHAPTER 6

Causal Inference by Surrogate Experiments

We address the problem of estimating the effect of intervening on a set of variables X from experiments on a different set, Z, that is more accessible to manipulation. This problem, which we call z-identifiability, reduces to ordinary identifiability when $Z = \emptyset$ and, like the latter, can be given syntactic characterization using the *do-calculus*. We provide a graphical necessary and sufficient condition for z-identifiability for arbitrary sets X, Z, and Y (the outcomes). We further develop a complete algorithm for computing the causal effect of X on Y using information provided by experiments on Z. Finally, we use our results to prove completeness of *do*-calculus relative to z-identifiability, a result that does not follow from completeness relative to ordinary identifiability.

6.1 Introduction

The relation between passive and experimental observations, and how they can aid the estimation of causal effects, is of central interest in the empirical sciences.

In this line of research, the *identification* problem ($I\mathcal{D}$, for short) asks whether causal effects can be computed from the joint distribution P over the observed variables, and theoretical knowledge encoded in the form of a causal diagram G.

This problem has been extensively studied in the literature, and (Pea95; Pea00) gave it rigorous mathematical treatment based on the structural semantics, and introduced several graphical conditions such as the "back-door" and "front-door" criteria, which was later generalized by his *do-calculus*. In the last decades, a number of conditions had emerged for non-parametric identifiability such as the ones given by (SGS93; GP95; PR95; Hal98; KM99). In a series of breakthrough results starting with the development of the concept of C-component (TP02), the *do-calculus* was finally shown to be complete (HV06a; SP06b). This result implies that there exists a finite sequence of applications of the rules of do-calculus that derives the target causal effect Q in terms of the observational distribution P if (and only if) Q is identifiable. It was also provided algorithms that return a mapping from P to Q whenever Q is identifiable.

In real world applications, it is not uncommon that the quantity Q is uniden-

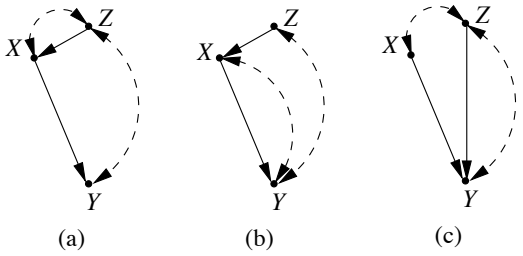


Figure 6.1: Causal diagrams illustrating $z\mathcal{ID}$ of the causal effect $Q = P(y|\hat{x})$. Q can be identified by experiments on Z in model (a), but not in (b) and (c).

tifiable, i.e., the distribution P together with the graph G are not able to unambiguously determine Q. A natural question arises whether the investigator could perform some auxiliary experiments (not necessary spelled out in Q), which would enable him/her to estimate the desired causal effects.

For instance, consider the diagram G in Fig. 6.1(a). Suppose one is interested in assessing the effect Q of cholesterol levels (X) on heart disease (Y), and data about subjects' diet (Z) is also collected. It is clear that Q is unidentifiable from the assumptions embodied in G, but it is infeasible in reality to control subjects' cholesterol level by intervention. Assume that an experiment can be conducted in which the subjects' diet (Z) is randomized; a natural question emerges whether Q is computable given this additional piece of experimental information?

Surprisingly, this ubiquitous problem has not received a thorough formal treatment. We introduce a variation of the $\mathcal{ID}$ problem to fill in this gap. Consider a setting in which, in addition to the information available in an ordinary $\mathcal{ID}$ instance (distribution P and graph G), further experiments can be performed over a set of variables Z; decide whether the target causal effects can be computed from the available information at hand. This extension generalizes the $\mathcal{ID}$ problem (when $Z = \emptyset$ the two problems coincide) and is called here the z-identification problem ($z\mathcal{ID}$, for short). The Z is called surrogate experiments, for obvious reasons.

Syntactically, the $z\mathcal{ID}$ problem amounts to transforming $P(y|\hat{x})$[1] into an equivalent expressions in *do*-calculus such that only members of Z may contain the hat symbol. Applying this rationale for the example above (Fig. 6.1(a)) entails the following reduction in *do*-calculus. First apply Rule 3 to add $\hat{z}$,

$$P(y|\hat{x}) = P(y|\hat{x},\hat{z}) \quad \text{since } (Y \perp\!\!\!\perp Z|X)_{G_{\overline{XZ}}}$$

[1]We will use $P(y|\hat{x})$ interchangeably with $P_x(y)$ or $P(y|do(x))$. We also will call the interventional operator *do()* as the "hat" operator.

Then apply Rule 2 to exchange $\hat{x}$ with x:

$$P(y|\hat{x},\hat{z}) = P(y|x,\hat{z}) \quad \text{since } (Y \perp\!\!\!\perp X|Z)_{G_{\underline{X}\overline{Z}}}$$

This last expression can be rewritten as,

$$P(y|x,\hat{z}) = \frac{P(y,x|\hat{z})}{P(x|\hat{z})} \tag{6.1}$$

This expression shows that performing an experiment on Z suffices to yield "identifiability" of the causal effect of X on Y without experimenting over X. [2]

The subtlety of this problem can be illustrated by noting that in the graph in Fig. 6.1(a) the effect is z-identifiable from $P(V)$ and $P(X,Y|\hat{Z})$ in G, whereas in the graph in Fig. 6.1(b) it is not (to be shown later). The only difference between these two graphs is the bidirected edge between the pairs (X,Z) and (X,Y).

One might surmise that $z\mathcal{ID}$ can be represented by a mutilated graph in which the edges incoming to Z are cut, and the problem would then be solved as ordinary identifiability. Unfortunately, this is not the case as shown in the graph in Fig. 6.1(c) where $Q = P(y|\hat{x})$. The option of manipulating Z does not enable us to compute the Z-specific causal effect of X on Y, $P(y|\hat{x},z)$ which , if available, would allow us to compute the overall causal effect by averaging over Z. Although $Q' = P(y|\hat{x},\hat{z})$ can be established from the mutilated graph, it does not help in establishing the Z-specific causal effect, or Q.

The first formal treatment of this problem (Pea95) led to the following sufficient condition for admitting a surrogate variable Z for the causal effect $P(y|\hat{x})$:

(i) X intercepts all directed paths from Z to Y, and

(ii) $P(y|\hat{x})$ is identifiable in $G_{\overline{Z}}$.

These conditions are satisfied indeed in the model of Fig. 6.1(a) but not in 6.1(b) or 6.1(c). Pearl's criterion is sufficient but was not shown to be necessary. Additionally, it was not extended to the case where Z and X are sets of variables. At the same time, the syntactic condition above, which requires the existence of a do-calculus transformation expression containing only do(z) terms is declarative, but is not computationally effective, since it does not specify the sequence of

[2] The expression also shows that only one level of Z suffices for the identification of $P(y|\hat{x})$ for any value of y and x. In other words, Z need not be varied at all; it can simply be held constant by external means and, if the assumptions embodied in G are valid, the r.h.s. of eq. (6.1) should attain the same value regardless of the (constant) level at which Z is being held constant. In practice, however, several levels of Z will be needed to ensure that enough samples are obtained for each desired value of X.

rules leading to the needed transformation, nor does it tell us if such a sequence exists. Even though do-calculus is complete for identifying causal effects, it is not immediately clear whether it is complete for $z\mathit{ID}$.

This chapter provides a systematic study of z-identifiability building on Pearl's condition and the previous results from the identifiability literature; our contributions are as follows:

- We provide a necessary and sufficient graphical condition for the problem of z-identification when Z is a set of variables.
- We then construct a complete algorithm for deciding z-identification of joint causal effects and returning the correct formula whenever those effects are z-identifiable.
- We further show that *do*-calculus is complete for the task of z-identification.

6.2 Notation and Definitions

The basic semantical framework in our analysis rests on *probabilistic causal models* as defined in Chapter 2. We build on the problem of identifiability, defined below for convenience, which expresses the requirement that causal effects must be computable from a combination of passive data P and the assumptions embodied in a causal graph G (*without* assuming any availability of additional experimental information).

Definition 27 (Causal Effects Identifiability (Pearl))**.** *Let X, Y be two sets of disjoint variables, and let G be the causal diagram. The causal effect of an action $do(X = x)$ on a set of variables Y is said to be identifiable from P in G if $P_x(y)$ is (uniquely) computable from $P(V)$ in any model that induces G.*

The following Lemma is the operational way to prove that a causal quantity is not identifiable given the assumptions embedded in G.

Lemma 6. *Let X, Y be two sets of disjoint variables, and let G be the causal diagram. $P_x(y)$ is not identifiable in G if there exist two causal models M^1 and M^2 compatible with G such that $P_1(V) = P_2(V)$, and $P_1(y|do(x)) \neq P_2(y|do(x))$.*

Proof. The latter inequality rules out the existence of a function from P to $P_x(y)$. □

Next, we formally introduce the problem of z-identifiability that generalizes the problem of identifiability whereas it is no longer assumed that experimental

information is not available at all, but there exists a set of variable Z in which experiments were performed and now is available for use. In other words, the explicit acknowledgement of the existence of the set Z adds a degree of freedom for the researcher, making the analysis more flexible and perhaps realistic.

Definition 28 (Causal Effects z-Identifiability)**.** *Let X, Y, Z be disjoint sets of variables, and let G be the causal diagram. The causal effect of an action $do(X = x)$ on a set of variables Y is said to be z-identifiable from P in G, if $P_x(y)$ is (uniquely) computable from $P(V)$ together with the interventional distributions $P(V \setminus Z'|do(Z'))$, for all $Z' \subseteq Z$, in any model that induces G.*

Armed with this new definition, we state next the sufficiency of the *do*-calculus for $z\mathcal{ID}$ that is analogous to (Pea00, Corol. 3.4.2) in respect to identification.

Theorem 29. *Let X, Y, Z be disjoint sets of variables, let G be the causal diagram, and $Q = P(y|do(x))$. Q is $z\mathcal{ID}$ from P in G if the expression $P(y|do(x))$ is reducible, using the rules of do-calculus, to an expression in which only elements of Z may appear as interventional variables.*

Proof. The result follows from soundness of do-calculus and the definition of z-identifiability. □

It is clear that if we have an efficient procedure to establish $z\mathcal{ID}$, we can immediately decide $\mathcal{ID}$ by setting $Z = \emptyset$. On the other hand, to be able to establish the converse of Theorem 29, we need to understand the conditions for non-$z\mathcal{ID}$, and so, we state next the analogous of Lemma 7 in this context.

Lemma 7. *Let X, Y, Z be disjoint sets of variables, and let G be the causal diagram. $P_x(y)$ is not z-identifiable in G if there exist two causal models M^1 and M^2 compatible with G such that $P^1(V) = P^2(V)$, $P^1(V \setminus Z'|do(Z')) = P^2(V \setminus Z'|do(Z'))$, for all $Z' \subseteq Z$, and $P^1_x(y) \neq P^2_x(y)$.*

Proof. Let I be the set of interventional distributions $P(V \setminus Z'|do(Z'))$, for any $Z' \subseteq Z$. The latter inequality rules out the existence of a function from P, I to $P_x(y)$. □

While Lemma 7 might appear convoluted, it is nothing more than a formalization of the statement "Q cannot be computed from information set S alone." Naturally, when S has two components, $\langle P, I\rangle$, the Lemma becomes lengthy. Even though the problems of $\mathcal{ID}$ and $z\mathcal{ID}$ are related, Lemma 7 indicates that proofs of non-$z\mathcal{ID}$ are at least as hard as the ones for non-$\mathcal{ID}$, given that to prove the former requires the construction of two models to agree on $\langle P, I\rangle$, while to prove the latter it is only required the two models to agree on the distribution P.

6.3 Characterizing zID Relations

The concept of confounded component (or *C-component*) was introduced in (TP02) to represent clusters of variables connected through bidirected edges, and was instrumental in establishing a number of conditions for ordinary identification (Def. 27). If G is not a C-component itself, it can be uniquely partitioned into a set $\mathcal{C}(G)$ of C-components. We state below this definition that will also play a key role in the problem of zID.

Definition 29 (C-component)**.** *Let G be a causal diagram such that a subset of its bidirected arcs forms a spanning tree over all vertices in G. Then G is a C-component (confounded component).*

A special subset of C-components that embraces the ancestral set of Y was noted by (SP06b) to play an important role in deciding identifiability – this observation can also be applied to z-identifiability, as formulated next.

Definition 30 (C-forest)**.** *Let G be a causal diagram, where Y is the maximal root set. Then G is a Y-rooted C-forest if G is a C-component and all observable nodes have at most one child.*

We next introduce a structure based on C-forests that witnesses unidentifiability characterized by a pair of C-forests. ID was shown by (SP06b) infeasible if and only if such structure exists as an edge subgraph of the given causal diagram.

Definition 31 (hedge)**.** *Let X, Y be set of variables in G. Let F, F' be R-rooted C-forests such that $F \cap X \neq 0$, $F' \cap X = 0$, $F' \subseteq F$, $R \subset An(Y)_{G_{\overline{X}}}$. Then F and F' form a hedge for $P_x(Y)$ in G.*

The presence of this structure will prove to be an obstacle to z-identifiability of causal effects in various scenarios. For instance, the p-graph in Fig. 6.1(b) is a Y-rooted C-forest in which $P_x(y)$ will show not to be z-identifiable. However, different than in the ID case, there is no sharp boundary here, since Fig. 6.1(a) also contains a Y-rooted C-forest but $P_x(y)$ was already shown to be zID.

We formally show next that there is a variation of this structure that is able to capture non-zID for a broad set of cases.

Theorem 30. *Let X, Y, Z be disjoint sets of variables and let G be the causal diagram. Then, the causal effects $Q = P_x(y)$ is not zID if there exists a hedge $\mathcal{F} = \langle F, F' \rangle$ for Q in $G_{\overline{Z}}$.*

Proof. The result is immediate. The existence of the hedge $\mathcal{F}$ for Q in $G_{\overline{Z}}$ implies that Z cannot help in the (ordinary) identification of Q. Let us assume that Q is

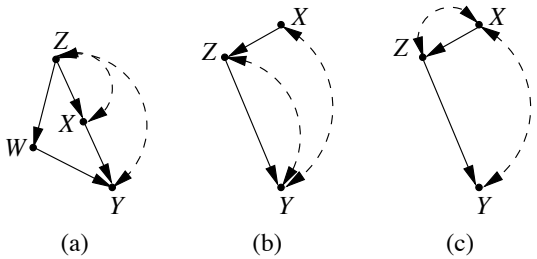


Figure 6.2: Graphs in which $P(y|\hat{x})$ is non-$z\mathcal{ID}$ from $do(Z)$ and there is no hedge in $G_{\overline{Z}}$.

$z\mathcal{ID}$. Note that Z does not participate in the hedge $\mathcal{F}$ since there is no bidirected edge going towards any of its elements in $G_{\overline{Z}}$, which is required by the definition of C-forest. Further, consider a parametrization such that all elements of Z are simply fair coins and disconnected from $V \setminus Z$ in G.

We can now use the same proof of non-$\mathcal{ID}$ based on $\mathcal{F}$ to prove non-$z\mathcal{ID}$ in G. The inequality of Q between the two models is obvious, and the agreement of the interventional distributions $do(Z)$ follows since Z is disconnected from $V \setminus Z$ by the chosen parametrization. This is a contradiction since $z\mathcal{ID}$ has to be valid for any parametrization compatible with G, which suffices to prove the result. □

Consider the next Corollary in regard to the p-graph, which is the smallest example in which Z could aid in the z-identification of Q but Q is still not z-identifiable from $do(Z)$. This and similar structures that prevent $z\mathcal{ID}$ will be one of the base cases for our proof of completeness, which requires a demonstration that whenever the algorithm fails to z-identify a causal relation, the relation is indeed non-$z\mathcal{ID}$.

Corollary 21. *$P_x(y)$ is not $z\mathcal{ID}$ in the p-graph.*

Proof. This follows directly from Theorem 30 since there exists a hedge in $G_{\overline{Z}}$. □

The result of Theorem 30 still does not characterize the $z\mathcal{ID}$ class, which suggests that the machinery used to prove completeness in the $\mathcal{ID}$ class is not immediately applicable to the $z\mathcal{ID}$ class.

For instance, consider the graph in Fig. 5.2(a) (called here *bv*-graph), which does not have a hedge for Q in $G_{\overline{Z}}$ but is still non-$z\mathcal{ID}$. The *bv*-graph coincides as an edge subgraph with Fig. 6.1(a) (note C-component induced over $\{X, Y, Z\}$), which turns out to be $z\mathcal{ID}$.

This is an interesting case, since up to this point, in ordinary identification, it was enough to locate a hedge for Q as an edge subgraph of the inputted diagram,

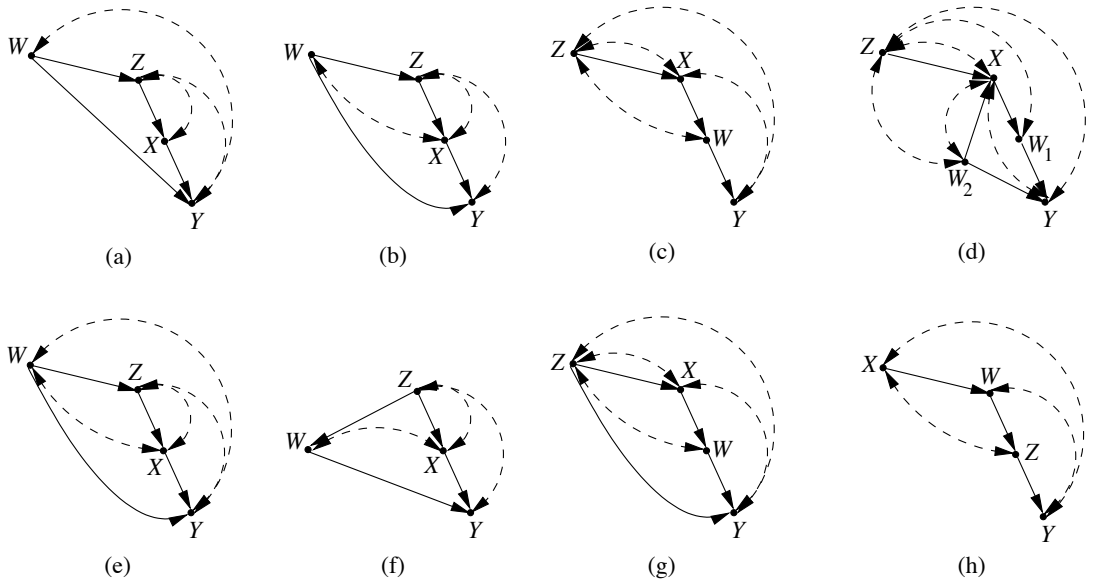


Figure 6.3: $P(y|\hat{x})$ is $z\mathcal{ID}$ from $\langle P, do(Z)\rangle$ in the graphs in the first row (a–d), but not in the the second row (e–h).

and all graphs sharing this substructure were equally unidentifiable (see Thm. 4 in (SP06b)) – this is no longer true here since Z needs to be taken into account. Mainly, note that the directed edges outside a C-component play a very critical role for the $z\mathcal{ID}$ problem as the *bv*-graph demonstrates.

Finally, we expand Pearl's condition (Pea00, pp. 87) in the following directions. We extend, in the intuitive way, his condition to consider when Z is a set of variables and, in turn, we supplement the sufficient part with its necessary counterpart. We finally have a complete characterization for the $z\mathcal{ID}$ class as shown below.

Theorem 31. *Let X, Y, Z be disjoint sets of variables and let G be the causal diagram. The causal effect $Q = P(y|do(x))$ is $z\mathcal{ID}$ in G if and only if one of the following conditions hold:*

a. Q is identifiable in G; or,

b. There exists $Z' \subseteq Z$ such that the following conditions hold,

(i) X intercepts all directed paths from Z' to Y, and

(ii) Q is identifiable in $G_{\overline{Z'}}$.[3]

Proof. See Appendix D. □

Let $Q = P(y|\hat{x})$ be the effect of interest and assume that experiments were performed over $\{Z\}$. Q is $z\mathcal{ID}$ from P and $do(Z)$ in the graphs in Fig. 6.3(a-d), while they are non-$z\mathcal{ID}$ in the graphs in Fig. 6.3(e-h). Except for the trivial case, Theorem 31 is existentially quantified and it is not immediately obvious how to efficiently select the covariates simultaneously satisfying both conditions of the Theorem. Clearly, a naive approach could lead to an exponential number of tests.

For example, consider the graph in Fig. 6.3(a) that is a variation of the *bv*-graph. In this graph, Q is $z\mathcal{ID}$ using experiments from $\{Z\}$. In turn, consider the graph in Fig. 6.3(e), which is the same as 6.3(a) but with the bidirected edge $W \longleftrightarrow X$ added. Now, Q is no longer $z\mathcal{ID}$ for $\{Z\}$ nor $\{Z, W\}$. If we further consider the graph in Fig. 6.3(b) with the bidirected edge $W \longleftrightarrow X$ removed from 3(e), not only Q becomes $z\mathcal{ID}$ for $\{Z\}$ but also for $\{Z, W\}$. This is a border case, note that if we input $\{Z, W\}$ as the surrogate variables for Pearl's criterion, it will not recognize Q as $z\mathcal{ID}$ given the existence of the directed path $W \rightarrow Y$. Finally, if we consider the graph in Fig. 6.3(f) in which the directed edge $W \rightarrow Z$ is flipped from 6.3(b), Q is no longer $z\mathcal{ID}$ for neither $\{Z, W\}$ nor $\{Z\}$.

This example can be extended indefinitely but it is clear that finding a set that satisfies both conditions of the Theorem (in structures more intricate than the 4-node example) does not follow immediately. The subject of the next section is about finding an efficient (and complete) algorithm to solve this problem.

But for now, consider the following result that confirms our intuition that surrogate experiments should not disturb the causal paths (non-descendents) of the variables that are being analyzed.

Corollary 22. *Let G be the causal diagram, X, $Y \subset V$ be disjoint sets of variables, and $Z \subseteq De(X)_{G_{An(Y)}}$. The causal effect $Q = P(y|do(x))$ is not $z\mathcal{ID}$ from P and $do(Z)$ in G, if Q is not $\mathcal{ID}$ from P in G.*

Proof. The result follows directly from Theorem 31. □

6.4 A Complete Algorithm for $z\mathcal{ID}$

In this section, we propose a simple extension of the ordinary identification algorithms to solve the problem of z-identifiability, which we call ID^z (Fig. 6.4).

We build on previous analysis of identifiability given in (Pea95; KM99; TP02; SP06b; HV06a), and we choose to start with the version called **ID** (SP06b) since the hedge structure is explicitly employed, which will show to be instrumental to prove completeness.

[3]This condition can be rephrased graphically as "There exists no hedge for Q as an edge subgraph in $G_{\overline{Z'}}$."

Before considering the technical results, we explain our strategy and how our version of the algorithm relates to the existent ones for ordinary identifiability.

(i) z-identifiability (sufficiency): Causal relations can be solved in our context through ordinary identifiability or identifiability relying on the experiments performed over Z. The current algorithms already operate on the first part, and they proceed exploring a sequence of equalities in do-calculus based on the C-component decomposition. (The idea is to apply a divide-and-conquer strategy breaking the problem into smaller, more manageable pieces, and then to assemble them back when it is possible.) It turns out that the equalities used by the algorithm are all in the interventional space (between interventional distributions except for the base cases), which is attractive for the $z\mathcal{ID}$ problem since certain interventional distributions Z are already available to use.

For instance, when steps 3 or 4 succeed in their tests and, at the same time, have non-empty intersection with Z, we exploit the common variables, updating the graph and respective data structures accordingly. We then continue solving an ordinary $\mathcal{ID}$ instance but no longer have to identify these variables and they possibly can help in the identifiability of others.

(ii) Non-z-identifiability (necessity): The algorithm proceeds until it is not able to resolve a certain subproblem, which implies the existence of a certain hedge. Note that the given hedge can be different than the one used for $\mathcal{ID}$ in the same graph since the experiments over Z possibly destroyed the original ones. Further, note that to use the given hedge to prove non-$z\mathcal{ID}$ is not immediate since, in the light of Lemma 7, more constraints need to be satisfied in order to support such claim. Still, it is clear that if Z is not involved in the hedge, it can be shown that the two problems coincide. The other cases in which Z has non-empty intersection with the hedge have to be handled more carefully.

We prove next soundness and completeness of ID^z.

Theorem 32 (soundness)**.** *Whenever ID^z returns an expression for $P_x(y)$, it is correct.*

Proof. The result is immediate since the soundness of **ID** was already established (SP06b, Thm. 5), which is inherited by ID^z by construction. Note that adding $Z' \subseteq Z$ as an interventional set and not trying to "identify" it later does not represent a problem, in the $z\mathcal{ID}$ sense, since by assumption we can use the interventional distributions $do(Z)$ in the final expression returned by the procedure. □

Note that the key difference between ID^z and the original **ID** implementation is in steps 3 and 4 in which possibly some $Z' \subseteq Z$ is added as an interventional set, and kept as so until the end of the execution. It is clear that these additions

function $ID^z(y, x, Z, \mathcal{I}, \mathcal{J}, P, G)$
INPUT: x, y: value assignments; Z: variables with interventions available; $\mathcal{I}, \mathcal{J}$: see caption; P: current probability distribution $do(\mathcal{I}, \mathcal{J}, x)$ (observational when $\mathcal{I} = \mathcal{J} = \emptyset$); G: causal graph.
OUTPUT: Expression for $P_x(y)$ in terms of P, P_z or $FAIL(F, F')$.

1 **if** $x = \emptyset$, **return** $\sum_{v \setminus y} P(v)$.
2 **if** $V \setminus An(Y)_G \neq \emptyset$,
return $ID^z(y, x \cap An(Y)_G, Z,$
$\mathcal{I}, \mathcal{J}, \sum_{v \setminus An(Y)_G} P, An(Y)_G)$.
3 Set $Z_w = ((V \setminus (X \cup \mathcal{I} \cup \mathcal{J})) \setminus An(Y)_{G_{\overline{X \cup \mathcal{I} \cup \mathcal{J}}}}) \cap Z$.
Set $W = ((V \setminus (X \cup \mathcal{I} \cup \mathcal{J})) \setminus An(Y)_{G_{\overline{X \cup \mathcal{I} \cup \mathcal{J}}}}) \setminus Z$.
if $(Z_w \cup W) \neq \emptyset$,
return $ID^z(y, x \cup w, Z \setminus Z_w, \mathcal{I} \cup z_w, \mathcal{J}, P, G)$.
4 **if** $\mathcal{C}(G \setminus (X \cup \mathcal{I} \cup \mathcal{J})) = \{S_0, S_1, ..., S_k\}$,
return $\sum_{v \setminus \{y, x, \mathcal{I}\}} \prod_i ID^z(s_i, (v \setminus s_i) \setminus Z,$
$Z \setminus (V \setminus S_i), \mathcal{I}, \mathcal{J} \cup (Z \cap (v \setminus s_i)), P, G)$.
if $\mathcal{C}(G \setminus (X \cup \mathcal{I} \cup \mathcal{J})) = \{S\}$,
5 **if** $\mathcal{C}(G) = \{G\}$, **FAIL**(G, S).
6 **if** $S \in \mathcal{C}(G)$,
return $\sum_{s \setminus y} \prod_{i | V_i \in S} P(v_i | v_G^{(i-1)} \setminus (\mathcal{I} \cup \mathcal{J}))$.
7 **if** $(\exists S') S \subset S' \in \mathcal{C}(G)$,
return $ID^z(y, x \cap S',\ Z,\ \mathcal{I},\ \mathcal{J},$
$\prod_{i | V_i \in C'} P(V_i | V_G^{(i-1)} \cap S', v_G^{(i-1)} \setminus (S' \cup \mathcal{I} \cup \mathcal{J})), S')$.

Figure 6.4: ID^z: Algorithm capable of recognizing $z\mathcal{ID}$; The variables $\mathcal{I}, \mathcal{J}$ represent indices for currently active Z-interventions introduced respectively by steps 3 or 4. Note that P is sensitive to current instantiations of $\mathcal{I}, \mathcal{J}$.

just can represent a benefit in computing the target Q since is always easier to identify a quantity in a subgraph of the original input.

Theorem 33. *Assume ID^z fails to z-identify $P_x(y)$ from P and $do(Z)$ in G (executes line 5). Then there exists $X' \subseteq X$, $Y' \subseteq Y$, $Z', Z'' \subseteq Z$ such that the graph pair G, S returned by the fail condition of ID^z contain as edge subgraphs C-forests F, F' that form a hedge for $P_{x',z'}(y', z'')$.*

Proof. This property is just partly inherited from the original **ID** since we can add $Z' \subseteq Z$ as interventional nodes along the execution of ID^z; we also keep track of $Z'' \subseteq Z$ that are related to $An(Y)$ during the execution of the procedure (to be specified below).

Consider G, Y_f, $\mathcal{I}$ and $\mathcal{J}$ local to the call in which ID^z exited with failure (line 5). It is true that the set Y_f is such that $Z'' = Y_f \cap Z$ and $Y' = Y_f \cap Y$. Let $Z' \subseteq Z$ be the active part of Z in the faulty call, which we kept track through $\mathcal{I} \cup \mathcal{J}$. The condition that triggered failure is that the whole graph was a single C-component. Let R be the root set of G. We can remove a set of directed arrows while keeping the root R such that the resulting F is an R-rooted C-forest.

Similarly to **ID**, note that since $F' = F \cap S$ is closed under descendent and only single directed arrows were removed from S to obtain F', F' is also a C-forest. Now, $F' \cap (X \cup Z') = \emptyset$ and $F' \cap (X \cup Z') \neq \emptyset$, by construction. Also, $R \subseteq An(Y', Z'')_{G_{\overline{X,Z'}}}$ and $Z'' \subseteq An(Y)_{G_{\overline{X,Z'}}}$, by line 2 and 3 of the algorithm. □

Theorem 34 (completeness)**.** *ID^z is complete.*

Proof. By Theorem 33, ID^z failure implies the existence of $X' \subseteq X$, $Y' \subseteq Y$, $Z', Z'' \subseteq Z$, and C-forests F, F' that form a hedge for $P_{x',z'}(y', z'')$. Let us proceed our analysis by cases:

Case $Z' = \emptyset, Z'' = \emptyset$. The construction provided by (SP06b, Corollary 2) can be used here since this case reduces to ordinary identifiability.

Case $Z' = \emptyset, Z'' \neq \emptyset$. Even though Z'' is in the root set of the hedge, and not related to the interventional part $(F \backslash F')$ where the asymmetry in the construction usually resides (to generate inequality in Q), the previous construction have to be used with certain caution, as given by case 1 of Thm. 31.

There is an interesting border subcase when $Y' = \emptyset$. We need to keep track of $\{\mathcal{I}, \mathcal{J}\}$ since if the Z-interventions are added in step 3, we should not be concerned with summing over the assignments of the variables added, but if the Z-interventions are added in step 4, we do have to take care of this case. Note that we would have some hedge in a do-equality in the form $Q = \sum_{z''} P_{x'}(z'') f(x, y, ...)$, in which if $f(.)$ is identifiable and uniformly distributed, Q would equate in both

models and spoil the counter-example. The problem is not difficult to fix, and we just have to create a map for $f()$ that is non-uniform. (See Thm. 31.)

Case $Z' \neq \emptyset, Z'' = \emptyset$. The construction provided in cases 2 and 3 of Thm. 31 were more involved since it was not know a priori which C-factor yielded the "faulty" call. In the ID^z case, we already located the hedge based on its trace, then we can essentially use the same construction to provide a counterexample.

Case $Z' \neq \emptyset, Z'' \neq \emptyset$. The construction in the two previous cases are not incompatible and can be combined to provide a counter-example to this case.

Moreover, the previous constructions were given over the subgraph H of G, and how to extend the counter-example to G is discussed in Theorem 31. □

Corollary 23. *The rules of do-calculus, together with standard probability manipulations are complete for determining z-identifiability of* $P_x(y)$.

Proof. It was already shown (SP06b, Thm. 7) that the operations of **ID** correspond to sequences of standard probability manipulations and application of the rules of do-calculus, which is true by construction for ID^z, the result follows. □

6.5 Conclusions

This chapter was concerned with a variation of the identifiability problem in which experiments can be conducted over a subset of the variables Z in addition to the assumptions embodied in a causal digram G and the statistical knowledge given as a probability distribution. (If Z is an empty set, the two problems coincide.)

We provide graphical and algorithmic conditions for the cases when the causal effect of an arbitrary set of variables on another arbitrary set can be determined uniquely from the available information. Furthermore, we use our results to prove completeness of *do*-calculus in respect to the z-identifiability class. Our results were developed in a non-parametric setting in the tradition of the do-calculus. For a future research direction, it would be interesting to explore how experimental data can aid the identification in the linear case and its relationship with instrumental variables.

This chapter complements the two previous chapters on generalizability in transportability and selection bias. The problem of experimental transportability deals with transferring causal information from an experimental to an observational environment, potentially different from the first. The problem of selection bias deals with extrapolation between an environment in which samples are selected preferentially and one in which no preferential sampling takes place. The extrapolation involved in z-Identification problems takes place between two different regimes; one in which experiments are performed over Z, and one in which future experiments are anticipated over X.

CHAPTER 7

Concluding Remarks

As we approach the age of "big data," researchers are becoming increasingly aware of the fact that traditional statistical techniques, including those based on machine learning, must be enriched with two additional ingredients:

1. the ability to integrate data from multiple, heterogeneous sources, and
2. the ability to distinguish causal from associational relationships.

The former becomes essential when mixing experimental and observational studies while the latter is necessary when constructing explanations for the data observed. This thesis develops a formal theory for handling these two components simultaneously. It builds on the modern theory of causation to develop a theoretical framework for understanding, representing, and algorithmizing causal generalizations in a mixture of experimental and observational studies.

The concepts and tools that emerge from this framework are applicable to a broad range of common, yet seemingly disparate problems in both AI and the empirical sciences. This thesis puts these problems under one theoretical umbrella, which include: transportability, selection bias, and general identification.

7.1 Contributions

Techniques for data analysis are usually dichotomized into two categories, experimental and observational, which are studied in isolation. We have shown that such dichotomy need not constrain the next generation of data science. Experimental studies (where interventions are feasible) and observational studies (where interventions are not allowed) are but two extremes of a rich spectrum of research designs that generate the bulk of the data available in practical, large scale situations. In typical medical explorations, for example, data from multiple observations and experiments are collected, coming from distinct experimental setups, different sampling conditions, and heterogeneous populations. This thesis develops a mathematical framework for handling such data-intensive explorations and, in this way, should become an indispensable tool in meeting the challenges presented by the "big data" generation.

Next, we briefly describe the research questions addressed in this thesis, followed by a summary of the technical solutions obtained.

Problem 1. Transportability (generalizing experimental findings across settings, populations, or domains). How can one reuse causal information acquired by experiments in one setting to answer causal queries in another, possibly different setting where only limited observations and experiments can be collected?

This problem has a long tradition in data-intensive fields, since experiments are invariably conducted with the intent of being generalized to an environment that is related to but different from the original experimental setup. Special instances of this problem are known in the literature under rubrics such as "external validity," "meta-analysis," "quasi-experiments," and "heterogeneity." For instance, a researcher may perform experiments on chimpanzees and aim to generalize the conclusions to human beings; or an engineer may train a robot in a simulator with the hope that it will perform well in the field. We asked: what mathematical principles support this leap of generalization? We showed that a key ingredient necessary to formalize this type of question is to identify areas of commonalities and disparities between the two settings (in the previous examples, species and environments). Given a coarse description of these areas, we proved it is possible to formally decide what knowledge is or is not transportable across settings, and if so, how. We then introduced a formal language for expressing qualitative differences between settings and, using this representation, we reduced the problem of transportability to an exercise in symbolic calculus. We further developed complete syntactic, graphical, and algorithmic conditions for deciding transportability and constructing the transported knowledge based on the available pieces of empirical evidence.

Problem 2. Selection Bias (generalizing statistical findings across sampling conditions). How can knowledge from a sampled subpopulation be generalized to the entire population when the sampling process is not random but discriminatory, depending on other variables in the analysis?

Selection bias is a threat to many data analyses and has been studied extensively in both parametric and semi-parametric forms. A non-random sampling process entails a distortion in the sample's proportions, and since the sample is no longer a faithful representation of the population, biased estimates will be produced regardless of how many samples were collected. We derived general, non-parametric conditions for deciding whether conditional distributions are recoverable from selection bias without resorting to external information. We then considered the case in which some external information might be available (e.g., a pilot study or census data), and derived sufficient conditions for deciding whether conditional distributions can be recovered from selection bias. We further analyzed selection bias in causal settings, and augmented the backdoor criterion to

recover causal effects under both selection and confounding conditions.

3. General (Experimental) Identification (generalizing experimental findings across experimental conditions in the same domain). How can some experimental knowledge be used as substitute for other experiments that are too difficult, expensive, or unethical to execute in practice?

In many cases, the variable of interest is not amenable to manipulation, while others (called auxiliary) are more easily accessible to experimentation. For instance, one can conduct an experiment on diet to estimate the effects of cholesterol level on heart attacks since it is infeasible to control cholesterol level directly. This setting generalizes the instrumental variable case, which relies on the parametric nature of relationships to derive target effects. This thesis considered a more involved problem in which linearity (or monotonicity) is not assumed, and the analysis is conducted in the most general non-parametric form. We derived a necessary and sufficient graphical condition for identifying causal effects in terms of auxiliary experiments. We proved the completeness of do-calculus for recognizing such identifiability conditions (i.e., if a causal effect cannot be expressed in terms of the auxiliary experiments by repeated application of the rules of do-calculus, such an expression does not exist). We developed a procedure for deciding general identification and for constructing a target estimator (whenever one exists).

Considering these three classes of problems, the theory developed in this thesis is general, in the sense that it takes as input any arbitrary set of assumptions and decides whether the specific instance admits solution. Moreover, the theory of causal generalizability delineates the formal boundary between computable and non-computable effects, and it provides conditions (algebraic, graphical, and algorithmic) for identifying which pieces of knowledge need to be collected in each environment to achieve a consistent estimate of the desired effects (when computable).

7.2 Future Work

There are other generalization tasks that were not investigated in this thesis and could benefit from the results obtained here. We next list a few of these tasks.

1. Automated Scientific Discovery. One of the challenges of our time is to make the best possible use of the vast amounts of data that has been generated, and ultimately we would like to automate the process of scientific discovery. Currently, we have an understanding and a formal language for expressing the causal and generalization principles pertaining to the core scientific method, so it will be interesting to explore how these principles can be integrated into intelligent systems to automate aspects of the process of scientific discovery. Eventually, we

would like these systems to fully support scientific inquiry, i.e., they should be able to help design and conduct experiments, collect and analyze data, test and refine theories, and then propose new research directions and create knowledge. This is a very challenging task, but it is reasonable to expect that with the advent of a causal language endowed with generalization capabilities, progress can be made towards this goal.

2. Domain Adaptation. The results developed in this work for transporting causal effect relationships across domains seems to have direct application to the sub-field of machine learning known as *domain adaptation*. While the machine learning literature is seriously concerned about discrepancies between training and test environments (DM06; Sto09), it has focused almost exclusively on predictive or classification tasks as opposed to effect-learning tasks. Even in these tasks machine learning researchers have rarely allowed a priori causal knowledge to guide the learning process and, as a result, have not sought theoretical guarantees in the form of sufficient conditions under which discrepancies between the training and test environments can be circumvented, or necessary conditions without which bias will persist regardless of sample size. Some work using representation equivalent to the one introduced here has been initiated, leveraging knowledge about invariances of the data-generating model across domains (SJP12; ZSM13), but additional work needs to be done to move the literature towards more general modalities of learning.

3. Surrogate Endpoints. There is a growing and unsettled literature on the problem of "surrogate endpoints" (Pre89; FGS92; BMB00), which considers a randomized clinical trial where one seeks a variable that would allow good predictability of an outcome for both treatment and control. In the words of (EH89): "investigators use surrogate endpoints when the endpoint of interest is too difficult and/or expensive to measure routinely and when they can define some other, more readily measurable, endpoint, which is sufficiently well correlated with the first to justify its use as a substitute." It is generally acknowledged in the literature that strong correlation is not sufficient for surrogacy, and the problem is still awaiting a formal characterization. There is a lot of interest in this problem because, for instance, it is usually infeasible to conduct follow-ups on subjects of a clinical trial for many years. We have initiated some exploration of how the theory of transportability can assist in the identification of valid surrogates in complex networks of cause-effect relationships (PB11a).

APPENDIX A

Proofs for Chapter 3

Theorem 9. *Let G be a selection diagram. Then for any node Y, the direct effect $P^*_{Pa(Y)}(y)$ is transportable if there is no subgraph of G which forms a Y-rooted sC-tree.*

Proof. We known from (Tia02, Theorem 22) that whenever there exists no subgraph G_T of G satisfying all of the following: (i) $Y \in T$; (ii) G_T has only one c-component, T itself; (iii) All variables in T are ancestors of Y in G_T, the direct effect on Y is identifiable, as sC-trees are structures of this type. Further (SP06b, Theorem 2) showed that the same holds for C-trees, which also implies the inexistence of a sC-trees. Since such structure does not show up in G, the target quantity is identifiable, and hence transportable.

It remains to show that the same holds whenever there exists a subgraph that is a C-tree and in which no S node points to Y, i.e., there is no Y-rooted sC-tree at all. It is true that $(S \perp\!\!\!\perp Y|Pa(Y))_{G_{\overline{Pa(Y)}}}$, given that all directed paths from S to Y are closed. This follows from the following facts: 1) all paths from S passing through Y's ancestors were cut in $G_{\overline{Pa(Y)}}$; 2) all bidirected paths were also closed given that the conditioning set contains only root nodes, and a connection from S must pass through at least one collider; 3) transportability does not depend on descendants of Y (by argument similar to (Tia02, Lemma 9)). Thus, it follows that we can write $P^*_{Pa(Y)}(Y) = P_{Pa(Y)}(Y|S) = P_{Pa(Y)}(Y)$, concluding the proof. □

Corollary 4. *Let G be a selection diagram. Then for any node Y, the direct effect $P^*_{Pa(Y)}(y)$ is transportable if there is no S node pointing to Y.*

Proof. Follows directly from Theorem 9. □

Lemma 8. *The exclusive OR (XOR) function is commutative and associative.*

Proof. Follows directly from the definition of the XOR function. □

Remark 1. The construction given below is a strict generalization of Theorem 8, and it is useful because it will provide a simplified construction of the same,

and also set the tone for proofs of generic graph structures which will in the sequel show to be instrumental in proving non-transportability in arbitrary structures.

Theorem 10. *Let G be a Y-rooted sC-tree. Then the effects of any set of nodes in G on Y are not transportable.*

Proof. The proof will proceed by constructing a family of counterexamples. For any such G and any set X, we will construct two causal models M_1 and M_2 that will agree on $\langle P, P^*, I\rangle$, but disagree on the interventional distribution $P^*_x(y)$.

Let the two models M_1, M_2 agree on the following features. All variables in $U \cup V$ are binary. All exogenous variables are distributed uniformly. All endogenous variables except Y are set to the bit parity (sum mod 2) of the values of their parents. The two models differ in respect to Y's definition. Consider the function for Y, $f_Y : U, Pa(Y) \rightarrow Y$ to be defined as follows:

$$\begin{cases} M_1 : Y = \big((pa(Y) \oplus u) \oplus s\big) \\ M_2 : Y = \big((pa(Y) \oplus u) \vee s\big) \end{cases}$$

Lemma 9. *The two models agree in the distributions $\langle P, P^*, I\rangle$.*

Proof. Since the two models agree on $P(U)$ and all functions except f_Y, it suffices to show that f_Y maintains the same input/output behavior in both models for each domains.

Subclaim 1: Let us show that both models agree in the observational and interventional distributions relative to domain Π, i.e., the pair $\langle P, I\rangle$. The index variable S is set to 0 in Π, and f_Y evaluates to $(pa(Y) \oplus u)$ in both models, which proves the subclaim.

Subclaim 2: Let us show that both models agree in the observational distribution relative to Π^*, i.e., P^*. The index variable S is set 1 in Π^*, and f_Y evaluates to $((pa(Y) \oplus u) \oplus 1)$ in M_1, and 1 in M_2. Since the evaluation in M_1 can be rewritten as $\neg((pa(Y) \oplus u)$, it remains to show that $(pa(Y) \oplus u)$ always evaluates to 0.

This fact is certainly true, consider the following observations: a) each variable in U has exactly two endogenous children; b) the given tree has Y as the root; c) all functions are *XOR* – these imply that Y is computing the bit parity of the sum of all U nodes, which turns out to be even, and so evaluates to 0 and proves the subclaim. □

Lemma 10. *For any set X, $P_1(Y|do(X), S = 1) \neq P_2(Y|do(X), S = 1)$.*

Proof. Given the functional description and the discussion in the previous Lemma, the function f_Y evaluates always to 1 in M_2.

Now let us consider M_1. Note that performing the intervention and cutting the edges going toward X creates an asymmetry on the sum of the bidirected edges departing from U, and consequently in the sum performed by Y. It will be the case that some U' will appear only once in the expression of Y. Therefore, depending on the assignment $X = x$, we will need to evaluate the sum (mod 2) over U' in Y or its negation, which given the uniformity of the distribution of U will yield $P_1(Y|do(X), S = 1) = 1/2$ in both cases. □

By Lemma 1, Lemmas 9 and 10 together prove Theorem 10. □

Corollary 5. *Let G be a selection diagram, let X and Y be set of variables. If there exists a node W which is an ancestor of some node $Y \in Y$ and such that there exists a W-rooted sC-tree which contains any variables in X, then $P^*_x(y)$ is not transportable.*

Proof. Fix a W-rooted sC-tree T, and a path p from W to Y. Consider the graph $p \cup T$. Note that in this graph $P^*_x(Y) = \sum_w P^*_x(w)P^*(Y|w)$. From the last Theorem $P^*_x(w)$ is not transportable, it is now easy to construct $P^*(Y|W)$ in such a way that the mapping from $P_x(W)$ to $P_x(Y)$ is one to one, while making sure all distributions are positive. □

Remark 2. The previous results comprised cases in which there exist sC-trees involved in the non-transportability of Y – i.e., Y or some of its ancestors were roots of a given sC-tree. In the problem of identifiability, the counterpart of sC-trees (i.e., C-trees) suffices to characterize non-identifiability for singleton Y. But transportability is more subtle and this is not the case here – it not only depends on X and Y "locations" in the graph, but also the relative position of the S-nodes. Consider Figures 3.8 and A.1(a) (called *sp*-graph). In these graphs there is no sC-tree but the effect of X on Y is still non-transportable.

The main technical subtlety here is that in sC-trees, a S-node combines its effect with a X-node intersecting in the root node (considering only the bidirected edges), which is not the case for non-transportability in general. Note that in the graphs in Figure 3.8, and the *sp*-graph, the nodes S and X intersect first through ordinary edges and meet through bidirected edges only on the Y node. This implies a certain "asynchrony" because, in the structural sense, the existence of a S-node implies a difference in the structural equations between domains, but only this difference does not imply non-transportability (for instance, $P^*_x(z)$ is transportable in the *sp*-graph even though the equations of Z being different in both models).

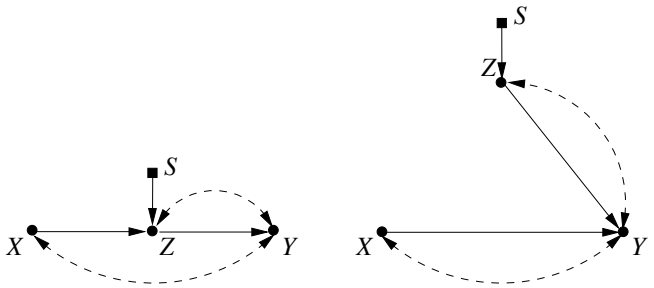


Figure A.1: Selection diagrams in which $P(y|do(x))$ is not transportable, there is no sC-tree but there is a sC-forest. These diagrams will be used as basis for the general case; the first diagram is named *sp*-graph and the second one *sb*-graph.

The key idea to produce a proof for non-transportability in these cases is to keep the effect of S-nodes after intersecting with X "dormant" until they reach the target Y, and then manifest. We implement this idea in the next two proofs, which can be seen as base cases, and should pavement the way for the most general problem.

Theorem 35. *$P^*_x(y)$ is not transportable in the sp-graph (Fig. A.1(a)).*

Proof. We will construct two causal models M_1 and M_2 compatible with the *sp*-graph that will agree on $\langle P, P^*, I\rangle$, but disagree on the interventional distribution $P^*_x(y)$.

Let us assume that all variables in $U \cup V$ are binary, and let U_1 be the common cause of X and Y, U_2 be the common cause of Z and Y, and U_3 be the random disturbance exclusive to Z. Let M_1 and M_2 be defined as follows:

$$M_1 = \begin{cases} X = U_1 \\ Z = \Big(\big((X \oplus U_2 \oplus 1) \oplus U_3\big) \vee S\Big) \oplus \Big(S \wedge (X \oplus U_2)\Big) \\ Y = Z \oplus U_1 \oplus U_2 \end{cases}$$

and:

$$M_2 = \begin{cases} X = U_1 \\ Z = \Big(\big((U_2 \oplus 1) \oplus U_3\big) \vee S\Big) \oplus \big(S \wedge U_2\big) \\ Y = Z \oplus U_2 \end{cases}$$

Both models agree in respect to $P(U)$, which is defined as follows: $P(U_1) = P(U_2) = P(U_3) = 1/2$.

Lemma 11. *The two models agree in the distributions $\langle P, P^*, I\rangle$.*

Proof. **Subclaim 1**: Let us show that both models agree in the observational and interventional distributions relative to domain Π, i.e., the pair $\langle P, I\rangle$. In both models X has the same expression, which entails the same (uniform) probabilistic behavior in both cases. The index variable S is set to 0 in Π, and Z evaluates to $(X \oplus U_2 \oplus 1 \oplus U_3)$ in M_1 and $(U_2 \oplus 1 \oplus U_3)$ in M_2. Clearly, for any value of $X = x$, since U is the same and uniformly distributed in both models, we obtain the same (uniform) input/output probabilistic behavior in M_1 and M_2 (note that U_2, U_3 can freely vary independently of X). In similar way, Y evaluates to $(1 + U_3)$ in both models, which entails the same (uniform) input/output probabilistic behavior in both models. In regard to $do(X = x)$, it is clear that Z did not depend (probabilistically) on the specific value of X, and so the equality between both models follows. For the case when we have $do(Z = z)$, Y evaluates to $(Z \oplus U_1 \oplus U_2)$ in M_1 and $(Z \oplus U_2)$ in M_2, and given the uniformity of U, they preserve the same (uniform) input/output probabilistic behavior. (For a more elaborated argument, see Theorem 5 below.)

Subclaim 2: Let us show that both models agree in the observational distribution P^* relative to Π^*. The index variable S is set 1 in Π^*, f_Z evaluates to $(X \oplus U_2 \oplus 1)$ in M_1, and $(U_2 \oplus 1)$ in M_2. Again, for any value of X, together with the uniformity of U, we obtain the same (uniform) input/output probabilistic behavior in both models (note again that U_2 can freely vary independently of variations of X, and so Z). Further, f_Y evaluates to 1 in both models, which yields the same (uniform) input/output behavior in both models. (To guarantee positivity, we can apply the trick of making a new $f'_Y()$ such that $f'_Y()$ returns 0 half the time, and f_Y the other half (i.e., set $f'_y() = [f_y() \wedge C]$, where C is a fair coin.) □

Lemma 12. *There exist values of X, Y such that $P_1(Y|do(X), S = 1) \neq P_2(Y|do(X), S = 1)$.*

Proof. Fix $X = 1, Y = 1$. First notice that f_Z evaluates to U_2 in M_1 and $(U_2 \oplus 1)$ in M_2. Given that U_2 is uniformly distributed, both quantities coincide (and they represent the effect of X on Z, which is transportable in G). Now the evaluation of f_Y in M_1 reduces to U_1, while it reduces to 1 in M_2, which show disagreement and finishes the proof of this Lemma. □

By Lemma 1, Lemmas 11 and 12 together prove Theorem 35. □

Remark 3. There exists a different sort of asymmetry in the case of Fig. A.1(b) (called *sb*-graph), and the nodes X and S do not intersect before meeting Y – i.e., they have disjoint paths and Y lies precisely in their intersection.

Still, this case is not the same of having a sC-tree because in sb-graphs we need to keep the equality from the S nodes to Y until S intersects X on Y. Employing a similar construct as in the sp-graph, we keep the effect of S dormant until it reaches Y and then emerges.

Theorem 36. *$P^*_x(y)$ is not transportable in the sb-graph (Fig. A.1(b)).*

Proof. We construct two causal models M_1 and M_2 compatible with the sb-graph that will agree on $\langle P, P^*, I\rangle$, but disagree on the interventional distribution $P^*_x(y)$.

Let us assume that all variables in $U \cup V$ are binary, and let U_1 be the common cause of X and Y, U_2 be the common cause of Z and Y, and U_3 be the random disturbance exclusive to X. Let M_1 and M_2 agree with the following definitions:

$$M_1, M_2 = \begin{cases} X = U_1 \\ Z = \big((U_3 \oplus U_2 \oplus 1) \vee S\big) \oplus \big(S \wedge U_2\big)\big) \end{cases}$$

and disagree in respect to Z as follows:

$$\begin{cases} M_1 : Y = Z \oplus U_2 \\ M_2 : Y = X \oplus Z \oplus U_1 \oplus U_2 \end{cases}$$

Both models also agree in respect to $P(U)$, which is defined as follows: $P(U_1) = P(U_2) = P(U_3) = 1/2$.

Lemma 13. *The two models agree in the distributions $\langle P, P^*, I\rangle$.*

Proof. **Subclaim 1**: Let us show that both models agree in the observational and interventional distributions relative to domain Π, i.e., the pair $\langle P, I\rangle$. The index variable S is set to 0 in Π, and $\{X, Z\}$ are defined in the same way in both models, and so it suffices to analyze Y, which in this case evaluates to $(U_3 \oplus 1)$ in both models, preserving the same (uniform) probabilistic behavior. Given that, it is not difficult to see that both models also evaluate in the same way when considering the interventions in I.

Subclaim 2: Let us show that both models agree in the observational distribution P^* relative to Π^*. The index variable S is set 1 in Π^*, given that $\{X, Z\}$ are defined in the same way in both models, together with the uniformity of U make them evaluate in the same way in both models, and Y evaluates to 1 in both models. (As in Lemma 11, the same trick to make the distribution positive could be applied here.) □

Lemma 14. *There exist values of X, Y such that $P_1(Y|do(X), S = 1) \neq P_2(Y|do(X), S = 1)$.*

Proof. Fix $X = 1, Y = 1$. First notice that f_Z evaluates to $(U_2 \oplus 1)$ in both models, and the evaluation of f_Y in M_1 reduces to 1, while it reduces to U_1 in M_2. It follows that in M_1, f_Y evaluates to 1 with probability 1, while in M_2 it evaluates to 1 with probability $P(U_1 = 1)$, which disagree by construction, finishing the proof of this Lemma. □

By Lemma 1, Lemmas 13 and 14 together prove Theorem 36. □

Remark 4. There are two complementary components to forge a general scheme to prove arbitrary non-transportability. First, the construct of Theorem 10 shows how to prove non-transportability for general structures such as sC-trees. In the sequel, the specific proofs of non-transportability for the *sp*-graph (Theorem 35) and *sb*-graph (Theorem 36) partition the possible interactions between X, S and Y. In the former, X and S intersect before meeting with Y, while in the latter they have disjoint paths and Y lies in their intersection. In the sequel, the proof for the general case combines these analyses, which we show below.

Theorem 11. *Assume there exist F, F' that form a s-hedge for $P_x^*(y)$ in Π and Π^*. Then $P_x^*(y)$ is not transportable from Π to Π^*.*

Proof sketch. We first consider counterexamples with the induced graph $H = De(F)_G \cap An(Y)_{G_{\overline{X}}}$, and assume, without loss of generality, that H is a forest. We construct two causal models M_1 and M_2 that will agree on $\langle P, P^*, I\rangle$, but disagree on the interventional distribution $P_x^*(y)$.

Let F be an R-rooted sC-forest, let V' be the set of observable variables and U' be the set of unobservable variables in F. Let us assume that all variables in $U' \cup V'$ are binary. Call W the set of variables pointed by S-nodes in F', which by the definition of sC-forest is guaranteed to be non-empty.

In model 1, let each $V_i \in V' \setminus W$ compute the bit parity of all its observable and unobservable parents (i.e., $f_i^{(1)} = \oplus(\bigcup_{V_j \in Pa_i} V_j)$, where the xor is applied for each element of the set and the result computed so far), while in model 2, let V_i compute the bit parity of all its parents except that any node in F' disregards the parents values if the parent is in F (i.e., $f_i^{(2)} = \oplus(\bigcup_{V_j \in Pa_i \cap F'} V_j)$ if V_i is in F', and $f_i^{(2)} = f_i^{(1)}$, otherwise).

Define $W \in W$ as follows:

$$\begin{cases} M_1 : W = \Big(\big(f_w^{(1)} \oplus U_w^*\big) \vee S\Big) \oplus \Big(S \wedge \big(1 \oplus f_w^{(1)}\big)\Big) \\ M_2 : W = \Big(\big(f_w^{(2)} \oplus U_w^*\big) \vee S\Big) \oplus \Big(S \wedge \big(1 \oplus f_w^{(2)}\big)\Big) \end{cases}$$

where f_w is constructed in similar way as f_i in M_1 and M_2 above, and U_w^* is an additional fair coin exclusively pointing to W. Let us call U_w the collection of such coins. Furthermore, let us assume that each $U_i \in \{U' \setminus U_w\}$ is also a fair coin (i.e., $P(U_i) = 1/2$).

Lemma 15. *The two models agree in the distribution of P^* and there exists a value assignment x for X such that $P_1(Y|do(x), S = 1) \neq P_2(Y|do(x), S = 1)$.*

Proof. For $S = 1$, the result follows directly since the systems of equations in both models reduce to the construction given in Theorem 4 at (SP06b). □

Lemma 16. *The two models agree in the distributions $\langle P, I \rangle$.*

Proof. Let us show that both models agree in the observational distribution P relative to domain Π. The selection variable S is set to 0 in Π, and note that both systems are the same as in Π^* except that now each variable $W \in W$ has an extra variable U_w^* pointing to it that should be taken into account in W's evaluation, and in turn in the whole system.

We have a forest over the endogenous nodes and all functions compute the bit parity of the value of their parents, and so we can view each node as computing the sum mod 2 of its exogenous ancestors in H. We want to show that the distribution of each family is equally likely for each possible assignment (i.e., $P(v_i|pa_i) = 1/2$, for all v_i, pa_i).

Let us partition the analysis in two cases. First consider the case of $V_i \in R$ in which there exists a S-node in the respective sC-tree. Note that that the evaluation of V_i relies only on the value of $U_w^* \in U_w$ in its respective tree since $U \in \{U' \setminus U_w\}$ has an even number of endogenous children in F, and it is counted twice, so evaluates to zero (i.e., it does not affect V_i's evaluation). For now, let us assume that there is only one U_w^* that affects the evaluation of V_i. Given the uniformity of U_w^*, it suffices to show that U_w^* can vary independently for any configuration of the parents of V_i.

For any configuration of $U' = (U_1 = u_1, ..., U_w^* = u_w^*, ...)$, consider the corresponding evaluation of $Pa_i = pa_i$, and also $V_i = u_w^*$. We want to show that it is possible to flip the current value of U_w^* from u_w^* to $\neg u_w^*$ while preserving the parents' evaluation pa_i. Assume this is not so. This implies that the evaluation of Pa_i and V_i count the same U's, contradiction.

To see why, consider $Pa_i^* \subseteq Pa_i$ the set of parents of V_i that are descendents of U_w^*. Now, for each of these parents flip the minimum number of variables from $U \setminus U_w$, and call this set U^*. (Note that this is always possible since we need at most one U for each parent, which should exist by construction of sC-forest.) Now, make $U_w^* = \neg u_w^*$, and note that $Pa_i = pa_i$ since flipping the values of U^*

compensates the flip of U_w^*. But it is also true now that V_i evaluates to $\neg u_w^*$ since, in the same way as before, all other variables in $\{U \setminus U_w\}$ are cancelled out in V_i's evaluation, including the ones in U^*. This proves the claim.

Consider the following two facts: **Subclaim 1**: Let X and Y be two binary variables such that $P(X = x) = p \neq 1/2$ and $P(Y = y) = q = 1/2$. Then the probabilistic input/output behavior of $Z = XOR(X, Y)$ is the same of Y. The variable $Z = 1$ whenever $\{(X = 1, Y = 0), (X = 0, Y = 1)\}$, which happens with probability $pq + (1 - p)(1 - q)$. Since $q = 1/2$, the expression reduces to $p * 1/2 + (1 - p) * 1/2 = 1/2$. **Subclaim 2**: Let X and Y be two binary variables such that $P(X = x) = P(Y = y) = p = 1/2$. Then the probabilistic input/output behavior of $Z = XOR(X, Y)$ is the same of X (or Y). This follows directly from Subclaim 1. It is clear that if there are multiple nodes from U_w in the evaluation of V_i, the same construction is also valid given the subclaim above. It is also not difficult to generalize this argument to consider root set that are not singleton, including roots in which there are not S-nodes as ancestors.

Finally, let us consider the case of $V_i \in \{F \setminus R\}$. It suffices to show that the function from $U' \setminus U_w$ to $V' \setminus R$ is 1-1 when we fix $U_w = u_w$. We use the same argument as Shpitser. Assume this is not so, and fix two instantiations of $U' \setminus U_w$ that map to the same value of $V' \setminus R$, and differ by the set $U^* = \{U_1, ..., U_k\}$. Since the bidirected edges form a spanning tree, there exists V^* with an odd number of parents in U^* (and were not in R, by construction). Order them topologically and let the topmost be called X. Note that if we flip all values in U^*, the value of X will also flip, contradiction. Given the uniformity of U', the claim follows. We can put this together with the previous claim, and the result follows. We can add fair coins as the input to all other variables outside F, which will imply the claim for the whole graph G.

Regarding the equality between I, note that given that the equality of both models holds for P, and removing edges due to interventions will just make some nodes from $U' \setminus U_w$ to have an odd number of children, it it not difficult to see based on the previous argument that this just creates more variables that are free to vary, which will entail the same probabilistic uniform behavior in both models. Another way to see this fact is to consider the new exogenous variables from $\{U \setminus U_w\}$ that have only one children after the intervention as analogous to U_w^*, and so the same argument follows. (For more details, see Appendix B.) □

Finally, Lemma 1 together with Lemmas 15 and 16 prove Theorem 11. □

Theorem 12 (soundness)**.** *Whenever **sID** returns an expression for $P_x^*(y)$, it is correct.*

Proof. Noting that the selection diagram inputted to **sID** is also a causal diagram

over Π^*, and trivial transportability is equivalent to identifiability in Π^*, the correctness of the identifiability calls were already established elsewhere (HV06a; SP06b).

It remains to show the correctness of the test in line 6 of **sID**. First note that, by construction, X' in each local call is always a set of pre-treatment covariates. But now the correctness follows directly by S-admissibility of X' together with Corollary 2. More specifically, note that the effect Q^* in each local call that uses line 6 can be expressed in its expanded form (using a typical C-component decomposition), and given that the independence imposed by S-admissibility holds, together with the fact that both populations share the same causal graph G, allow that the functions of Π^* to be replaced with the respective functions in Π, which implies the result. □

Remark 5. The next results are similar to the identification counterparts given in (TP02) and (SP06a).

Theorem 13. *Assume **sID** fails to transport $P^*_x(y)$ (executes line 7). Then there exists $X' \subseteq X$, $Y' \subseteq Y$, such that the graph pair D, C_0 returned by the fail condition of **sID** contain as edge subgraphs sC-forests F, F' that form a s-hedge for $P_{x'}(y')$.*

Proof. Before failure **sID** evaluated false consecutively at line 5 and 6, so D local to this call is a sC-component, and let R be its root set. We can remove some directed arrows from D while preserving R as root, yielding a R-rooted sC-forests F. Since by construction $F' = F \cap C_0$ is closed under descendants and only directed arrows were removed, both F, F' are sC-forests. Also by construction, $R \subset An(Y)_{D_{\overline{X}}}$ together with the fact that X and Y from the recursive call are clearly subsets of the original input, finish the proof. □

Corollary 6 (completeness)**.** ***sID*** *is complete.*

Proof. The result follows from Theorem 13 where $P_{x'}(y')$ is not transportable in H. But now, it is easy to add the remaining variables from G, making them independent of H (e.g., as random coins). So, the models in the counterexample induce G, and witness the non-transportability of $P_x(y)$. □

Corollary 7. *$P^*_x(y)$ is transportable from Π to Π^* in G if and only if there is not s-hedge for $P_{x'}(y')$ in G for any $X' \subseteq X$ and $Y' \subseteq Y$.*

Proof. Follows directly from the previous Corollary. □

Theorem 14. *The rules of do-calculus, together with standard probability manipulations are complete for establishing transportability of all effects of the form* $P^*_x(y)$.

Proof. It was shown elsewhere (SP06a) that the steps of **sID** but line 6 correspond to sequences of standard probability manipulations and applications of the rules of do-calculus. The line 6 is constituted by a conditional independence judgement, and standard probability operations for the replacement of the functions based on the invariance allowed by the S-admissibility of the local X' in each recursive call (as discussed above in the proof of correctness). □

Corollary 8. *Theorem 7 is not complete.*

Proof. Figure 3.6(c) demonstrates a selection diagram in which the relation $R = P^*(y|do(x))$ is transportable, but Theorem 7 is not capable of recognizing it.

Let us test the applicability of each of its conditions:

Step 1. R is not trivially transportable due to the confounding arc $X \leftrightarrow Z$ due to Tian's identifiability criterion (TP02);

Step 2. There is no S-admissible set because the confounding arc $V \leftrightarrow Y$ and Verma's inducing path condition (VP90);

Step 3. There is no set W which makes $(X \perp\!\!\!\perp Y|W)$ to hold, this is due to the confounding arc $X \leftrightarrow Y$;

Since there is no remaining actions to be taken, the algorithm exits without returning any expression. □

Next, we derive the transport formula for the causal effect in the model of Fig. 3.4(d) based on Theorem 7 (i.e., eq. (3.14)),

$$\begin{aligned} P^*(y|do(x)) &= P(y|do(x),s) \\ &= \sum_z P(y|do(x),s,z)P(z|do(x),s) \\ &= \sum_z P(y|do(x),z)P(z|do(x),s) \\ &\qquad \big(\text{2nd cond. of Theorem 7, } S\text{-admissibility of} \\ &\qquad Z \text{ of } CE(X,Y)\big) \end{aligned}$$

$$
\begin{aligned}
&= \sum_z P(y|do(x),z) \sum_w P(z|do(x),w,s) P(w|do(x),s) \\
&= \sum_z P(y|do(x),z) \sum_w P(z|w,s) P(w|do(x),s) \\
&\qquad \big(\text{3rd cond. of Theorem 7, } (X \perp\!\!\!\perp Z|S,W)\big) \\
&= \sum_z P(y|do(x),z) \sum_w P(z|w,s) P(w|do(x)) \\
&\qquad \big(\text{2nd cond. of Theorem 7, } S\text{-admissibility of} \\
&\qquad \text{the empty set } \{\} \text{ of } CE(X,W)\big) \\
&= \sum_z P(y|do(x),z) \sum_w P^*(z|w) P(w|do(x)) \qquad \text{(A.1)}
\end{aligned}
$$

In turn, we show the derivation of the transport formula for the causal effect in the model of Fig. 3.5 (Eq. (3.15)):

$$
\begin{aligned}
P^*(y|do(x)) &= P(y|do(x),s,s') = \sum_z P(y|do(x),s,s',z) P(z|do(x),s,s') \\
&= \sum_z P(y|do(x),z) P(z|do(x),s,s') \\
&\qquad \big(\text{2nd cond. of Theorem 7,} \\
&\qquad\quad S\text{-admissibility of } Z \text{ of } CE(X,Z)\big) \\
&= \sum_z P(y|do(x),z) \sum_w P(z|do(x),s,s',w) P(w|do(x),s,s') \\
&= \sum_z P(y|do(x),z) \sum_w P(z|s,s',w) P(w|do(x),s,s') \\
&\qquad \big(\text{3rd cond. of Theorem 7, } (X \perp\!\!\!\perp Z|S,S',W)\big) \\
&= \sum_z P(y|do(x),z) \sum_w P(z|s,s',w) \sum_t P(w|do(x),s,s',t) P(t|do(x),s,s') \\
&= \sum_z P(y|do(x),z) \sum_w P(z|s,s',w) \sum_t P(w|do(x),t) P(t|do(x),s,s') \\
&\qquad \big(\text{2nd condition of Theorem 7,} \\
&\qquad\quad S\text{-admissibility of } T \text{ on } CE(X,W)\big) \\
&= \sum_z P(y|do(x),z) \sum_w P(z|s,s',w) \sum_t P(w|do(x),t) P(t|s,s') \\
&\qquad \big(\text{1st condition of Theorem 7,} \\
&\qquad \text{3rd rule of } do\text{-calculus, } (X \perp\!\!\!\perp T|S,S')_{G_{\bar{X}}}\big) \\
&= \sum_z P(y|do(x),z) \sum_w P^*(z|w) \sum_t P(w|do(x),t) P^*(t) \qquad \text{(A.2)}
\end{aligned}
$$

APPENDIX B

Proofs for Chapter 4

We first show the two models relative to Fig. 4.3 in which Eqs. 4.6 and 4.7 are satisfied. Let V be the set of observable variables and U be the set of unobservable variables in $\mathcal{D}$. Let us assume that all variables in $U \cup V$ are binary. Let $U_1, U_2 \in U$ be the common causes of X and Y and Z and Y, respectively; let $U_3, U_4 \in U$ be the random disturbances exclusive to Z and Y, respectively, and $U_5, U_6 \in U$ be extra random disturbances exclusive to Y. Let S_a and S_b index the model in the following way: the tuples $\langle S_a = 1, S_b = 0\rangle$, $\langle S_a = 0, S_b = 1\rangle$, $\langle S_a = 0, S_b = 0\rangle$ represent domains π_a, π_b, and π^*, respectively. Define the two models as follows:

$$M_1 = \begin{cases} X = U_1 \\ Z = U_2 \oplus (U_3 \wedge S_a) \\ Y = ((X \oplus Z \oplus U_1 \oplus U_2 \oplus (U_4 \wedge S_b)) \\ \qquad \wedge U_5) + (\neg U_5 \wedge U_6) \end{cases} \quad M_2 = \begin{cases} X = U_1 \\ Z = U_2 \oplus (U_3 \wedge S_a) \\ Y = ((Z \oplus U_2 \oplus (U_4 \wedge S_b)) \\ \qquad \wedge U_5) \oplus (\neg U_5 \wedge U_6) \end{cases}$$

Both models agree in respect to $P(U)$, which is defined as $P(U_i) = 1/2$, $i = 1, ..., 6$. It is not difficult to evaluate these models and note that the constraints given in Eqs. (4.29) and (4.30) are indeed satisfied (including positivity), the result follows.

Theorem 15. *Let $\mathcal{D} = \{D^{(1)}, ..., D^{(n)}\}$ be a collection of selection diagrams relative to source domains $\Pi = \{\pi_1, ..., \pi_n\}$, and target domain π^*, respectively, and S_i represents the collection of S-variables in the selection diagram $D^{(i)}$. Let $\{\langle P^i, I^i_z\rangle\}$ and $\langle P^*, I^*_z\rangle$ be respectively the pairs of observational and interventional distributions in the sources Π and target π^*. The effect $R = P^*(y|do(x))$ is mz-transportable from Π to π^* in $\mathcal{D}$ if the expression $P(y|do(x), S_1, ..., S_n)$ is reducible, using the rules of the do-calculus, to an expression in which (1) do-operators that apply to subsets of I^i_z have no S_i-variables or (2) do-operators apply only to subsets of I^*_z.*

Proof. Consider the following encoding for the domains. Let S_i-variables be an index corresponding to the source domain $\pi_i \in \Pi$, and let the tuple $\langle S_1 = 0, ..., S_i = 1, ..., S_n = 0\rangle$ represent the distributions in this domain. Let the tuple $\langle S_1 = 0, S_2 = 0, ..., S_n = 0\rangle$ represent the distributions in the target domain π^*. The result is now direct since every relation satisfying the conditions of Theorem

15 can be written as a combination of terms computable from the model. The terms without S_i-variables (or $\langle S_1 = 0, ..., S_i = 1, ..., S_n = 0\rangle$) can be written as an experimental distribution in π_i (I_z^i). The terms containing S_i-variables, for all i (or, $\langle S_1 = 0, S_2 = 0, ..., S_n = 0\rangle$), are experimentally identifiable from π^* (I_z^*); all other terms can be written in terms of the observational distribution in π^* and are estimable, therefore, from observations in π^*, the result follows. □

We build on the positive and negative examples given in the section and consider first mz^*-shedges in the one dimensional case when no experimental information is available.

Theorem 16. *If there is a hedge for $P_x^*(y)$ in G and no experimental data is available (i.e., $I_z^* = \{\}$), there exists an mz^*-shedge for $P_x^*(y)$ in G.*

Proof. Consider the hedge $\mathcal{F} = \langle F, F'\rangle$ for the effect $P_x^*(y)$ in G such that F, F' are the respective R-rooted C-forests, and let $\mathcal{F}^*$ be the structure relative to the corresponding mz^*-shedge. Then, consider the following operations:

(i) Initially, set $\mathcal{F}^*$ equal to $\mathcal{F}$, so the root set R^* is the same as R;

(ii) Let $R_x = R \cap De(X)_{\mathcal{F}}$. If there are directed paths from $R \setminus R_x$ to some element of R_x (not passing through X) in G, add the respective paths to F'^* (including all intermediate nodes); alternatively, find a root set $R^* \subseteq An(Y)_{G_{\overline{X}}}$ such that each node in R has a directed path to R^* or $R^* \setminus Y = \emptyset$;

(iii) Remove unnecessary directed paths from F^* enforcing that each node has at most one outgoing directed edge in F^*.

We need to show that a mz^*-shedge can be obtained from a hedge given the construction above, which entail three different scenarios based on condition (ii). First, it might be possible to add to F^* directed paths (not passing through X) from each element of $R \setminus R_x$ to some element of R_x that is present in G. Second, a legitimate root set R is part of $An(Y)_{G_{\overline{X}}}$, so the current set could be extended until reaching some node such that R^* is descendent of R. Eventually, it might be necessary to stretch the paths from the original set R until reaching Y itself. If there are nodes with multiple outgoing edges, it is also possible to remove some of them as needed. It remains to be shown that one of the conditions of Def. 5 is satisfied for the domain represented by G. But this is direct since there are is experimental information to be used ($I_z^* = \emptyset$), which necessarily satisfies cond. 2 of the definition, the result follows. □

Remark 6. The graph G in Fig. 4.7(d) is an example in which the root set needs to be relocated and ultimately stretched reaching Y itself. Also, if there

is an additional edge $Z \to U$ in G, an edge would need to be removed to satisfy the requirement of one outgoing edge per node. More intricate scenarios can be contrived but the construction above suffices for this paper.

Theorem 17. *Let $\mathcal{D} = \{D^{(1)}, ..., D^{(n)}\}$ be a collection of selection diagrams relative to source domains $\Pi = \{\pi_1, ..., \pi_n\}$, and target domain π^*, respectively, and $\{I_z^i\}$, for $i = \{*, 1, ..., n\}$ defined appropriately. If there is an mz^*-shedge for the effect $R = P_x^*(y)$ relative to experiments $(I_z^*, I_z^1, ..., I_z^n)$ in $\mathcal{D}$, R is not mz-transportable from Π to π^* in $\mathcal{D}$ (relative to all experiments I_z^i).*

Proof sketch. We first consider the case in which the mz^*-shedge already satisfies the condition relative to the root set and need not to be augmented. Let F be the R-rooted C-forest (basis), let V' be the set of observable variables and U' be the set of unobservable variables in F. We first consider counterexamples with the induced graph $H = De(F)_D \cap An(Y)_D$, and enforcing the conditions of the definition. Assume, without loss of generality, that H is a forest. We construct two causal models M_1 and M_2 that will agree on the collection of distributions $\{\langle P^i, I_z^i\rangle\}$, $\langle P^*, I_z^*\rangle$, but disagree on the interventional distribution $P_x^*(y)$.

We consider cases in which there are S-nodes in the inputted mz^*-shedge, otherwise the problem reduces to previously studied cases. Let us assume that all variables in $U' \cup V'$ are binary. Call W the set of variables pointed by S-nodes in F'.

Consider the following encoding for the domains. Let S_i be the index variable corresponding to the source domain $\pi_i \in \Pi$, and let the tuple $\langle S_1 = 0, ..., S_i = 1, ..., S_n = 0\rangle$ represent the index for the functional model relative to this domain. Let the tuple $\langle S_1 = 0, S_2 = 0, ..., S_n = 0\rangle$ represent the index for functional model relative to the target domain π^*.

In model 1, let each $V_i \in V' \setminus W$ compute the bit parity of all its observable and unobservable parents (i.e., $f_i^{(M_1)} = \oplus(\bigcup_{V_j \in Pa_i} V_j)$. In model 2, let V_i compute the bit parity of all its parents except that any node in F' disregards the parents values if the parent is in F (i.e., $f_i^{(M_2)} = \oplus(\bigcup_{V_j \in Pa_i \cap F'} V_j)$ if V_i is in F', and $f_i^{(M_2)} = f_i^{(M_1)}$, otherwise).

Finally, define $W \in W$ as follows:

$$W = f_{w^*}^{(M_j)} \oplus \left(U_w^* \wedge S_i\right),$$

where $f_{w^*}^{(M_j)}$ is constructed as in the previous case for the variables in $V' \setminus W$ (for both M_1 and M_2); U_w^* is an additional fair coin pointing exclusively to W, and S_i is the S-node relative to domain π_i. Let us call U_w the collection of such

coins. Furthermore, let us assume that each $U_i \in \{U' \setminus U_w\}$ is also a fair coin (i.e., $P(U_i) = 1/2$).

Lemma 17. *The two models M_1 and M_2 are compatible with the selection diagrams $\mathcal{D}$.*

Proof. The result is immediate. Consider the functional model that generates any domain π_i, in both models M_1 and M_2. By construction, the index tuple is set to $\langle S_1 = 0, ..., S_i = 1, ..., S_n = 0\rangle$ in π_i, and $\langle S_1 = 0, ..., S_i = 0, ..., S_n = 0\rangle$ in π^*. So, it is obvious that in both models, the only structural differences between π_i and π^* are the equations of $W \in W$ in which S_i appears. □

Lemma 18. *The two models agree in the distribution of P^* and there exists an assignment for X and Y such that $P_{M_1}(Y|do(X), S = 1) \neq P_{M_2}(Y|do(X), S = 1)$.*

Proof. The index tuple is set to $\langle S_1 = 0, ..., S_n = 0\rangle$ in π^*, and the result follows directly since in both models, the systems of equations reduce to the construction given by Theorem 4 of (SP06c). □

Lemma 19. *The two models agree in the collection of observational distributions ($\{P^i\}$) in the source domains π_i, $i = 1, ..., n$.*

Proof. For the domains with selection diagrams without S-nodes, the result follows directly from the previous lemma. Let us show that both models agree in the observational distribution P^i relative to source domain π_i with S-nodes. The index tuple is set to $\langle S_1 = 0, ..., S_i = 1, ..., S_n = 0\rangle$ in π_i, and note that both systems are the same as in π^* except that now there exist some variables $W \in W$ with an extra variable U_w^* pointing to it that should be taken into account in W's evaluation, and in turn in the whole system.

We have a forest over the endogenous nodes and all functions compute the bit parity of the value of their parents, and so we can view each node as computing the sum mod 2 of its exogenous ancestors in H. We want to show that the distribution of each family is equally likely for each valid assignment (i.e., $P(v_i|pa_i) = 1/2$, for all valid v_i, pa_i).

Let us partition the analysis in two cases. First consider the case of $V_i \in R$ and when R is a singleton. Note that that the evaluation of V_i relies only on the value of $U_w^* \in U_w$ in its respective tree since $U \in \{U' \setminus U_w\}$ has an even number of endogenous children in F, which is counted twice, so evaluates to zero (i.e., does not affect V_i's evaluation). Without loss of generality, assume that there is only one U_w^* that affects the evaluation of V_i (see subclaim 2). Given the uniformity of U_w^*, it suffices to show that U_w^* can vary independently for any configuration of V_i's parents.

For any configuration of $U' = (U_1 = u_1, ..., U_w^* = u_w^*, ...)$, consider the corresponding evaluation of $Pa_i = pa_i$, and also $V_i = u_w^*$. We want to show that it is possible to flip the current value of U_w^* from u_w^* to $\neg u_w^*$ while preserving the parents' evaluation pa_i. Assume this is not so. This implies that the evaluation of Pa_i and V_i count the same U's, contradiction.

To see why, consider $Pa_i^* \subseteq Pa_i$ the set of parents of V_i that are descendents of U_w^*. Now, flip the minimum number of $U \setminus U_w$ such that Pa_i^* flips while $Pa_i \setminus Pa_i^*$'s configuration is preserved, call this set U^*. (Note that this is always possible since we need at most one U for each parent, which should exist by construction of C-forest.) Now, make $U_w^* = \neg u_w^*$, and note that $Pa_i = pa_i$ since flipping the values of U^* was compensated by the flip of the value of U_w^*. But it is also true now that V_i evaluates to $\neg u_w^*$ since the values of the variables in $\{U \setminus U_w\}$ are washed away in V_i's evaluation, including the ones in U^*, which proves the subclaim.

Consider the following two facts: **Subclaim 1**: Let X and Y be two binary variables such that $P(X = x) = p \neq 1/2$ and $P(Y = y) = q = 1/2$. Then the probabilistic input/output behavior of $Z = XOR(X, Y)$ is the same of Y. The variable $Z = 1$ whenever $\{(X = 1, Y = 0), (X = 0, Y = 1)\}$, which happens with probability $pq + (1 - p)(1 - q)$. Since $q = 1/2$, the expression reduces to $p * 1/2 + (1 - p) * 1/2 = 1/2$. **Subclaim 2**: Let X and Y be two binary variables such that $P(X = x) = P(Y = y) = p = 1/2$. Then the probabilistic input/output behavior of $Z = XOR(X, Y)$ is the same of X (or Y). This follows directly from Subclaim 1. It is clear that if there are multiple nodes from U_w in the evaluation of V_i, the same construction is also valid given the subclaim above. It is not difficult to generalize this argument to consider root set that are not singleton, including roots in which there are not S-nodes as ancestors. The main observation for this case is that the nodes outside R (i.e., $F \setminus R$) exhaust the configurations of the preceding U-nodes (no degrees of freedom available to change R), and the only U-variables that are free to vary are the ones in U_w and bidirectedly connecting the elements of R. But these variables are present in the same way in both models, so the induced distributions can be shown to be same.

Finally, let us consider the case of $V_i \in \{F \setminus R\}$. It suffices to show that the function from $U' \setminus U_w$ to $V' \setminus R$ is 1-1 when we fix $U_w = u_w$. We use essentially the same argument as Shpitser. Assume this is not so, and fix two instantiations of $U' \setminus U_w$ that map to the same value of $V' \setminus R$, and differ by the set $U^* = \{U_1, ..., U_k\}$. Since the bidirected edges form a spanning forest, there exists V^* with an odd number of parents in U^* (and were not in R, by construction). Order them topologically and let the topmost be called X. Note that if we flip all values in U^*, the value of X will also flip, contradiction. Given the uniformity of U', the claim follows. We can put this together with the previous subclaims, and the

result follows. □

Remark 7. The next lemma builds on the idea given above and is based on the simple observation that more U-variables will be free to vary after an intervention on Z.

Lemma 20. *The two models agree in the collection of interventional distributions* ($\{I_z^i\}$) *in the respective source domains* π_i, $i = 1, ..., n$, *and target domain* π^*.

Proof. First consider a pair of domains (π_a, π_b) such that $Z_a, Z_b \cap (F \setminus F') \neq \emptyset$ and (i) condition 1 of Def. 5 does not hold in π_a but cond. 3 does hold, (ii) condition 1 of the definition does hold in π_b but cond. 3 does not hold. It is not possible that condition 1 will not hold for π_b in $G \setminus (Z_a \cap (F \setminus F'))$ since S_b still points to F' in π_b. This implies that we can disregard the variables in Z_i from π_i such that cond. 1 is false and 3 is true, and so we simply disconnect these variables from F; note that intervention on these variables will have no effect on the remaining variables in the system.

Let us show the equality between the interventional distributions I_z^i in both models for domains $\{\pi_i\}$ such that S-variables are not separable from F'. We first consider domains π_i in which Z_i intersects with F' (and not $F \setminus F'$). By construction, the functional models of the nodes in $F \setminus F'$ are the same, so the induced distributions coincide (and do not change with interventions on F'). Let us consider $V_i \in F'$, and its respective function can be rewritten as

$$f_i(.) = f'(Pa_i \cap F') \oplus f''(Pa_i \setminus F') \oplus M_i, \tag{B.1}$$

where f' and f'' are the functions representing the application of the exclusive-or for all elements in the argument, and M_i is an indicator such that if V_i is pointed by a S-node, $M_i = (U_i^* \wedge S_i)$, otherwise, $M_i = 0$.

Consider the following procedure. For each $Z \in Z$, add all its bidirected neighbors V_j to queue L, i.e., each V_j such that $Z \overset{U_{zj}}{\dashleftarrow\!\!\dashrightarrow} V_j$. While $L \neq \emptyset$, pick the element in front $V_j \in L$, and do the following: (1) mark U_{zj} as U_j if $Z \overset{U_{zj}}{\dashleftarrow\!\!\dashrightarrow} V_j$, for some $Z \in Z$; or, (2) mark U_{kj} as U_j, if $V_k \overset{U_{kj}}{\dashleftarrow\!\!\dashrightarrow} V_j$ and $U_k \dashrightarrow V_k$, for some $V_k \neq V_j$ that was previously relabelled. In both cases, add its unseen bidirected neighbors to L.

We want to show that each node in F' (1) has at least one incoming U-node that was marked; and, (2) this node can freely vary independently of its parents within the C-forest. In turn, consider the following subclaims:

Subclaim 3: The procedure finishes and reaches all vertices in F'. This follows directly from the fact that F' is a C-forest, which is finite and all nodes

are bidirectedly connected, together with the construction of the algorithm that implements a BFS-like search.

Subclaim 4: For each $V_j \in V(F')$, there exists at least one marked U_j associated with it. We show the subclaim by induction in the depth of the search induced by the given procedure. For $d = 1$, the claim is certainly true since the respective nodes are directly connected to some $Z \in \mathbf{Z}$, so the marking condition (1) is immediately triggered. Assume that the statement is true for depth $d = l - 1$, and we show next that it holds for $d = l$. Let us call V_l the first node at depth l that is not marked. By inductive hypothesis, all its neighbors at level $l - 1$ were marked as well as its neighbors at level l (since it is the first node in this level not marked). Therefore, by construction, we should mark V_l when we test condition (2) of the procedure, so the result follows.

Subclaim 5: Consider a topological order over the nodes in F'; it is true that each $V^{(i)} \in F'$ $(i = 1, ..., |F'|)$ is such that $P_z(V^{(i)}|pa_i)$ is equally likely, for any configuration $Pa_i = pa_i$. Given that F' is a c-forest, we consider only the parents within the component F', i.e., we evaluate only the first factor (f') in Eq. (B.1) (the remaining parents relative to f'' will be consider later on). Let us show that it is possible to flip the value of $V^{(i)}$ from $v^{(i)}$ to $\neg v^{(i)}$ while preserving the assignment $Pa_i = pa_i$. Assume that U_i is the U-node associated with $V^{(i)}$ given by the previous subclaim, and further that $U^{(i)} = (u_1, ..., u_i)$ is an assignment of the U-nodes previous to $V^{(i)}$ in the order that is compatible with the configuration pa_i. If none of the parents of $V^{(i)}$ is connected to U_i, we are done. Assume that this is not the case. Let $Pa^*_{i,1}$ be the parent of $V^{(i)}$ that is adjacent to U_i, and let U^*_1 be the U-node given by the previous subclaim associated with $Pa^*_{i,1}$. Let $Pa^*_{i,j}$ be defined similarly, the parent of $V^{(i)}$ that is possibly adjacent to U^*_{j-1}, and let U^*_j be the U-node given by the previous subclaim exclusive to $Pa^*_{i,j}$. Now we keep $U^{(i)} \setminus \{U_i, U^*_1, U^*_2, ...\}$ and flip the values of $\{U_i, U^*_1, U^*_2, ...\}$, contradiction, the result follows within F'. But note that if we add the parents that are in $F \setminus F'$ (i.e, f''), the result follows by subclaim 2.

Remark 8. The previous subclaims are intuitive and also expected. First note that previous to the intervention (i.e., in P^i), there was an one-to-one relation from U to V, and after the intervention on Z, the U-nodes that were adjacent to Z are "hanging" (can vary independently of any other nodes). Given that the U-nodes induce a spanning forest over the $V(F')$, there is at least as many U-nodes as $V(F')$ nodes after the intervention. In other words, each variable in F' (including V_i) has an U-node that is free to vary which together with f_i imply that V_i can freely vary independently of all other values in F'. This same intuition can be extended for different arrangements of the root set.

Consider now the case of interventions in which Z_i intersects with $F \setminus F'$.

Note that the variables in $F \setminus F'$ are the same in both models, so the equality between the respective distributions follows. It is the case that similar argument as the one given in the previous case can be applied here as well considering the following two additional facts about the topology of F'. First, there exists still a bijection from $U_{F'}$ to $V(F') \setminus R$ (disregarding the nodes S and in $F \setminus F'$ in the evaluation of f), and so the distribution of such families are all equally likely. If we add the factors relative to f'' (i.e., the vertices in $F \setminus F'$), the distributions are the same by subclaim 2. Second, there are $|R| - 1$ bidirected edges connecting the elements in the root set, which means that the value of one of the nodes (R^*) is implied by the others. But now we can use in both models the S-nodes in F' to make R^* to vary, and so the entire root set varies uniformly in both models, which implies the result. □

Lemmas 17–20 prove Theorem 17 for when there is no need to extend the c-forests with paths from $R \setminus R_x$ to R_x not passing through X. We outline a simple extension on how to use the previous results for this specific case, which will follow the ideas of Proposition **??** and Thm. 16.

Let T be the set of nodes added to make R_x accessible from $R \setminus R_x$, add for each $T_i \in T$ two exogenous variables B_i and U_i, and define T_i in both models as follows:

$$T_i = \Big(\big(\oplus (Pa_i) \oplus U_i \big) \vee B_i \Big) \oplus \Big(B_i \wedge \big((\oplus(Pa_i) \oplus 1) \big) \Big),$$

with $P(U_i) = 1/2$, $P(B_1 = 1, ..., B_{|T|} = 1) = 1/2$, and $P(B_1 = 0, ..., B_{|T|} = 0) = 1/2$. By the construction of the mz^*-shedge, T is in F'.

It is tedious but not difficult to show that the result will follow based on the following observation. If $B = \langle B_1 = 1, ..., B_{|T|} = 1 \rangle$, the two systems will behave as given in the previous case, which will entail the desired equalities and inequalities between the distributions across domains. Alternatively, if $B = \langle B_1 = 0, ..., B_{|T|} = 0 \rangle$, randomness is being added to the systems to assure positivity, and the two models will coincide. I.e., the properties of the previous cases are essentially preserved for when the root set has to be stretched towards Y, just extra attention is required to avoid determinism.

□

Theorem 18 (soundness)**.** *Whenever TR^{mz} returns an expression for $P^*_x(y)$, it is correct.*

Proof. First note that the selection diagrams inputted to TR^{mz} can be viewed as a causal diagram over π^*, and trivial mz-transportability is equivalent to identifiability in π^*. The correctness of the identifiability calls were already established

in (HV06a; SP06b). Following the encoding given in Theorem 17, note that the process of identification of the target relation without the Z-nodes that were considered in lines 10-11 is sound since, by assumption, the distribution $do(Z)$ can be used after testing for direct transportability in the respective local call. □

Corollary 8 (completeness)**.** ***TR***^{mz} *is complete.*

Proof. Follows from Theorems 17,18, and 19. □

Corollary 9 (mz^*-shedge characterization)**.** $P^*_x(y)$ *is* mz*-transportable from* Π *to* π^* *in* $\mathcal{D}$ *if and only if there is not* mz^**-shedge for* $P^*_{x'}(y')$ *in* $\mathcal{D}$ *for any* $X' \subseteq X$ *and* $Y' \subseteq Y$*.*

Proof. Follows directly from Corollary 8. □

Corollary 10 (do-calculus characterization)**.** *The rules of do-calculus together with standard probability manipulations are complete for establishing* mz*-transportability of causal effects.*

Proof. It was shown elsewhere that the steps of the ID algorithm (HV06a; SP06c) (based on (Tia02)) but lines 10 and 11 correspond to sequences of standard probability manipulations and applications of the rules of do-calculus. Line 10 is constituted by a conditional independence test, and standard probability operations for the replacement of the functions based on the invariance allowed by the S-admissibility of the local X' in each recursive call, the result follows. □

APPENDIX C

Proofs for Chapter 5

Theorem 20. *The distribution $P(y|x)$ is s-recoverable from G_s if and only if $(S \perp\!\!\!\perp Y|X)$.*

Proof sketch. (if) It is obvious that if X d-separates S from Y in G_s, $P(y|x)$ is s-recoverable.

(only if) We show that whenever there exists an open path between S and Y that is not blocked by X, two distributions P_1, P_2 compatible with the causal model can be constructed such that they agree in the probability distribution under selection bias, $P_1(V \mid S = 1) = P_2(V \mid S = 1)$, and disagree in the target distribution $Q = P(Y \mid X)$, i.e., $P_1(Y \mid X) \neq P_2(Y \mid X)$.

Let P_1 be compatible with the graph $G_1 = G_s$, and P_2 with the subgraph G_2 where the edges pointing to S are removed (see (Tia02, Lemma 8)). Notice that P_2 harbors an additional independence relative $(V \perp\!\!\!\perp S)_{P_2}$, where V represents all variables in G_s but the selection mechanism S. We will set the parameters of P_1 through its factors and then compute the parameters of P_2 by enforcing $P_2(V \mid S = 1) = P_1(V \mid S = 1)$. Since $P_2(V|S = 1) = P_2(V)$, we will have $P_1(V|S = 1) = P_2(V)$.

Given a Markovian data-generating model (Pea00), P_1 can be parametrized through its factors in the Markovian decomposition $P_1(S = 1 \mid Pa_s), P_1(X \mid Pa_x), \ldots$, more generally, $P_1(V_i \mid PA_i)$ for each family in the graph. Recoverability should hold for any parametrization, so we assume that all variables are binary. In turn, we examine the possible ways of how S is connected to Y while conditioned on X.

Case 1. Firstly, let us consider the case in which $Y \in Pa_S$, which implies that S is not separable from Y in G_s. We follow the construction given in Lemma 1. Let U be the set of nodes that connect X to Y. The distribution of Y is a function of the values of X if we sum out all variables in U, $P_1(Y|X) = \sum_U \prod_{X,U,Y} P_1(V_i|Pa_i))$, so without loss of generality we can parametrize this distribution directly. Now, we can write the conditional distribution in the second

causal model as follows:

$$P_2(Y|X) = P_1(Y|X, S=1) = \frac{P_1(Y, X, S=1)}{P_1(X, S=1)} \tag{C.1}$$

$$= \frac{P_1(S=1|Y)P_1(Y|X)}{P_1(S=1|Y)P_1(Y|X) + P_1(S=1|\overline{Y})P_1(\overline{Y}|X)}, \tag{C.2}$$

where the first equality is enforced by construction, the second and third follow from the axioms of probability.

Consider the subgraph G' such that all $V \setminus \{X, Y, U, S\}$ are disconnected from $\{X, Y, U, S\}$, where we can parametrize the complete model as in (Tia02, Lemma 8). Now we compare $P_2(Y|X)$ with $P_1(Y|X)$. The equality constraint imposed over these quantities can be seen as a line in the parameter space of higher dimension, which has measure zero. This implies that for almost all parametrizations, $P_1(Y|X)$ and eq. (2) will not be the same. For instance, we can set the distribution of every family in G' but the selection node equal to $1/2$, and set the distribution $P_1(S=1|Y) = \alpha, P_1(S=1|\overline{Y}) = \beta$, for $0 < \alpha, \beta < 1$ and $\alpha \neq \beta$. The result follows since the other parameters of P_2 are free and can be chosen to match P_1, and $P_2(Y|X) = \alpha/(\alpha+\beta)$ and $P_1(Y|X) = 1/2$.

Case 2. Let us consider the case in which there exists an open directed path from Y to S, which means that it does not pass through X (i.e., only the values of X will end up being used in the construction). Let Z be the immediate child of Y in this path and assume the distance from Z to S is arbitrary. Let W be the set of nodes that connect Z to S and U be the set of nodes that connect X to Y. Consider the induced subgraph G' such that all nodes in G_s but $V \setminus \{X, U, Y, Z, W, S\}$ are disconnected from $\{X, U, Y, Z, W, S\}$.

Following eq. (1), $P_2(Y|X) = \frac{P_1(Y,X,S=1)}{P_1(X,S=1)}$, we can rewrite the numerator of the r.h.s. in expanded form as

$$\begin{aligned}
P_1(Y, X, S=1) &= \sum_{U,Z,W} P_1(X, U, Y, W, Z, S=1) \\
&= \sum_{U,Z,W} P_1(X|Pa_x) \; ... \; P_1(S=1 \mid Pa_s) \\
&= \sum_{U,Z,W} \prod_{V \cup S} P_1(V_i \mid Pa_i) \\
&= \sum_{U} \prod_{U} P_1(V_i \mid Pa_i) \sum_{Z,W} \prod_{V \cup S \setminus U} P_1(V_i \mid Pa_i)
\end{aligned} \tag{C.3}$$

Given a topological order compatible with G', the families in U are functions of X but not of Z, W, Y, S, and since the same value of X is instantiated in the

numerator and the denominator in eq. (1), these factors cancel out. So, we consider only the second sum in eq. (3). Now, we can rewrite

$$\sum_{Z,W} \prod_{V \cup S \setminus U} P_1(V_i \mid Pa_i)$$
$$= \sum_{Z} \prod_{V \setminus U \cup W} P_1(V_i \mid Pa_i) \sum_{W} \prod_{W \cup S} P_1(V_i \mid Pa_i) \tag{C.4}$$

The sum over the factors relative to W in eq. (4) is a function of Z (since $Z \in An(S)$), so define $f(Z) = \sum_W \prod_{S \cup W} P_1(V_i|Pa_i)$. The distribution of Y is a function of the value of X since we sum out all values of U, let us call it $P(Y|\tilde{X})$. Define $\alpha_z(Y) = P(Z|Y)$, and since Y is not affected by Z, we can rewrite eq. (4) as $P(Y|\tilde{X}) \sum_Z \alpha_z(Y) f(Z)$. Given these observations, we rewrite $P_2(Y|X)$ (eq. (1)) as follows

$$\frac{P_1(Y|\tilde{X}) \sum_Z \alpha_z(Y) f(Z)}{\big(P_1(Y|\tilde{X}) \sum_Z \alpha_z(Y) f(Z)\big) + \big(P_1(\overline{Y}|\tilde{X}) \sum_Z \alpha_z(\overline{Y}) f(Z)\big)},$$

which we want to compare with $P_1(Y|\tilde{X})$.

By construction of G', $f(Z)$ and $\alpha_z(Y)$ as a convolution, it is the case that the expressions for Q_1 and Q_2 cannot be simplified in the general case. We explore the fact that the equality constraint between these two quantities (for all values of X and Y) imposes weak constraints in the high dimensional parameter space and valid parametrizations have Lebesgue mass zero; i.e., for almost all parameters that we chose the equality between Q_1 and Q_2 will not hold, we chose explicitly one of such parameters. So, first make $P_1(Y|\tilde{X}) = 1/2$ for all values of $Y, \tilde{X}$, which implies

$$P_2(Y|\tilde{X}) = \frac{\sum_Z \alpha_z(Y) f(Z)}{\big(\sum_Z \alpha_z(Y) f(Z)\big) + \big(\sum_Z \alpha_z(\overline{Y}) f(Z)\big)} \tag{C.5}$$

We can compose the linear transformations encoded in $f(Z)$, which is from the parameter space of $W \cup S$ to Z, that is, $[0,1]^{2^{|W|+1}} \to [0,1]^{|Z|}$. Consider a topological order $W_1 < W_2 < ... < W_{|W|} < S$ relative to $W \cup S$. We re-arrange the product $\sum_W \prod_{S \cup W} P_1(V_i|Pa_i)$ as 2×2 matrices relative to each factor $P(W_i|W_{i+1})$ (each row sums to 1 satisfying the integrality constraint) and $P(S = 1|W_{|W|})$ is a column-vector 2×1 for each value of $W_{|W|}$.

Let the matrix of the first distribution relative to W_1 be $M = [p, 1-p; 1-q, q]$, for some $0 < p, q < 1$, which will be instantiated below. We can decompose M in its canonical form, i.e., in terms of its eigenvectors, $[1, -(p-1)/(q-1)], [1, 1]$, and eigenvalues $[1, p+q-1]$. The product in $f(Z)$ is a composition of linear

transformations, which is also a linear transformation. We make each distribution to follow the same form given by M, so this composition is equivalent to the product of the matrix with the eigenvectors times the power to $k = |W|$ of the matrix with the eigenvalues in the diagonal times the inverse of the matrix with the eigenvectors, let us call it M_c. After some trivial (but tedious) algebra, we obtain:

$$\begin{aligned} M_c(1,1) &= 1 - \frac{(1-p)\big((p+q-1)^k - 1\big)}{p+q-2} \\ M_c(1,2) &= \frac{(1-p)\big((p+q-1)^k - 1\big)}{p+q-2} \\ M_c(2,1) &= \frac{(1-q)\big((p+q-1)^k - 1\big)}{p+q-2} \\ M_c(2,2) &= 1 - \frac{(1-q)\big((p+q-1)^k - 1\big)}{p+q-2} \end{aligned} \tag{C.6}$$

Set $(p = 3/5, 1-q = 2/5)$, it is not difficult to check that this assignment yields a valid parametrization for the distribution, we have

$$\begin{aligned} M_c(1,1) = M_c(2,2) &= 1 - \frac{1}{2}(1 - \big(\frac{1}{5}\big)^k) \\ M_c(1,2) = M_c(2,1) &= \frac{1}{2}(1 - \big(\frac{1}{5}\big)^k) \end{aligned} \tag{C.7}$$

Now, let $(P(S=1|W_{|W|}) = 2/3, P(S=1|\overline{W_{|W|}}) = 1/2)$, and we can see that $f(Z) = 7/12 + \epsilon, f(\overline{Z}) = 7/12 - \epsilon$, where $\epsilon = (1/5)^k$. We can chose $\alpha_z(y) = 1/3, \alpha_z(\overline{y}) = 3/4$. Finally, we can evaluate eq. (5) and note that $Q_2 = 1/2 - (2/7)\epsilon$, which is never equal to $1/2$ $(= Q_1)$ given that the graph is finite.

Case 3. Let us consider the case in which the path from Y to S pass through an ancestor of Y. Let us call $A = An(Y) \setminus \{Y\}$. Since $A \setminus X$ is not d-separated from Y given X in G_s, there is a path p from $Z \in A \setminus X$ to Y that is not blocked by X. Without loss of generality, let us consider the closest Z in this path. There are two possible cases to consider: p might be a directed path from Z to Y that does not contain X as an intermediate (e.g., $Z \to \ldots \to Y$); or, p might contain converging arrows into X $(Z \to \ldots \to X \leftarrow \ldots \to Y)$.

Subcase 3a. We start with when p is a directed path. Let U be the set of nodes that connect X to Y, W the nodes that connect Z to Y (given X), and R the nodes that connect Z to S. Consider the induced subgraph G' of G_s such that all nodes except $\{X, U, Z, W, R, Y\}$ are removed from G_s (i.e., $V \setminus \{X, U, Z, W, R, Y\}$ can be parametrized as random coins, see (Tia02, Lemma 8)).

Since $Z \in An(S)$, let us call p' the path connecting Z to $An(S) \setminus An(Y)$ in G_s (i.e., $Z \to \ldots \to S$). Add p' with all its nodes to G'. Note that Z is such that it blocks the concatenation of p and p'. Note that this concatenation is such that it has two emanating arrows from Z (i.e., $p \leftarrow Z \to p'$). Now, we can transform G_s while staying in the same equivalence class . In order to do so, reverse the direction of all arrows in p such that Z is no longer in $An(Y)$. Now, the same parametrization as discussed in case 2 is valid for this case.

Subcase 3b. Consider the case in which p contain converging arrows into X. Let us consider the variables X, Y, Z, and let L be the common ancestor that, together with Z, has converging arrows into X in p. The construction here will be similar to the previous case except for two main differences.

First, the path p can be seen as the concatenation of four segments $p_1, ..., p_4$ such that p_1 is the segment $L \to \ldots \to Y$, p_2 is the segment $L \to \ldots \to X$, p_3 is the segment $Z \to \ldots \to X$, and p_4 the segment $Z \to \ldots \to S$. Note that by construction, there might exist only chains along each of these segments, so to avoid algebraic clutter we assume that those are segments of length one, but it is trivial to stretch those segments following the same structure given in case 2 for $f(Z)$. When we have multiple X's in p, we will have the concatenation of several segments p_3 and p_4, and it will also be simple to extend the construction given for $f(Z)$ for this case. Remarkably, these segments capture precisely the forbidden subgraph that precludes s-recoverability when p has converging arrows to X. Second, no directed path between X and Y is used in the construction of the counterexample and the induced subgraph G' without these paths can also be generated by the original model (Tia02, Lemma 8).

We follow similar structure as in case 2. Following eq. (1), $P_2(Y|X) = \frac{P_1(Y,X,S=1}{P_1(X,S=1)}$, we can rewrite the numerator as

$$\sum_L P_1(Y|L)P_1(L) \sum_Z P_1(X|Z,L)P_1(Z)P_1(S=1|Z) \tag{C.8}$$

Define $\alpha_L(Y) = P_1(Y|L)P_1(L)$ and note that the second sum is not affected by Y but it is a function of L, so define $f(L) = \sum_Z P_1(X|Z,L)P_1(Z)P_1(S=1|Z)$, and write

$$P_2(Y|X) = \frac{\sum_L \alpha_L(Y)f(L)}{\left(\sum_L \alpha_L(Y)f(L)\right) + \left(\sum_L \alpha_L(\overline{Y})f(L)\right)} \tag{C.9}$$

Define another function of L that sums out S, $g(L) = \sum_Z P_1(X|Z,L)P_1(Z)$, and note that $P_1(Y|X)$ is the same as eq. (9) with the function f replaced with g. This expression cannot be simplified in general since there is a dependence across the two functions. To see that, consider the following parametrization:

$\alpha_L(Y) = \alpha_{\overline{L}}(Y) = 1/3$, $\alpha_L(\overline{Y}) = 1/9, \alpha_{\overline{L}}(\overline{Y}) = 2/9$, $P(Z) = 1/2$; $P_1(X|Z, L) = 1/2 + \epsilon, P_1(X|\overline{Z}, L) = 1/2 - \epsilon$, $P_1(X|Z, \overline{L}) = P_1(X|\overline{Z}, \overline{L}) = 1/2$, for $0 < \epsilon < 1/2$. Call $P(S = 1|Z) = \alpha, P(S = 1|\overline{Z}) = \beta$, and pick any α, β such that $\alpha > \beta$. After some trivial (but tedious) algebra, we have $P_1(Y|X) = 2/3$ and

$$P_2(Y|X) = \frac{2}{3}\left(\frac{\alpha + \beta + \epsilon(\alpha - \beta)}{\alpha + \beta + \frac{8}{9}\epsilon(\alpha - \beta)}\right), \tag{C.10}$$

which are always different. □

Remark 9. We considered Markovian models in Theorem 1, but the extension for Semi-Markovians is straightforward. This is so because the latent variables impose no constraints over the distribution of the observables, which means that there are even more degrees of freedom that can be used to produce a parametrization following the lack of separability.

Theorem 21. *If there is a set C that is measured in the biased study with $\{X, Y\}$ and in the population level with X such that $(Y \perp\!\!\!\perp S|\{C, X\})$, then $P(y|x)$ is recoverable as*

$$P(y|x) = \sum_c P(y|x, c, S = 1)P(c|x). \tag{C.11}$$

Proof. We can condition $P(y|x)$ on the set C and write

$$\begin{aligned} P(y|x) &= \sum_c P(y|x, c)P(c|x) && \text{(C.12)} \\ &= \sum_c P(y|x, c, S = 1)P(c|x), && \text{(C.13)} \end{aligned}$$

where the last line follows since C is such that $(Y \perp\!\!\!\perp S|\{C, X\})$. QED. □

Lemma 21. *If $Y \perp\!\!\!\perp S|(C, X)$, then $Y \perp\!\!\!\perp S|(C', X)$, where $C' = C \cap An(Y \cup S \cup X)$ (AC96).*

Lemma 22. *Given three sets of nodes X, Y, and Z, and a set $C \subseteq An(X \cup Y \cup Z)$, $X \perp\!\!\!\perp Y|(Z \cup C)$ if and only if $Z \cup C$ separates X from Y in undirected graph $(G_{An(X \cup Y \cup Z)})^m$, the moral graph of $G_{An(X \cup Y \cup Z)}$ (AC96).*

Theorem 22. *There exists some set $C \subseteq T \cap M$ such that $Y \perp\!\!\!\perp S|\{C, X\}$ if and only if the set $(C' \cup X)$ d-separates S from Y where $C' = [(T \cap M) \cap An(Y \cup S \cup X)] \setminus (Y \cup S \cup X)$.*

Proof. The "if" part is trivial as it gives a set that d-separates S from Y.

(only if) If there exists a set $C \subseteq T \cap M$ (that is disjoint from Y, S, X) such that $Y \perp\!\!\!\perp S|(C, X)$ then the set $C'' = C \cap An(Y \cup S \cup X)$ satisfies $Y \perp\!\!\!\perp S|(C'', X)$ based on Lemma 21. From Lemma 22 we have that $C'' \cup X$ separates S from Y in the undirected graph $(G_{An(Y \cup S \cup X)})^m$. In an undirected graph, if $(C'' \cup X) \subseteq (C' \cup X)$, is a separator, then $C' \cup X$ must be a separator. Using Lemma 22 again, we obtain that $(C' \cup X)$ d-separates S from Y in G. □

Lemma 23. *Let C_1 be a minimal set satisfying $Y \perp\!\!\!\perp S|(C_1, X)$, C_1^o be any subset of C_1 (including empty set), and $C_1^m = C_1 \setminus C_1^o$. If C_2 is a minimal set satisfying $C_1^m \perp\!\!\!\perp S|(C_1^o, X, C_2)$, then we must have $Y \perp\!\!\!\perp S|(C_2, C_1, X)$ and $Y \perp\!\!\!\perp S|(C_2, C_1^o, X)$.*

Proof. Since C_1 is minimal, by Lemma 21 we obtain $C_1 \subseteq An(Y \cup S \cup X)$. Similarly we have $C_2 \subseteq An(S \cup X \cup C_1)$, and therefore $C_2 \subseteq An(Y \cup S \cup X)$. Since $Y \perp\!\!\!\perp S|(C_1, X)$, by Lemma 22 we have that $C_1 \cup X$ separates S from Y in the undirected graph $(G_{An(Y \cup S \cup X)})^m$. Since $C_2 \subseteq An(Y \cup S \cup X)$ we have that $C_1 \cup X \cup C_2$ separates S from Y in the undirected graph $(G_{An(Y \cup S \cup X)})^m$. Then by Lemma 22 we obtain $Y \perp\!\!\!\perp S|(C_2, C_1, X)$. Given $Y \perp\!\!\!\perp S|(C_2, C_1^m, C_1^o, X)$, and $C_1^m \perp\!\!\!\perp S|(C_1^o, X, C_2)$, we obtain $Y \perp\!\!\!\perp S|(C_2, C_1^o, X)$ by the contraction axiom. □

Lemma 24. *For sets W, X, let C_1 be a nonempty minimal set satisfying $W \perp\!\!\!\perp S|(C_1, X)$. Let C_1^o be any subset of C_1, and $C_1^m = C_1 \setminus C_1^o$. We have*

$$P(w|x) = \sum_{c_1} P(w|x, c_1, S = 1)P(c_1|x). \tag{C.14}$$

Then

1. *$C_1 \perp\!\!\!\perp S|X$ does not hold.*
2. *Let $C_2 \subseteq M$ be a minimal set satisfying $C_1 \perp\!\!\!\perp S|(X, C_2)$. Then $W \perp\!\!\!\perp S|(C_2, X)$. Therefore,*

$$P(c_1|x) = \sum_{c_2} P(c_1|x, c_2, S = 1)P(c_2|x). \tag{C.15}$$

$$P(w|x) = \sum_{c_2} P(w|x, c_2, S = 1)P(c_2|x). \tag{C.16}$$

   That is, if $P(c_1|x)$ is recovered via Theorem 21, then $P(w|x)$ must be recovered via Theorem 21.

3. *$C_1^m \perp\!\!\!\perp S|(C_1^o, X)$ does not hold.*

4. *Let $C_2 \subseteq M$ be a minimal set satisfying $C_1^m \perp\!\!\!\perp S|(C_1^o, X, C_2)$. Then $W \perp\!\!\!\perp S|(C_2, C_1^o, X)$. Therefore,*

$$P(c_1^m|c_1^o, x) = \sum_{c_2} P(c_1^m|c_1^o, x, c_2, S = 1)P(c_2|c_1^o, x). \tag{C.17}$$

$$P(w|x) = \sum_{c_1^o, c_2} P(w|c_1^o, x, c_2, S = 1)P(c2, c_1^o|x). \tag{C.18}$$

That is, if $P(c_1^m|c_1^o, x)$ is recovered via Theorem 21, then $P(w|x)$ must be recovered via Theorem 21.

Proof. 1. If $C_1 \perp\!\!\!\perp S|X$, from $W \perp\!\!\!\perp S|(C_1, X)$ and the contraction graphoid axiom, we obtain $W \perp\!\!\!\perp S|X$. This contradicts with C_1 being minimal.

2. Given $C_2 \subseteq M$ being a minimal set satisfying $C_1 \perp\!\!\!\perp S|(X, C_2)$, we obtain $W \perp\!\!\!\perp S|(C_2, X)$ by Lemma 23.

3. If $C_1^m \perp\!\!\!\perp S|(C_1^o, X)$, from $W \perp\!\!\!\perp S|(C_1^m, C_1^o, X)$ and the contraction graphoid axiom, we obtain $S \perp\!\!\!\perp W|(C_1^o, X)$. This contradicts with C_1 being minimal.

4. Given $C_2 \subseteq M$ being a minimal set satisfying $C_1^m \perp\!\!\!\perp S|(C_1^o, X, C_2)$, we obtain $W \perp\!\!\!\perp S|(C_2, C_1^o, X)$ by Lemma 23.

□

Theorem 23. *For $X \subseteq T$, $Y \notin T$, $Q = P(y|x)$ is C-recoverable if and only if it is recoverable by Theorem 21, that is, if and only if there exists a set $C \subseteq T \cap M$ such that $(Y \perp\!\!\!\perp S|C, X\})$ (where C could be empty). If s-recoverable, $P(y|x)$ is given by $P(y|x) = \sum_c P(y|x, c, S = 1)P(c|x)$.*

Proof sketch. (if) If there exists a set $C \subseteq T \cap M$ such that $Y \perp\!\!\!\perp S|(C, X)$, then it is clear $RC(Y, X)$ will recover $P(y|x)$.

(only if) Assume there exists no set $C \subseteq T \cap M$ such that $Y \perp\!\!\!\perp S|(C, X)$. If there exists no set $C \subseteq M$ such that $Y \perp\!\!\!\perp S|(C, X)$, then $RC(Y, X)$ will output FAIL. Assume for every minimal set $C_1 \subseteq M$ satisfying $Y \perp\!\!\!\perp S|(C_1, X)$, there exist some variables in C_1 that are not in T. We need to prove $RC(Y, X)$ will not recover $P(y|x)$.

The only way for $RC(Y, X)$ to recover $P(y|x)$ is by the following

$$P(y|x) = \sum_{c_1} P(y|x, c_1, S = 1)P(c_1|x), \tag{C.19}$$

such that $R(C_1, X)$ recovers $P(c_1|x)$ for some C_1. By Lemma 24, the only way for $R(C_1, X)$ to recover $P(c_1|x)$ is that there exists some $C_1^o \subset C_1$ (C_1^o could

be empty set) for which there exists a minimal set $C_2 \subseteq M$ satisfying $C_1^m \perp\!\!\!\perp S|(C_1^o, X, C_2)$ where $C_1^m = C_1 \setminus C_1^o$, such that either $C_2 \cup C_1^o \cup X \subseteq T$ rendering $P(c_1^m|c_1^o, x)$ being recovered via Theorem 21 or $R(C_2, C_1^o \cup X)$ recovers $P(c_2|c_1^o, x)$ (and $R(C_1^o, X)$ recovers $P(c_1^o|x)$). But $P(c_1^m|c_1^o, x)$ being recovered via Theorem 21 would contradict with our assumption since by Lemma 24 it means $P(y|x)$ will be recovered via Theorem 21.

These same arguments apply to $R(C_2, C_1^o \cup X)$. By repeated application of Lemma 24, we have that if $RC(Y, X)$ succeeds in recovering $P(y|x)$, then there exist a sequence of function calls $R(C_1, X), R(C_2, C_1^o \cup X), \ldots, R(C_k, C_{k-1}^o \cup \ldots \cup C_1^o \cup X)$ that ends with $R(C_k, C_{k-1}^o \cup \ldots \cup C_1^o \cup X)$ succeeding in computing $R(C_k|C_{k-1}^o \cup \ldots \cup C_1^o \cup X)$ by recovering $R(C_k^m|C_k^o, C_{k-1}^o \cup \ldots \cup C_1^o \cup X)$ via Theorem 21. Then by reasoning backwards using Lemma 24, we have that $R(C_{k-1}^m|C_{k-1}^o, C_{k-2}^o \cup \ldots \cup C_1^o \cup X)$ must be recovered via Theorem 21, and so on, until we obtain $P(c_1^m|c_1^o, x)$ must be recovered via Theorem 21 and finally $P(y|x)$ must be recovered via Theorem 21. This would contradict with our assumption. Therefore $RC(Y, X)$ will not recover $P(y|x)$.

Theorem 24 (Selection-backdoor adjustment)**.** *If a set Z satisfies the s-backdoor criterion relative to the pairs (X, Y) and (M, T) (as in def. 2), then the causal effect of X on Y is identifiable and recoverable and is given by the formula*

$$P(y|do(x)) = \sum_{z^+} P(y|x, z, S = 1)P(z) \tag{C.20}$$

Proof. We first condition the effect of X on Y on Z^+ and write

$$P(y|do(x)) = \sum_{z^+} P(y|do(x), z^+)P(z^+|do(x)) \tag{C.21}$$

We can rewrite the effect in eq. (C.21) as

$$\begin{aligned} P(y|do(x)) &= \sum_{z^+} P(y|do(x), z^+)P(z^+) && \text{(C.22)} \\ &= \sum_{z^+} P(y|x, z^+)P(z^+), && \text{(C.23)} \end{aligned}$$

where eq. (C.22) follows by the third rule of the do-calculus together with the fact that $(Z^+ \perp\!\!\!\perp X)_{G_{\overline{X}}}$ (since by construction Z^+ contains only non-descendants of Y), and eq. (C.23) follows by the second rule of the do-calculus together with condition (i).

We can rewrite the second term in eq. (C.23) summing over Z^- and pull the sum out, which yield

$$P(y|do(x)) = \sum_{z^+, z^-} P(y|x, z^+)P(z^+, z^-) \tag{C.24}$$

By the contraction graphoid axiom conditions (ii) and (iii) entail $(Y \perp\!\!\!\perp S, Z^- | X, Z^+)$, so we can add $\{Z^-, S\}$ to the first term of eq. (C.24) and obtain

$$P(y|do(x)) = \sum_{z^+, z^-} P(y|x, z^+, z^-, S = 1) P(z^+, z^-). \tag{C.25}$$

Note that eq. (C.25) is identifiable and its recoverability is given by condition 4. □

In the sequel, we provide a procedure for listing all recoverable distributions in the form of $P(y, B|A)$. Note that $P(y, B|A)$ being recoverable implies other distributions such as $P(y|A, D)$ is recoverable for all $D \subseteq B$.

Procedure Sink-Recover(G, Y)

1. Remove $V \setminus An(Y \cup S)$ from G.

2. Eliminate S.

   (a) If $Y \in Pa_S$, exit with failure.

   (b) Otherwise, $P(Y, V \setminus Pa_S \setminus \{Y\} | Pa_S)$ is s-recoverable, and remove S from G.

3. Eliminate non-ancestors of Y from the graph one by one. Given $P(Y, B|A)$ s-recoverable, iterate in reverse topological order, for each sink node Z.

   (a) If $Y \notin Pa_Z$, $P(Y, B \setminus Pa_Z | A \cup Pa_Z \setminus Z)$ is s-recoverable, and remove Z from G.

   (b) Otherwise, exit if no non-ancestors of Y can be removed.

4. Now all non-ancestors of Y have been removed and we have $P(Y, B|A)$ s-recoverable.

   (a) For $C \subseteq An(Y) \setminus \{Y\}$,
   if $(Y \perp\!\!\!\perp A - C | C)$, then $P(Y|C)$ s-recoverable.

The procedure operates traversing the graph and trying to recover distributions in the form $P(y, B|A)$ until the current node can no longer be separated from Y given its parents (and respective ancestors), or it ends listing all distributions and reaching Y itself. It is not difficult to see that whenever the algorithm exits with failure, one of the separability conditions discussed in the proof of Theorem 1 is violated, which means that a counterexample for *s*-recoverability can be produced.

Interestingly, the Sink-Recover() can be easily modified to list odds ratios (OR), extending the query-specific treatment given in (BP12b). Note that the symmetry of the functional form of the OR can be exploited in this case so that the separability test in the procedure can be relaxed. Under this relaxation the current Z must be separated from X *or* Y rather than always Y.

Theorem 25 (OR G-recoverability)**.** *Let graph G contain the arrow $X \to Y$ and a set C of measured X-independent covariates. The c-specific odds ratio $OR(Y, X \mid C)$ is G-recoverable from s-biased data if and only if there exists an additional set Z of measured variables such that the following conditions hold in G:*

1. $(X \perp\!\!\!\perp S \mid \{Y, Z, C\})_G$ *or* $(Y \perp\!\!\!\perp S \mid \{X, Z, C\})_G$.
2. *Z is OR-admissible relative to (X, Y, C).*

Moreover, $OR(Y, X \mid C) = OR(Y, X \mid C, Z, S = 1)$.

Proof. (if part) Our target quantity is $OR(X, Y \mid C)$ and given that Z is OR-admissible relative to (X, Y, C), Corollary 17 permits us to add Z and rewrite it as $OR(X, Y \mid C, Z)$. Given that the first condition of the theorem holds, Corollary 16 implies $OR(X, Y \mid C, Z) = OR(X, Y \mid C, Z, S = 1)$. This establishes G-recoverability since the r.h.s. is estimable from the available s-biased data.

(only if part) If the conditions of the theorem cannot be satisfied, then $OR(X, Y \mid C)$ is not G-recoverable, that is, there exist two distributions P_1, P_2 compatible with G such that they agree in the probability under selection, $P_1(V \setminus \{S\} \mid S = 1) = P_2(V \setminus \{S\} \mid S = 1)$, and disagree in the odds ratio, $OR_1(X, Y \mid C) \neq OR_2(X, Y \mid C)$. We first consider the case when $C = \{\}$, and we will construct two such distributions. Let P_1 be compatible with the graph $G_1 = G$, and P_2 with the subgraph G_2 where all edges pointing to S are removed. Both are compatible with G, since compatibility with a subgraph assures compatibility with the graph itself (Pea88). Notice that P_2 harbors an additional independence $(V \setminus \{S\} \perp\!\!\!\perp S)_{P_2}$. By construction $P_1(X, Y \mid S = 1) = P_2(X, Y \mid S = 1)$, but since

$$P_2(X, Y | S = 1) = P_2(X, Y),$$

we have:

$$P_1(X, Y | S = 1) = P_2(X, Y)$$

We can then simplify OR_2 rewriting it as follows

$$OR_2 = \frac{P_1(X, Y, S = 1) P_1(\overline{X}, \overline{Y}, S = 1)}{P_1(\overline{X}, Y, S = 1) P_1(X, \overline{Y}, S = 1)}, \tag{C.26}$$

and similarly for OR_1,

$$OR_1 = \frac{P_1(X,Y)P_1(\overline{X},\overline{Y})}{P_1(\overline{X},Y)P_1(X,\overline{Y})} \tag{C.27}$$

We want to show that it is possible to produce a parametrization of P_1 in such way that $OR_1(X,Y) \neq OR_2(X,Y)$. First, let us consider the class of Markovian models. Accordingly, P_1 can be parametrized through its factors in the Markov decomposition $P_1(S = 1 \mid PA_s), P_1(X \mid PA_x), \ldots$, or more generally, $P_1(V_i \mid PA_i)$ for each family in the graph. This choice of parameters induces a valid parameterization for P_2 as well. Firstly, let us consider the case in which condition 1 of the theorem fails, i.e., $\{X,Y\}$ are not separable from S. Thus, eq. (C.26) can be rewritten using the identity $P_1(X,Y,S=1) = P_1(S=1 \mid X,Y)P_1(X,Y)$, yielding:

$$OR_2 = OR_1\left(\frac{P_1(S=1|X,Y)P_1(S=1 \mid \overline{X},\overline{Y})}{P_1(S=1 \mid \overline{X},Y)P_1(S=1 \mid X,\overline{Y})}\right) \tag{C.28}$$

Note that making the multiplier of OR_1 in eq. (C.28) different than 1 entails $OR_2 \neq OR_1$, which will happen for *almost all* parametrizations of $P_1(S=1 \mid .)$ independently of the one chosen for $P_1(X,Y)$. In case there are additional nodes pointing to S, we can just make them independent of S in this new parametrization given that compatibility with the subgraph is enough to ensure compatibility with G.

Now, let us consider the case in which condition 2 of the theorem fails, i.e., there is no OR-admissible sequence in relation to $(X,Y,\{\})$. Let $Z = V \setminus \{X,Y,S\}$, and expand $P_1(X,Y,S=1)$ in the following way[1]:

$$\begin{aligned} P_1(X,Y,S=1) &= \sum_Z P_1(X,Y,S=1,Z) \\ &= \sum_Z P_1(X \mid PA_x) \ldots P_1(S=1 \mid PA_s) \\ &= \sum_Z \prod_{V \cap S=1} P_1(V_i \mid PA_i) \end{aligned} \tag{C.29}$$

Notice that each term in eq. (C.26) can be rearranged for each assignment of S'

[1]It clear that we should consider in the expression above (in respect to Z) just the nodes that are somehow related to S, i.e., its ancestors, otherwise we could just sum these vertices out because they do not offer any additional constraint over the distribution of interest related to OR, and then in its respective parameterization.

parents (i.e., $PA_s = pa_s^{(j)}$), for instance, we can write based on eq. (C.29):

$$\begin{aligned}
&P_1(X, Y, S = 1) = \\
&\quad P_1(S = 1 \mid PA_s = pa_s^{(1)}, \lambda)\Big(\sum_{Z, PA_s = pa_s^{(1)}} \prod_{V \setminus S} P_1(V_i \mid PA_i)\Big) + \\
&\quad P_1(S = 1 \mid PA_s = pa_s^{(2)}, \lambda)\Big(\sum_{Z, PA_s = pa_s^{(2)}} \prod_{V \setminus S} P_1(V_i \mid PA_i)\Big) + \\
&\quad \ldots \\
&\quad P_1(S = 1 \mid PA_s = pa_s^{(k)}, \lambda)\Big(\sum_{Z, PA_s = pa_s^{(k)}} \prod_{V \setminus S} P_1(V_i \mid PA_i)\Big)
\end{aligned} \tag{C.30}$$

where k is the number of configurations of S' parents, and λ indexes configurations of X or Y whenever one of them is a parent of S. Given eq. (C.30), let us call $P_1(S = 1 \mid PA_s = pa_s^{(1)}, \lambda) = \alpha_1^\lambda$, $P_1(S = 1 \mid PA_s = pa_s^{(2)}, \lambda) = \alpha_2^\lambda, \ldots$, and also call $\sum_{Z, PA_s = pa_s^{(j)}} \prod_{V \setminus S} P_1(V_i \mid PA_i) = f_j(x, y)$ for each configuration $X = x, Y = y, PA_s = pa_s^{(j)}$. Then, we can write eq. (C.30) in the following simplified manner:

$$P_1(X, Y, S = 1) = \alpha_1^\lambda f_1(x, y) + \alpha_2^\lambda f_2(x, y) + \ldots \tag{C.31}$$

for all values of X and Y. We can then rewrite OR_2 based on eq. (C.31) as

$$\begin{aligned}
OR_2 = &\frac{(\alpha_1^\lambda f_1(x, y) + \alpha_2^\lambda f_2(x, y) + \ldots)}{(\alpha_1^{\overline{\lambda}} f_1(\overline{x}, y) + \alpha_2^{\overline{\lambda}} f_2(\overline{x}, y) + \ldots)} \\
&\times \frac{(\alpha_1^{\overline{\lambda}} f_1(\overline{x}, \overline{y}) + \alpha_2^{\overline{\lambda}} f_2(\overline{x}, \overline{y}) +)}{(\alpha_1^\lambda f_1(x, \overline{y}) + \alpha_2^\lambda f_2(x, \overline{y}) + \ldots)}
\end{aligned} \tag{C.32}$$

and similarly for OR_1:

$$OR_1 = \frac{(f_1(x, y)) + f_2(x, y) + \ldots)(f_1(\overline{x}, \overline{y}) + f_2(\overline{x}, \overline{y}) + \ldots)}{(f_1(\overline{x}, y)) + f_2(\overline{x}, y) + \ldots)(f_1(x, \overline{y}) + f_2(x, \overline{y}) + \ldots)} \tag{C.33}$$

There is an important observation here. Given that there is no admissible sequence relative to $(X, Y, \{\})$, there exists a set W such that W is needed to separate S from X or Y, but also $(W \not\perp\!\!\!\perp \{X, Y\} \mid Z')$, for Z' non-descendents of W and in $Anc(S)$, otherwise there will exist an admissible sequence. If W is different than $\{S\}$, it is the case that, by construction, W is contained in the factor $f_i(x, y)$. Thus, we have an asymmetry given that W, and so $f_i()$,

change depending *simultaneously* on the specific instantiation of X and Y, and consequently eq. (C.32) cannot be simplified in the general case. I.e., the linear combinations encoded in $f_i()$'s at eq. (C.32) do not deteriorate, factoring out independently of the given parametrization given that there is a different element in each one of them.

Now let us consider the following parametrization for P_1: set $P_1(V_i \mid PA_i) = 1/2$ for all families except for the family of the S node (i.e., $P(S = 1 \mid PA_s)$) and the exclusive families included in the factor $f_i(x, y)$ (i.e., for when $X = x, Y = y$). Thus, rewrite OR_2 based on eq. (C.32):

$$OR_2 = \frac{(\alpha_1^\lambda f_1(x,y) + \alpha_2^\lambda f_2(x,y) + \ldots)}{(1/2)^l(\alpha_1^\lambda + \alpha_2^\lambda + \ldots)} \tag{C.34}$$

where l is equal to k minus the number of summands in the respective expression (eq. (C.30)). Let us also rewrite eq. (C.33) accordingly with this given parametrization, which yields:

$$OR_1 = \frac{(f_1(x,y) + f_2(x,y) + \ldots)}{k(1/2)^l} \tag{C.35}$$

After applying some simplifications on eqs. (C.34) and (C.35), we obtain, respectively,

$$OR_2 = \frac{(\alpha_1^\lambda f_1(x,y) + \alpha_2^\lambda f_2(x,y) + \ldots)}{(\alpha_1^\lambda + \alpha_2^\lambda + \ldots)} \tag{C.36}$$

and

$$OR_1 = \frac{(f_1(x,y) + f_2(x,y) + \ldots)}{k} \tag{C.37}$$

Notice that OR_2 in eq. (C.36) is the weighted arithmetic mean of $f_i(.)$'s averaged by α_i^λ's, and OR_1 in eq. (C.37) is the arithmetic mean of $f_i(.)$'s. After simplifications, the remaining parameters lie in the space $[0, 1]^{m+k}$, where m is the number of free parameters in $f_i(.)$'s. Note that $OR_1 - OR_2 = 0$ adds a constraint in this space, and in order to satisfy it we should choose any point in a surface in $[0, 1]^{m+k-1}$ inside $[0, 1]^{m+k}$, i.e., which has Lebesgue measure zero. Consequently, if we randomly choose parameters the equality will *almost never* hold (and the inequality $OR_1 \neq OR_2$ *almost always*), and then just randomly draw the parameters from $[0, 1]^{m+k}$ until this is the case, which finishes this part of the proof. The case of the conditional OR is similar, and we basically have to write appropriately eqs. (C.26) and (C.27) considering C, and exactly the same reasoning applies.

For the case when the graph contains unobservable variables, the proof is essentially the same except that an appropriate parametrization of the underlying generating model should be used – for such, consider the factorization given in (ER11). □

Theorem 26. *Let graph G contain the arrow $X \to Y$, a necessary condition for G to permit the G-recoverability of $OR(Y, X \mid C)$ for a given set C of pre-treatment covariates is that S and every ancestor A_i of S that is also a descendant of X have a separating set T_i that either d-separates A_i from X given Y, or d-separates A_i from Y given X.*

Proof. For the necessity of the condition, we need to show that the failure of any ancestor A_i of S that is also a descendant of X (including S itself) to be separated (from either X or Y) prevents recoverability of $OR(Y, X \mid C)$. Indeed, A_i cannot be part of admissible sequence nor can any of its children be part of an admissible sequence, because in order to separate any such child from either X or Y we would need to condition on the father A_i, and then, the sequence will become non-admissible. Proceeding by induction, we eventually reach S itself, whose failure to enter an admissible sequence renders the existence of such sequence impossible. By Theorem 25, the inexistence of admissible sequence implies the not G-recoverability of $OR(X, Y, C)$. □

Theorem 27. *Let G be a DAG containing the arrow $X \to Y$ and two sets of variables, measured V and unmeasured U. A necessary and sufficient condition for G to permit the G-recoverability of $OR(Y, X \mid C)$ for a given set C of pre-treatment variables is when the sink-procedure below terminates. Moreover, $OR(Y, X \mid C) = OR(Y, X \mid C, Z, T, S = 1)$, where $Z = \big(An(S) \setminus An(Y)\big) \cap V$ and T is given by the sink-procedure (shown in the chapter).*

Proof. We use along the proof some graphoid axioms and other DAG properties as shown in (Pea88). Let us first consider the correctness of the algorithm. The main idea of the reduction sequence is to use each conditional independence (CI) in step 2 of the sink-procedure to substantiate an OR reduction, creating a mapping starting from the s-biased data $OR(X, Y \mid C, Z_1, ..., Z_k, S = 1)$ and reaching the target (unbiased) expression $OR(X, Y \mid C)$. If nodes are not added in step 3 of the algorithm, it is obvious that the sequence induces a valid step-OR reduction, which witnesses the OR G-recoverability. So, let us consider the case when nodes have to be added to T along the execution of the algorithm. At each step i, we reduce $OR(X, Y \mid C, T, Z_1, ..., Z_i)$ to $OR(X, Y \mid C, T, Z_1, ..., Z_{i-1})$ allowed by the CI in step 2. But given that T_i can be added to T along the execution of the algorithm, we need to show that this operation is allowed, i.e., it does not invalidate the construction of the desired mapping between the unbiased

OR and the s-biased one. Towards contradiction, consider an arbitrary node Z_j such that

$$(Z_j \perp\!\!\!\perp X \mid C, T, Y, Z_1, ..., Z_{j-1}) \text{ or } (Z_j \perp\!\!\!\perp Y \mid C, T, X, Z_1, ..., Z_{j-1}) \tag{C.38}$$

Now, consider the first Z_k such that $k < j$ and, in order to satisfy step 2 in the sink-procedure, W has to be added to the conditioning set, then

$$(Z_k \perp\!\!\!\perp X \mid C, T, Y, Z_1, ..., Z_{k-1}, W) \text{ or } (Z_k \perp\!\!\!\perp Y \mid C, T, X, Z_1, ..., Z_{k-1}, W) \tag{C.39}$$

but also

$$(Z_j \perp\!\!\!\perp X \mid C, T, Y, Z_1, ..., Z_{j-1}, W) \text{ or } (Z_j \perp\!\!\!\perp Y \mid C, T, X, Z_1, ..., Z_{j-1}, W) \tag{C.40}$$

is false. If the sink-procedure ends, it is also true that

$$(T \perp\!\!\!\perp Y \mid C, X) \tag{C.41}$$

From eq. (C.38), all paths from Z_j to X or Y (including the ones passing through W) are closed after conditioning on $\{C, T, Y, Z_1, ..., Z_{j-1}\}$. From eq. (C.39) and the minimal choice of T_i in step 3, it must be the case that there is a path p from Z_k to X or Y such that p is blocked by some $W \in \boldsymbol{W}$. From eq. (C.40), there exists a path p' that has to be open after condition on W, and therefore there exists a collider U such that $U = W$ or $W \in Desc(U)$. Let us consider two possible scenarios for p', the first when it goes from Z_j to Y, and the second when it goes from Z_j to X. In the former case, there is an open path from W to Y, which is a contradiction with eq. (C.41) given that $W \subseteq T$. Then it must be the case that W only blocks paths ending in X, so let us assume the case in which the end node in p' is X. From (C.39), p is such that $Z_k \leftarrow\!\rightarrow\!\leftarrow ... - W - ... \rightarrow\!\leftarrow\!\rightarrow X$, where we are condition on all intermediate converging arrows and W must be a chain or a common cause (i.e., $\rightarrow W \rightarrow$ or $\leftarrow W \rightarrow$). Split p into $p_1 : Z_k \dots W$, and $p_2 : W \dots X$. From eq. (C.40), p' is such that W opens a collider U, then the path from Z_j to X. Split p' into $p'_1 : Z_j ... \rightarrow U$ and $p'_2 : U \leftarrow ... X$. Now we have two possibilities. If p_2 is such that $W \rightarrow \dots X$, we can concatenate $Z_k \overset{p'_1}{\rightarrow} U \rightarrow W \overset{p_2}{\rightarrow} X$, which shows an open path from Z_k to X even before conditioning on W, contradiction. If p_2 is such that $W \leftarrow \dots X$, p_1 must be $W \rightarrow \dots Z_k$, and we have two possibilities: (a) Z_k can be a descendent of W, and in this case the collider in U is already open even without conditioning on W, contradiction; (b) W is connected to Z_k through some collider, for instance,

p_1 could be $W \rightarrow \ldots \rightarrow C \leftarrow \ldots Z_k$, but similarly as before, given that we condition on C, which is a descendent of W, and so of U, the collider was already conditioned as well as the path from Z_k to X open, contradiction. Therefore, it cannot be the case that after adding $T_k \subseteq NonDesc(X)$ to block paths from Z_k to X or Y, there is a node Z_j such that $k < j$, and which previously had its paths to X or Y blocked, turned to have them open after conditioning on T_k. Thus, we are allowed to modify each CI obtained in step 2 before Z_k in the sequence adding T_k, and then based on the admissible sequence starting from $OR(X, Y \mid C, T, Z_1, ..., Z_n)$, we can reduce it through this new augmented CIs of step 2 until reaching the desired expression $OR(X, Y \mid C)$.

Now we consider the complexity of the algorithm, and we show that it runs in polynomial time. Notice that only the step 3 of the algorithm could imply some backtracking – i.e., when it chooses a (minimal) set T_i of non-descendants of X that renders the equality in step 2 to be true. The choice of separating set *per se* is polynomial, see footnote 11.

Consider that the choice of T_i implies failure in step 5 when it tests the validity of $(T \perp\!\!\!\perp Y \mid X, C)$. Assume that it exists a sequence Q of ancestors of S and not ancestors of X, $(Z_1, ..., Z_k, ..., Z_n)$ such that for each Z_i there is a separating set T_i which makes the independence test valid. Let $T = \bigcup T_i$, and assume that $(T \perp\!\!\!\perp Y \mid X, C)$ holds. Assume now that in round k, the sink procedure chooses a different (minimal) separating set than T_k, and call this new set T'_k, and subsequently $(T'_{k+1}, ..., T'_n)$. We have the new sequence Q' with additional separators $(T_1, ..., T_{k-1}, T'_k, ..., T'_n)$. Call $T' = \bigcup T'_i$, and $\Delta = T' \setminus (T \cap T')$.

We have that $(T' \not\perp\!\!\!\perp Y \mid X, C)$ holds, or just $(\Delta \not\perp\!\!\!\perp Y \mid X, C)$. (This follows from $(\Delta \perp\!\!\!\perp Y \mid X, C)$, which by composition yields $(T' \perp\!\!\!\perp Y \mid X, C)$, contradiction. See also (PP10).) Let $\delta \in \Delta$ be the first node such that that Q and Q' disagree and which make step 5 to fail. δ blocks at least one path from Z_k to X (after condition on $\{C, Y, T, Z_1, ..., Z_{k-1}, T_i \setminus \delta\}$) or from Z_k to Y (after condition on $\{C, X, T, Z_1, ..., Z_{k-1}, T_i \setminus \delta\}$), otherwise the sequence will not be admissible (pass in the test of step 2). By construction, it must be the case that there is an open path from Z_k to Y passing through δ (after cond. on $\{C, X, Q, Z_1, ..., Z_{k-1}, T_i \setminus \delta\}$).

Let p be part of this path from δ to Y (or, $\delta - ... - Y$). There must exist in Q a vertex v which blocks this same path from Z_k to $\{X, Y\}$ or $\{Y\}$ in the test of step 2. But v is in p or connected through an open path p' to δ (i.e., $p : \delta - ... - v - ... - Y$ or $v - ... - p' - ... - \delta - ... - p - ... - Y$), otherwise we would not need δ in the first place, contradicting minimality. In both cases, there is an open path from v to Y, which contradicts the assumption about Q validating $(T \perp\!\!\!\perp Y \mid X, C)$ as true, and therefore it cannot exist such δ. Applying the same reasoning for the whole sequence Q' inductively, we conclude that it cannot exist

M	1	2	3	4	5	6	7	8	9	10	11	12
1	$(c_1-1)b_1$	c_1b_2	c_1b_3	c_1b_4								
2	c_2b_1	$(c_2-1)b_2$	c_2b_3	c_2b_4								
3	c_3b_1	c_3b_2	$(c_3-1)b_3$	c_3b_4								
4					$(c_4-1)b_1$	c_4b_2	c_4b_3	c_4b_4				
5					c_5b_1	$(c_5-1)b_2$	c_5b_3	c_5b_4				
6					c_6b_1	c_6b_2	$(c_6-1)b_3$	c_6b_4				
7									$(c_7-1)b_1$	c_7b_2	c_7b_3	c_7b_4
8									c_8b_1	$(c_8-1)b_2$	c_8b_3	c_8b_4
9									c_9b_1	c_9b_2	$(c_9-1)b_3$	c_9b_4
10	$(1-c_{10})b_1$	$-c_{10}b_2$	$-c_{10}b_3$	$-c_{10}b_4$	$(1-c_{10})b_1$	$-c_{10}b_2$	$-c_{10}b_3$	$-c_{10}b_4$	$(1-c_{10})b_1$	$-c_{10}b_2$	$-c_{10}b_3$	$-c_{10}b_4$
11	$-c_{11}b_1$	$(1-c_{11})b_2$	$-c_{11}b_3$	$-c_{11}b_4$	$-c_{11}b_1$	$(1-c_{11})b_2$	$-c_{11}b_3$	$-c_{11}b_4$	$-c_{11}b_1$	$(1-c_{11})b_2$	$-c_{11}b_3$	$-c_{11}b_4$
12	$-c_{12}b_1$	$-c_{12}b_2$	$(1-c_{12})b_3$	$-c_{12}b_4$	$-c_{12}b_1$	$-c_{12}b_2$	$(1-c_{12})b_3$	$-c_{12}b_4$	$-c_{12}b_1$	$-c_{12}b_2$	$(1-c_{12})b_3$	$-c_{12}b_4$

such sequence. Therefore, step 5 does not imply any backtracking.

Similarly, let us consider the case when the choice of T_j implies failure in a subsequent step 2. In the sequence Q', it is true that when the algorithm chooses T_j to satisfy the admissibility of Z_j, it blocks some paths from Z_j to X. Now, assume that for Z_k, $k < j$, there is an open path through T_j, i.e., $Z_k \longleftrightarrow U \longleftrightarrow X$, where $U = T_j$ or $T_j \in Desc(U)$. But if you do not choose T_j (or any other node that blocks this path), we would have an open path from Z_k to X through T_j, contradiction.

We now argue about the completeness of the procedure. Let us first consider the case in which there is not X-independent variable in the admissible sequence, the sink-procedure will return an admissible sequence whenever one exists. Notice that the sink-procedure performs a search for an admissible sequence in reverse topological order, and this only makes the conditional independence's tests easier than in any other order. This is so because in each step, we are adding all non-descendents of Z_k (are non-colliders for Z_k), which completely disconnects Z_k from X or Y except for paths passing through non-descendents of X. (Also, non step-wise reductions can be converted to step-wise one through the graphoids decomposition and weak union.)

Assume that there is a sequence $(A_1, ..., A_m)$ called A that does not follow the order given by the sink-procedure and it is admissible. Now, let us call Q the sequence $(Z_1, ..., Z_n)$ given by the sink-procedure, and further assume that Q is not admissible. It is true that the last element of both sequences is S, and in Q we would have the blocking set $\{Z_1, ..., Z_{n-1}\}$ while in A we would have $\{A_1, ..., A_{m-1}\}$. It is true that $\{A_1, ..., A_{m-1}\} \subseteq \{Z_1, ..., Z_{n-1}\}$, and this is an invariant along the algorithm for all nodes in A. Recall two facts: (a) for now, we are assuming that there are not disagreements between T_Q and T_A; (b) adding descendents of Z_k in each step can only open some paths and spoil separation. It must be the case for the sink-procedure to fail, there exists $Z_k \in Q$ such that $(Z_k \perp\!\!\!\perp X \mid Y, C, Z_1, ..., Z_{k-1})$ and $(Z_k \perp\!\!\!\perp Y \mid X, C, Z_1, ..., Z_{k-1})$ are both false. Thus, there is at least one path from Z_k to X and from Z_k to Y that are not blocked by $\{Z_1, ..., Z_{k-1}\} \cup \{C\}$ (and respectively, $\{Y\}$ and $\{X\}$); call the set of these paths P_1 and P_2, respectively.

Assume that A also chooses Z_k at some point along its execution, and Z_k is labeled there A_m. It must be the case that all paths from A_m to X or all paths from A_m to Y are blocked by $\{A_1, ..., A_{m-1}\} \cup \{C\}$ (and respectively, $\{Y\}$ and $\{X\}$). But if $\{A_1, ..., A_{m-1}\} \subseteq \{Z_1, ..., Z_{k-1}\}$, this is a contradiction. Now assume that A does not choose Z_k along its execution. There are ancestors of S which have to block P_1 from S to X or P_2 from S to Y, and we consider without loss of generality the subset $\{A_1, ..., A_l\}$ that renders this separation to hold. Consider A_j the first descendant of Z_k in G^* that is in $\{A_1, ..., A_l\}$. If such node is S, we reach a contradiction. Assume that A_j is not S but some of its ancestors. To separate A_j from X or Y, we need to block the paths from it to X or Y, but there are unblockable paths P_1 and P_2 passing through Z_k ($A_j \leftarrow ... - Z_k - P_1 - X$ or $A_j \leftarrow ... - Z_k - P_2 - Y$), and therefore A_j cannot be part of an admissible sequence, contradiction. Then, it is the case that if both algorithms do not disagree in the choice of the non-descendents of X, there is indeed not admissible sequence. For the case when we add X-independent variables along the sequence, the result also follows, and this is so based on the fact shown previously that there is no backtracking in the choice of T_i, and any algorithm that chooses T_i consistently obtains the same outcome in terms of separation. Each time that the sink-procedure does not return any sequence, we can produce a counter-example for the G-recoverability of the triplet (X, Y, C) based on the construction of Theorem 25. □

Theorem 28. *The joint distribution of $P(X, Y, Z)$ is recoverable from s-biased data whenever the following conditions hold: (i) the S node is affected by the set Z only through $\{X, Y\}$; (ii) the set Z is d-connected to $\{X, Y\}$ (and combinations); (iii) the dimensionality of Z matches the dimensionality of $\{X, Y\}$; (iv) the marginal probability of Z is known. In other words, the distribution $P(X, Y, Z)$ is recoverable from s-biased data whenever $(S \perp\!\!\!\perp Z \mid X, Y)$, $(Z \not\perp\!\!\!\perp \{X, Y\})$, $(Z \not\perp\!\!\!\perp X \mid Y)$,$(Z \not\perp\!\!\!\perp Y \mid X)$, the dimensionality of Z and $X \cup Y$ matches, and the marginal distribution of $P(Z)$ is given.*

Proof. Let us first show the result for the binary case. To match the dimensionality requirement, we assume that $Z = Z_1 \cup Z_2$ and both Z_1 and Z_2 are binary satisfying:

$$P(Z_1, Z_2 \mid X, Y, S) \;=\; P(Z_1, Z_2 \mid X, Y) \tag{C.42}$$

To simplify the notation, let us write:

- $P(X = x, Y = y \mid Z_1 = z_1, Z_2 = z_2) = \alpha_{xy,z_1z_2}$
- $P(Z_1 = z_1, Z_2 = z_2) = \beta_{z_1z_2}$
- $P(Z_1 = z_1, Z_2 = z_2 \mid X = x, Y = y) = \gamma_{z_1z_2,xy}$

Note that the parameters $\gamma_{z_1z_2,xy}$ and $\beta_{z_1z_2}$ impose constraints on the distribution α_{xy,z_1z_2}, which can be made explicit by the following equation,

$$\gamma_{z_1z_2,xy} = \frac{\alpha_{xy,z_1z_2}\beta_{z_1z_2}}{\sum_{z'_1,z'_2}\alpha_{xy,z'_1z'_2}\beta_{z'_1z'_2}} \tag{C.43}$$

Now, for a given assignment $< X = 0, Y = 0 >$, let us list all independent parameters $\gamma_{z_1z_2,00}$,

$$\begin{aligned}
\gamma_{00,00} &= \frac{\alpha_{00,00}\beta_{00}}{\sum_{z'_1,z'_2}\alpha_{00,z'_1z'_2}\beta_{z'_1z'_2}} \\
\gamma_{01,00} &= \frac{\alpha_{00,01}\beta_{01}}{\sum_{z'_1,z'_2}\alpha_{00,z'_1z'_2}\beta_{z'_1z'_2}} \\
\gamma_{10,00} &= \frac{\alpha_{00,10}\beta_{10}}{\sum_{z'_1,z'_2}\alpha_{00,z'_1z'_2}\beta_{z'_1z'_2}}
\end{aligned} \tag{C.44}$$

Note that $\gamma_{11,00}$ is not an independent parameter because it is completely determined by the other three equations in (C.44) given the integrality constraint. For now, we have 3 equations and 4 unknown variables ($\{\alpha_{00,00}, \alpha_{00,01}, \alpha_{00,10}, \alpha_{00,11}\}$.)

Similarly, we write the constraints for the assignments $< X = 1, Y = 0 >$ and $< X = 0, Y = 1 >$, respectively,

$$\gamma_{00,10} = \frac{\alpha_{10,00}\beta_{00}}{\sum_{z'_1,z'_2}\alpha_{10,z'_1z'_2}\beta_{z'_1z'_2}}, ... \tag{C.45}$$

$$\gamma_{00,01} = \frac{\alpha_{01,00}\beta_{00}}{\sum_{z'_1,z'_2}\alpha_{01,z'_1z'_2}\beta_{z'_1z'_2}}, ... \tag{C.46}$$

Now, we can write the equations for the constraints relative to the variables α_{11,z_1z_2} as a function of the previous variables $\{\alpha_{00,z_1z_2}, \alpha_{01,z_1z_2}, \alpha_{10,z_1z_2}\}$,

$$\begin{aligned}
\gamma_{00,11} = \Big((1-\big(\alpha_{00,00}+\alpha_{01,00}+\alpha_{10,00}\big))\beta_{00}\Big)\Big/ \\
\Big(\sum_{Z'_1,Z'_2}\Big(1-\big(\alpha_{00,z'_1z'_2}+\alpha_{01,z'_1z'_2}+\alpha_{10,z'_1z'_2}\big)\Big)\beta_{z'_1z'_2}\Big), ...
\end{aligned} \tag{C.47}$$

Notice that the parameters $\gamma_{z_1z_2,11}$ are independent, and we have 12 equations and 12 unknowns, but it remains to show that the equations are all independent (notice that the last three constraints in eq. (C.47) involve variables of the other constraints). Another fact to observe is that the system is indeed linear. We show that the matrix M, induced by the eqs. (C.44, C.45, C.46, C.47), is linear and (almost surely) invertible, and generates an unique solution. M is invertible if and only if its determinant is non-zero. For convenience, let us display the variables α_{xy,z_1z_2} column-wise, renaming $\beta_{z_1z_2}$ as constants $b_1 - b_4$, and $\gamma_{z_1z_2,xy}$ as constants $c_1 - c_{12}$. The matrix is shown on the top of page 86.

In what follows, we exploit the block structure of M and apply the following transformations to better visualize its determinant.

1. First note that all columns $\{1, 5, 9\}$ are multiplied by b_1, which can be factored out by the determinant property. Similarly for the other columns in respect to $\{b_2, b_3, b_4\}$, which can be expressed as $det(M) = (b_1 b_2 b_3 b_4)^3 det(M^{(1)})$, where $M^{(1)}$ is the resultant matrix.
2. Let us sum lines $\{1, 4, 7\}$ to line 10, lines $\{2, 5, 8\}$ to line 11, and $\{3, 6, 9\}$ to line 12, which generate matrix $M^{(2)}$.
3. We now sum the columns of $M^{(2)}$, -1 times column 4 to column 1, -1 times column 4 to column 2, and -1 times column 4 to column 3 (similarly for the other blocks), which yields $M^{(3)}$.
4. Sum the columns of $M^{(3)}$, c_1 times column 1, c_2 times column 2 and c_3 times column 3 to column 4 (similarly for the other blocks), yielding $M^{(4)}$.
5. Now, reorder the columns, "pushing" column 4 and 8 towards the end, call the resultant matrix $M^{(5)}$.

Now we are done, notice that the $det(M) = (b_1 b_2 b_3 b_4)^3 det(M^{(5)})$, and the determinant of $M^{(5)}$ is the determinant of two block matrices, the square matrix $M_1^{(5)}$ from lines 1-9 multiplied by another square matrix $M_2^{(5)}$ from lines 10-12. Note that $det(M_1^{(5)}) = -1$, and remains to show that $det(M_2^{(5)})$ is almost always different than zero. The parameters c_1 to c_{12} are independent, and given the form obtained to $M_2^{(5)}$ where all entries are independent, this implies that $M_2^{(5)}$ is non-singular almost surely, and so it is $M^{(5)}$ – coincidental cancellations will occur with Lebesgue measure zero.

Therefore, we consider M as full rank, which can be solved algebraically with standard techniques yielding the solution $\alpha = M^{-1}\gamma$. This result, together with $P(Z)$ yields the joint distribution $P(Y, X, Z)$. The case for non-binary variables follows in a straightforward way, just noticing the requirement for agreement between the dimensions of the IV set Z and $\{X, Y\}$. □

Corollary 20. *The causal effect of Oestrogen (X) on Endometrial Cancer (Y) as studied in (HF78; HHR04) (Fig. 5.7(c)) is recoverable from s-biased data whenever there is an IV set Z pointing to X, and the conditions of the Theorem 28 hold. Moreover, the same holds without relying on Z whenever the following conditions hold: (i) X has the same dimensionality of $\{W, Y\}$; (ii) the marginal distribution of $P(X)$ is available.*

Proof. First, apply Theorem 28 to the variables $\{W, Y\}$ replacing X with W, and obtain $P(W, Y)$. Further note that $P(X \mid Y, W, S = 1) = P(X \mid Y, W)$, which together with the first observation finishes this part of proof. The proof for when we do not rely on Z is essentially the same. □

APPENDIX D

Proofs for Chapter 6

Theorem 31. *Let X, Y, Z be disjoint sets of variables and let G be the causal diagram. The causal effect $Q = P(y|do(x))$ is $\boldsymbol{zI\mathcal{D}}$ from $do(Z)$ in G if and only if one of the following conditions hold:*

a. Q is identifiable in G; or,

b. There exists $Z' \subseteq Z$ such that the following conditions hold,

(i) X intercepts all directed paths from Z' to Y, and

(ii) Q is identifiable in $G_{\overline{Z'}}$.

Proof sketch. The sufficiency part is direct. If condition a. holds, the result follows trivially. If condition b. holds, consider the set $Z' \subseteq Z$ that satisfies both conditions of the Theorem. Using condition b.(i), we can apply Rule 3 of *do*-calculus in Q since X intercepts all directed paths from Z' to Y, $(Z' \perp\!\!\!\perp X)_{G_{\overline{X}}}$, yielding $P(Y|do(X)) = P(Y|do(X), do(Z'))$. It is not difficult to see that this last expression, together with the fact that Z' is a root set and b.(ii) hold, imply the result.

It is more involved to prove necessity, and we consider its contrapositive.

Lemma 25. *If conditions a. or b. of Theorem 31 do not hold for all $Z' \subseteq Z$, Q is not z-identifiable from G and $do(Z)$.*

Proof. Consider an arbitrary $Z' \subseteq Z$ and the following Lemmas.

Lemma 26. *If conditions a. and b.(ii) of Theorem 31 do not hold for Z', Q is not z-identifiable from G and $do(Z')$.*

Proof. The result follows from Theorem 30. □

Lemma 27. *If conditions a. and b.(i) of Theorem 31 do not hold, and condition b.(ii) does hold for Z', Q is not z-identifiable from G and $do(Z')$.*

Proof. Let I be the set of interventional distributions $P(V \setminus Z|do(Z))$. The proof will show a counter-example to the z-identifiability of $P_x(y)$ through two models M_1 and M_2 that agree upon $\langle P, I\rangle$ and disagree on $Q = P_x(y)$. We need to find *any* two models where $\boldsymbol{zID}$ fails, and in particular, we are free to pick models not faithful (SGS93).

Since condition a. does not hold, Q is not $\boldsymbol{ID}$ in G, and so there exists a hedge $\mathcal{F} = \langle F, F'\rangle$ for $Q^* = P_{x^*}(y^*)$ in G, for $X^* \subseteq X$, $Y^* \subseteq Y$ (SP06b, Corol. 3).

The fact that condition b.(ii) holds implies that there exists no hedge for Q in $G_{\overline{Z'}}$. This, in turn, implies the existence of at least one bidirected edge in $\mathcal{F}$ that is broken by the virtue of the mutilation $do(Z')$. Therefore, it is clear that Z' must be part of the R-rooted C-forest F, where $R = An(Y^*)_{\overline{X^*}}$. Without loss of generality, we consider Z' that is related to the troublesome structure $\mathcal{F}$ in its respective induced subgraph (to be defined next).

Since condition b.(i) does not hold, there exist directed paths from Z' to R not blocked by X. Without loss of generality, let us consider the set of paths π together with the C-forest $De(F)_G \cap An(Y^*)_{G_{\overline{X}}}$, and call this structure H. Note that the existence of π prevents us from adding to Q elements of Z' as interventions using the 3rd rule of *do*-calculus, but the interventional distribution of Z' will eventually appear based on the C-decomposition as shown next.

When a graph is not a C-component itself, it can be decomposed (uniquely) into C-components (Tia02). Consider the decomposition $C(H \setminus X^*) = \{S^*, S^0\}$, where S^0 is a shorthand for the C-components $\{S_1, ..., S_k\}$, and $F' \cap S^* \neq 0$. Based on this decomposition, it is an equivalence in *do*-calculus $Q = f(H, X^*, S^*, S^0)$, where f is the C-factorization. Rewriting in a convenient way the usual Tian's $f(.)$, we have the following equality in *do*-calculus,

$$Q = \sum_{v \setminus x \cup y} P_{v \setminus s^*}(s^*) \prod_{S_i \in S^0} P_{v \setminus s_i}(s_i) \tag{D.1}$$

Consider a partition of Z' such that $Z^* = Z' \cap S^*$ and $Z^0 = Z' \cap S^0$. For simplicity, we also use S^0 as the union of the elements of the respective components, since the context is clear.

Case 1. We first show that the Lemma is valid for $Z^0 = \emptyset$. We construct two models to show non-$\boldsymbol{zID}$ as discussed above.

Let V be the set of observable variables and U be the set of unobservable variables in F. Consider a parametrization in which all variables are boolean and the exogenous variables are fair coins (i.e., distributed uniformly). In M_1 each variable computes the bit parity of its respective parent values (observables and unobservables), while in M_2 each variable does the same except for nodes in F'

that compute the bit parity but ignore the values of the parents in F.

Note that these two models are essentially the same as the ones used to prove ordinary non-$I\!D$ given by Shpitser. It is not difficult to see that the two requirements for non-$I\!D$ hold in these models; we briefly outline the argument in the sequel.

The equality of P between the two models follows since the variables in $F \setminus F'$ are equally computing the bit parity of the ancestral exogenous variables and have identical functional and probabilistic form in both models. For each variable $V_i \in (F' \setminus R)$, there is always at least one exogenous variable $U \in U$ that can freely vary in the evaluation of $f_i(.)$, which together with the uniformity of U imply that they are all uniformly distributed. For the elements of the root set $V_i \in R$, in both models, they compute their respective bit parity as even, which happens with the same probability also given the uniformity of U. For these variables in R, the same trick as shown in Theorem 3 in (SP06b) to yield positivity can be applied here.

For the inequality of Q, it is clear that in M_1, R continues varying uniformly given the intervention on $do(X^*)$, while in M_2 it is insensitive to what happens in $F \setminus F'$, and so it computes the bit parity as even independent of the value of X^*. Given space constraints and to avoid redundancy, we refer to Theorem 3 in (SP06b) for a more detailed proof of these two facts.

It remains to show that both models agree on the distributions of I. Consider the following subclaim that will help to establish this fact.

Subclaim. Let X and Y be two binary variables such that $P(X = x) = p$ and $P(Y = y) = q = 1/2$. Then the probabilistic input/output behavior of $Z = (X \oplus Y)$ is the same of Y and uniformly distributed. The variable $Z = 1$ whenever $\{(X = 1, Y = 0), (X = 0, Y = 1)\}$, which happens with probability $pq+(1-p)(1-q)$. Since $q = 1/2$, the expression reduces to $p*1/2+(1-p)*1/2 = 1/2$.

Note that $Z' \subset S^*$, and so both models agree on the input/output functional description in the C-forests F' except for some additional incoming bidirected edges from F. It turns out that this does not affect the distribution $do(Z')$ since there is always a free variable U in the evaluation of the function, and so by the previous subclaim, both models induce the same distribution I in F', finishing the proof of this case.

We consider now the complementary case when $Z^0 \neq \emptyset$. There are two subcases to consider here. By assumption Q is not $I\!D$, and we are trying to evaluate whether the C-factor $Q^* = P_{v \setminus s^*}(s^*)$ is or is not $I\!D$ when experiments over Z' are available. Note that the original hedge is related to this factor Q^* since $R \subseteq F'$ and (SP06b)[Corol. 3].

Case 2. Consider first the case in which Q^* is $\mathcal{ID}$ from $do(Z')$. The idea is to find a hedge $\mathcal{F}^*$ (different than $\mathcal{F}$) relative to the elements of Z' that helped to yield identifiability for Q (breaking the original $\mathcal{F}$), and show that the identification of this new hedge is not recoverable using other elements of Z'. The following subclaim implements this strategy.

Subclaim. If the conditions of the Lemma hold, $Z^0 \neq \emptyset$, and Q^* is $\mathcal{ID}$ from $do(Z')$, there exist a pair of R^*-rooted C-forests F^*, F'^* forming a hedge $\mathcal{F}^*$ for Q such that $F^* \subset F$ and $(F^* \setminus F'^*) \cap Z^0 = \emptyset$.

Since Q is $\mathcal{ID}$ in $G_{\overline{Z'}}$, there is no hedge there. We have already argued that Z^0 participates in the hedge for this factor since the mutilation $do(Z')$ was able to destroy $\mathcal{F}$. Since Z^0 is non-empty, it is also the case that its elements (and the other variables in the respective components) are not in the same component of R in $H \setminus X^*$.

For simplicity, define $S^z = \{S_k | Z^0 \cap S_k \neq 0\}$, and we call S^z interchangeably as the union of the elements of the respective components when the context is clear. Define $W = H \setminus X^* \setminus S^z$, and then it is true that W is not in the same component of S^z after $do(X^*)$. Since Q was non-$\mathcal{ID}$ in G but became $\mathcal{ID}$ in $G_{\overline{Z'}}$, together with the fact that W and S^z are not bidirectely connected in $H_{\overline{X^*}}$, imply the existence of a bidirected connection between Z^0 and X^*.

Note that S^z partition the C-components involving Z^0, and we would be done if there is a hedge $\mathcal{F}^*$ for the C-factor $S_0^z \in S^z$ in eq. (D.1). Assume that this is not the case. We can apply the same argument as before, and consider now the most interesting case when we reach the last component $S_m^z \in S^z$. Assume that there is no hedge for the respective C-factor for S_m^z. But this implies that there exists an outgoing directed edge in $\mathcal{F}$ connecting X to W, contradiction since Q was not $\mathcal{ID}$. So, we can certainly construct a hedge $\mathcal{F}^*$ for Q^* removing unnecessary bidirected edges from F while keeping the directed edges from X, which proves the subclaim.

Finally, the existence of $\mathcal{F}^*$ implies that Q^{**} is not $z\mathcal{ID}$ by similar reason of case 1. And so, we can use the same construction in the respective induced graph for S^z and X^*.

Case 3. Consider the case in which $Z^0 \neq \emptyset$ and Q^* is not $\mathcal{ID}$ from $do(Z')$. Recall that we need to construct two models that agree on the distribution P, disagree in Q, but also agree on I. It is not immediately clear how to use the construction given case 1 since Z^0 was not in $F \setminus F'$, which we relied on to produce the two models witnessing non-$z\mathcal{ID}$ for Q.

The idea is to fix the original hedge $\mathcal{F}$ producing a new hedge $\mathcal{F}^*$ that matches the requirement of the construction of the case 1, which proved to be feasible given the assumptions of this case. The following subclaim helps to solve this problem.

Subclaim. If the conditions of the Lemma hold, $Z^0 \neq \emptyset$, and Q^* is not $\mathcal{ID}$ given $do(Z')$, there exist a pair of R^*-rooted C-forests F^*, F'^* forming a hedge $\mathcal{F}^*$ for Q such that $F^* \subset F$ and $(F^* \setminus F'^*) \cap Z^0 = \emptyset$.

Similar observations as in the previous case apply here, but note that since Q^* is not $\mathcal{ID}$ from $do(Z')$ now, the hedge $\mathcal{F}$ contains also Z^0 besides X^*, Y^*; without loss of generality, assume $Z^* \subseteq Z'$ is the relevant subset for this hedge.

By construction, Z^0 and Y^* are in different components in $H \setminus X^*$, and so there are no bidirected edges connecting them in $H_{\overline{X^*}}$. This implies again X^* is in the same induced component of Y^* after $do(Z')$ since Q is not $\mathcal{ID}$ from $do(Z')$.

Now set the graph F^* as a copy of the C-forest F, and then remove from F^* the components involving Z^0, which are just some unnecessary edges and nodes relative to them. Note that F^* is still a R-rooted C-forest since R, W and X^* remain there; it is also true that there exists a directed path from X^* to R^* not passing through Z^0 that was kept in F^*. Assume that this is not the case. Then, we could remove X^* using the third rule of do-calculus in the first place, but since Z^0 is in a different component of Y^*, we would obtain $\mathcal{ID}$ from Z^0. Contradiction. The subclaim and the lemma follow.

Finally, there are just two remaining observations in regard to the function created to generate the counter-example for the previous cases. First, if $R \cap Y^* = \emptyset$, the sum in eq. (2) can degenerate when the other C-factors are identifiable, which happens because the chosen parametrization induces the same uniform distribution over observables in both models. Note that these terms would be factored out in the expression for Q, and since we are summing over all configurations of Z^* (and all variables different than $\{X^*, Y^*, Z^*\}$), the target effect will collapse to the same value in M_1 and M_2, spoiling the counter-example. It is not difficult to solve this problem by equally perturbing the parameters in both models and making the observational distributions agree (which is required by $z\mathcal{ID}$), but not uniformly distributed. The same reasoning applies when other C-factors are non-identifiable, i.e., we also have to make sure that they do not map to the same value.

The other observation is that the strategy employed induces a counter-example over H, which need to be extended to the inputted diagram G. But this is also not difficult since by construction, $X \cap H = X^*$, and we can just extend the parametrization making the other variables independent of H. Now, we have a witness for non-$z\mathcal{ID}$ of $P_x(y)$ from $\langle P, do(Z)\rangle$ in G. □

The Lemmas 26 and 27 can be combined to obtain Lemma 25. □

Lemma 25 suffices to prove necessity, completing the proof of Theorem 31. □

Bibliography

[AC96] S. Acid and L.D. Campos. "An Algorithm for Finding Minimum d-Separating Sets in Belief Networks." In *Proceedings of the 12th Annual Conference on Uncertainty in Artificial Intelligence (UAI-96)*, pp. 3–10, San Francisco, CA, 1996. Morgan Kaufmann.

[Ald89] J. Aldrich. "Autonomy." *Oxford Economic Papers*, **41**:15–34, 1989.

[Ang97] Joshua D Angrist. "Conditional independence in sample selection models." *Economics Letters*, **54**(2):103–112, 1997.

[Bak06] S.G. Baker. "Surrogate Endpoints: Wishful Thinking or Reality?" *Journal of the National Cancer Institute*, **98**(8):502–503, 2006.

[Ber46] J. Berkson. "Limitations of the application of fourfold table analysis to hospital data." *Biometrics Bulletin*, **2**:47–53, 1946.

[BLH13] E. Bareinboim, S. Lee, V. Honavar, and J. Pearl. "Transportability from Multiple Environments with Limited Experiments." In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, pp. 136–144. Curran Associates, Inc., 2013.

[BMB00] M. Buyse, G. Molenberghs, T. Burzykowski, D. Renard, , and H. Geys. "The validation of surrogate endpoints in meta-analyses of randomized experiments." *Biostatistics*, (1):49–68, 2000.

[Bol89] K.A. Bollen. *Structural Equations with Latent Variables*. John Wiley, New York, 1989.

[BP97] A. Balke and J. Pearl. "Bounds on treatment effects from studies with imperfect compliance." *Journal of the American Statistical Association*, **92**(439):1172–1176, September 1997.

[BP12a] E. Bareinboim and J. Pearl. "Causal Inference by Surrogate Experiments: z-Identifiability." In Nando de Freitas and Kevin Murphy, editors, *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence (UAI 2012)*, pp. 113–120. AUAI Press, 2012.

[BP12b] E. Bareinboim and J. Pearl. "Controlling Selection Bias in Causal Inference." In M. Girolami and N. Lawrence, editors, *Proceedings of The Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2012)*, pp. 100–108. JMLR (22), 2012.

[BP12c] E. Bareinboim and J. Pearl. "Transportability of Causal Effects: Completeness Results." In Joerg Hoffmann. and Bart Selman, editors, *Proceedings of The Twenty-Sixth Conference on Artificial Intelligence (AAAI 2012)*, pp. 698–704, 2012.

[BP13a] E. Bareinboim and J. Pearl. "Causal Transportability with Limited Experiments." In M. desJardins and M. Littman, editors, *Proceedings of the Twenty-Seventh National Conference on Artificial Intelligence (AAAI 2013)*, pp. 95–101, Menlo Park, CA, 2013. AAAI Press.

[BP13b] E. Bareinboim and J. Pearl. "Meta-Transportability of Causal Effects: A Formal Approach." In C. Carvalho and P. Ravikumar, editors, *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2013)*, pp. 135–143. JMLR W&CP 31, 2013.

[BP14] E. Bareinboim and J. Pearl. "Transportability from Multiple Environments with Limited Experiments: Completeness Results." Technical Report Technical Report R-435, Cognitive Systems Laboratory, Department of Computer Science, UCLA, 2014.

[BTP14] E. Bareinboim, J. Tian, and J. Pearl. "Recovering from Selection Bias in Causal and Statistical Inference." In C. Brodley and P. Stone, editors, *Proceedings of the Twenty-Eight National Conference on Artificial Intelligence (AAAI 2014)*, Menlo Park, CA, 2014. AAAI Press.

[CMR08] C. Cortes, M. Mohri, M. Riley, and A. Rostamizadeh. "Sample Selection Bias Correction Theory." In *Proceedings of the 19th International Conference on Algorithmic Learning Theory*, ALT '08, pp. 38–53, Berlin, Heidelberg, 2008. Springer-Verlag.

[Coo95] G. Cooper. "Causal discovery from data in the presence of selection bias." *Artificial Intelligence and Statistics*, pp. 140–150, 1995.

[Cor51] J. Cornfield. "A method of estimating comparative rates from clinical data; applications to cancer of the lung, breast, and cervix." *Journal of the National Cancer Institute*, **11**:1269–1275, 1951.

[Cox58] D.R. Cox. *The Planning of Experiments*. John Wiley and Sons, NY, 1958.

[CS63] D. Campbell and J. Stanley. *Experimental and Quasi-Experimental Designs for Research*. Wadsworth Publishing, Chicago, 1963.

[CS10] S.R. Cole and E.A. Stuart. "Generalizing Evidence From Randomized Clinical Trials to Target Populations." *American Journal of Epidemiology*, **172**:107–115, 2010.

[Daw02] A.P. Dawid. "Influence diagrams for causal modelling and inference." *International Statistical Review*, **70**:161–189, 2002.

[DKK10] V. Didelez, S. Kreiner, and N. Keiding. "Graphical Models for Inference Under Outcome-Dependent Sampling." *Statistical Science*, **25(3)**:368–387, 2010.

[DM06] H. Daume III and D. Marcu. "Domain Adaptation for Statistical Classifiers." *Journal of Artificial Intelligence Research*, **26**:101–126, 2006.

[Dun75] O.D. Duncan. *Introduction to Structural Equation Models*. Academic Press, New York, 1975.

[EH89] S.S. Ellenberg and J.M. Hamilton. "Surrogate endpoints in clinical trials: Cancer." *Statistics in Medicine*, **8**:405–413, 1989.

[Elk01] C. Elkan. "The Foundations of Cost-sensitive Learning." In *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'01, pp. 973–978, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[ER11] R.J. Evans and T.S. Richardson. "Marginal log-linear parameters for graphical Markov models." arXiv:1105.6075 [stat.ME], 2011.

[FGS92] L.S. Freedman, B.I. Graubard, and A. Schatzkin. "Statistical validation of intermediate endpoints for chronic diseases." *Statistics in Medicine*, **8**:167–178, 1992.

[FR02] C.E. Frangakis and D.B. Rubin. "Principal Stratification in Causal Inference." *Biometrics*, **1**(58):21–29, 2002.

[Gen92] Z. Geng. "Collapsibility of relative risk in contingency tables with a response variable." *Journal Royal Statistical Society*, **54**(2):585–593, 1992.

[GG08] M.M. Glymour and S. Greenland. "Causal Diagrams." In K.J. Rothman, S. Greenland, and T.L. Lash, editors, *Modern Epidemiology*, pp. 183–209. Lippincott Williams & Wilkins, Philadelphia, PA, 3rd edition, 2008.

[Gol72] A.S. Goldberger. "Structural Equation Models in the Social Sciences." *Econometrica: Journal of the Econometric Society*, **40**:979–1001, 1972.

[GP95] D. Galles and J. Pearl. "Testing identifiability of causal effects." In P. Besnard and S. Hanks, editors, *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI 1995)*, pp. 185–195. Morgan Kaufmann, San Francisco, 1995.

[GP11] S. Greenland and J. Pearl. "Adjustments and their Consequences – Collapsibility Analysis using Graphical Models." *International Statistical Review*, **79**(3):401–426, 2011.

[GPR99] S. Greenland, J. Pearl, and J.M Robins. "Causal diagrams for epidemiologic research." *Epidemiology*, **10**(1):37–48, 1999.

[GRB09] S. Geneletti, S. Richardson, and N. Best. "Adjusting for selection bias in retrospective, case-control studies." *Biostatistics*, **10(1)**, 2009.

[GVP90] D. Geiger, T.S. Verma, and J. Pearl. "Identifying Independence in Bayesian Networks." In *Networks*, volume 20, pp. 507–534. John Wiley, Sussex, England, 1990.

[Haa43] T. Haavelmo. "The statistical implications of a system of simultaneous equations." *Econometrica*, **11**:1–12, 1943. Reprinted in D.F. Hendry and M.S. Morgan (Eds.), *The Foundations of Econometric Analysis*, Cambridge University Press, 477–490, 1995.

[Hal98] J.Y. Halpern. "Axiomatizing Causal Reasoning." In G.F. Cooper and S. Moral, editors, *Uncertainty in Artificial Intelligence*, pp. 202–210. Morgan Kaufmann, San Francisco, CA, 1998. Also, *Journal of Artificial Intelligence Research* 12:3, 17–37, 2000.

[HCS03] L. Hayduk, G. Cummings, R. Stratkotter, M. Nimmo, K. Grygoryev, D. Dosman, M. Gillespie, and H. Pazderka-Robinson. "Pearls D-Separation: One More Step Into Causal Thinking." *Structural Equation Modeling*, **10**(2):289–311, 2003.

[Hec79] J.J. Heckman. "Sample Selection Bias as a Specification Error." *Econometrica*, **47**(1):pp. 153–161, 1979.

[Hec92] J.J. Heckman. "Randomization and Social Policy Evaluation." In C. Manski and I. Garfinkle, editors, *Evaluations: Welfare and Training Programs*, pp. 201–230. Harvard University Press, Cambridge, MA, 1992.

[Hec00] J.J. Heckman. "Causal parameters and policy analysis in economics: A twentieth century retrospective." *The Quarterly Journal of Economics*, **115**(1):45–97, 2000.

[Hec05] J.J. Heckman. "The Scientific Model of Causality." *Sociological Methodology*, **35**:1–97, 2005.

[Hei09] M. Hein. "Binary classification under sample selection bias." In J. Candela, M. Sugiyama, A. Schwaighofer, and N.D. Lawrence, editors, *Dataset Shift in Machine Learning*, pp. 41–64. MIT Press, Cambridge, MA, 2009.

[HF78] R. Horwitz and A. Feinstein. "Alternative analytic methods for case-control studies of estrogens and endometrial cancer." *New England Journal of Medicine*, **299**:368–387, 1978.

[HGH10] M. Höfler, A.T. Gloster, and J. Hoyer. "Causal effects in psychotherapy: Counterfactuals counteract overgeneralization." *Psychotherapy Research*, p. DOI: 10.1080/10503307.2010.501041, 2010.

[HHR04] M.A. Hernán, S. Hernández-Díaz, and J.M. Robins. "A structural approach to selection bias." *Epidemiology*, **15**(5):615–625, 2004.

[HIM05] V. Joseph Hotz, Guido Imbens, and Julie Holland Mortimer. "Predicting the efficacy of future training programs using past experiences at other locations." *Journal of Econometrics*, **125**(1-2):241–270, 2005.

[HO85] L. V. Hedges and I. Olkin. *Statistical Methods for Meta-Analysis*. Academic Press, 1985.

[Hur50] L. Hurwicz. "Generalization of the Concept of Identification." In T.C. Koopmans, editor, *Statistical Inference in Dynamic Economic Models*, Cowles Commission, Monograph 10, pp. 245–257. Wiley, New York, 1950.

[HV06a] Y. Huang and M. Valtorta. "Identifiability in Causal Bayesian Networks: A Sound and Complete Algorithm." In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI 2006)*, pp. 1149–1156. AAAI Press, Menlo Park, CA, 2006.

[HV06b] Y. Huang and M. Valtorta. "Pearl's Calculus of Intervention Is Complete." In R. Dechter and T.S. Richardson, editors, *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence (UAI 2006)*, pp. 217–224. AUAI Press, Corvallis, OR, 2006.

[Jew91] Nicholas P. Jewell. "Some Surprising Results about Covariate Adjustment in Logistic Regression Models." *International Statistical Review*, **59**(2):227–240, 1991.

[JG09] M.M. Joffe and T. Green. "Related Causal Frameworks for Surrogate Outcomes." *Biometrics*, **65**:530–538, 2009.

[KC06] Manabu Kuroki and Zhihong Cai. "On recovering a population covariance matrix in the presence of selection bias." *Biometrika*, **93**(3):601–611, 2006.

[KF09] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[KM99] M. Kuroki and M. Miyakawa. "Identifiability criteria for causal effects of joint interventions." *Journal of the Royal Statistical Society*, **29**:105–117, 1999.

[Koo53] T.C. Koopmans. "Identification problems in econometric model construction." In W.C. Hood and T.C. Koopmans, editors, *Studies in Econometric Method*, pp. 27–48. Wiley, New York, 1953.

[KSC84] H. Kiiveri, T.P. Speed, and J.B. Carlin. "Recursive causal models." *Journal of Australian Math Society*, **36**:30–52, 1984.

[Lau01] S.L. Lauritzen. "Causal inference from graphical models." In D.R. Cox and C. Kluppelberg, editors, *Complex Stochastic Systems*, pp. 63–107. Chapman and Hall/CRC Press, Boca Raton, FL, 2001.

[LN82] P.W. Lane and J.A. Nelder. "Analysis of covariance and standardization as instances of prediction." *Biometrics*, **38**:613–621, 1982.

[LR86] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc., New York, NY, USA, 1986.

[LR08] S. L. Lauritzen and T. S. Richardson. "Discussion of McCullagh: Sampling bias and logistic models." *J. Roy. Statist. Soc. Ser. B*, **70**:140–150, 2008.

[Man07] C. Manski. *Identification for Prediction and Decision*. Harvard University Press, Cambridge, Massachusetts, 2007.

[Mar50] J. Marschak. "Statistical inference in economics." In T. Koopmans, editor, *Statistical Inference in Dynamic Economic Models*, pp. 1–50. Wiley, New York, 1950. Cowles Commission for Research in Economics, Monograph 10.

[Mat07] R. L. Matzkin. "Nonparametric identification." In J.J. Heckman and E.E. Leamer, editors, *Handbook of Econometrics*, volume 6 of *Handbook of Econometrics*, chapter 73. Elsevier, 2007.

[Mul09] S.A. Mulaik. *Linear Causal Modeling with Structural Equations*. Chapman & Hall/Crc Statistics in the Social and Behavioral Sciences. CRC Press, 2009.

[MW12] J. Mefford and J. S. Witte. "The Covariate's Dilemma." *PLoS Genet*, **8**(11):e1003096, 11 2012.

[PB11a] J. Pearl and E. Bareinboim. "Transportability across studies: A formal approach." Technical Report R-372, Cognitive Systems Laboratory, Department of Computer Science, UCLA, 2011.

[PB11b] J. Pearl and E. Bareinboim. "Transportability of Causal and Statistical Relations: A Formal Approach." In W. Burgard and D. Roth, editors, *Proceedings of the Twenty-Fifth National Conference on Artificial Intelligence (AAAI 2011)*, pp. 247–254. AAAI Press, Menlo Park, CA, 2011.

[PDS12] M. Pirinen, P. Donnelly, and C. Spencer. "Including known covariates can reduce power to detect genetic effects in case-control studies." *Nature Genetics*, **44**:848–851, 2012.

[Pea88] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA, 1988.

[Pea93a] J. Pearl. "Aspects of graphical models connected with causality." In *Proceedings of the 49th Session of the International Statistical Institute*, pp. 391–401, Tome LV, Book 1, Florence, Italy, 1993.

[Pea93b] J. Pearl. "Comment: Graphical Models, Causality, and Intervention." *Statistical Science*, **8**(3):266–269, 1993.

[Pea95] J. Pearl. "Causal diagrams for empirical research." *Biometrika*, **82**(4):669–710, 1995.

[Pea00] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2000. Second ed., 2009.

[Pea01] J. Pearl. "Direct and indirect effects." In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI 2001)*, pp. 411–420. Morgan Kaufmann, San Francisco, CA, 2001.

[Pea09a] J. Pearl. "Causal Inference in Statistics: An Overview." *Statistics Surveys*, **3**:96–146, 2009.

[Pea09b] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, second edition, 2009.

[Pea11] J. Pearl. "Principal Stratification - a Goal or a Tool?" *The International Journal of Biostatistics*, **7**(1):1–14, 2011.

[Pea12a] J. Pearl. "The Causal Foundations of Structural Equation Modeling." In R.H̃. Hoyle, editor, *Handbook of Structural Equation Modeling*. Guilford Press, New York, 2012.

[Pea12b] J. Pearl. "The Mediation Formula: A guide to the assessment of causal pathways in nonlinear models." In C. Berzuini, P. Dawid, and L. Bernardinell, editors, *Causality: Statistical Perspectives and Applications*, p. Chapter 12. Wiley, New York, 2012.

[Pea13] J. Pearl. "Lindear models: A useful "Microscope" for causal analysis." *Journal of Causal Inference*, **1**:155–170, 2013.

[PP10] J. Pearl and A. Paz. "Confounding Equivalence in Causal Equivalence." In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pp. 433–441. AUAI, Corvallis, OR, 2010.

[PR95] J. Pearl and J.M. Robins. "Probabilistic evaluation of sequential plans from causal models with hidden variables." In P. Besnard and S. Hanks, editors, *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI 1995)*, pp. 444–453. Morgan Kaufmann, San Francisco, 1995.

[Pre89] R.L. Prentice. "Surrogate endpoints in clinical trials: definition and operational criteria." *Statistics in Medicine*, **8**:431–440, 1989.

[PV91] J. Pearl and T. Verma. "A Theory of Inferred Causation." In J.A. Allen, R. Fikes, and E. Sandewall, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, pp. 441–452. Morgan Kaufmann, San Mateo, CA, 1991.

[Rob86] J.M. Robins. "A new approach to causal inference in mortality studies with a sustained exposure period – applications to control of the healthy workers survivor effect." *Mathematical Modeling*, **7**:1393–1512, 1986.

[Rob01] J.M. Robins. "Data, design, and background knowledge in etiologic inference." *Epidemiology*, **12**(3):313–320, 2001.

[Rub74] D.B. Rubin. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology*, **66**:688–701, 1974.

[SCC02] W.R. Shadish, T.D. Cook, and D.T. Campbell. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton-Mifflin, Boston, second edition, 2002.

[SE07] A. T. Smith and C. Elkan. "Making generative classifiers robust to selection bias." In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '07, pp. 657–666, New York, NY, USA, 2007. ACM.

[SGS93] P. Spirtes, C.N. Glymour, and R. Scheines. *Causation, Prediction, and Search*. Springer-Verlag, New York, 1993.

[SGS00] P. Spirtes, C.N. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2nd edition, 2000.

[Sim53] H.A. Simon. "Causal ordering and identifiability." In Wm. C. Hood and T.C. Koopmans, editors, *Studies in Econometric Method*, pp. 49–74. Wiley and Sons, Inc., New York, NY, 1953.

[SJP12] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. "On Causal and Anticausal Learning." In J Langford and J Pineau, editors, *Proceedings of the 29th International Conference on Machine Learning (ICML 2012)*, pp. 1255–1262, New York, NY, USA, 2012. Omnipress.

[SP06a] I. Shpitser and J Pearl. "Identification of Conditional Interventional Distributions." In R. Dechter and T.S. Richardson, editors, *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence (UAI 2006)*, pp. 437–444. AUAI Press, Corvallis, OR, 2006.

[SP06b] I. Shpitser and J. Pearl. "Identification of Joint Interventional Distributions in Recursive Semi-Markovian Causal Models." In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI 2006)*, pp. 1219–1226. AAAI Press, Menlo Park, CA, 2006.

[SP06c] I. Shpitser and J. Pearl. "Identification of Joint Interventional Distributions in Recursive Semi-Markovian Causal Models." In *Proceedings*

of the Twenty-First National Conference on Artificial Intelligence, pp. 1219–1226. AAAI Press, Menlo Park, CA, 2006.

[Sto09] A.J. Storkey. "When training and test sets are different: characterising learning transfer." In J. Candela, M. Sugiyama, A. Schwaighofer, and N.D. Lawrence, editors, *Dataset Shift in Machine Learning*, pp. 3–28. MIT Press, Cambridge, MA, 2009.

[SW60] R.H. Strotz and H.O.A. Wold. "Recursive versus nonrecursive systems: An attempt at synthesis." *Econometrica*, **28**:417–427, 1960.

[Tia02] J. Tian. *Studies in Causal Reasoning and Learning*. PhD thesis, Computer Science Department, University of California, Los Angeles, CA, November 2002.

[TL11] J. Textor and M. Liskiewicz. "Adjustment Criteria in Causal Diagrams: An Algorithmic Perspective." In Avi Pfeffer and Fabio Cozman, editors, *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, pp. 681–688. AUAI Press, 2011.

[TP02] J. Tian and J. Pearl. "A general identification condition for causal effects." In *Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI 2002)*, pp. 567–573. AAAI Press/The MIT Press, Menlo Park, CA, 2002.

[TPP98] J. Tian, A. Paz, and J. Pearl. "Finding Minimal Separating Sets." Technical Report R-254, Cognitive Systems Laboratory, Department of Computer Science, UCLA, CA, 1998.

[VP88] T. Verma and J. Pearl. "Causal Networks: Semantics and Expressiveness." In *Proceedings of the Fourth Workshop on Uncertainty in Artificial Intelligence*, pp. 352–359, Mountain View, CA, 1988. Also in R. Shachter, T.S. Levitt, and L.N. Kanal (Eds.), *Uncertainty in AI 4*, Elesevier Science Publishers, 69–76, 1990.

[VP90] T. Verma and J. Pearl. "Equivalence and Synthesis of Causal Models." In *Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence (UAI 1990)*, pp. 220–227, Cambridge, MA, July 1990. Also in P. Bonissone, M. Henrion, L.N. Kanal and J.F. Lemmer (Eds.), *Uncertainty in Artificial Intelligence 6*, Elsevier Science Publishers, B.V., 255–268, 1991.

[Wes16] H. Westergaard. "Scope and Method of Statistics." *Publications of the American Statistical Association*, **15**(115):229–276, 1916.

[Whi78] A.S. Whittemore. "Collapsibility of Multidimensional Contingency Tables." *Journal of the Royal Statistical Society, B*, **40**(3):328–340, 1978.

[Yul34] G.U. Yule. "On Some Points Relating to Vital Statistics, More Especially Statistics of Occupational Mortality." *Journal of the Royal Statistical Society*, **97**(1):1–84, 1934.

[Zad04] B. Zadrozny. "Learning and Evaluating Classifiers Under Sample Selection Bias." In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, pp. 114–, New York, NY, USA, 2004. ACM.

[Zha08] J. Zhang. "On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias." *Artif. Intell.*, **172**:1873–1896, November 2008.

[ZSM13] K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang. "Domain adaptation under Target and Conditional Shift." In *Proceedings of the 30th International Conference on Machine Learning (ICML)*. JMLR: W&CP volume 28, 2013.

Generalizability in Causal Inference: Theory and Algorithms
Errata

Elias Bareinboim

March 2019

The following is a revision of some results presented in [1]. Section, definition, theorem and lemma numbers match the manuscript in mention. Changes are highlighted in red.

4 Transportability from Multiple Studies with Limited Experiments

4.1 Characterizing mz-Transportable Relations

The following is a revised definition of mz^*-shedge, a graphical structure that witnesses the non-transportability of a causal distribution. The removal of condition 3 of the original definition is not strictly needed but since it's entailed by conditions 1 and 2, we prefer to phrase in this way for the sake of clarity.

Definition 18 (mz^*-shedge)**.** *Let $\mathcal{D} = (D^{(1)}, \ldots, D^{(n)})$ be a collection of selection diagrams relative to source domains $\Pi = (\pi_1, \ldots, \pi_n)$ and target domain π^*, respectively, $\mathbf{S_i}$ represents the collection of S-variables in the selection diagram $D^{(i)}$, and let $D^{(*)}$ be the causal diagram of π^*. Let $\{\langle P^i, I^i_z\rangle\}$ be the collection of pairs of observational and interventional distributions of $\{\pi_i\}$, where $I^i_z = \bigcup_{\mathbf{Z}'\subseteq\mathbf{Z}_i} P^i(\mathbf{v}|do(\mathbf{z}'))$, and in an analogous manner, $\langle P^*, I^*_z\rangle$ be the observational and interventional distributions of π^*, for $\mathbf{Z_i}$ the set of experimental variables in π_i. Consider a pair of $\mathbf{R}$-rooted C-~~forests~~components $\mathcal{F} = \langle F, F'\rangle$ such that $F' \subset F$, $F' \cap \mathbf{X} = \emptyset$, $F \cap \mathbf{X} \neq \emptyset$, and $\mathbf{R} \subseteq An(\mathbf{Y})_{G_{\overline{\mathbf{X}}}}$ ~~(called hedge)~~. We say that ~~the induced~~ a collection of pairs of $\mathbf{R}$-rooted C-forests over each diagram, $\langle \mathcal{F}^{(*)}, \mathcal{F}^{(1)}, \ldots, \mathcal{F}^{(n)}\rangle$, with $\mathcal{F}^{(i)} = \langle F^{(i)}, F'^{(i)}\rangle$, $F^{(i)} \subseteq F$, $i = \{*, 1, \ldots, n\}$, $\bigcup_i F'^{(i)} = F'$, is an mz^*-shedge for $P^*_{\mathbf{x}}(\mathbf{y})$ relative to experiments $(I^*_z, I^1_z, \ldots, I^n_z)$ if they are all hedges for $P_{\mathbf{x}}(\mathbf{y})$, and one of the following conditions hold for each domain π_i, $i = \{*, 1, \ldots, n\}$:*

1. *There exists at least one variable of $\mathbf{S_i}$ pointing to the induced diagram $F'^{(i)}$, or*

2. $(F^{(i)} \setminus F'^{(i)}) \cap \mathbf{Z_i}$ *is an empty set.*

3. ~~*The collection of pairs of C-forests induced over diagrams,* $\langle \mathcal{F}^{(*)}, \mathcal{F}^{(1)}, \ldots, F^{(i)} \setminus \mathbf{Z}^*_{\mathbf{i}}, \ldots, \mathcal{F}^{(n)} \rangle$*, is also an mz-shedge relative to* $(I^*_z, I^1_z, \ldots, I^i_{z \setminus z^*_i}, \ldots, I^n_z)$*, where* $\mathbf{Z}^*_{\mathbf{i}} = (F^{(i)} \setminus F'^{(i)}) \cap \mathbf{Z_i}$*.*~~

~~*We call mz*-shedge the mz-shedge in which there exist one directed path from* $\mathbf{R} \setminus (\mathbf{R} \cap De(\mathbf{X})_F)$ *to* $(\mathbf{R} \cap De(\mathbf{X})_F)$ *not passing through* $\mathbf{X}$*.*~~

With a revised definition, we provide a new proof for Theorem 17 and related Lemmas 17, 18 and 19.

Theorem 17. *Let* $\mathcal{D} = \{D^{(1)}, \ldots, D^{(n)}\}$ *be a collection of selection diagrams relative to source domains* $\Pi = \{\pi_1, \ldots, \pi_n\}$*, and target domain* π^**, respectively, and* $\{I^i_z\}$*, for* $i = \{*, 1, \ldots, n\}$ *defined appropriately. If there is an* mz^**-shedge for the effect* $R = P^*_{\mathbf{x}}(\mathbf{y})$ *relative to experiments* $(I^*_z, I^1_z, \ldots, I^n_z)$ *in* $\mathcal{D}$*,* R *is not* mz*-transportable from* Π *to* π^* *in* $\mathcal{D}$ *(relative to all experiments* I^i_z*).*

Proof sketch. Let F be the $\mathbf{R}$-rooted C-component (basis). Without loss of generality, we will consider a structure with a maximal root-set. That is, one that when subjected to the following procedure remains unchanged:

1. let $\mathbf{B} = An(\mathbf{Y})_{G_{\overline{\mathbf{X}}}} \cap (F \setminus \mathbf{X})$,
2. consider the subgraph $F \setminus \mathbf{X}$ and let $\mathbf{R}'$ be the set of variables in $\mathbf{B}$ that are also in the same C-component as any element of $\mathbf{R}$ in that subgraph.
3. Then, remove from $\mathcal{F}$ the edges outgoing from $\mathbf{R}'$ and let $\mathbf{R} = \mathbf{R} \cup \mathbf{R}'$.

After the previous steps, we obtain a new mz^*-shedge with a maximal root-set, where the variables in F' are exactly those in the root-set $\mathbf{R}$. To witness, assume for the sake of contradiction there exists a variable V in F' not in $\mathbf{R}$, by definition F' is an $\mathbf{R}$-rooted C-component containing no variables in $\mathbf{X}$, and since V belongs to F' it must fall into $\mathbf{B}$ in step one and also satisfied step two. Hence, V can be added to the root-set as in step three, contradicting the fact that F had maximal root-set.

Let $\mathbf{T} = \mathbf{F} \setminus \mathbf{R}$ be the observable variables in F that are not in $\mathbf{R}$. Let $\mathbf{U}'$ be the set of unobservable variables in F and partition it into the sets:

- $\mathbf{U_T} = \{U \in \mathbf{U}' \mid T_1 \leftarrow U \rightarrow T_2 \text{ and } T_1, T_2 \in \mathbf{T}\}$,
- $\mathbf{U_R} = \{U \in \mathbf{U}' \mid R_1 \leftarrow U \rightarrow R_2 \text{ and } R_1, R_2 \in \mathbf{R}\}$, and
- $\mathbf{U}_\times = \{U \in \mathbf{U}' \mid T \leftarrow U \rightarrow R \text{ and } T \in \mathbf{T}, R \in \mathbf{R}\}$.

Let $\mathbf{U}^i_{\mathbf{T}} = \mathbf{U_T} \cap F^{(i)}, \mathbf{U}^i_{\mathbf{R}} = \mathbf{U_R} \cap F^{(i)}$ and $\mathbf{U}^i_\times = \mathbf{U}_\times \cap F^{(i)}$.

We construct two causal models M_1 and M_2 that will agree on the collection of distributions $\{\langle P^i, I^i_z \rangle\}$, $\langle P^*, I^*_z \rangle$, but disagree on the interventional distribution $P^*_{\mathbf{x}}(\mathbf{y})$.

Let k_t be the number of $\mathcal{F}^{(i)}$s in which a variable $T \in \mathbf{T}$ appears. Then, we will parametrize T as a k_t-bit variable with $T_{[i]}$ representing the bit in T

corresponding to $\mathcal{F}^{(i)}$. Similarly, define k_u for $U \in \mathbf{U_T} \cup \mathbf{U}_\times$, then U is a k_u-bit variable where $U_{[i]}$ stands for the bit associated with $\mathcal{F}^{(i)}$.

Call $\mathbf{W}$ the set of variables pointed by S-nodes in F' and consider the following encoding for the domains: let S_i be the index variable corresponding to the source domain $\pi_i \in \Pi$, and let the tuple $\langle S_1 = 0, \ldots, S_i = 1, \ldots, S_n = 0\rangle$ represent the index for the functional model relative to this domain. Let the tuple $\langle S_1 = 0, S_2 = 0, \ldots, S_n = 0\rangle$ represent the index for functional model relative to the target domain π^*.

Let $\mathbf{Pa}_v$ stand for the set of observable and unobservable parents of variable V in F and $\mathbf{Pa}_v^i$ for the set of parents of the same variable in $F^{(i)}$. For a set of variables $\mathbf{V}$, let $\mathbf{Pa_v} = \bigcup_{V \in \mathbf{V}} \mathbf{Pa}_v$ and $\mathbf{Pa_v^i} = \bigcup_{V \in \mathbf{V}} \mathbf{Pa}_v^i$.

In both models, let each bit $T_{[i]}$ of $T \in \mathbf{T}$ be governed by the function

$$f_{t_{[i]}} = \bigoplus_{A \in \mathbf{Pa}_t^i} A_{[i]}. \tag{1}$$

Variables in $\mathbf{R} \cup \mathbf{U_R}$ are binary. Pick an arbitrary variable $R^* \in \mathbf{R}$. For any $R \in \mathbf{R} \setminus \mathbf{W}$ in model 1 and 2 except for R^* in model 2, let

$$f_r = \left(\bigwedge_{\pi_i \in \Pi, T \in \mathbf{Pa}_r^i \cap \mathbf{T}} g_i(T) \quad \wedge \bigwedge_{\pi_i \in \Pi, U \in \mathbf{Pa}_r^i \cap \mathbf{U}_\times} U_{[i]} \right) \wedge \left(\bigoplus_{U \in \mathbf{Pa}_r \cap \mathbf{U_R}} U \right); \tag{2}$$

where $g_i(.)$ is defined as follows:

$$g_i(T) = \begin{cases} \overline{T_{[i]}} & \text{if } |\mathbf{Pa_r^i} \cap \mathbf{T}| \text{ is odd and } |\mathbf{U}_\times^i| \text{ is odd, or} \\ & \text{if } |\mathbf{Pa_r^i} \cap \mathbf{T}| \text{ is even and } |\mathbf{U}_\times^i| \text{ is even and } T = T^{(i)}. \\ T_{[i]} & \text{otherwise.} \end{cases} \tag{3}$$

Where $T^{(i)}$ is any variable chosen from the set $\mathbf{Pa_r^i} \cap \mathbf{T}$ for each domain π_i.

For R^* in model 2:

$$f_{r^*} = \left(\bigwedge_{T \in \mathbf{Pa}_{r^*}^i \cap \mathbf{T}} g_i(T) \quad \wedge \bigwedge_{U \in \mathbf{Pa}_{r^*}^i \cap \mathbf{U}_\times} U_{[i]} \right) \wedge \overline{\left(\bigoplus_{U \in \mathbf{Pa}_{r^*} \cap \mathbf{U_R}} U \right)}. \tag{4}$$

For $R \in (\mathbf{R} \cap \mathbf{W})$ let

$$R \leftarrow f_r \wedge \bigwedge_{S_i | (S_i \to R) \in \mathcal{D}} \overline{S_i}, \tag{5}$$

where f_r is constructed as in the previous case and S_i is an S-node pointing to R, relative to domain π_i.

Every bit of the $\mathbf{U}$-variables is set to behave as a fair coin.

Lemma 17. *The two models M_1 and M_2 are compatible with the selection diagrams $\mathcal{D}$.*

Proof. The result is immediate. Consider the functional model that generates any domain π_i, in both models M_1 and M_2. By construction, the index tuple is set to $\langle S_1 = 0, \ldots, S_i = 1, \ldots, S_n = 0\rangle$ in π_i, and $\langle S_1 = 0, \ldots, S_i = 0, \ldots, S_n = 0\rangle$ in π^*. So, it is obvious that in both models, the only structural differences between π_i and π^* are the equations of $W \in \mathbf{W}$ in which S_i appears. □

Lemma 18. *The two models agree in the distribution of* $P^i(\mathbf{t}, \mathbf{r}), i = \{*, 1, \ldots, n\}$ *and there exists an assignment for* $\mathbf{X}$ *and* $\mathbf{Y}$ *such that* $P^*_{M_1}(\mathbf{Y}|do(\mathbf{X})) \neq P^*_{M_2}(\mathbf{Y}|do(\mathbf{X}))$.

Proof. (*Matching observational distributions*)
First consider any particular domain π_i and a particular assignment $\mathbf{u}$ of the variables in $\mathbf{U}$. We have that in both models the value of $\mathbf{T}$ has to be the same since the functions are the same in those models (with fixed π_i all S_i have the same value).

Let $\mathbf{R}_0$ be the set of nodes in $\mathbf{R}$ for which the expression in the first parenthesis of equation (2) evaluates to 0 in both models. Note that the set $\mathbf{R}_0$ is determined by the variables in $\mathbf{U_T} \cup \mathbf{U}_\times$, because those determine the values of $\mathbf{T}$, and the variables in $\mathbf{U_R}$ only appear in the second part of equation (2) which is not taken into account in the definition of $\mathbf{R}_0$. We will show that $\mathbf{R}_0$ is not empty in the context of the non-intervened models corresponding to π_i. Consider any $U \in \mathbf{U}^i_\times$ such that $U_{[i]} = 0$, then any R that is pointed by U will have value 0 in both models due to the construction of f_r, and we are done. We continue with the situation where all such U have $U_{[i]} = 1$. Consider the quantity C_i defined as

$$C_i = \bigoplus_{T \in \mathbf{Pa}^i_\mathbf{r} \cap \mathbf{T}} T_{[i]}, \tag{6}$$

and note that due to the forestness of $F^{(i)}$ and the parametrization; C_i computes the xor of all the unobservable variables in $\mathbf{U_T}$ and $\mathbf{U}_\times$, having those in $\mathbf{U_T}$ accounted twice. Together with the fact that for any $U \in \mathbf{U}^i_\times$, $U_{[i]} = 1$, it follows that

$$C_i = \bigoplus_{U \in \mathbf{U}^i_\times} U_{[i]} = |\mathbf{U}^i_\times| \ mod\ 2. \tag{7}$$

Note that the set of parents of variables in $\mathbf{R}$ in $\mathbf{T}$ (i.e. $\mathbf{Pa}^i_\mathbf{r} \cap \mathbf{T}$) must be non-empty for any given hedge, then consider each one of the following four scenarios:

1. $|\mathbf{Pa}^i_\mathbf{r} \cap \mathbf{T}|$ is odd and $|\mathbf{U}^i_\times|$ is odd: We have $C_i = 1$ which implies that at least one of $T_{[i]}$ has to be 1. Since g_i negates all $T_{[i]}$ in this case, we have that at least one R (with T as a parent) will have 0 as value.

2. $|\mathbf{Pa}^i_\mathbf{r} \cap \mathbf{T}|$ is odd and $|\mathbf{U}^i_\times|$ is even: We have $C_i = 0$ which implies that at least one of $T_{[i]}$ has to be 0. Since g_i leaves each $T_{[i]}$ the same, we have that at least one R will have 0 as value.

3. $|\mathbf{Pa}^i_{\mathbf{r}} \cap \mathbf{T}|$ is even and $|\mathbf{U}^i_{\times}|$ is odd: We have $C_i = 1$ which implies that at least one of $T_{[i]}$ has to be 0. As in the previous case g_i leaves $T_{[i]}$ the same so at least one $R = 0$ in both models.

4. $|\mathbf{Pa}^i_{\mathbf{r}} \cap \mathbf{T}|$ is even and $|\mathbf{U}^i_{\times}|$ is even: We have $C_i = 0$, so for all combinations but for all $T_{[i]} = 1$, there are always at least two $T_{[i]} = 0$. Since g_i negates only one $T_{[i]}$, it follows that there is always at least one $g_i(T) = 0$ and any R pointed by T will have value 0.

Due to the previous analysis we have that $\mathbf{R}_0$ is non-empty. Pick $\hat{R} \in \mathbf{R}_0$ that is closest to R^* in terms of the length of the bidirected path $\overline{p}$ made of edges in $\mathbf{U_R}$ between them (the length of the path is 0 if $R^* \in \mathbf{R}_0$. Make $\mathbf{u}^1 = \mathbf{u}$ (the considered assignment) and $\mathbf{u}^2$ equal to $\mathbf{u}$ for all $\mathbf{U}$ except those in $\overline{p}$ for which their negation is taken. By definition $\overline{p}$ intersects with $\mathbf{R}_0$ only at the endpoints. Also, for every intermediate node R of $\overline{p}$, there are two parents in $\mathbf{U_R}$ being negated; from the parametrization of f_r we can tell that the value of R remains the same because this change does not affect the parity being computed by the xor. We have then that $\mathbf{u}^1$ corresponds to $\mathbf{u}^2$ and repeating the reasoning for any other assignment of the $\mathbf{u}$ we get a bijective relationship between assignments producing the same observation in both models, hence the distributions over the observed variables is the same.

(*Different interventional distribution*)
For the second part of the claim, consider the distribution $P(\mathbf{r}|do(\mathbf{X} = \hat{\mathbf{x}}))$, where $\hat{\mathbf{x}}$ is an assignment where each bit of $X \in \mathbf{X}$ is given by

$$\hat{x}_{[i]} = \begin{cases} 0 & \text{if } X \in F^{(i)} \text{ and } X \notin \mathbf{Pa}^i_{\mathbf{r}}, \\ g_i(1) & \text{if } X \in \mathbf{Pa}^i_{\mathbf{r}}, \end{cases} \tag{8}$$

Start by noting when this intervention on $\mathbf{X}$ is performed, every $F^{(i)}$ is affected (because by definition every $\mathcal{F}^{(i)}$ intersects $\mathbf{X}$). We want to show that under this circumstance, there exists at least one assignment $\mathbf{u}$ such that $\mathbf{R}_0$ is empty. Start with an assignment where $\mathbf{u}_{\times} = 1$, if every $g_i(T) = 1$ for $i = \{*, 1, \ldots, n\}$, $T \in \mathbf{Pa}^i_{\mathbf{r}}$ we are done. Otherwise, for every i, T such that $g_i(T) = 0$ find a path $\overline{p}$, in $F^{(i)}$, between T and a variable in $An(\mathbf{X})_{F^{(i)}}$ (that includes $\mathbf{X}$) made of bidirected edges corresponding to variables in $\mathbf{U_T}$. Such path must exists due to the fact that the mz^*-shedge under consideration has a maximal root-set and T is in $An(\mathbf{Y})_{G_{\overline{\mathbf{X}}}}$ (because it is a parent of some $R \in \mathbf{R}$).

We can flip the bit associated with π_i for all $\mathbf{u}$s in $\overline{p}$, which preserve the parity (hence the bit value) of every intermediate observable, while the value of the observable in the endpoint is either fixed by intervention (if the path ends in some $X \in \mathbf{X}$) or can change without affecting any variable in $\mathbf{Pa}^i_{\mathbf{r}} \setminus \{T\}$ (and hence R as well) because it is an ancestor of $\mathbf{X}$ that has been intervened and $F^{(i)}$ is a forest. Changing the unobservables in the path also changes the parity of T and since $g_i(\overline{T}) = \overline{g_i(T)}$ we have that now $g_i(T) = 1$.

This process only affects bits associated with π_i, by repeating it for every other T, i such that $g_i(T) = 0$, we get an assignment where $\mathbf{R}_0$ is empty. Under

these circumstances, the value of variables in $\mathbf{R}$ is determined by the xor in the second parenthesis of equation (2) that depends only on variables in $\mathbf{U_R}$, that free so far. Then, in M_1 we have

$$\bigoplus_{R\in\mathbf{R}} R = \bigoplus_{R\in\mathbf{R}}\bigoplus_{U\in\mathbf{Pa}_r\cap\mathbf{U_R}} U = \bigoplus_{U\in\mathbf{U_R}} (U\oplus U) = 0, \tag{9}$$

since every $U \in \mathbf{U_R}$ appears exactly twice. As for M_2

$$\bigoplus_{R\in\mathbf{R}} R = \left(\bigoplus_{R\in\mathbf{R},R\neq R^*}\bigoplus_{U\in\mathbf{Pa}_r\cap\mathbf{U_R}} U\right) \oplus \overline{\left(\bigoplus_{U\in\mathbf{Pa}_{r^*}\cap\mathbf{U_R}} U\right)} \tag{10}$$

$$= \left(\bigoplus_{R\in\mathbf{R},R\neq R^*}\bigoplus_{U\in\mathbf{Pa}_r\cap\mathbf{U_R}} U\right) \oplus \left(1\oplus \bigoplus_{U\in\mathbf{Pa}_{r^*}\cap\mathbf{U_R}} U\right) \tag{11}$$

$$= 1\oplus \bigoplus_{R\in\mathbf{R}}\bigoplus_{U\in\mathbf{Pa}_r\cap\mathbf{U_R}} U \tag{12}$$

$$= 1\oplus \left(\bigoplus_{U\in\mathbf{U_R}} (U\oplus U)\right) \tag{13}$$

$$= 1. \tag{14}$$

Then, from the first part of this proof we have that for any $\mathbf{u}$ for which $\mathbf{R}_0$ the distributions both models produce the same observations, however, for the intervention $do(\mathbf{X}=\hat{\mathbf{x}})$ there are $\mathbf{u}$ for which $\mathbf{R}_0$ is empty and we have that model 2 produces more observations where $\bigoplus\mathbf{R}=1$ hence the different observations which implies $P^*_{M_1}(\bigoplus\mathbf{r}=1|do(\hat{\mathbf{x}})) \neq P^*_{M_2}(\bigoplus\mathbf{r}=1|do(\hat{\mathbf{x}}))$.

(Mapping $\mathbf{R}$ to $\mathbf{Y}$)
By definition, there is a directed path in G from every $R\in\mathbf{R}$ to $\mathbf{Y}$ (could be zero-length) not intersecting $\mathbf{X}$. Augment M_1 and M_2 such that for any non-zero-length path $\overline{q}$ from R to $Y\in\mathbf{Y}$, each variable except for R let the function be an xor of its parents. If the path contains an intermediate variable $R'\in\mathbf{R}$ in $\overline{q}$ add an extra bit to it, such that the original bit computes the original function and the new one the xor of its parents.

In this new models $\bigoplus\mathbf{Y}=\bigoplus\mathbf{R}$, then the second part of the lemma follows.

Lemma 19. *The two models agree in the collection of interventional distributions ($\{I^i_z\}$) in the respective source domains π_i, $i=1,\ldots,n$, and target domain π^*.*

Proof. Consider a domain π_i and a set $\mathbf{Z}\subseteq\mathbf{Z_i}$. From the definition of mz^*-shedge we have that either condition 1 or 2 are true for $F^{(i)}$. In the former case we have some indicator in $\mathbf{S_i}$ pointing to a variable $R\in\mathbf{R}$ that will be set to 0 in both models in domain π_i. In the latter case, and by the same argument

used in the proof for lemma 18, we have that any variable $R \in \mathbf{R}$ that is a child of a variable $T \in \mathbf{T} \cap F^{(i)}$ belongs to $\mathbf{R}_0$.

In any case $\mathbf{R}_0$ is not empty and as in the proof for lemma 18 this implies that the observed distributions match in both models for the variables in F. The mapping described in lemma 18 modifies both models in the same way and may only change the functions of $\mathbf{R}$ by adding extra bits without changing the fact that the observations match between distributions. □

6 Causal Inference by Surrogate Experiments

6.1 Characterizing *zID* Relations

Theorem 31. *Let* $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$ *be disjoint sets of variables and let* G *be the causal diagram. The causal effect* $Q = P(\mathbf{y}|do(\mathbf{x}))$ *is zID in* G *if and only if one of the following conditions hold:*

a. Q *is identifiable in* G*; or,*

b. There is no hedge $\mathcal{F} = \langle F, F' \rangle$ *for* Q *in* G *such that* $(F \setminus F') \cap \mathbf{Z}$ *is empty.*

~~*(i)* $\mathbf{X}$ *intercepts all directed paths from* $\mathbf{Z}'$ *to* $\mathbf{Y}$*, and*~~

~~*(ii)* Q *is identifiable in* $G_{\overline{\mathbf{Z}'}}$.~~[6]

Proof. (only if) Suppose there exists a hedge as described in condition b, Suppose Q is not identifiable in G (condition a) and there is a hedge $\mathcal{F}$ as described in condition b. Note that $\mathcal{F}$ satisfies the definition for mz^*-shedge, hence by Theorem 17 it follows that Q is not identifiable from $P(\mathbf{v})$, $\{P_{\mathbf{z}'}(\mathbf{v}|do(\mathbf{z}'))\}_{\mathbf{Z}' \subseteq \mathbf{Z}}$, which equates to Q not being *zID*.

(if) Suppose Q is not *zID*, then it easy to see that Q is not identifiable from $P(\mathbf{v})$ (which is considered by *zID*) therefore condition a is not satisfied. Let $\mathcal{F} = \langle F, F' \rangle$ be the hedge in G witnessing that some factor Q' associated with Q is not identifiable from $P(\mathbf{v})$. Let $\mathbf{Z}' = (F \setminus F') \cap \mathbf{Z}$, if $\mathbf{Z}' = \emptyset$, condition b does not hold and we are done. Otherwise, we can consider the distribution $P(\mathbf{v}|do(\mathbf{z}'))$ associated with $G_{\overline{\mathbf{z}'}}$ where $\mathcal{F}$ cannot be a hedge (every variable in $\mathbf{Z}'$ belongs to a different C-component in that graph). Then, Q' is identifiable from $P(\mathbf{v}|do(\mathbf{z}'))$ and there has to be another Q'' that is not identifiable from $P(\mathbf{v})$ else Q is *zID*. Let $\mathcal{F}'$ be the hedge associated with Q'' and by repeating the reasoning above, we have that either we end up with a hedge as forbidden by condition b or a contradiction. Therefore, Q being not *zID* implies that both conditions a and b are false; which entails the forward direction of this theorem. □

The corollary below followed from the original condition in Theorem 31.

[3] This condition can be rephrased graphically as "There exists no hedge for Q as an edge subgraph in $G_{\overline{\mathbf{Z}'}}$."

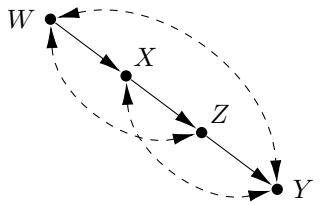


Figure 1: Graph in which $P(y|do(x))$ is not $I\!D$ from $P(\mathbf{v})$ and G, but it is $zI\!D$ with experiments on Z, that is in $De(X)_{G_{An(Y)}}$.

Corollary 22. *Let G be the causal diagram, $\mathbf{X}$, $\mathbf{Y} \subset \mathbf{V}$ be disjoint sets of variables, and $\mathbf{Z} \subseteq De(\mathbf{X})_{G_{An(\mathbf{Y})}}$. The causal effect $Q = P(\mathbf{y}|do(\mathbf{x}))$ is not $zI\!D$ from P and $do(\mathbf{Z})$ in G, if Q is not $I\!D$ from P in G.*

This corollary is not valid. To understand the subtlety with this statement, consider the graph in Fig. 1 where the query to be z-identified is $Q = P(y|do(x))$ and the available distributions are $P(\mathbf{v})$ and $P(y, x, w|do(z))$. According with the corollary, if Q is not identifiable from G and $P(\mathbf{v})$, it would not be identifiable even with experiments on Z because $\{Z\} \subseteq De(X)_{G_{An(Y)}}$. However, the following derivation follows:

$$P(y|do(x)) = \sum_z P(y|do(x), z)P(z|do(x)) \tag{15}$$

$$= \sum_z P(y|do(x), do(z))P(z|do(x)) \tag{16}$$

$$= \sum_z P(y|do(z))P(z|do(x)) \tag{17}$$

$$= \sum_z P(y|do(z)) \sum_w P(z|do(x), w)P(w|do(x)) \tag{18}$$

$$= \sum_z P(y|do(z)) \sum_w P(z|x, w)P(w|do(x)) \tag{19}$$

$$= \sum_z P(y|do(z)) \sum_w P(z|x, w)P(w), \tag{20}$$

that certifies that Q is $zI\!D$. The key point missed in Corollary 22 is that if Q is decomposable into more than one factor, some of them could be identified from the observational distribution and others from experimental distributions, fact that cannot be captured in a non-recursive condition.

6.2 A Complete Algorithm for $zI\!D$

Below the algorithm $\mathbf{ID^z}$ is reestated to make some recursive calls more explicit.

function $\mathbf{ID^z}(\mathbf{y}, \mathbf{x}, \mathbf{Z}, \mathcal{I}, \mathcal{J}, P, G)$
INPUT: $\mathbf{x}, \mathbf{y}$: value assignments; $\mathbf{Z}$: variables with interventions available; $\mathcal{I}, \mathcal{J}$: see caption; P: current probability distribution $do(\mathcal{I}, \mathcal{J}, x)$ (observational when $\mathcal{I} = \mathcal{J} = \emptyset$); G: causal graph.
OUTPUT: Expression for $P_{\mathbf{x}}(\mathbf{y})$ in terms of $P, P_{\mathbf{z}}$ or $\mathbf{FAIL}(F, F')$.

1 **if** $\mathbf{x} = \emptyset$, **return** $\sum_{\mathbf{v} \setminus \mathbf{y}} P(\mathbf{v})$.
2 **if** $\mathbf{V} \setminus An(\mathbf{Y})_G \neq \emptyset$,
return $\mathbf{ID^z}(\mathbf{y}, \mathbf{x} \cap An(\mathbf{Y})_G, \mathbf{Z},$
$\mathcal{I}, \mathcal{J}, \sum_{\mathbf{v} \setminus An(\mathbf{Y})_G} P, An(\mathbf{Y})_G)$.
3 Set $\mathbf{Z_w} = ((\mathbf{V} \setminus (\mathbf{X} \cup \mathcal{I} \cup \mathcal{J})) \setminus An(\mathbf{Y})_{G_{\overline{\mathbf{X} \cup \mathcal{I} \cup \mathcal{J}}}}) \cap \mathbf{Z}$.
Set $\mathbf{W} = ((\mathbf{V} \setminus (\mathbf{X} \cup \mathcal{I} \cup \mathcal{J})) \setminus An(\mathbf{Y})_{G_{\overline{\mathbf{X} \cup \mathcal{I} \cup \mathcal{J}}}}) \setminus \mathbf{Z}$.
if $(\mathbf{Z_w} \cup \mathbf{W}) \neq \emptyset$,
return $\mathbf{ID^z}(\mathbf{y}, \mathbf{x} \cup \mathbf{w}, \mathbf{Z} \setminus \mathbf{Z_w}, \mathcal{I} \cup \mathbf{z_w}, \mathcal{J}, P_{\mathcal{I}, \mathbf{z_w}, \mathcal{J}}, G \backslash \mathbf{Z_w})$.
4 **if** $\mathcal{C}(G \setminus (\mathbf{X} \cup \mathcal{I} \cup \mathcal{J})) = \{S_0, S_1, ..., S_k\}$,
return $\sum_{\mathbf{v} \setminus \{\mathbf{y}, \mathbf{x}, \mathcal{I}\}} \prod_i \mathbf{ID^z}(s_i, (\mathbf{v} \setminus s_i) \setminus \mathbf{Z},$
$\mathbf{Z} \setminus (\mathbf{V} \setminus S_i), \mathcal{I}, \mathcal{J} \cup (\mathbf{Z} \cap (\mathbf{v} \setminus \mathbf{s_i})), P_{\mathcal{I}, \mathcal{J}, \mathbf{Z} \cap (\mathbf{V} \setminus \mathbf{S_i})}, G \backslash (\mathbf{Z} \cap (\mathbf{V} \backslash \mathbf{S_i})))$.
if $\mathcal{C}(G \setminus (\mathbf{X} \cup \mathcal{I} \cup \mathcal{J})) = \{S\}$,
5 **if** $\mathcal{C}(G) = \{G\}$, $\mathbf{FAIL}(G, S)$.
6 **if** $S \in \mathcal{C}(G)$,
return $\sum_{s \setminus \mathbf{y}} \prod_{i | V_i \in S} P(v_i | v_G^{(i-1)} \setminus (\mathcal{I} \cup \mathcal{J}))$.
7 **if** $(\exists S') S \subset S' \in \mathcal{C}(G)$,
return $\mathbf{ID^z}(\mathbf{y}, \mathbf{x} \cap S', \mathbf{Z}, \mathcal{I}, \mathcal{J},$
$\prod_{i | V_i \in C'} P(V_i | V_G^{(i-1)} \cap S', v_G^{(i-1)} \setminus (S' \cup \mathcal{I} \cup \mathcal{J})), S')$.

Figure 2: $\mathbf{ID^z}$: Algorithm capable of recognizing $z\mathcal{ID}$; The variables $\mathcal{I}, \mathcal{J}$ represent indices for currently active Z-interventions introduced respectively by steps 3 or 4. Note that P is sensitive to current instantiations of $\mathcal{I}, \mathcal{J}$.

References

[1] Elias Bareinboim. *Generalizability in Causal Inference: Theory and Algorithms.* PhD thesis, UCLA, 2014.