# OBSERVATIONAL CAUSAL INFERENCE FOR NETWORK DATA SETTINGS

by
Eli Sherman

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
October, 2022

# Abstract

Observational causal inference (OCI) has shown significant promise in recent years, both as a tool for improving existing machine learning techniques and as an avenue to aid decision makers in applied areas, such as health and climate science. OCI relies on a key notion, identification, which links the counterfactual of interest to the observed data via a set of assumptions. Historically, OCI has relied on unrealistic assumptions, such as the 'no latent confounders' assumption. To address this, Huang and Valtorta (2006) and Shpitser and Pearl (2006) provided sound and complete algorithms for identification of causal effects in causal directed acyclic graphs with latent variables. Nevertheless, these algorithms can only handle relatively simple causal queries.

In this dissertation, I will detail my contributions which generalize identification theory in key directions. I will describe theory which enables identification of causal effects when i) data do not satisfy the 'independent and identically distributed' assumption, as in vaccine or social network data, and ii) the intervention of interest is a function of other model variables, as in off-line, off-policy learning, iii) when these two complicated settings intersect. Additionally, I will highlight some novel ways to conceive of interventions in networks. I will conclude with a discussion of future directions.

**Readers:** Ilya Shpitser (advisor), Mark Dredze, David Arbour, Anqi Liu

# Acknowledgments

Perhaps my favorite part of reading another researcher's thesis is getting to read the acknowledgments section. No researcher makes it through a PhD without a great deal of support from colleagues, friends, and family. In a way, the acknowledgments section humanizes the writer and reminds the reader that he or she is more than just a paper-writing machine.

I've been anxiously anticipating the chance to write this section since before I entered grad school. I hope what follows lives up to my expectations about the gravity and importance of expressing gratitude to those who have helped me (and continue to help me!) along the way.

First and foremost, I owe a depth of gratitude to my advisor, Ilya. Over the years, Ilya has been extraordinarily generous with his time. We spent long hours puzzling at chalk boards together as he shared his wisdom about causal inference, research, and academia. Ilya helped me grow from a mistake-prone, naive student fresh from undergrad into a valued colleague, with the ability to contribute independent ideas and perspective to the broader research community. Perhaps most importantly, Ilya has helped me to realize the great value of deliberate, careful, and intentional work ethic. This perspective, which will enhance the quality and impact of my work as I progress in my career, is more important than any paper or project we worked on together during my grad school years. Thank you, Ilya.

Next, thank you to committee member, mentor, and dear friend David Arbour.

David took a chance on me, hiring me as an intern at Adobe several years ago. Since then he has been like a second advisor, serving as a research sounding board, putting up with my precociousness and grandiose ideas. He guided me in navigating 'the real world' (i.e. industry) and I have him to thank for my successful job search. I look forward to continuing our friendship, professionally and personally.

Thank you also to my other committee members: Mark Dredze, Anqi Liu, and Ben Langmead. Mark's mentorship on my mental health research project was invaluable. He taught me a great deal about carrying out applied research, how to think about real world problems, and how to develop skills in areas beyond one's primary area of expertise. Anqi and Ben's advice and guidance brought a fresh perspective to my thesis and GBO, respectively. They challenged me to provide a comprehensive treatment of my work in the service of communicating my contributions to a wide technical audience.

Next, I'd like to acknowledge my many fantastic collaborators. Thank you to Glenn Whitman, Marc Sussman, Stefano Schena, Chin-Siang Ong, Eric Etchill, and Diane Alejo from JHMI Cardiac Surgery and Scott Owens, Hitinder Gurm, and Ulysses Balis of Michigan Medicine. Talking to clinical collaborators has always been one of my favorite parts of research because of their depth of knowledge in an area so different from my own expertise; there's always a different perspective to learn from. Among this group, I would like to give a special thanks to Glenn. Glenn is everything one could ask for in a collaborator: an expert and a teacher, yet unpretentious and always working to identify and address his own blind spots. Glenn guided me in carrying out impactful clinical research and continues to provide valuable feedback towards improving my professional interpersonal skills.

Thank you to Avi Feller, Alex Franks, Betsy Ogburn, Dan Scharfstein, and Michael Rosenblum. All played foundational roles in my exposure to causal inference beyond identification theory. Avi and Alex taught me about the world (and challenges!)

of finite sample inference. Betsy, Dan, and Michael were frequent interlocutors at Bloomberg School of Public Health, providing valuable feedback on research projects over the years.

Thank you to Sara Magliacane, Caleb Miles, Izzie Fulcher, Eli Ben-Michael, Bonnie Smith, Lamar Hunt, and Ben Ackerman, informally the Causal Conference Crew. Many of them are only few years ahead of me in their careers and they have been fantastic informal teachers and mentors as I traverse the causal world. They taught me about their research, navigating academic politics, the job market, and have been fantastic friends.

Thank you to Aurora Schmidt and Vlad Barash both for your insights as collaborators on the ground truth project and for continuing to teach me about your work.

Thank you to Sridhar Mahadevan, Zach Shahn, Yoonyoung Park, Stacy Hobson, and Raya Horesh for hosting me for internships during my graduate studies. Thank you also to Malvern Madondo, Yash Chandak, Aahlad Manas Puli, and Kiran Shiragur whose companionship began during those internships and continues to this day.

Thank you also to Google and the Johns Hopkins Institute for Data Science for supporting my research. Fellowships from those organizations allowed me to spend the final 2.5 years of my PhD exploring my interests, unencumbered by most of the complicated academic incentive structures that typically limit PhD student innovation.

Next, I'd like to give thanks to Amy Cohn. Amy was my first ever research mentor. When I was a sophomore, she brought me into the CHEPS family and introduced me to the worlds of academic and clinical research. Even though I only worked with Amy for a year, she has continue to graciously give me her time, serving as a GBO member and, as recently as last summer, spending an hour on Zoom listening to me speak gibberish as I sought her expertise on optimization. Though I won't be a faculty member (in the immediate term, at least), I hope that one day my students or direct reports look up to me in the way I look up to Amy.

Alongside Amy, I must also acknowledge Jenna Wiens, my other Michigan mentor and my first machine learning mentor. Jenna's tutelage provided me with a strong, foundational understanding of machine learning and clinical data science research. She taught me how to read and write papers, how to give a talk, how to teach a class, and how to engage deeply with a diverse set of stakeholders. As with Ilya, David, and Amy, I owe my career to Jenna and I am so glad that she took me under her wing.

Now, on a more personal level, I want to thank the many close friends who helped me, intellectually and emotionally, through grad school. First, thank you to my lab mates: Dan Malinsky, Razieh Nabi, Rohit Bhattacharya, Jaron Lee, Ranjani Srinivasan, Noam Finkelstein, Zach Wood-Doughty, Numair Sani, Amir Ghassami, and Trung Phung. Thank you to this group for patiently explaining concepts to me (sometimes for a second, third, or seventh time), listening to my rants about academia, causal inference, sports, and a host of other topics, counseling me on career and personal decisions, celebrating good times, comforting me in bad times, being drinking buddies, being meme buddies, and, most of all, just being there. This list is obviously inexhaustive, but I nevertheless hope that they will read this and understand how much they mean to me.

Thank you also to the broader JHU PhD community. Thank you to Adarsh Subbaswamy, Keith Harrigian, Carlos Aguirre, Rachel Sherman (#NotRelated), Benj Shapiro, April Kim, Yasamin Nazari, Thia Steinhardt, Ravi Shankar, Ayushi Sinha, Gabby Beck, Meghana Madhyastha, Jie Ying Wu, Sam Kovaka, Alishah Chator, Arka Choudhuri, Enayat Ullah, and countless others. Thank you also to Max Smith at Michigan. Each of these individuals materially and substantially improved my experience during grad school, from serving as sounding boards for research ideas, to fighting together for a better student experience and commiserating about the existential dread of grad life. I am thankful to all for sharing this time with me and I look forward to continue these friendships as I move into the next phase.

Outside of academia[1], thank you to Yvonne and Paul for being the most dynamic of duos, and the dearest of friends. Thank you to Steph for being the best roommate in the world (sorry Ben), a charming guide to Charm City, my belay partner, my relationship coach, my shoulder to cry on, and my big sister.

Thank you to Yogi and Olive. I know you can't read this, but maybe when AGI becomes a viable product, we'll be able to translate this thesis into bark-speak and then you'll understand how much the warmth of your cold, wet noses supported me through these last few years.

Thank you to Brendan for being my rock since week 6. Thank you for being my foodie competition, my fitness inspiration, and my lifeline to the mitten. Thank you for helping inspire my interest in policy, urban development, and the climate. Our lives will forever be intertwined and I can't wait to see how we'll inspire each other next. I love you, brother.

Thank you to Ben Cher[2]. You've been integral to who I have become as a musician, an outdoorsman, and traveler, and also as a researcher and intellectual. Thank you for spending long nights and early mornings with me, arguing about everything from the pros and cons of managed care organizations and the ethics of centralized resource allocation to the most effective way to construct a PB&J (hint: don't put trail mix on it instead of actual peanut butter). Thank you for keeping an open mind and opening my mind. Thank you for teaching me how to admit when I'm wrong. Thank you for being the Jacques to my Pierre. I can't wait to listen to Brahms 1 atop Mount Brunswick with you.

[1] Last names omitted for privacy. I was deliberate about including last names for the academic crowd because, after all, academia is all about cults of personality: if you're reading this, especially at some point in the distant future, I encourage you to look up these fine folks. See what they're up to and read and cite their papers. Maybe you could even reach out to collaborate!

[2] Yes, Ben is an academic. Go read his papers

Thank you to Nam and Ashleigh Le. Thank you for looking after Jessica while she waited for me, and thank you for looking after both of us now that I'm here. Thank you for stand up and board games and CAVA nights. Thank you, Nam, for being my soccer and classical concert buddy and thank you, Ashleigh, for being my personal makeup and fashion consultant. Your love and support over the past several months in particular has been critical in helping me cross the finish line.

Thank you to my parents. Invoking Billy Collins[3], this thesis is for you. You gave me life, love, support, food, a roof, a bed, and an education. In return, I give you this thesis. Hopefully, "this is enough to make us even".

Thank you to my brother, David, my oldest friend. Thank you for letting me bask in your shadow when we were little and for helping me to step out from under it to create my own as we grew up. Thank you for talking me off the Master-out ledge. Thank you for being my primary source of procrastinatory reading and musical ecclecticism and for sharing your love of sports, Michigan or otherwise. I cannot envision getting through my PhD without our daily correspondence and your love.

Finally, thank you to my beloved wife, Jessica Bonnie[4]. As I graduate *from* Hopkins, you are the most important thing I will take with me, the most important thing I have gained. Thank you for being my best friend, my couch co-potato, and partner. Thank you for picking up the slack for me as I struggled to balance work, mental health, and tasks as simple as folding the laundry. Thank you for coaching me through my self-doubt and helping me to appreciate the value of slowing down and taking stock of what's important in life. Thank you also for opening your life to me so that I can be part of *your* growth story. I can't wait to see what you become and to support you as you have supported me.

[3] https://www.poetryfoundation.org/poems/50975/the-lanyard

[4] Also an academic! Don't call her 'Jessica Sherman'.

# Contents

# List of Figures

# Chapter 1

# Introduction

## 1.1 Background: Observational Causal Inference in Empirical Science

Practitioners in fields ranging from economics to education to healthcare are constantly searching for better ways to formulate, implement, and measure the impacts of policy. Dating back at least to the Industrial Revolution[1], policy- and decision-making has increasingly relied on empirical analyses which depend on our ability to reason about *counterfactual* scenarios. Decision-makers need to answer questions like, 'what would be the *effect* on voter turnout if this automatic voter registration policy were enacted?', or 'how much would the conditions of patients of different subtypes improve *had* they been prescribed the trial drug?' The field of causal inference is concerned with developing methods to rigorously evaluate such counterfactual queries.

The gold standard for determining the presence and magnitude of causal relationships is the randomized control trial. In an idealized case, the study population can be randomly split into 'treated' and 'control' groups, and then average outcomes in each group can be compared directly to determine the 'effect' of receiving treatment. Key to this type of analysis is the assumption that the randomization scheme is sufficient; the different study groups must be reasonably similar so that any observed

[1] Consider, for instance, the writings and epidemiological work of John Snow (Snow, 1849).

differences in outcomes can be interpreted as being a result of the treatment alone. Formalizing the assumptions necessary to ensure sufficient randomization is the domain of 'experimental design', a sub-field of causal inference. By applying methods from experimental design, practitioners can confidently draw causal conclusions and address the counterfactual queries that motivate their analyses.

Unfortunately, experimentation is often infeasible due to expense or ethical concerns. For instance, suppose we randomly select a subset of the population, and instruct these individuals to smoke cigarettes so that pulmonologists can establish a firm estimate of the effect of smoking on lung cancer risk before the age of 60. This is clearly unethical since we already have a strong prior about smoking's adverse effects and we'd be consciously harming the smoking-assigned group.

On the other hand, non-randomized, observational data (such as follow-up surveys given to patients, or electronic medical records) is typically abundant. Following the smoking example above, we might obtain *observational* data on a large population of individuals who either did or did not choose to smoke and likewise did or did not develop lung cancer. Observational causal inference (OCI) seeks to develop methods that use these non-randomized data to emulate a target randomized control trial and obtain a reliable estimate of the causal relationships the idealized trial would have revealed.

The first step in an observational causal analysis is to identify or collect a data set from a population similar to the population of interest. The data set will contain information on the outcome of interest, variable(s) signifying whether the patient was treated or untreated[2], and possibly other variables that help characterize the study population (e.g., demographics or clinical measurements). A key property of *observational* studies, however, is that the analyst will have little or no control over

[2] I am implicitly describing a study performed with a binary treatment. The causal inference community has also defined approaches for handling continuous treatments

the data collection process. He or she may not be able to specify what variables are observed and, by definition, will not be able to manipulate the data generating process or control treatment assignments as in an experimental study.

Once a suitable data set has been selected, the analyst will then specify a parameter of interest, such as the average difference in outcomes between the treated and control groups, known as the 'average causal effect'. Because this parameter of interest is constructed from (inherently unmeasurable) counterfactual quantities, the analyst must find a way to establish a link between the observed data she has access to and the counterfactuals. This linking process, known as *identification*, entails making assumptions about the observed data generating process. Observational analyses of this sort have been used to study phenomena in fields as varied as epidemiology (Robins, Hernan, and Brumback, 2000; Robins et al., 1992), air pollution (Papadogeorgou, Choirat, and Zigler, 2019), and algorithmic advertising (Nabi et al., 2022).

## 1.2 Motivation: The Limitations of Observational Causal Inference Assumptions

The OCI community has developed a number of formalisms for identifying causal effects and leveraging observational data to estimate those effects[3] Over time, these formalisms have become increasingly general. Wright, 1934 serves as an early example of using graphical models to analyze causal paths, motivated by questions in animal husbandry. Assumptions in that era were typically highly restrictive: Wright assumed linear models, for instance, while his contemporary, Ronald Fisher, studied experimental design using randomized trial data (Fisher, 1935). Later, Rubin, Pearl, and Robins (Rubin, 1974; Pearl, 2009; Robins, 1986) developed more comprehensive causal frameworks that rely on fewer assumptions, deriving, for instance, general theories of non-parametric causal

[3] I will review the fundamental points in detail in Chapter 2 and direct the reader to Pearl (2009) for a more complete treatment.

models. In turn, these frameworks enabled further explorations and insights, such as general identification theories (e.g., the works of Shpitser (Shpitser and Pearl, 2006; Shpitser, 2013) and Barenboim (Bareinboim and Pearl, 2016; Pearl and Bareinboim, 2011)) and recent work on automatic derivation of robust and efficient semi-parametric estimators from causal models (Bhattacharya, Nabi, and Shpitser, 2020; Rotnitzky and Smucler, 2019).

In this thesis, I will describe my efforts to generalize the observational causal inference framework in two areas where modeling assumptions remain relatively restrictive. The first concerns the study of causal dynamics in networked systems, where there is dependence between study subjects or data samples. The second concerns the study of policies, such as clinical treatment guidelines and means-tested welfare programs. Here I will describe the shortcomings of existing approaches with respect to these two areas. In the main body of this thesis I will present methods and theory that expands the bounds of what is possible when performing causal analyses in these more general domains.

### 1.2.1 Policy Analysis

Much of the causal inference literature focuses on analyses that give insight into *population-level* average effects. This is carried out by running, or emulating, a trial with a finite number of treatment arms and with little dependence between baseline variables and treatment arm assignment. For instance, we might simply assign patients to take one of two drugs (Drug A vs. Drug B), which might allow us to draw conclusions about the relative efficacy of the drugs in the study population *on average*.

In practice, however, we know that many treatments affect each patient differently and so the 'best' treatment for a given patient might differ from the 'best' treatment for the population average. It seems reasonable, then, to conclude that to make decisions optimally, we should derive methods that enable us to tailor treatments to individual subjects. In causal inference, this corresponds to finding a mapping between characteristics of the study subject and an action that optimizes the outcome of interest. This mapping is known as a 'treatment rule', 'dynamic treatment regime', or 'policy' and has been studied in both the causal inference literature (Chakraborty and Moodie, 2013) and the reinforcement learning literature (Bertsekas and Tsitsiklis, 1996). Despite a well-developed literature in these areas, there remains a gap in characterizing, formally, when the effects of policies can be estimated from observed data under a variety of assumptions and policy setups. I help fill that gap with work I will present in this thesis.

### 1.2.2 Network Analysis

The assumption of independent and identically distributed (iid) samples is ubiquitous in data analysis. In many research areas, however, this assumption simply does not hold. For instance, social media data often exhibits dependence due to homophily and contagion (Shalizi and Thomas, 2011). Similarly, in epidemiology, data exhibiting herd immunity is likely dependent across units. Likewise, signal processing and sequence learning often consider data that are spatially (Mnih et al., 2015) or temporally (Sutskever, Vinyals, and Le, 2014) dependent.

The difficulties of inter-sample dependence also arise in causal inference. For example, consider the notion of a 'gifted & talented' (G/T) program in public education (Davis and Rimm, 1989; Hodges et al., 2018). G/T programs aim to identify 'gifted' children and *intervene* on their educational trajectory by moving them to an alternate classroom for part of the school day. A key obstacle to analyzing the effectiveness of such as program is that fact that each student's treatment – inclusion or exclusion from the program – can affect his peers' educational outcomes: the student being included could change the learning environment in *both* the gifted and non-gifted classrooms. This phenomenon, where some units' treatments may causally affect other units' outcomes is known as *interference* in the causal inference literature (Cox, 1958; Hudgens and Halloran, 2008; Ogburn, VanderWeele, et al., 2014).

Even under the iid assumption, making causal inferences from observed data is difficult due to the presence of unobserved confounding. This difficulty is worsened when the data are subject to interference and other inter-sample dependences, as described in detail in Shalizi and Thomas (2011) and Ogburn, Shpitser, and Lee (2018). These difficulties prevent identification of causal parameters of interest (i.e. establishing a link from the observed world to the counterfactual world), and pose estimation challenges such as reducing effective sample size.

To properly evaluate counterfactual queries, such as 'what would the average student's standardized test scores have been had they been included in the G/T program?', it is necessary to develop methods specifically designed to handle the complexities of data dependence. In this thesis, I will describe some of my own proposals in that direction.

## 1.3 Outline and Contributions

This thesis is organized as follows. Chapter 2 provides a thorough review of graphical causal inference, a research area that strongly influenced the work presented in the other chapters. The subsequent chapters contain original research (each corresponding to a previously-published paper on which I was a primary author) that tackles some of the issues highlighted in the Motivation section above.

In Chapter 2, I will fix the majority of the notation used throughout the thesis and so this chapter should be viewed as a reference. Additional notation will introduced in individual chapters as necessary.

In Chapters 3 and 4, I will present novel causal identification algorithms and supporting theory that proves their correctness. As detailed in Chapter 2, identification algorithms enable researchers to understand what is 'possible' with regard to causal effect estimation. They can help determine when an effect of interest is estimable from observed data or, conversely, when more data or assumptions are necessary.

In Chapter 3, I focus on the question of identification of policies, while ignoring the complexities of non-iid data. The work I present in Chapter 3 is much more far-reaching than considering solely policy effects. The work also considers identification in the context of 'mediation analysis', where a researcher might wish to assess the direct and indirect effects of adopting an action or treatment rule[4]. While this broader subject matter is important and challenging, I have focused my framing (e.g., in the Motivation section above) on policies, specifically, because of how policies play an integral part in my other work.

As a complement, Chapter 4 focuses on identification theory in the interference and dependent data setting. I describe how a certain type of graphical model called a 'segregated graph' (Shpitser, 2015) can be used to represent network data. I illustrate how segregated graphs can be used to construct a general representation for interference. I then give an algorithm for establishing when causal effects arising from simulated trials are, or are not, identifiable.

Chapter 5 serves as a synthesis, bridging the gap between Chapters 4 and 3. I present an identification algorithm for evaluating whether the effects of policies can be estimated from observed data when there is interference or data dependence present.

Finally, in Chapter 6, I consider a different use-case of the policy analysis framework presented in earlier chapters. I study how we might estimate the effects of modifying social network ties (adding or removing connections), rather than modifying the values of individual variables. This work gives some perspective on both the dependent data

[4] For instance, an analyst might wish to separately evaluate the chemical effect of a drug (likely positive effect on the outcome) and the adherence effect (possibly a negative effect if the drug produces severe side effects).

and policy analysis problem spaces, and also touches on the philosophy of interventions which underlies much of the mathematical literature on causality.

I conclude with a reflection on the impact of these works in the research community and provide a discussion for how this theory-grounded work could be translated into practical use. As a model, I refer the reader to my appendix chapters, which cover applied healthcare research I carried out during my PhD studies that does not directly relate to the work presented in the main body of the dissertation.

# Chapter 2

# Causal Inference Foundations

## 2.1 Background and Notation

In this chapter, I fix notation, describe key assumptions, and review foundational causal inference results. I will not cover *all* concepts used throughout the thesis in this chapter. Instead, I will provide the basic background here and introduce other concepts as necessary in the ensuing body chapters.

### 2.1.1 Preliminaries

I will adopt a formalism of causal inference that relies on graphical models, potential outcomes, and probability theory. This formalism relies heavily on the random variables and their distributions. Variables and their realizations will be denoted in upper and lower case, respectively: $V$ and $v$. Sets will be denoted in boldface: $\mathbf{V}$ and $\mathbf{v}$. The state space of a variable $V$ will be denoted by $\mathfrak{X}_V$

As described in Chapter 1, the goal of observational causal analyses is to emulate a hypothetical randomized control trial so that we can evaluate a counterfactual query. By convention, I will represent the outcome (a random variable) with the letter $Y$ and the treatment (also a random variable) with the letter $A$. The expression $Y(A = a)$, often shortened to $Y(a)$, represents the *counterfactual random variable*[1],

[1] Typically just ‘counterfactual’ or ‘counterfactual variable’

'what would the value of the outcome have been had we, possibly contrary to observed fact, set the treatment to the value $a$?', while $p(Y(a))$ represents the *distribution* of that counterfactual variable.

These notions extend to sets of treatments and outcomes. Let $\mathbf{V}$ be the set of variables in a model. Let $\mathbf{Y} \subseteq \mathbf{V}$ and $\mathbf{A} \subseteq \mathbf{V} \setminus \mathbf{Y}$. Then, we can define $\mathbf{Y}(\mathbf{a}) \equiv \{\mathbf{Y}\}(\mathbf{a}) \equiv \{Y(\mathbf{a}) \mid Y \in \mathbf{Y}\}$. The distribution $p(\mathbf{Y}(\mathbf{a}))$ is sometimes written as $p(\mathbf{Y}|\text{do}(\mathbf{a}))$ (Pearl, 2009).

### 2.1.2 Fundamental Assumptions

As hinted in the previous chapter, these counterfactual objects are not generally observable. Critically, this means that $p(Y(A = a)) \neq p(Y|A = a)$ in general and using the latter as an estimate of the former would yield biased results. That said, if we make assumptions about the data generating process, it *is* possible to obtain equality between these two distributions, or otherwise express $p(Y(A = a))$ as a function of the observed data distributions, and thus make unbiased estimates. A counterfactual distribution that is a functional of the observed data distribution is said to be *identified*, and the process which establishes whether such a functional exists is called *identification*.

There are many combinations of assumptions that can help identify effects and thus most sets of assumptions that lead to identification are *sufficient* rather than *necessary*. One well-studied approach is to assume *positivity*, *consistency*, and *ignorability*. Positivity states that $0 < p(A = a) < 1$ for all $a$. A violation of this assumption would mean trying to perform an analysis in which one of the treatments of interest was never prescribed and making it virtually impossible to draw conclusions about the impact choosing that treatment would have. Throughout this thesis, I will assume distributions are positive.

Consistency states that $A = a \implies Y(a) = Y$. That is, if the observed value of the treatment $A$ was $a$ then the counterfactual $Y(a)$ is equal to the observed value $Y$. Critically, this says nothing about the value of our target $Y(a)$ if the treatment variable $A$ were equal to a value other than $a$. Returning to the smoking example from the Background section of Chapter 1, suppose that $A$ denotes whether the subject smokes and $Y$ denotes whether the subject has received a lung cancer diagnosis before age 60[2]. For each subject in the study we will observe *either* $Y(\text{smoke})$ *or* $Y(\text{no smoke})$. For the subjects where $A = \text{smoke}$ we know the value of $Y(\text{smoke})$ by application of consistency, but not $Y(\text{no smoke})$.

*Ignorability* places assumptions on the relationship between the treatment and the counterfactual outcomes: $Y(a) \perp\!\!\!\perp A$. This assumption is analogous to randomization. In a randomized control trial, the intervention value $a$ is of primary interest, rather than the random variable $A$. When we randomize subjects to a study arm, we are removing the dependence between the random variable and the counterfactual under the intervention.

Under these two assumptions, it can be proven that $p(Y(a)) = p(Y|A = a)$. In the following, the first equality holds by ignorability and the second holds by consistency.

$$p(Y(a)) = p(Y(a)|A = a) = p(Y|A = a),$$

In practice, however, the existence of confounding variables means that ignorability typically does not hold outside of randomized studies. In these situations, it may be the case that the weaker assumption, conditional ignorability, does hold. Formally, conditionally ignorability holds if $Y(a) \perp\!\!\!\perp A|\mathbf{C}$, where $\mathbf{C}$ represents the set of confounding variables. In this case, the above derivation is modified in the following way, with the first equality by chain rule, the second by conditional ignorability, and the

[2] Let's also assume that all patients are 60 years or older so that we can ignore issues relating to censoring for the purposes of this simple example

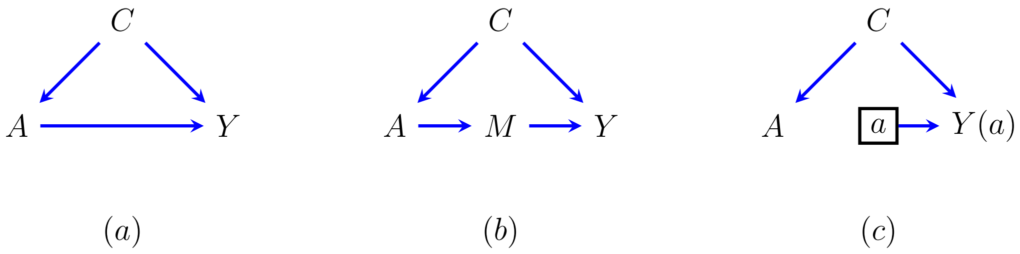


Figure 2-1. a) A DAG representing the classic conditionally ignorable model. All variables are assumed to be observed. b) A DAG similar to that in (a), but with a mediating variable $M$ such that there is no direct connection between $A$ and $Y$. c) A post-intervention graph obtained by intervening on $A$ in the graph in (a). This graph contains counterfactuals and thus does not represent the observed world.

last (as before) by consistency.

$$\begin{aligned} p(Y(a)) &= \sum_{\mathbf{C}} p(Y(a)|\mathbf{C})p(\mathbf{C}) \\ &= \sum_{\mathbf{C}} p(Y(a)|A = a, \mathbf{C})p(\mathbf{C}) \\ &= \sum_{\mathbf{C}} p(Y|A = a, \mathbf{C})p(\mathbf{C}) \end{aligned}$$

If (conditional) ignorability and consistency do not hold, identification may not be possible. Much of the causal inference literature studies identification in settings where conditional ignorability does not hold. Before discussing that theory in more detail, I will introduce causal graphical models, which are used heavily in developing general identification theories

## 2.2 Graphical Models

Towards constructing a general criteria for determining when a counterfactual is or is not identifiable, we will turn to graphical models. Graphs consist of nodes and edges. In a graph representing a statistical model, nodes correspond to variables in a probability distribution while edges represent the relationships between variables. As a simple case, consider the class of models known as directed acyclic graphs (DAG).

In DAGs, all edges have a single arrowhead and the graph as a whole does not have any cycles (i.e., no $V_1 \rightarrow V_2 \rightarrow \cdots \rightarrow V_1$ structures allowed). For example, Figure 2-1 (a) represents the joint distribution $p(Y, A, C)$ for our conditionally ignorable example above. The joint distribution $p(\mathbf{V})$ is said to be *Markov* to a graph $\mathcal{G}$ (where $\mathbf{V}$ denotes the set of variables in the distribution) if we can rewrite it as:

$$p(\mathbf{V}) = \prod_{V \in \mathbf{V}} p(V | pa_{\mathcal{G}}(V)), \tag{2.1}$$

where $pa_{\mathcal{G}}(V) \equiv \{W \in \mathbf{V} \mid W \rightarrow V\}$ denote the *parents* of $V$ in the $\mathcal{G}$.

That is, the joint distribution *factorizes* as a product of conditional distributions. One conditional distribution appears in the factorization for each variable in the graph. In Figure 2-1 (a) the factorization is simply the chain rule of probability: $p(Y, A, C) = p(Y|A, C)p(A|C)p(C)$.

In addition to parents, we can define a variety of genealogic sets to describe relationships in a graph. For DAGs, these additional genealogic sets include

$$\begin{aligned}
\text{children: } \mathrm{ch}_{\mathcal{G}}(V) &\equiv \{W \in \mathbf{V} \mid V \rightarrow W\} \\
\text{ancestors: } \mathrm{an}_{\mathcal{G}}(V) &\equiv \{W \in \mathbf{V} \mid W \rightarrow \ldots \rightarrow V\} \\
\text{descendants: } \mathrm{de}_{\mathcal{G}}(V) &\equiv \{W \in \mathbf{V} \mid V \rightarrow \ldots \rightarrow W\} \\
\text{non-descendants: } \mathrm{nd}_{\mathcal{G}}(V) &\equiv \mathbf{V} \setminus \mathrm{de}_{\mathcal{G}}(V).
\end{aligned}$$

These notions generalize disjunctively to sets of variables. For instance, for $\mathbf{V}' \subseteq \mathbf{V}$, we have $\mathrm{pa}_{\mathcal{G}}(\mathbf{V}') \equiv \bigcup_{V \in \mathbf{V}'} \mathrm{pa}_{\mathcal{G}}(V)$. We also define *strict parents*: for $\mathbf{A} \subseteq \mathbf{V}$, $\mathrm{pa}^s_{\mathcal{G}}(\mathbf{A}) \equiv \mathrm{pa}_{\mathcal{G}}(\mathbf{A}) \setminus \mathbf{A}$. Finally, for a variable $\mathbf{A} \subseteq \mathbf{V}$ and a graph $\mathcal{G}$, $\mathcal{G}_{\mathbf{A}}$ will denote the subgraph of $\mathcal{G}$ containing only the vertices in $\mathbf{A}$ and edges between them. These represent the basic notions necessary to work with DAG models. I will introduce other concepts and notation later as I describe more general models.

Through rules known as 'd-separation' (Pearl, 1988), we can determine which variables are (conditionally) independent in any distribution Markov to a DAG. When evaluating d-separation between two (possibly singleton) sets $\mathbf{A}$ and $\mathbf{B}$ of variables, with a (possibly empty) conditioning set $C$, we consider all *paths* between the two sets[3]. A path is said to be *closed* or *blocked* if it contains one of the following substructures with the corresponding condition:

- a 'chain': $X \to Y \to Z$ with $Y \in \mathbf{C}$
- a 'fork': $X \leftarrow Y \to Z$ with $Y \in \mathbf{C}$
- a 'collider': $X \to Y \leftarrow Z$ with $Y \notin \mathbf{C}$ and $\mathrm{de}_{\mathcal{G}}(Y) \cap \mathbf{C} = \emptyset$

If *all* paths from $\mathbf{A}$ to $\mathbf{B}$ are blocked, then $\mathbf{A} \perp\!\!\!\perp \mathbf{B}|\mathbf{C}$. Using d-separation, we can see that $Y \perp\!\!\!\perp A|\{C, M\}$ and $M \perp\!\!\!\perp C|A$ in Figure 2-1(b).

## 2.3 Causal Graphical Models

*Causal* graphical models expand upon these ideas by placing a causal interpretation on the edges in a graph. Formally, I will assume the 'structural causal model' (Pearl, 2009): each variable $V$ in the model is a function $f_V(pa_{\mathcal{G}}(V), \epsilon_V)$ of i) that variable's parents, and ii) an error term $\epsilon_V$[4]. These functions are sometimes referred to as *structural equations*.

To illustrate, consider Figure 2-1(a). $C$ is determined by some function $f_C(\epsilon_C)$ since it has no parents. $A$ is a function $f_A(C, \epsilon_A)$ of $C$ and an independent error term $\epsilon_A$. Lastly, $Y$ is a function $f_Y(A, C, \epsilon_Y)$ of $C$, (random) $A$, and an independent error term $\epsilon_Y$.

[3] We look at all acyclic sequences of nodes and edges with the starting node in $\mathbf{A}$ or $\mathbf{B}$ and the ending node in the other

[4] Keep in mind: these are stochastic models, even though they are not 'statistical' in the traditional sense

### 2.3.1 Interventions

Causal models of a DAG $\mathcal{G}$ describe sets of distributions defined on counterfactual random variables of the form $V(\mathbf{a})$, where $\mathbf{a}$ are values of $\text{pa}_{\mathcal{G}}(V)$. That is, $\mathbf{V}$ is the set containing any joint distribution over all potential outcome random variables, where the sets of variables

$$\{\{V(\mathbf{a}_V) \mid \mathbf{a}_V \in \mathfrak{X}_{\text{pa}_{\mathcal{G}}(V)}\} \mid V \in \mathbf{V}\}$$

are mutually independent (Pearl, 2009).

These *atomic counterfactuals* model the relationship between $\text{pa}_{\mathcal{G}}(V)$, representing direct causes of $V$, and $V$ itself. We can explicitly connect counterfactuals to the notion of *interventions*. As defined previously, the counterfactual $V(\text{pa}_{\mathcal{G}}(V))$ represents the value of $V$ when $\text{pa}_{\mathcal{G}}(V)$ is *set* to $\mathbf{a}_V$. In the structural equation model, when set $\text{pa}_{\mathcal{G}}(V) \leftarrow \mathbf{a}_V$, we use the *intervention values* $\mathbf{a}_V$ in all downstream evaluations of $\text{pa}_{\mathcal{G}}(V)$. That is, $V$'s structural equation is evaluated as $f_V(\text{pa}_{\mathcal{G}}(V) = \mathbf{a}_V, \epsilon_V)$ post-intervention (Pearl, 2009) and so we have equality:

$$V(\mathbf{a}_V) = f_V(\text{pa}_{\mathcal{G}}(V) = \mathbf{a}_V, \epsilon_V)$$

From the atomic counterfactuals, all other counterfactuals may be defined using recursive substitution. For any $\mathbf{A} \subseteq \mathbf{V} \setminus \{V\}$,

$$V(\mathbf{a}) \equiv V(\mathbf{a}_{\text{pa}_{\mathcal{G}}(V)\cap\mathbf{A}}, \{\text{pa}_{\mathcal{G}}(V) \setminus \mathbf{A}\}(\mathbf{a})). \tag{2.2}$$

I will refer this basic intervention type, where there is a single, constant intervention value specified for each variable, as the *node intervention*.

Graphically, node interventions are represented by a procedure known as 'graph manipulation' or 'graph surgery'. When we intervene on $A$, setting it to $a$, we add a new node (represented with a square) for the intervention value $a$ as in Figure 2-1(c). Since the value of (random variable) $A$ was not changed, the node for $A$ keeps its *incoming* edges. In turn, since downstream values are affected by the intervention, $A$'s *outgoing* edges are shifted to the new $a$ node. replace the random $A$ with $a$ in the downstream structural equations and replace those downstream variables by counterfactuals. $C$'s node doesn't change since $C$ is pre-treatment. $Y$ becomes a counterfactual $Y(a)$ since it is downstream of the intervention. The graph in Figure 2-1(c) is known as a single-world intervention graph (SWIG) (Richardson and Robins, 2013). By applying the rules of d-separation to this graph, we observe that $Y(a) \perp\!\!\!\perp A|C$, which matches the conditional ignorability assumption introduced above.

### 2.3.2 Identification Theory in Causal DAGs

Recall from above that a causal parameter is said to be *identified* in a causal model if it is a function of the observed data distribution $p(\mathbf{V})$ and *non-identified* otherwise.

In all causal models of a DAG $\mathcal{G}$ considered in the literature, all interventional distributions $p(\{\mathbf{V} \setminus \mathbf{A}\}(\mathbf{a}))$ are identified by the *g-formula* (Robins, 1986):

$$p(\mathbf{V} \setminus \mathbf{A}) = \prod_{V \in \mathbf{V} \setminus \mathbf{A}} p(V|pa_{\mathcal{G}}(V))|_{\mathbf{A}=\mathbf{a}} \tag{2.3}$$

This formula describes the counterfactual for the entire distribution; the formula for a specific variable, like the outcome $Y$, can be obtained by marginalizing. The formula also permits intervening on sets of treatment variables, rather than singletons. The $|_{\mathbf{A}=\mathbf{a}}$ notation means 'evaluated at', encoding the above ideas regarding evaluation of structural equations that are downstream of the intervention variable.

Not all interventional distributions are identified when there are hidden variables present in the causal model. I discuss identification theory in hidden variable DAGs next.

## 2.4 Latent Variable Models

Suppose that we are studying a clinical phenomenon and know the true causal DAG associated with that phenomenon. This DAG might contain some variables which cannot be measured (e.g., socioeconomic status is rarely fully captured in observational studies). This non-observability often leads to non-identifiability. As an intuitive example, suppose $C$ is latent in Figure 2-1(a). We would not be able to determine whether changes we observe in $Y$ (the patient receiving a lung cancer diagnosis) are due to the action taken $A$ (the patient smoking) or some latent common cause $C$ (e.g., patients with lower socioeconomic status are more likely to smoke (Hiscock et al., 2012) and separately are more likely to be exposed to cancer-causing pollution (Cohen and Pope 3rd, 1995)).

In order to handle latent variables, I will introduce a new type of edge: the bidirected edge $\leftrightarrow$. If $A \leftrightarrow B$ in a graph, this signifies that $A$ and $B$ share a common cause[5]. Verma and Pearl (1990) proposed a procedure, known as the *latent projection* operation, for converting a DAG in which some variables are latent to a mixed graph with latent variables replaced by either directed ($\rightarrow$) or bidirected edges as described below.

Formally, given a DAG $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$, where $\mathbf{V}$ are observed and $\mathbf{H}$ are latent, define a latent projection $\mathcal{G}(\mathbf{V})$ to be an acyclic directed mixed graph (ADMG) with the vertex set $\mathbf{V}$ and $\rightarrow$ and $\leftrightarrow$ edges. An edge $A \rightarrow B$ exists in $\mathcal{G}(\mathbf{V})$ if there is a directed path from $A$ to $B$ in $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$ with all intermediate vertices in $\mathbf{H}$. Similarly, an edge $A \leftrightarrow B$ exists in $\mathcal{G}(\mathbf{V})$ if there is a path without consecutive edges $\rightarrow \circ \leftarrow$ from $A$ to $B$ with the first edge on the path of the form $A \leftarrow$ and the last edge on the path of the form $\rightarrow B$, and all intermediate vertices on the path in $\mathbf{H}$.

For instance, the two graphs in 2-2 contain a latent variable $U$ which affects both

[5] Or that they share multiple common causes

the treatment and outcome variables $A$ and $Y$. Performing the latent projection operation on these graphs yields Figure 2-4, where we've removed the latent variable in each graph and replaced it with a bidirected edge ($\leftrightarrow$) between the causal children of the latent variable. This leads to another graphical set definition, the *district*:

$$\text{dis}_{\mathcal{G}}(V) \equiv \{W \in \mathbf{V} | V \leftrightarrow \cdots \leftrightarrow W\}$$

By convention, $\text{an}_{\mathcal{G}}(V) \cap \text{de}_{\mathcal{G}}(V) \cap \text{dis}_{\mathcal{G}}(V) = \{V\}$. Observe that the set of districts in a graph $\mathcal{G}$ partition the graph. This set is denoted by $\mathcal{D}(\mathcal{G})$. In a DAG, the set of districts is trivially the set of all singleton vertices, since $\leftrightarrow$ edges are not present.

Observe that multiple DAGs can be mapped to the same ADMG via the latent projection operation and so ADMGs represent *equivalence classes* of latent-variable DAGs. It has been shown (see, for instance, Richardson et al., 2017, though other proofs exist) that for any two DAGs that yield the same ADMG when the latent projection operation is applied the identification theory will be the same. This property is critical for identification theory: it enables developing theories in terms of latent projections directly rather than needing to reason about the DAGs in a given equivalence class separately.

A general algorithm for identification of interventional distributions was proposed and proven sound in Tian and Pearl (2002). The soundness property states that if the effect of interest is identifiable then the algorithm will output the correct function of the observed data. Shpitser and Pearl (2006) simplified that original algorithm[6] and proved the completeness property of the identification criterion. The completeness property states that if the algorithm fails to identify the target effect, then no method can successfully yield identification for that effect. The Shpitser and Pearl (2006) algorithm is re-printed in Appendix C for reference, however, I will not provide a

[6] The original Tian algorithm resembles a complicated computer program, spanning several pages of a paper. The 2006 Shpitser version leverages the recursiveness of Tian's algorithm in order to consolidate it to just 7 lines.

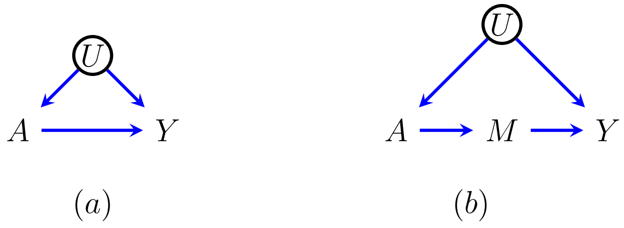


Figure 2-2. a) a DAG representing the classic conditionally ignorable model but with a latent variable confounder $U$ rather than having all variables assumed observed. b) A graph similar to that in (a), but with an intermediate variable $M$ on a causal path from the treatment $A$ to the outcome $Y$.

walkthrough[7]. Instead, I will focus on a more recent reformulation of the ID algorithm posited in Richardson et al. (2017). This reformulation serves as the basis for much of the novel identification theory presented in this thesis and so I will now review key concepts from Richardson et al. (2017).

## 2.5 Identification in Latent Variable DAGs

Above, I highlighted how identified interventional distributions in a DAG can be expressed as a truncated factorization via the g-formula (Eq. 2.3). I will now describe how identifiable interventional distributions can be expressed as a truncated *nested* factorization (Richardson et al., 2017) of a latent projection ADMG. The nested factorization of $p(\mathbf{V})$ with respect to an ADMG $\mathcal{G}(\mathbf{V})$ is defined on *Markov kernel* objects derived from $p(\mathbf{V})$ and conditional ADMGs derived from $\mathcal{G}(\mathbf{V})$. I will first formally define kernels and conditional ADMGs, and a *fixing operation* which is used to derive these objects. These concepts will build up to a formal definition of the nested factorization.

[7] A detailed explanation of the mapping from the 2006 algorithm to the 2017 one line algorithm is given in Section 4 of Shpitser and Sherman (2018)

### 2.5.1 Conditional ADMGs and Markov Kernels

A Markov kernel $q_{\mathbf{V}}(\mathbf{V}|\mathbf{W})$ is a mapping from $\mathfrak{X}_{\mathbf{W}}$ to normalized densities over $\mathbf{V}$. Conditioning and marginalization are defined in kernels in the usual way:

$$q_{\mathbf{V}}(\mathbf{A}|\mathbf{W}) \equiv \sum_{\mathbf{V}\setminus\mathbf{A}} q_{\mathbf{V}}(\mathbf{V}|\mathbf{W}); \qquad q_{\mathbf{V}}(\mathbf{V}\setminus\mathbf{A}|\mathbf{A}\cup\mathbf{W}) \equiv \frac{q_{\mathbf{V}}(\mathbf{V}|\mathbf{W})}{q_{\mathbf{V}}(\mathbf{A}|\mathbf{W})},$$

for $\mathbf{A} \subseteq \mathbf{V}$. A conditional distribution is one type of kernel, but others are possible.

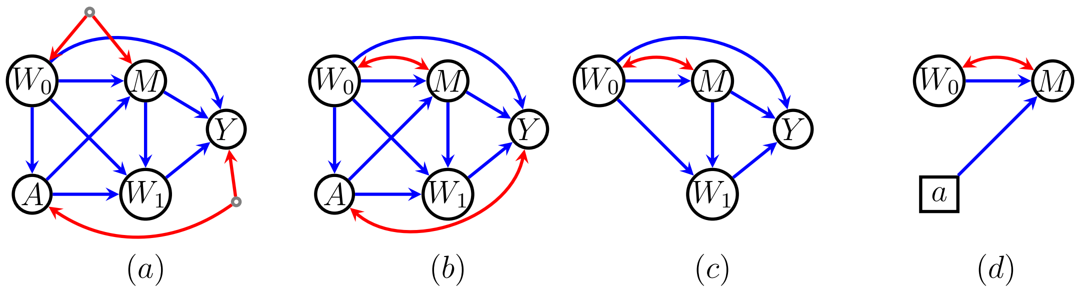


Figure 2-3. (a) A causal model with a treatment $A$ and outcome $Y$. (b) A latent projection of the DAG in (a). (c) The graph derived from (b) corresponding to $\mathcal{G}_{\mathbf{Y}^*} = \mathcal{G}_{\{Y,M,W_0,W_1\}}$. (d) A CADMG corresponding to $p(M, W_0|\text{do}(a))$.

A conditional ADMG (CADMG) $\mathcal{G}(\mathbf{V}, \mathbf{W})$ is a type of ADMG where nodes are partitioned into two sets. The set $\mathbf{W}$ corresponds to *fixed* constants, and the set $\mathbf{V}$ corresponds to *random variables*. A CADMG has the property that no edges with an arrowhead into an element of $\mathbf{W}$ may exist. Intuitively, a CADMG represents a situation where some variables have already been intervened on. Pearl introduced a similar concept called the 'mutilated graph' in Pearl (2009). For example, the graph in Fig. 2-3 (d) is a CADMG $\mathcal{G}(\{W_0, M\}, \{A\})$ corresponding to the situation where $W_0, M$ are random variables and $A$ is fixed to a constant. Just as a distribution may be associated with a DAG via factorization, so may a kernel be associated with a CADMG in a particular way Richardson et al. (2017). The CADMG in Fig. 2-3 (d) may be associated with $p(W_0, M|\text{do}(a)) = p(M|a, W_0)p(W_0)$. Genealogic definitions, such as $\text{pa}_{\mathcal{G}}(.)$, carry over identically to CADMGs. Districts in a CADMG are defined as subsets of $\mathbf{V}$.

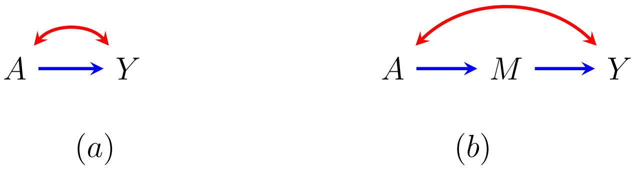


Figure 2-4. a) an ADMG in which $Y(a)$ is known to be non-identifiable; known as the 'bow-arc' graph. b) an ADMG in which, despite $A$ and $Y$ sharing a common cause, is known to yield an identifiable $Y(a)$; known as the 'front-door' graph.

### 2.5.2 The Fixing Operator

Given a CADMG $\mathcal{G}(\mathbf{V}, \mathbf{W})$, a variable $V \in \mathbf{V}$ is *fixable* if $\text{de}_{\mathcal{G}}(V) \cap \text{dis}_{\mathcal{G}}(V) = \emptyset$. Put differently, $V$ is fixable if paths $V \leftrightarrow \cdots \leftrightarrow B$ and $V \to \cdots \to B$ do not *both* exist in $\mathcal{G}$ for any $B \in \mathbf{V} \setminus V$. For example, in Fig. 2-3 (b), $M$ is fixable, while $W_0$ is not. Intuitively, $V$ is fixable in a CADMG $\mathcal{G}(\mathbf{V}, \mathbf{W})$ if, in a causal graph representing a hypothetical situation $p(\mathbf{V}|\text{do}(\mathbf{w}))$, where variables in $\mathbf{W}$ were already intervened on, $p(\mathbf{V} \setminus \{V\}|\text{do}(\mathbf{w}, v))$ is identified by the application of the g-formula to $p(\mathbf{V}|\text{do}(\mathbf{w}))$. Whenever a variable $V$ is fixable, a fixing operator may be applied to both the CADMG and the kernel to yield a new causal graph and a new kernel representing the situation where $V$ is also intervened on.

Given $V \in \mathbf{V}$ fixable in a CADMG $\mathcal{G}(\mathbf{V}, \mathbf{W})$, the fixing operator $\phi_V(\mathcal{G})$ yields a new CADMG $\widetilde{\mathcal{G}}(\mathbf{V} \setminus \{V\}, \mathbf{W} \cup \{V\})$, where all vertices and edges in $\mathcal{G}(\mathbf{V}, \mathbf{W})$ are kept, *except* $V$ is viewed as fixed, and all edges with arrowheads into $V$ are removed. Given $V \in \mathbf{V}$ fixable in a CADMG $\mathcal{G}(\mathbf{V}, \mathbf{W})$, and a kernel $q_{\mathbf{V}}(\mathbf{V}|\mathbf{W})$ associated with $\mathcal{G}$, the fixing operator $\phi_V(q_{\mathbf{V}}; \mathcal{G})$ yields a new kernel

$$\tilde{q}_{\mathbf{V}\setminus\{V\}}(\mathbf{V} \setminus \{V\}|\mathbf{W} \cup \{V\}) \equiv \frac{q_{\mathbf{V}}(\mathbf{V}|\mathbf{W})}{q_{\mathbf{V}}(V|\mathbf{W} \cup \text{nd}_{\mathcal{G}}(V))},$$

where the denominator is defined as above by marginalization and conditioning within the kernel $q_{\mathbf{V}}$. If $\text{ch}_{\mathcal{G}}(V) = \emptyset$, division by $q_{\mathbf{V}}(V|\,\text{nd}_{\mathcal{G}}(V))$ is equivalent to marginalizing $V$ from $q_{\mathbf{V}}$. In this way, the fixing operator unifies applications of the g-formula in lines 6 and 7 of the Shpitser and Pearl (2006) formulation of the ID Algorithm (see

Appendix C) and marginalization of irrelevant variables in line 2 of the algorithm. The recursive operation of the ID algorithm can be expressed concisely as repeated invocations of the operator. This allows us to concisely express functionals returned by ID algorithm as a one line formula. Many of the novel theoretical contributions in this thesis are extensions of this one line algorithm: to new intervention types, more general graphical models, or both.

### 2.5.3 The Nested Factorization

A set $\mathbf{V}^\dagger \subseteq \mathbf{V}$ is said to be *fixable* in a latent projection $\mathcal{G}(\mathbf{V})$ if there exists a *valid* sequence $\sigma_{\mathbf{V}^\dagger} = \langle V_1, V_2, \ldots, V_k \rangle$ of variables in $\mathbf{V}^\dagger$ such that $V_1$ is fixable in $\mathcal{G}$, $V_2$ is fixable in $\phi_{V_1}(\mathcal{G})$, and so on. If $\mathbf{V}^\dagger$ is fixable, $\mathbf{V} \setminus \mathbf{V}^\dagger$ is called a *reachable* set. A reachable set $\mathbf{V}$ is said to be *intrinsic* if $\mathcal{G}_{\mathbf{V}}$ has a single district. Define $\phi_{\sigma_{\mathbf{V}^\dagger}}(\mathcal{G})$ and $\phi_{\sigma_{\mathbf{V}^\dagger}}(q; \mathcal{G})$ via function composition to be the operators that fix all elements in $\mathbf{V}^\dagger$ in the order dictated by $\sigma_{\mathbf{V}^\dagger}$.

The distribution $p(\mathbf{V})$ is said to obey the nested factorization for an ADMG $\mathcal{G}$ if there exists a set of kernels $\{q_{\mathbf{C}}(\mathbf{C} \mid \text{pa}_{\mathcal{G}}(\mathbf{C})) \mid \mathbf{C} \text{ is intrinsic in } \mathcal{G}\}$ such that for every fixable $\mathbf{V}^\dagger$, and any valid sequence $\sigma_{\mathbf{V}^\dagger}$,

$$\phi_{\sigma_{\mathbf{V}^\dagger}}(p(\mathbf{V}); \mathcal{G}) = \prod_{\mathbf{D} \in \mathcal{D}(\phi_{\sigma_{\mathbf{V}^\dagger}}(\mathcal{G}))} q_{\mathbf{D}}(\mathbf{D} | \, \text{pa}^s_{\mathcal{G}}(\mathbf{D})).$$

All valid fixing sequences for $\mathbf{V}^\dagger$ yield the same CADMG $\mathcal{G}(\mathbf{V} \setminus \mathbf{V}^\dagger, \mathbf{V}^\dagger)$, and if $p(\mathbf{V})$ obeys the nested factorization for $\mathcal{G}$, all valid fixing sequences for $\mathbf{V}^\dagger$ yield the same kernel (Richardson et al., 2017). As a result, for any valid sequence $\sigma$ for $\mathbf{V}^\dagger$, we can redefine the operator $\phi_\sigma$, for both graphs and kernels, to be $\phi_{\mathbf{V}^\dagger}$, meaning "apply the fixing operator to elements of $\mathbf{V}^\dagger$ in some valid sequence". In addition, it can be shown (Richardson et al., 2017) that the above kernel set is characterized as:

$$\{q_{\mathbf{C}}(\mathbf{C} \mid \text{pa}_{\mathcal{G}}(\mathbf{C})) \mid \mathbf{C} \text{ is intrinsic in } \mathcal{G}\} = \{\phi_{\mathbf{V} \setminus \mathbf{C}}(p(\mathbf{V}); \mathcal{G}) \mid \mathbf{C} \text{ is intrinsic in } \mathcal{G}\}.$$

Thus, we can re-express the above nested factorization as stating that for any fixable set $\mathbf{V}^{\dagger}$, we have

$$\phi_{\mathbf{V}^{\dagger}}(p(\mathbf{V}); \mathcal{G}) = \prod_{\mathbf{D} \in \mathcal{D}(\phi_{\mathbf{V}^{\dagger}}(\mathcal{G}))} \phi_{\mathbf{V} \setminus \mathbf{D}}(p(\mathbf{V}); \mathcal{G}).$$

Since fixing is defined on CADMGs and kernels, the definition of nested Markov models generalizes in a straightforward way to a kernel $q(\mathbf{V}|\mathbf{W})$ being in the nested Markov model for a CADMG $\mathcal{G}(\mathbf{V}, \mathbf{W})$. This holds if for every $\mathbf{V}^{\dagger}$ fixable in $\mathcal{G}(\mathbf{V}, \mathbf{W})$,

$$\phi_{\mathbf{V}^{\dagger}}(q(\mathbf{V}|\mathbf{W}); \mathcal{G}) = \prod_{\mathbf{D} \in \mathcal{D}(\phi_{\mathbf{S}}(\mathcal{G}))} \phi_{\mathbf{V} \setminus \mathbf{D}}(q(\mathbf{V}|\mathbf{W}); \mathcal{G}).$$

An important result in Richardson et al. (2017) states that if $p(\mathbf{V} \cup \mathbf{H})$ obeys the factorization for a DAG $\mathcal{G}$ with vertex set $\mathbf{V} \cup \mathbf{H}$, then $p(\mathbf{V})$ obeys the nested factorization for the latent projection ADMG $\mathcal{G}(\mathbf{V})$.

### 2.5.4 The One Line ID Algorithm

This leads to the one line ID algorithm proposed in Richardson et al. (2017) as a reformulation of the Shpitser and Pearl (2006) ID algorithm. Let $\mathbf{A} \subseteq \mathbf{V}$ and $\mathbf{Y} \subseteq \mathbf{V} \setminus \mathbf{A}$. Define $\mathbf{Y}^{\star} \equiv \text{an}_{\mathcal{G}_{\mathbf{V} \setminus \mathbf{A}}}(\mathbf{Y})$. Then we have the following sound and complete formula for the identification of $p(\mathbf{Y}(\mathbf{a}))$:

$$p(\mathbf{Y}(\mathbf{a})) = \sum_{\mathbf{Y}^{*} \setminus \mathbf{Y}} \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}_{\mathbf{Y}^{*}})} \phi_{\mathbf{V} \setminus \mathbf{D}}(p(\mathbf{V}); \mathcal{G}(\mathbf{V}))|_{\mathbf{A}=\mathbf{a}}. \tag{2.4}$$

Whenever $\mathbf{V} \setminus \mathbf{D}$ for every $\mathbf{D}$ is fixable, the formula (2.4) yields the correct expression for $p(\mathbf{Y}|\text{do}(\mathbf{a}))$ in terms of the observed data[8]. If some $\mathbf{V} \setminus \mathbf{D}$ is not fixable, the algorithm fails, and $p(\mathbf{Y}|\text{do}(\mathbf{a}))$ is not identified[9]. See Richardson et al. (2017) for a detailed proof.

For instance, consider the example graphs in Figures 2-2 and 2-4. Figure 2-4(a) depicts a 'bow-arc' graph, a classic example of a non-identifiable $Y(a)$ (see above for

[8] This is the soundness property, formally stated. See also: the exposition in §2.4.

[9] This relates to the completeness property. Formally, Shpitser and Pearl (2006) and Richardson et al. (2017) showed that if this algorithm fails, then no algorithm can successfully identify the effect.

an intuitive explanation of why this intervention is not identifiable) (Shpitser and Pearl, 2006).

In Figure 2-4(b), on the other hand, $Y(a)$ *is* identifiable even though $A$ and $Y$ share a common cause. This non-intuitive result was first proven using rules from probability theory by Pearl and the graph in Figure 2-4(b) is known as the 'front-door' graph (Pearl, 2009). The ID algorithm (Eq. 2.4) will verify this results (i.e. that the intervention is identifiable) and enables derivation of an identifying functional, the expression exclusively in terms of observed data: $p(Y(a)) = \sum_M p(M|A = a) \sum_{A'} p(Y|A', M)p(A')$.

As a more complicated example, in Fig. 2-3 (a), the ID algorithm yields the following identifying formula for $p(Y|\text{do}(a))$:

$$\begin{gathered} p(Y(a)) = \sum_{W_0,A,M,W_1} p(W_1|M, A = a, W_0)\times \\ p(M|A = a, W_0)p(W_0) \sum_{W_0,A} p(Y|W_1, M, A, W_0)p(W_0, A). \end{gathered} \tag{2.5}$$

See Appendix C for a complete derivation.

# Chapter 3

# Identification of Personalized Effects Associated With Causal Pathways

## 3.1 Introduction

Establishing causal relationships between actions and outcomes is fundamental to rational decision-making. As described in Chapter 1, the gold standard for establishing causal relationships is the randomized controlled trial (RCT), which may be used to establish *average* causal effects within a population. While average treatment effects reported from RCTs (real or emulated via OCI methods) establish whether a particular action is helpful *on average*, optimal decision-making must tailor decisions to specific situations by, for instance, using ideas from the dynamic treatment regime (Chakraborty and Moodie, 2013) or reinforcement learning (Bertsekas and Tsitsiklis, 1996) literatures.

If an action is known to have a beneficial effect on some outcome, it is often desirable to understand the causal *mechanism* behind this effect. A popular type of mechanism analysis is *mediation analysis*, which seeks to decompose average treatment effects into direct and indirect components, or more generally into components associated with specific causal pathways. These components of the average causal effect are known as direct, indirect, and path-specific effects, and are also defined as population averages (Avin, Shpitser, and Pearl, 2005; Pearl, 2001; Robins and Greenland, 1992).

In this chapter, derived from work originally appearing in Shpitser and Sherman (2018), I first supplement the review from Chapter 2 with background concepts and notation specific to this chapter's developments. I then define counterfactual outcomes necessary to personalize effects associated with causal pathways, give an algorithm for non-parametric identification of these outcomes and prove that it is complete for arbitrary policies. Estimation methods for identified outcomes of this type were studied contemporaneously in Nabi and Shpitser (2018).

### 3.1.1 Why Personalize Effects Along Causal Pathways?

It often makes sense to structure decision-making such that the *overall* effect of an action on the outcome is maximized for a given unit. However, in some cases it is appropriate to choose an action such that only a part of the effect of an action on the outcome is maximized. Consider management of HIV patients' care. Since HIV is a chronic disease, care for HIV patients involves designing a long-term treatment plan to minimize the chance of viral failure (an undesirable outcome). In designing such a plan, an important choice is when to initiate primary therapy, and when to switch to a second line therapy. Initiating or switching too early risks unneeded side effects and ”wasting” treatment efficacy, while initiating or switching too late risks viral failure (Hernan et al., 2006).

In the context of HIV, however, *treatment adherence* is an important component of the overall effect of the drug on the outcome. Patients who do not take prescribed doses compromise the efficacy of the drug, and different drugs may have different levels of adherence. Thus, for HIV patients, the overall effect of the drug can be viewed as a combination of the chemical effect and the adherence effect (Miles et al., 2017). Therefore, choosing an action that maximizes the overall effect of HIV treatment on viral failure entangles these two very different causal mechanisms. One approach to tailoring treatments to patients in a way that disentangles these mechanisms is to find a policy that optimizes a part of the effect, say the chemical (direct) effect of the drug, while hypothetically keeping the adherence levels to some reference level. Finding such a policy yields information on how best to assign drugs to maximize their chemical efficacy in settings where adherence levels can be controlled to that of a reference treatment – even if the only data available is one where patients have differential adherence.

## 3.2 Preliminaries

I will first give graph theoretic preliminaries. Next, I describe the more general *edge intervention* that sets variables to different values for different outgoing edges in a graph. Edge interventions are used to formulate direct, indirect, and path-specific effects in mediation analysis. Then, I define counterfactual responses to policies that set variables not to *constant* values but to values that potentially depend on other sets of variables. Extending these notions, I describe counterfactuals that generalize both responses to edge interventions, and responses to policies, namely responses to *edge-specific policies*. I briefly describe identification theory for these counterfactuals in causal models with no hidden variables, and note this theory is based on variations of a truncated factorization known as the g-formula (Robins, 1986).

I next consider identification theory for these counterfactuals in hidden variable causal models. This theory is more complex, and is based on the ID algorithm (Shpitser and Pearl, 2006; Tian and Pearl, 2002). Using the reformulated ID algorithm described in Chapter 2 as a base, I describe ways to express any functional corresponding to a counterfactual distribution identifiable in a hidden variable causal model as a truncated factorization formula. I posit algorithms for the identification of the above counterfactual types. Finally, I describe a completeness result for the identification algorithm for responses to unrestricted edge-specific policies in hidden variable causal models.

The primary contribution of this chapter lies in the presentation of counterfactuals and identification theory for policies, edge-specific interventions, and edge-specific policies. Nevertheless, given the heavy reliance on prior theory, I will continue to develop the discussion of past work that I began in Chapter 2 as a means of building towards the primary results.

### 3.2.1 Graph Theory

Throughout this chapter, I will rely on causal graphs to build and analyze causal models. See Chapter 2 for a complete introduction to graphical models including definitions and notation for variables, genealogic sets, and subgraphs.

### 3.2.2 Edge Interventions

Recall from Chapter 2, that we can perform *node* interventions in graphs. A more general type of intervention in a graphical causal model is the *edge intervention* (Shpitser and Tchetgen Tchetgen, 2016), which maps a set of directed edges in $\mathcal{G}$ to values of their source vertices. Edge interventions have a natural interpretation in cases where a treatment variable has multiple components that a) influence the outcome in different ways, b) occur or do not occur together in observed data, and c) may in principle be intervened on separately. For instance, smoking leads to poor health outcomes due to two components: smoke inhalation and exposure to nicotine. A smoker would be exposed to both of these components, while a non-smoker to neither. However, one might imagine exposing someone selectively only to nicotine but not smoke inhalation (via a nicotine patch), or only smoke inhalation but not nicotine (via smoking plant matter not derived from tobacco leaves). These types of hypothetical experiments correspond precisely to edge interventions, and have been used to conceptualize direct and indirect effects (Pearl, 2001; Robins and Greenland, 1992), often on the mean difference scale.

Formally, I will write the mapping of a set of edges to values of their source vertices using the following shorthand: $(a_1 W_1)_{\to}, (a_2 W_2)_{\to}, \ldots, (a_k W_k)_{\to}$ to mean that edge $(A_1 W_1)_{\to}$ is assigned to value $a_1$, $(A_2 W_2)_{\to}$ is assigned to value $a_2$, and so on until $(A_k W_k)_{\to}$ is assigned to value $a_k$. Alternatively, I will write $\mathfrak{a}_\alpha$ to mean edges in $\alpha$ are mapped to values in the *multiset* $\mathfrak{a}$ (since multiple edges may share the same source vertex, and be assigned to different values). For a subset $\beta \subseteq \alpha$, and an assignment $\mathfrak{a}_\alpha$ denote $\mathfrak{a}_\beta$ to be a restriction of $\mathfrak{a}_\alpha$ to edges in $\beta$.

I will write counterfactual responses to edge interventions as $Y(\mathfrak{a}_\alpha)$ or, for simple cases, as: $Y((aY)_{\to}, (a'M)_{\to})$ meaning the response to $Y$ where $A$ is set to value $a$ for the purposes of the edge $(AY)_{\to}$ and to $a'$ for the purposes of the edge $(AM)_{\to}$. An edge intervention that sets a set of edges $\alpha$ to values in the multiset $\mathfrak{a}$ is defined via the following generalization of recursive substitution (Eq. 2.2):

$$Y(\mathfrak{a}_\alpha) \equiv Y(\mathfrak{a}_{\{(ZY)_{\to} \in \alpha\}}, \{\text{pa}_{\mathcal{G}}^{\bar{\alpha}}(Y)\}(\mathfrak{a}_\alpha)), \tag{3.1}$$

where $\text{pa}_{\mathcal{G}}^{\bar{\alpha}}(Y) \equiv \{W \mid (WY)_{\to} \notin \alpha\}$. For example, in the DAG in Fig. 3-1 (a), $Y((a'Y)_{\to}, (aM)_{\to})$ is defined as $Y(a', M(a, W), W)$.

For simplicity of presentation, I will restrict attention to edge interventions with the property that if $(AW)_{\to} \in \alpha$, then for any $V \in \text{ch}_{\mathcal{G}}(A)$, $(AV)_{\to} \in \alpha$. These types of edge interventions set values for all causal pathways for a set of treatment variables. This is the convention in the majority of existing mediation literature as these interventions are most relevant in practical mediation analysis problems. Specifically, in the HIV example, we are interested in the effect of a drug along all pathways that start with a particular edge, while the effect of the drug via pathways that begin with other edges is kept to a reference level. This assumption may be relaxed, at the price of complicating the theory (Shpitser and Tchetgen Tchetgen, 2016).

Edge interventions are used to define direct and indirect effects. For example, in the model given by the DAG in Fig 3-1 (a), the direct effect of $A$ on $Y$ is defined as $\mathbb{E}[Y((aY)_{\rightarrow}, (aM)_{\rightarrow})] - \mathbb{E}[Y((a'Y)_{\rightarrow}, (aM)_{\rightarrow})]$ which is equal to $\mathbb{E}[Y(a)] - \mathbb{E}[Y(a', M(a))]$. The indirect effect may be defined similarly as $\mathbb{E}[Y((a'Y)_{\rightarrow}, (aM)_{\rightarrow})] - \mathbb{E}[Y((a'Y)_{\rightarrow}, (a'M)_{\rightarrow})]$, which is equal to $\mathbb{E}[Y(a', M(a))] - \mathbb{E}[Y(a')]$. The direct and indirect effects add up to the ACE.

Note that while direct, indirect, and path-specific effects may be defined directly as nested counterfactuals (Pearl, 2001; Shpitser, 2013), this notation quickly becomes unreadable for complicated interventions applied at multiple time points. The edge intervention notation may be viewed as a generalization of the do(.) operator notation of Pearl to mediation problems, which avoids having to specify the entire nested counterfactual, and instead directly ties interventions and sets of causal pathways to which these interventions apply (as represented by the first edge shared by all pathways in the set).

Identification of edge interventions in graphical causal models without hidden variables corresponds quite closely with identification of regular (node) interventions, as follows. Let $\mathbf{A}_\alpha \equiv \{A \mid (AB)_{\rightarrow} \in \alpha\}$. Consider an edge intervention given by the mapping $\mathfrak{a}_\alpha$. Then, under the functional model of a DAG $\mathcal{G}$, the joint distribution of counterfactual responses $p(\{\mathbf{V} \setminus \mathbf{A}_\alpha\}(\mathfrak{a}_\alpha))$ is identified via the following generalization of (2.3) called the *edge g-formula*:

$$\prod_{V \in \mathbf{V} \setminus \mathbf{A}_\alpha} p(V | \mathfrak{a}_{\{(ZV)_{\rightarrow} \in \alpha\}}, \text{pa}^{\bar{\alpha}}_{\mathcal{G}}(V)). \tag{3.2}$$

For example, in Fig 3-1 (a), $p(Y((aY)_{\rightarrow}, (a'M)_{\rightarrow})) = \sum_{W,M} p(Y|a, M, W)p(M|a', W)p(W)$, which is obtained by marginalizing $W, M$ from the edge g-formula.

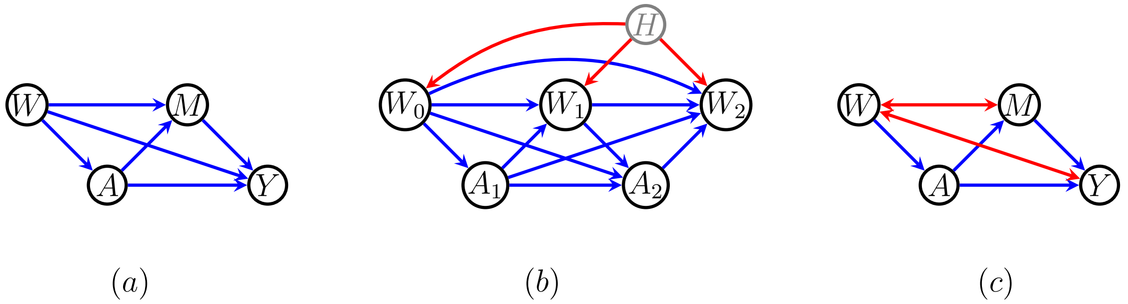
Edge interventions represent a special case of the more general notion of a *path intervention* (Shpitser and Tchetgen Tchetgen, 2016). Responses to both of these interventions are used to define *path-specific effects* (Pearl, 2001), however responses to edge interventions are precisely those that are always identified under the functional model of a DAG, via (3.1). Responses to path interventions that cannot be rephrased as responses to edge interventions are not identified even in a DAG model, including the functional model, due to the presence of *recanting witnesses* (Avin, Shpitser, and Pearl, 2005). For this reason, in this chapter I restrict attention only to edge interventions and responses to edge-specific policies.

### 3.2.3 Responses To Treatment Policies

In personalized medicine settings, counterfactual responses to conditional interventions that set treatment values in response to other variables via a known function are of interest. As an example, assume the graph in Fig. 3-1 (b) represents an observational study of cancer patients where $W_0$ represents baseline patient metrics, $A_1$ is the primary therapy, $W_1$ is the measured intermediate response to the primary therapy, $A_2$ is a decision to either continue primary therapy or switch to a secondary therapy in the event of a poor response to $A_1$, and $W_2$ is the outcome of interest. In this setting, we might be interested in evaluating policies in the set $\{f_{A_1} : \mathfrak{X}_{W_0} \mapsto \mathfrak{X}_{A_1}, f_{A_2} : \mathfrak{X}_{\{W_0,W_1\}} \mapsto \mathfrak{X}_{A_2}\}$ that map patient characteristics to decisions about therapies $A_1$ and $A_2$. We evaluate the efficacy of these policies via the counterfactual variable $W_2(f_{A_1}, f_{A_2})$, representing patient outcomes had treatment decisions been made according to those policies.

These types of variables are defined via a generalization of (2.2), where instead of setting values of parents in $A_1, A_2$ to values fixed by the intervention, values of parents in $A$ are instead set according to $f_{A_1}$ and $f_{A_2}$. In particular, $W_2(f_{A_1}, f_{A_2})$ is defined as

$$W_2[f_{A_2}(W_1[f_{A_1}(W_0), W_0], W_0), W_1[f_{A_1}(W_0), W_0], f_{A_1}(W_0), W_0]. \tag{3.3}$$



$(a)$ $(b)$ $(c)$

Figure 3-1. (a) A simple causal DAG, with a treatment $A$, an outcome $Y$, a vector $W$ of baseline variables, and a mediator $M$. (b) A more complex causal DAG with two treatments $A_1, A_2$, an intermediate outcome $W_1$, and the final outcome $W_2$. $H$ is a hidden common cause of the $W$ variables. (c) A graph where $p(Y(a, M(a')))$ is identified, but $p(Y(f_A(W), M(a)))$ is not.

The distribution of this variable is identified under the functional model via the natural generalization of (2.3) as

$$\begin{aligned}\sum_{W_0, W_1} & p(W_2 | W_0, f_{A_1}(W_0), W_1, f_{A_2}(W_0, W_1)) \times \\ & \qquad p(W_1 | W_0, f_{A_1}(W_0)) p(W_0). \end{aligned} \tag{3.4}$$

More generally, given a DAG $\mathcal{G}$, a topological ordering $\prec$, and a set $\mathbf{A} \subseteq \mathbf{V}$, for each $A \in \mathbf{A}$, define $\mathbf{W}_A$ to be some subset of predecessors of $A$ according to $\prec$. Then, given a set of functions $\mathbf{f}_{\mathbf{A}}$ of the form $f_A : \mathfrak{X}_{\mathbf{W}_A} \mapsto \mathfrak{X}_A$, define $Y(\mathbf{f}_{\mathbf{A}})$, the counterfactual response $Y \in \mathbf{V}$ to $\mathbf{A}$ being intervened on via $\mathbf{f}_{\mathbf{A}} \equiv \{f_A \mid A \in \mathbf{A}\}$, as

$$Y(\{f_A(\mathbf{W}_A(\mathbf{f}_{\mathbf{A}})) | A \in \text{pa}_{\mathcal{G}}(Y) \cap \mathbf{A}\}, \{\text{pa}_{\mathcal{G}}(Y) \setminus \mathbf{A}\}(\mathbf{f}_{\mathbf{A}})). \tag{3.5}$$

In a functional model of a DAG $\mathcal{G}$, the effect of $\mathbf{f}_{\mathbf{A}}$ on the set of variables not being intervened upon, $\mathbf{V} \setminus \mathbf{A}$, represented by the distribution $p(\{\mathbf{V} \setminus \mathbf{A})\}(\mathbf{f}_{\mathbf{A}}))$, is identified by the following modification of (2.3) Tian (2008):

$$\prod_{V \in \mathbf{V} \setminus \mathbf{A}} p(V | \{f_A(\mathbf{W}_A) | A \in \mathbf{A} \cap \text{pa}_{\mathcal{G}}(V)\}, \text{pa}_{\mathcal{G}}(V) \setminus \mathbf{A}). \tag{3.6}$$

## 3.3 Edge-Specific Policies

I now give a general definition of counterfactual responses to edge-specific policies that generalize both responses to edge interventions (where a variable is set to different constants for different outgoing edges) and responses to policies, where a variable is set according to a single known function for all causal pathways at once.

As an example, we can view Fig. 3-1 (a) as representing a cross-sectional study of HIV patients of the kind described in Miles et al. (2017), where $W$ is a set of baseline characteristics, $A$ is one of a set of possible antiretroviral treatments, $M$ is adherence to treatment, and $Y$ is a binary outcome variable signifying viral failure. In this type of study, we may wish to find $f_A(W)$ that maximizes the expected outcome $Y$ had $A$ been set according to $f_A(W)$ for the purposes of the direct effect of $A$ on $Y$, and $A$ were set to some reference level $a$ for the purposes of the effect of $A$ on $M$. In other words, we may wish to find $f_A(W)$ to maximize the counterfactual mean $\mathbb{E}[Y(f_A(W), M(a, W), W)]$. This would correspond to finding a treatment policy that maximizes the direct (chemical) effect, if it were possible to keep adherence to a level $M(a)$ as if a reference (easy to adhere to) treatment $a$ were given.

I now give a general definition for responses to such edge-specific policies. Fix a set of directed edges $\alpha$, and define $\mathbf{A}_\alpha \equiv \{A \mid (AB)_\rightarrow \in \alpha\}$. As before, I assume if $(AW)_\rightarrow \in \alpha$, then for all $V \in \mathrm{ch}_{\mathcal{G}}(A)$, $(AV)_\rightarrow \in \alpha$. Define $\mathfrak{f}_\alpha \equiv \{f_A^{(AW)_\rightarrow} : \mathfrak{X}_{\mathbf{W}_A} \mapsto \mathfrak{X}_A \mid (AW)_\rightarrow \in \alpha\}$ as the set of policies associated with edges in $\alpha$. Note that $\mathfrak{f}_\alpha$ may contain multiple policies for a given treatment variable $A$.

Define $Y(\mathfrak{f}_\alpha)$, the counterfactual response of $Y$ to the set of edge-specific policies $\mathfrak{f}_\alpha$, as the following generalization of (3.1) and (3.5):

$$Y(\{f_A^{(AY)_\rightarrow}(\mathbf{W}_A(\mathfrak{f}_\alpha)) | (AY)_\rightarrow \in \alpha\}, \{\mathrm{pa}_{\mathcal{G}}^{\bar{\alpha}}(Y)\}(\mathfrak{f}_\alpha)). \tag{3.7}$$

In my earlier example, if $\mathfrak{f}_{\{(AY)_\rightarrow,(AM)_\rightarrow\}} \equiv \{f_A^{(AY)_\rightarrow}(W), \tilde{f}_A^{(AM)_\rightarrow}\}$, where $\tilde{f}_A$ assigns $A$ to a constant value $a$, then $Y(\mathfrak{f}_{\{(AY)_\rightarrow,(AM)_\rightarrow\}}) \equiv Y(f_A(W), M(a, W), W)$.

The joint counterfactual distribution for responses to edge-specific policies, $p(\{V(\mathfrak{f}_\alpha)|V \in \mathbf{V} \setminus \mathbf{A}_\alpha\})$, is identified under the functional model, and generalizes (3.2) and (3.4) as follows:

$$\prod_{V \in \mathbf{V} \setminus \mathbf{A}_\alpha} p(V|\{f_A^{(AV)_\to}(\mathbf{W}_A)|(AV)_\to \in \alpha\}, \text{pa}_{\mathcal{G}}^{\bar{\alpha}}(V)). \tag{3.8}$$

This is a consequence of the fact that (3.2) holds regardless of how edge interventions are set. In Fig. 3-1 (a), for example,

$$p(Y(f_A(W), M(a, W), W)) = \sum_{W,M} p(Y|f_A(W), M, W)p(M|a, W)p(W)$$

## 3.4 Identification in Hidden Variable DAG Models

As highlighted in Chapter 2, in a causal model of a DAG where some variables are hidden, not every causal parameter is a function of the observed data distribution. In that chapter's review, I highlighted the background necessary to develop a general identification algorithm for *node* interventions in hidden variable DAGs (Tian and Pearl, 2002; Shpitser and Pearl, 2006; Huang and Valtorta, 2006; Richardson et al., 2017). In this section, I will extend the ideas introduced in Chapter 2 to edge, policy, and edge-specific policy interventions. Important concepts from Chapter 2 include latent projections and bidirected edges, Markov kernels, ADMGs and Conditional ADMGs, the fixing operator $\phi$, and the one line ID algorithm (Eq. 2.4).

### 3.4.1 Reformulations of Generalized ID Algorithms

In the previous chapter, I described a reformulation of the ID algorithm, proposed in Richardson et al. (2017), as a one line formula. This formula can be viewed as a modified g-formula or truncated factorization. In this subsection I will describe how existing identification theory for edge and policy effects can similarly be reformulated as one line algorithms.

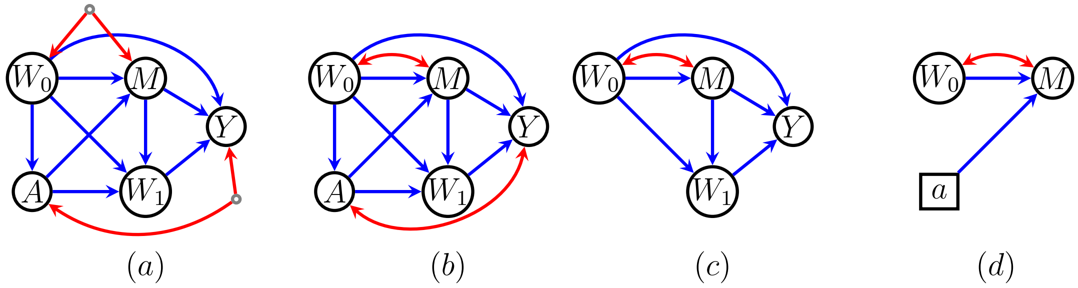


Figure 3-2. (a) A causal model with a treatment $A$ and outcome $Y$. (b) A latent projection of the DAG in (a). (c) The graph derived from (b) corresponding to $\mathcal{G}_{\mathbf{Y}^*} = \mathcal{G}_{\{Y,M,W_0,W_1\}}$. (d) A CADMG corresponding to $p(M, W_0|\text{do}(a))$.

**Edge Interventions** Identification of path-specific effects where each path is associated with one of two possible value sets $\mathbf{a}, \mathbf{a}'$ was given a general characterization in Shpitser (2013) via the *recanting district criterion*. Here, I reformulate this result in terms of the fixing operator in a way that generalizes (2.4), and applies to the response of any edge intervention, including those that set edges to multiple values rather than two. This result can also be viewed as a generalization of *node consistency* of edge interventions in DAG models, found in Shpitser and Tchetgen Tchetgen (2016).

Given $\mathbf{A}_\alpha \equiv \{A \mid (AB)_\to \in \alpha\}$, and an edge intervention given by the mapping $\mathfrak{a}_\alpha$, define $\mathbf{Y}^* \equiv \text{an}_{\mathcal{G}_{\mathbf{V}\setminus\mathbf{A}_\alpha}}(\mathbf{Y})$. The joint distribution of the counterfactual response $p(\{\mathbf{V} \setminus \mathbf{A}_\alpha\}(\mathfrak{a}_\alpha))$ is identified if $p(\{\mathbf{V} \setminus \mathbf{A}_\alpha\}(\mathbf{a}))$ is identified via (2.4), and for every $\mathbf{D} \in \mathcal{D}(\mathcal{G}_{\mathbf{Y}^*})$, for every $A \in \mathbf{A}_\alpha$, $\mathfrak{a}_\alpha$ has the same value assignment for every directed edge out of $A$ into $\mathbf{D}$. Under these assumptions, we have the following result.

**Theorem 1** *$p(\mathbf{Y}(\mathfrak{a}_\alpha))$ is identified and equal to*

$$\sum_{\mathbf{Y}^*\setminus\mathbf{Y}} \prod_{\mathbf{D}\in\mathcal{D}(\mathcal{G}_{\mathbf{Y}^*})} \phi_{\mathbf{V}\setminus\mathbf{D}}(p(\mathbf{V});\mathcal{G})\Big|_{\mathfrak{a}_{\{(AD)_\to\in\alpha|D\in\mathbf{D},A\in\mathbf{A}_\alpha\}}} \tag{3.9}$$

*Proof:* This follows directly from results in Shpitser (2013) and Richardson et al. (2017). Identifying edge interventions entails identifying $\prod_{\mathbf{D}\in\mathcal{D}(\mathcal{G}_{\mathbf{Y}^*})} p(\mathbf{D}|\text{do}(\mathbf{a}_\mathbf{D}))$, where $\mathbf{a}_\mathbf{D}$ is an assignment for $\text{pa}^s_\mathcal{G}(\mathbf{D})$, and $\mathbf{a}_\mathbf{D}$ possibly assigns different values to elements of $\mathbf{A}$

with respect to different districts. The fact that this identification algorithm can be rephrased as (3.9) follows directly by Theorem 60 in Richardson et al. (2017). □

Consider the example in Fig. 3-2 (a)[1]. Assume we set $A = a$ for the edge $(AM)_{\rightarrow}$ and $A = a'$ for the edge $(AW_1)_{\rightarrow}$. The identifying functional for $p(Y((aW_1)_{\rightarrow}, (a'M)_{\rightarrow}))$ has a form nearly identical to that in Eq. 2.5, which was the identifying functional for $p(Y(a))$ in this graph. In this functional, however, some terms are evaluated at $A = a$, and some at $A = a'$:

$$\begin{aligned} \sum_{W_0, A, M, W_1} \Big[ & p(W_1 | M, A = a, W_0) \\ & \times p(M | A = a', W_0) p(W_0) \\ & \times \Big[ \sum_{W_0, A} p(Y | W_0, A, M, W_1) p(W_0, A) \Big] \Big] \end{aligned} \tag{3.10}$$

### Policy Interventions (Dynamic Treatment Regimes)

A general algorithm for identification of responses to a set of policies $\mathbf{f}_{\mathbf{A}}$ was given in Tian (2008). I now reformulate that algorithm in terms of the fixing operator. Define a graph $\mathcal{G}_{\mathbf{f}_{\mathbf{A}}}$ to be a graph obtained from $\mathcal{G}$ by removing all edges into $\mathbf{A}$, and adding for any $A \in \mathbf{A}$, directed edges from $\mathbf{W}_A$ to $A$. By definition of $\mathbf{W}_A$, $\mathcal{G}_{\mathbf{f}_{\mathbf{A}}}$ is guaranteed to be acyclic. Define $\mathbf{Y}^* \equiv \text{an}_{\mathcal{G}_{\mathbf{f}_{\mathbf{A}}}}(\mathbf{Y}) \setminus \mathbf{A}$. Assume $p(\mathbf{Y}^*(\mathbf{a}))$ is identified in $\mathcal{G}$. Then, under the above assumptions, we have the following result.

**Theorem 2** *$p(\mathbf{Y}(\mathbf{f}_{\mathbf{A}}))$ is identified in $\mathcal{G}$. Moreover, the identification formula is*

$$\sum_{(\mathbf{Y}^* \cup \mathbf{A}) \setminus \mathbf{Y}} \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}_{\mathbf{Y}^*})} \phi_{\mathbf{V} \setminus \mathbf{D}}(p(\mathbf{V}); \mathcal{G}) \Big|_{\tilde{\mathbf{a}}_{\text{pa}^{\mathbf{s}}_{\mathcal{G}}(\mathbf{D}) \cap \mathbf{A}}} \tag{3.11}$$

*where $\tilde{\mathbf{a}}_{\text{pa}^{\mathbf{s}}_{\mathcal{G}}(\mathbf{D}) \cap \mathbf{A}}$ is defined as*

$$\begin{cases} \{A = f_A(\mathbf{W}_A) \mid A \in \text{pa}_{\mathcal{G}}(\mathbf{D}) \cap \mathbf{A}\} & \text{pa}_{\mathcal{G}}(\mathbf{D}) \cap \mathbf{A} \neq \emptyset \\ \emptyset & \textit{otherwise} \end{cases}$$

[1] This graph is identical to that in Figure 2-3. It is re-printed here for the reader's convenience.

*Proof:* This follows from the fact that identification of $p(\mathbf{Y}(\mathbf{f_A}))$ can be rephrased as identification of $p(\mathbf{Y}^*(\mathbf{a}))$, with values $\mathbf{a}$ set according to $\{\mathbf{W}_A | A \in \mathbf{A}\}$, where all $\mathbf{W}_A$ in the set are subsets of $\mathbf{Y}^*$. Identification of $p(\mathbf{Y}^*(\mathbf{a}))$ may be rephrased as (3.11) follows by Theorem 60 in Richardson et al. (2017). □

The outer sum over $\mathbf{A}$ in (3.11) is vacuous if $\mathbf{f_A}$ is a set of deterministic policies. To illustrate (3.11), in the example in Fig. 3-2 (b), $p(Y(A = f_A(W_0)))$ is identified as

$$\begin{aligned} \sum_{W_0,A,M,W_1} \Big[ & \Big[ p(W_1|M, A{=}f(W_0), W_0) \Big] \\ & \times \Big[ p(M|A{=}f(W_0), W_0) p(W_0) \Big] \\ & \times \Big[ \sum_{W_0,A} p(Y|W_1, M, A, W_0) p(W_0, A) \Big] \Big]. \end{aligned} \tag{3.12}$$

### 3.4.2 Identification Of Edge-Specific Policies

Having reformulated existing identification results on responses to policies (3.11) and responses to edge interventions arising in mediation analysis (3.9) in terms of the fixing operator, I generalize these results for identification of responses to edge-specific policies.

Given $\mathbf{A}_\alpha \equiv \{A | (AB)_\rightarrow \in \alpha\}$, and a set of edge-specific policies given by the set of mappings $\mathfrak{f}_\alpha$, define the graph $\mathcal{G}_{\mathfrak{f}_\alpha}$ to be one where all edges with arrowheads into $\mathbf{A}_\alpha$ are removed, and directed edges from any vertex in $\mathbf{W}_A$ to $A \in \mathbf{A}_\alpha$ added. Fix a set $\mathbf{Y}$ of outcomes of interest, and define $\mathbf{Y}^*$ equal $\mathrm{an}_{\mathcal{G}_{\mathfrak{f}_\alpha}}(\mathbf{Y}) \setminus \mathbf{A}_\alpha$. We have the following result.

**Theorem 3** *$p(\mathbf{Y}(\mathfrak{f}_\alpha))$ is identified if $p(\mathbf{Y}^*(\mathbf{a}))$ is identified, and for every $\mathbf{D} \in \mathcal{D}((\mathcal{G}_{\mathfrak{f}_\alpha})_{\mathbf{Y}^*})$, $\mathfrak{f}_\alpha$ yields the same policy assignment for every edge from $A \in \mathbf{A}_\alpha$ to $\mathbf{D}$. Moreover, the identifying formula is*

$$\sum_{(\mathbf{Y}^* \cup \mathbf{A}_\alpha) \setminus \mathbf{Y}} \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}_{\mathbf{Y}^*})} \phi_{\mathbf{V} \setminus \mathbf{D}}(p(\mathbf{V}); \mathcal{G})|_{\tilde{\mathbf{a}}_{\mathrm{pa}^{\mathbf{s}}_{\mathcal{G}}(\mathbf{D}) \cap \mathbf{A}_\alpha}} \tag{3.13}$$

*where* $\tilde{\mathbf{a}}_{\text{pa}^{\mathbf{s}}_{\mathcal{G}}(\mathbf{D})\cap\mathbf{A}_\alpha}$ *is defined to be* $\{A = f_A(\mathbf{W}_A) \in \mathfrak{f}_\alpha \mid A \in \text{pa}_{\mathcal{G}}(\mathbf{D}) \cap \mathbf{A}_\alpha\}$*, if* $\text{pa}_{\mathcal{G}}(\mathbf{D}) \cap \mathbf{A}_\alpha \neq \emptyset$*, and is defined to be the* $\emptyset$ *otherwise.*

*Proof:* This is a straightforward generalization of the proofs of Theorems 1 and 2. □

Responses to edge-specific policies are identified in strictly fewer cases compared to responses to edge interventions. This is because $\mathbf{Y}^*$ is a larger set in the former case. As an example, consider the graph in Fig. 3-1 (c), where we are interested either in the counterfactual $p(Y(a, M(a')))$, used to define pure direct effects, or the counterfactual $p(Y(f_A(W), M(a')))$.

For the former counterfactual, we have $\mathbf{Y}^* = \{Y, M\}$, and $p(Y(a, M(a')))$ equal to

$$\sum_m \left( \frac{\sum_w p(Y, m|a, w)p(w)}{\sum_w p(m \mid a, w)p(w)} \right) \sum_w p(m \mid a', w)p(w)$$

Observe that for the latter counterfactual, the set $\mathbf{Y}^* = \{Y, M, W\}$ forms a single district in $\mathcal{G}_{\mathbf{Y}^*}$, and the edge-specific policy set $\mathfrak{f}_{\{(AM)_\rightarrow,(AY)_\rightarrow\}}$ sets edges from $A$ to this district to different policies. As a result, Theorem 3 is insufficient to conclude identification.

Generalizations of the example in Fig. 3-1 (b) are the most relevant in practice, as their causal structure corresponds to longitudinal observational studies, of the kind considered in Robins (1986), and many other papers. However, I illustrate complications that may arise in identifiability of responses to edge-specific policies with our running example in Fig. 3-2 (b), where we are interested in the response of $Y$ to edge-specific policies $\mathfrak{f}_{\{(AM)_\rightarrow,(AW_1)_\rightarrow\}} = \{f_A^{(AM)_\rightarrow}(W_0), f_A^{(AW_1)_\rightarrow}(W_0)\}$. Theorem 3 yields the following identifying formula:

$$\begin{aligned}\sum_{W_0,A,M,W_1} \Big[ &\Big[p(W_1|M, A = f_A^{(AM)_\rightarrow}(W_0), W_0)\Big] \\ &\times \Big[p(M|A = f_A^{(AW_1)_\rightarrow}(W_0), W_0)p(W_0)\Big] \\ &\times \Big[\sum_{W_0,A} p(Y|W_1, M, A, W_0)p(W_0, A)\Big]\Big]. \end{aligned} \quad (3.14)$$

Note that (3.14) generalizes both (3.10), which sets $A$ to different constants in different terms, and (3.12), which sets $A$ to the output of a function that depends on $W_0$. I give a detailed derivation of this functional in the appendix (Appendix C).

## 3.5 On Completeness

An identification algorithm for a class of parameters is said to be *complete* relative to a class of causal models if, whenever the algorithm fails to identify a parameter within a model class, the parameter is in fact not identified within that class.

The ID algorithm is known to be complete for the class of interventional distributions in the class of functional models (Huang and Valtorta, 2006; Shpitser and Pearl, 2006). I restate this result here, and give a sequence of increasingly general completeness results for the identification algorithms described so far. Completeness results on policies and edge-specific policies are new. For completeness results pertaining to policies, I assume a completely unrestricted class of policies. If the set of policies of interest, $\mathbf{f}_{\mathbf{A}}$ or $\mathfrak{f}_{\alpha}$ is restricted, or alternatively if the causal model has parametric restrictions, completeness results presented here may no longer hold.

**Theorem 4** *Given disjoint subsets $\mathbf{Y}, \mathbf{A}$ of $\mathbf{V}$ in an ADMG $\mathcal{G}$, define $\mathbf{Y}^* \equiv \text{an}_{\mathcal{G}_{\mathbf{V}\setminus\mathbf{A}}}(\mathbf{Y})$. Then $p(\mathbf{Y}(\mathbf{a}))$ is not identified if there exists $\mathbf{D} \in \mathcal{D}(\mathcal{G}_{\mathbf{Y}^*})$ that is not a reachable set in $\mathcal{G}$.*

**Corollary 1** *The algorithm for identification of $p(\mathbf{Y}(\mathbf{a}))$, as phrased in (2.4), is complete.*

**Theorem 5** *Given $\mathbf{A}_{\alpha} \equiv \{A \mid (AB)_{\rightarrow} \in \alpha\}$, and an edge intervention given by the mapping $\mathfrak{a}_{\alpha}$, define $\mathbf{Y}^* \equiv \text{an}_{\mathcal{G}_{\mathbf{V}\setminus\mathbf{A}_{\alpha}}}(\mathbf{Y})$. The joint distribution of the counterfactual response $p(\{\mathbf{V} \setminus \mathbf{A}_{\alpha}\}(\mathfrak{a}_{\alpha}))$ is not identified if $p(\{\mathbf{V} \setminus \mathbf{A}_{\alpha}\}(\mathbf{a}))$ is not identified, or there exists $\mathbf{D} \in \mathcal{D}(\mathcal{G}_{\mathbf{Y}^*})$ and $A \in \mathbf{A}_{\alpha}$, such that $\mathfrak{a}_{\alpha}$ has the different value assignments for a pair of directed edges out of $A$ into $\mathbf{D}$.*

**Corollary 2** *The algorithm for identification of* $p(\mathbf{Y}(\mathfrak{a}_\alpha))$*, as phrased in (3.9), is complete.*

**Theorem 6** *Define* $\mathcal{G}_{\mathbf{f_A}}$ *to be a graph obtained from* $\mathcal{G}$ *by removing all edges into* $\mathbf{A}$*, and adding for any* $A \in \mathbf{A}$*, directed edges from* $\mathbf{W}_A$ *to* $A$*. Define* $\mathbf{Y}^* \equiv \text{an}_{\mathcal{G}_{\mathbf{f_A}}}(\mathbf{Y}) \setminus \mathbf{A}$*. Then if* $p(\mathbf{Y}^*(\mathbf{a}))$ *is not identified in* $\mathcal{G}$*,* $p(\mathbf{Y}(\mathbf{f_A}))$ *is not identified in* $\mathcal{G}$ *if* $\mathbf{f_A}$ *is the unrestricted class of policies.*

**Corollary 3** *The algorithm for identification of* $p(\mathbf{Y}(\mathbf{f_A}))$*, as phrased in (3.11), is complete for unrestricted policies.*

**Theorem 7** *Define the graph* $\mathcal{G}_{\mathfrak{f}_\alpha}$ *to be one where all edges with arrowheads into* $\mathbf{A}_\alpha$ *are removed, and directed edges from any vertex in* $\mathbf{W}_A$ *to* $A \in \mathbf{A}_\alpha$ *added. Fix a set* $\mathbf{Y}$ *of outcomes of interest, and define* $\mathbf{Y}^*$ *equal* $\text{an}_{\mathcal{G}_{\mathfrak{f}_\alpha}}(\mathbf{Y}) \setminus \mathbf{A}_\alpha$*. Then if* $p(\mathbf{Y}^*(\mathbf{a}))$ *is not identified, or there exists* $\mathbf{D} \in \mathcal{D}((\mathcal{G}_{\mathfrak{f}_\alpha})_{\mathbf{Y}^*})$*, such that* $\mathfrak{f}_\alpha$ *yields different policy assignments for two edges from* $A \in \mathbf{A}_\alpha$ *to* $\mathbf{D}$*,* $p(\mathbf{Y}(\mathfrak{f}_\alpha))$ *is not identified.*

**Corollary 4** *The algorithm for identification of* $p(\mathbf{Y}(\mathfrak{f}_\alpha))$*, as phrased in (3.13), is complete for unrestricted policies.*

Detailed proofs of these results are in Appendix C. Corollaries are immediate consequences of the preceding Theorems.

## 3.6 Conclusion

In this chapter, I defined counterfactual responses to policies that set treatment values in such a way that they affect outcomes with respect to certain causal pathways only. Such counterfactuals arise when we wish to personalize only some portion of the causal effect of a treatment, while keeping other portions set to some reference values.

An example might be optimizing the chemical effect of a drug, while keeping drug adherence to a reference value.

I gave a general algorithm for identifying these responses from data, which generalizes similar algorithms due to Tian (2008) and Shpitser (2013) for dynamic treatment regimes, and edge-specific effects, respectively. Further, I showed that given an unrestricted class of policies the algorithm is complete. As a corollary, this established that the identification algorithm for dynamic treatment regimes in Tian (2008) is complete for unrestricted policies.

Given a fixed set of policies associated with a set of causal pathways, and assuming (3.13) yields a functional containing only conditional densities, as is the case in the functional (3.14), the counterfactual mean under those policies $\mathbb{E}[Y(\mathfrak{f}_\alpha)]$ may be estimated using the maximum likelihood plug-in estimator. Such an estimator can be viewed as a generalization of the parametric g-formula (Robins, 1986) to edge-specific policies. More general estimation strategies, and approaches to learning the optimal set of policies are the subject of the contemporaneous paper Nabi and Shpitser (2018) published as a companion to Shpitser and Sherman (2018).

# Chapter 4

# Identification of the Effects of Node Interventions in Segregated Graph Models

## 4.1 Introduction

As discussed in the introductory Chapter 1, the iid assumption is ubiquitous in statistics and causal inference, but nevertheless does not hold in a variety of situations. In Chapters 2 and 3, I highlighted the fact that even when the iid assumption holds, obtaining valid causal estimates from observational data is challenging due to latent confounding. When we move to the non-iid world, as in social networking and infectious disease settings, identification becomes *more* challenging. As a tangible example, interference (Cox, 1958; Hudgens and Halloran, 2008; Ogburn, VanderWeele, et al., 2014) can induce additional complexities by opening confounding paths that would not be present if we assumed inter-subject independence. If we do not properly account for this network-induced confounding (i.e., by pretending it isn't there), we can obtain arbitrarily biased estimates of causal effects.

Towards accounting for network dependence when performing causal analyses, it is useful to think about how network dynamics impact our ability to perform the identification process, linking target counterfactuals to available data. As described in previous chapters, an extensive literature on identification of causal parameters (under the iid assumption) has been developed. To briefly re-summarize, the *g-formula* (Robins, 1986) identifies any interventional distribution in directed acyclic graph-based (DAG) causal models without latent variables, while a complete identification theory in hidden variable DAG models was developed in Tian and Pearl (2002), Shpitser and Pearl (2006), and Huang and Valtorta (2006).

Beyond identification theory, an extensive theory of estimation of identified causal parameters has also been developed. Some approaches are described in Robins (1986) and Robins, Hernan, and Brumback (2000), although this is far from an exhaustive list. While work on identification and estimation of causal parameters under interference exists (Hudgens and Halloran, 2008; Tchetgen Tchetgen and VanderWeele, 2012; Ogburn, VanderWeele, et al., 2014; Peña, 2018; Peña, 2016; Maier, Marazopoulou, and Jensen, 2013; Arbour, Garant, and Jensen, 2016), no general theory that unifies identification and estimation has been developed up to now. In this chapter, derived from novel work originally published in Sherman and Shpitser (2018), I will describe an identification result that extends the above identification theory to a more general type of graphical model, the latent-variable chain graph Lauritzen, 1996; Lauritzen and Richardson, 2002, which permits a more parsimonious representation of network data. I will also provide a method that answers the question of how to estimate causal effects when both latent variables are present and the iid assumption does not hold.

## 4.2 A Motivating Example

To motivate subsequent developments, I introduce the following example application. Consider a large group of internet users, belonging to a set of online communities, perhaps based on shared hobbies or political views. For each user $i$, their time spent online $A_i$ is influenced by their observed vector of baseline factors $C_i$, and unobserved factors $U_i$. In addition, each user maintains a set of friendship ties with other users via

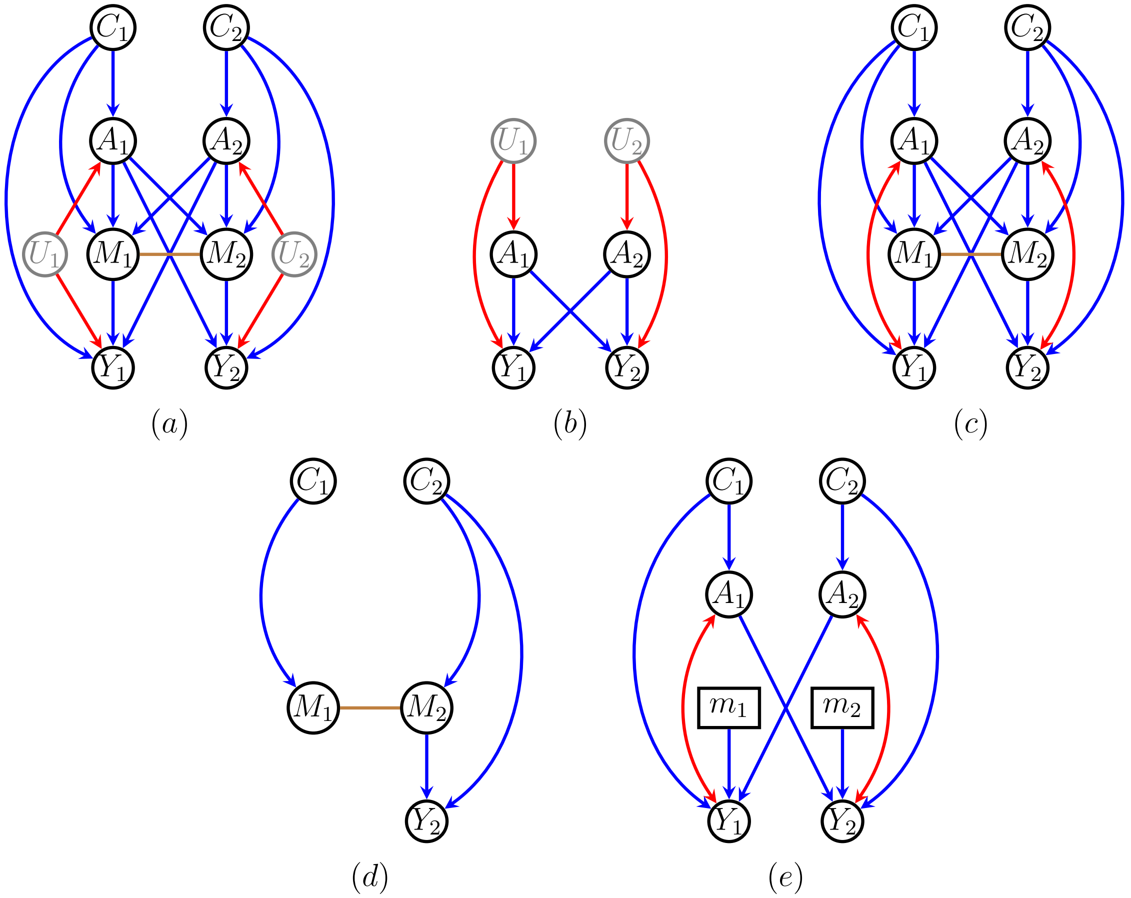


Figure 4-1. (a) A causal model representing the effect of community membership on article sharing, mediated by social network structure. (b) A causal model on dyads which is a variation of causal models of interference considered in Ogburn, VanderWeele, et al. (2014). (c) A latent projection of the CG in (a) onto observed variables. (d) The graph representing $\mathcal{G}_{\mathbf{Y}^*}$ for the intervention operation $\text{do}(a_1)$ applied to (c). (e) The ADMG obtained by fixing $M_1, M_2$ in (c).

an online social network. The user's activity level in the network, $M_i$, is potentially dependent on the user's friends' activities, meaning that for users $j$ and $k$, $M_j$ and $M_k$ are potentially dependent. The dependence between $M$ variables is modeled as a stable symmetric relationship that has reached an equilibrium state. Furthermore, activity level $M_i$ for user $i$ is influenced by observed factors $C_i$, time spent online $A_i$, and the time spent online $A_j$ of any unit $j$ who is a friend of $i$. Finally, denote user $i$'s sharing behavior by $Y_i$. This behavior is influenced by the social network activity of the unit, and possibly the unit friends' time spent online.

A crucial assumption in this example is that for each user $i$, purchasing behavior $Y_i$ is causally influenced by baseline characteristics $C_i$, social network activity $M_i$, and unobserved characteristics $U_i$, but time spent online $A_i$ does not *directly* influence sharing $Y_i$, except as mediated by social network activity of the users. While this might seem like a rather strong assumption, it is more reasonable than standard "front-door" assumptions (Pearl, 2009) in the literature, since we allow the entire social network structure to mediate the influence $A_i$ on $Y_i$ for every user.

We are interested in predicting how a counterfactual change in a set of users' time spent online influences their purchasing behavior. Note that solving this problem from observed data on users as we described is made challenging both by the fact that unobserved variables causally affect both community membership and sharing, creating spurious correlations, and because social network membership introduces dependence among users. In particular, for realistic social networks, every user's activity potentially depends on every other user's activity (even if indirectly). This implies that a part of the data for this problem may effectively consist of a single dependent sample (Tchetgen, Fulcher, and Shpitser, 2017).

In the remainder of the chapter, I formally describe how causal inference may be performed in examples like above, where both unobserved confounding and data dependence are present. In section 4.3 I review relevant terminology and notation that was not described in previous chapters. I also introduce the dependent data setting I will consider. In section 4.4 I describe more general *nested* factorizations (Richardson et al., 2017) applicable to marginals obtained from hidden variable DAG models, and describe identification theory in causal models with hidden variables in terms of a modified nested factorization. In section 4.5, I introduce causal chain graph models (Lauritzen and Richardson, 2002) as a way of modeling causal problems with interference and data dependence, and pose the identification problem for interventional distributions in such models. In section 4.6 I give a sound and complete identification algorithm for interventional distributions in a large class of causal chain graph models with hidden variables, which includes the above example, but also many others. I describe experiments, which illustrate how identified functionals given by our algorithm may be estimated in practice, even in *full interference* settings where all units are mutually dependent, in section 4.7. Concluding remarks are found in section 4.8.

## 4.3 Background on Causal Inference And Interference Problems

### 4.3.1 Graph Theory

As before, I will consider causal models represented by mixed graphs. In addition to directed ($\rightarrow$) and bidirected ($\leftrightarrow$) edges, in this chapter I introduce a new type of edge: the undirected ($-$) edge. Please see Chapter 2 for a thorough review of graphical model concepts. Beyond that review, this chapter requires introduction of the following notation and concepts.

For a mixed graph $\mathcal{G}$ of the above type, the standard graphical sets (e.g., parents) for a variable $V \in \mathbf{V}$ are as previously defined, with the following additions:

$$\text{siblings: } \text{sib}_{\mathcal{G}}(V) \equiv \{W \in \mathbf{V} | W \leftrightarrow V\}$$

$$\text{neighbors: } \text{nb}_{\mathcal{G}}(V) \equiv \{W \in \mathbf{V} | W - V\}$$

Define the *anterior* of $V$, or $\text{ant}_{\mathcal{G}}(V)$, to be the set of all vertices with a partially directed path (a path containing only $\rightarrow$ and $-$ edges such that no $-$ edge can be oriented to induce a directed cycle) into $V$. As with the other relations, these new relations generalize disjunctively to sets.

Consider a mixed graph $\mathcal{G}$. Analogous to districts, defined in Chapter 2, define a *block* $\mathbf{B}$ to be a maximal set of vertices, where every vertex pair in $\mathcal{G}_{\mathbf{B}}$ is connected by an undirected path (a path containing only $-$ edges). Any block of size at least 2 is called a non-trivial block. Define a *maximal clique* as a maximal set of vertices pairwise connected by undirected edges. The set of districts in $\mathcal{G}$ is denoted by $\mathcal{D}(\mathcal{G})$, the set of blocks is denoted by $\mathcal{B}(\mathcal{G})$, the set non-trivial blocks is denoted by $\mathcal{B}^{nt}(\mathcal{G})$, and the set of cliques is denoted by $\mathcal{C}(\mathcal{G})$. The district of $V$ is denoted by $\text{dis}_{\mathcal{G}}(V)$. By convention, for any $V$, $\text{dis}_{\mathcal{G}}(V) \cap \text{de}_{\mathcal{G}}(V) \cap \text{an}_{\mathcal{G}}(V) \cap \text{ant}_{\mathcal{G}}(V) = \{V\}$.

A mixed graph is called *segregated (SG)* if it contains no partially directed cycles, and no vertex has both neighbors and siblings, Fig. 4-1 (c) is an example. In a SG $\mathcal{G}$, $\mathcal{D}(\mathcal{G})$ and $\mathcal{B}^{nt}(\mathcal{G})$ partition $\mathbf{V}$. A SG without bidirected edges is called a chain graph (CG) (Lauritzen, 1996). A SG without undirected edges is called an acyclic directed mixed graph (ADMG) (Richardson, 2003). A CG without undirected edges or an ADMG without bidirected edges is a directed acyclic graph (DAG) (Pearl, 1988). A CG without directed edges is called an undirected graph (UG). Given a CG $\mathcal{G}$, the augmented graph $\mathcal{G}^a$ is the UG where any adjacent vertices in $\mathcal{G}$ or any elements in $\text{pa}_{\mathcal{G}}(\mathbf{B})$ for any $\mathbf{B} \in \mathcal{B}(\mathcal{G})$ are connected by an undirected edge.

### 4.3.2 Graphical Models

Recall from prior chapters that a DAG model is a set of distributions associated with a DAG $\mathcal{G}$ that can be written in terms of a DAG factorization (Lauritzen, 1996):

$$p(\mathbf{V}) = \prod_{V \in \mathbf{V}} p(V \mid \text{pa}_{\mathcal{G}}(V)).$$

Here I also introduce the notion of an undirected graph (UG) model, or a Markov random field (MRF). A UG is a set of distributions associated with a UG $\mathcal{G}$ that can be written in terms of a UG factorization:

$$p(\mathbf{V}) = Z^{-1} \prod_{\mathbf{C} \in \mathcal{C}(\mathcal{G})} \psi_{\mathbf{C}}(\mathbf{C}),$$

where $Z$ is a normalizing constant. A CG model is a set of distributions associated with a CG $\mathcal{G}$ that can be written in terms of the following two level factorization:

$$p(\mathbf{V}) = \prod_{\mathbf{B} \in \mathcal{B}(\mathcal{G})} p(\mathbf{B} | \operatorname{pa}_{\mathcal{G}}(\mathbf{B})),$$

where for each $\mathbf{B} \in \mathcal{B}(\mathcal{G})$,

$$p(\mathbf{B} | \operatorname{pa}_{\mathcal{G}}(\mathbf{B})) = Z(\operatorname{pa}_{\mathcal{G}}(\mathbf{B}))^{-1} \prod_{\mathbf{C} \in \mathcal{C}((\mathcal{G}_{\mathbf{B} \cup \operatorname{pa}_{\mathcal{G}}(\mathbf{B})})^{a}); \mathbf{C} \not\subseteq \operatorname{pa}_{\mathcal{G}}(\mathbf{B})} \psi_{\mathbf{C}}(\mathbf{C}).$$

In words, the above product can be read as follows. Consider the induced subgraph on $\mathbf{B}$ and the parents of $\mathbf{B}$ in $\mathcal{G}$. Then augment (or 'moralize') that graph (Lauritzen, 1996). Then consider all the *cliques* in the graph which are not subsets of the parents of $\mathbf{B}$. There should be a clique potential for each such clique appearing in the product.

Please see Chapter 2 for a description of how to place a causal interpretation on DAG models and for formal definitions of related terms such as counterfactual, identifiability, and the g-formula. An analogous description of how to place a causal interpretation on a chain graph model is provided below.

### 4.3.3 Modeling Dependent Data

So far, the causal and statistical models I have introduced assumed data generating process that produce independent samples. To capture examples of the sort I introduced in section 4.2, I must generalize these models. Suppose we analyze data with $M$ blocks with $N$ units each. It is not necessary to assume that blocks are equally sized for the kinds of problems we consider, but we make this assumption to simplify our notation. Denote the variable $Y$ for the $i$'th unit in block $j$ as $Y_i^j$. For each block $j$, let $\mathbf{Y}^j \equiv (Y_1^j, \ldots, Y_N^j)$, and let $\mathbf{Y} \equiv (\mathbf{Y}^1, \ldots, \mathbf{Y}^M)$. In some cases units' block memberships will not be a primary concern. In these cases I will omit the superscript and the subscript will index the unit with respect to all units in the network.

We are interested in counterfactual responses to interventions on $\mathbf{A}$, treatments on all units in all blocks. For any $\mathbf{a} \in \mathfrak{X}_{\mathbf{A}}$, define $Y_i^j(\mathbf{a})$ to be the potential response of unit $i$ in block $j$ to a hypothetical treatment assignment of $\mathbf{a}$ to $\mathbf{A}$. Define $\mathbf{Y}^j(\mathbf{a})$ and $\mathbf{Y}(\mathbf{a})$ in the natural way as vectors of responses, given a hypothetical treatment assignment to $\mathbf{a}$, either for units in block $j$ or for all units, respectively. Let $\mathbf{a}^{(j)}$ be a vector of values of $\mathbf{A}$, where values assigned to units in block $j$ are *free variables*, and other values are *bound variables*. Furthermore, for any $\mathbf{\tilde{a}^j} \in \mathfrak{X}_{\mathbf{A^j}}$, let $\mathbf{a}^{(j)}[\mathbf{\tilde{a}^j}]$ be a vector of values which agrees on all bound values with $\mathbf{a}^{(j)}$, but which assigns $\mathbf{\tilde{a}^j}$ to all units in block $j$ (e.g. which binds free variables in $\mathbf{a}^{(j)}$ to $\mathbf{\tilde{a}^j}$).

A common assumption is *interblock non-interference*, also known as *partial interference* in Sobel (2006) and Tchetgen Tchetgen and VanderWeele (2012), where for any block $j$, treatments assigned to units in a block other than $j$ do not affect the responses of any unit in block $j$. Formally, this is stated as $(\forall j, \mathbf{a}^{(j)}, \mathbf{a}'^{(j)}, \mathbf{\tilde{a}^j}), \mathbf{Y^j}(\mathbf{a^{(j)}}[\mathbf{\tilde{a}^j}]) = \mathbf{Y^j}(\mathbf{a'^{(j)}}[\mathbf{\tilde{a}^j}])$. Counterfactuals under this assumption are written in a way that emphasizes they only depend on treatments assigned within that block. That is, for any $\mathbf{a}^{(j)}$, $\mathbf{Y}^j(\mathbf{a}^{(j)}[\mathbf{\tilde{a}^j}]) \equiv \mathbf{Y^j}(\mathbf{\tilde{a}^j})$.

In this chapter and the ensuing chapters, I will largely follow the convention of Ogburn, VanderWeele, et al. (2014), where variables corresponding to distinct units within a block are shown as distinct vertices in a graph. As an example, Fig. 4-1 (b) represents a causal model with observed data on multiple realizations of *dyads* or blocks of two dependent units (Kenny, Kashy, and Cook, 2020). Note that the arrow from $A_2$ to $Y_1$ in this model indicates that the treatment of unit 2 in a block influences the outcome of unit 1, and similarly for treatment of unit 1 and outcome of unit 2. In this model, a variation of models considered in Ogburn, VanderWeele, et al. (2014), the interventional distributions $p(Y_2|\text{do}(a_1)) = p(Y_2|a_1)$ and $p(Y_1|\text{do}(a_2)) = p(Y_1|a_2)$ even if $U_1, U_2$ are unobserved.

## 4.4 Causal Inference with Hidden Variables

If a causal model contains hidden variables, only data on the observed marginal distribution is available. As discussed in previous chapters, when hidden variables are present, not every interventional distribution is identified, and identification theory becomes more complex. In the original version of this section (appearing in Sherman and Shpitser (2018)), we reviewed hidden variable DAG identification theory. That review has been largely removed here to avoid a redundant presentation. I again refer the reader to Chapter 2 for a full review.

### 4.4.1 Latent Projection ADMGs

Recall that we can perform a latent projection operation on a latent variable DAG $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$ to obtain an ADMG on the observed margin $\mathcal{G}(\mathbf{V})$. I will prove later that an analogous operation can be performed on latent variable chain graphs to obtain a latent projection on the chain graph's observed margin. As an example, the graph in Fig. 4-1 (c) is the latent projection of Fig. 4-1 (a). Note that a variable pair in a latent projection $\mathcal{G}(\mathbf{V})$ may be connected by both a directed and a bidirected edge, and that multiple distinct hidden variable DAGs or chain graphs $\mathcal{G}_1(\mathbf{V} \cup \mathbf{H}_1)$ and $\mathcal{G}_2(\mathbf{V} \cup \mathbf{H}_2)$ may share the same latent projection acyclic mixed graph.

## 4.5 Chain Graphs For Causal Inference With Dependent Data

I will now generalize causal models to represent settings with data dependence, specifically to cases where variables may exhibit stable but symmetric relationships. These may correspond to friendship ties in a social network, physical proximity, or rules of infectious disease spread. These stand in contrast to causal relationships which are also stable, but asymmetric. I represent settings with both of these kinds of relationships using causal CG models under the Lauritzen-Wermuth-Freydenburg (LWF) interpretation. Though there are alternative conceptions of chain graphs (Drton, 2009), we concentrate on LWF CGs here. This is because LWF CGs yield observed data distributions with smooth parameterizations. In addition, LWF CGs yield Markov properties where each unit's friends (and direct causes) screen the unit from other units in the network. This sort of independence is intuitively appealing in many network settings. Extensions of our results to other CG models are likely possible, but we leave them to future work.

LWF CGs were given a causal interpretation in Lauritzen and Richardson (2002). In a causal CG, the distribution $p(\mathbf{B} | \operatorname{pa}_{\mathcal{G}}(\mathbf{B}))$ for each block $\mathbf{B}$ is determined via a computer program that implements a Gibbs sampler on variables $B \in \mathbf{B}$, where the conditional distribution $p(B|\mathbf{B} \setminus \{B\}, \operatorname{pa}_{\mathcal{G}}(\mathbf{B}))$ is determined via a structural equation of the form $f_B(\mathbf{B} \setminus \{B\}, \operatorname{pa}_{\mathcal{G}}(\mathbf{B}), \epsilon_B)$. This interpretation of $p(\mathbf{B} | \operatorname{pa}_{\mathcal{G}}(\mathbf{B}))$ allows the implementation of a simple intervention operation $\operatorname{do}(b)$. The operation sets $B$ to $b$ by replacing the line of the Gibbs sampler program that assigns $B$ to the value returned by $f_B(\mathbf{B} \setminus \{B\}, \operatorname{pa}_{\mathcal{G}}(\mathbf{B}), \epsilon_B)$ (given a new realization of $\epsilon_B$), with an assignment of $B$ to the value $b$. It was shown (Lauritzen and Richardson, 2002) that in a causal CG model, for any disjoint $\mathbf{Y}, \mathbf{A}$, $p(\mathbf{Y}|\operatorname{do}(\mathbf{a}))$ is identified by the CG version of the g-formula (2.3): $p(\mathbf{Y}|\operatorname{do}(\mathbf{a})) = \prod_{\mathbf{B} \in \mathcal{B}(\mathcal{G})} p(\mathbf{B} \setminus \mathbf{A} | \operatorname{pa}(\mathbf{B}), \mathbf{B} \cap \mathbf{A})|_{\mathbf{A}=\mathbf{a}}$.

In the example above, stable symmetric relationships inducing data dependence, represented by undirected edges, coexist with hidden variables. To represent causal inference in this setting, I generalize earlier developments for hidden variable causal DAG models to hidden variable causal CG models. Specifically, I first define a latent projection analogue called the segregated projection for a large class of hidden variable CGs using segregated graphs (SGs). I then define a factorization for SGs that generalizes the nested factorization and the CG factorization, and show that if a distribution $p(\mathbf{V} \cup \mathbf{H})$ factorizes given a CG $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$ in the class, then $p(\mathbf{V})$ factorizes according to the segregated projection $\mathcal{G}(\mathbf{V})$. Finally, I derive identification theory for hidden variable CGs as a generalization of (2.4) that can be viewed as a truncated SG factorization.

### 4.5.1 Segregated Projections Of Latent Variable Chain Graphs

Fix a chain graph CG $\mathcal{G}$ and a vertex set $\mathbf{H}$ such that for all $H \in \mathbf{H}$, $H$ does not lie in $\mathbf{B} \cup \text{pa}_{\mathcal{G}}(\mathbf{B})$, for any $\mathbf{B} \in \mathcal{B}^{nt}(\mathcal{G})$. We call such a set $\mathbf{H}$ *block-safe.*

**Definition 1** *Given a CG $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$ and a block-safe set $\mathbf{H}$, define a segregated projection graph $\mathcal{G}(\mathbf{V})$ with a vertex set $\mathbf{V}$. Moreover, for any collider-free path from any two elements $V_1, V_2$ in $\mathbf{V}$, where all intermediate vertices are in $\mathbf{H}$, $\mathcal{G}(\mathbf{V})$ contains an edge with end points matching the path. That is, we have $V_1 \mathcal{G}ets \circ \ldots \circ \to V_2$ leads to the edge $V_1 \leftrightarrow V_2$, $V_1 \to \circ \ldots \circ \to V_2$ leads to the edge $V_1 \to V_2$, and in $\mathcal{G}(\mathbf{V})$.*

As an example, the SG in Fig. 4-1 (c) is a segregated projection of the hidden variable CG in Fig. 4-1 (a). While segregated graphs preserve conditional independence structure on the observed marginal of a CG for *any* $\mathbf{H}$ (Shpitser, 2015), we chose to further restrict the set $\mathbf{H}$ in order to ensure that the directed edges in the segregated projection retain an intuitive causal interpretation of edges in a latent projection (Verma and Pearl, 1990). That is, whenever $A \to B$ in a segregated projection, $A$

is a causal ancestor of $B$ in the underlying causal CG. SGs represent latent variable CGs, meaning that they allow causal systems that model feedback that leads to network structures, of the sort considered in Lauritzen and Richardson (2002), but simultaneously allow certain forms of unobserved confounding in such causal systems.

### 4.5.2 Segregated Factorization

The segregated factorization of an SG can be defined as a product of two kernels which themselves factorize, one in terms of a CADMG (a conditional graph with only directed and bidirected arrows), and another in terms of a *conditional chain graph (CCG)* $\mathcal{G}(\mathbf{V}, \mathbf{W})$, a CG with the property that the only type of edge adjacent to any element $W$ of $\mathbf{W}$ is a directed edge out of $W$. A kernel $q(\mathbf{V}|\mathbf{W})$ is said to be Markov relative to the CCG $\mathcal{G}(\mathbf{V}, \mathbf{W})$ if $q(\mathbf{V}|\mathbf{W}) = Z(\mathbf{W})^{-1} \prod_{\mathbf{B} \in \mathcal{B}(\mathcal{G})} q(\mathbf{B}| \operatorname{pa}_{\mathcal{G}}(\mathbf{B}))$, and $q(\mathbf{B}| \operatorname{pa}_{\mathcal{G}}(\mathbf{B})) = Z(\operatorname{pa}_{\mathcal{G}}(\mathbf{B}))^{-1} \prod_{\mathbf{C} \in \mathcal{C}((\mathcal{G}_{\mathbf{B} \cup \operatorname{pa}_{\mathcal{G}}(\mathbf{B})})^a); \mathbf{C} \not\subseteq \operatorname{pa}_{\mathcal{G}}(\mathbf{B})} \psi_{\mathbf{C}}(\mathbf{C})$, for each $\mathbf{B} \in \mathcal{B}(\mathcal{G})$.

I now show, given $p(\mathbf{V})$ and an SG $\mathcal{G}(\mathbf{V})$, how to construct the appropriate CADMG and CCG, and the two corresponding kernels. Given a SG $\mathcal{G}$, let *district variables* $\mathbf{D}^*$ be defined as $\bigcup_{\mathbf{D} \in \mathcal{D}(\mathcal{G})} \mathbf{D}$, and let *block variables* $\mathbf{B}^*$ be defined as $\bigcup_{\mathbf{B} \in \mathcal{B}^{nt}(\mathcal{G})} \mathbf{B}$. Since $\mathcal{D}(\mathcal{G})$ and $\mathcal{B}^{nt}(\mathcal{G})$ partition $\mathbf{V}$ in a SG, $\mathbf{B}^*$ and $\mathbf{D}^*$ partition $\mathbf{V}$ as well. Let the induced CADMG $\mathcal{G}^d$ of a SG $\mathcal{G}$ be the graph containing the vertex sets $\mathbf{D}^*$ as $\mathbf{V}$ and $\operatorname{pa}^s_{\mathcal{G}}(\mathbf{D}^*)$ as $\mathbf{W}$, and which inherits all edges in $\mathcal{G}$ between $\mathbf{D}^*$, and all directed edges from $\operatorname{pa}^s_{\mathcal{G}}(\mathbf{D}^*)$ to $\mathbf{D}^*$ in $\mathcal{G}$. Similarly, let the induced CCG $\mathcal{G}^b$ of $\mathcal{G}$ be the graph containing the vertex set $\mathbf{B}^*$ as $\mathbf{V}$ and $\operatorname{pa}^s_{\mathcal{G}}(\mathbf{B}^*)$ as $\mathbf{W}$, and which inherits all edges in $\mathcal{G}$ between $\mathbf{B}^*$, and all directed edges from $\operatorname{pa}_{\mathcal{G}}(\mathbf{B}^*)$ to $\mathbf{B}^*$. We say that $p(\mathbf{V})$ obeys the factorization of a SG $\mathcal{G}(\mathbf{V})$ if $p(\mathbf{V}) = q(\mathbf{D}^*| \operatorname{pa}^s_{\mathcal{G}}(\mathbf{D}^*))q(\mathbf{B}^*| \operatorname{pa}_{\mathcal{G}}(\mathbf{B}^*))$, $q(\mathbf{B}^*| \operatorname{pa}_{\mathcal{G}}(\mathbf{B}^*))$ is Markov relative to the CCG $\mathcal{G}^b$, and $q(\mathbf{D}^*| \operatorname{pa}^s_{\mathcal{G}}(\mathbf{D}^*))$ is in the nested Markov model of the CADMG $\mathcal{G}^d$.

The following theorem gives the relationship between a joint distribution that factorizes given a hidden variable CG $\mathcal{G}$, its marginal distribution, and the corresponding segregated factorization. This theorem is a generalization of the result proven in Richardson et al. (2017) relating hidden variable DAGs and latent projection ADMGs. The proof is deferred to Appendix D

**Theorem 8** *If $p(\mathbf{V}\cup\mathbf{H})$ obeys the CG factorization relative to $\mathcal{G}(\mathbf{V}\cup\mathbf{H})$, and $\mathbf{H}$ is block-safe then $p(\mathbf{V})$ obeys the segregated factorization relative to the segregated projection $\mathcal{G}(\mathbf{V})$.*

## 4.6 A Complete Identification Algorithm for Latent Variable Chain Graphs

With Theorem 8 in hand, we are ready to characterize general non-parametric identification of interventional distributions in hidden variable causal chain graph models, where hidden variables form a block-safe set. This result can be viewed on the one hand as a generalization of the CG g-formula derived in Lauritzen and Richardson (2002), and on the other hand as a generalization of the ID algorithm (2.4).

**Theorem 9** *Assume $\mathcal{G}(\mathbf{V}\cup\mathbf{H})$ is a causal CG, where $\mathbf{H}$ is block-safe. Fix disjoint subsets $\mathbf{Y},\mathbf{A}$ of $\mathbf{V}$. Let $\mathbf{Y}^* = \text{ant}_{\mathcal{G}(\mathbf{V})_{\mathbf{V}\setminus\mathbf{A}}}\mathbf{Y}$. Then $p(\mathbf{Y}|do(\mathbf{a}))$ is identified from $p(\mathbf{V})$ if and only if every element in $\mathcal{D}(\widetilde{\mathcal{G}}^d)$ is reachable in $\mathcal{G}^d$, where $\widetilde{\mathcal{G}}^d$ is the induced CADMG of $\mathcal{G}(\mathbf{V})_{\mathbf{Y}^*}$.*

*Moreover, if $p(\mathbf{Y}|do(\mathbf{a}))$ is identified, it is equal to*

$$\begin{aligned}\sum_{\mathbf{Y}^*\setminus\mathbf{Y}} &\left[\prod_{\mathbf{D}\in\mathcal{D}(\widetilde{\mathcal{G}}^d)} \phi_{\mathbf{D}^*\setminus\mathbf{D}}(q(\mathbf{D}^*|\,\text{pa}_{\mathcal{G}(\mathbf{V})}(\mathbf{D}^*));\mathcal{G}^d)\right] \\ &\times\left[\prod_{\mathbf{B}\in\mathcal{B}(\widetilde{\mathcal{G}}^b)} p(\mathbf{B}\setminus\mathbf{A}|\,\text{pa}_{\mathcal{G}(\mathbf{V})_{\mathbf{Y}^*}}(\mathbf{B}),\mathbf{B}\cap\mathbf{A})\right]\Bigg|_{\mathbf{A}=\mathbf{a}}\end{aligned}$$

*where*

$$q(\mathbf{D}^*|\,\text{pa}_{\mathcal{G}(\mathbf{V})}(\mathbf{D}^*)) = \frac{p(\mathbf{V})}{(\prod_{\mathbf{B}\in\mathcal{B}^{nt}(\mathcal{G}(\mathbf{V}))} p(\mathbf{B}|\,\text{pa}_{\mathcal{G}(\mathbf{V})}(\mathbf{B}))},$$

*and $\widetilde{\mathcal{G}}^b$ is the induced CCG of $\mathcal{G}(\mathbf{V})_{\mathbf{Y}^*}$.*

To illustrate the application of this theorem, consider the SG $\mathcal{G}$ in Fig. 4-1 (c), where we are interested in $p(Y_2|\text{do}(a_1, a_2))$. It is easy to see that $\mathbf{Y}^* = \{C_1, C_2, M_1, M_2, Y_2\}$ (see $\mathcal{G}_{\mathbf{Y}^*}$ in Fig. 4-1 (d)) with $\mathcal{B}(\mathcal{G}_{\mathbf{Y}^*}) = \{\{M_1, M_2\}\}$ and $\mathcal{D}(\mathcal{G}_{\mathbf{Y}^*}) = \{\{C_1\}, \{C_2\}, \{Y_2\}\}$. The chain graph factor of the factorization in Theorem 9 is $p(M_1, M_2|A_1 = a_1, A_2, C_1, C_2)$. Note that this expression further factorizes according to the (second level) undirected factorization of blocks in a CCG. For the three district factors $\{C_1\}, \{C_2\}, \{Y_2\}$ in Fig. 4-1 (d), we must fix variables in three different sets $\{C_2, A_1, A_2, Y_1, Y_2\}$, $\{C_1, A_1, A_2, Y_1, Y_2\}$, $\{C_1, C_2, A_1, Y_1, A_2\}$ in $\mathcal{G}^d$, shown in Fig. 4-1 (e). We defer the full derivation involving the fixing operator to the supplementary material (Appendix D). The resulting identifying functional for $p(Y_2|\text{do}(a_1, a_2))$ is:

$$\sum_{\{C_1, C_2, M_1, M_2\}} p(M_1, M_2|a_1, a_2, C_1, C_2) \sum_{A_2} p(Y_2|a_1, A_2, M_2, C_2) p(A_2|C_2) p(C_1) p(C_2) \tag{4.1}$$

## 4.7 Experiments

I now illustrate how identified functionals given by Theorem 9 may be estimated from data. Specifically I consider network average effects (N.E.), the network analogue of the average causal effect (ACE), as defined in Hudgens and Halloran (2008):

$$\text{NE}^i(\mathbf{a}_{-i}) = \frac{1}{N} \sum_i E[Y_i(A_i = 1, \mathbf{A}_{-1} = 1)] - E[Y_i(A_i = 0, \mathbf{A}_{-i} = 0)]$$

in the article sharing example described in section 4.2, and shown in simplified form (for two units) in Fig. 4-1 (a). The experiments and results I present here generalize easily to other network effects such as direct and spillover effects (Hudgens and Halloran, 2008), although I do not consider this here. For purposes of illustration I consider a simple setting where the social network is a 3-regular graph, with networks of size $N = [400, 800, 1000, 2000]$. Under the hidden variable CG model I described in section 4.2, the above effect is identified by a functional which generalizes (4.1) from a network of size 2 to a larger network. Importantly, since I assume a single connected network of $M$ variables, we are in the *full interference setting* where only a single sample from $p(M_1, \ldots M_N | A_1, \ldots, A_N, C_1, \ldots, C_N)$ is available. This means that while the standard maximum likelihood plug-in estimation strategy is possible for models for $Y_i$ and $A_i$ in (4.1), the strategy does not work for the model for $M$. Instead, I adapt the auto-g-computation approach based on the pseudo-likelihood and coding estimators proposed in Tchetgen, Fulcher, and Shpitser (2017), which is appropriate for full interference settings with a Markov property given by a CG, as part of our estimation procedure. Note that the approach in Tchetgen, Fulcher, and Shpitser (2017) was applied for a special case of the set of causal models considered here, in particular those with no unmeasured confounding. Here I use the same approach for estimating general functionals in models that may include unobserved confounders between treatments and outcomes. In fact, our example model is analogous to the model in Tchetgen, Fulcher, and Shpitser (2017), in the same way that the front-door criterion is to the backdoor criterion in causal inference under the assumption of iid data (Pearl, 2009).

The detailed estimation strategy, along with a more detailed description of our results, is described in Appendix D. I performed 1000 bootstrap samples of the 4 different networks. Since calculating the true causal effects is intractable even if true model parameters are known, I calculate the approximate 'ground truth' for each intervention by sampling from our data generating process under the intervention 5 times and averaging the relevant effect. I calculated the (approximation of) the bias of each effect by subtracting the estimate from the 'ground truth.' The 'ground truth' network average effects range from $-.453$ to $-.456$. As shown in Tables 4-I and 4-II, both estimators recover the ground truth effect with relatively small bias. Estimators for effects which used the pseudo-likelihood estimator for $M$ generally have lower variance than those that used the coding estimator for $M$, which is expected due to the greater efficiency of the former. This behavior was also observed in Tchetgen, Fulcher, and Shpitser (2017). In both estimators, bias decreases with network size. This is also expected intuitively, although detailed asymptotic theory for statistical inference in networks is currently an open problem, due to dependence of samples.

| 95% Confidence Intervals of Bias of Network Average Effects | | | | | |
|---|---|---|---|---|---|
| | $N$ | 400 | 800 | 1000 | 2000 |
| Estimator | Coding | (-.157, .103) | (-.129, .106) | (-.100, .065) | (-.086, .051) |
| | Pseudo | (-.133, .080) | (-.099, .089) | (-.116, .074) | (-.070, .041) |

Table 4-I. 95% confidence intervals for the bias of each estimating method for the network average effects. All intervals cover the approximated ground truth since they include 0

| Bias of Network Average Effects | | | | | |
|---|---|---|---|---|---|
| | $N$ | 400 | 800 | 1000 | 2000 |
| Estimator | Coding | -.000 (.060) | -.020 (.051) | -.024 (.052) | -.022 (.034) |
| | Pseudo | .006 (.052) | -.023 (.042) | -.023 (.042) | -.021 (.026) |

Table 4-II. The biases of each estimating method for the network average effects. Standard deviation of the bias of each estimate is given in parentheses.

## 4.8 Conclusion

In this chapter, I generalized existing non-parametric identification theory for hidden variable causal DAG models to hidden variable causal chain graph models, which can represent both causal relationships, and stable symmetric relationships that induce data dependence. Specifically, I gave a representation of all identified interventional distributions in such models as a truncated factorization associated with *segregated graphs*, mixed graphs containing directed, undirected, and bidirected edges which represent marginals of chain graphs.

I also demonstrated how statistical inference may be performed on identifiable causal parameters, by adapting a combination of maximum likelihood plug in estimation, and methods based on coding and pseudo-likelihood estimators that were adapted for full interference problems in Tchetgen, Fulcher, and Shpitser (2017). I illustrated my approach with an example of calculating the effect of community membership on article sharing if the effect of the former on the latter is mediated by a complex social network of units inducing full dependence.

# Chapter 5

# General Identification of Dynamic Treatment Regimes Under Interference

## 5.1 Introduction

To this point, I have highlighted the challenges that arise when attempting to make valid causal inferences from observational data. Chief among the obstacles is latent confounding, which can lead to biased estimates of effects. As discussed, prior work has given positive results towards overcoming the latent confounding issue; not by circumventing confounding, but rather by making it possible to clearly delineate when confounding is or is not problematic enough to render inference impossible.

Chapters 3 and 4 extended the prior work on identification theory in two key ways. The former enables the analysis of a broader class of interventions in the form of several sound and complete algorithms for the identification of the effects of policies, path-specific effects, and path-specific policies. The latter established a principled approach to handling network dependence when attempting to make causal inferences from non-iid data.

The present chapter serves as a synthesis between these two lines of inquiry. Derived from novel research originally published in Sherman, Arbour, and Shpitser (2020), I describe theory for the identification of policy effects in dependent data settings. Before diving into the motivation for this direction below, I wish to point out the existence of two other works that were developed by other authors in parallel. First, Ogburn, Shpitser, and Lee (2018) provided substantial commentary on the efficacy of using chain graphs to represent network dynamics when performing causal inference. The second, Viviano (2019) similarly considers policies under interference. The work covered in this chapter differs substantially: Viviano focuses on welfare maximization and assumes units are identically distributed. The characterization of policy interventions studied here generalizes welfare maximization and, as in Ogburn, Shpitser, and Lee (2018) and prior chapters of this thesis, the present network representation is non-parametric.

**Motivating Policies in Networks.** In this chapter, we consider identification of DTRs in the interference setting. As motivation, consider the following example from psephology (the study of elections) (Blackwell, 2013): candidates running for public office target voters by purchasing television advertisements; each candidate must decide how many ads to buy and whether they should be positive ("my record is stellar") or negative ("my opponent is scandalous").

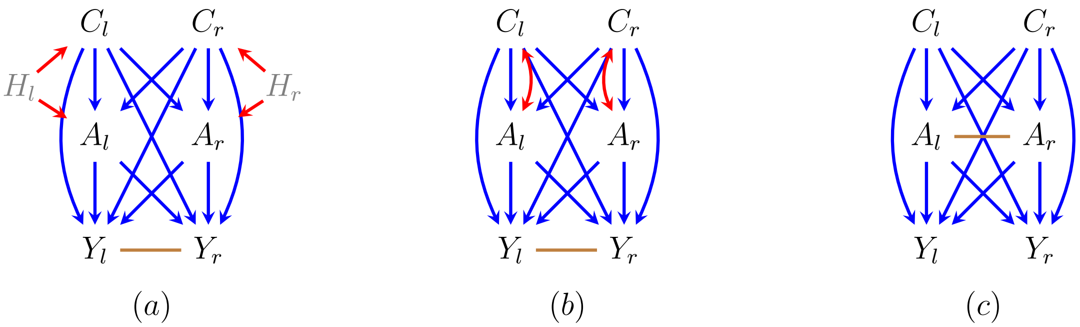


Figure 5-1. Graphical representations of competitive dynamics in an election campaign, where (a) $H$'s represent latent confounders, (b) is a latent projection with $H$'s replaced by bi-directed edges, and (c) is an alternative model where $A$'s exhibit best-response dynamics.

These dynamics can be represented via the causal graphs in Fig. 5-1. For each candidate, $C$ denotes observed pre-decision covariates, such as prior polling performance, previous advertising, and cash on hand, $A$ represents the candidate's advertising decision, $Y$ represents polling performance in the current decision time frame, and $H$ represents unobserved confounders that affect the candidate's pre-decision covariates and decision but don't directly affect the outcome. $l$ and $r$ index the variables for a left- and right-leaning candidate respectively. Directed edges denote a direct causal relationship, while undirected edges denote non-causal dependence (e.g. $A_l - A_r$ could be interpreted as candidates acting based on beliefs about what each other will do). While we use this two-candidate example as motivation throughout this manuscript, our contributions apply to networks of arbitrary size and topology.

The remainder of this chapter is organized as follows: I fix notation and discuss relevant background work in Secs. 5.2 and 5.3. We characterize the variety of possible policy interventions in Sec. 5.4. We then give a novel identification result for effects of policy interventions in Lauritzen-Wermuth-Freydenburg (LWF) latent-variable chain graphs (Lauritzen, 1996; Lauritzen and Richardson, 2002) in Sec. 5.5. We demonstrate estimation of these effects via a simulation study in Sec. 5.6 and conclude with a discussion of ongoing work.

## 5.2 Notation

As in Chapter 4, I will employ segregated graphs (SGs) (Shpitser, 2015) to represent causal network dynamics. I will not review all the background notation for this chapter as we did in our original paper Sherman, Arbour, and Shpitser (2020). Instead, as in past chapters, I will only introduce concepts and notation that are new in this chapter and refer to reader to Chapters 2-4 for the full background.

Recall from Chapter 4 that the anterior $\text{ant}_{\mathcal{G}}(V)$ is the set of nodes with a partially directed path – a path containing only $\rightarrow$ and $-$ edges such that no set of undirected edges can be oriented to form a directed cycle – *into* $V$. Here, we also define the exterior $\text{ext}_{\mathcal{G}}(V)$ is the set of nodes with a partially directed path *out of* $V$. In turn, the strict exterior $\overline{\text{ext}}_{\mathcal{G}}(V) \subseteq \text{ext}_{\mathcal{G}}(V)$ omits $V$ and the set $\{W \in \mathbf{V} : W - \cdots - V\}$. By convention, $\text{ext}_{\mathcal{G}}(V) \cap \text{ant}_{\mathcal{G}}(V) \cap \text{dis}_{\mathcal{G}}(V) = \{V\}$. As with other genealogical notions, the exterior and strict exterior can be extended to sets. When the relevant graph is clear from context, we drop the $\mathcal{G}$ subscript.

Also recall from Chapter 3 that for graphs with a partial ordering $\prec$ on $\mathbf{V}$, let $\mathbf{V}_{\prec A}$ denote $A$'s predecessors in the ordering. Additionally, for a set $\mathbf{S} \subseteq \mathbf{V}$ in $\mathcal{G}$, let $\mathcal{G}_{\mathbf{S}}$ refers to the subgraph of $\mathcal{G}$ containing only $\mathbf{S}$ and edges connecting nodes in $\mathbf{S}$.

### 5.2.1 Causal Graphical Models

As before, this chapter will assume Pearl's functional model. As a reminder, in DAGs counterfactuals $V(\mathbf{a})$ are determined by structural equations $f_V(\mathbf{a}, \epsilon_V)$, which remain invariant under an intervention $\mathbf{a}$; $\epsilon_V$ denotes an exogenous random variable for $f_V$. By *recursive substitution*, we can define all other variables in the model: for $\mathbf{A} \subseteq \mathbf{V} \setminus \{V\}$ and $\mathbf{a}$ in the state space of $\mathbf{A}$, $p(V(\mathbf{a}))$ (sometimes written as $p(\mathbf{V}|\text{do}(\mathbf{a})$ (Pearl, 2009)) is defined as $V(\mathbf{a}_{\text{pa}(V)}, \{W(\mathbf{a}) : W \in \text{pa}(V) \setminus \mathbf{A}\})$.

In this chapter we will focus primarily on causal chain graphs (CGs), which follow similar semantics. Each variable $B$ in a block $\mathbf{B}$ is determined by a structural equation $f_B(\mathbf{B} \setminus \{B\}, \text{pa}(\mathbf{B}), \epsilon_B)$, a function of other variables in $\mathbf{B}$, the parents of $\mathbf{B}$, and an exogenous variable. Each $\mathbf{B}$'s joint distribution $\mathbf{B}$ is obtained by Gibbs sampling over the structural equations for $\mathbf{B}$ until equilibrium (trivial blocks equilibrate instantly). Assuming an ordering on blocks in $\mathcal{G}$, but not on variables in each block, and iid realizations of $\epsilon_{B_i}$, the data generating process for CGs is given by Procedure 1 (Lauritzen and Richardson, 2002).

**Procedure 1** CG Data Generating Process

```
procedure CG-DGP($\mathcal{G}, \{f_B : B \in \mathbf{V}\}$)
    for each block $\mathbf{B}_i \in \mathcal{B}(\mathcal{G})$ do
        repeat
            for each variable $B_j \in \mathbf{B}_i$ do
                $B_j \leftarrow f_{B_j}(\mathbf{B}_i \setminus B_j, \text{pa}_{\mathcal{G}}(\mathbf{B}_i), \epsilon_{B_j})$
            end for
        until equilibrium
    end forreturn V
end procedure
```

| Graph Type | Latents | Intervention Type | $\mathbf{Y}^\star$ | Modified Factorization |
|---|---|---|---|---|
| DAG | No | Node – $\mathbf{a}$ | N/A | $\prod_{V \in \mathbf{V} \setminus \mathbf{A}} p(V \mid \text{pa}(V))\vert_{\mathbf{A}=\mathbf{a}}$ |
| CG | No | Node – $\mathbf{a}$ | N/A | $\prod_{\mathbf{B} \in \mathcal{B}(\mathcal{G})} p(\mathbf{B} \setminus \mathbf{A} \mid \text{pa}(\mathbf{B}), \mathbf{B} \cap \mathbf{A})\vert_{\mathbf{A}=\mathbf{a}}$ |
| ADMG | Yes | Node – $\mathbf{a}$ | $\text{an}_{\mathcal{G}_{\mathbf{V} \setminus \mathbf{A}}}(\mathbf{Y})$ | $\prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}_{\mathbf{Y}^\star})} \phi_{\mathbf{V} \setminus \mathbf{D}}(p(\mathbf{V}); \mathcal{G})\vert_{\mathbf{A}=\mathbf{a}}$ |
| SG | Yes | Node – $\mathbf{a}$ | $\text{ant}_{\mathcal{G}_{\mathbf{V} \setminus \mathbf{A}}}(\mathbf{Y})$ | $\prod_{\mathbf{D} \in \mathcal{D}(\tilde{\mathcal{G}}^d)} \phi_{\mathbf{D}^\star \setminus \mathbf{D}}(q(\mathbf{D}^\star \mid \text{pa}^s_{\mathcal{G}}(\mathbf{D}^\star)); \mathcal{G}^d) \quad \times$ <br> $\prod_{\mathbf{B} \in \mathcal{B}(\tilde{\mathcal{G}}^b)} p(\mathbf{B} \setminus \mathbf{A} \mid \text{pa}_{\mathcal{G}_{\mathbf{Y}^\star}}(\mathbf{B}), \mathbf{B} \cap \mathbf{A})\vert_{\mathbf{A}=\mathbf{a}}$ |
| ADMG | Yes | Policy – $\mathbf{f_A}$ | $\text{an}_{\mathcal{G}_{\mathbf{f_A}}}(\mathbf{Y})$ | $\prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}_{\mathbf{Y}^\star})} \phi_{\mathbf{V} \setminus \mathbf{D}}(p(\mathbf{V}); \mathcal{G})\vert_{\mathbf{A}=\tilde{\mathbf{a}}}$ |

Table 5-I. Summary of existing identification approaches. The first two rows use standard g-formulas, the third row is the ID algorithm, and the final two extend ID. The present work generalizes the last two rows. In the fifth row, $\tilde{\mathbf{a}} = \{A = f_A(\mathbf{W}_A) \mid A \in \text{pa}_{\mathcal{G}}(\mathbf{D}) \cap \mathbf{A}\}$ if $\text{pa}_{\mathcal{G}}(\mathbf{D}) \cap \mathbf{A} \neq \emptyset$ and $\tilde{\mathbf{a}} = \emptyset$ otherwise.

## 5.3 Identification in Latent-Variable Causal Graphical Models

In this section, I briefly review identification theory in latent variable causal models. See Chapters 3-4 for complete details. The current chapter bridges these literatures: I posit a sound and complete algorithm for the identification of responses to policies in latent variable (LV) causal CGs.

### 5.3.1 Re-expressing the ID Algorithm

Richardson et al. (2017) makes clear the connections between the ID algorithm, which is a modified nested factorization of acyclic directed mixed graphs (ADMGs), and the g-formula (Table 5-I, first row), which is a modified DAG factorization.

As described in prior chapters, the sequence of identification papers from Tian and Pearl (2002) to Richardson et al. (2017) provided an increasingly comprehensive and general framework for developing identification theory for a variety of graph and intervention types. In particular, their shared formalism enables straightforward generalizations to other identification settings. For these reasons, the SG policy identification results described in this chapter are based on this framework. The framework's base concepts (such as kernels, fixing, and the nested model) are covered in detail elsewhere. For reference, each existing ID approach is summarized in Table 5-I.

**Latent Projections.** The details of latent projection ADMGs can be found elsewhere. I will, however, remind the reader that segregated graphs are the chain graph analogue of ADMGs, where SGs represent an equivalence class of LV-CGs. For a latent variable CG $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$, $\mathbf{H}$ is *block-safe* (Sherman and Shpitser, 2018) if no $V \in \mathbf{V}$ has a latent parent and no latent $H \in \mathbf{H}$ has an incident undirected edge. By applying the same latent projection operation mentioned above to a LV-CG with block-safe $\mathbf{H}$, one obtains the corresponding SG.

Again, without rehashing too many details, I will provide a brief example of how the ID algorithm can be used in the elections example from above (Fig. 5-1(c)). Suppose we assume each candidate's decision is independent of other decisions given covariates (i.e., no $A_l - A_r$ edge). We can use the ID algorithm (Eq. 2.4) formula to consider the effect on a candidate's polling of advertising positively and negatively in fixed proportion (say, equally, $a = .5$).

As another example, consider the subgraph on $C_1, A_1, M_1, Y_1$ in Fig. 5-2(a); $p(Y_1|\text{do}(a_1))$ is not identified (Shpitser and Pearl, 2006). In the $C_2, A_2, M_2, Y_2$ subgraph, however, $p(Y_2|\text{do}(a_2))$ is identified by the front-door formula:

$$\sum_{M_2,C_2} p(M_2|a_2,C_2)p(C_2)\sum_{A_2'} p(Y_2|M_2,C_2,A_2')p(A_2'|C_2)$$

### 5.3.2 Identification in Segregated Graphs

**The Segregated Factorization.** Again, recall from Chapter 4 that we can extending the factorizations for ADMGs and CGs, Sherman and Shpitser (2018) to define the segregated factorization for SGs.

An SG $\mathcal{G}$ is partitioned by variables that lie in non-trivial blocks, denoted $\mathbf{B}^\star = \cup_{\mathbf{B}\in\mathcal{B}^{nt}(\mathcal{G})}\mathbf{B}$, and those that don't, denoted $\mathbf{D}^\star = \cup_{\mathbf{D}\in\mathcal{D}(\mathcal{G})}\mathbf{D}$. An SG satisfying the segregated factorization can be expressed as the product of kernels for these two sets.

The first kernel, $q(\mathbf{B}^\star|\,\text{pa}^s_{\mathcal{G}}(\mathbf{B}^\star)) = \prod_{\mathbf{B}\in\mathcal{B}^{nt}(\mathcal{G})} p(\mathbf{B}|\,\text{pa}_{\mathcal{G}}(\mathbf{B}))$, factorizes with respect to a conditional chain graph (CCG) $\mathcal{G}(\mathbf{V},\mathbf{W})$, which we denote by $\mathcal{G}^b$ with $\mathbf{V}$ corresponding to $\mathbf{B}^\star$ and $\mathbf{W}$ to $\text{pa}^s(\mathbf{B}^\star)$. $\mathcal{G}^b$ contains edges between nodes in $\mathbf{B}^\star$ and between nodes in $\text{pa}^s_{\mathcal{G}}(\mathbf{B}^\star)$ that exist in $\mathcal{G}$.

The second kernel, $q(\mathbf{D}^\star|\,\text{pa}^s_{\mathcal{G}}(\mathbf{D}^\star)) = \frac{p(\mathbf{V})}{q(\mathbf{B}^\star|\,\text{pa}^s_{\mathcal{G}}(\mathbf{B}^\star))}$, nested factorizes with respect to a CADMG denoted $\mathcal{G}^d$, with random nodes $\mathbf{D}^\star$ and fixed nodes $\text{pa}^s(\mathbf{D}^\star)$. Like $\mathcal{G}^b$, $\mathcal{G}^d$ contains edges between nodes in $\mathbf{D}^\star$ and between nodes in $\text{pa}^s_{\mathcal{G}}(\mathbf{D}^\star)$ that are present in $\mathcal{G}$.

For example, in the graph in Fig. 5-2(a), we have

$$\begin{aligned} q(\mathbf{D}^\star|\,\text{pa}^s_{\mathcal{G}}(\mathbf{D}^\star)) &= p(Y_2,Y_3,A_2|C_2,M_2,M_3) \\ &\quad\times p(Y_1,A_1,C_1|M_1)p(A_3|C_3) \end{aligned}$$

$$q(\mathbf{B}^\star|\,\text{pa}^s_{\mathcal{G}}(\mathbf{B}^\star)) = p(M_1,M_2,M_3|A_1,A_2,A_3)p(C_2,C_3)$$

which correspond to Fig. 5-2(b) and (c) respectively.

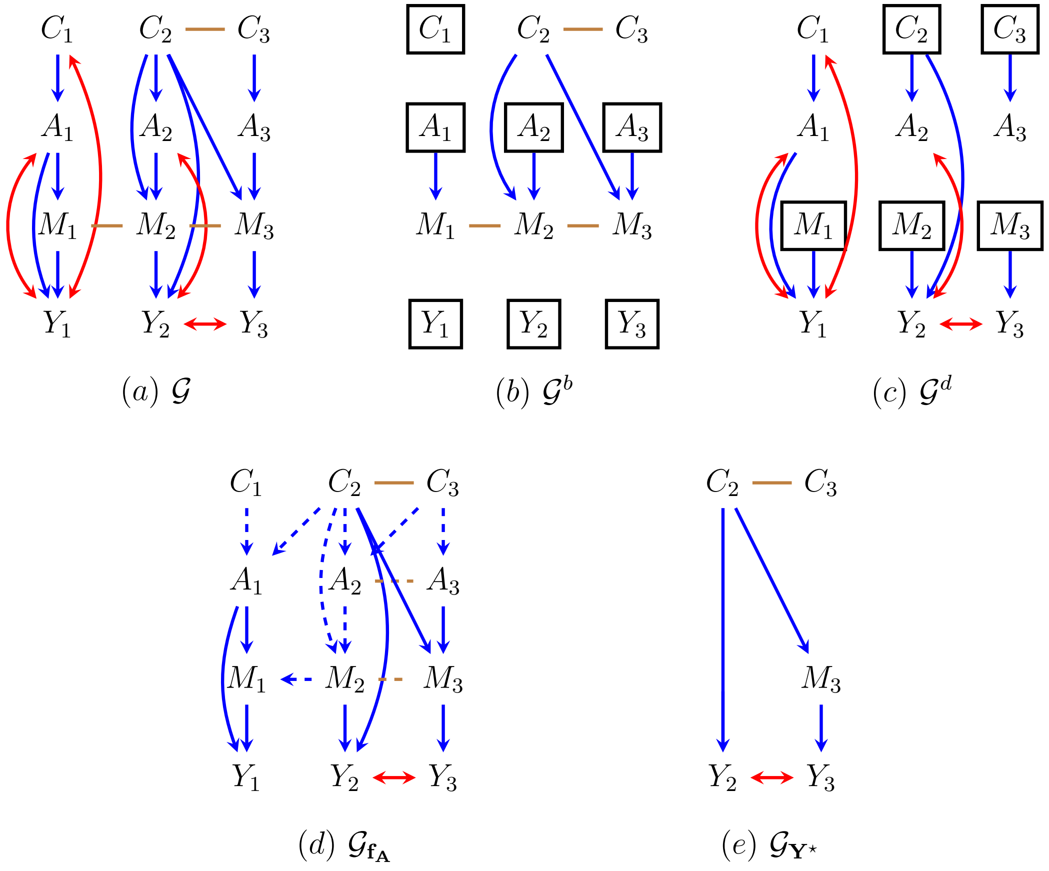


Figure 5-2. (a) An SG $\mathcal{G}$ where bi-directed edges signify the presence of latent confounders. (b) and (c) The conditional chain graph $\mathcal{G}^b$ and conditional ADMG $\mathcal{G}^d$ obtained from $\mathcal{G}$. (d) The post-intervention graph $\mathcal{G}_{\mathbf{f_A}}$ induced by the policy intervention $\mathbf{f_A}$ as described in Sec. 5.4. Nodes with changed structural equations have dashed incoming edges. (e) The corresponding $\mathcal{G}_{\mathbf{Y}^\star}$ for $\mathcal{G}_{\mathbf{f_A}}$ in 5-2(b) with outcome $\mathbf{Y} = \{Y_2, Y_3\}$.

**The Segregated Graph ID Algorithm.** The above leads to a segregated graph ID algorithm. See 4.6. Returning to this chapter's running elections example, Fig. 5-1(b), the SG ID formula is applicable when considering the effect of the left-leaning candidate taking a fixed action $a_l$, with the right-leaning candidate's action still having an impact on the left's poll standing. $p(Y_l(a_l))$ is identified by:

$$\sum_{C_l, C_r, A_r, Y_r} p(Y_l, Y_r | C_l, C_r, A_r, a_l) p(A_2 | C_l, C_r) p(C_l) p(C_r)$$

### 5.3.3 Policy Interventions in ADMGs

Again, recall from Chapter 3 that policy interventions represent an extension of classical node interventions.. For an ADMG $\mathcal{G}(\mathbf{V})$ with topological ordering $\prec$ on $\mathbf{V}$ and an intervention set $\mathbf{A} \subseteq \mathbf{V}$, let $\mathbf{f_A}$ be the set of policies $\{f_A : A \in \mathbf{A}\}$. Each $f_A$ is a stochastic function of some $\mathbf{W}_A \subseteq \mathbf{V}_{\prec A}$, where $f_A(\mathbf{W}_A)$ maps the state space of $\mathbf{W}_A$ to the state space of $A$. Intervening with $f_A$ corresponds to removing edges *into* $A$ in $\mathcal{G}$ and adding edges from $\mathbf{W}_A$ to $A$, yielding a new graph $\mathcal{G}_{\mathbf{f_A}}$. A policy-analogue of the ID algorithm follows was proved sound and complete in Shpitser and Sherman (2018) and presented in Chapter 3. As with other ID results, this result can be viewed as a novel modified factorization, as shown in row five of Table 5-I.

In our elections example, assume candidates' decisions and outcomes are independent of each other. This formula (Eq. 3.11) can be used to consider the effect on a candidate's polling of advertising based on the relevant covariates, e.g., if the election is less than 2 months away, advertise negatively, and buy positive ads until then.

## 5.4 Varieties of Policy Interventions

We now describe extensions of policy interventions to network data representable by SGs. These interventions correspond to replacing structural equations in Procedure 1 with new equations, under conditions we describe below such that the resulting data generating process yields a new SG. As we discuss, these policy interventions induce a variety of edge changes in SGs.

### 5.4.1 Inducing Direct Causation

As in the latent-variable DAG case (Shpitser and Sherman, 2018), we can intervene by inducing a parent-child relationship between the treatment node and other variables in the graph or modify the nature of existing relationship. In our elections example from

Sec. 5.1, this might correspond to intervening on the left candidate's decision $A_l$ such that she adopts a new strategy for responding to her competitor's characteristics $C_r$ relative to her (observed) status quo strategy. For illustrative purposes, this type of intervention is demonstrated by the addition of the $C_2 \rightarrow A_1$ edge and the modification to the $C_1 \rightarrow A_1$ edge between Fig. 5-2(a) and 5-2(d).

### 5.4.2 Inducing or Modifying Undirected Dependence

We can also consider changing the block structure of the SG. There are two types of such interventions:

1. *Modifying the functional form encoded by an existing undirected edge.* In Fig. 5-1 (b), we can think of the undirected edge $A_l - A_r$ as representing each candidates' beliefs about the other candidate's actions. In the observed data, candidates will best-respond to each other according to these beliefs. We can imagine changing the way one (or both) of the candidates reasons about their opponent's possible actions, such as making one candidate hyper-responsive to their opponent's anticipated action. Mechanically, we intervene on $A_l$ (analogously $A_r$) with a function $f_{A_l}$ that takes $A_r$ as an argument. We needn't intervene on the other candidate to maintain the undirected edge between the $A$'s. This type of intervention is demonstrated by the change to the $M_2 - M_3$ edge from Fig. 5-2(a) to 5-2(d).

2. *Inducing co-dependence by adding a new undirected edge between two nodes.* This might correspond to having a third candidate $c$ join the race and intervening such that $A_c - A_l$ and $A_c - A_r$. In this case, it is necessary to intervene on both endpoint nodes for the new undirected edge in order; we modify the respective structural equations to take the other endpoint as an argument. We further restrict these interventions by requiring that they do not induce a partially directed cycle, which would violate the segregation property of the graph. We formalize this requirement below. We note

that this type of intervention can be thought of as a chain graph generalization of connection interventions, proposed in Sherman and Shpitser (2019) (See also Chapter 6). As an example, consider the addition of the $A_2 - A_3$ edge in Fig. 5-2(d) relative to 5-2(a).

### 5.4.3 Removing Dependence

Finally, we can consider removing undirected dependence between nodes. Once again there are two types:

1. *Partial removal.* We intervene on a single node to make its structural equation no longer a function of the other end point of the undirected edge. In our elections example (Fig. 5-1(c)), this corresponds to a 'first mover' scenario where $A_l$ is made to not depend on $A_r$ and thus candidate $l$ makes her decision before candidate $r$. Graphically, we change the undirected edge $A_l - A_r$ to a directed edge $A_l \rightarrow A_r$ since $A_r$ is still determined by candidate $l$'s decision; see, for instance, the $M_1 - M_2$ and $M_1 \leftarrow M_2$ edges in Fig. 5-2(a) and 5-2(d).

2. *Complete removal.* We remove both dependences by intervening on both endpoints of an undirected edge so that the structural equations are no longer functions of each other. This corresponds to a candidate dropping out of the race in our elections example. Like dependence-inducing interventions above, this intervention type can be viewed as an SG analogue of severance interventions (Sherman and Shpitser, 2019) (again, see Chapter 6).

## 5.5 Identification of Policies in Segregated Graphs

In this section we formalize policy interventions and provide a procedure for obtaining the post-intervention graph from $\mathcal{G}$. We then give a criterion for the identification of policy interventions in SGs (Shpitser, 2015) and demonstrate application of this

**Procedure 2** Obtaining $\mathcal{G}_{\mathbf{f_A}}$ from $\mathcal{G}$

```
procedure InterveneGraph($\mathcal{G}, \mathbf{f_A}(\mathbf{Z_A})$)
    Initialize $\mathcal{G}_{\mathbf{f_A}} \leftarrow \mathcal{G}$
    for each $A \in \mathbf{A}$ do
        Replace all $V - A$ with $A \to V$ in $\mathcal{G}_{\mathbf{f_A}}$
        Remove all $\cdot \to A$, $\cdot \leftrightarrow A$ from $\mathcal{G}_{\mathbf{f_A}}$
        Add edges $\mathbf{Z}_A \to A$ in $\mathcal{G}_{\mathbf{f_A}}$
    end for
    for each $V_i, V_j \in \mathbf{V}$ do
        if $V_i \to V_j$ and $V_j \to V_i$ in $\mathcal{G}_{\mathbf{f_A}}$ then
            Remove $V_i \to V_j$ and $V_j \to V_i$ from $\mathcal{G}_{\mathbf{f_A}}$
            Add $V_i - V_j$ in $\mathcal{G}_{\mathbf{f_A}}$
        end if
    end forreturn $\mathcal{G}_{\mathbf{f_A}}$
end procedure
```

criterion to Fig. 5-2 and to our electoral example, Fig. 5-1. We defer proofs and derivations to Appendix E.

### 5.5.1 Formalizing Policy Interventions in Segregated Graphs

Before providing identification conditions, we first formally define policy interventions in SGs. Recall that in ADMGs a policy $f_A(\mathbf{W}_A) \in \mathbf{f_A}$ was required to be a function of variables $\mathbf{W}_A$ preceding $A$ in a topological ordering on the nodes in $\mathcal{G}$. In SGs we loosen this restriction such that $f_A$ operates as a structural equation that can also be a stochastic function of variables in the same block as $A$. For an intervention inducing a block or modifying the structural equations in a block, we use Procedure 1 to obtain a new block distribution.

For $f_A(\mathbf{Z}_A)$ to be a valid policy in an SG $\mathcal{G}(\mathbf{V})$, we require $\mathbf{Z}_A \subseteq \mathbf{V} \setminus \overline{\text{ext}}(A)$. In turn, for $\mathbf{f_A}$ to be valid, all constituent policies must be valid and they may not collectively violate the CG property by inducing a partially directed cycle. We formalize this notion as follows: let $A_i \triangle A_j$ denote that variable $A_i$ is made (either directly or indirectly) a function of $A_j$ for $A_i, A_j \in \mathbf{A}$. To prevent partially directed cycles, we stipulate that if $A_i \triangle A_j$ and $A_j \triangle A_i$ then we require $A_i \in \mathbf{Z}_{A_j}$ and vice versa. This motivates the following definition.

**Definition 2** *A policy intervention $\mathbf{f_A}(\mathbf{Z_A})$ is 'segregation preserving' if (a) for each $A \in \mathbf{A}$, $\mathbf{Z}_A \subseteq \mathbf{V} \setminus \overline{\text{ext}}(A)$, and (b) for any $A_i, A_j \in \mathbf{A}$ if $A_i \triangle A_j$ and $A_j \triangle A_i$, we have that $A_i \in \mathbf{Z}_{A_j}$ and $A_j \in \mathbf{Z}_{A_i}$.*

For a given intervention set $\mathbf{f_A}$, we can construct a post-intervention graph $\mathcal{G}_{\mathbf{f_A}}$ according to Procedure 2, which follows from the analogous procedure for policy identification in LV-DAGs. In Lemma 1, we show that $\mathcal{G}_{\mathbf{f_A}}$ is an SG when $\mathbf{f_A}$ is segregation-preserving. As an example of this procedure's application, consider Fig. 5-2(a). Suppose we wish to perform an intervention $\mathbf{f_A}(\mathbf{Z_A})$ as in Table 5-II. Then $\mathcal{G}_{\mathbf{f_A}}$ is given by Fig. 5-2.

| $A \in \mathbf{A}$ | $A_1$ | $A_2$ | $A_3$ | $M_2$ |
|---|---|---|---|---|
| $\mathbf{Z}_A$ | $C_2$ | $C_2, C_3, A_3$ | $A_2, C_3$ | $A_2, C_2, M_3$ |

Table 5-II. Intervention variables $A \in \mathbf{A}$ and induced dependences $\mathbf{Z}_A$ for the intervention in Fig. 5-2

### 5.5.2 Identification Results

First, we show that the post-intervention $\mathcal{G}_{\mathbf{f_A}}$ is an SG.

**Lemma 1** *Given an SG $\mathcal{G}(\mathbf{V})$ and a segregation-preserving intervention $\mathbf{f_A}(\mathbf{Z_A})$, the post-intervention graph $\mathcal{G}_{\mathbf{f_A}}$ obtained via Procedure 2 is an SG.*

We now present the main result of this paper. This theorem provides sufficient conditions for the identification of the effects of policy interventions in SGs.

**Theorem 10** *Let $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$ be a causal LV-CG with $\mathbf{H}$ block-safe, and a topological order $\prec$. Fix disjoint $\mathbf{Y}, \mathbf{A} \subseteq \mathbf{V}$. Let $\mathbf{f_A}(\mathbf{Z_A})$ be a segregation preserving policy set. Let $\mathbf{Y}^\star \equiv \text{ant}_{\mathcal{G}_{\mathbf{f_A}}}(\mathbf{Y}) \setminus \mathbf{A}$. Let $\mathcal{G}^d, \tilde{G}^d$ be the induced CADMGs on $\mathcal{G}_{\mathbf{f_A}}$ and $\mathcal{G}_{\mathbf{Y}^\star}$, and*

*$\tilde{G}^b$ the induced CCG on $\mathcal{G}_{\mathbf{Y}^\star}$. Let $q(\mathbf{D}^\star|\operatorname{pa}^s_{\mathcal{G}_{\mathbf{f_A}}}(\mathbf{D}^\star)) = \prod_{\mathbf{D}\in\mathcal{G}_{\mathbf{f_A}}} q(\mathbf{D}|\operatorname{pa}^s_{\mathcal{G}_{\mathbf{f_A}}}(\mathbf{D}))$, where $q(\mathbf{D}|\operatorname{pa}^s_{\mathcal{G}_{\mathbf{f_A}}}(\mathbf{D})) = \prod_{D\in\mathbf{D}} p(D|\mathbf{V}_{\prec D})$ if $\mathbf{D}\cap\mathbf{A} = \emptyset$ and $q = f_A(\mathbf{Z}_A)$ if $\mathbf{D}\cap\mathbf{A} \neq \emptyset$. $p(\mathbf{Y}(\mathbf{f_A}(\mathbf{Z_A})))$ is identified in $\mathcal{G}$ if and only if $p(\mathbf{Y}^\star(\mathbf{a}))$ is identified in $\mathcal{G}$ for the unrestricted class of policies. If identified, $p(\mathbf{Y}(\mathbf{f_A}(\mathbf{Z_A}))) =$*

$$\sum_{\{\mathbf{Y}^\star\cup\mathbf{A}\}\setminus\mathbf{Y}}\left[\prod_{\mathbf{B}\in\mathcal{B}(\tilde{\mathcal{G}}^b)} p^\star(\mathbf{B}|\operatorname{pa}_{\mathcal{G}_{\mathbf{f_A}}}(\mathbf{B}))\right] \times\left[\prod_{\mathbf{D}\in\mathcal{D}(\tilde{\mathcal{G}}^d)} \phi_{\mathbf{D}^\star\setminus\mathbf{D}}(q(\mathbf{D}^\star|\operatorname{pa}^s_{\mathcal{G}_{\mathbf{f_A}}}(\mathbf{D}^\star));\mathcal{G}^d)\right]\Bigg|_{\mathbf{A}=\tilde{\mathbf{a}}} \tag{5.1}$$

*where (a) $\tilde{\mathbf{a}} = \{A = f_A(\mathbf{Z}_A) : A \in \operatorname{pa}_{\mathcal{G}_{\mathbf{f_A}}}(\mathbf{D})\cap\mathbf{A}\}$ if $\operatorname{pa}_{\mathcal{G}_{\mathbf{f_A}}}(\mathbf{D})\cap\mathbf{A} \neq \emptyset$ and $\tilde{\mathbf{a}}_{\mathbf{D}} = \emptyset$ otherwise, and (b) $p^\star$ is obtained by running Procedure 1 over functions $g_{B_i}(B_{-i}, \operatorname{pa}_{\mathcal{G}_{\mathbf{f_A}}}(B_i), \epsilon_{B_i})$ where $g_{B_i} \in \mathbf{f_A}$ if $B_i \in \mathbf{A}$ and $g_{B_i}$ is given by the observed distribution if $B_i \notin \mathbf{A}$[1].*

The outer sum over $\mathbf{A}$ is extraneous if $\mathbf{f_A}$ corresponds to a set of deterministic policies.

### 5.5.3 Estimands and Optimal Policy Selection

We now demonstrate how to obtain identified functionals via Eq. 5.1. We describe identification of the effect on $\{Y_2, Y_3\}$ in Fig. 5-2(a) of the intervention in Table 5-II, and then give the functional for our elections example, Fig. 5-1(b), which we estimate in the next section.

From Fig. 5-2(a), we obtain $\mathcal{G}_{\mathbf{f_A}}$ in Fig. 5-2(d) by applying the intervention detailed in Table 5-II. In turn, from this post-intervention graph we observe that $\mathbf{Y}^\star = \operatorname{ant}_{\mathcal{G}_{\mathbf{f_A}}}(\mathbf{Y}) \setminus \mathbf{A} = \{C_2, C_3, M_3, Y_2, Y_3\}$ and obtain the induced subgraph $\mathcal{G}_{\mathbf{Y}^\star}$ in Fig. 5-2(e).

[1] This distribution is identified from univariate terms but it cannot be obtained in closed-form.

$\mathcal{G}_{\mathbf{Y}^\star}$ factorizes into kernels relating to district nodes and block nodes:

$$q_{\mathcal{D}}(C_1, A_1, M_1, Y_1, Y_2, Y_3 | C_2, M_2, M_3)$$

and

$$q_{\mathcal{B}}(M_2, M_3, A_2, A_3, C_2, C_3 | \emptyset).$$

The block nodes factorize as a product of blocks, as in the first term of Eq. 5.1. Separately, we must fix sets for each $\mathcal{G}_{\mathbf{Y}^\star}$ district $\{\{M_3\}, \{Y_2, Y_3\}\}$ in $q_{\mathcal{D}}$. This yields the functional (full derivation in Appendix E) for $p(\{Y_2, Y_3\}(\mathbf{f_A}))$:

$$\sum_{\{A_1,A_2,A_3,M_2,M_3,C_2,C_3\}} p^\star(A_2, A_3 | C_2, C_3) p^\star(M_2, M_3 | A_2, A_3, C_2) \\ \times p(Y_2, Y_3 | Y_1, A_1, M_1, M_3, C_1, C_2) p^\star(C_2, C_3)$$

Similarly, we consider the effect on $Y_l$ of intervening with a policy $f_{A_l}(C_l)$ in our electoral example, Fig. 5-1 (b). $f_{A_l}(C_l)$ corresponds to a myopic strategy in which the candidate makes decisions based only on their own covariates. Applying Eq. 5.1,

$$p(Y_l(f_{A_l}(C_l, C_r))) = \sum_{C_l, C_r, A_r, Y_r} p(A_r | C_l, C_r) p(C_l) p(C_r) \\ \times p(Y_l, Y_r | C_l, C_r, A_r, f_{A_l}(C_l, C_r)) \tag{5.2}$$

To choose an optimal action for the left candidate, we select $f_{A_l}(C_l)$ from a set of candidate policies $\mathcal{F}_{A_l}(C_l)$:

$$f_{A_l}(C_l, C_r) = \underset{\tilde{f}_{A_l}(C_l, C_r) \in \mathcal{F}_{A_l}(C_l, C_r)}{\arg\max} \; p(Y_l(\tilde{f}_{A_l}(C_l, C_r))) \tag{5.3}$$

## 5.6 Estimation

We now demonstrate how functionals identified by Eq. 5.1 can be estimated from observed data. Specifically, we seek optimal $f_A(\mathbf{C})$'s for versions of the functional in Eq. 5.2. To do so, we fit nuisance models and utilize the plug-in principle to perform indirect Q-learning for policy optimization. This approach yields consistent estimates

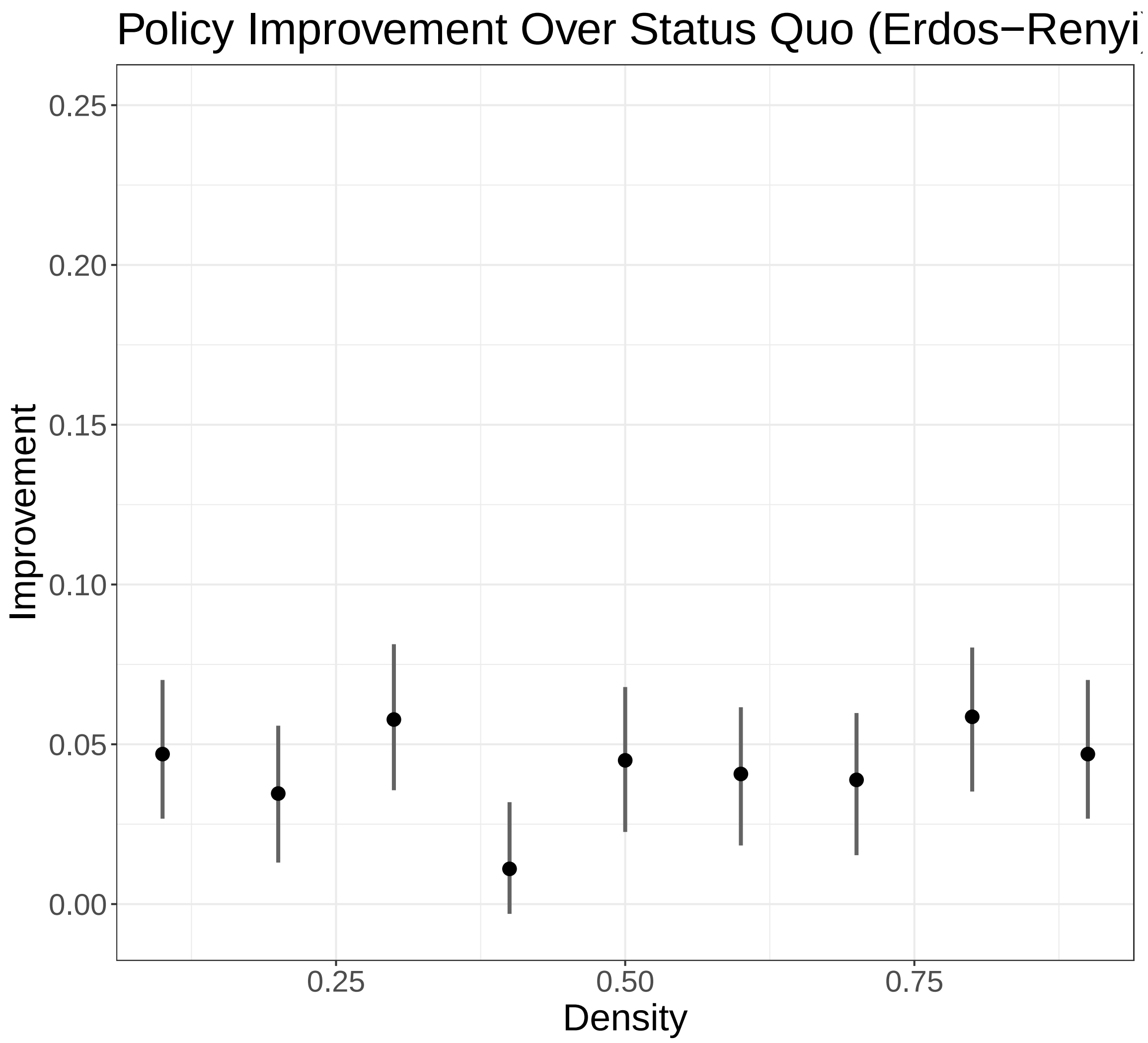


Figure 5-3. Bias of estimates obtained using single-unit modeling, ignoring interference. The presence of bias suggests ignoring interference is highly problematic.

of the optimized outcome under regularity conditions, assuming correctly specified nuisance models (Chakraborty and Moodie, 2013).

For our experiments we first generate 10-node network graphs using three popular network generators: Erdős and Rényi (1960), Watts and Strogatz (1998), and Albert and Barabási (2002). In-unit and cross-unit structures are identical to the 2-node graph in Fig. 5-1(b). We then generate data for each $C$, $A$, and $Y$ using the following densities (note that $C_i$ is a 3-dimensional vector):

$$\begin{aligned} C_{i,j} &\sim \text{Beta}(\alpha_j, \beta_j) \\ p(A_i = 1 | C_i, C_{-i}) &= \text{expit}(\sum_{j=1}^{3} \gamma_j C_{i,j} + \frac{\tau_{AC}}{|\mathcal{N}_i|} \sum_{k \in \mathcal{N}_i} \sum_{j=1}^{3} C_{k,j}) \\ p(Y_i = 1 | A_i, A_{-i}, C_i, C_{-i}, Y_{-i}) &= \text{expit}\Bigg( \eta A_i + \sum_{j=1}^{3} \delta_j C_{i,j} \\ &\quad + \frac{1}{|\mathcal{N}_i|} \sum_{k \in \mathcal{N}_i} \Big( \tau_{YA} A_k + \tau_{YY} Y_k + \sum_{j=1}^{3} \tau_{YC} C_{k,j} \Big) \Bigg) \end{aligned}$$

where $\mathcal{N}_i$ denote unit $i$'s neighbors in $\mathcal{G}_{\mathbf{f_A}}$.

We use Gibbs sampling to approximate undirected edges between $Y$'s (Tchetgen, Fulcher, and Shpitser, 2017). We assume *partial* interference: we generate 1000 samples of each network topology and use these to fit nuisance models. We run the following experiments by obtaining $1,000$ bootstrap replications of the generated data and calculating a 95% confidence interval of the relevant effect:

1. *Bias from incorrectly assuming iid.* As a demonstration of the importance of using interference-aware modeling, we consider performing *node* interventions on each $A_i$ obtained from our Erdős-Rényi samples, setting $A_i$ to 1 and 0. We estimate the average causal effect (ACE) of these node interventions ($E[Y_i(1) - Y_i(0)]$) using models implied by ID (Table 5-I, row three), which provides sound functionals when data *are* iid, as well as models implied by the SG ID algorithm (Table 5-I, row four) which respect the dependent nature of the data. We treat the latter models as 'ground truth' and calculate the bias of the ACE induced by inappropriately assuming data are iid. These results are given in Fig. 5-3. Observing that bias is universally bounded away from 0 in these results, it's clear that it's imperative to respect network dependence in causal modeling.

2. *Benefit of optimizing interventions.* Here we demonstrate the efficacy of policy interventions for picking tailored interventions that optimize a subject's outcome, by estimating the 10-unit version of the identified functional in Eq. 5.2. From our

generated samples, we fit logistic regression models for $E[Y_{-i}|\mathbf{A},\mathbf{C}]$ and $E[Y_i|\mathbf{A},\mathbf{C},Y_{-i}]$, where $i$ denotes the unit we wish to optimize for. Because the true model for $Y$ is an expit model, logistic regression qualifies as a correctly specified model and thus using this procedure is sufficient to ensure consistency of the Q-learning procedure (Chakraborty and Moodie, 2013; Murphy and Russell, 2002). Models for $p(\mathbf{A},\mathbf{C})$ are estimated using the empirical distribution.

For each sample we estimate the effect of intervening with a policy $f_{A_i}(C_i) \in \mathcal{F}_{A_i}(C_i) = \{|C_i|^{-1}\sum_{j\in[|C_i|]} k_j C_{ij} : k_j \in \mathbb{R}\}$ (i.e. $\mathcal{F}_{A_i}$ is the set of means of linear combinations of $C_i$'s components). For this experimental setup, the optimization in 5.3 can be viewed as simply picking the value of the free parameter $\mathbf{k}$[2]. We initialized $\mathbf{k}$ to random values in $[0,1]^3$. We also enforced constraints on $Y_i$ and $A_i$ such that the chosen (for $A$) and predicted (for $Y$) values remain in the state space of those variables: the range $[0,1]$. Optimization was performed using the LBFGS solver.

We report the *difference* between the optimized and observed ('status quo') $Y_i$'s. The results for the Erdős-Rényi generator can be found in Figs. 5-3 and 5-4. Results for the other generators can be found in the supplementary material, Appendix E. Since $Y$ is binary, an expected difference of .05 corresponds to a 5.0% increase in $Y$ over the status quo. Fig. 5-4 demonstrates that the proposed approach virtually guarantees an improved outcome over the status quo.

## 5.7 Conclusion

In this chapter I discussed identification of policy intervention effects in the interference setting. I characterized interpretations of possible interventions and gave criteria for identifying their effects in latent-variable causal chain graph models. Further, I demonstrated estimation via a simulation study. Future directions include exploring the intersection of policies, interference, and game theory, and developing robust

[2] $\mathbf{k}$ is 3-dimensional since $\mathbf{C}_i$ is 3-dimensional.

estimation strategies for this setting.

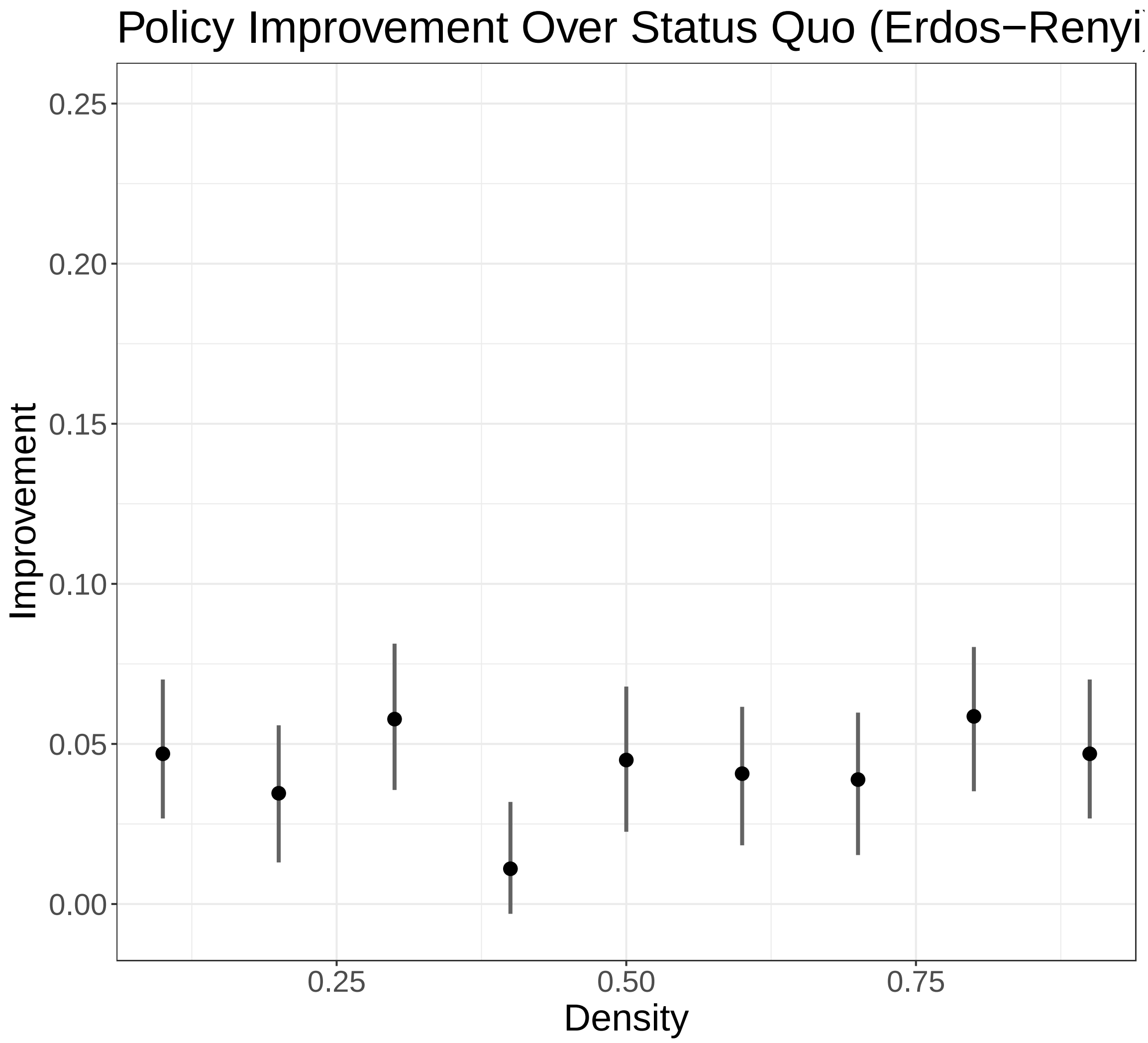


Figure 5-4. Difference in expected outcomes between an optimized strategy and the status quo. We analyze several network densities to demonstrate the generality of this approach.

# Chapter 6

# Intervening on Network Ties

## 6.1 Introduction

As described in preceding chapters, interventions play a foundational role in the observational causal framework. As we have seen, in the classical intervention approaches (e.g., node interventions), researchers select one or more ‘treatment’ variables and outcomes of interest. The value of the outcome of interest is estimated under the hypothetical scenario in which the *value* of the treatment variable is changed to a specific, researcher-chosen value.

In this chapter I will, once again, focus on domains pertaining to networks of interacting study subjects such as infectious disease spread and social networks. As previously discussed, several recent papers have proposed methods for obtaining inferences from dependent data (Tchetgen, Fulcher, and Shpitser, 2017; Sherman and Shpitser, 2018; Ogburn, VanderWeele, et al., 2014; Ogburn et al., 2017). As in the iid setting, these papers emulate RCTs by *intervening on variables* and estimating the effects on downstream outcomes.

Unfortunately, existing methods for network inference, including both those cited above and those presented earlier in this thesis, are ill-suited to consider more general changes to the network. For instance, in urban development economics authors have proposed housing vouchers as a ‘treatment’ to incent families to move to neighborhoods with greater opportunity for upward social mobility (Chetty, Hendren, and Katz, 2016). Evaluating the effect of extracting a family from one neighborhood and placing them in a new neighborhood with new social connections isn't possible by considering changes to values of variables alone: the network itself changes.

In this chapter, which is derived from original research first published in Sherman and Shpitser (2019), I extend the classical causal inference framework to consider changes to *social network structure*. First, I review different network representations in the causal inference literature as well as notions of interventions that have departed from conventional variable interventions. Next, I give a motivating example based on the global political economy. Extending Malinsky (2018), I propose *network interventions*; interventions on the *structure* of a network where ties between units are formed or broken. I define the *individual participant* and *average bystander* effects of these interventions, analogous to the network effects described in Hudgens and Halloran (2008) and discuss identification. I then demonstrate that post-severance distributions satisfy independence constraints for the severed units while remaining minimally KL-divergent from the pre-intervention distributions. Finally, I demonstrate estimation of network intervention effects from observational data via a simulation study.

## 6.2 Review

### Causal Networks

Since networks are of interest to a variety of fields, there are numerous representations, each with their own advantages and limitations. These representations were developed as a means of studying interference – the phenomenon that arises when neighbors' treatments causally affect each other's outcomes. While the present work doesn't focus explicitly on interference, we discuss it here since our work is complementary to that literature.

A widely used approach, characterized in Ogburn, VanderWeele, et al. (2014), represents networks with directed acyclic graphs (DAGs), where network connections appear as directed edges from one individual's variables to another's. This approach lends a natural causal interpretation that follows from a rich literature on causal DAGs. Importantly, the relationships between individuals are encoded in the functional relationships represented by edges connecting different units; when two individuals are not friends, edges will be absent.

Recent work (Tchetgen, Fulcher, and Shpitser, 2017; Sherman and Shpitser, 2018; Ogburn, Shpitser, and Lee, 2018) advocates representing networks with Lauritzen-Wermuth-Freydenburg (LWF) chain graphs (CGs), which were given a causal interpretation in Lauritzen and Richardson (2002) (see also Chapters 4-5). CGs extend DAGs by permitting representation of symmetric relationships (i.e. stable-state equilibria) via undirected edges. Ogburn, Shpitser, and Lee (2018) argued that CGs under the LWF interpretation can approximate feedback processes when those processes are slow. While CGs provide a more general representation, their interpretation in the context of the present work is somewhat complicated.

Beyond these notions, there is a substantial literature on probabilistic relational models (Koller et al., 2007; Friedman et al., 1999). These models generalize conventional graphical models by employing first-order logic to describe the nature of relationships between entities. These models have been extended to causal inference in network settings (Arbour, Garant, and Jensen, 2016), however, similar to chain graphs, their use in public health contexts like those considered here is not yet well established. For these reasons, we will restrict attention to graphical models.

Aside from graphical representations, a large subset of the interference literature formalizes inter-unit relationships algebraically as in Hudgens and Halloran (2008). Many of these formulations could be reformulated using graphical models.

## Structural Interventions

The majority of the causal inference literature has focused on hypothetical experiments wherein interventions are made upon *variables* (e.g. smoking status) and the effects of interventions are considered with respect to some outcome (e.g. lung cancer). The two dominating frameworks, the Neyman-Rubin potential outcomes framework (Neyman, 1923; Rubin, 1974) and Pearl's graph-based framework (Pearl, 2009) differ primarily in their philosophical approach, and recently researchers have begun to use their terminology and mechanics interchangeably. See, for example Richardson and Robins (2013). The causal interpretation of variable interventions under these (and other) frameworks is the subject of literature at the intersection of applied fields and the philosophy of science. These discussions are broad and have a lengthy history. For an incomplete survey, we encourage the interested reader to consult the works of Halpern, Tian, and Pearl (Tian and Pearl, 2001; Halpern and Pearl, 2005), and Woodward (Woodward, 2001).

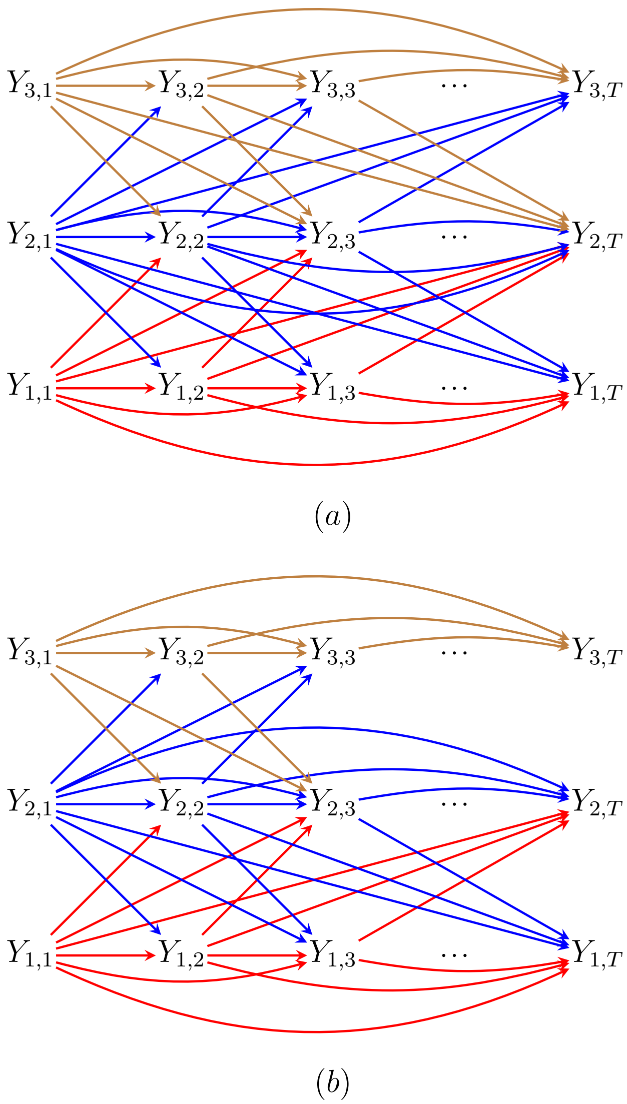
In the past two decades, there has been a movement towards defining more general notions of intervention. Korb proposed several generalizations, including interventions that are stochastic with respect to the treatment variable (Korb et al., 2004). Eberhardt and Scheines discussed similar ideas, contrasting 'hard' and 'soft' interventions, corresponding to changing causal structure (e.g. removing edges) and parametric form respectively (Eberhardt and Scheines, 2007). They proposed using this continuum of interventions to aid in causal discovery efforts (see also Tian and Pearl (2001)). Malinsky proposed a framework for considering the effects of changes to the structure of a causal model on a 'macro level' (Malinsky, 2018). In this framework, one modifies structural equations or manipulates parameters in order to evaluate counterfactuals pertaining to the world in which macro level features are different. Finally, unrelated to philosophy, Ogburn et al. (2017) proposes a type of edge intervention in social networks as a means of understanding changes in network ties. In addition, interventions on paths and edges were considered in the context of mediation analysis in Shpitser and Tchetgen Tchetgen (2016). In the current work, we build on these ideas to evaluate general interventions on network ties, enabling us to envision the counterfactual world in which two units are severed or connected.

## 6.3 Motivating Example: The Political Economy

In this section we give a motivating example: a model of trade relations between countries, in which network interventions provide a means of understanding counterfactual changes to network structure.

In global economics, policies made by one country, such as treaties, trade deals, and tariffs, have a direct impact on the nations geographically and diplomatically connected to the policymaker. In light of current events, we refer to the interventions represented in Fig. 6-1 as the 'Brexit' scenario (for severing two or more countries) and the 'Turkey joins the EU' scenario (for connecting two or more countries). These types of temporal DAG models, also known as dynamic Bayes nets (Murphy and Russell, 2002), correspond to time series cross sectional data from the political economy literature Beck and Katz (2011). Each country $i$ is represented by temporally sequential observations $Y_{i,1}, Y_{i,2}, \dots Y_{i,T}$ where each $Y$ is a vector of economic variables (GDP, unemployment rate, open-market funds rate, etc.).

As a generalization of Fig. 6-1, we can imagine the network having several countries, each with multiple neighbors. We can then consider the hypothetical effect of a 'clean break' at time $t$ between one country and some or all of its neighbors (introducing non-stationarity (Robinson and Hartemink, 2009)). This is represented by moving from Fig. 6-1 (a) to Fig. 6-1 (b) by severing the connection between countries 2 and 3 at $t = 3$. We can also consider the reverse intervention, where two previously unconnected countries are connected, corresponding to the signing of a trade agreement.



$(a)$

$(b)$

Figure 6-1. (a) A DAG representing time series cross sectional data on three countries where country 2 has a trade agreement with countries 1 and 3; (b) the DAG in (a) after an intervention is performed, severing the alliance between countries 2 and 3 at $t = 3$.

Graphically, this corresponds to moving from Fig. 6-1 (b) to (a). Using the framework we propose in this paper, a decision-maker could evaluate these hypothetical policies *prior to* implementation and ensure that they have the intended effect.

## 6.4 Representing Networks with Causal DAGs

Throughout this chapter, I will consider performing causal inference in social networks represented by *DAGs*. In this section I formalize the notation needed to define network interventions and their associated effects. As in previous chapters, I will not fully review all of the relevant concepts and notations. Instead, I will highlight new concepts and notation needed for this work and re-emphasize previously discussed concepts where appropriate.

### Causal DAG Prerequisites

Recall from previous chapters that a parameter in a model is said to be *identified* if it is expressible as a function of the observed data. In causal DAGs with no unobserved variables, all counterfactual distributions $p(\mathbf{V}(\mathbf{a}))$ are identified by the *g-formula* (Robins, 1986). For notational convenience, in the original version of this work, we expressed the g-formula slightly differently. It nevertheless has an identical meaning and use:

$$p(\{W(\mathbf{a}) : W \in \mathbf{V} \setminus \mathbf{A}\}) = \prod_{W \in \mathbf{V} \setminus \mathbf{A}} p(W \,|\, \mathrm{pa}_{\mathcal{G}}(W))|_{\mathbf{A}=\mathbf{a}}$$

As an example of this formula's application, consider a single-unit version of Fig. 6-2. If we are interested in the effect of setting $A = a$, the interventional distribution $p(\mathbf{V}(a))$ is given by $p(Y|A = a, C)p(C)$.

## Edge Interventions

As we saw in Chapter 3, while classical interventions set a variable to a value, there are often cases where we are interested in how an intervention affects an outcome along multiple pathways, such as the separate effects of smoking, smoke inhalation and nicotine, on a patient's risk of lung cancer. In these cases, it is natural to think of interventions in which we intervene on the treatment node with different values for each edge out of the node. For instance, we might consider setting smoking status to 0 for the sake of the smoke inhalation edge and to a reference value for the nicotine exposure edge, corresponding to having the patient smoke e-cigarettes.

Here, I review some of the formalities of edge interventions since they serve as an important basis for the novel work that follows. For a complete treatment, see Chapter 3.

Formally, for a set of treatment variables $\mathbf{A}$ the set of edges out of $\mathbf{A}$ is denoted by $\alpha$. Interventions are performed with a multiset $\mathfrak{a}_\alpha$ which maps edges to constant values for $A$ or to the natural value of $A$ for each $A \in \mathbf{A}$. As with node interventions, for $\mathbf{A}_\alpha = \{A|(AB)_\rightarrow \in \alpha\}$, where $(AB)_\rightarrow \in \alpha$ signifies that an edge $A \rightarrow B$ is in $\alpha$, edge interventions given by $p(\{W(\mathfrak{a}_\alpha) : W \in \mathbf{V} \setminus \mathbf{A}_\alpha\})$ are identified by the edge g-formula (Shpitser and Tchetgen Tchetgen, 2016) with $\text{pa}_{\mathcal{G}}^{\bar{\alpha}}(V) = \{W|(WV)_\rightarrow \notin \alpha\}$:

$$\prod_{W \in \mathbf{V} \setminus \mathbf{A}_\alpha} p(W|\mathfrak{a}_{(ZW)_\rightarrow \in \alpha}, \text{pa}_{\mathcal{G}}^{\bar{\alpha}}(W)). \tag{6.1}$$

If we again consider a single-unit version of Fig. 6-2, when we intervene with $\mathfrak{a}_\alpha = \{(CA)_\rightarrow = c, (CY)_\rightarrow = c'\}$, the distribution $p(\{W(\mathfrak{a}_\alpha) : W \in \mathbf{V} \setminus \mathbf{A}_\alpha\})$ is given by $p(Y|A, C = c')p(A|C = c)$.

## Stochastic Interventions

Echoing Chapters 3 and 5, as an alternative generalization to classical interventions, we might be interested in customizing treatments according to unit-specific characteristics.

For instance, we might want to choose a cancer patient's chemotherapy regimen according to the specific characteristics of their tumor. Rather than setting treatments to fixed values, we set them to analyst-specified functions of pre-treatment covariates. This type of policy intervention is the subject of the dynamic treatment regime (DTR) literature (Tian, 2008; Shpitser and Sherman, 2018).

Formally, for a set of treatment variables $\mathbf{A}$, the set of pre-treatment covariates we wish to use to set each $A \in \mathbf{A}$ is denoted $\mathbf{C}_A$. Policy interventions entail setting $\mathbf{A}$ to the set of functions $\mathbf{f_A}$, where $f_A \in \mathbf{f_A}$ maps $\mathfrak{X}_{\mathbf{C}_A} \to \mathfrak{X}_A$. Responses to policy interventions, $p(\{W(\mathbf{f_A}) : W \in \mathbf{V} \backslash \mathbf{A}\})$, are identified by the policy g-formula (Shpitser and Sherman, 2018):

$$\prod_{W \in \mathbf{V} \backslash \mathbf{A}} p(W | \{f_A(\mathbf{C}_A) : A \in \mathbf{A} \cap \text{pa}_{\mathcal{G}}(W)\}, \text{pa}_{\mathcal{G}}(W) \setminus \mathbf{A}) \tag{6.2}$$

Continuing with our single-unit example for Fig. 6-2, suppose we are interesting in setting $A$ to a policy that is a function of $C$: $f_A(C)$. Then the counterfactual distribution $p(\mathbf{V}(f_A(C)))$ is given by $p(Y|A = f_A(C), C)p(C)$.

## DAG Representation of Network Data

In this chapter, I will represent networks of interacting agents with DAGs following Ogburn, VanderWeele, et al. (2014). I will assume each network $\mathcal{G}$ is associated with a probability distribution $p(\mathbf{V})$ and that $\mathcal{G}$ has a causal interpretation as described in Chapters 2-3. Denote the set of agents ('units' or 'subjects') in $\mathcal{G}$ by $\mathcal{A}$. $\mathcal{G}$ can be partitioned into sub-graphs $\mathcal{G}_i$ with variables $\mathbf{V}_i \subset \mathbf{V}$ for each agent $i \in \mathcal{A}$. The marginal distribution for agent $i$ is therefore denoted $p(\mathbf{V}_i)$. The notation $-i$ will refer to $\mathcal{A} \setminus i$. Analogously, $\mathcal{G}_{-i}$ denotes the subgraph of $\mathcal{G}$ where $\mathbf{V}_i$ and its associated edges have been removed.

We define the notion of *unit homogeneity*. This assumption has two parts: a) if there exists a unit $i \in \mathcal{A}$ with variable $V_i \in \mathbf{V}_i$, then there is a corresponding $V_k \in \mathbf{V}_k$

for all $k \in \mathcal{A}$ with an analogous interpretation; and b) if there exists a unit $i \in \mathcal{A}$ with variables $V_i, U_i \in \mathbf{V}_i$ such that $V_i \in \text{pa}_{\mathcal{G}_i}(U_i)$, then $V_k \in \text{pa}_{\mathcal{G}_k}(U_k)$ for all $k \in \mathcal{A}$. The first part ensures that units are all of the same 'type' (e.g. all agents have the same demographic variables, and the same outcome variable). The second part ensures that the existence of a relationship between one unit's variables implies the same relationship exists for all other units.

For an example of these definitions, consider Fig. 6-2. Each unit has a variable of each 'type' (e.g. $C, A, Y$) and the connections between variables are the same for each unit (e.g. $C_i \to A_i$ in all units).

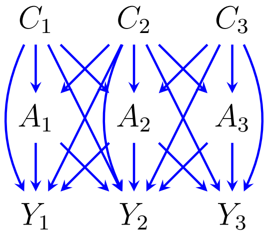


Figure 6-2. A simple social network represented by a DAG. The network exhibits unit homogeneity, symmetric connections, and homogeneous connections.

On the network level, we define the notions of connectedness, symmetry of connections, and homogeneity of connections. Two units $i, j \in \mathcal{A}$, with $i \neq j$, are said to be *connected* if for some $V_i \in \mathbf{V}_i$ and some $U_j \in \mathbf{V}_j$ it is the case that $V_i \in \text{pa}_{\mathcal{G}}(U_j)$. The connection between $i$ and $j$ is said to be *symmetric* if the vice-versa relationship holds. That is, if $i$ and $j$ are connected and the connection is symmetric then for all $V_i \in \text{pa}_{\mathcal{G}}(U_j)$, we have $V_j \in \text{pa}_{\mathcal{G}}(U_i)$, where $V_i$ is analogous to $V_j$ and $U_i$ is analogous to $U_j$. The set of units connected to unit $i$ in $\mathcal{G}$, also referred to as $i$'s *neighbors*, will be denoted $\mathcal{N}_{\mathcal{G}}(i)$. Finally, we define homogeneity of connections, which ensures that the relationships across the network are similar. If $i$ and $j$ are connected and there is an edge between some $V_i \in \mathbf{V}_i$ and some $V_j \in \mathbf{V}_j$ then network connections are *homogeneous* if for *all* connected units $k, l$ in the network, an edge is present between the analogous $V_k \in \mathbf{V}_k$ and $V_l \in \mathbf{V}_l$.

We further define homogeneity of functional form which strengthens the notion of homogeneity for connections by imposing that, for any pair of connected nodes, the marginal distribution with respect to those two nodes is the same as the marginal distribution for any other pair of connected nodes (e.g. $p(\mathbf{V}_i, \mathbf{V}_j) = p(\mathbf{V}_k, \mathbf{V}_l)$ for all $i \neq j$ and $k \neq l$)). Under this assumption, pairwise relationships between units are the same, regardless of the type of unit. This assumption is reasonable in certain applied contexts, such as infectious disease spread, which is governed by a process that operates in the same way for any unit in the population.

For an example of these definitions, once again consider Fig. 6-2. Connections are symmetric (e.g. $C_1 \rightarrow A_2$ and $C_2 \rightarrow A_1$) and homogeneous (e.g. $C_1 \rightarrow A_2$ and likewise $C_3 \rightarrow A_2$).

## 6.5 Network Interventions

In this section we introduce the notion of *network interventions* where we intervene on the *structure* of a network by adding or removing edges, changing relationships between units. We define effects of these interventions and give identification criteria in §6.6, describe appealing properties of certain network interventions with respect to KL-divergence in §6.7, and discuss estimation in §6.8.

### Severance Interventions

We will call interventions in which we sever two individuals in a network 'severance interventions'. For a graph $\mathcal{G}$ with pre-intervention distribution $p(\mathbf{V})$, where $\mathbf{V}$ is partitioned by $\{\mathbf{V}_i | i \in \mathcal{A}\}$, we denote the intervention severing units $i$ and $j$ by $i \bowtie j$. Graphically, this corresponds to removing *all* edges between $\mathbf{V}_i$ and $\mathbf{V}_j$, yielding the graph $\mathcal{G}_{i \bowtie j}$. We will define responses to severances with respect to individual units (e.g. $p(V_i(i \bowtie j))$. The joint response is simply the joint distribution over these counterfactuals.

We propose two different types of severance intervention. Each formulation has a corresponding causal interpretation and one could use either formulation depending on the application.

The first formulation, which we will call 'value-based' severance and is closely tied to classical mediation analysis, generalizes edge interventions (Shpitser and Tchetgen Tchetgen, 2016) to networks. We intervene on variables in an edge-specific manner, replacing cross-unit edges into a unit, say $i$, with synthetic edges into $i$ that represent fixed relationships no longer dependent on variables in the previously connected unit.

For $\mathbf{V}_i \subset \mathbf{V}$, let $\mathbf{A}_{\mathbf{V}_i} = \text{pa}_{\mathcal{G}_j}(\mathbf{V}_i)$, the parents of $\mathbf{V}_i$ in $\mathbf{V}_j$. We consider setting $\mathbf{A}_{\mathbf{V}_i}$ to $\mathbf{a}_{\mathbf{V}_i}$ for the sake of edges from $\mathbf{A}_{\mathbf{V}_i}$ to $\mathbf{V}_i$. All other edges out of $\mathbf{A}_{V_i}$ maintain the observed values of their source node so that for all $V_j \in \mathbf{V} \setminus \mathbf{V}_i$, $V_j$'s pre- and post-intervention distributions are the same. Since the intervention values are constant, $i$ and $j$ are no longer connected. Returning to our diplomacy example, one might choose $\mathbf{a}_{\mathbf{V}_i}$ to be a reference value in the network, such as network averages of economic variables. Formally, $p(V_i(i \bowtie j; \mathbf{a}_{\mathbf{V}_i})) =$

$$p\Big(V_i(\mathbf{A}_{\mathbf{V}_i} = \mathbf{a}_{\mathbf{V}_i}, \{V_j(\mathbf{a}_{\mathbf{V}_i}) : V_j \in \text{pa}_{\mathcal{G}_{-j}}(V_i)\})\Big)$$

The second formulation, which we call 'stochastic' severance, entails marginalizing out the parents from the severed unit. We phrase these as policy interventions.

Consider $V_i \in \mathbf{V}_i$ and let $\mathbf{A} = \{A \in \mathbf{V}_i | \, \text{pa}_{\mathcal{G}_j}(A) \neq \emptyset\}$ (i.e. $\mathbf{A}$ is the set of unit-$i$ variables with parents in unit $j$). The counterfactual $p(V_i(i \bowtie_{\mathbf{f_A}} j))$ corresponds to selecting a set of stochastic policies $\mathbf{f_A}$ where each $f_A$ is *unit-structure preserving* (see below). The counterfactual is given by the recursive formula:

$$p(V_i(i \bowtie_{\mathbf{f_A}} j)) = f_{V_i}(\{C(\mathbf{f_A}) : C \in \text{pa}_{\mathcal{G}}(A) \setminus \text{pa}_{\mathcal{G}_j}(A)\})$$

For a policy $f_{V_i}$ to be unit-structure preserving, $p(V_i|\operatorname{pa}_{\mathcal{G}_{-j}}(V_i))$ must be the same in the pre- and post-intervention distributions. This ensures that unit $i$'s causal structure is maintained. Formally, $f_{V_i}(\{W : W \in \operatorname{pa}_{\mathcal{G}_{-j}}(V_i)\}) =$

$$\int_{\operatorname{pa}_j(V_i)} p(V_i|\operatorname{pa}_{\mathcal{G}}(V_i)) d(\operatorname{pa}_j(V_i)),$$

where $\operatorname{pa}_j(V_i) = \operatorname{pa}_{\mathcal{G}}(V_i) \cap \mathbf{V}_j$.

We will argue in §6.7 that post-severance distributions are minimally KL-divergent from $p(\mathbf{V})$ among the class of distributions corresponding to the DAG with reduced edge set. Specifically, this holds for value-based severances if, instead of fixing values, we allow the source nodes of edge interventions to vary and average over those nodes. Likewise, for stochastic severances, the KL result holds if we pick $f_{V_i}$ such that $V_i$ and it's remaining parents $\operatorname{pa}_{\mathcal{G}_{i \bowtie j}}(V_i)$ have a particular relationship.

## Connection Interventions

We will call interventions in which we adjoin two previously unconnected individuals in a network *connection interventions*. We will denote the intervention where units $i$ and $j$ are joined by $i \diamond j$. Graphically, this corresponds to inserting one or more edges from $\mathcal{G}_i$ to $\mathcal{G}_j$ or vice-versa, yielding $\mathcal{G}_{i \diamond j}$. As before, we will define responses to connection interventions with respect to individual units (e.g. $p(V_i(i \diamond j))$. The joint response $p(\mathbf{V}(i \diamond j))$ is simply the joint distribution over these counterfactual variables. We describe three separate and increasingly general formulations of connection interventions.

### Interventions Under Functional Form Homogeneity

If we assume that the functional forms of network ties are homogeneous, and further assume that each structural equation in the network aggregates arbitrarily many inputs, then the new structural equation for each variable is determined by the equations for the analogous variables in the network.

We might be interested in counterfactual situations that are not present in the observed data, such as the case when connecting two units results in one unit having more neighbors than any unit in the observed data. Because we assume homogeneity of functional form, we can only allow for classes of policies that can flexibly handle an arbitrary number of neighbor nodes.

For the intervention to be well-defined, we must have $f_V \in \mathcal{F}$, where $\mathcal{F}$ is a class of aggregator functions of the form $f(h_U(U_1, U_2, \ldots), h_W(W_1, W_2, \ldots), \ldots)$. Each $h_Z$ maps $\mathcal{Z} \to \mathbb{R}$ where $\mathcal{Z}$ is an arbitrary-sized multiset of $Z$-type variables. In turn, $f$ maps $\mathcal{H} \to \mathbb{R}$ where $\mathcal{H}$ is the arbitrary-sized multiset of outputs from the $h$ functions. For instance, if $V_i$ has parents of types $\mathbf{U}, \mathbf{W} \subset \mathbf{V}$, we might select $h_U$ to output the mean of the $U$'s, $h_W$ to output the median of the $W$'s, and $f$ to output the sum of those two values.

Suppose $i$ and $m$, and $i$ and $k$ are connected in $\mathcal{G}$. Then under functional form homogeneity, the relationships between $V_i \in \mathbf{V}_i$ and $U_m \in \text{pa}_{\mathcal{G}_m}(V_i)$ and $V_i$ and the analogous $U_k \in \text{pa}_{\mathcal{G}_k}(V_i)$ are governed by a function $f_V$. In the post-intervention distribution, $p(V_i(i \diamond_{f_V} j))$, where units $i$ and $j$ are connected, the relationship between $V_i$ and $U_j$ is also governed by $f_V$. The associated counterfactual is given by:

$$p(V_i(i \diamond_{f_V} j)) = p(V_i = f_V(\{V(i \diamond j) : V \in \text{pa}_{\mathcal{G}_{i \diamond j}}(V_i)\}))$$

### Intervening With Known Policies

We can relax the assumption of homogeneous network ties by intervening with a known functional form. As with the previous formulation, the analyst is interested in understanding the effect of inducing a specific relationship. Continuing our diplomacy example from §6.3, consider Turkey as a candidate for EU membership. Since Turkey has a large, robust economy, it may be able to negotiate a more favorable entrance with specific parameters, similar to Switzerland's non-member bilateral treaties. This formulation represents the inverse operation of function form-based severance interventions.

We wish to evaluate the effect of connecting units $i$ and $j$ with a known induced relationship. In the pre-intervention distribution, $V_i \in \mathbf{V}_i$ is determined by $f_{V_i}(\text{pa}_{\mathcal{G}}(V_i), \epsilon_{V_i}) \in \mathcal{F}_{V_i}$. For the intervention to be valid, the analyst must specify $f'_{V_i} \in \mathcal{F}'_{V_i}$ where $\mathcal{F}'_{V_i}$ is a family of unit-structure preserving functions. The counterfactual is defined as:

$$p(V_i(i \diamond_{f'_{V_i}} j)) = p(V_i = f'_{V_i}(\{V(i \diamond j) : V \in \text{pa}_{\mathcal{G}_{i\diamond j}}(V_i)\}))$$

In this context, the notion of a unit-structure preserving policy is the same as before, however for notational clarity we define $\mathbf{S} = \text{pa}_{\mathcal{G}_{i\diamond j}}(V_i) \setminus \text{pa}_{\mathcal{G}}(V_i)$, $V_i$'s new parents in the post-intervention graph, and rephrase the definition as:

$$p(V_i | \text{pa}_{\mathcal{G}}(V_i)) = \int_{\mathbf{S}} f'_{V_i}(\text{pa}_{\mathcal{G}_{i\diamond j}}(V_i)) p(\mathbf{S}) d\mathbf{S} \tag{6.3}$$

### Intervening with Unknown Policies

In the most general formulation, we do not assume the analyst knows the interventional policy in advance. Instead, we formalize a procedure for picking an optimal policy to govern the relationship between connected units subject to some known constraints. In the example where we consider Turkey joining the EU, this corresponds to the EU and Turkey negotiating a treaty that jointly optimizes their outcomes (e.g. mean per-capita GDP).

Building on the preceding subsection, we can simply express this type of intervention as an optimization on some jointly defined criterion, such as utility, within a class of policies. Let $\mathcal{F}'_{V_i}$ and $\mathcal{F}'_{V_j}$ be families of unit-structure preserving candidate policies for $V_i$ and $V_j$. Let $\mathcal{C}$ be a known set of constraints that the solution must satisfy (e.g. Turkey cannot trade away more natural resources than it has). Let $g((V_i, V_j)(f_{V_i}, f_{V_j}))$ be a known function that captures the joint outcome for units $i$ and $j$ under a given pair of $f$'s. Then the optimal $f$'s are given by:

$$\operatorname*{arg\,max}_{f_{V_i} \in \mathcal{F}'_{V_i}, f_{V_j} \in \mathcal{F}'_{V_j}} E[g(V_i, V_j)(f_{V_i}, f_{V_j})] \text{ subject to } \mathcal{C}$$

Solving this optimization corresponds to evaluating $p(\mathbf{V}(i \diamond j)(f_{V_i}, f_{V_j}))$ for each pair of candidate $f$'s that satisfy $\mathcal{C}$ and picking the best pair.

## 6.6 Effects and Identification of Network Interventions

Hudgens and Halloran (2008) defined the direct, spillover, and network average effects for interference settings. Respectively, these correspond to the effect on unit $i$'s outcome when $i$'s treatment is modified, the effect on $i$'s outcome when $i$'s neighbor's treatment is modified, and the average effect on all units' when someone's treatment is modified (i.e. the sum of the direct and spillover effects). Since these effects are defined for a particular type of node intervention, it is necessary to define analogous effects for network interventions.

We define two new effects: the individual participant effect (IPE), and the average bystander effect (ABE). The IPE is defined for units $i$ and $j$ when they are the subjects of a network intervention. $IPE_i$ is the contrast between $i$'s observed and interventional outcomes. For severances (with connections defined analogously), this contrast is given by $IPE_i(i \bowtie j) = Y_i - E[Y_i(i \bowtie j)]$. We can also define the average participant effect (APE) as the mean of $IPE_i$ and $IPE_j$.

The ABE captures the contrast for units not directly involved in a network intervention. By the Markov property of DAGs, for a network intervention on $i$ and $j$, the ABE is non-trivial for $i$ and $j$'s pre-intervention neighbors $\mathcal{N}_{\mathcal{G}}(i) \cup \mathcal{N}_{\mathcal{G}}(j) \setminus \{i, j\}$ (e.g. the other countries $i$ and $j$ have treaties with). For severances, $ABE(i \bowtie j) =$

$$\frac{1}{|(\mathcal{N}_i \cup \mathcal{N}_j) \setminus \{i, j\}|} \sum_{k \in (\mathcal{N}_i \cup \mathcal{N}_j) \setminus \{i, j\}} Y_k - E[Y_k(i \bowtie j)]$$

Connections are defined analogously. Following Hudgens and Halloran (2008), the average effect on the network (e.g. the effect on the 'global' economy) is the sum of APE and ABE.

## Identification

For a given intervention type, if the IPE is identified then the ABE is also identified and vice versa. We therefore focus on the criteria for identification of each type of intervention we've discussed.

Under our setup, value-based severance interventions are the network analogue of edge interventions in mediation settings. For a severance of units $i$ and $j$, let $\alpha$ be the set of edges out of $\text{pa}_{\mathcal{G}}(\mathbf{V}_i) \cap \mathbf{V}_j$. If $\mathfrak{a}_\alpha$ specifies a constant value for each edge $\mathbf{V}_j \to \mathbf{V}_i$ and that the source nodes for all other edges in $\alpha$ are random, then $p(\mathbf{V}_i(i \bowtie j; \mathfrak{a}_\alpha))$ is identified by the edge g-formula (Eq. 6.1).

For instance, in Fig. 6-2, if we are interested in the effect on $Y_2$ of severing units 2 and 3 by setting $\mathbf{A}_{V_2} = \mathbf{a}_{V_2} = \{C_3 = c_3, A_3 = a_3\}$ for the sake of the edges $(C_3Y_2)_\to$ and $(A_3Y_2)_\to$, then:

$$
\begin{aligned}
&p(V_2(2 \bowtie 3); \mathbf{a}_{\mathbf{V}_3}) = \\
&\qquad p(Y_2|A_1, A_2, C_1, C_2, A_3 = a_3, C_3 = c_3) \\
&\qquad \times p(A_1|C_1, C_2)p(A_2|C_1, C_2, C_3 = c_3)p(C_1)p(C_2)
\end{aligned}
$$

The other interventions we define entail a change in the functional form of the variables of interest. Suppose we wish to join units $i$ and $j$ with $\mathbf{A} = \{V \in \mathbf{V}_i | \, \text{pa}_{\mathcal{G}_{i\diamond j}}(V) \cap \mathbf{V}_j \neq \emptyset\}$ and $\mathbf{C}_A = \{V \in \mathbf{V}_j | A \in \text{ch}_{\mathcal{G}_{i\diamond j}}(V)\}$. Then, if $\mathbf{f_A}$ are all functions that either satisfying the aggregator properties for the homogeneous case, or are unit-structure preserving for the non-homogeneous case, the counterfactual $p(\mathbf{V}_i(i \diamond_{\mathbf{f_A}} j))$ is identified by the policy g-formula (Eq. 6.2). Stochastic severances (e.g. of units $i$ and $j$) are also identified under our setup, with $\mathbf{A} = \{A \in \mathbf{V}_i | \, \text{pa}_{\mathcal{G}_j}(A) \neq \emptyset\}$ and

$\mathbf{C}_A = \mathrm{pa}_{\mathcal{G}}(A) \setminus \mathrm{pa}_{\mathcal{G}_j}(A)$ for each $A \in \mathbf{A}$ and $\mathbf{f}_{\mathbf{A}}$ satisfying unit-structure preservation for each $f_A$. For homogeneous connections, we must also estimate the parameters of each aggregator function ($h_V, h_W$, etc.) from observed data. These are identified by maximum likelihood from $\mathcal{G}$

As an example, if we are interested in performing a stochastic severance on units 2 and 3 in Fig. 6-2, suppose we set $f_{V_i}(\mathrm{pa}_{\mathcal{G}_{2\bowtie 3}}(V_i)) = p(V_i | \mathrm{pa}_{\mathcal{G}}(V_i) \setminus \mathrm{pa}_{\mathcal{G}_3}(V_i))$ for each $V_i \in \mathbf{V}_i$. Then the identifying functional for the effect on $\mathbf{V}_2$ is given by:

$$\begin{aligned} p(\mathbf{V}_2(2 \bowtie_{f'_{V_i}} 3)) &= p(Y_2 | A_1, A_2, C_1, C_2) \\ &\times p(A_1 | C_1, C_2) p(A_2 | C_1, C_2) p(C_1) p(C_2) \end{aligned}$$

## Latent-Variable Network Interventions

Throughout this chapter, I have assumed that data is representable by a DAG where all variables are *observed*. We can relax this assumption to allow for models in which some variables are latent. In these cases, the interpretation of the proposed interventions remains the same, however, identification conditions will be modified slightly.

Consider a latent-variable DAG $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$ with $\mathbf{V}$ observed and $\mathbf{H}$ hidden. From $\mathcal{G}$, we can obtain a acyclic directed mixed graph (ADMG) $\mathcal{G}'(\mathbf{V})$ via a latent projection operation (Richardson et al., 2017). $\mathcal{G}'$ represents an equivalence class of graphs that share the same observed variables and set of independence constraints (Richardson et al., 2017).

Identification of network interventions in an ADMG $\mathcal{G}'$ relies on the assumptions described in the previous section, existing non-parametric identification theory for ADMGs, and the requirement that the network intervention operates only on edges that are present in both $\mathcal{G}$ and $\mathcal{G}'$. As pointed out previously, value-based severances in DAGs can be identified by the edge g-formula. Under the relaxation allowing for latent variables, value-based severances are instead identifiable according to a version of the ID algorithm adapted to edge interventions, proven sound and complete in Shpitser and Sherman (2018). Likewise, for stochastic severances and for connection interventions, if the identification conditions described in the previous sub-section hold, then the respective interventions are identifiable according to a version of the ID algorithm adapted to policy interventions, proven sound and complete in Shpitser and Sherman (2018)).

## 6.7 Optimal Choice of Post-Severance Distribution

In this section we prove a series of results regarding the KL-divergence from a distribution $p(\mathbf{V})$, corresponding to a known DAG $\mathcal{G}$, to another distribution $\tilde{p}(\mathbf{V})$, corresponding to a DAG in which edges have been removed. The results demonstrate that the KL-divergence from $p$ to $\tilde{p}$ is minimized when $\tilde{p}$ takes on a form similar to the g-formula (Robins, 1986). These *probabilistic* results help justify the g-formula and edge g-formula as intuitive tools for analyzing causal queries in DAGs. Moreover, these results motivate the manner in which we perform severances.

The first result demonstrates that when removing edges between a node $A$ and its parents, a simple modification to the factorization of $\mathcal{G}$, removing $A$'s parents from the term for $A$ yields the KL-minimal distribution satisfying the independence constraints implied by the severance.

**Theorem 11** *Let $\mathbf{V}$ be a set of random variables with $p(\mathbf{V})$ corresponding to a DAG $\mathcal{G}$. Let $A \in \mathbf{V}$. Let $\mathcal{P}(\mathbf{V})$ be the set of probability distributions that factorize according to $\mathcal{G}$. Then*

$$p(A) \prod_{V \in \mathbf{V} \setminus A} p(V \,|\, \mathrm{pa}_{\mathcal{G}}(V)) = \underset{\tilde{p} \in \mathcal{P}(\mathbf{V})}{\arg\min} D_{KL}(p || \tilde{p}) \quad s.t.\ A \perp\!\!\!\perp \mathrm{pa}_{\mathcal{G}}(A)$$

The second result generalizes the first by allowing for edge removal between $A$ and a *subset* of its parents.

**Theorem 12** *Let* $\mathbf{V}$ *be a set of random variables with* $p(\mathbf{V})$ *corresponding to a DAG* $\mathcal{G}$*. Let* $A \in \mathbf{V}$ *and* $\mathbf{B} \subseteq \mathbf{V}$ *such that* $\mathbf{B} \subseteq \mathrm{pa}_{\mathcal{G}}(A)$*. Let* $\mathcal{P}(\mathbf{V})$ *be the set of probability distributions that factorize according to* $\mathcal{G}$*. Then*

$$\begin{aligned} &p(A|\,\mathrm{pa}_{\mathcal{G}}(A) \setminus \mathbf{B}) \prod_{V \in \mathbf{V} \setminus A} p(V|\,\mathrm{pa}_{\mathcal{G}}(V)) \\ &\quad = \operatorname*{arg\,min}_{\tilde{p} \in \mathcal{P}(\mathbf{V})} D_{KL}(p||\tilde{p}) \text{ s.t. } A \perp\!\!\!\perp \mathbf{B}|\,\mathrm{pa}_{\mathcal{G}}(A) \setminus \mathbf{B} \end{aligned}$$

The following result generalizes the previous theorem to allow for removal of *any* set of edges in $\mathcal{G}$. This result corresponds to directly to severance interventions. If we remove the dependence of each variable on the parents for which we remove edges, and otherwise keep the variable functionally consistent with its original structural equation, the result is the minimally KL-divergent distribution from the original distribution that reflects the severance.

**Theorem 13** *Let* $\mathbf{V}$ *be a set of random variables with* $p(\mathbf{V})$ *corresponding to a DAG* $\mathcal{G}$*. Let* $\mathbf{A} \in \mathbf{V}$ *and for each* $A \in \mathbf{A}$ *define* $\mathrm{In}(A) \subseteq \mathrm{pa}_{\mathcal{G}}(A)$*, the set of parents of* $A$ *whose edges into* $A$ *we wish to remove. Let* $\mathcal{P}(\mathbf{V})$ *be the set of probability distributions that factorize according to* $\mathcal{G}$*. Then*

$$\begin{aligned} &\prod_{A \in \mathbf{A}} p(A|\,\mathrm{pa}_{\mathcal{G}}(A) \setminus \mathrm{In}(A)) \prod_{V \in \mathbf{V} \setminus \mathbf{A}} p(V|\,\mathrm{pa}_{\mathcal{G}}(V)) \\ &\quad = \operatorname*{arg\,min}_{\tilde{p} \in \mathcal{P}(\mathbf{V})} D_{KL}(p||\tilde{p}) \text{ s.t. } A \perp\!\!\!\perp \mathrm{In}(A)|\,\mathrm{pa}_{\mathcal{G}}(A) \setminus \mathrm{In}(A) \;\; \forall A \in \mathbf{A} \end{aligned}$$

The final two results are corollaries of Thm. 13 and are closely related to classical causal inference. The first corresponds to variable interventions where we fix some $\mathbf{A} \subseteq \mathbf{V}$ to a value $\mathbf{a}$. The KL-closest distribution to $p(\mathbf{V})$ is given by the g-formula, where terms for each $A \in \mathbf{A}$ are removed and variables with parents in $\mathbf{A}$ are evaluated with those parents set to $\mathbf{a}$.

**Theorem 14** *Let* $\mathbf{V}$ *be a set of random variables with* $p(\mathbf{V})$ *corresponding to a DAG* $\mathcal{G}$*. Let* $\mathbf{A} \subseteq \mathbf{V}$ *and assume that for some* $\mathbf{a}$ *we have* $p(\mathbf{A} = \mathbf{a}) > 0$*. Let* $\mathcal{P}(\mathbf{V})$ *be the*

| 95% Confidence Intervals of Bias | | | |
|---|---|---|---|
| *Intervention* | *Erdős-Rényi* | *Barabasi-Albert* | *Watts-Strogatz* |
| Homogeneous Connection | (-.0049, .0020) | (-.0021, .0006) | (-.0024, .0010) |
| Known Connection | (-.0014, .0010) | (-.0004, .0016) | (-.0018, .0020) |
| Unknown Connection | (-.0035, .0025) | (-.0134, .0124) | (-.0280, .0093) |
| Stochastic Severance | (-.0015, .0043) | (-.0096, .0066) | (-.0032, .0020) |
| Value Severance | (-.0088, .0112) | (-.0010, .0020) | (-.0048, .0016) |

Table 6-I. 95% confidence intervals for the bias of estimates of each type of network intervention.

*set of probability distributions that factorize according to $\mathcal{G}$. Then*

$$\begin{aligned}&\prod_{V \in \mathbf{V} \setminus \mathbf{A}} p(V \,|\, \mathrm{pa}_{\mathcal{G}}(V))|_{\mathbf{A}=\mathbf{a}} \\ &\qquad = \arg\min_{\tilde{p} \in \mathcal{P}(\mathbf{V})} D_{KL}(p || \tilde{p}) \;\; s.t. \\ &\qquad\qquad \tilde{p}(A_i \,|\, \mathrm{nd}_{\mathcal{G}}(A_i)) = I(A_i = a_i) \;\; \textit{for } i = \{1, \ldots, |\mathbf{A}|\}\end{aligned}$$

The final result, which can be found in Appendix F, generalizes the above theorem to edge interventions (Shpitser and Tchetgen Tchetgen, 2016). This result corresponds to the value-based formulation of severances. When we fix a set of edges to constant values, the resulting distribution is given by the edge g-formula and is the KL-closest distribution to the pre-intervention distribution that reflects the fact that those edges have been fixed.

## 6.8 Experiments

I now describe a set of simulation studies which demonstrate the feasibility of obtaining unbiased estimates of the effects of network interventions. In these experiments we assume *partial interference*: we observe $M$ samples of a network, each with $N$ units. While we do not consider full interference scenarios, in which the analyst has access to only a single sample of the network, similar results could be obtained in that setting using the auto-g-computation algorithm (Tchetgen, Fulcher, and Shpitser, 2017). We also assume that all pre-intervention networks satisfy symmetry of connections, and homogeneity of units, connections, and functional form.

We consider a social network graph resembling Fig. 6-2 where all variables $C, A$, and $Y$ are binary. In four separate experiments we demonstrate estimation across varying social network generators, varied attachment probabilities for the Erdős-Rényi generator, varied network sizes, and varied sample sizes. For the latter three experiments we restrict attention to the stochastic severance intervention. For each unit $i$ we generate values for $\mathbf{V}_i$ according to log-linear models with parameters $\tau_C, \tau_A, \tau_Y$. For the detailed setup, please see Appendix F.

For each experiment we estimate the average IPE by separately applying the intervention to each unit in the network. For severances we remove the connection between the unit of interest and it's highest degree neighbor while for connections we connect the unit to it's highest degree *non*-neighbor.

## Estimation and Evaluations

For each experiment we first fit models for each variable type given it's parents via MLE where features for neighbor variables are sums of those variables. We estimate values of endogenous nodes using Monte Carlo sampling using these fit models and exogenous nodes via the empirical distribution. We estimate values in the pre- and post-intervention worlds and report the mean difference between these estimates across all units and all samples of the network. For specific details on the mechanics of each intervention type, please see Appendix F.

To evaluate the performance of this estimation technique, we generate 'ground truth' graphs corresponding to the result of each intervention and generate values for the $Y_i$'s of interest. For each simulated network we generated 1000 bootstrap samples. We compare the intervention effects to the ground truth effects and obtain the bias of our approach. As presented in Tables 6-I, and F-II - F-IV, the 95% confidence interval for each experiment covers the ground truth bias and thus shows that the effects of network interventions can be consistently estimated.

## 6.9 Discussion

In this chapter, I proposed a framework for intervening on the *structure* of a social network graph by severing or creating connections between subjects. I defined effects that extend the network effects defined in Hudgens and Halloran (2008). I then proved that for severances, and causal interventions generally, the g-formula and edge g-formula obtain distributions that are minimally KL-divergent from the pre-intervention distribution subject to the independence constraint imposed by the intervention. Finally, I demonstrated that these effects can be estimated from observational data via a simulation study.

In the future, this framework could be generalized to chain graph models to allow for more flexibility of network representation.

# Chapter 7

# Conclusion

This thesis has pushed forward the field of observational causal inference in several key areas. In this final chapter, I will summarize the specific contributions I've made and talk about their broader impacts and potential impacts. I will conclude with a discussion of future research directions that follow from this work.

## 7.1 Summary

While each chapter in this work was derived from distinct research projects, each of which has previously been the subject of a publication, there are several themes that span most, if not all of these works. I've proposed novel approaches to representing causal models, posited and proven theoretical results describing when queries are or are not estimable from observed data (i.e., identifiable), and expanded the notion of an intervention. Each of these stands as an important contribution to the area academically, and, hopefully, will soon pay practical dividends as well.

### 7.1.1 Identification Theory

The chief contribution of this thesis are the identification theory results. In Chapters 3-5, I presented a series of identification results that generalize the seminal ID algorithm (Tian and Pearl, 2002; Shpitser and Pearl, 2006; Huang and Valtorta, 2006; Richardson et al., 2017) to a broader class of settings. The results in Chapter 3 are perhaps the most consequential, as they relate directly to reinforcement learning and sequential decision making problems, which have attracted great attention in recent years. In light of the recent popularity of algorithmic advertising, the results on identification in non-iid and network settings in Chapters 4 and 5 are also of significant interest in the broader community (Nabi et al., 2022)

### 7.1.2 Representation of Causal Models and Counterfactual Queries

The fundamental framework I relied upon to carry out this work was primarily formulated elsewhere. Nevertheless, my work added to this framework. In Chapter 4, I expounded on a novel use of chain graph models (including latent variable chain graphs) to represent social networking dynamics and other non-iid settings. This representation is more parsimonious for situations where there are symmetric or equilibrium relationships in the observed data in addition to causal relationships which are symmetric. As highlighted in Chapter 5, this modeling paradigm is also useful in situations where simultaneous decisions are being made, such as in multi-agent reinforcement learning. Since publication, this modeling approach has also been studied in Ogburn, Shpitser, and Lee (2018).

### 7.1.3 Varieties of Interventions

Closely related to the issues of how we represent causal models and what is estimable from those models is the question 'what can we ask of our causal model?' In multiple chapters in this thesis, I sought to answer that question. In Chapters 3 and 5, I gave novel algorithms for the identification of causal effects associated with stochastic and policy interventions. Policy interventions in the way I've conceived permit the analysis of a more general class of hypothetical experiments than those represented by classical interventions which set variables to constant values. In practice, this form of intervention could allow applied researchers to perform more realistic analyses of the effects, or efficacy, of policies they are considering adopting. This thesis (nor the papers it is derived from) is not the first to posit this type of intervention (Eberhardt and Scheines, 2007; Tian, 2008). Nevertheless, this work expands the utility of that line of research by characterizing when and how responses to such interventions can be estimated.

In addition to stochastic interventions, the structural interventions considered in Chapter 6 are relatively novel. This work is stands much closer to the philosophy of causation than the other works presented in this thesis. As such the chapter has a different flavor: rather than characterizing, formally, what is possible, I sought to both pose and answer a 'what if?' question. Ultimately, the answer I provided only scratched the surface, as that work has since inspired several further inquiries into the particulars of structural interventions in the presence of non-iid data (Subbaswamy, Chen, and Saria, 2019; Witty et al., 2019; Galhotra et al., 2022).

### 7.1.4 Other Contributions

While the body chapters of this thesis cover a coherent and sequential (though the publication order of the original papers does not match the ordering of the chapters) story of causal modeling in the dependent data setting, I'd like to momentarily highlight the other contributions I made during my PhD studies. These are the subject of the appendix chapters which follow.

The first appendix chapter is derived from an applied paper on the topic of mental health surveillance from social media data. In recent years, social media-derived surveillance tools have come into relatively widespread use. In this chapter I sought to answer the question of whether demographic characteristics (specifically gender) can confound predictions of mental health status (i.e., depression). This is inherently a causal question and so it closely relates to the work from the main body of the thesis. Unfortunately, as in most real world settings, this question was not *truly* answerable from observed data because of the presence of too many latent confounders. I did, however, use predictive importance of social media terms different users used in their posts as a proxy and found that the answer is ‘yes’. In particular, the performance of models designed to evaluate whether an individual is or is not depressed differs depending on gender. This finding can be used to help tailor models to users based on demographics so as to provide better population-level mental health care.

The second appendix, derived from a paper originally published in the Annals of Thoracic Surgery, concerns the development of a tool for predicting the risk of readmission to the hospital following cardiac surgery. The model, fit from electronic medical record data and a curated data source, dramatically outperformed the state of the art for the task. This work’s contribution, however, was more in the vein of communication: we sought to demonstrate to a *clinical* audience the feasibility and importance of using modern machine learning techniques to develop risk prediction models. From a (pseudo-) causal perspective, we were able to derive some useful clinical insight: we analyzed the Shapley values of our top-performing model to identify that several *post*-discharge variables are highly important for predicting readmission. This finding has since spurred discussions (internal to our collaboration team) about the possibility of running an in-clinic follow-up study to evaluate the real-world importance of the post-discharge phase as it relates to readmission.

## 7.2 Discussion

### 7.2.1 Limitations

As I close, it seems appropriate to also discuss some of the limitations of this work.

As in any theory-grounded research area, translating theory into real world use is very challenging. The work presented in this thesis relies on a number of assumptions that will not hold in many real world studies. It will therefore often lack generality, despite being explicitly crafted as a generalized version of existing theory. Beyond this issue, communication regarding the potential efficacy of these contributions to applied audiences is an important factor in their broader adoption.

### On Assumptions and Practical Considerations

The identification results I presented, including those in Chapter 2 that predate this work, generally make a variety of formal assumptions. Take, for instance, the assumption that the causal graph used as an input to the identification algorithm is correctly specified (i.e. latent variables in the right spots). This assumption is often unrealistic. In practice, graphs are typically specified in consultation with a subject-matter expert who may be prone to biases that lead to misspecification.

Violations of these assumptions can often be addressed by other areas of the causal inference literature. For instance, the causal discovery literature (Spirtes et al., 2000) proposes to learn graphs algorithmically from observational data. This approach is not without problems, of course. Those methods rely on their own assumptions and are often reliant on parametric methods[1] that are prone to misspecification. Nevertheless, causal discovery serves as a nice example of an approach to help ensure the assumptions necessary to use identification theory are satisfied.

Satisfying assumptions for using the algorithm may not be enough, however. For a variety of reasons, graphical models remain relatively under-utilized in fields like epidemiology and economics[2]. There, identification generally relies on much more restrictive assumptions, such as ignorability (see Chapter 2) or specific parametric

[1] Such as parametric statistical independence testing or parametric scoring functions.

[2] For instance, in the current editions (as of this writing) of the popular epidemiology journals, Epidemiology and The American Journal of Epidemiology, about 5% of papers mention causal graphs. About 1 in 50 describes their use of graphs in reproducible way. None considers a graphical model more complex than a fully-observed DAG

models. It will therefore be necessary for the causal graphical model community to convince applied researchers of the value of graphs before more general concepts, like identification theory, can receive widespread attention.

As an example of interaction between the theoretical and applied communities, consider the fact that in many real world observational studies, there is simply too much latent confounding[3] to obtain point identification. A cynic might then say "Yes, you can *use* the ID algorithm, but in the real world it is not *useful* since it will always fail to give me a functional to estimate." Clearly, such a comment is not entirely without merit: the ID algorithm formally embodies a conservative, if not pessimistic, view of what is knowable (Greenland and Finkle, 1995; Greenland, 1996). This comment may nevertheless be overly critical. Communicating the potential uses (some of which I delineate below) of the identification algorithms is an important and under-explored element of the research community's ongoing work.

### 7.2.2 Future Work

As discussed previously, applied researchers and real-world decision makers tend to make strong assumptions and then reason about the potential implications to violations of those assumptions. This stands in contrast to the a more conservative approach implied by identification theory: start with a model known to be unworkable (e.g. a model that is fully agnostic about confounder relationships and will therefore always lead to non-identifiability) and then incrementally add assumptions (e.g. remove edges, parametrize part or all of the model, assume the existence of proxies, etc.) until the model yields non-trivial results. Towards bridging the gap between these opposing approaches, it would be productive to think about ways that the identification framework presented in this thesis can be adapted and explained in a way that makes

[3] In most cases, when presented with a graph, an analyst will not be able to definitively (or comfortably) rule out the existence of enough causal relationships between latent variables and observed variables to make the effect of interest identifiable

it appealing to applied researchers.

One potential solution I've envisioned is to use ID-style algorithms as interactive tools in empirical science. The following points, at a high level, towards a framework similar to that described exclusively for DAGs in Haber et al. (2022). When a counterfactual query is non-identifiable, there are four potential paths forward:

1. Choose a more restrictive model
2. Choose a more general parameter space
3. Collect more data
4. Pick a different parameter

The first option might entail dropping causal edges or making parametric assumptions. The standard approach in applied work generally falls into this first category. This option also captures classes of methods like instrumental variables (Angrist, Imbens, and Rubin, 1996) and the broader proximal inference literature (Shpitser, Wood-Doughty, and Tchetgen, 2021). The second option includes evaluating partial identifiability (Manski, 2003; Duarte et al., 2021), which yields a range of effect sizes rather than a point estimate. The third option is often the most intuitive: most latent confounding arises due to a (naive or deliberate) lack of measurement, rather than an inability to measure. Therefore, in many cases non-identification can be addressed by modifying collection protocols to include more variables. Finally, it may be the case that 'giving up' on the target parameter is preferable and a similar, but identifiable, parameter should be estimated instead. As an example, the mediation literature has taken this approach to handling confounding. In some settings, the natural direct and indirect effects (Robins and Greenland, 1992) are not identifiable due to confounding between the mediator and outcome that is downstream of the intervention variable (the *recanting witness criterion*, see Avin, Shpitser, and Pearl, 2005 and Chapter 3).

VanderWeele, Vansteelandt, and Robins (2014) proposed addressing this by deriving alternative notions of direct and indirect effects which are identifiable in different settings from those defined in (Robins and Greenland, 1992).

Each of these options is viable in some scenarios and impractical in others. The ID algorithm cannot explicitly tell us which option is best for a given scenario, since that insight is grounded in a practical, scenario-specific understanding of the model. Nevertheless, a hypothetical synthesis between ID-style algorithms and the standard applied causal workflow would involve using the identification theory to *help* narrow down which option may be best.

As an example, suppose a researcher has posited a causal model with multiple latent variables, observed confounders, mediators, and a treatment and an outcome. Moreover, suppose that only some of the latent variables confound the treatment and outcome. In such a setting, the ID algorithm would help illuminate *which* latent variables are problematic. The researcher would be able to consider these variables critically and make a decision among the 4 options described above. For instance, if the offending latent variable is a basic demographic variable, the researcher might conclude that the easiest path is not to make parametric assumptions, but rather that collecting the variable is easy enough to be the best path forward.

The above example could make up part of a larger iterative process of using ID-style algorithms to evaluate effects. The researcher might choose one option (e.g. dropping edges) and find that the effect of interest is still not identifiable. She might then find that a different option is more reasonable, based on the variables that lead to non-identification in the subsequent application of the identification algorithm. Such a process would help make transparent the impact assumptions have on modeling. Moreover, it could be used to characterize the uncertainty of the overall modeling process, similar to sensitivity analysis, by observing how effect estimates change when different options are used to move from a non-identifiable effect to an identifiable effect.

This potential partnership between identification theory and applied science has many open areas for future research.

On the theory and methods end, this process could be viewed as a new angle on the sub-field of experimental design. For instance, the process of collecting additional variables in order to move a query from ‘non-identifiable’ to ‘identifiable’ is not trivial. Extending the model space by adding a variable could lead to positivity violations in finite sample settings and the community would need to find ways to test when this issue arises and propose solutions to handle it.

On the applied side of the issue, this line of inquiry opens many possibilities for improving scientists’ process for deriving new insights. For instance, consider the readmission work covered in Appendix B. In that work we studied the Shapley values of our machine learning models[4]. That analysis led to hypotheses about which variables causally affect readmission. Moreover, while the use of Shapley values in clinical research is relatively new, analyzing the parameters of an associative model is commonplace. Unfortunately, neither Shapley values nor the parameters of a trained predictive model can provide clear insight into the causal relationships at play. In the Appendix B example, a more comprehensive analysis is required in order to draw conclusions about these candidate causes of readmission.

The iterative process I laid out above could be adopted to address the shortcomings of these associational methods. For instance, using the predictive models we trained as a basis, we could formally evaluate the identifiability of the effect of various observed variables on readmission and, based on the results, either collect more data or think critically about the viability of making stronger modeling assumptions.

[4] Shapley values are pseudo-causal. They are calculated by perturbing the values of test features and evaluating how those perturbations affect model predictions. Essentially, they provide a measure of the causal impact of the variables on *predictions* rather than on the true value of the outcome

# Appendix A

# Towards Understanding the Role of Gender in Deploying Social Media-Based Mental Health Surveillance Models

## A.1 Introduction

The United States Centers for Disease Control and Prevention estimates that 8% of American adults suffer from major depression at a given time (Brody, Pratt, and Hughes, 2018). This represents a critical public health threat, as depression is associated with downstream physical health complications (Rush, 2007; Alboni et al., 2008) and an increased risk of suicide (Richards and O'Hara, 2014). Among the many efforts to address this crisis is a line of research at the intersection of language modeling, social media analysis, and mental health. The seminal papers De Choudhury et al. (2013) and Coppersmith, Dredze, and Harman (2014) demonstrated the general feasibility of predicting mental health status from social media data.

A major obstacle to the practical use of mental health surveillance models is differential performance for different subgroups of the population. This behavior can arise either because the training data is not sufficiently representative of the population, or because some groups are simply harder to predict given the same data.

The former case is well-studied in the machine learning literature and can be addressed by careful data collection and training regimes. The latter case, however, is often more subtle and harder to address. *Not* identifying and addressing these differences in performance degrades the utility of the models. In particular, if the performance is worse for historically marginalized populations it can reinforce existing inequities such as under-diagnosis of depression (Elazar and Goldberg, 2018).

In this work we aim to assess the scope of the differential performance problem by studying the relationship between gender and predictions of depression. The most useful insight we could gain would be determining whether or not gender is a confounder for depression predictions; that is, whether gender both causally affects the way in which users post on Reddit and causally affects our predictions of the user's depression status. Unfortunately, testing whether this causal dynamic is true is very difficult with the purely observational data available to us. Towards testing this phenomena, we will instead test the slightly weaker hypotheses i) that depression predictions exhibit gender bias (i.e., there are differences in performance across genders) and ii) that these differences are due, at least in part, to differing uses of language between men and women in talking about their mental state. Together these hypotheses serve as a sort of associational version of the causal phenomenon we'd like to study. They can tell us whether depression predictions are correlated with gender and whether certain terms are likely to have different meanings based on the gender of the author.

We test hypothesis (i) quantitatively by fitting depression prediction models to a novel data set collected from Reddit with *ground truth* genders, derived from self-disclosures, and comparing the performances across genders. We test hypothesis (ii) qualitatively by looking at features strongly predictive of depression for each gender. We identify themes that are concordant across genders and consistent with the literature (De Choudhury et al., 2016) as well as themes that are discordant across genders and support our hypothesis that men and women use many terms differently to talk about (non-) depression. We follow these analyses with a discussion of open questions that follow from this work. In particular, we discuss the use of causal methodologies to assess our stronger hypothesis that gender confounds depression prediction. We highlight the types of methods that could be used and the data that is necessary to test the causal hypothesis. We conclude with a discussion of limitations and the ethical implications of this work.

## A.2 Related Work

Several existing papers have considered the role of demographics in mental health prediction. Elazar and Goldberg (2018) demonstrated that demographics are implicitly encoded in text data. Wood-Doughty et al. (2017) and Loveys et al. (2018) both studied differing language use across cultures. The former used a Twitter data set with *inferred* demographic labels, while the latter used a carefully-curated proprietary data set from 7 Cups of Tea. Amir, Dredze, and Ayers (2019) explored the role of cohort selection in assessing mental health disorder prevalence. Aguirre, Harrigian, and Dredze (2021) is the closest to the present work. The authors characterized the biases present in depression prediction models by showing there are differences in performance for different demographic subgroups. This work studied biases that arise due to the specific data set used for training,focusing on the popular, publicly available data sets CLPsych (Coppersmith et al., 2015) and MULTITASK (Benton, Mitchell, and Hovy, 2017).

The present work differs from those cited in that we seek to quantify demographic bias in depression prediction using self-disclosures in a publicly available data set. This approach improves scalability and reproducibility compared to hand-labeled and proprietary data sets. Additionally, while self-disclosures are not perfect, they are not subject to the same degree of noise and error that is induced when using genders inferred by using a pre-trained model, trained on an auxiliary data source.

Our estimates of the depression prediction performance across genders are therefore likely to be of a higher quality. Moreover, our analyses of features that are predictive of depression for each gender are also likely to be less noisy than they would be if we were also inferring genders from those same features.

## A.3 Data Collection

To obtain a dataset with ground truth gender, we mined all posts and comments from the r/AskMen and r/AskWomen subreddits between January 1, 2019 and December 31, 2019 using the Pushshift API (Baumgartner et al., 2020). In total, we collected 251,487 original submissions and 4,481,354 comments.

For each post, we consider the flair – an optional tag users can apply to their posts to reveal information about themselves or the content of their post – to determine the ground truth gender of the post author. We considered the author of a post to be true-male if they used one of ‘Male’, ‘male’, ‘Dude’, or ♂ for their flair, and true-female if they used one of ‘Female’, ‘female’, ♀, or ♀♡. Of the mined posts, 1,002,079 had some sort of flair, while 660,684 had one of the male or female indicator flairs. This process yielded a data set of 15,140 unique male and 11,241 unique female users, as well as 59 users whose gender-related flair use was inconsistent (i.e. at least one post each with a male- and female-indicating flair). While people who identify as non-binary are known to have higher rates of depression (Budge, Adelson, and Howard, 2013; Wolohan et al., 2018) and thus could benefit from the studies like this one, we did not have a reliable method for identifying non-binary users beyond the list of inconsistent users and the sub-population in our cohort was too small to yield meaningful analysis. For the remainder of the paper we restrict attention to binary genders under the folk conception of gender (Larson, 2017).

For each of the 26,381 gender-binary users, we collected the user’s entire Reddit posting and commenting history from January 1, 2019 to December 31, 2019, totaling 1,035,782 original submissions and 19,029,981 comments across 64,162 subreddits. Following the literature on social media-driven mental health surveillance (De Choudhury et al., 2013; Yates, Cohan, and Goharian, 2017), we defined a user as true-depressed if they authored an original submission or comment in r/depression during the study period and true-control otherwise. The breakdown of gender and depression classes is 721 and 713 depressed males and females respectively, and 14,416 and 10,526 control males and females respectively.

## A.4 Methods

We fit user-level models to predict depression status from our harvested Reddit data. To enable analysis of the impact of gender as a confounder, we fit separate models on two separate data sets: a random sample of the true-men users in our data set, and a random sample of the true-women users. To reduce noise induced by 'throwaway' or 'lurker' accounts, we excluded users who made fewer than 5 posts (submissions + comments) during the study period. This decision could reduce our results' generalizability since throwaway accounts may be owned by users with separate primary accounts and post with the throwaway differently (e.g. posting more personal information).

Because depression is a rare outcome in our data, our initial train and test sets had very few depressed individuals (109 train, 26 test). This proved too few to draw meaningful conclusions about the role of gender in depression prediction. We therefore report the performance of our models trained on data sets constructed by performing *balanced sampling* from the full data. The resulting class breakdowns are: 721 and 613 depressed males and females respectively, and 820 and 712 control males and females respectively.

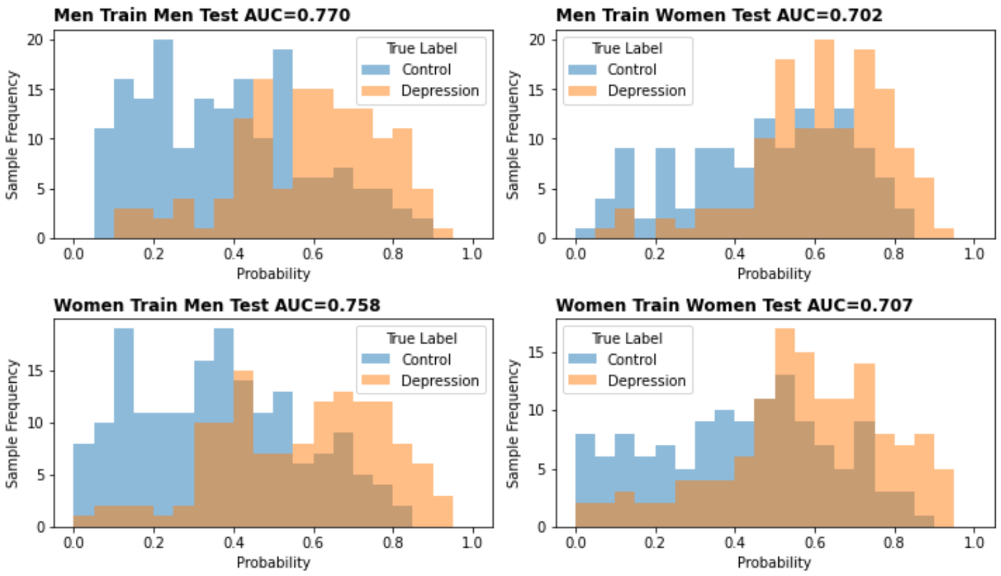


Figure A-1. Performance of each model, trained to predict depression on either male users or female users only, when evaluated on each test set

We split each of these sampled data sets 80-20 into train and test sets, stratifying by user. We then constructed a Bag-of-Words (BoW) vocabulary from the submissions and comments for each user in the training sets. We included 1-, 2-, and 3-grams, as well as LIWC (Pennebaker, Booth, and Francis, 2007) and TF-IDF (Jones, 1972) features. We imposed that features must be used by a minimum of 25 users to be included in the vocabulary. We also removed posts from the r/depression subreddit from each user's BoW vector and filtered out terms and subreddits commonly associated with self-disclosure of mental health disorders using the SMHD dataset (Cohan et al., 2018). To model depression, we used the scikit-learn implementation of regularized logistic regression (Pedregosa et al., 2011). At the end of training, we discarded all but the top 100,000 features using the pairwise mutual information criterion as an additional regularization step.

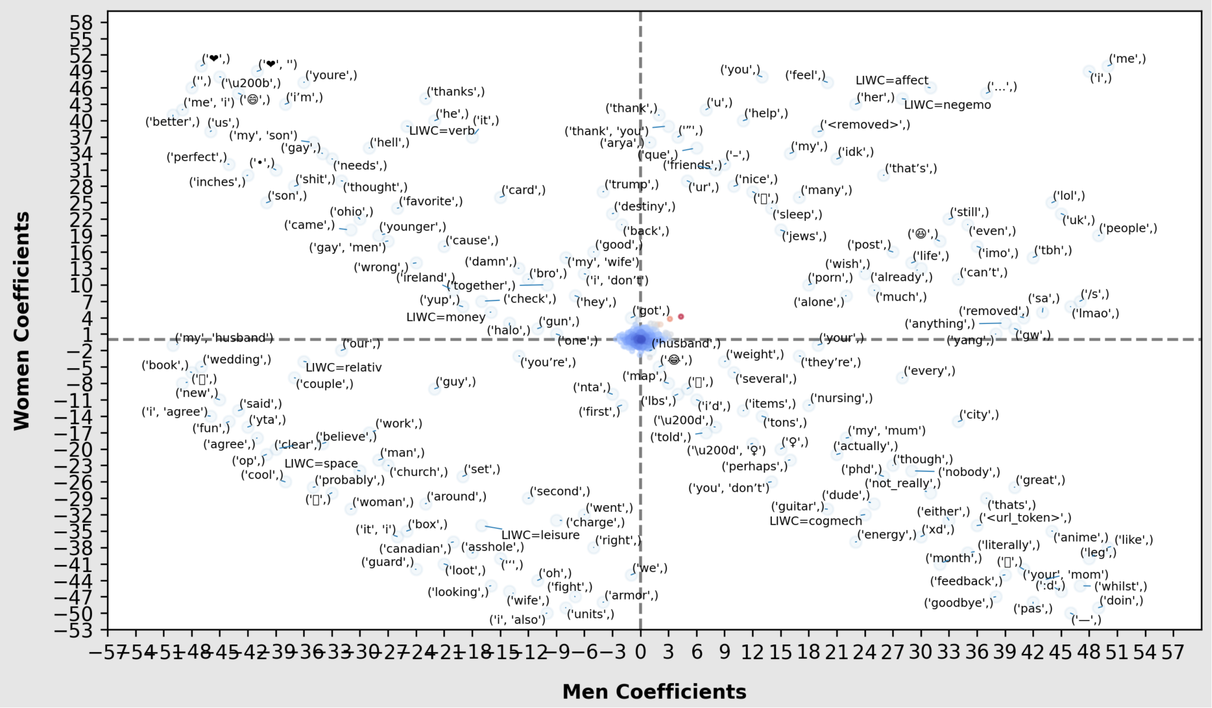


Figure A-2. Features in common between the male- and female- trained models with the 50 highest scoring features in each quadrant labeled

# A.5 Results

## A.5.1 Model Performance

The performance of each model on each test set is shown in Figure A-1. The most striking result is that the performance of both models is considerably higher on the men-only test set than on the women-only test set (.770 vs. .702 and .758 vs. .707 respectively). This difference indicates that predicting depression among men is easier than among women. Looking at the distribution graphs, it appears that women are *over diagnosed* as depressed. Mechanically, this difference in predictions likely arises due to the existence of a few key features that indicate depression for one gender but not the other. We identify candidate features in the analysis below.

### A.5.2 Feature Analysis

We extracted the regression coefficients from each of our models and generated a scatter plot in Figure A-2 of the 50,967 features the two models had in common. Towards identifying strongly predictive features, we scored each feature using the sum of the absolute value of the coefficient from each model for that feature. In the figure, we labeled the 50 highest-scoring features in each plot quadrant.

**Concordant Depression Features (top right)** Even though we filtered out self-disclosure tokens (e.g. 'depression' and 'depressed'), we see that many of the most predictive features are consistent with themes discussed in the mental health surveillance literature (De Choudhury et al., 2016): emotion ('feel', LIWC affect, LIWC negemo), physical symptoms of depression ('sleep'), and indicators of social isolation ('alone', 'porn', and personal pronouns 'me', 'my', and 'I'). One notable feature is the token 'jews'. This feature could indicate that many depressed Jewish people of both genders frequently discuss their religious identity on Reddit, possibly in the context of their peoples' historically marginalized status (McCullough and Larson, 1999). Also plausible is that the token is indicative of anti-semitic tendencies which are correlated with depression (e.g. blaming one's personal struggles on a scapegoat minority group). This phenomenon has been documented in the largely-male ‘incel’ community (Hoffman, Ware, and Shapiro, 2020) but we could not find a clear connection between anti-semitism and depression among women in the psychology or sociology literature.

**Concordant Control Features (lower left)** These feature themes are also consistent with findings in the literature. Features indicative of social interactions are quite common ('church', 'wedding', 'couple') as well as features that suggest positive affect regarding life activities ('fun', 'cool', LIWC leisure).

**Discordant Features (top left, lower right)** These features are of primary interest for identifying potential gender-based confounding. Here we find features that are predictive of depression in women but control in men or vice-versa. We observe that there are several terms that likely have different meanings for men and women users. Many of these pertain to social interactions.

For instance, ‘gay’, ‘gay men’ and ‘my husband’ are all strongly predictive of control for men. This suggests that men who are comfortable discussing non-straight sexualities online are also in a relatively healthy mental state. In contrast, these terms (along with ’my wife’) indicate increased mental health struggles for (possibly gay) women. We suspect ’my husband’ is neutral for women because there are roughly equal numbers of users praising and condemning their husbands.

Beyond sexuality, we see that some familial terms have differing predictive interpretations across genders. ’my mum’ is predictive of depression for men and control for women, while the reverse is true for ’my son’. This suggests a substantial difference in parent-child relationships depending on the gender of each: each gender appears to have an affinity for family members of the same gender.

We also highlight a few features with broader societal interpretations. ’trump’ is strongly predictive of depression among women but neutral for men. This is consistent with the well-known ‘gender gap’ phenomenon and could also indicate that mental health is in part a function of political climate. The LIWC category ’money’ is slightly depression predictive for women and control-predictive for men. Similar to the above, this could be an artifact of the wage gap: money topics may be more stressful for women because they tend to earn less money for the same amount or more work.

## A.6 Discussion

In this paper we showed that depression predictions do indeed exhibit gender bias. This was evidenced by a substantially better performance when predicting depression among males than when predicting among females. We also identified terms that are used differently between men and women, providing insight into the manifestations of depression beyond modeling dynamics.

### A.6.1 Open Questions and Future Work

As hinted in the introduction, the key open question is **does gender confound depression predictions?** In other words, does gender *both* affect depression predictions *and* the features we use to predict it? There are numerous plausible explanations for why both of these causal relationships may hold or not hold, but without a rigorous causal analysis, it is not possible to rule any one explanation out in favor of another.

To properly evaluate whether a associational relationship is in fact causal, the causal framework requires 'intervening' on an independent variable while holding other variables in the system constant to see whether there are changes in the dependent variable. Here, that means intervening on gender, which is infeasible to carry out directly.

There may however, be some viable proxy approaches for simulating the intervention on gender. One such approach would entail fitting a model to predict the ground truth gender and then using a clustering algorithm to find male and female centroids based on the most predictive features in the gender prediction model. The analyst could then simulate an intervention on gender for the purposes of analyzing changes to depression prediction by replacing the user's feature vector in the depression inference model with each gender centroid vector. This approach will not permit a true causal interpretation but it could provide insights into the relationship between gender and depression prediction beyond those gained from the simple models studied in this work. Unfortunately this approach cannot be applied to analyzing the relationship between gender and the text features since it entails changing those text features.

Outside of the explicit question of confounding, we can ask **how do we correct for the performance differentials across demographic groups when predicting depression?**. As hinted earlier, an obvious approach with support in the literature (Amir, Dredze, and Ayers, 2019) is to simply collect 'better' data. This is an unsatisfying answer, however, since good data is often hard to come by or expensive to collect. Instead, we can again turn to causal inference ideas to try to address data quality issues. We can potentially use methods from the causal fairness literature to impose constraints on depression models to ensure negligible differences in prediction performance. For instance, following Nabi, Malinsky, and Shpitser (2019), we could impose a constraint that requires that the total effect of gender on depression predictions is zero, or, plainly, that there is no difference in model performance when we do or don't condition on gender.

### A.6.2 Limitations

Aside from the limitations described above, i) all users in our cohort posted in r/AskMen or r/AskWomen (which we used to derive ground truth) and ii) we rebalanced our data sets due to insufficient numbers of depressed users in the 'representative' population. These decisions could reduce the generalizability of our results. One way to address this would be to collect data on more users by expanding the study period and by consulting other subreddits with gender self-disclosure such as r/relationships (Wang and Jurgens, 2018).

Additionally, while our use of self-disclosed genders increases scalability, this could induce bias in two ways. Users could be dishonest in their disclosure and, even if they aren't, users who choose to self-disclose could be fundamentally different from the general population. It's likely that the only solution is to collect data external to Reddit about Reddit users' genders as a more reliable supplement to our data.

Finally, our depression labels were not obtained via self-disclosures. Rather, they were defined based on whether the user posted in the r/depression subreddit. While this approach is consistent with data collection approaches from the literature (De Choudhury et al., 2013), it is likely to induce some noise. For instance, a user could post in the subreddit to seek support for a friend or relative, rather than for themself and would therefore be incorrectly labeled as depressed. One way to address this would be to take a more nuanced approach to labeling. For instance, we could use regular expressions matched on the text of r/depression posts to develop a more exclusive labeling policy that filters out users who are not seeking personal support.

### A.6.3 Ethics

As in any applied setting it is necessary to weigh the potential advantages and harms of carrying out our research agenda. This work has the potential to cause harm in a couple key ways.

First, as previously mentioned, we restrict attention to users satisfying a narrow and dated 'folk' definition of gender in line with much of the existing research in the space of computational psychology. This is done at the cost of excluding non-binary individuals, who potentially stand to benefit the most from this work due to the increased prevalence of depression in gender non-conforming populations. Furthermore, excluding any marginalized population from a study of this type has the potential to reinforce existing biases. For instance, if our model had demonstrated improved prediction performance for the binary genders, that could lead to an incorrect assumption that the model will perform well on the general population, which includes non-binary genders. This could lead to *worse* performance for the unstudied groups.

Second, while we infer depression status from Reddit users with the goal of alleviating harms, these approaches could be harnessed with malice to identify and target already vulnerable individuals whose screen names and posting behavior are public.

On the other hand, there is great potential in this study and the work that will follow it. Identifying obstacles to model deployment for a restricted population will likely aid in correcting those obstacles for the entire population. This would substantially improve the performance and, more importantly, the clinical utility of mental health surveillance models. Given the potential benefits of this study we feel it is better to proceed, with care and transparency, rather than sit idle for lack of perfect answers to address the issues the work poses.

# Appendix B

# Leveraging Machine Learning to Predict 30-Day Hospital Readmission after Cardiac Surgery

## B.1 Introduction

Hospital readmission is a key outcome in clinical medicine and health policy. Constituting approximately 13.9% of hospital stays in the U.S. (Bailey et al., 2019), readmissions result in billions of dollars of additional direct healthcare expenditures each year (Commission, 2007) and numerous indirect costs. These issues are particularly relevant in cardiac surgery, where two of the highest volume procedures, coronary artery bypass graft (CABG) surgery and heart valve surgery, also have high readmission rates (18.5% and 15.1% of admissions, 14th and 20th highest respectively) (Weiss, Elixhauser, and Steiner, 2006). Towards limiting readmissions, the Patient Protection and Affordable Care Act mandated the creation of the Hospital Readmissions Reduction Program, which penalizes hospitals for exceeding hospital-specific all-cause readmission rates for several conditions and procedures, including CABG.

Predicting 30-day readmission after cardiac surgery, however, is notoriously challenging. Existing approaches generally rely on less-flexible methods or expensive prospective data collection (Kilic et al., 2017). Moreover, even the best approaches in the literature have impractically poor performance, suggesting a reliable approach to predict readmissions following cardiac surgery does not yet exist. A reliable prediction tool would enable clinicians to adopt better treatment plans for high-risk patients and reduce downstream costs.

Machine learning (ML) is a promising avenue for developing reliable clinical prediction models. The field of machine learning seeks to develop algorithms capable of making decisions that improve as they encounter more data. Recent years have seen dramatic advances in ML performance with applications in clinical settings such as sepsis prediction (Henry et al., 2015), mental-health surveillance (Sherman et al., 2021), healthcare-associated infection prediction (Wiens, Horvitz, and Guttag, 2012), and medical imaging (Lundervold and Lundervold, 2019). These performance improvements can be attributed in part to the fact that ML algorithms are designed to flexibly model data; they are characterized by their ability to handle high dimensional data (hundreds or thousands of variables, rather than dozens) and they have mechanisms for automatic feature selection. Considering these algorithms' potential, we sought to develop a model to predict 30-day all-cause readmission following cardiac surgery.

## B.2 Patients and Methods

### B.2.1 Study Population and Variable Details

We analyzed a cohort of all adult patients who underwent a procedure performed by a cardiac surgeon at the Johns Hopkins Hospital between January 1, 2011, and June 30, 2016. From this initial cohort, we excluded patients who died in hospital during the index admission and patients whose records contained no lab test results (see also Figure B-1). This study was approved by the Johns Hopkins University Institutional Review Board, study number IRB00128973.

We leveraged patient data extracted from the electronic medical record (EMR)

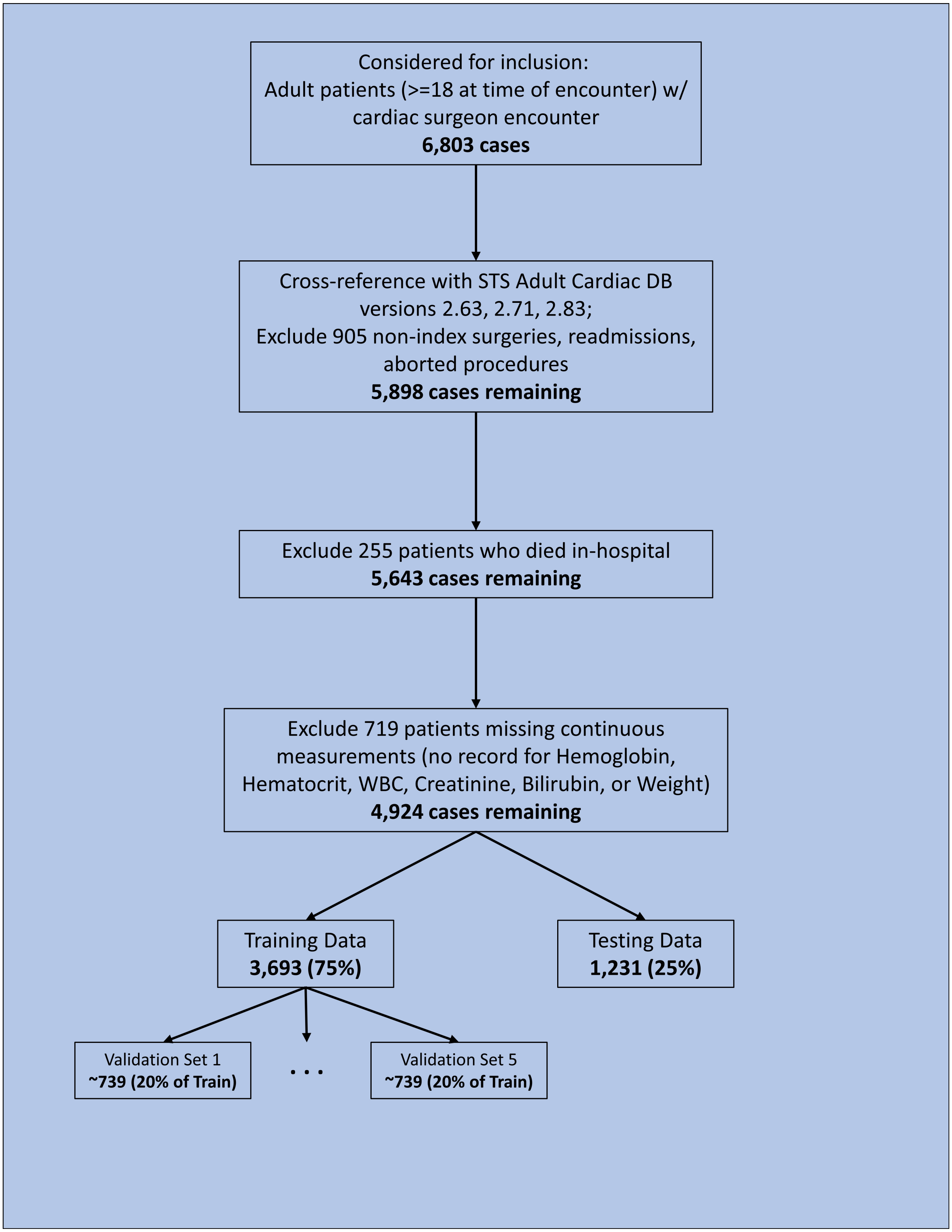

Considered for inclusion:
Adult patients (>=18 at time of encounter) w/ cardiac surgeon encounter
**6,803 cases**
Cross-reference with STS Adult Cardiac DB versions 2.63, 2.71, 2.83;
Exclude 905 non-index surgeries, readmissions, aborted procedures
**5,898 cases remaining**
Exclude 255 patients who died in-hospital
**5,643 cases remaining**
Exclude 719 patients missing continuous measurements (no record for Hemoglobin, Hematocrit, WBC, Creatinine, Bilirubin, or Weight)
**4,924 cases remaining**
Training Data
**3,693 (75%)**
Testing Data
**1,231 (25%)**
Validation Set 1
**~739 (20% of Train)**
. . .
Validation Set 5
**~739 (20% of Train)**

Figure B-1. Diagram detailing the process we took in applying our exclusion criteria and splitting our data to obtain analysis and evaluation data.

| **Variable Name** | **Categories** | **Missing** | **Overall** |
|---|---|---|---|
| n | | | 5643 |
| Patient Age, mean (SD) | | 0 | 61.1 (14.4) |
| Gender, n | Female | 0 | 1843 (32.7%) |
| Length of Stay (hr), median [Q1, Q3] | | 0 | 214.4 [152.8, 342.5] |
| LatinX ('Ethnicity'), n | Yes | 61 | 73 (1.3%) |
| Caucasian, n | Yes | 47 | 4157 (74.3%) |
| Black, n | Yes | 47 | 916 (16.4%) |
| Pre-op Dialysis, n | Yes | 8 | 155 (2.8%) |
| Pre-op Hypertension, n | Yes | 13 | 3845 (68.3%) |
| Pre-op Diabetes, n | Yes | 9 | 1608 (28.5%) |
| ICU Readmission, n | Yes | 2 | 209 (3.7%) |
| Post-op Renal Failure, n | Yes | 3 | 134 (2.4%) |
| Post-op Sepsis, n | Yes | 1888 | 26 (0.7%) |
| Weight Delta (kg), mean (SD) | | 55 | -3.1 (21.5) |
| Discharge Loc., n | Ext./Transitional Care or Rehab | 54 | 880 (15.7%) |
| | Home | | 4643 (83.1%) |

Table B-I. Summary Statistics for Select Variables (pre-exclusion).

and linked these to patient records from versions 2.61-2.81 of the Society of Thoracic Surgeons Adult Cardiac Surgery Database (STS). From the STS data, we extracted static variables collected during the pre-, intra-, and post-operative periods, making sure to only select variables whose definitions were stable across the three versions. From the EMR we derived several time series variables using patients' lab test results and weight, including the minimum and maximum values of each of these measurements throughout the admission and a variable describing the trend from admission to discharge. Summary statistics for a selection of the variables are provided in Table B-I.

We applied three standard transformations to the collected variables: 1) we handled missing values by adding a 'missing' category for categorical variables and mean-filling continuous variables, 2) we converted categorical variables to binary variables to avoid having our algorithms incorrectly interpret categories as ordinal, a process known as ‘one-hot encoding’ (Garavaglia and Sharma, 1998), and 3) we normalized continuous variables to lie in the unit interval. In addition to this “full data” cohort, which includes a large variety of procedures, we fit models on two restricted cohorts: cases during which the patient was placed on cardiopulmonary bypass, and cases that constitute the 7 major procedures for which the STS has published risk scores (isolated CABG, isolated aortic valve replacement, aortic valve replacement + CABG, mitral valve replacement, mitral valve repair, mitral valve replacement + CABG, and mitral valve repair + CABG). We refer to these cohorts as “On-Pump” and “7-Majors” respectively.

### B.2.2 Machine Learning Algorithms

We fit (‘trained’) and evaluated three different machine learning algorithms: a Random Forest (RF) model (Breiman, 2001), XGBoost (Chen and Guestrin, 2016), and a Support Vector Classifier (SVC) (Boser, Guyon, and Vapnik, 1992). An overview of these models is provided below. As benchmarks representative of existing approaches, we also fit a standard logistic regression model on the full variable set and an RF model on a restricted STS-only variable set. Analysis was performed using standard Python libraries.

Random Forests and XGBoost are tree-based algorithms. Trees consist of a hierarchy of nodes which are used to sort data points (e.g., patients) into groups (e.g., high/low readmission risk) based on the values of the variables for each data point. The RF and XGBoost algorithms automatically choose which variables to use and employ automated feature selection through a random model search process. Support vector classifiers are a binary classifier that constructs a linear decision rule to separate positive from negative examples, like the logistic regression classifier. Unlike logistic regression, however, SVC maps features to a high dimensional space, prior to constructing the decision rule. This allows more flexible decision rules to be constructed. All three models have several ‘hyperparameters’ set by the analyst. These parameters control fitting behavior, such as the restrictiveness of automatic feature selection procedures.

### B.2.3 Machine Learning Pipeline

Though the training algorithm for each model is different, all follow the same general process, referred to as a ‘pipeline’ in the ML literature. After pre-processing the data, we randomly split the data into two sets: a ‘training’ set, constituting 75% of the data, and a ‘test’ set, containing the other 25% of the data. Critically, we hold out the test set until the very end of the pipeline so that we can apply the trained model to it and obtain a realistic assessment of how the model would perform if applied to unseen data as in a clinical deployment.

After we split the data into training and testing sets, we further randomly split the training data into five equal-sized sets. We use these ‘validation’ sets to perform cross-validation to choose the model’s hyperparameters. We hold out one validation set and fit the model on the other four. We calculate the area under the receiver operating curve (AUC) for each hyperparameter setting of the model. We repeat this process such that each validation set is held out once. We then refit the model on all the entire training set using the hyperparameters that had the highest average validation AUC, yielding our final model.

We apply this final model to the held-out test set, calculating both an AUC score and a Brier score for the test set. To construct 95% confidence intervals, we perform bootstrap sampling. We construct 1,000 test sets by sampling the original test set with replacement and apply the model to each sampled test set to obtain an AUC and Brier score for each. We calculate the 2.5th and 97.5th percentile scores among of the bootstraps to construct the confidence interval.

### B.2.4 Univariate Analysis

Beyond calculating our models' performance using the AUC and Brier score on a held-out test set, we also analyzed the parameters of our best-performing model to determine which features were most important for predicting readmission. For RFs (the best-performing model; see Results section), we analyzed the models' Gini importance (Breiman, 2001) and Shapley values (Lundberg and Lee, 2017). Gini importance, like weights in a logistic regression, describes the relative importance of a variable in a tree-based regression model. Unlike logistic regression weights, Gini importance values cannot characterize whether the correlations between variables are positive or negative. In contrast, Shapley values can provide such characterizations (i.e., "higher patient age has a positive association with readmission"). Both metrics add interpretability to the non-parametric RFs model.

## B.3 Results

### B.3.1 Data Details

6,803 adult cardiac surgery encounters were extracted from the EMR. Excluding non-index surgeries and readmissions, which did not appear in the STS database (905), cases where the patient died in the hospital (255), and cases which were missing one or more EMR lab/weight variables (719) resulted in an analysis cohort of 4,924 cases (see Figure B-1). 723 (14.7%) were readmitted to our institution within 30 days of discharge. Since readmission definitions are inconsistent across versions of the STS database, we defined readmission using admission and discharge dates in the EMR. Many of the cases excluded due to missing lab/weight measurements were missing those values because they represented minor procedures (e.g., lead extractions, sternal wire removals). Our data collection and processing steps yielded a set of 165 variables for each case to use for model development. The On-Pump and 7-Majors cohorts had

| Model | Full Data | On-Pump | 7-Majors |
|---|---|---|---|
| Random Forest | .76 (.72, .79) | .76 (.72, .80) | .80 (.75, .86) |
| XGBoost | .75 (.71, .78) | .76 (.73, .80) | .80 (.75, .85) |
| Support Vector Classification | .75 (.72, .78) | .75 (.71, .79) | .79 (.73, .85) |
| Logistic Regression | .75 (.71, .79) | .74 (.70, .78) | .76 (.70, .82) |

Table B-II. Held-Out AUCs (95% Confidence Interval) for Readmission Prediction Trained Using Various Data Sets

4,516 (625 readmissions; 13.8%) and 2,951 (361 readmission; 12.2%) respectively.

### B.3.2 Model Performance

The AUC and Brier score of each model is in Tables B-II and B-III. AUC confidence intervals for the ML models overlapped, indicating the performances were quite close. The best performing model, according to AUC point estimates, was a Random Forest. On the held-out test set, the model achieved an AUC of .76 (95% CI: (.72, .79)) with Brier .16 (.15, .17). The RF model trained and evaluated solely on STS data achieved an AUC of .64 (95% CI: (.60, .68)), which is comparable to models from the literature.

The SVM had the best Brier .12 (.11, .14) suggesting it might perform better than the RF under certain conditions. We treat RF as the “best” model for the purposes of further analysis since AUC is a more appropriate score given the intended use case (Hernández-Orallo, Flach, and Ferri Ramírez, 2012) and because comparison models from the literature generally only report AUC. Performance differences between the full and On-Pump cohorts were negligible while “7-Majors” models tended to have higher performance (e.g., full cohort Random Forest AUC .76 (.72, .79) vs. .80 (.75, .86) for the 7-Majors Random Forest).

### B.3.3 Feature Importance

The average Gini importance across all 165 variables in the top performing Random Forest model for the full data cohort was .006. The five most important variables were

| **Model** | **Full Data** | **On-Pump** | **7-Majors** |
|---|---|---|---|
| Random Forest | .16 (.15, .17) | .16 (.15, .16) | .14 (.13, .15) |
| XGBoost | .20 (.19, .21) | .20 (.19, .21) | .19 (.18, .19) |
| Support Vector Classification | .12 (.11, .14) | .11 (.10, .12) | .09 (.07, .10) |
| Logistic Regression | .20 (.19, .21) | .21 (.19, .22) | .19 (.18, .20) |

Table B-III. Held-Out Brier Score (95% Confidence Interval) for Readmission Prediction Trained on Various Data Sets

length of stay (Gini importance .076, 12.5x more important than the average variable, positively predictive), discharge location: rehab (.068, 11.2x, positively predictive) discharge location: home (.047, 7.7x, negatively predictive), ICU length of stay (.042, 7.0x, positively predictive), and the lowest hemoglobin measurement throughout the admission (.035, 5.8x, negatively predictive). Two of our time series EMR-derived variables, measuring the delta in serum creatinine and patient weight, had relatively high Gini importance values of .024 and .022, ranking 9th and 22nd among all variables, respectively.

Figure B-2 exhibits the Shapley values of the 20 most important variables. Each row corresponds to a variable. Red and blue dots in a row denote high and low values of that variable, while the x-axis represents the predictive value of the variable. For instance, the first feature, “discharge location (rehab)”, is binary and thus can only take on the values 1 (red) and 0 (blue). In the figure, we see that red values (discharged to rehab) tend to fall on the positive side of the x-axis and vice-versa for blue values (not discharged to rehab); thus, “discharge location (rehab)” is positively correlated with 30-day readmission. Similarly, “discharge location (home)” is negatively correlated with 30-day readmission.

## B.4 Comment

Predicting readmissions following cardiac surgery is notoriously challenging. In this study, we leveraged data collected during standard care, augmenting traditional STS

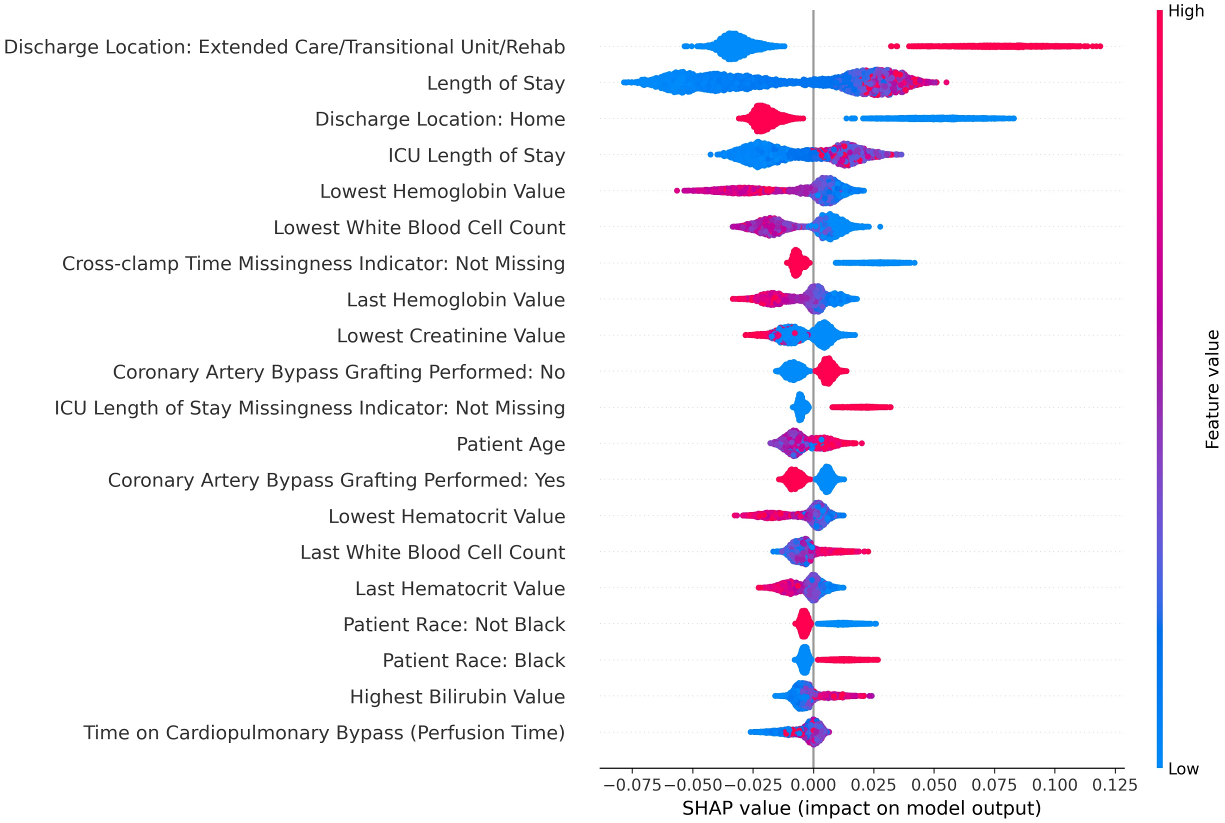


Figure B-2. Shapley values for high-importance variables in RF model on full data.

| **Paper** | **Model Variables** | **Train Perf.** | **Test Perf.** |
|---|---|---|---|
| Kilic et al., 2017 | STS Variables | .64 (N/A) | N/A |
| Fanari et al., 2017 | Curated risk factors | .74 (.65, .75) | .60 (N/A) |
| Brown et al., 2018 | Curated risk factors, prospectively-collected biomarkers | .74 (.68, .79) | .48 (.42, .54) |
| Benuzillo et al., 2018 | Curated risk factors | .63 (N/A) | N/A |
| Tam et al., 2018 | Curated risk factors | .63 (N/A) | N/A |

Table B-IV. Representative Summary of Readmission Prediction Performance in the Literature (AUC, 95% Confidence Intervals)

variables with EMR-derived variables, to train machine learning models for readmission prediction. Our models outperformed the state of the art (see below). These results are promising and suggest machine learning techniques for readmission prediction should be further explored. Nevertheless, the fact that even state-of-the-art methods do not yet perform well enough to be deployed in practice supports criticisms of the use of readmissions to evaluate quality of care and determine hospital reimbursements.

### B.4.1 Comparison to Existing Approaches

We provide a representative summary of recent efforts to develop a prediction model for post-cardiac surgery readmission in Table B-IV. Kilic et al. (2017), Fanari et al. (2017), Benuzillo et al. (2018), and Tam et al. (2018) all have a similar approach: hand-select a small set of risk factors, perform variable selection, and fit a logistic or Cox regression. Brown et al. (2018) is unique in that it included six prospectively-collected biomarkers. Except for Fanari et al. (2017) and Brown et al. (2018), the AUCs reported in these papers describe model performance on the training data. As described in the Methods section, reporting training performance is not informative: any sufficiently powerful regression method can achieve 100% training accuracy with no guarantee the model will generalize to unseen data (Liu et al., 2019). Training performance cannot provide insight into how the model might perform when applied prospectively to new patients, the ultimate intended use case. That distinction aside, our best-performing model yielded better test performance than the reported performance of all published models we identified. Since Fanari et al. (2017) and Brown et al. (2018) did report held-out test performance, we can make direct comparisons. Fanari et al. (2017) had a held-out AUC of .595 while Brown et al. (2018) had a held-out AUC of .48 (95% CI (.42, .54)). Both perform substantially worse on held out data. For context, AUC $< .50$ suggests a model performs worse than a "model" that makes predictions by simply tossing a coin. Clearly, the performance gap, a difference of test AUCs of .16, between the literature and our models is stark. We also note that several comparison papers did not include confidence intervals for their performance estimates, making it difficult to properly assess the efficacy of those approaches.

### B.4.2 Choice of Cohort

The higher performance of the 7-Majors-trained models relative to the full- and On-Pump-trained models suggests that it is easier to predict readmission risk for these procedures. We argue that this supports our use of the full cohort for our other analyses: it is necessary to study and develop models that can make accurate predictions for the full spectrum of patients a clinician might see. It is possible that developing separate models for the 7-Majors and non-7-Majors would yield better performance on each task. We leave this question for a future study. Note that some of the comparison models from the literature (Tam et al., 2018; Brown et al., 2018) were trained using 7-Majors data, suggesting the outperformance of our models may be even more stark than what was highlighted above: a difference in AUC of as much as .32.

### B.4.3 Insights from Univariate Analysis

It is not surprising that overall and ICU length of stay are important for predicting readmissions. Discharge location being highly predictive, on the other hand, is notable.

Murphy et al. (2008) suggests that a related variable, living alone, is predictive of readmission. Our models are associational rather than causal and so they cannot identify the causes of readmission, but this finding and Murphy et al. (2008) suggest that some key factors leading to readmission may not be directly related to in-patient care. Future research should be devoted to understanding how post-discharge care affects readmission risk. On a different note, the relative importance of our creatinine and patient weight trend variables suggests that time series variables can be beneficial for readmission prediction. Removing the time series variables from the Random Forest model decreased AUC by .12. Other areas of clinical prediction modeling, outside of readmissions and cardiac surgery, would likely see similar performance increases by adding time series and EMR variables that are not always available in curated data sources like the STS database.

Turning attention to the Shapley values, most of the results are consistent with existing clinical knowledge. We highlight, however, that being black is highly correlated with readmission. This suggests readmission studies merit the same critical look at disparities that has been applied elsewhere in medicine in recent years (Mazzeffi et al., 2020).

### B.4.4 Limitations

As mentioned in the Methods section, the STS database has inconsistent readmission definitions across versions, so we defined readmissions according to the EMR. Unfortunately, this approach is imperfect. Our data likely contains instances where a patient was readmitted to another hospital system, which was not captured by our EMR system during the study period. In such a case, we would incorrectly label the patient as 'not readmitted'. A more accurate labeling scheme would likely improve reported prediction performance relative to our results, since better labeling would reduce noise. Recently, the Maryland Cardiac Surgery Quality Initiative began reporting data on these inter-system readmissions (Mazzeffi et al., 2020). Unfortunately, the Initiative's data are too recent to inform the present study since they do not overlap temporally with our data. Better readmission labels obtained in this way would likely improve our study's external validity.

Our ability to model readmissions is also limited by the nature of EMR data. Medical data are often described in the ML community 'messy' because they have heterogeneous variables, missing values, and measurement error. For instance, we had to exclude a large portion of our the available data due to variable missingness and limit our variables to those which had stable definitions over time. We also were not able to include several meaningful factors, such as medications, compliance, health literacy, and measures of the familial support. We strongly suspect that including these variables would improve our models.

### B.4.5 Broader Implications and Future Work

We envision models like those studied will be incorporated into clinical practice in a manner that maximizes the decision-support to clinician-effort ratio. While we used STS variables (often harvested post-discharge) in this study, the information contained in those variables can be collected prospectively from an institution's EMR. A well-validated model could be integrated into EMR software to automatically display readmission risk estimates, making use of the tool effectively costless to clinicians. Similar tools already exist for other adverse events (Henry et al., 2015). This would help clinicians identify patients who require attentive post-discharge care or enrollment in readmission reduction programs. Those resources could translate to reducing readmissions and cost savings.

Our RF model's performance of AUC .76 is a two-sided coin. We used modern ML algorithms and easily outperformed the all previously published approaches for the task, demonstrating that ML approaches are worthy of further study. Moreover, ML methodologies could be applied to a broad variety of diagnoses to predict readmission risk, yielding further cost savings and reducing adverse outcomes. On the other hand, an AUC of .76 may not be accurate enough for clinician adoption. The fact that our relatively sophisticated approach requires further development and validation lends support to existing criticisms (Wadhera et al., 2018) of the use of readmissions to evaluate quality-of-care, especially since existing care quality metrics (such as the Hospital Readmissions Reduction Program model) are based on simple regressions characterized by lower performance. This study presents a possible avenue for improving quality-of-care metrics which could further reduce costs and adverse outcome risk.

Finally, we meditate on more specific future directions. Recent work (Healy et al., 2020) suggests mobility is highly predictive of readmission: the proposed regression model, based on prospectively-collected ambulation profiles, achieved test set AUC of .93 (95% CI (.84, 1.0)). While Healy et al. (2020) had a small study population (n=100) and mobility was not available in our EMR, it is a strong candidate for inclusion in an improved version of our models. We also believe that it is necessary to study allocation of post-discharge resources. The literature has paid little attention to post-discharge factors like discharge location and yet our study showed it is a key predictor. It would also be valuable to determine which of the relationships we observed are causal or merely associational. A causality-minded study could help focus model improvements and further developments toward reducing readmissions.

# Appendix C

# Supplementary Material for Chapter 3

## C.1 The ID Algorithm

```
function $\mathbf{ID}(\mathbf{y}, \mathbf{x}, \mathrm{P}, \mathrm{G})$:
INPUT: $\mathbf{x}$,$\mathbf{y}$ value assignments, P a probability distribution, G a causal diagram.
OUTPUT: Expression for $P_{\mathbf{x}}(\mathbf{y})$ in terms of P 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 $\mathbf{ID}(\mathbf{y}, \mathbf{x} \cap An(\mathbf{Y})_G, \sum_{\mathbf{v} \setminus An(\mathbf{Y})_G} P, An(\mathbf{Y})_G)$.
    3) let $\mathbf{W} = (\mathbf{V} \setminus \mathbf{X}) \setminus An(\mathbf{Y})_{G_{\overline{\mathbf{x}}}}$.
       if $\mathbf{W} \neq \emptyset$, return $\mathbf{ID}(\mathbf{y}, \mathbf{x} \cup \mathbf{w}, P, G)$.
    4) if $C(G \setminus \mathbf{X}) = \{S_1, ..., S_k\}$,
       return $\sum_{\mathbf{v} \setminus (\mathbf{y} \cup \mathbf{x})} \prod_i \mathbf{ID}(s_i, \mathbf{v} \setminus s_i, P, G)$.
       if $C(G \setminus \mathbf{X}) = \{S\}$:
          5) if $C(G) = \{G\}$, throw $\mathbf{FAIL}(G, G \cap S)$.
          6) if $S \in C(G)$,
             return $\sum_{s \setminus \mathbf{y}} \prod_{\{i | V_i \in S\}} P(v_i | v_G^{(i-1)})$.
          7) if $(\exists S')S \subset S' \in C(G)$,
             return $\mathbf{ID}(\mathbf{y}, \mathbf{x} \cap S',$
             $\prod_{\{i | V_i \in S'\}} P(V_i | V_G^{(i-1)} \cap S', v_G^{(i-1)} \setminus S'), S')$.
```

Figure C-1. ID Algorithm as it appears in Shpitser and Pearl, 2006.

## C.2 Example Derivation For A Response To An Edge-Specific Policy

We seek to identify the distribution $p(Y(f_A^{(AM)\to}(W_0), f_A^{(AW_1)\to}(W_0)))$ in Fig. 3-2 (b). $\mathbf{Y}^* = \{Y, W_1, M_1, W_0\}$, and $\mathcal{D}(\mathcal{G}_{\mathbf{Y}^*}) = \{\{Y\}, \{W_0, M_1\}, \{W_1\}\}$ (the graph $\mathcal{G}_{\mathbf{Y}^*}$ is shown in Fig. 3-2 (c)). Thus, we have three terms, a term $\phi_{\{W_0,M_1,A,W_1\}}(p;\mathcal{G})$ for $Y$, a term $\phi_{\{W_0,A,M_1,Y\}}(p;\mathcal{G})$ for $W_1$, and a term $\phi_{\{A,W_1,Y\}}(p;\mathcal{G})$ for $\{W_0, M_1\}$. We have

$$\begin{aligned}
\phi_{\{W_0,A,M_1,Y\}}(p;\mathcal{G}) &= \phi_{\{W_0,A,M_1\}}\left(\sum_Y p; \mathcal{G}^{(a)}\right) \\
&= \phi_{\{W_0,A\}}\left(\frac{p(W_0,A,M_1,W_1)}{p(M_1 \mid A, W_0)}; \mathcal{G}^{(b)}\right) \\
&= \phi_{\{W_0\}}\left(\frac{p(W_0,A,M_1,W_1)}{p(M_1, A \mid W_0)}; \mathcal{G}^{(c)}\right) \\
&= p(W_1 \mid M_1, A, W_0),
\end{aligned}$$

where $\mathcal{G}^{(a)}, \mathcal{G}^{(b)}, \mathcal{G}^{(c)}$ are CADMGs in Figs. C-2 (a), (b), and (c), respectively. Similarly, $\phi_{\{W_0,M_1,A,W_1\}}(p;\mathcal{G})$ is equal to

$$\begin{aligned}
&\phi_{\{W_0,M_1,A\}}\left(\frac{p(W_0,A,M_1,W_1,Y)}{p(W_1 \mid M_1, A, W_0)}; \mathcal{G}^{(d)}\right) \\
&= \phi_{\{W_0,A\}}\left(\frac{p(W_0,A,M_1,W_1,Y)}{p(W_1, M_1 \mid A, W_0)}; \mathcal{G}^{(e)}\right) \\
&= \phi_{\{W_0\}}\left(\sum_A \frac{p(W_0,A,M_1,W_1,Y)}{p(W_1, M_1 \mid A, W_0)}; \mathcal{G}^{(f)}\right) \\
&= \sum_{W_0,A} p(W_2 \mid W_1, M_1, A, W_0) p(A, W_0),
\end{aligned}$$

where $\mathcal{G}^{(d)}, \mathcal{G}^{(e)}, \mathcal{G}^{(f)}$ are CADMGs in Figs. C-2 (d), (e), and (f), respectively. Finally,

$$\begin{aligned}
\phi_{\{A,W_1,Y\}}(p;\mathcal{G}) &= \phi_{\{A,W_1\}}\left(\sum_Y p; \mathcal{G}^{(a)}\right) \\
&= \phi_{\{A\}}\left(\sum_{Y,W_1} p; \mathcal{G}^{(g)}\right) \\
&= \frac{p(W_0,A,M_1)}{p(A \mid W_0)} = p(M_1 \mid A, W_0) p(W_0),
\end{aligned}$$

where $\mathcal{G}^{(a)}, \mathcal{G}^{(g)}$ are CADMGs in Figs. C-2 (a), and (g), respectively. Note that whenever the fixing operation for a kernel $q_{\mathbf{V}}(\mathbf{V}|\mathbf{W})$ that fixes $V \in \mathbf{V}$ is such that $\mathbf{V} \setminus \{V\} \subseteq \text{nd}_{\mathcal{G}(\mathbf{V},\mathbf{W})}(V)$, the resulting kernel can be viewed as $\tilde{q}_{\mathbf{V}\setminus\{V\}}(\mathbf{V} \setminus \{V\}|\mathbf{W} \cup \{V\}) = \sum_V q_{\mathbf{V}}(\mathbf{V}|\mathbf{W})$. We now combine these terms, evaluating $A$ to either $f_A^{(AW_1)\rightarrow}(W_0)$ or $f_A^{(AM)\rightarrow}(W_0)$, as appropriate, yielding the functional in (3.14) for $p(Y(f_A^{(AW_1)\rightarrow}(W_0), f_A^{(AM)\rightarrow}(W_0)))$, namely:

$$\begin{aligned}&\sum_{W_0,A,M,W_1} \Big[\Big[p(W_1|M, A = f_A^{(AM)\rightarrow}(W_0), W_0)\Big]\\ &\times \Big[p(M|A = f_A^{(AW_1)\rightarrow}(W_0), W_0)p(W_0)\Big]\\ &\times \Big[\sum_{W_0,A} p(Y|W_1, M, A, W_0)p(W_0, A)\Big]\Big].\end{aligned}$$

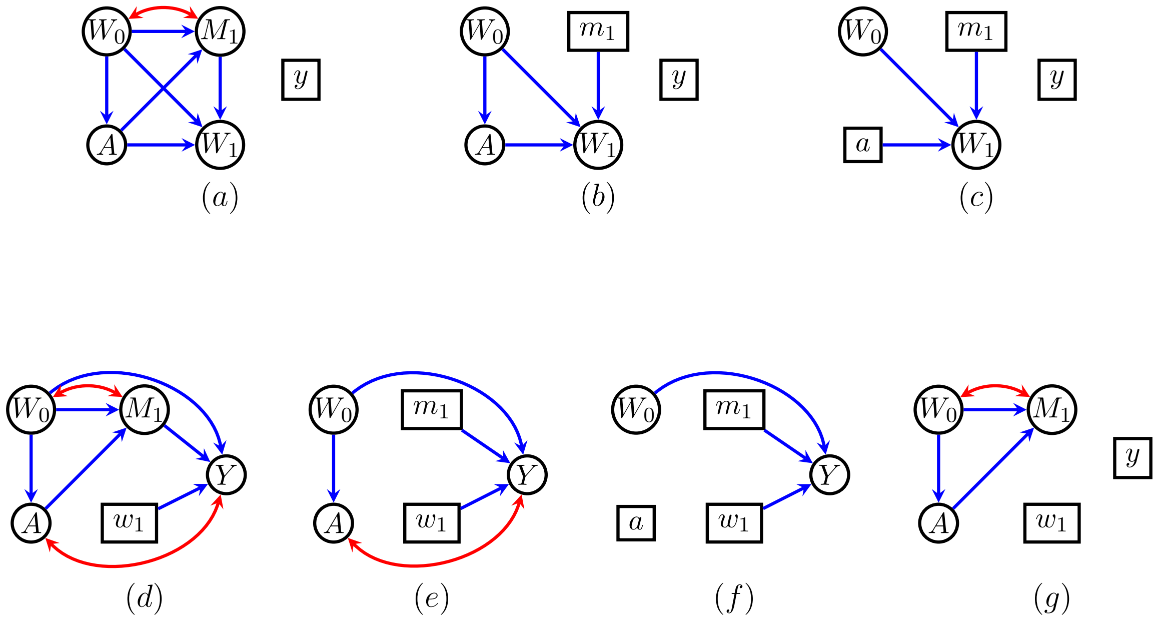


Figure C-2. CADMGs obtained from fixing in $\mathcal{G}$ shown in Fig. 3-2 (b): (a) $\phi_{\{Y\}}(\mathcal{G})$, (b) $\phi_{\{Y,M_1\}}(\mathcal{G})$, (c) $\phi_{\{Y,M_1,A\}}(\mathcal{G})$, (d) $\phi_{\{W_1\}}(\mathcal{G})$, (e) $\phi_{\{W_1,M_1\}}(\mathcal{G})$, (f) $\phi_{\{W_1,M_1,A\}}(\mathcal{G})$, (g) $\phi_{\{Y,W_1\}}(\mathcal{G})$.

## C.3 Proofs

Before giving proofs of our main results, we state the following utility lemma which will be useful throughout subsequent developments.

**Lemma 2** *Let $\mathcal{G}$ be a DAG with vertex set $\mathbf{V}$. Fix $A, B \in \mathbf{V}$ such that $B \notin \text{de}_{\mathcal{G}}(A) \setminus \text{ch}_{\mathcal{G}}(A)$ and $A \notin \text{de}_{\mathcal{G}}(B) \setminus \text{ch}_{\mathcal{G}}(B)$. Let $\mathcal{G}_{A\times B}$ be a directed graph containing vertices $(\mathbf{V} \setminus \{A, B\}) \cup Z$, and the following set of edges. First, all edges between vertices in $\mathbf{V} \setminus \{A, B\}$ in $\mathcal{G}$ also are in $\mathcal{G}_{A\times B}$. Second, for every $C \neq A, B$, for every edge of the form $C \to A$ or $C \to B$ in $\mathcal{G}$, there is an edge $C \to Z$ in $\mathcal{G}_{A\times B}$, and for every edge of the form $A \to C$ or $B \to C$ in $\mathcal{G}$, there is an edge $Z \to C$ in $\mathcal{G}_{A\times B}$. Then*

*(a) $\mathcal{G}_{A\times B}$ is a DAG.*

*(b) Any element in the causal model for $\mathcal{G}$ is an element of the causal model for $\mathcal{G}_{A\times B}$, if we interpret the Cartesian product of variables $A$ and $B$ in this element as the variable $Z$.*

*Proof:* If $\mathcal{G}_{A\times B}$ is not a DAG, there is a directed cycle involving $Z$, e.g. $W \to \ldots \to \circ \to Z \to \circ \to \ldots \to W$. Since $\mathcal{G}$ is a DAG, this implies either $\mathcal{G}$ has a pair of paths $W \to \ldots \to \circ \to A$ and $B \to \circ \to \ldots \to W$, or a pair of paths $W \to \ldots \to \circ \to B$ and $A \to \circ \to \ldots \to W$. This violates our assumption on the genealogical relationship between $A$ and $B$.

We construct the element in the causal model for $\mathcal{G}_{A\times B}$ as follows. Given the structural equation $f_A(\text{pa}_{\mathcal{G}}(A), \epsilon_A)$ for $A$, and the structural equation $f_B(\text{pa}_{\mathcal{G}}(B), \epsilon_B)$ for $B$ in some element of a causal model in $\mathcal{G}$, define the structural equation $f_Z(\text{pa}_{\mathcal{G}_{A\times B}}(Z), \epsilon_Z)$ to be the function that sets the component of $Z$ corresponding to $A$ via $f_A(\text{pa}_{\mathcal{G}}(A), \epsilon_A)$, the component of $Z$ corresponding to $B$ via $f_B(\text{pa}_{\mathcal{G}}(B), \epsilon_B)$, and where $\epsilon_Z = \epsilon_A \times \epsilon_B$.

The structural equations and independent error terms for variables other than $Z$ are inherited from the element of the causal model for $\mathcal{G}$. By construction, all error terms are independent. By definition of the structural equation model with independent errors, this gives an element in the causal model of $\mathcal{G}_{A\times B}$. □

**Corollary 5** *Fix $\mathcal{G}, A, B$ with the properties in Lemma 2. Fix any causal parameter*

*$\beta$ that is not identified in $\mathcal{G}$. If $A, B$ is reintepreted to refer to $Z = A \times B$, then $\beta$ is also not identified in $\mathcal{G}_{A \times B}$.*

*Proof:* If $\beta$ is not identified, there exist two elements in the causal model for $\mathcal{G}$ which agree on the observed data distribution, but disagree on $\beta$. The construction in the proof of Lemma 2 allows us to reinterpret those elements as elements of the causal model for $\mathcal{G}_{A \times B}$, and $\beta$ as a parameter in the causal model for $\mathcal{G}_{A \times B}$. This immediately yields two elements in the model for $\mathcal{G}_{A \times B}$ which disagree on $\beta$, but agree on the observed data distribution. □

We now give the proofs of the main results. The proof of the following result is already known. We give a version of it here to show the close relationship between proofs of other the results in this paper, and the method for proving this result.

**Theorem 4** *Given disjoint subsets $\mathbf{Y}, \mathbf{A}$ of $\mathbf{V}$ in an ADMG $\mathcal{G}$, define $\mathbf{Y}^* \equiv \text{an}_{\mathcal{G}_{\mathbf{V} \setminus \mathbf{A}}}(\mathbf{Y})$. Then $p(\mathbf{Y}(\mathbf{a}))$ is not identified if there exists $\mathbf{D} \in \mathcal{D}(\mathcal{G}_{\mathbf{Y}^*})$ that is not a reachable set in $\mathcal{G}$.*

*Proof:* Assume there exists $\mathbf{D} \in \mathcal{D}(\mathcal{G}_{\mathbf{Y}^*})$ that is not a reachable set in $\mathcal{G}$. Let $\mathbf{R} = \{D \in \mathbf{D} | \text{ch}_{\mathcal{G}}(D) \cap \mathbf{D} = \emptyset\}$, and $\mathbf{A}^* = \mathbf{A} \cap \text{pa}_{\mathcal{G}}(\mathbf{D})$. Then there exists a hedge consisting of $\mathbf{D}$ and a superset of $\mathbf{D}$ for $p(\mathbf{R}|\text{do}(\mathbf{a}^*))$, and $p(\mathbf{R}|\text{do}(\mathbf{a}^*))$ is not identified via a construction based on hedges in Shpitser and Pearl (2006).

Let $\mathbf{Y}'$ be the minimal subset of $\mathbf{Y}$ such that $\mathbf{R} \subseteq \text{an}_{\mathcal{G}_{\mathbf{V} \setminus \mathbf{A}}}(\mathbf{Y}')$. Consider an edge subgraph $\mathcal{G}^\dagger$ of $\mathcal{G}$ consisting of all edges in $\mathcal{G}$ in the hedge above, and a subset of edges on directed paths in $\mathcal{G}_{\mathbf{V} \setminus \mathbf{A}}$ from $\mathbf{R}$ to $\mathbf{Y}'$ that form a forest. Note that if $p(\mathbf{Y}'|\text{do}(\mathbf{a}^*))$ is not identified in $\mathcal{G}^\dagger$, $p(\mathbf{Y}|\text{do}(\mathbf{a}))$ is also not identified in $\mathcal{G}$, since by construction, $p(\mathbf{Y}'|\text{do}(\mathbf{a}^*)) = p(\mathbf{Y}'|\text{do}(\mathbf{a}))$, and if the marginal $p(\mathbf{Y}'|\text{do}(\mathbf{a}))$ is not identified, the joint $p(\mathbf{Y}|\text{do}(\mathbf{a}))$ is also not identified. Since $\mathcal{G}^\dagger$ is an edge subgraph of $\mathcal{G}$, $p(\mathbf{Y}|\text{do}(\mathbf{a}))$ is also not identified in $\mathcal{G}$.

We now show that $p(\mathbf{Y}'|\text{do}(\mathbf{a}^*))$ is not identified in $\mathcal{G}^\dagger$. If $\mathbf{R} \subseteq \mathbf{Y}'$, our conclusion is trivial.

If not, pick a vertex $\widetilde{Y}$ in $\mathcal{G}^{\dagger}$ such that $\mathrm{pa}_{\mathcal{G}^{\dagger}}(\widetilde{Y}) \subseteq \mathbf{R}$, and $\mathrm{pa}_{\mathcal{G}^{\dagger}}(\widetilde{Y}) \setminus \mathbf{Y}' \neq \emptyset$. Such a vertex is guaranteed to exist, since $\mathcal{G}^{\dagger}$ is acyclic and $\mathbf{R} \setminus \mathbf{Y}' \neq \emptyset$. We want to show the following subclaim: if $p(\mathbf{R}|\mathrm{do}(\mathbf{a}^*))$ is not identifiable, then $p(\mathbf{R} \setminus (\mathrm{pa}_{\mathcal{G}^{\dagger}}(\widetilde{Y}) \setminus \mathbf{Y}') \cup \widetilde{Y}|\mathrm{do}(\mathbf{a}^*))$ is also not identified. Note that in the model given by $\mathcal{G}^{\dagger}$,

$$p(\mathbf{R} \setminus (\mathrm{pa}_{\mathcal{G}^{\dagger}}(\widetilde{Y}) \setminus \mathbf{Y}') \cup \widetilde{Y}|\mathrm{do}(\mathbf{a}^*)) = \sum_{\mathrm{pa}_{\mathcal{G}^{\dagger}}(\widetilde{Y}) \setminus \mathbf{Y}'} p(\mathbf{R}|\mathrm{do}(\mathbf{a}^*)) p(\widetilde{Y}|\,\mathrm{pa}_{\mathcal{G}^{\dagger}}(\widetilde{Y}))$$

Since $p(\mathbf{R}|\mathrm{do}(\mathbf{a}^*))$ is not identified in the model corresponding to the subgraph of $\mathcal{G}^{\dagger}$ pertaining to the hedge for $p(\mathbf{R}|\mathrm{do}(\mathbf{a}^*))$, there exist two elements in this model that agree on the observed data distribution, but disagree on $p_1(\mathbf{R}|\mathrm{do}(\mathbf{a}^*))$ and $p_2(\mathbf{R}|\mathrm{do}(\mathbf{a}^*))$. In fact, the two elements constructed in Shpitser and Pearl (2006) used discrete state space variables.

Note that the right hand side expression above can be viewed, for discrete state space variables, as a linear mapping from vectors representing probabilities $p(\mathbf{R}|\mathrm{do}(\mathbf{a}^*))$ to vectors representing probabilities $p(\mathbf{R} \setminus (\mathrm{pa}_{\mathcal{G}^{\dagger}}(\widetilde{Y}) \setminus \mathbf{Y}'), \widetilde{Y}|\mathrm{do}(\mathbf{a}^*))$. To prove the subclaim, it suffices to extend the above two elements with the same distribution $p(\widetilde{Y}|\,\mathrm{pa}_{\mathcal{G}^{\dagger}}(\widetilde{Y}))$ in such a way that this linear mapping is one to one. This will ensure the two elements still agree on the observed data distribution but disagree on $p_1(\mathbf{R} \setminus (\mathrm{pa}_{\mathcal{G}^{\dagger}}(\widetilde{Y}) \setminus \mathbf{Y}'), \widetilde{Y}|\mathrm{do}(\mathbf{a}^*))$ and $p_2(\mathbf{R} \setminus (\mathrm{pa}_{\mathcal{G}^{\dagger}}(\widetilde{Y}) \setminus \mathbf{Y}'), \widetilde{Y}|\mathrm{do}(\mathbf{a}^*))$. Many such choices for $p(\widetilde{Y}|\,\mathrm{pa}_{\mathcal{G}^{\dagger}}(\widetilde{Y}))$ are possible. For example, any appropriate stochastic matrix of full column rank will suffice.

We now redefine $\mathbf{R} \equiv \mathbf{R} \setminus (\mathrm{pa}_{\mathcal{G}^{\dagger}}(\widetilde{Y}) \setminus \mathbf{Y}') \cup \widetilde{Y}$, and apply the above subclaim inductively until $\mathbf{R} \subseteq \mathbf{Y}'$. Note that if $\widetilde{Y} = Y \in \mathbf{Y}'$, we may first apply the induction to $\widetilde{Y}$ as an artificial "copy" of $Y$, and then redefine $Y$ as a Cartesian product of $Y$ and $\widetilde{Y}$, with the conclusion following by Corollary 5.

This proves the claim. □

To illustrate the operation of the proof, consider the graph in Fig. C-3 (a), where we want to show $p(Y_2|\text{do}(a))$ is not identified. First, note that $\mathbf{Y}^* = \{Y_2, M, Y_1\}$, with $\{Y_1, Y_2\}$ not reachable. This entails the hedge structure composed of two "C-forests" shown in Fig. C-3 (b) and (c), see Shpitser and Pearl (2006) for further details on how hedges are defined. The presence of the hedge structure immediately implies $p(Y_1, Y_2|\text{do}(a))$ is not identified. The inductive argument in the proof proceeds as follows. First a distribution $p(M|Y_1)$ is constructed such that $p(Y_2, M|\text{do}(a)) = \sum_{Y_1} p(M|Y_1)p(Y_1, Y_2|\text{do}(a))$ is not identified in Fig. C-3 (d). Next, a distribution $p(\widetilde{Y}_2|M)$ is constructed such that $p(\widetilde{Y}_2, Y_2|\text{do}(a)) = \sum_M p(\widetilde{Y}_2|M)p(Y_2, M|\text{do}(a))$ is not identified in Fig. C-3 (e). Finally, we use Corollary 5 to conclude non-identifiability of $p(Y_2|\text{do}(a))$ in Fig. C-3 (a) by redefining $Y_2$ in Fig. C-3 (a) to be a Cartesian product of $Y_2$ and $\widetilde{Y}_2$ in Fig. C-3 (e). This construction corresponds to Fig. C-3 (f). Note that Fig. C-3 (a) and Fig. C-3 (f) are identical up to vertex relabeling.

We next prove an analogous theorem for edge interventions. A similar proof for a closely related claim (not involving edge interventions) appeared in Shpitser (2013).

**Theorem 5** *Given $\mathbf{A}_\alpha \equiv \{A \mid (AB)_\rightarrow \in \alpha\}$, and an edge intervention given by the mapping $\mathfrak{a}_\alpha$, define $\mathbf{Y}^* \equiv \text{an}_{\mathcal{G}_{\mathbf{V}\setminus\mathbf{A}_\alpha}}(\mathbf{Y})$. The joint distribution of the counterfactual response $p(\{\mathbf{V} \setminus \mathbf{A}_\alpha\}(\mathfrak{a}_\alpha))$ is not identified if $p(\{\mathbf{V} \setminus \mathbf{A}_\alpha\}(\mathbf{a}))$ is not identified, or there exists $\mathbf{D} \in \mathcal{D}(\mathcal{G}_{\mathbf{Y}^*})$ and $A \in \mathbf{A}_\alpha$, such that $\mathfrak{a}_\alpha$ has the different value assignments for a pair of directed edges out of $A$ into $\mathbf{D}$.*

*Proof:* Assume there exists $\mathbf{D} \in \mathcal{D}(\mathcal{G}_{\mathbf{Y}^*})$ that is not a reachable set in $\mathcal{G}$, or $\mathfrak{a}_\alpha$ has different value assignments for a pair of directed edges out of $A$ into $\mathbf{D}$. Let $\mathbf{R} = \{D \in \mathbf{D}| \, \text{ch}_{\mathcal{G}}(D) \cap \mathbf{D} = \emptyset\}$, and $\mathbf{A}^* = \mathbf{A} \cap \text{pa}_{\mathcal{G}}(\mathbf{D})$. Then we have one of two cases. Either there exists a hedge consisting of $\mathbf{D}$ and a superset of $\mathbf{D}$ for $p(\mathbf{R}|\text{do}(\mathbf{a}^*))$, and $p(\mathbf{R}|\text{do}(\mathbf{a}^*))$ is not identified via a construction based on hedges in Shpitser and Pearl (2006). Or $p(\mathbf{R}(\mathfrak{a}_{\{(AD)_\rightarrow | A \in \mathbf{A}, D \in \mathbf{D}\}}))$ is not identified by counterexamples in Shpitser (2013).

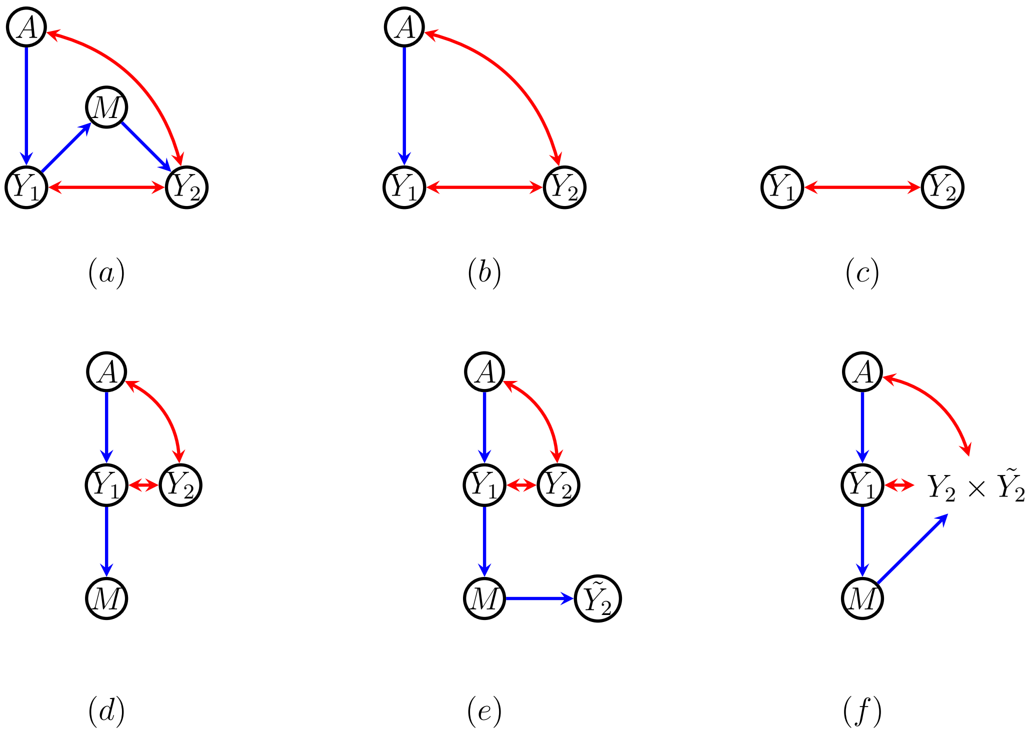


Figure C-3. (a) A graph in which we are interested in the effect of $A$ on $Y_2$, $p(Y_2|do(a))$; (b) and (c) Two forests that form a hedge with the root set $\{Y_1, Y_2\}$, $p(Y_1, Y_2|do(a))$ is not identified; (d) A subgraph illustrating the injectivity argument: $p(M, Y_2|do(a))$ is not identified; (e) Adding an artificial variable $\tilde{Y}_2$, $p(Y_2, \tilde{Y}_2|do(a))$ is not identified; (f) Joining $Y_2$ and $\tilde{Y}_2$ via the Cartesian product, $p(Y_2 \times \tilde{Y}_2|do(a))$ is not identified.

Note that $p(\mathbf{R}|\text{do}(\mathbf{a}^*))$ is equal to $p(\mathbf{R}(\mathfrak{a}^\dagger_{\{(AD)_\rightarrow|A\in\mathbf{A},D\in\mathbf{D}\}}))$, where $\mathfrak{a}^\dagger$ assigns all edges from $\mathbf{A}$ to $\mathbf{D}$ to a consistent value. As a result, in the discussions below we will unify the above two cases by assuming non-identifiability of $p(\mathbf{R}(\mathfrak{a}_{\{(AD)_\rightarrow|A\in\mathbf{A},D\in\mathbf{D}\}}))$, for some $\mathfrak{a}$.

We now proceed as before. Let $\mathbf{Y}'$ be the minimal subset of $\mathbf{Y}$ such that $\mathbf{R} \subseteq \text{an}_{\mathcal{G}_{\mathbf{V}\setminus\mathbf{A}}}(\mathbf{Y}')$. Consider an edge subgraph $\mathcal{G}^\dagger$ of $\mathcal{G}$ consisting of all edges in $\mathcal{G}$ in the recanting district or hedge above, and a subset of edges on directed paths in $\mathcal{G}_{\mathbf{V}\setminus\mathbf{A}}$ from $\mathbf{R}$ to $\mathbf{Y}'$ that form a forest. Note that if $p(\mathbf{Y}'(\mathfrak{a}_{\{(AD)_\rightarrow|A\in\mathbf{A},D\in\mathbf{D}\}}))$ is not identified in $\mathcal{G}^\dagger$, $p(\mathbf{Y}(\mathfrak{a}_\alpha))$ is also not identified in $\mathcal{G}$, since by construction, $p(\mathbf{Y}'(\mathfrak{a}_{\{(AD)_\rightarrow|A\in\mathbf{A},D\in\mathbf{D}\}})) = p(\mathbf{Y}'(\mathfrak{a}_\alpha))$, and if the marginal $p(\mathbf{Y}'(\mathfrak{a}_\alpha))$ is not identified, the joint $p(\mathbf{Y}(\mathfrak{a}_\alpha))$ is also not identified. Since $\mathcal{G}^\dagger$ is an edge subgraph of $\mathcal{G}$, $p(\mathbf{Y}(\mathfrak{a}_\alpha))$

is also not identified in $\mathcal{G}$.

We now show that $p(\mathbf{Y}'(\mathfrak{a}_{\{(AD)_{\rightarrow}|A\in\mathbf{A},D\in\mathbf{D}\}}))$ is not identified in $\mathcal{G}^{\dagger}$. If $\mathbf{R} \subseteq \mathbf{Y}'$, our conclusion is trivial.

If not, pick a vertex $\widetilde{Y}$ in $\mathcal{G}^{\dagger}$ such that $\mathrm{pa}_{\mathcal{G}^{\dagger}}(\widetilde{Y}) \subseteq \mathbf{R}$, and $\mathrm{pa}_{\mathcal{G}^{\dagger}}(\widetilde{Y}) \setminus \mathbf{Y}' \neq \emptyset$. Such a vertex is guaranteed to exist, since $\mathcal{G}^{\dagger}$ is acyclic and $\mathbf{R} \setminus \mathbf{Y}' \neq \emptyset$. We want to show the following subclaim: if $p(\mathbf{R}(\mathfrak{a}_{\{(AD)_{\rightarrow}|A\in\mathbf{A},D\in\mathbf{D}\}}))$ is not identifiable, then $p(\{\mathbf{R} \setminus (\mathrm{pa}_{\mathcal{G}^{\dagger}}(\widetilde{Y}) \setminus \mathbf{Y}') \cup \widetilde{Y}(\mathfrak{a}_{\{(AD)_{\rightarrow}|A\in\mathbf{A},D\in\mathbf{D}\}}))$ is also not identified. Note that in the model given by $\mathcal{G}^{\dagger}$,

$$p(\{\mathbf{R} \setminus (\mathrm{pa}_{\mathcal{G}^{\dagger}}(\widetilde{Y}) \setminus \mathbf{Y}') \cup \widetilde{Y}\}(\mathfrak{a}_{\{(AD)_{\rightarrow}|A\in\mathbf{A},D\in\mathbf{D}\}})) = \sum_{\mathrm{pa}_{\mathcal{G}^{\dagger}}(\widetilde{Y}) \setminus \mathbf{Y}'} p(\mathbf{R}(\mathfrak{a}_{\{(AD)_{\rightarrow}|A\in\mathbf{A},D\in\mathbf{D}\}}))p(\widetilde{Y} |\, \mathrm{pa}_{\mathcal{G}^{\dagger}}(\widetilde{Y}))$$

Since $p(\mathbf{R}(\mathfrak{a}_{\{(AD)_{\rightarrow}|A\in\mathbf{A},D\in\mathbf{D}\}}))$ is not identified in the model corresponding to the appropriate subgraph of $\mathcal{G}^{\dagger}$ pertainining to $p(\mathbf{R}(\mathfrak{a}_{\{(AD)_{\rightarrow}|A\in\mathbf{A},D\in\mathbf{D}\}}))$, there exist two elements in this model that agree on the observed data distribution, but disagree on $p_1(\mathbf{R}(\mathfrak{a}_{\{(AD)_{\rightarrow}|A\in\mathbf{A},D\in\mathbf{D}\}}))$ and $p_2(\mathbf{R}(\mathfrak{a}_{\{(AD)_{\rightarrow}|A\in\mathbf{A},D\in\mathbf{D}\}}))$. In fact, the two elements constructed in Shpitser (2013) and Shpitser and Pearl (2006) used discrete state space variables.

Note that the right hand side expression above can be viewed, for discrete state space variables, as a linear mapping from vectors representing probabilities $p(\mathbf{R}(\mathfrak{a}_{\{(AD)_{\rightarrow}|A\in\mathbf{A},D\in\mathbf{D}\}}))$ to vectors representing probabilities $p(\{\mathbf{R}\setminus(\mathrm{pa}_{\mathcal{G}^{\dagger}}(\widetilde{Y})\setminus\mathbf{Y}'), \widetilde{Y}\}(\mathfrak{a}_{\{(AD)_{\rightarrow}|A\in\mathbf{A},D\in}$ To prove the subclaim, it suffices to extend the above two elements with the same distribution $p(\widetilde{Y} |\, \mathrm{pa}_{\mathcal{G}^{\dagger}}(\widetilde{Y}))$ in such a way that this linear mapping is one to one. This will ensure, the two elements still agree on the observed data distribution, but disagree on $p_1(\{\mathbf{R} \setminus (\mathrm{pa}_{\mathcal{G}^{\dagger}}(\widetilde{Y}) \setminus \mathbf{Y}'), \widetilde{Y}\}(\mathfrak{a}_{\{(AD)_{\rightarrow}|A\in\mathbf{A},D\in\mathbf{D}\}}))$ and $p_2(\{\mathbf{R} \setminus (\mathrm{pa}_{\mathcal{G}^{\dagger}}(\widetilde{Y}) \setminus \mathbf{Y}'), \widetilde{Y}\}(\mathfrak{a}_{\{(AD)_{\rightarrow}|A\in\mathbf{A},D\in\mathbf{D}\}}))$. Many such choices for $p(\widetilde{Y} |\, \mathrm{pa}_{\mathcal{G}^{\dagger}}(\widetilde{Y}))$ are possible. For example, any appropriate stochastic matrix of full column rank will suffice.

We now redefine $\mathbf{R} \equiv \mathbf{R} \setminus (\mathrm{pa}_{\mathcal{G}^{\dagger}}(\widetilde{Y}) \setminus \mathbf{Y}') \cup \widetilde{Y}$, and apply the above subclaim inductively until $\mathbf{R} \subseteq \mathbf{Y}'$. As before, whenever $\widetilde{Y} = Y \in \mathbf{Y}'$, we redefine $Y$ as a Cartesian product of $\widetilde{Y}$ and $Y$, with the conclusion following by Corollary 5.

This proves the claim. $\square$

We illustrate the two problematic structures that create non-identifiability of $p(Y((aY)_{\rightarrow}, (a'M)_{\rightarrow})) = p(Y(a, M(a')))$ in Fig. C-4 (a) and (b). In (a), the recanting district criterion does not hold, however, $p(Y|\text{do}(a))$ is not identified. In (b), $p(Y|\text{do}(a))$ is identified, but the recanting district criterion fails, since $Y$ and $M$ form a district, but the edge intervention assigns $A$ to different values for different edges from $A$ into the district. The inductive part of the argument in Theorem 5 is identical to that in Theorem 4.

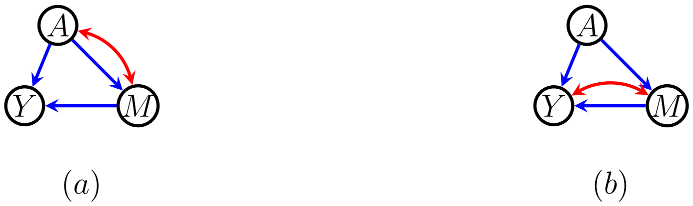


Figure C-4. An example of the two problematic structures that prevent identification of $p(Y((aY)_{\rightarrow}, (a'M)_{\rightarrow}))$. (a) There is a hedge structure preventing identification of $p(Y|\text{do}(a))$. (b) The recanting district criterion holds.

Next, we give a completeness results for responses to arbitrary, possibly stochastic policies. This result is new and shows the algorithm in Tian (2008) is complete for unrestricted policies.

**Theorem 6** *Define $\mathcal{G}_{\mathbf{f_A}}$ to be a graph obtained from $\mathcal{G}$ by removing all edges into $\mathbf{A}$, and adding for any $A \in \mathbf{A}$, directed edges from $\mathbf{W}_A$ to $A$. Define $\mathbf{Y}^* \equiv \text{an}_{\mathcal{G}_{\mathbf{f_A}}}(\mathbf{Y}) \setminus \mathbf{A}$. Then if $p(\mathbf{Y}^*(\mathbf{a}))$ is not identified in $\mathcal{G}$, $p(\mathbf{Y}(\mathbf{f_A}))$ is not identified in $\mathcal{G}$ if $\mathbf{f_A}$ is the unrestricted class of policies.*

*Proof:* Assume there exists $\mathbf{D} \in \mathcal{D}(\mathcal{G}_{\mathbf{Y}^*})$ that is not a reachable set in $\mathcal{G}$. Let $\mathbf{R} = \{D \in \mathbf{D} | \text{ch}_{\mathcal{G}}(D) \cap \mathbf{D} = \emptyset\}$, and $\mathbf{A}^* = \mathbf{A} \cap \text{pa}_{\mathcal{G}}(\mathbf{D})$. Then there exists a hedge consisting of $\mathbf{D}$ and a superset of $\mathbf{D}$ for $p(\mathbf{R}|\text{do}(\mathbf{a}^*))$, and $p(\mathbf{R}|\text{do}(\mathbf{a}^*))$ is not identified via a construction based on hedges in Shpitser and Pearl (2006).

Because $\mathcal{G}_{\mathbf{V} \setminus \mathbf{A}}$ is an edge subgraph of $\mathcal{G}_{\mathbf{f_A}}$, there is some element $\mathbf{D}' \in \mathcal{D}(\mathcal{G}_{\text{an}_{\mathcal{G}_{\mathbf{V} \setminus \mathbf{A}}}(\mathbf{Y})})$ that is a subset of $\mathbf{D}$. If $\mathbf{D} = \mathbf{D}'$, it suffices to consider policies that set $\mathbf{A}^*$ to constants, and our proof is immediate by the argument in Theorem 4.

Otherwise, we proceed as follows. Let $\mathbf{Y}'$ be the minimal subset of $\mathbf{Y}$ such that $\mathbf{R} \subseteq \text{an}_{\mathcal{G}_{\mathbf{f_A}}}(\mathbf{Y}')$. Consider an edge subgraph $\mathcal{G}^\dagger$ of $\mathcal{G}_{\mathbf{f_A}}$ consisting of all edges in $\mathcal{G}$ in the hedge above, and a subset of edges on directed paths in $\mathcal{G}_{\mathbf{f_A}}$ from $\mathbf{R}$ to $\mathbf{Y}'$ that form a forest. Note that unlike previous proofs, these directed paths may intersect $\mathbf{A}$ due to the addition of edges to $\mathcal{G}_{\mathbf{f_A}}$ from $\mathbf{W}_A$ to $A \in \mathbf{A}$. Let $\mathbf{A}^\dagger$ be the set $\mathbf{A}^*$ and all elements in $\mathbf{A}$ in $\mathcal{G}^\dagger$.

For every $A^\dagger \in \mathbf{A}^\dagger$, we restrict attention to policies that map from $\mathbf{W}^\dagger_{A^\dagger}$ to $A^\dagger$, where $\mathbf{W}^\dagger_{A^\dagger}$ is $\mathbf{W}_{A^\dagger}$ intersected with vertices in $\mathcal{G}^\dagger$.

Note that if $p(\mathbf{Y}'(\{A^\dagger = f_{A^\dagger}(\mathbf{W}^\dagger_{A^\dagger}) | A^\dagger \in \mathbf{A}^\dagger\}))$ is not identified in $\mathcal{G}^\dagger$, $p(\mathbf{Y}(\mathbf{f_A}))$ is also not identified in $\mathcal{G}$, since by construction, $p(\mathbf{Y}'(\{A^\dagger = f_{A^\dagger}(\mathbf{W}^\dagger_{A^\dagger}) | A^\dagger \in \mathbf{A}^\dagger\})) = p(\mathbf{Y}'(\mathbf{f_A}))$ in $\mathcal{G}^\dagger$, and if the marginal $p(\mathbf{Y}'(\mathbf{f_A}))$ is not identified, the joint $p(\mathbf{Y}(\mathbf{f_A}))$ is also not identified. Since $\mathcal{G}^\dagger$ is an edge subgraph of $\mathcal{G}$, $p(\mathbf{Y}(\mathbf{f_A}))$ is also not identified in $\mathcal{G}$.

We now show that $p(\mathbf{Y}'(\{A^\dagger = f_{A^\dagger}(\mathbf{W}^\dagger_{A^\dagger}) | A^\dagger \in \mathbf{A}^\dagger\}))$ is not identified in $\mathcal{G}^\dagger$.

If $\mathbf{R} \subseteq \mathbf{Y}'$, it immediately implies the case above where $\mathbf{D} = \mathbf{D}'$, and we are done by Theorem 4. If not, we proceed inductively, as before. Pick a vertex $\widetilde{Y}$ in $\mathcal{G}^\dagger$ such that $\text{pa}_{\mathcal{G}^\dagger}(\widetilde{Y}) \subseteq \mathbf{R}$, and $\text{pa}_{\mathcal{G}^\dagger}(\widetilde{Y}) \setminus \mathbf{Y}' \neq \emptyset$. Such a vertex is guaranteed to exist, since $\mathcal{G}^\dagger$ is acyclic and $\mathbf{R} \setminus \mathbf{Y}' \neq \emptyset$. We now have two cases, $\widetilde{Y} \notin \mathbf{A}^*$ or $\widetilde{Y} \in \mathbf{A}^*$. In the former case, we use the inductive argument from Theorem 4.

Note, in particular, that if $\widetilde{Y} \in \mathbf{A}^\dagger \setminus \mathbf{A}^*$, we simply treat $\widetilde{Y}$ as an ordinary variable, and it's policy as an ordinary conditional distribution. A special argument isn't necessary here since $\widetilde{Y}$ does not intersect the original hedge structure for $\mathbf{D}$.

Now consider the latter case, where $\widetilde{Y} \in \mathbf{A}^*$. This case we simply create copies of variables on the path $\widetilde{Y} \to W_1 \to \ldots \to W_k \to \widetilde{Y}' \in \mathbf{Y}'$ in $\mathcal{G}^\dagger$, yielding a graph $\widetilde{\mathcal{G}}^\dagger$. We extend the previous inductive argument by considering an "extended" observed data joint distribution where conditional distributions of $\{W_1, \ldots, W_k, \widetilde{Y}\} \cap \mathbf{A}^*$ given their parents are specified by appropriate policies in $\mathbf{f}_{\mathbf{A}}$. For the unrestricted policy class, the inductive argument again implies that

$$p(\{\mathbf{R} \setminus (\text{pa}_{\mathcal{G}^\dagger}(\widetilde{Y}) \setminus \mathbf{Y}'), \widetilde{Y}'\}(\mathbf{a}^*_{\mathbf{A}^* \setminus \{\widetilde{Y}\}})) = \sum_{(\mathbf{a}^*_{\widetilde{Y}} \cup \text{pa}_{\mathcal{G}^\dagger \cup \{W_1, \ldots W_k\}}(\widetilde{Y})) \setminus \mathbf{Y}'} p(\mathbf{R}|\text{do}(\mathbf{a}^*)) p(\widetilde{Y}'|W_k) p(W_1|\widetilde{Y}) \prod_{i=2}^{k} p(W_i|W_{i-1}) \tilde{p}(\widetilde{Y} = \mathbf{a}^*_{\widetilde{Y}} | \text{pa}_{\mathcal{G}^\dagger}(\widetilde{Y}))$$

is not identified in $\widetilde{\mathcal{G}}^\dagger$ if $p(\mathbf{R}|\text{do}(\mathbf{a}^*))$ is not identified in $\widetilde{\mathcal{G}}^\dagger$.

We now inductively apply Lemma 2 to construct elements in $\mathcal{G}^\dagger$ where $p(\{\mathbf{R} \setminus (\text{pa}_{\mathcal{G}^\dagger}(\widetilde{Y}) \setminus \mathbf{Y}'), Y'\}(\mathbf{a}^*_{\mathbf{A}^* \setminus \{\widetilde{Y}\}}))$ is not identified by Corollary 5.

We now redefine $\mathbf{R} \equiv \mathbf{R} \setminus (\text{pa}_{\mathcal{G}^\dagger}(\widetilde{Y}))$, and $\mathbf{A}^* \equiv \mathbf{A}^* \setminus \{\widetilde{Y}\}$. The induction terminates when $\mathbf{A}^* = \emptyset$ and $\mathbf{R} \subseteq \mathbf{Y}'$, yielding our conclusion.

□

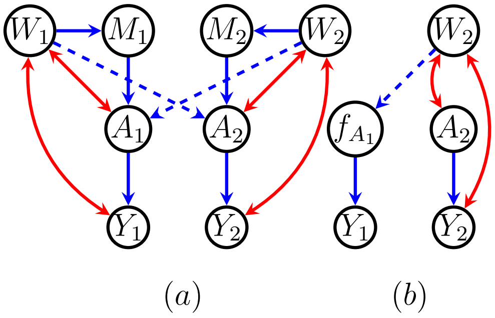


Figure C-5. (a) A graph in which we're interested in the distribution $p(\{Y_1, Y_2\}(A_1 = f_{A_1}(W_2), A_2 = f_{A_2}(W_1)))$. (b) A subgraph of $\mathcal{G}_{\mathbf{Y}^*}$ for the given counterfactual which shows the hedge structure, and the form of the inductive argument which yields non-identification.

We illustrate the novel ideas in this proof via Fig. C-5 (a) and (b), where we are interested in identification of $p(\{Y_1, Y_2\}(A_1 = f_{A_1}(W_2), A_2 = f_{A_2}(W_1)))$. First, note that the fact that $A_1$ is determined by $W_2$ via $f_{A_1}$ and $A_2$ is determined by $W_1$ via $f_{A_2}$ implies the set $\mathbf{Y}^* = \{Y_1, Y_2, A_1, A_2, W_1, W_2\}$ is larger than it would have been had we been interested in $p(Y_1, Y_2|\text{do}(a_1, a_2))$, in which case $\mathbf{Y}^*$ would be equal to $\{Y_1, Y_2\}$. Second, note that $p(Y_1, Y_2|\text{do}(a_1, a_2))$ is identified in this graph, while $p(\{Y_1, Y_2\}(A_1 = f_{A_1}(W_2), A_2 = f_{A_2}(W_1)))$ is not. Specifically, the subgraph shown in Fig. C-5 (b) contains the hedge structure for $p(Y_2, W_2|\text{do}(a_2))$, along with a path $W_2 \to f_{A_1} \to Y_1$ which yields the inductive argument showing non-identification.

For this example, it sufficed to consider a trivial policy for $A_2$ which always sets $A_2$ to a constant. However, the policy $f_{A_1}$ needed to dependent on $W_2$ in order to allow the inductive argument to go through showing that if $p(Y_2, W_2|\text{do}(a_2))$ is not identified, $p(\{Y_2, Y_1\}(a_2, A_1 = f_{A_1}(W_2)))$ is also not identified.

Finally, we give an argument for completeness, for unrestricted policies, of the identification algorithm for responses to edge-specific policies. The following proof can be viewed as a generalization of the arguments in Theorems 5 and 6. This result is also new.

**Theorem 7** *Define the graph $\mathcal{G}_{\mathfrak{f}_\alpha}$ to be one where all edges with arrowheads into $\mathbf{A}_\alpha$ are removed, and directed edges from any vertex in $\mathbf{W}_A$ to $A \in \mathbf{A}_\alpha$ added. Fix a set $\mathbf{Y}$ of outcomes of interest, and define $\mathbf{Y}^*$ equal $\text{an}_{\mathcal{G}_{\mathfrak{f}_\alpha}}(\mathbf{Y}) \setminus \mathbf{A}_\alpha$. Then if $p(\mathbf{Y}^*(\mathbf{a}))$ is not identified, or there exists $\mathbf{D} \in \mathcal{D}((\mathcal{G}_{\mathfrak{f}_\alpha})_{\mathbf{Y}^*})$, such that $\mathfrak{f}_\alpha$ yields different policy assignments for two edges from $A \in \mathbf{A}_\alpha$ to $\mathbf{D}$, $p(\mathbf{Y}(\mathfrak{f}_\alpha))$ is not identified.*

*Proof:* Assume there exists $\mathbf{D} \in \mathcal{D}(\mathcal{G}_{\mathbf{Y}^*})$ that is not a reachable set in $\mathcal{G}$, or $\mathfrak{f}_\alpha$ has the different policy assignments for a pair of directed edges out of $A$ into $\mathbf{D}$. Let $\mathbf{R} = \{D \in \mathbf{D}|\, \text{ch}_{\mathcal{G}}(D) \cap \mathbf{D} = \emptyset\}$, and $\mathbf{A}^* = \mathbf{A} \cap \text{pa}_{\mathcal{G}}(\mathbf{D})$. Then we have one of two cases. Either there exists a hedge consisting of $\mathbf{D}$ and a superset of $\mathbf{D}$ for $p(\mathbf{R}|\text{do}(\mathbf{a}^*))$, and $p(\mathbf{R}|\text{do}(\mathbf{a}^*))$ is not identified via a construction based on hedges in Shpitser and Pearl (2006); or $p(\mathbf{R}(\mathfrak{f}_{\{(AD)_\to | A \in \mathbf{A}, D \in \mathbf{D}\}}))$ is not identified by counterexamples in Shpitser

(2013) (i.e., the "recanting district criterion").

Because $\mathcal{G}_{\mathbf{V}\setminus\mathbf{A}}$ is an edge subgraph of $\mathcal{G}_{\mathfrak{f}_{\mathbf{A}}}$, there is some element $\mathbf{D}' \in \mathcal{D}(\mathcal{G}_{\mathrm{an}_{\mathcal{G}_{\mathbf{V}\setminus\mathbf{A}}}(\mathbf{Y})})$ that is a subset of $\mathbf{D}$. If $\mathbf{D} = \mathbf{D}'$, it suffices to consider interventions that set the all edges out of $\mathbf{A}^*$ to the same policy and our proof follows from the argument in Theorem 6.

Additionally, note that $p(\mathbf{R}|\mathrm{do}(\mathbf{a}^*))$ is equal to $p(\mathbf{R}(\mathfrak{f}^{\dagger}_{\{(AD_{\rightarrow}|A\in\mathbf{A},D\in\mathbf{D}\})}))$, where $\mathfrak{f}^{\dagger}$ assigns all edges from $\mathbf{A}$ to $\mathbf{D}$ to a consistent value. As a result, we can unify the two cases above (hedge and recanting district) by assuming non-identifiability of $p(\mathbf{R}(\mathfrak{f}_{\{(AD_{\rightarrow}|A\in\mathbf{A},D\in\mathbf{D}\})}))$ for some policy set $\mathfrak{f}$.

We now proceed as before. Let $\mathbf{Y}'$ be the minimal subset of $\mathbf{Y}$ such that $\mathbf{R} \subseteq \mathrm{an}_{\mathcal{G}_{\mathfrak{f}_\alpha}}(\mathbf{Y}')$. Consider an edge subgraph $\mathcal{G}^{\dagger}$ of $\mathcal{G}_{\mathfrak{f}_\alpha}$ consisting of all edges in $\mathcal{G}_{\mathfrak{f}_\alpha}$ in the hedge above, and a subset of edges on directed paths in $\mathcal{G}_{\mathfrak{f}_\alpha}$ from $\mathbf{R}$ to $\mathbf{Y}'$ that form a forest. As in Theorem 6, these directed paths may intersect $\mathbf{A}$ due to the addition of edges in $\mathcal{G}_{\mathfrak{f}_\alpha}$ from $\mathbf{W}_{\mathbf{A}}$ to $A \in \mathbf{A}$. Let $\mathbf{A}^{\dagger}$ be the union of the set $\mathbf{A}^*$ and all elements that are in $\mathbf{A}$ in $\mathcal{G}^{\dagger}$. For every $A^{\dagger} \in \mathbf{A}^{\dagger}$ we restrict attention to policies that map values of $\mathbf{W}^{\dagger}_{A^{\dagger}}$ to $A^{\dagger}$, where $\mathbf{W}^{\dagger}_{A^{\dagger}}$ is $\mathbf{W}_{A^{\dagger}}$ intersected with the vertices in $\mathcal{G}^{\dagger}$.

Note that if $p(\mathbf{Y}'(\{A^{\dagger} = \mathfrak{f}_{\{(AD)_{\rightarrow}|A\in\mathbf{A},D\in\mathbf{D}\}}(\mathbf{W}^{\dagger}_{A^{\dagger}})|A^{\dagger} \in \mathbf{A}^{\dagger}\}))$ is not identified in $\mathcal{G}^{\dagger}$, $p(\mathbf{Y}(\mathfrak{f}_\alpha))$ is also not identified in $\mathcal{G}$. This is because, by construction, $p(\mathbf{Y}'(\{A^{\dagger} = \mathfrak{f}_{\{(AD)_{\rightarrow}|A\in\mathbf{A},D\in\mathbf{D}\}}(\mathbf{W}^{\dagger}_{A^{\dagger}})|A^{\dagger} \in \mathbf{A}^{\dagger}\})) = p(\mathbf{Y}'(\mathfrak{f}_\alpha))$ in $\mathcal{G}^{\dagger}$, and if the marginal $p(\mathbf{Y}'(\mathfrak{f}_\alpha))$ is not identified the joint $p(\mathbf{Y}(\mathfrak{f}_\alpha))$ is also not identified in $\mathcal{G}^{\dagger}$. Because $\mathcal{G}^{\dagger}$ is an edge subgraph of $\mathcal{G}$, $p(\mathbf{Y}(\mathfrak{f}_\alpha))$ is also not identifiable in $\mathcal{G}$.

We now show that

$$p(\mathbf{Y}'(\{A^{\dagger} = \mathfrak{f}_{\{(AD)_{\rightarrow}|A\in\mathbf{A},D\in\mathbf{D}\}}(\mathbf{W}^{\dagger}_{A^{\dagger}})|A^{\dagger} \in \mathbf{A}^{\dagger}\}))$$

is not identified in $\mathcal{G}^{\dagger}$. Note that if $\mathbf{R} \subseteq \mathbf{Y}'$, we are done since this implies $\mathbf{D} = \mathbf{D}'$ which implies we can simply apply Theorem 6 as described above.

If $\mathbf{R} \not\subseteq \mathbf{Y}'$, pick a vertex $\tilde{Y}$ in $\mathcal{G}^{\dagger}$ such that $\mathrm{pa}_{\mathcal{G}^{\dagger}}(\tilde{Y}) \subseteq \mathbf{R}$ and $\mathrm{pa}_{\mathcal{G}^{\dagger}}(\tilde{Y}) \setminus \mathbf{Y}' \neq \emptyset$.

Such a vertex is guaranteed to exist since $\mathcal{G}^\dagger$ is acyclic and $\mathbf{R} \setminus \mathbf{Y}' \neq \emptyset$. We now have two cases, $\widetilde{Y} \notin \mathbf{A}^*$ or $\widetilde{Y} \in \mathbf{A}^*$. In the former case, we use the inductive argument from Theorem 5. In particular, if $\widetilde{Y} \in \mathbf{A}^\dagger \setminus \mathbf{A}^*$, we treat $\widetilde{Y}$ as an ordinary variable, and the element of $\mathfrak{f}$ pertaining to $\widetilde{Y}$ and its outgoing edge in $\mathcal{G}^\dagger$ as an ordinary distribution with the properties that yield an injective map. This element of $\mathfrak{f}$ is then used to obtain non-identification in the inductive step corresponding to $\widetilde{Y}$. A special argument isn't necessary here since $\widetilde{Y}$ does not intersect the original hedge structure for $\mathbf{D}$.

Now consider the latter case, where $\widetilde{Y} \in \mathbf{A}^*$. We apply the same argument as in Theorem 6. We create copies of variables on the path $\widetilde{Y} \to W_1 \to \ldots \to W_k \to \widetilde{Y}' \in \mathbf{Y}'$ in $\mathcal{G}^\dagger$, yielding a graph $\widetilde{\mathcal{G}}^\dagger$. We extend the previous inductive argument by considering an "extended" observed data joint distribution where conditional distributions of $\{W_1, \ldots, W_k, \widetilde{Y}\} \cap \mathbf{A}^*$ given their parents are specified by appropriate policies in $\mathfrak{f}_{\mathbf{A}}$. For the unrestricted policy class, the inductive argument again implies that

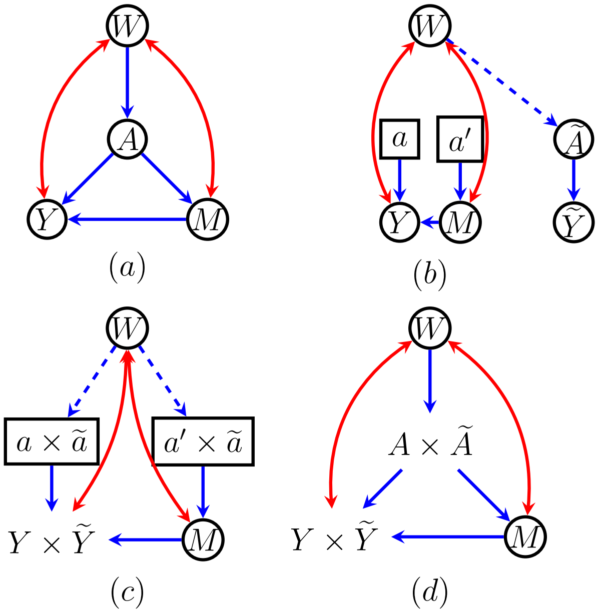
$$\begin{aligned} p(\{\mathbf{R} \setminus (\text{pa}_{\mathcal{G}^\dagger}(\widetilde{Y}) \setminus \mathbf{Y}'), \widetilde{Y}'\}(\mathbf{a}^*_{\mathbf{A}^* \setminus \{\widetilde{Y}\}})) = \\ \sum_{(\mathbf{a}^*_{\widetilde{Y}} \cup \text{pa}_{\mathcal{G}^\dagger \cup \{W_1, \ldots W_k\}}(\widetilde{Y})) \setminus \mathbf{Y}'} p(\mathbf{R} | \text{do}(\mathbf{a}^*)) p(\widetilde{Y}' | W_k) p(W_1 | \widetilde{Y}) \\ \prod_{i=2}^{k} p(W_i | W_{i-1}) \tilde{p}(\widetilde{Y} = \mathbf{a}^*_{\widetilde{Y}} | \operatorname{pa}_{\mathcal{G}^\dagger}(\widetilde{Y})) \end{aligned}$$

is not identified in $\widetilde{\mathcal{G}}^\dagger$ if $p(\mathbf{R}|do(\mathbf{a}^*))$ is not identified in $\widetilde{\mathcal{G}}^\dagger$ by Corollary 5.

Note that this construction yields a composite variable $Z$ corresponding to $\widetilde{Y}$ and its copy, where the original version of the variable has a policy that unconditionally assigns outgoing edges to different values, while the copied version of the variable has a policy that conditionally assigns a value based on $\text{pa}_{\mathcal{G}^\dagger}(\widetilde{Y})$ that is consistent across all outgoing edges in $\mathcal{G}^\dagger$. This somewhat unnatural policy is nevertheless within the unrestricted class of edge-specific policies.

We redefine $\mathbf{R} \equiv \mathbf{R} \setminus (\text{pa}_{\mathcal{G}^\dagger}(\widetilde{Y}))$, and $\mathbf{A}^* \equiv \mathbf{A}^* \setminus \{\widetilde{Y}\}$. The induction terminates when $\mathbf{A}^* = \emptyset$ and $\mathbf{R} \subseteq \mathbf{Y}'$, yielding our conclusion. □

We illustrate the novel ideas in this proof via the example in Fig. C-6 (a), where we are interested in $p(Y(\mathfrak{f}_{\{(AY)_{\rightarrow},(AM)_{\rightarrow}\}}))$, where $\mathfrak{f}$ sets $A$ according to $f^{(AY)_{\rightarrow}}(W)$ for the purposes of $(AY)_{\rightarrow}$, and to $f^{(AM)_{\rightarrow}}(W)$ for the purposes of $(AM)_{\rightarrow}$. In this example, it suffices to construct a subgraph, shown in Fig. C-6 (b), containing a recanting district along with a path from $W$ to $\widetilde{Y}$, a copy of $Y$. Note that in this subgraph there are three versions of the $A$ variable. Two versions represent conflicting value settings corresponding to different edges from $A$ into a district $\{W, M, Y\}$. This is necessary to demonstrate the existence of the recanting district structure. The third version of $A$ is set according to the mapping from $W$, and it's necessary in order to run the inductive argument which says if $p(Y, M, W((aY)_{\rightarrow}, (aM)_{\rightarrow}))$ is not identified in Fig. C-6 (b), neither is $p(Y, M, \widetilde{Y}((aY)_{\rightarrow}, (aM)_{\rightarrow}))$. Merging the appropriate variables yields Fig. C-6 (c), which demonstrates the edge-specific policy for $A$ that is not identified. Finally, the observed data version of the graph in Fig. C-6 (c) is Fig. C-6 (d), which is identical to Fig. C-6 (a) up to vertex relabeling.



$(a)$ $(b)$ $(c)$ $(d)$

Figure C-6. (a) A graph in which we are interested in $p(Y(\mathfrak{f}_{\{(AY)_{\rightarrow},(AM)_{\rightarrow}\}}))$, where $\mathfrak{f}$ sets $A$ according to $f^{(AY)_{\rightarrow}}(W)$ for the purposes of $(AY)_{\rightarrow}$, and to $f^{(AM)_{\rightarrow}}(W)$ for the purposes of $(AM)_{\rightarrow}$. (b) The graph demonstrating the problematic recanting district structure $\{Y, M, W\}$ where $A$ is set to different values unconditionally for different edges into the district, along with a path from $W$ to $\widetilde{Y}$, yielding an inductive argument of non-identification. (c) A version of the graph in (b) where variables are merged, and the effect of the $A$ edge-specific policy on $Y$ is still not identified. (d) The graph isomorphic to (a) up to vertex relabeling which shows non-identification of $p(Y(\mathfrak{f}_{\{(AY)_{\rightarrow},(AM)_{\rightarrow}\}}))$.

# Appendix D

# Supplementary Material for Chapter 4

## D.1 Proofs

**Theorem** 8 *If $p(\mathbf{V} \cup \mathbf{H})$ obeys the CG factorization relative to $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$, and $\mathbf{H}$ is block-safe then $p(\mathbf{V})$ obeys the segregated factorization relative to the segregated projection $\mathcal{G}(\mathbf{V})$.*

*Proof:* Assume the premise of the theorem. Then, $p(\mathbf{O} \cup \mathbf{H}) = \prod_{\mathbf{B} \in \mathcal{B}(\mathcal{G})} p(\mathbf{B} | \operatorname{pa}_{\mathcal{G}}(\mathbf{B}))$.

For every $\mathbf{D} \in \mathcal{D}(\mathcal{G}(\mathbf{V}))$, let $\mathbf{H}_{\mathbf{D}} \equiv \mathbf{H} \cap \operatorname{an}_{\mathcal{G}_{\mathbf{D} \cup \mathbf{H}}}(\mathbf{D})$. Then $p(\mathbf{V})$ is equal to

$$\begin{aligned}
&\sum_{\mathbf{H}} \left( \prod_{\mathbf{B} \in \mathcal{B}^{nt}(\mathcal{G})} p(\mathbf{B} | \operatorname{pa}_{\mathcal{G}}(\mathbf{B})) \right) \left( \prod_{\{B\} \notin \mathcal{B}^{nt}(\mathcal{G})} p(B | \operatorname{pa}_{\mathcal{G}}(B)) \right) \\
&= \left( \prod_{\mathbf{B} \in \mathcal{B}^{nt}(\mathcal{G})} p(\mathbf{B} | \operatorname{pa}_{\mathcal{G}}(\mathbf{B})) \right) \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}(\mathbf{V}))} \sum_{\mathbf{H}_{\mathbf{D}}} \left( \prod_{B \in \mathbf{D}} p(B | \operatorname{pa}_{\mathcal{G}}(B)) \right) \\
&= \left( \prod_{\mathbf{B} \in \mathcal{B}^{nt}(\mathcal{G})} p(\mathbf{B} | \operatorname{pa}_{\mathcal{G}}(\mathbf{B})) \right) \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}(\mathbf{V}))} q(\mathbf{D} | \operatorname{pa}^{s}_{\mathcal{G}(\mathbf{V})}(\mathbf{D})) \\
&= q(\mathbf{B}^* | \operatorname{pa}_{\mathcal{G}(\mathbf{V})}(\mathbf{B}^*)) q(\mathbf{D}^* | \operatorname{pa}^{s}_{\mathcal{G}(\mathbf{V})}(\mathbf{D}^*)).
\end{aligned}$$

The fact that $q(\mathbf{B}^* | \operatorname{pa}_{\mathcal{G}(\mathbf{V})}(\mathbf{B}^*))$ factorizes according to the CCG $\mathcal{G}^b$ follows by construction.

Let $\widetilde{\mathbf{B}} \equiv \{B \in \mathbf{V} \cup \mathbf{H} \mid \{B\} \notin \mathcal{B}^{nt}(\mathcal{G})\}$. Then

$$q(\widetilde{\mathbf{B}} | \operatorname{pa}^{s}_{\mathcal{G}}(\widetilde{\mathbf{B}})) = \prod_{B : \{B\} \notin \mathcal{B}^{nt}(\mathcal{G})} p(B | \operatorname{pa}_{\mathcal{G}}(B))$$

factorizes according to the CADMG (in fact a conditional DAG) $\mathcal{G}(\widetilde{\mathbf{B}}, \mathrm{pa}^s_{\mathcal{G}}(\widetilde{\mathbf{B}}))$ obtained from $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$ by making all elements in $\mathrm{pa}^s_{\mathcal{G}}(\widetilde{\mathbf{B}})$ fixed, and all elements $\widetilde{\mathbf{B}}$ random, keeping all edges among $\widetilde{\mathbf{B}}$ in $\mathcal{G}$, and all outgoing directed edges from $\mathrm{pa}^s_{\mathcal{G}}(\widetilde{\mathbf{B}})$ to $\widetilde{\mathbf{B}}$ in $\mathcal{G}$. The fact that $q(\mathbf{D}^*| \mathrm{pa}_{\mathcal{G}(\mathbf{V})}(\mathbf{D}^*))$ factorizes according $\mathcal{G}^d$, the latent projection CADMG obtained from $\mathcal{G}(\widetilde{\mathbf{B}}, \mathrm{pa}^s_{\mathcal{G}}(\widetilde{\mathbf{B}}))$ by treating $\mathbf{H}$ as hidden variables now follows by the inductive application of Lemmas 46 and 49 in Richardson et al. (2017) to $q(\widetilde{\mathbf{B}}| \mathrm{pa}^s_{\mathcal{G}}(\widetilde{\mathbf{B}}))$ and $\mathcal{G}(\widetilde{\mathbf{B}}, \mathrm{pa}^s_{\mathcal{G}}(\widetilde{\mathbf{B}}))$. □

**Theorem 9** *Assume $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$ is a causal CG, where $\mathbf{H}$ is block-safe. Fix disjoint subsets $\mathbf{Y}, \mathbf{A}$ of $\mathbf{V}$. Let $\mathbf{Y}^* = \mathrm{ant}_{\mathcal{G}(\mathbf{V})_{\mathbf{V}\setminus\mathbf{A}}} \mathbf{Y}$. Then $p(\mathbf{Y}|do(\mathbf{a}))$ is identified from $p(\mathbf{V})$ if and only if every element in $\mathcal{D}(\widetilde{\mathcal{G}}^d)$ is reachable in $\mathcal{G}^d$, where $\widetilde{\mathcal{G}}^d$ is the induced CADMG of $\mathcal{G}(\mathbf{V})_{\mathbf{Y}^*}$.*

*Moreover, if $p(\mathbf{Y}|do(\mathbf{a}))$ is identified, it is equal to*

$$\sum_{\mathbf{Y}^*\setminus\mathbf{Y}} \left[ \prod_{\mathbf{D}\in\mathcal{D}(\widetilde{\mathcal{G}}^d)} \phi_{\mathbf{D}^*\setminus\mathbf{D}}(q(\mathbf{D}^*| \mathrm{pa}_{\mathcal{G}(\mathbf{V})}(\mathbf{D}^*)); \mathcal{G}^d) \right] \tag{D.1}$$

$$\times \left[ \prod_{\mathbf{B}\in\mathcal{B}(\widetilde{\mathcal{G}}^b)} p(\mathbf{B} \setminus \mathbf{A}| \mathrm{pa}_{\mathcal{G}(\mathbf{V})_{\mathbf{Y}^*}}(\mathbf{B}), \mathbf{B} \cap \mathbf{A}) \right] \Bigg|_{\mathbf{A}=\mathbf{a}} \tag{D.2}$$

*where*

$$q(\mathbf{D}^*| \mathrm{pa}_{\mathcal{G}(\mathbf{V})}(\mathbf{D}^*)) = \frac{p(\mathbf{V})}{(\prod_{\mathbf{B}\in\mathcal{B}^{nt}(\mathcal{G}(\mathbf{V}))} p(\mathbf{B}| \mathrm{pa}_{\mathcal{G}(\mathbf{V})}(\mathbf{B}))},$$

*and $\widetilde{\mathcal{G}}^d$ is the induced CCG of $\mathcal{G}(\mathbf{V})_{\mathbf{Y}^*}$.*

*Proof:* We proceed by proving a series of subclaims.

**Claim 1**: *If $p(\mathbf{O})$ obeys the segregated factorization relative to $\mathcal{G}(\mathbf{O})$, then $p(\mathbf{A})$ obeys the segregated factorization relative to $\mathcal{G}(\mathbf{O})_{\mathbf{A}}$* for any subset $\mathbf{A} \subseteq \mathbf{O}$ anterial in $\mathcal{G}(\mathbf{O})$. A set $\mathbf{A}$ is anterial if, whenever $X \in \mathbf{A}$, $\mathrm{ant}_{\mathcal{G}}(X) \subseteq \mathbf{A}$.

We show this by induction. Assume $p(\mathbf{O})$ obeys the segregated factorization relative to $\mathcal{G}(\mathbf{O})$, and $\mathbf{A}$ consists of all elements in $\mathbf{O}$ other than those in $\mathbf{B} \in \mathcal{B}^{nt}(\mathcal{G}(\mathbf{O}))$. Then by writing $p(\mathbf{A}) = \sum_{\mathbf{B}} p(\mathbf{O})$ as a segregated factorization for $p(\mathbf{O})$, we note that the nested factorization remains unchanged by the marginalization, and the block factorization remains unchanged, except the factor corresponding to $\mathbf{B}$ is removed.

Similarly, assume $p(\mathbf{O})$ obeys the segregated factorization relative to $\mathcal{G}(\mathbf{O})$, and $\mathbf{A}$ consists of all elements in $\mathbf{O}$ other than some element $B$ not in any $\mathbf{B} \in \mathcal{B}^{nt}(\mathcal{G}(\mathbf{O}))$ such that $\text{ch}_{\mathcal{G}}(B)$ is empty. Then by writing $p(\mathbf{A}) = \sum_{\mathbf{B}} p(\mathbf{O})$ as a segregated factorization for $p(\mathbf{O})$, we note that the block factorization remains unchanged by the marginalization, and the kernel

$$q(\mathbf{B}^* \setminus \{B\} \mid \text{pa}^s_{\mathcal{G}(\mathbf{O})}(\mathbf{B}^*)) = \sum_B \frac{p(\mathbf{V})}{\prod_{\mathbf{B} \in \mathcal{B}^{nt}(\mathcal{G}(\mathbf{V}))} p(\mathbf{B} | \text{pa}_{\mathcal{G}(\mathbf{V})}(\mathbf{B}))}$$

is nested Markov relative to the CADMG $\tilde{\mathcal{G}}(\mathbf{O})^\ulcorner$ obtained from $\mathcal{G}(\mathbf{O})^d$ by removing $B$ and all edges adjacent to $B$. To see this, note that reachable sets in $\tilde{\mathcal{G}}(\mathbf{O})^\ulcorner$ are a strict subset of reachable sets in $\mathcal{G}(\mathbf{O})^d$, since $B$ is fixable in $\mathcal{G}(\mathbf{O})^d$, and moreover all kernels corresponding to reachable sets in $\tilde{\mathcal{G}}(\mathbf{O})^\ulcorner$ may be obtained from $q(\mathbf{B}^* \mid \text{pa}^s_{\mathcal{G}(\mathbf{O})}(\mathbf{B}^*))$ by marginalizing $B$ first, and applying the fixing operator to remaining variables in $\mathcal{B}^* \setminus \{B\}$. As a result, the nested global Markov property for the former graph is implied by the nested global Markov property of the latter graph, proving our claim.

**Claim 2**: *The algorithm specified by the equation (D.2) is sound for identification of $p(\mathbf{Y} | do(\mathbf{a}))$.*

Per claim 1, without loss of generality assume $\mathbf{Y}$ has no children in $\mathcal{G}(\mathbf{O})$. Consider the chain graph g-formula:

$$p(\mathbf{Y}(\mathbf{a})) = \prod_{\mathbf{B} \in \mathcal{B}(\mathcal{G}(\mathbf{O} \cup \mathbf{H}))} p(\mathbf{B} \setminus \mathbf{A} | \text{pa}_{\mathcal{G}}(\mathbf{B}), \mathbf{B} \cap \mathbf{A})|_{\mathbf{A}=\mathbf{a}}.$$

We can decompose this into factors relating to the non-trivial blocks and districts in the graph:

$$\begin{aligned} p(\mathbf{Y}(\mathbf{a})) = &\prod_{\mathbf{B} \in \mathcal{B}^{nt}(\mathcal{G}(\mathbf{O} \cup \mathbf{H}))} p(\mathbf{B} \setminus \mathbf{A} | \text{pa}_{\mathcal{G}}(\mathbf{B}), \mathbf{B} \cap \mathbf{A})|_{\mathbf{A}=\mathbf{a}} \\ &\times \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}(\mathbf{O} \cup \mathbf{H}))} p(\mathbf{D} \setminus \mathbf{A} | \text{pa}_{\mathcal{G}}(\mathbf{D}), \mathbf{D} \cap \mathbf{A})|_{\mathbf{A}=\mathbf{a}}. \end{aligned}$$

Since $\mathbf{H}$ is block-safe, the factors in the first term – those that correspond to non-trivial blocks – are the same in the segregated graph as in the original chain graph and thus we can re-write the above as:

$$\begin{aligned} p(\mathbf{Y}(\mathbf{a})) = &\prod_{\mathbf{B} \in \mathcal{B}^{nt}(\mathcal{G}_{\mathbf{Y}^*})} p(\mathbf{B} \setminus \mathbf{A} | \operatorname{pa}_{\mathcal{G}}(\mathbf{B}), \mathbf{B} \cap \mathbf{A})|_{\mathbf{A}=\mathbf{a}} \\ &\times \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}(\mathbf{O} \cup \mathbf{H}))} p(\mathbf{D} \setminus \mathbf{A} | \operatorname{pa}_{\mathcal{G}}(\mathbf{D}), \mathbf{D} \cap \mathbf{A})|_{\mathbf{A}=\mathbf{a}}. \end{aligned}$$

Meanwhile the factors in the second term describe a kernel $q(\mathbf{D}^* | \operatorname{pa}_{\mathcal{G}_{(\mathbf{O} \cup \mathbf{H})}}(\mathbf{D}^*))$ associated with a CADG $\mathcal{G}(\mathbf{O} \cup \mathbf{H}, \mathbf{B}^*)$ which we can manipulate to obtain the desired result by following the argument in the proof of Theorem 60 in Richardson et al. (2017).

Let $\mathbf{A}^* = \mathbf{O} \setminus \mathbf{Y}^* \supseteq \mathbf{A}$. By the global Markov property of conditional DAGs (CDAGs) proven in Richardson et al. (2017), $p(\mathbf{Y}^* | do_{\mathcal{G}(\mathbf{O} \cup \mathbf{H}, \mathbf{B}^*)}(\mathbf{a})) = p(\mathbf{Y}^* | do_{\mathcal{G}(\mathbf{O} \cup \mathbf{H}, \mathbf{B}^*)}(\mathbf{a}^*))$.

Let $\mathcal{G}^*((\mathbf{O} \setminus \mathbf{A}^*) \cup \mathbf{H}, \mathbf{B}^* \cup \mathbf{A}^*) = \phi_{\mathbf{A}^*}(\mathcal{G}(\mathbf{O} \cup \mathbf{H}, \mathbf{B}^*))$. Let $\sigma_{\mathbf{H}}$ denote the *latent projection operation* such that $\sigma_{\mathbf{H}}(\mathcal{G}(\mathbf{O} \cup \mathbf{H}) = \mathcal{G}(\mathbf{O})$. Then, by commutativity of $\sigma_{\mathbf{H}}$ and the fixing operator (Corollary 53 in Richardson et al. (2017)), $\sigma_{\mathbf{H}}(\phi_{\mathbf{A}^*}(\mathcal{G}(\mathbf{O} \cup \mathbf{H}, \mathbf{B}^*))) = \phi_{\mathbf{A}^*}(\sigma_{\mathbf{H}}(\mathcal{G}(\mathbf{O} \cup \mathbf{H}, \mathbf{B}^*))) = \mathcal{G}^*(\mathbf{Y}^*, \mathbf{B}^* \cup \mathbf{A}^*)$. By definition of induced subgraphs, $\mathcal{G}(\mathbf{O}, \mathbf{B}^*)_{\mathbf{Y}^*} = (\phi_{\mathbf{A}^*}(\mathcal{G}(\mathbf{O}, \mathbf{B}^*)))_{\mathbf{Y}^*}$. By these two equalities, we have $\mathcal{G}(\mathbf{O}, \mathbf{B}^*)_{\mathbf{Y}^*} = \mathcal{G}^*(\mathbf{O}, \mathbf{B}^* \cup \mathbf{A}^*)_{\mathbf{Y}^*}$ and thus $\mathcal{D}(\mathcal{G}(\mathbf{O}, \mathbf{B}^*)_{\mathbf{Y}^*}) = \mathcal{D}(\mathcal{G}^*(\mathbf{Y}^*, \mathbf{B}^* \cup \mathbf{A}^*))$.

For each $\mathbf{D} \in \mathcal{D}(\mathcal{G}^*(\mathbf{Y}^*, \mathbf{B}^* \cup \mathbf{A}^*))$, let $\mathbf{H}_{\mathbf{D}} \equiv \mathbf{H} \cap \operatorname{an}_{\mathcal{G}(\mathbf{O} \cup \mathbf{H}, \mathbf{B}^*)_{\mathbf{D} \cup \mathbf{H}}}(\mathbf{D})$ and $\mathbf{H}^* \equiv \bigcup_{\mathbf{D} \in \mathcal{D}(\mathcal{G}^*(\mathbf{Y}^*, \mathbf{B}^* \cup \mathbf{A}^*))} \mathbf{H}_{\mathbf{D}}$. Then, by construction, if $\mathbf{D}, \mathbf{D}' \in \mathcal{D}(\mathcal{G}^*(\mathbf{Y}^*, \mathbf{B}^* \cup \mathbf{A}^*)$ and $\mathbf{D} \neq \mathbf{D}'$ then $\mathbf{H}_{\mathbf{D}} \cap \mathbf{H}_{\mathbf{D}'} = \emptyset$. Additionally, for all $\mathbf{D} \in \mathcal{D}(\mathcal{G}^*(\mathbf{Y}^*, \mathbf{B}^* \cup \mathbf{A}^*)$, it is the case that $\operatorname{pa}_{\mathcal{G}(\mathbf{O} \cup \mathbf{H}, \mathbf{B}^*)}(\mathbf{D} \cup \mathbf{H}_{\mathbf{D}}) \cap \mathbf{H}^* = \mathbf{H}_{\mathbf{D}}$. And $\mathbf{Y}^* \cup \mathbf{H}^*$ is ancestral in $\mathcal{G}(\mathbf{O} \cup \mathbf{H}, \mathbf{B}^*)$ which implies that if $v \in \mathbf{Y}^* \cup \mathbf{H}^*$, then $\operatorname{pa}_{\mathcal{G}(\mathbf{O} \cup \mathbf{H}, \mathbf{B}^*}(v) \cap \mathbf{H} \subseteq \mathbf{H}^*$.

By the DAG g-formula and the above features of the construction,

$$
\begin{aligned}
&p(\mathbf{Y}^*| do_{\mathcal{G}(\mathbf{O}\cup\mathbf{H},\mathbf{B}^*)}(\mathbf{a}^*)) \\
&\quad= \sum_{\mathbf{H}} \prod_{v\in(\mathbf{H}\cup\mathbf{Y}^*)} p(v|\operatorname{pa}_{\mathcal{G}(\mathbf{O}\cup\mathbf{H},\mathbf{B}^*)}(v)) \\
&\quad= \sum_{\mathbf{H}^*} \prod_{v\in(\mathbf{H}^*\cup\mathbf{Y}^*)} p(v|\operatorname{pa}_{\mathcal{G}(\mathbf{O}\cup\mathbf{H},\mathbf{B}^*)}(v)) \cdot \sum_{\mathbf{H}\setminus\mathbf{H}^*} \prod_{v\in(\mathbf{H}\setminus\mathbf{H}^*)} p(v|\operatorname{pa}_{\mathcal{G}(\mathbf{O}\cup\mathbf{H},\mathbf{B}^*)}(v)) \\
&\quad= \sum_{\mathbf{H}^*} \prod_{\mathbf{D}\in\mathcal{D}(\mathcal{G}^*(\mathbf{Y}^*,\mathbf{A}^*\cup\mathbf{B}^*))} \prod_{v\in(\mathbf{D}\cup\mathbf{H}_\mathbf{D})} p(v|\operatorname{pa}_{\mathcal{G}(\mathbf{O}\cup\mathbf{H},\mathbf{B}^*)}(v)) \\
&\quad= \prod_{\mathbf{D}\in\mathcal{D}(\mathcal{G}^*(\mathbf{Y}^*,\mathbf{A}^*\cup\mathbf{B}^*))} \left( \sum_{\mathbf{H}_\mathbf{D}} \prod_{v\in(\mathbf{D}\cup\mathbf{H}_\mathbf{D})} p(v|\operatorname{pa}_{\mathcal{G}(\mathbf{O}\cup\mathbf{H},\mathbf{B}^*)}(v)) \right).
\end{aligned}
\tag{D.3}
$$

For any district $\mathbf{D} \in \mathcal{D}(\mathcal{G}^*(\mathbf{Y}^*, \mathbf{B}^* \cup \mathbf{A}^*))$,

$$
\begin{aligned}
&\sum_{\mathbf{H}_\mathbf{D}} \prod_{v\in\mathbf{D}\cup\mathbf{H}_\mathbf{D}} p(v|\operatorname{pa}_{\mathcal{G}(\mathbf{O}\cup\mathbf{H},\mathbf{B}^*)}(v)) \\
&\quad= \sum_{\mathbf{H}_\mathbf{D}} \prod_{v\in(\mathbf{D}\cup\mathbf{H}_\mathbf{D})} p(v|\operatorname{pa}_{\mathcal{G}(\mathbf{O}\cup\mathbf{H},\mathbf{B}^*)}(v)) \cdot \sum_{\mathbf{H}\setminus\mathbf{H}_\mathbf{D}} \prod_{v\in(\mathbf{H}\setminus\mathbf{H}_\mathbf{D})} p(v|\operatorname{pa}_{\mathcal{G}(\mathbf{O}\cup\mathbf{H},\mathbf{B}^*)}(v)) \\
&\quad= \sum_{\mathbf{H}} \prod_{v\in\mathbf{D}\cup\mathbf{H}_\mathbf{D}} p(v|\operatorname{pa}_{\mathcal{G}(\mathbf{O}\cup\mathbf{H},\mathbf{B}^*)}(v)) \\
&\quad= \sum_{\mathbf{H}} \phi_{\mathbf{D}^*\setminus\mathbf{D}}(q(\mathbf{D}^*|\operatorname{pa}_{\mathcal{G}(\mathbf{O}\cup\mathbf{H},\mathbf{B}^*)}(\mathbf{D}^*))); \mathcal{G}(\mathbf{O}\cup\mathbf{H},\mathbf{B}^*))
\end{aligned}
\tag{D.4}
$$

Once again, these equalities are a result of the above constructions of $\mathbf{H}$ and $\mathbf{H}^*$. By commutativity (Lemma 55 in Richardson et al. (2017)), we can remove references to $\mathbf{H}$:

$$
\begin{aligned}
&p(\mathbf{Y}^*| do_{\mathcal{G}(\mathbf{O}\cup\mathbf{H},\mathbf{B}^*)}(\mathbf{A}^*)) \\
&\quad= \prod_{\mathbf{D}\in\mathcal{D}(\mathcal{G}(\mathbf{Y}^*,\mathbf{B}^*\cup\mathbf{A}^*))} \phi_{\mathbf{D}^*\setminus\mathbf{D}} q(\mathbf{D}^*|\operatorname{pa}_{\mathcal{G}(\mathbf{O},\mathbf{B}^*)}(\mathbf{D}^*)); \mathcal{G}(\mathbf{O},\mathbf{B}^*)) \\
&\quad= \prod_{\mathbf{D}\in\mathcal{D}(\mathcal{G}(\mathbf{Y}^*,\mathbf{B}^*\cup\mathbf{A}^*))} \phi_{\mathbf{D}^*\setminus\mathbf{D}} q(\mathbf{D}^*|\operatorname{pa}_{\mathcal{G}}(\mathbf{D}^*)); \mathcal{G}^d) \\
&\quad= \prod_{\mathbf{D}\in\mathcal{D}(\mathcal{G}_{\mathbf{Y}^*})} \phi_{\mathbf{D}^*\setminus\mathbf{D}} q(\mathbf{D}^*|\operatorname{pa}_{\mathcal{G}}(\mathbf{D}^*)); \mathcal{G}^d)
\end{aligned}
$$

The second equality is true because $\operatorname{pa}_{\mathcal{G}}(\mathbf{D}^*) \subseteq \operatorname{pa}_{\mathcal{G}(\mathbf{O},\mathbf{B}^*)}(\mathbf{D}^*)$ and by the assumption of a block-safe chain graph. The final equality is true by block-safeness and the definition of induced subgraphs.

Finally by the fact that $p(\mathbf{Y}| do_{\mathcal{G}(\mathbf{O}\cup\mathbf{H},\mathbf{B}^*)}(\mathbf{A})) = \sum_{\mathbf{Y}^*\setminus\mathbf{Y}} p(\mathbf{Y}^*| do_{\mathcal{G}(\mathbf{O}\cup\mathbf{H},\mathbf{B}^*)}(\mathbf{A}^*))$, we can re-write the above as:

$$p(\mathbf{Y}|do_{\mathcal{G}(\mathbf{O}\cup\mathbf{H},\mathbf{B}^*)}(\mathbf{A})) = \sum_{\mathbf{Y}^*\setminus\mathbf{Y}} \prod_{\mathbf{D}\in\mathcal{D}(\mathcal{G}_{\mathbf{Y}^*})} \phi_{\mathbf{D}^*\setminus\mathbf{D}} q(\mathbf{D}^*|\operatorname{pa}_{\mathcal{G}}(\mathbf{D}^*)); \mathcal{G}^d)$$

We combine this with the block portioned derived above via chain-graph g-formula to obtain the result of the sub-claim

**Claim 3**: *If there is a district in $\mathcal{D}(\mathcal{G}(\mathbf{O})_{\mathbf{Y}^*})$ that is not reachable in $\mathcal{G}^d$, then $p(\mathbf{Y}|do(\mathbf{a}))$ is not identifiable.*

Let $\mathbf{D} \in \mathcal{D}(\mathcal{G}(\mathbf{O})_{\mathbf{Y}^*})$ be unreachable. Let $\mathbf{R} = \{D \in \mathbf{D}| \operatorname{ch}_{\mathcal{G}}(D) \cap \mathbf{D} = \emptyset\}$. Let $\mathbf{A}^* = A \cap \operatorname{pa}_{\mathcal{G}}(D)$. Then there exists a superset of $\mathbf{D}$, $\mathbf{D}'$, such that $\mathbf{D}$ and $\mathbf{D}'$ form a hedge for $p(\mathbf{R}|do(\mathbf{a}^*))$ and thus $p(\mathbf{R}|do(\mathbf{a}^*))$ is not identified (Shpitser and Pearl, 2006).

Let $\mathbf{Y}'$ be the minimal subset of $\mathbf{Y}$ such that $\mathbf{R} \subseteq \operatorname{ant}_{\mathcal{G}(\mathbf{O})_{\mathbf{O}\setminus\mathbf{A}}}(\mathbf{Y}')$. Consider an edge subgraph $\mathcal{G}^\dagger$ of $\mathcal{G}$ consisting of all edges in $\mathcal{G}$ in the hedge formed by $\mathbf{D}, \mathbf{D}'$ and edges on partially directed paths in $\mathcal{G}(\mathbf{O})_{\mathbf{O}\setminus\mathbf{A}}$ from every element in $\mathbf{R}$ to some element in $\mathbf{Y}'$, such that the edge subgraph does not contain any cycles (directed or otherwise).

We proceed as follows. We first define an ADMG $\tilde{\mathcal{G}}^\dagger$ from $\mathcal{G}^\dagger$ as follows. The vertices and edges making up the hedge structure (Shpitser and Pearl, 2006) in $\mathcal{G}^\dagger$ are also present in $\tilde{\mathcal{G}}^\dagger$. For every partially directed path $\sigma$ from an element in $\mathbf{R}$ to an element in $\mathbf{Y}'$, we construct a directed path from $\mathbf{R}$ in $\tilde{\mathcal{G}}^\dagger$ containing vertex copies of vertices on the undirected path $\sigma$, and which orients all undirected edges in $\sigma$ away from $\mathbf{R}$ and towards the element copy in $\tilde{\mathcal{G}}^\dagger$ of the appropriate element of $\mathbf{Y}'$ in $\mathcal{G}^\dagger$.

We then prove non-identifiability of $p(\tilde{\mathbf{Y}}'|\text{do}(\mathbf{a}^*))$ in $\tilde{\mathcal{G}}^\dagger$, where $\tilde{\mathbf{Y}}'$ is the set of all vertex copies in $\tilde{\mathcal{G}}^\dagger$ of vertices in $\mathbf{Y}'$ in $\mathcal{G}^\dagger$, using standard techniques for ADMGs. In particular, we follow the proof of Theorem 4 in the supplement of Shpitser and Sherman (2018).

We next show that $p(\mathbf{Y}' \mid \text{do}(\mathbf{a}^*))$ is not identified in $\mathcal{G}^\dagger$. For the two counterexamples in the causal model given by $\tilde{\mathcal{G}}^\dagger$ witnessing non-identifiability of $p(\tilde{\mathbf{Y}}' \mid \text{do}(\mathbf{a}^*))$

in the above proof, we will construct two counterexamples in the causal model given by $\mathcal{G}^\dagger$ witnessing non-identifiability of $p(\mathbf{Y}' \mid \text{do}(\mathbf{a}^*))$.

To do so, we define new variables along all partially directed paths from $\mathbf{R}$ to $\mathbf{Y}'$ in $\mathcal{G}^\dagger$ as Cartesian products of variable copies in counterexamples constructed. Note that any such variable containing only a single element in $\mathbf{R}$ in its anterior in $\mathcal{G}^\dagger$ will only have a single copy, while a variable containing two elements in $\mathbf{R}$ in its anterior in $\mathcal{G}^\dagger$ will contain two copies, and so on. It's clear that the two resulting elements contain vertices in $\mathcal{G}^\dagger$, agree on the observed data distribution, and disagree on $p(\mathbf{Y}' \mid \text{do}(\mathbf{a}^*))$.

What remains to show is that the distributions so constructed obey one of CG Markov properties associated with a CG $\mathcal{G}^\dagger$. Fix a (possibly trivial) block $\mathbf{B}$ in $\mathcal{G}^\dagger$. We must show for each $B \in \mathbf{B}$ that $p(B \mid \mathbf{B} \setminus B, \text{pa}_{\mathcal{G}^\dagger}(\mathbf{B})) = p(B \mid \text{nb}_{\mathcal{G}^\dagger}, \text{pa}_{\mathcal{G}}(B))$.

For any $B \in \mathbf{B}$ in $\mathcal{G}^\dagger$, there exists a set $B_1, \ldots, B_k$ of variables in $\tilde{\mathcal{G}}^\dagger$ such that $B$ is defined as $B_1 \times \ldots \times B_k$. Moreover, any variable $A \in \text{nb}_{\mathcal{G}^\dagger}(B) \cup \text{pa}_{\mathcal{G}^\dagger}(B)$ corresponds to a Cartesian product $A_1 \times A_m$ of variables where $A_i$ is a child or a parent of some variables $B_j$. The result then follows by d-separation in $\tilde{\mathcal{G}}^\dagger$, and the fact that the part of $\tilde{\mathcal{G}}^\dagger$ outside of the hedge structure does not contain any colliders by construction. □

## D.2 Derivations

Consider Figure 4-1 (c). We are interested in identifying $p(Y_2(a_1, a_2))$. We set $\mathbf{Y}^*$ to the anterior of $\mathbf{Y}$ in $\mathcal{G}_{\mathbf{V}\setminus\mathbf{A}}$: $\mathbf{Y}^* \equiv \{C_1, C_2, M_1, M_2, Y_2\}$ (see $\mathcal{G}_{\mathbf{Y}^*}$ shown in Fig. 4-1 (d)) with $\mathcal{B}(\mathcal{G}_{\mathbf{Y}^*}) = \{\{M_1, M_2\}\}$ and $\mathcal{D}(\mathcal{G}_{\mathbf{Y}^*} = \{\{C_1\}, \{C_2\}, \{Y_2\}\}$. We can now proceed with the version of the ID algorithm for SGs. The CCG portion of the algorithm simply yields $p(M_1, M_2 | A_1 = a_1, A_2, C_1, C_2)$. Note that this expression further factorizes according to the factorization of blocks in a chain graph. For the ADMG portion of the algorithm, we must fix variables in three different sets $\{C_2, A_1, A_2, Y_1, Y_2\}$, $\{C_1, A_1, A_2, Y_1, Y_2\}$, $\{C_1, C_2, A_1, A_2, Y_1\}$ in $\mathcal{G}^d$, shown in Fig. 4-1 (e), corresponding to three districts in Fig. 4-1 (d). We have:

$$\begin{aligned}
&\phi_{\{C_2,A_1,A_2,Y_1,Y_2\}}(p(Y_1,Y_2|A_1,A_2,M_1,M_2,C_1,C_2)p(A_1,A_2,C_1,C_2))\\
&= \phi_{\{C_2,A_1,A_2,Y_1\}}(p(Y_1|A_1,A_2,M_1,M_2,C_1,C_2,Y_2)p(A_1,A_2,C_1,C_2))\\
&= \phi_{\{C_2,A_1,A_2\}}(p(A_1,A_2,C_1,C_2))\\
&= \phi_{\{C_2,A_2\}}(p(A_2,C_1,C_2))\\
&= \phi_{\{C_2\}}(p(C_1,C_2))\\
&= p(C_1)
\end{aligned} \tag{D.5}$$

$$\begin{aligned}
&\phi_{\{C_1,A_1,A_2,Y_1,Y_2\}}(p(Y_1,Y_2|A_1,A_2,M_1,M_2,C_1,C_2)p(A_1,A_2,C_1,C_2))\\
&= \phi_{\{C_1,A_1,A_2,Y_1\}}(p(Y_1|A_1,A_2,M_1,M_2,C_1,C_1,Y_2)p(A_1,A_2,C_1,C_2))\\
&= \phi_{\{C_1,A_1,A_2\}}(p(A_1,A_2,C_1,C_2))\\
&= \phi_{\{C_1,A_2\}}(p(A_2,C_1,C_2))\\
&= \phi_{\{C_1\}}(p(C_1,C_2))\\
&= p(C_2)
\end{aligned} \tag{D.6}$$

$$\begin{aligned}
&\phi_{\{C_1,C_2,A_1,A_2,Y_1\}}(p(Y_1,Y_2|A_1,A_2,M_1,M_2,C_1,C_2)p(A_1,A_2,C_1,C_2))\\
&= \phi_{\{A_1,Y_1,A_2\}}(p(Y_1,Y_2|A_1,A_2,M_1,M_2,C_1,C_2)p(A_1,A_2|C_1,C_2))\\
&= \phi_{\{A_1,A_2\}}(p(Y_2|A_1,A_2,M_1,M_2,C_1,C_2)p(A_1,A_2|C_1,C_2))\\
&= \sum_{A_2} p(Y_2|A_1,A_2,M_1,M_2,C_1,C_2)p(A_2|C_2)\\
&= \sum_{A_2} p(Y_2|A_1,A_2,M_2,C_2)p(A_2|C_2)
\end{aligned} \tag{D.7}$$

with the last term evaluated at $A_1 = a_1$. Thus, the identifying functional is:

$$\begin{aligned}
p(Y_2(a_1,a_2)) = \sum_{\{C_1,C_2,M_1,M_2\}} \Bigg[ & p(M_1,M_2|a_1,a_2,C_1,C_2)\\
& \times \Bigg[ \sum_{A_2} p(Y_2|a_1,A_2,M_2,C_2)p(A_2|C_2)p(C_1)p(C_2) \Bigg] \Bigg]
\end{aligned} \tag{D.8}$$

# D.3 Simulation Study

## D.3.1 The Auto-G-Computation Algorithm

To estimate identifying functionals corresponding to causal effects given dependent data, we generally use maximum likelihood plug in estimation. The exception is the factor $p(\mathbf{M} \mid \text{pa}_{\mathcal{G}}(\mathbf{M}))$, which may not be estimated if $M_i$ variables for all units $i$ are dependent, as is the case in our simulation study. In this case, the above density must be estimated from a single sample. Thus, standard statistical methods such as maximum likelihood estimation fail to work. We adapt the auto-g-computation algorithm method in Tchetgen, Fulcher, and Shpitser (2017), which exploits Markov assumptions embedded in our CG model, as well as the pseudo-likelihood or coding estimation methods introduced in Besag (1975). We briefly describe the approach here.

The auto-g-computation algorithm is a generalization of the Monte Carlo sampling version of the standard g-computation algorithm for classical causal models (represented by DAGs) (Westreich et al., 2012) to causal models represented by CGs. Auto-g-computation proceeds by generating samples from a block using Gibbs sampling. The parameters for Gibbs factors used in the sampler (which, by the global Markov property for CGs, take the form of $p(X_i \mid \text{pa}_{\mathcal{G}}(X_i) \cup \text{nb}_{\mathcal{G}}(X_i)))$ are learned via parameter sharing and coding or pseudo-likelihood based estimators. For any block $\mathbf{B}$, the Gibbs sampler draws samples from $p(\mathbf{X} \mid \text{pa}_{\mathcal{G}}(\mathbf{X}))$, given a fixed set of samples drawn from all blocks with elements in $\text{pa}_{\mathcal{G}}(\mathbf{X})$, or specific values of $\text{pa}_{\mathcal{G}}(\mathbf{X})$ we are interested in, as follows.

Gibbs Sampler for $\mathbf{X}$:

```
for $t = 0$, let $\mathbf{x}^{(0)}$ denote initial values ;
for $t = 1, ..., T$
    draw value of $X_1^{(t)}$ from $p(X_1|\mathbf{x}^{(t-1)}_{\text{pa}_{\mathcal{G}}(X_1)\cup\text{nb}_{\mathcal{G}}(X_1)}))$;
    draw value of $X_2^{(t)}$ from $p(X_2|\mathbf{x}^{(t-1)}_{\text{pa}_{\mathcal{G}}(X_2)\cup\text{nb}_{\mathcal{G}}(X_2)}))$;
    $\vdots$
    draw value of $X_m^{(t)}$ from $p(X_m|\mathbf{x}^{(t-1)}_{\text{pa}_{\mathcal{G}}(X_m)\cup\text{nb}_{\mathcal{G}}(X_m)}))$;
```

Since we are interested in estimating a functional similar to (D.8), we use observed values of $\mathbf{C}$, and intervened on values $a_i, a_j$ as the values of $\text{pa}_{\mathcal{G}}(\mathbf{M})$ in the Gibbs sampler.

The coding-likelihood and pseudo-likelihood estimators we use are described in more detail in Tchetgen, Fulcher, and Shpitser (2017). Both estimators rely on parameter sharing for densities $p(M_i \mid \text{pa}_{\mathcal{G}}(M_i) \cup \text{nb}_{\mathcal{G}}(M_i))$ across different units $i$, and for the network to be sufficiently sparse such that each $M_i$ depends on only a few other variables in the model, relative to the total number of units.

The coding estimator uses a subset of the data that corresponds to units that form independent sets in the network adjacency graph (where units are adjacent of they are friends in the network, and not adjacent otherwise). A set of units is a *maximal* independent set in the network adjacency graph if a) no two vertices in the set are adjacent, and b) it is impossible to add another unit to the set without violating the adjacency constraint. A *maximum* independent set is a maximal independent set such that there does not exist a larger maximal independent set in the same graph. Finding maximum independent sets is a classic NP-complete problem; in practice we find several *maximal* independent sets and pick the one with largest cardinality as a heuristic. See Table D-I below for the size of $S_{max}$ for each network size in our experiments. The coding likelihood estimator was proven consistent and asymptotically

| $N$ | 400 | 800 | 1000 | 2000 |
|---|---|---|---|---|
| $\lvert S_{max}\rvert$ | 159 | 309 | 384 | 763 |

Table D-I. The size of $S_{max}$ used for the coding-likelihood estimator in each network

normal in Tchetgen, Fulcher, and Shpitser (2017) whereas pseudo-likelihood estimation is, under mild assumptions, consistent but not asymptotically normal. On the other hand, pseudo-likelihood estimation is more efficient than coding likelihood estimation since it makes use of all of the data.

### D.3.2 Simulation Specifics

For data generation we use the following densities for $A_i, M_i, Y_i$, parameterized by $\tau_A = \{\gamma_0, \gamma_{C_1}, \ldots, \gamma_{C_p}, \gamma_{U_1}, \ldots, \gamma_{U_q}\}, \tau_M = \{\beta_0, \beta_A, \beta_{C_1}, \ldots, \beta_{C_p}\beta_{A_{nb}}, \beta_{M_{nb}}\}, \tau_Y = \{\alpha_0, \alpha_{C_1}, \ldots, \alpha_{C_p}, \alpha_{U_1}, \ldots, \alpha_{U_q}, \alpha_{A_{nb}}, \alpha_M\}$:

$$\begin{aligned}
&p(A_i = 1|\mathbf{C}_i, \mathbf{U}_i; \tau_A) = expit(\gamma_0 + \Big(\sum_{l=1}^{p} \gamma_{C_l} C_{il}\Big) + \Big(\sum_{l=1}^{q} \gamma_{U_l} U_{il}\Big))\\
&p(M_i = 1|A_i, \mathbf{C}_i, \{A_j, M_j|j \in \mathcal{N}_i\}; \tau_M)\\
&\qquad\qquad = expit(\beta_0 + \beta_A A_i + \Big(\sum_{l=1}^{p} \beta_{C_l} C_{il}\Big) + \Big(\sum_{j\in\mathcal{N}_i} (\beta_{A_{nb}} A_j + \beta_{M_{nb}} M_j)\Big))\\
&p(Y_i = 1|\mathbf{C}_i, \mathbf{U}_i, M_i, \{A_j|j \in \mathcal{N}_j\}; \tau_Y)\\
&\qquad\qquad = expit(\alpha_0 + \Big(\sum_{l=1}^{p} \alpha_{C_l} C_{il}\Big) + \Big(\sum_{l=1}^{q} \alpha_{U_l} U_{il}\Big) + \Big(\sum_{j=\mathcal{N}_i} \alpha_{A_{nb}} A_j\Big) + \alpha_M M_i).
\end{aligned}$$

The values of the parameters for the beta distributions we use to generate $\mathbf{C}_i, \mathbf{U}_i$ can be found in Table D-IIa while the values of $\tau_A, \tau_M, \tau_Y$ can be found in Table D-IIb.

### D.3.3 Extended Results

In the main paper we gave confidence intervals and the mean and standard deviation of the bias of our estimators. All results were calculated by averaging over 1000 simulated networks.

As discussed in the main body of the paper, the estimators we use are able to

| Variable | a | b |
|---|---|---|
| $C_1$ | 1.5 | 3 |
| $C_2$ | 6 | 2 |
| $C_3$ | 0.8 | 0.8 |
| $U_1$ | 2.3 | 1.1 |
| $U_2$ | 0.9 | 1.1 |
| $U_3$ | 2 | 2 |

(a) Parameters for **C** and **U**

| Parameter | Value |
|---|---|
| $\tau_A$ | (-1, 0.5, 0.2, 0.25, 0.3, -0.2, 0.25) |
| $\tau_M$ | (-1, -0.3, 0.4, 0.1, 1, -0.5, -1.5) |
| $\tau_Y$ | (-0.3, -0.2, 0.2, -0.05, 0.1, -0.2, 0.25, -1, 3) |

(b) Parameters for $\tau_A, \tau_M, \tau Y$

Table D-II. The parameters for each generating distribution

| Ground Truth Network Average Effects | | | | |
|---|---|---|---|---|
| $N$ | 400 | 800 | 1000 | 2000 |
| Ground Truth | -.455 | -.453 | -.455 | -.456 |

Table D-III. The ground truth effects for each network, calculated by averaging over 5 samples of the data generating process for each network under the relevant interventions

recover the effects of interest reasonably well. The approximate ground truth values for these effects can be found in Table D-III. The fact that the coding estimator restricts the network to a small fraction of its total units means it is considerably less efficient than the pseudo-likelihood estimator.

Though the pseudo-likelihood estimator is not in general asymptotically normal, it does not perform substantially worse than the provably asymptotically normal coding-likelihood estimator. In both cases, the true effect is covered by the 95% confidence interval of the estimator.

# Appendix E

# Supplementary Material for Chapter 5

## E.1 Proofs

**Lemma 1** *Given a segregated graph $\mathcal{G}(\mathbf{V})$ and a segregation-preserving policy intervention $\mathbf{f_A}(\mathbf{Z_A})$, the post-intervention graph $\mathcal{G}_{\mathbf{f_A}}$ obtained via Procedure 2 is a segregated graph.*

*Proof:* In order for $\mathcal{G}_{\mathbf{f_A}}$ to be a segregated graph, it must not have a node with both an incident bi-directed and undirected edge (the 'segregation' property) and it must not have any partially directed cycles (the 'chain' property).

We first show that $\mathcal{G}_{\mathbf{f_A}}$ satisfies the segregation property. First we consider edges that appear in both $\mathcal{G}$ and $\mathcal{G}_{\mathbf{f_A}}$ (potentially with a modified functional form). Since we do not add any $\leftrightarrow$ edges when constructing $\mathcal{G}_{\mathbf{f_A}}$, and since we assumed $\mathcal{G}$ is a segregated graph, these edges are all incident to nodes that do not also have incident directed edges.

We can therefore restrict attention to undirected edges that were newly created when constructing $\mathcal{G}_{\mathbf{f_A}}$. These edges correspond to connecting two previously unconnected nodes. This requires intervening on both end points, which entails removing all incident $\leftrightarrow$ edges, as described in Procedure 2. This accounts for all possible undirected edges. In particular, we cannot convert a directed edge $X \to Y$ to an undirected edge $X - Y$: this would require intervening on $X$ with $f_X(\mathbf{Z}_X)$ where $Y \in \mathbf{Z}_X$ which violates our construction that $\mathbf{Z} \subseteq \mathbf{V} \setminus \overline{\text{ext}}_{\mathcal{G}}(X)$.

Since no undirected edge is incident to a node that also has an incident bi-directed edge, $\mathcal{G}_{\mathbf{f_A}}$ satisfies the segregation property.

We now show that $\mathcal{G}_{\mathbf{f_A}}$ satisfies the chain property. We argue by contradiction: suppose $\mathcal{G}_{\mathbf{f_A}}$ *does* have a newly induced (relative to $\mathcal{G}$) partially directed cycle. Then, without loss of generality, one of the following sub-structures appears in $\mathcal{G}_{\mathbf{f_A}}$ but not in $\mathcal{G}$: (1) $W \to X \to Y \to W$, (2) $W \to X - Y \to W$, or (3) $W \to X - Y - W$.

Sub-structure (1) contradicts our assumption that $\mathbf{f_A}$ is segregation-preserving. Specifically, we have that $W \triangle X$ directly and $X \triangle W$ through $Y$, however $W \notin \mathbf{Z}_Y$.

In sub-structure (2), consider scenarios where two edges were present in $\mathcal{G}$ and we seek to add the third edge. When adding either the $W \to X$ or $Y \to W$ edge, we have that $W \triangle X$ and $X \triangle W$ (analogously for $W, Y$) but $X \notin \mathbf{Z}_W$ ($W \notin \mathbf{Z}_Y$) which is a contradiction. Meanwhile, adding the $X - Y$ edge requires that $Y \in \mathbf{Z}_X$, however $Y \in \overline{\text{ext}}_{\mathcal{G}}(X)$ which yields a contradiction. A similar argument involving $\triangle$ applies when only one of the three edges was present in $\mathcal{G}$ and we seek to add the other two.

In sub-structure (3) a similar argument applies. Suppose we seek to add the $W \to X$ edge with the two undirected edges present. $X \triangle W$ in the post-intervention graph but it is not the case that $W \triangle X$, yielding a contradiction. Adding the $Y - X$ edge yields a contradiction since $X \in \overline{\text{ext}}_{\mathcal{G}}(Y)$. Similarly, adding the $Y - W$ edge yields a contradiction since $Y \in \overline{\text{ext}}_{\mathcal{G}}(W)$. Again, we can make a similar argument for adding two of the three edges.

The above argument generalizes trivially to larger sub-structures in the graph (e.g., 4-cycles) and so $\mathcal{G}_{\mathbf{f_A}}$ will not have any partially directed cycles. Since $\mathcal{G}_{\mathbf{f_A}}$ satisfies both the chain property and the segregation property, it is a segregated graph. □

**Theorem 10** *Let $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$ be a causal LV-CG with $\mathbf{H}$ block-safe, and a topological order $\prec$. Fix disjoint $\mathbf{Y}, \mathbf{A} \subseteq \mathbf{V}$. Let $\mathbf{f}_{\mathbf{A}}(\mathbf{Z}_{\mathbf{A}})$ be a segregation preserving policy set. Let $\mathbf{Y}^\star \equiv \operatorname{ant}_{\mathcal{G}_{\mathbf{f}_{\mathbf{A}}}}(\mathbf{Y}) \setminus \mathbf{A}$. Let $\mathcal{G}^d, \tilde{G}^d$ be the induced CADMGs on $\mathcal{G}_{\mathbf{f}_{\mathbf{A}}}$ and $\mathcal{G}_{\mathbf{Y}^\star}$, and $\tilde{G}^b$ the induced CCG on $\mathcal{G}_{\mathbf{Y}^\star}$. Let $q(\mathbf{D}^\star | \operatorname{pa}^s_{\mathcal{G}_{\mathbf{f}_{\mathbf{A}}}}(\mathbf{D}^\star)) = \prod_{\mathbf{D} \in \mathcal{G}_{\mathbf{f}_{\mathbf{A}}}} q(\mathbf{D} | \operatorname{pa}^s_{\mathcal{G}_{\mathbf{f}_{\mathbf{A}}}}(\mathbf{D}))$, where $q(\mathbf{D} | \operatorname{pa}^s_{\mathcal{G}_{\mathbf{f}_{\mathbf{A}}}}(\mathbf{D})) = \prod_{D \in \mathbf{D}} p(D | \mathbf{V}_{\prec D})$ if $\mathbf{D} \cap \mathbf{A} = \emptyset$ and $q = f_A(\mathbf{Z}_A)$ if $\mathbf{D} \cap \mathbf{A} \neq \emptyset$. $p(\mathbf{Y}(\mathbf{f}_{\mathbf{A}}(\mathbf{Z}_{\mathbf{A}})))$ is identified in $\mathcal{G}$ if and only if $p(\mathbf{Y}^\star(\mathbf{a}))$ is identified in $\mathcal{G}$ for the unrestricted class of policies. If identified, $p(\mathbf{Y}(\mathbf{f}_{\mathbf{A}}(\mathbf{Z}_{\mathbf{A}}))) =$*

$$\sum_{\{\mathbf{Y}^\star \cup \mathbf{A}\} \setminus \mathbf{Y}} \left[ \prod_{\mathbf{B} \in \mathcal{B}(\tilde{\mathcal{G}}^b)} p^\star(\mathbf{B} | \operatorname{pa}_{\mathcal{G}_{\mathbf{f}_{\mathbf{A}}}}(\mathbf{B})) \right] \times \left[ \prod_{\mathbf{D} \in \mathcal{D}(\tilde{\mathcal{G}}^d)} \phi_{\mathbf{D}^\star \setminus \mathbf{D}}(q(\mathbf{D}^\star | \operatorname{pa}^s_{\mathcal{G}_{\mathbf{f}_{\mathbf{A}}}}(\mathbf{D}^\star)); \mathcal{G}^d) \right] \Bigg|_{\mathbf{A} = \tilde{\mathbf{a}}} \tag{E.1}$$

*where (a) $\tilde{\mathbf{a}} = \{A = f_A(\mathbf{Z}_A) : A \in \operatorname{pa}_{\mathcal{G}_{\mathbf{f}_{\mathbf{A}}}}(\mathbf{D}) \cap \mathbf{A}\}$ if $\operatorname{pa}_{\mathcal{G}_{\mathbf{f}_{\mathbf{A}}}}(\mathbf{D}) \cap \mathbf{A} \neq \emptyset$ and $\tilde{\mathbf{a}}_{\mathbf{D}} = \emptyset$ otherwise, and (b) $p^\star$ is obtained by running Procedure 1 over functions $g_{B_i}(B_{-i}, \operatorname{pa}_{\mathcal{G}_{\mathbf{f}_{\mathbf{A}}}}(B_i), \epsilon_{B_i})$ where $g_{B_i} \in \mathbf{f}_{\mathbf{A}}$ if $B_i \in \mathbf{A}$ and $g_{B_i}$ is given by the observed distribution if $B_i \notin \mathbf{A}$*[1].

*Proof:* We prove two subclaims.

**Claim 1: The segregated graph policy ID formula, equation E.1, is sound**

We first note that each variable in $\mathbf{V} \cup \mathbf{H}$ is defined by a structural equation model. Since $\mathbf{f}_{\mathbf{A}}$ is assumed to be segregation preserving, lemma 1 implies that all variables in $\mathbf{H}$ have an unchanged structural equation in $\mathbf{f}_{\mathbf{A}}$. Among $\mathbf{V}$ there exist two types of variables: those that have a symmetric functional dependence with another variable (i.e., for $V_i, V_j \in \mathbf{V}$ the structural equations $f_{V_i}, f_{V_j}$ are functions of each other), and those without symmetric dependence.

We impose an ordering on the variables in $\mathcal{G}_{\mathbf{f}_{\mathbf{A}}}$ in order of their dependence on other variables in the graph: we first evaluate variables $\mathbf{V} \in (\mathbf{V} \cup \mathbf{H})$ with structural equations that don't depend on other variables ($V \sim f_V(\epsilon_V)$) and then variables

[1] This distribution is identified from univariate terms but it cannot be obtained in closed-form.

that are functions of those variables and so on. Following Lauritzen and Richardson, 2002, groups of variables that have symmetrically dependent structural equations are chain components corresponding to $\mathcal{B}^{nt}(\mathcal{G}_{\mathbf{f_A}})$. Variables that do not exhibit symmetric dependence are trivial chain components. Our ordering therefore implies a DAG on chain components (it is acyclic aside from in-component cycles by lemma 1).

It's clear that for trivial chain components the functions $f_V$ immediately reach an equilibrium. We can normalize these functions, and write the margin over their corresponding variables as:

$$\prod_{V \in \mathbf{D} : \mathbf{D} \in \mathcal{D}(\mathcal{G}_{\mathbf{f_A}}(\mathbf{V} \cup \mathbf{H}))} p(V | \operatorname{pa}_{\mathcal{G}_{\mathbf{f_A}}(\mathbf{V} \cup \mathbf{H})}(V))|_{\mathbf{A} = \mathbf{f_A}}$$

Now, for each non-trivial chain component $\mathbf{B}$, the structural equations for each constituent variable treats inputs that are not in the component as known (this can be done since those variables are evaluated earlier in the ordering on the DAG of components) and evaluates each variable in the component via a Gibbs sampling process. The values obtained upon convergence can then be passed to components later in the ordering. This follows by application of proposition 6 in Lauritzen and Richardson, 2002, and so we can express the DAG factorization over chain components as:

$$\begin{aligned} p(\mathbf{V} \cup \mathbf{H}(\mathbf{f_A})) = & \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}_{\mathbf{f_A}}(\mathbf{V} \cup \mathbf{H}))} p(\mathbf{D} | \operatorname{pa}_{\mathcal{G}_{\mathbf{f_A}}(\mathbf{V} \cup \mathbf{H})}(\mathbf{D}))|_{\mathbf{A} = \mathbf{f_A}} \\ & \times \prod_{\mathbf{B} \in \mathcal{B}^{nt}(\mathcal{G}_{\mathbf{f_A}}(\mathbf{V} \cup \mathbf{H}))} p^{\star}(\mathbf{B} | \operatorname{pa}_{\mathcal{G}_{\mathbf{f_A}}(\mathbf{V} \cup \mathbf{H})}(\mathbf{B})) \end{aligned}$$

$\mathcal{G}_{\mathbf{f_A}}$ is a proper latent-variable chain graph.

We derive the remainder of the proof via the argument in the proof of theorem 2 in Sherman and Shpitser, 2018. We assume without loss of generality that $\mathbf{Y}$ has no children in $\mathcal{G}(\mathbf{V})$.

Consider the chain graph factorization of $\mathcal{G}_{\mathbf{f_A}}$ derived above. Because $\mathbf{H}$ is block-safe in $\mathcal{G}$, the non-trivial blocks term can be re-written as follows:

$$\begin{aligned}\prod_{\mathbf{B}\in\mathcal{B}^{nt}(\mathcal{G}_{\mathbf{f_A}}(\mathbf{V}\cup\mathbf{H}))} p^\star(\mathbf{B}|\operatorname{pa}_{\mathcal{G}_{\mathbf{f_A}}(\mathbf{V}\cup\mathbf{H})}(\mathbf{B})) &= \prod_{\mathbf{B}\in\mathcal{B}^{nt}(\mathcal{G}_{\mathbf{f_A}}(\mathbf{V}))} p(\mathbf{B}|\operatorname{pa}_{\mathcal{G}_{\mathbf{f_A}}(\mathbf{V})}(\mathbf{B})) \\ &= \prod_{\mathbf{B}\in\mathcal{B}^{nt}(\tilde{\mathcal{G}}^b)} p^\star(\mathbf{B}|\operatorname{pa}_{\mathcal{G}_{\mathbf{f_A}}(\mathbf{V})}(\mathbf{B}))|_{\mathbf{A}=\tilde{\mathbf{a}}_\mathbf{B}}\end{aligned}$$

We are now left with the following factorization for the overall graph:

$$\begin{aligned}p(\{\mathbf{V}\cup\mathbf{H}\}(\mathbf{f_A})) = &\prod_{\mathbf{B}\in\mathcal{B}^{nt}(\tilde{\mathcal{G}}^b)} p^\star(\mathbf{B}|\operatorname{pa}_{\mathcal{G}_{\mathbf{f_A}}(\mathbf{V})}(\mathbf{B})) \\ \times &\prod_{\mathbf{D}\in(\mathbf{V}\cup\mathbf{H})\setminus\left(\bigcup_{\mathbf{B}\in\mathcal{B}^{nt}(\mathcal{G}_{\mathbf{f_A}})}\mathbf{B}\right)} \prod_{V\in\mathbf{D}\setminus\mathbf{A}} p(V|\operatorname{pa}_{\mathcal{G}_{\mathbf{f_A}}(\mathbf{V}\cup\mathbf{H})}(V)) \prod_{V\in\mathbf{D}\cap\mathbf{A}} f_V(\mathbf{Z}_V)|_{\mathbf{A}=\mathbf{f_A}}\end{aligned}$$

The factors in the second term are singleton nodes by construction and so they are defined by either observed $p(V|\operatorname{pa}_{\mathcal{G}_{\mathbf{f_A}}(\mathbf{V}\cup\mathbf{H})}(V))$ if $V\notin\mathbf{A}$ and $f_V\in\mathbf{f_A}(\mathbf{Z}_V)$ if $V\in\mathbf{A}$.

If we marginalize $\mathbf{H}$ from this second set of terms, using standard procedures Tian and Pearl, 2002, then the resulting expression is the kernel described in the statement of the theorem: $q(\mathbf{D}^\star|\operatorname{pa}^s_{\mathcal{G}(\mathbf{V})}(\mathbf{D}^\star)) = \prod_{\mathbf{D}\in\mathcal{D}(\mathcal{G}_{\mathbf{f_A}})} q(\mathbf{D}|\operatorname{pa}^s_{\mathcal{G}_{\mathbf{f_A}}(\mathbf{V})}(\mathbf{D}))$, where $q(\mathbf{D}|\operatorname{pa}^s_{\mathcal{G}_{\mathbf{f_A}}(\mathbf{V})}(\mathbf{D})) = \prod_{D\in\mathbf{D}} p(\mathbf{D}|\mathbf{V}_{\prec\mathbf{D}})$ if $\mathbf{D}\cap\mathbf{A}=\emptyset$ and $q(\mathbf{D}|\operatorname{pa}^s_{\mathcal{G}_{\mathbf{f_A}}(\mathbf{V})}(\mathbf{D})) = f_A(\mathbf{Z}_A)$ if $\mathbf{D}\cap\mathbf{A}\neq\emptyset$.

Since $\mathbf{Z_A}$ are all observed by assumption, we can manipulate this kernel as in the proof of soundness for theorem 2 in Sherman and Shpitser, 2018. Whereas in Sherman and Shpitser, 2018 the authors fixed $\mathbf{A}$ to constants, here we can express setting $\mathbf{A}$ to stochastic values according to $\mathbf{f_A}$. The claim is then immediate.

**Claim 2: The segregated graph policy ID formula is complete**

We adapt the proof techniques in Shpitser and Sherman, 2018; Sherman and Shpitser, 2018. At a high level, we will use the fact that $p(\mathbf{Y}^\star(\mathbf{a}))$ is not identified to demonstrate that there is a hedge in $\mathcal{G}$. We will then extend the hedge down the graph to reach $\mathbf{Y}$ via $\operatorname{ext}_{\mathcal{G}_{\mathbf{Y}^\star}}(\text{hedge})$ and $\operatorname{ant}_{\mathcal{G}_{\mathbf{Y}^\star}}(\mathbf{Y})$ to show non-identification. We do this by arguing along the partially directed paths from the hedge to $\mathbf{Y}$, which requires considering subgraphs of $\mathcal{G}_{\mathbf{Y}^\star}$. We show non-identifiability in each of an increasingly restricted submodel of $\mathcal{G}_{\mathbf{Y}^\star}$ and then show that non-identification in the submodels yields non-identification in $\mathcal{G}_{\mathbf{Y}^\star}$. More concretely, there are two complications that must be dealt with for showing completeness of policy interventions: the hedge might intersect $\mathbf{Y}$ and we must extend the hedge down to $\mathbf{Y}$ via partially directed paths. We construct a subgraph for demonstrating the latter case and then a subgraph of that for the former case. We now proceed with the proof.

Suppose $p(\mathbf{Y}^\star(\mathbf{a}))$ is not identified in $\mathcal{G}$. Then there is a district $\mathbf{D} \in \mathcal{D}(\mathcal{G}_{\mathbf{Y}^\star})$ that is not reachable in $\mathcal{G}$. Let $\mathbf{R} = \{D \in \mathbf{D} | \operatorname{ch}_{\mathcal{G}}(D) \cap \mathbf{D} = \emptyset\}$. Let $\mathbf{A}^\star = \mathbf{A} \cap \operatorname{pa}_{\mathcal{G}}(D)$. Then there exists $\mathbf{D}' \supset \mathbf{D}$, such that $\mathbf{D}$ and $\mathbf{D}'$ form a hedge for $p(\mathbf{R}|\operatorname{do}(\mathbf{a}^\star))$ and thus $p(\mathbf{R}|\operatorname{do}(\mathbf{a}^\star))$ is not identified by Shpitser and Pearl, 2006.

Let $\mathbf{Y}'$ be the minimal subset of $\mathbf{Y}$ such that $\mathbf{R} \subseteq \operatorname{ant}_{\mathcal{G}_{\mathbf{f_A}}}(\mathbf{Y}')$. Consider a subgraph $\mathcal{G}^\dagger$ of $\mathcal{G}_{\mathbf{f_A}}$, with vertices $\mathbf{V}' \subseteq \mathbf{V}$, consisting of all edges in $\mathcal{G}$ in the hedge on $\mathbf{D}, \mathbf{D}'$ described above, and edges that lie in partially directed paths in $\mathcal{G}_{\mathbf{f_A}}$ from $\mathbf{R}$ to $\mathbf{Y}'$. We restrict attention, without loss of generality, to at most one child per node in each partially directed path such that our paths form a forest from $\mathbf{R}$ to $\mathbf{Y}'$. By Lemma 1, $\mathcal{G}^\dagger$ does not contain any directed, nor partially directed cycles. Let $\mathbf{A}^\dagger = \{\mathbf{A}^\star \cup A | A \in \mathbf{A} \text{ in } \mathcal{G}^\dagger\}$. For each $A^\dagger \in \mathbf{A}^\dagger$, we restrict attention to policies that map from $\mathbf{Z}^\dagger_{A^\dagger}$ to $A^\dagger$, where $\mathbf{Z}^\dagger_{A^\dagger} = \mathbf{Z}_{A^\dagger} \cap \mathbf{V}'$.

Now, following the proof of theorem 2 in the supplement of Sherman and Shpitser, 2018, we define an ADMG $\tilde{\mathcal{G}}^\dagger$ which has the same vertices and edges as the $\mathbf{D}, \mathbf{D}'$ hedge in $\mathcal{G}^\dagger$, and has a copy of each vertex in each partially directed path from $\mathbf{R}$ to $\mathbf{Y}'$ in $\mathcal{G}^\dagger$ but replaces all the undirected edges on those partially directed paths with directed edges oriented away from $\mathbf{R}$ towards $\mathbf{Y}'$. We denote the variable copies in $\tilde{\mathcal{G}}^\dagger$ corresponding to $\mathbf{Y}'$ in $\mathcal{G}^\dagger$ by $\tilde{\mathbf{Y}}'$. This orientation is possible because each undirected edge either corresponds to a (known) policy in the intervention set, or to an observed structural equation. In either case, the observed distribution continues to argree between the two counterexamples witnessing non-identifiability. For $\mathbf{A}^\dagger$ in

$\tilde{\mathcal{G}}^{\dagger}$, we further restrict attention to policies inducing directed edges from $\mathbf{R}$ to $\tilde{\mathbf{Y}}'$ (i.e. ignoring policies going the opposite direction that induce undirected edges). We denote these nodes by $\tilde{\mathbf{A}}^{\dagger}$.

We now show that $p(\tilde{\mathbf{Y}}'(\{\tilde{A}^{\dagger} = f_{\tilde{A}^{\dagger}} | \tilde{A}^{\dagger} \in \tilde{\mathbf{A}}^{\dagger}\}))$ is not identified in $\tilde{\mathcal{G}}^{\dagger}$ following the argument in the proof of theorem 6 in the supplement of Shpitser and Sherman, 2018. Observe that for $\mathbf{R} \subseteq \tilde{\mathbf{Y}}'$, the subclaim is immediate by the recursive argument in the proof of theorem 4 in Shpitser and Sherman, 2018. Otherwise, pick a node $\tilde{Y}'$ in $\tilde{\mathcal{G}}^{\dagger}$ such that $\text{pa}_{\tilde{\mathcal{G}}^{\dagger}}(\tilde{Y}') \subseteq \mathbf{R}$ and $\text{pa}_{\tilde{\mathcal{G}}^{\dagger}}(\tilde{Y}') \setminus \tilde{\mathbf{Y}}' \neq \emptyset$ (as in Shpitser and Sherman, 2018, such a vertex must exist since $\tilde{\mathcal{G}}^{\dagger}$ is acyclic and $\mathbf{R} \setminus \tilde{\mathbf{Y}}' \neq \emptyset$). If this $\tilde{Y}' \in \tilde{\mathbf{A}}^{\dagger} \setminus \mathbf{A}^{\star}$, the subclaim is immediate since $\tilde{Y}'$ does not intersect our hedge and we can extend down the graph using the argument in theorem 4 of Shpitser and Sherman, 2018.

If $\tilde{Y}' \in \mathbf{A}^{\star}$ then we can create a graph $\bar{\mathcal{G}}$ by copying the variables on the path $\tilde{Y}' \to V_1 \to \cdots \to \bar{Y} \in \tilde{\mathbf{Y}}'$ in $\tilde{\mathcal{G}}^{\dagger}$. We then apply the argument in theorem 4 of Shpitser and Sherman, 2018 to show that $p(\bar{Y}(\mathbf{a}^{\star}))$ is not identified along this path when we set $\mathbf{a}^{\star}$ according to the policies specified by $\mathbf{f}_{\mathbf{A}^{\star}}$. This follows since, by assumption, $\mathbf{f}_{\mathbf{A}^{\star}} \subseteq \mathbf{f}_{\mathbf{A}}$ lies in an unrestricted policy class. Now, as $p(\bar{Y}(\mathbf{a}^{\star}))$ is not identified in $\bar{\mathcal{G}}$, we can use the two counterexamples witnessing non-identifiability in $\bar{\mathcal{G}}$ to obtain non-identifiability for $p(\tilde{\mathbf{Y}}'(\mathbf{f}_{\tilde{\mathbf{A}}^{\dagger}}))$. To do so, we define new variables in $\tilde{\mathcal{G}}^{\dagger}$ that are the Cartesian product of variable copies created in $\bar{\mathcal{G}}$ and their corresponding variables in $\tilde{\mathcal{G}}^{\dagger}$. Non-identifiability follows via the standard argument in lemma 1 of Shpitser and Sherman, 2018.

Now that we have shown that $p(\tilde{\mathbf{Y}}'(\{\tilde{A}^{\dagger} = f_{\tilde{A}^{\dagger}} | \tilde{A}^{\dagger} \in \tilde{\mathbf{A}}^{\dagger}\}))$, we have two counterexamples witnessing non-identifiability in $\tilde{\mathcal{G}}^{\dagger}$ which agree on the observed data distribution but disagree on the counterfactual distribution. We use these counterexamples to demonstrate non-identifiability of $p(\mathbf{Y}'(\{A = f_A | A \in \mathbf{A}^{\star}\}))$ in $\mathcal{G}^{\dagger}$. To do so, we define variables along the partially directed paths from $\mathbf{R}$ to $\mathbf{Y}'$ in $\mathcal{G}^{\dagger}$. These variables are created by taking the Cartesian product of variable copies in $\tilde{\mathcal{G}}^{\dagger}$ and the corresponding variables in $\mathcal{G}^\dagger$. As before, the counterexamples continue to agree on the observed data distribution and disagree on the counterfactual distribution. Thus $p(\mathbf{Y}'(\{A = f_A | A \in \mathbf{A}^\star\}))$ is not identified in $\mathcal{G}^\dagger$. Since $\mathbf{Y}' \subseteq \mathbf{Y}^\star$, the result is immediate, subject to the remaining argument on the chain graph properties of $\mathcal{G}^\dagger$ and $\tilde{\mathcal{G}}^\dagger$ below.

Following Sherman and Shpitser, 2018, fix a block $\mathbf{B}$ in $\mathcal{G}^\dagger$. For any $B \in \mathbf{B}$, there exists a set of variables $B_1, \ldots, B_k$ in $\tilde{\mathcal{G}}^\dagger$ such that $B$ is defined as the Cartesian product of $B_1, \ldots, B_k$. Any variable $A \in \text{nb}_{\mathcal{G}^\dagger} \cup \text{pa}_{\mathcal{G}^\dagger}(B)$ is similarly a Cartesian product of $A$ variables. Then it follows that $B \perp\!\!\!\perp ((\text{pa}_{\mathcal{G}^\dagger}(\mathbf{B}) \cup \mathbf{B}) \setminus (\text{nb}_{\mathcal{G}^\dagger}(B) \cup \text{pa}_{\mathcal{G}^\dagger}(B))) | (\text{nb}_{\mathcal{G}^\dagger}(B) \cup \text{pa}_{\mathcal{G}^\dagger}(B))$ by d-separation rules in the ADMG $\tilde{\mathcal{G}}^\dagger$ and that there are no colliders in $\tilde{\mathcal{G}}^\dagger$. These both follow from our vertex copy argument which separates out $B$ from the rest of the block and eliminates the possibility of colliders by making every path from $\mathbf{R}$ to $\mathbf{Y}'$ a partially directed chain. This demonstrates that $\mathcal{G}^\dagger$ and $\tilde{\mathcal{G}}^\dagger$ (and trivially $\bar{\mathcal{G}}$) satisfy the independence constraints implied by the CG Markov property, thus proving the claim. □

## E.2 Derivation of the Figure 5-2 Functional

From Fig. 5-2(a), we obtain $\mathcal{G}_{\mathbf{f_A}}$ in Fig. 5-2(b) by applying the intervention detailed in Table 5-II. In turn, from this post-intervention graph we observe that $\mathbf{Y}^\star = \text{ant}_{\mathcal{G}_{\mathbf{f_A}}}(\mathbf{Y}) \setminus \mathbf{A} = \{C_2, C_3, M_3, Y_2, Y_3\}$ and obtain the induced subgraph $\mathcal{G}_{\mathbf{Y}^\star}$ in Fig. 5-2(c).

$\mathcal{G}_{\mathbf{Y}^\star}$ factorizes into kernels relating to district nodes and block nodes:

$$q_{\mathcal{D}}(C_1, A_1, M_1, Y_1, Y_2, Y_3 | C_2, M_2, M_3)$$

and

$$q_{\mathcal{B}}(M_2, M_3, A_2, A_3, C_2, C_3 | \emptyset).$$

The block nodes factorize as a product of blocks, as in the first term of Eq. E.1:

$$\begin{aligned}
q_{\mathcal{B}}(\mathbf{B}^{\star} | \operatorname{pa}_{\mathcal{G}_{\mathbf{f_A}}}(\mathbf{B}^{\star})) &= \prod_{\mathbf{B} \in \mathcal{B}(\tilde{\mathcal{G}}^b)} p^{\star}(\mathbf{B} | \operatorname{pa}_{\mathcal{G}_{\mathbf{f_A}}}(\mathbf{B})) \\
&= p^{\star}(M_2, M_3 | A_2, A_3, C_2) p^{\star}(A_2, A_3 | C_2, C_3) p^{\star}(C_2, C_3)
\end{aligned}$$

Note that $p^{\star}(C_2, C_3) = p(C_2, C_3)$ since the $C_2 - C_3$ block is unchanged relative to the observed data.

Separately, we must fix sets for each $\mathcal{G}_{\mathbf{Y}^{\star}}$ district $\{\{M_3\}, \{Y_2, Y_3\}\}$ in $q_{\mathcal{D}(\mathcal{G})}$. The derivations of these pieces is as follows:

$$\begin{aligned}
&\phi_{\mathbf{D}^{\star} \setminus \{M_3\}}(q(C_1, A_1, M_1, Y_1, Y_2, Y_3 | C_2, M_2, M_3); \mathcal{G}^d) = \\
&\qquad \phi_{\mathbf{D}^{\star}}(q(C_1, A_1, M_1, Y_1, Y_2, Y_3 | C_2, M_2, M_3); \mathcal{G}^d)
\end{aligned}$$

This follows since $M_3$ is already fixed in this kernel and subgraph. Since we must fix all variables in the kernel and all variables in the kernel are fixable, this term simplifies to $p(\emptyset) = 1$.

For the second kernel, we have: $\phi_{\mathbf{D}^{\star} \setminus \{Y_2, Y_3\}}(q(C_1, A_1, M_1, Y_1, Y_2, Y_3 | C_2, M_2, M_3); \mathcal{G}^d)$

$$\begin{aligned}
&= \phi_{C_1, A_1, M_1, Y_1}(q(C_1, A_1, M_1, Y_1, Y_2, Y_3 | C_2, M_2, M_3); \mathcal{G}^d) \\
&= \phi_{A_1, M_1, Y_1}(\frac{q(C_1, A_1, M_1, Y_1, Y_2, Y_3 | C_2, M_2, M_3)}{p(C_1)}; \phi_{C}(\mathcal{G}^d)) \\
&= \phi_{A_1, M_1, Y_1}(q(A_1, M_1, Y_1, Y_2, Y_3 | C_2, M_2, M_3, C_1); \phi_{C_1}(\mathcal{G}^d)) \\
&= \phi_{M_1, Y_1}(\frac{q(A_1, M_1, Y_1, Y_2, Y_3 | C_2, M_2, M_3, C_1)}{p(A_1 | C_1)}; \phi_{C_1, A_1}(\mathcal{G}^d)) \\
&= \phi_{M_1, Y_1}(q(M_1, Y_1, Y_2, Y_3 | C_2, M_2, M_3, C_1, A_1); \phi_{C_1, A_1}(\mathcal{G}^d)) \\
&= \phi_{Y_1}(\frac{q(M_1, Y_1, Y_2, Y_3 | C_2, M_2, M_3, C_1, A_1)}{p(M_1 | A_1)}; \phi_{C_1, A_1, M_1}(\mathcal{G}^d)) \\
&= \phi_{Y_1}(q(Y_1, Y_2, Y_3 | C_2, M_2, M_3, C_1, A_1, M_1); \phi_{C_1, A_1, M_1}(\mathcal{G}^d)) \\
&= \frac{q(Y_1, Y_2, Y_3 | C_2, M_2, M_3, C_1, A_1, M_1)}{p(Y_1 | A_1, M_1)}; \phi_{C_1, A_1, M_1, Y_1}(\mathcal{G}^d)) \\
&= p(Y_2, Y_3 | C_1, C_2, M_1, M_2, M_3, A_1, Y_1)
\end{aligned}$$

This yields the functional for $p(\{Y_2, Y_3\}(\mathbf{f_A}))$:

$$\sum_{\{A_1,A_2,A_3,M_2,M_3,C_2,C_3\}} \Big( p^\star(M_2, M_3|A_2, A_3, C_2) p^\star(A_2, A_3|C_2, C_3) p^\star(C_2, C_3) \\ \times p(Y_2, Y_3|C_1, C_2, M_1, M_2, M_3, A_1, Y_1) \Big)$$

## E.3 Experimental Details and Extended Results

The parameters for the Beta distribution for $C$ for both types of experiments (policy and bias) are given by:

| $\alpha$ | $\beta$ |
|---|---|
| 1.5 | 3 |
| 6 | 2 |
| .8 | .8 |

Table E-I. Parameters for generating $C_i$

The parameters for $A_i$ and $Y_i$ differ between the bias and policy experiments. For $A$ we have:

| Parameter | Bias | Policy |
|---|---|---|
| $\gamma_1$ | 1 | .5 |
| $\gamma_2$ | 0 | .2 |
| $\gamma_3$ | 0 | .25 |
| $\tau_{AC}$ | 0 | .15 |

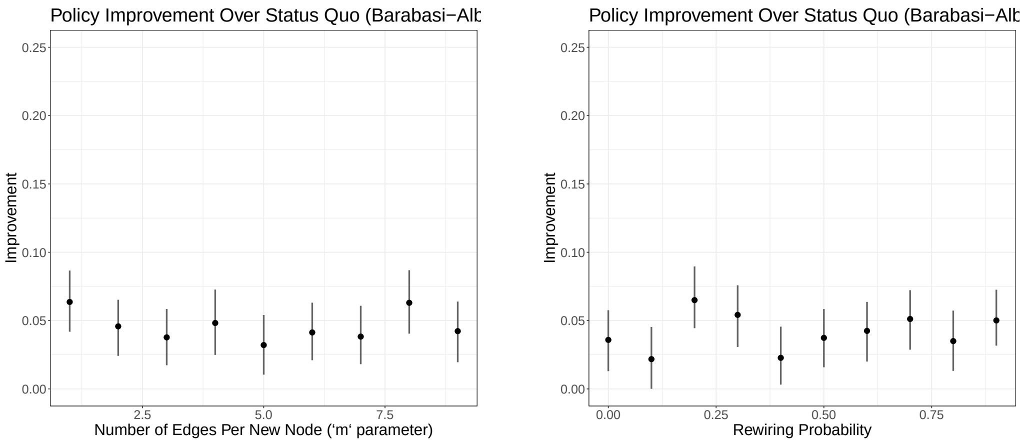
Table E-II. Parameters for generating $A_i$

And for $Y$ we have:

Finally, for the policy experiment we have results (Figure E-1b) similar to those in the main draft, which demonstrate the efficacy of policy interventions in selection actions that yield a more optimal outcome.

| Parameter | Bias | Policy |
|---|---|---|
| $\eta$ | -3 | .6 |
| $\delta_1$ | 1 | -.3 |
| $\delta_2$ | 0 | .4 |
| $\delta_3$ | 0 | .1 |
| $\tau_{YA}$ | 3 | .2 |
| $\tau_{YY}$ | .1 | .3 |
| $\tau_{YC}$ | 0 | -.2 |

Table E-III. Parameters for generating $Y_i$



(a) (b)

Figure E-1. Difference in expected outcomes between adopting an optimal strategy and using the status quo strategy in the Barabási-Albert model E-1a and the Watts-Strogatz small world model E-1b. We perform these analyses at several network densities to demonstrate the general efficacy of this approach.

# Appendix F

# Supplementary Material for Chapter 6

## F.1 Additional Network Intervention Examples

### Housing Vouchers

In urban development economics, housing vouchers have been proposed as a means of inducing families to move from areas with less opportunity for upward socioeconomic mobility to areas of with greater opportunity (Chetty, Hendren, and Katz, 2016). As pointed out in Hudgens and Halloran (2008), oftentimes these proposals ignore the social network-related implications of such a policy (e.g. decision is impacted by talking to neighbors). In Fig. F-1 we consider the specific impacts of transplanting a family from one neighborhood to the other.

Each unit $i$ in Fig. F-1 has a variable $C_i$ representing the unit's demographic information, and an outcome $Y_i$ which represents the socioeconomic outcome targeted by the housing voucher, such as annual income. In Fig F-1 (a) we see the pre-intervention network, where unit 3 is neighbors with unit 2, while after the intervention (Fig. F-1 (b)) unit 3 has been moved to a new neighborhood and is now neighbors with unit 4 instead. This example represents both a severance and a connection intervention.

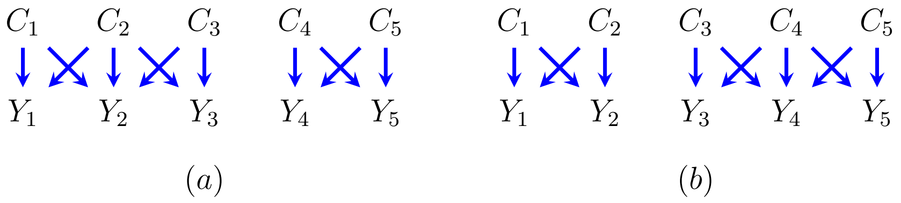


Figure F-1. (a) A DAG representing hypothetical connections between family units in two separate neighborhoods; (b) the DAG resulting from moving unit 3 from the first neighborhood to the second neighborhood via both a severance and connection intervention.

### Influencer Networks

Understanding social influence in networks is of interest to a wide variety of fields. Researchers who study infectious diseases and intravenous drug abuse often attempt to identify major influencers that impact many people. More recently, determining and leveraging knowledge of influence has become the focus of the algorithmic marketing community. If one can identify the strongest influencer in a network (not necessarily the individual with the most connections), then understanding the effects of removing that individual from the network (e.g. by arresting a drug-kingpin) might be useful for policymakers.

In Fig. F-2, we represent this phenomenon via an undirected graph where each unit, represented by $Y_i$, is internally a DAG and the undirected edges between units encode symmetric directed relationships similar to those in our other examples. Fig. F-2 (b) depicts the result of a hypothetical intervention on Fig. F-2 (a) in which unit 3 is effectively *removed* from the network by severing all connections with it's friends. As a side-effect unit 4 is also effectively removed from the network.

## F.2 Proofs

We first prove two utility results on factorizations of joint densities following Chen (2007). The first is a simple lemma which is needed to prove the corollary that follows.

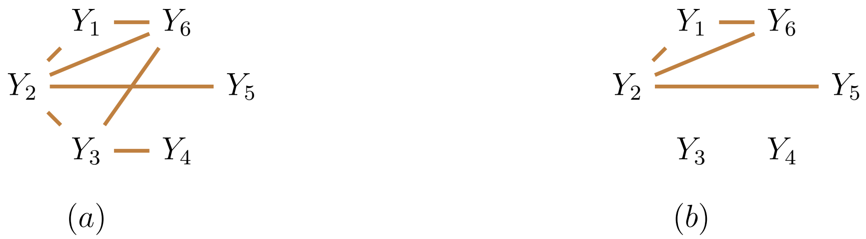


Figure F-2. (a) An undirected graph representing connections in an influence network between 6 agents; (b) the undirected graph resulting from intervening on the network in (a) such that unit 3 is removed from the network.

We will use Corollary 6 in each of the results that follow.

Let $\mathbf{a}$ be a set of fixed values. Then for a two-variable conditional density $f(Y_1, Y_2 \mid \mathbf{a})$, we have:

$$\frac{f(Y_1 \mid Y_2^0, \mathbf{a}) OR(Y_1, Y_2 \mid \mathbf{a}) f(Y_2 \mid Y_1^0, \mathbf{a})}{\sum_{Y_1, Y_2} f(Y_1 \mid Y_2^0, \mathbf{a}) OR(Y_1, Y_2 \mid \mathbf{a}) f(Y_2 \mid Y_1^0, \mathbf{a})},$$

where the odds ratio is given by

$$\begin{aligned} OR(Y_1, Y_2 \mid \mathbf{a}) &= \frac{f(Y_1 \mid Y_2, \mathbf{a}) f(Y_1^0 \mid Y_2^0, \mathbf{a})}{f(Y_1 \mid Y_2^0, \mathbf{a}) f(Y_1^0 \mid Y_2, \mathbf{a})} \\ &= \frac{f(Y_1, Y_2 \mid \mathbf{a}) f(Y_1^0, Y_2^0 \mid \mathbf{a})}{f(Y_1, Y_2^0 \mid \mathbf{a}) f(Y_1^0, Y_2 \mid \mathbf{a})}. \end{aligned}$$

and $Y_1^0$ and $Y_2^0$ signify reference values of $Y_1$ and $Y_2$.

**Lemma 3**

$$\frac{f(Y \mid Z, X)}{f(Y_0 \mid Z, X)} = \frac{f(Y \mid X) OR(Z, Y \mid X)}{f(Y_0 \mid X) \mathbb{E}[OR(Z, Y \mid X) \mid Y_0, X]}$$

*Proof:*

$$
\begin{aligned}
\frac{f(Y \mid Z, X)}{f(Y_0 \mid Z, X)} &=^1 \frac{f(Y \mid X) f(Z \mid Y, X)/f(Z \mid X)}{f(Y_0 \mid X) f(Z \mid Y_0, X)/f(Z \mid X)} \\
&=^2 \frac{f(Y \mid X) f(Z \mid Y, X)}{f(Y_0 \mid X) f(Z \mid Y_0, X)} \\
&=^3 \frac{f(Y \mid X) \frac{f(Z|Y_0,X) OR(Z,Y|X)}{\sum_Z f(Z|Y_0,X) OR(Z,Y|X)}}{f(Y_0 \mid X) f(Z \mid Y_0, X)} \\
&=^4 \frac{f(Y \mid X) OR(Z, Y \mid X)}{f(Y_0 \mid X) \mathbb{E}[OR(Z, Y \mid X) \mid Y_0, X]},
\end{aligned}
$$

where equality 1 is by Bayes rule, 2 by cancellation, 3 by the Chen factorization of a conditional density, and 4 by definition. □

**Corollary 6** $f(Y_1, Y_2 \mid \mathbf{a}) = \frac{\mathbf{X}}{\sum_{Y_1,Y_2} \mathbf{X}}$ *where* $\mathbf{X} =$

$$
\frac{f(Y_1 \mid \mathbf{a}) \; OR(Y_1, Y_2 \mid \mathbf{a}) OR(Y_1^0, Y_2|\mathbf{a}) OR(Y_1, Y_2^0|\mathbf{a}) \; f(Y_2 \mid \mathbf{a})}{\mathbb{E}[OR(Y_1, Y_2^0 \mid \mathbf{a}) \mid Y_1^0, \mathbf{a}] \mathbb{E}[OR(Y_1^0, Y_2 \mid \mathbf{a}) \mid Y_2^0, \mathbf{a}]}
$$

*Proof:* We have the following for $f(Y_1, Y_2 \mid \mathbf{a})$ (from Chen (2007)):

$$
\begin{aligned}
&\frac{f(Y_1 \mid Y_2^0, \mathbf{a}) OR(Y_1, Y_2 \mid \mathbf{a}) f(Y_2 \mid Y_1^0, \mathbf{a})}{\sum_{Y_1,Y_2} f(Y_1 \mid Y_2^0, \mathbf{a}) OR(Y_1, Y_2 \mid \mathbf{a}) f(Y_2 \mid Y_1^0, \mathbf{a})} \\
&=^1 \frac{\frac{f(Y_1|Y_2^0,\mathbf{a})}{f(Y_1^0|Y_2^0,\mathbf{a})} OR(Y_1, Y_2 \mid \mathbf{a}) \frac{f(Y_2|Y_1^0,\mathbf{a})}{f(Y_2^0|Y_1^0,\mathbf{a})}}{\sum_{Y_1,Y_2} \frac{f(Y_1|Y_2^0,\mathbf{a})}{f(Y_1^0|Y_2^0,\mathbf{a})} OR(Y_1, Y_2 \mid \mathbf{a}) \frac{f(Y_2|Y_1^0,\mathbf{a})}{f(Y_2^0|Y_1^0,\mathbf{a})}} \\
&=^2 \left[ \frac{f(Y_1 \mid \mathbf{a})}{f(Y_1^0 \mid \mathbf{a})} \frac{OR(Y_1, Y_2^0 \mid \mathbf{a})}{\mathbb{E}[OR(Y_1, Y_2^0 \mid \mathbf{a}) \mid Y_1^0, \mathbf{a}]} \right. \\
&\qquad \times OR(Y_1, Y_2 \mid \mathbf{a}) \times \\
&\qquad \left. \frac{OR(Y_1^0, Y_2 \mid \mathbf{a})}{\mathbb{E}[OR(Y_1^0, Y_2 \mid \mathbf{a}) \mid Y_2^0, \mathbf{a}]} \frac{f(Y_2 \mid \mathbf{a})}{f(Y_2^0 \mid \mathbf{a})} \right] \\
&\times \left[ \sum_{Y_1,Y_2} \left[ \frac{f(Y_1 \mid \mathbf{a})}{f(Y_1^0 \mid \mathbf{a})} \frac{OR(Y_1, Y_2^0 \mid \mathbf{a})}{\mathbb{E}[OR(Y_1, Y_2^0 \mid \mathbf{a}) \mid Y_1^0, \mathbf{a}]} \right. \right. \\
&\qquad \times OR(Y_1, Y_2 \mid \mathbf{a}) \times \\
&\qquad \left. \left. \frac{OR(Y_1^0, Y_2 \mid \mathbf{a})}{\mathbb{E}[OR(Y_1^0, Y_2 \mid \mathbf{a}) \mid Y_2^0, \mathbf{a}]} \frac{f(Y_2 \mid \mathbf{a})}{f(Y_2^0 \mid \mathbf{a})} \right] \right]^{-1} \\
&=^3 \frac{\mathbf{X}}{\sum_{Y_1,Y_2} \mathbf{X}}.
\end{aligned}
$$

Equality 1 holds by probability rules since $Y_1^0$ and $Y_2^0$ are fixed. Equality 2 holds by application of Lemma 1. Equality 3 holds by reverse application of the identity for $Y_1^0$ and $Y_2^0$. □

We now prove the main results of the paper. Each applies Corollary 6 to show increasingly general results pertaining to the KL-divergence of a distribution $\tilde{p}$ with additional independence constraints relative to an observational distribution $p$.

**Theorem 11** *Let* $\mathbf{V}$ *be a set of random variables with* $p(\mathbf{V})$ *corresponding to a DAG* $\mathcal{G}$. *Let* $A \in \mathbf{V}$. *Let* $\mathcal{P}(\mathbf{V})$ *be the set of probability distributions that factorize according to* $\mathcal{G}$. *Then*

$$p(A) \prod_{V \in \mathbf{V} \setminus A} p(V \mid \text{pa}_{\mathcal{G}}(V)) = \underset{\tilde{p} \in \mathcal{P}(\mathbf{V})}{\arg\min} D_{KL}(p||\tilde{p}) \\ \qquad\qquad\qquad\qquad s.t.\ A \perp\!\!\!\perp \text{pa}_{\mathcal{G}}(A)$$

*Proof:* Applying Corollary 6, we can express the KL-divergence of $\tilde{p}(A, \text{pa}_{\mathcal{G}}(A))$ from $p(A, \text{pa}_{\mathcal{G}})$ as proportional to:

$$\begin{aligned}
&\log \frac{p(A, \text{pa}_{\mathcal{G}}(A))}{\tilde{p}(A, \text{pa}_{\mathcal{G}}(A))} \\
&\quad = \log \frac{\left[\frac{p(A) \frac{OR_{num}}{OR_{den}} p(\text{pa}_{\mathcal{G}}(A))}{\sum_{A, \text{pa}_{\mathcal{G}}(A)} p(A) \frac{OR_{num}}{OR_{den}} p(\text{pa}_{\mathcal{G}}(A))}\right]}{\left[\frac{\tilde{p}(A) \frac{\widetilde{OR}_{num}}{\widetilde{OR}_{den}} \tilde{p}(\text{pa}_{\mathcal{G}}(A))}{\sum_{A, \text{pa}_{\mathcal{G}}(A)} \tilde{p}(A) \frac{\widetilde{OR}_{num}}{\widetilde{OR}_{den}} \tilde{p}(\text{pa}_{\mathcal{G}}(A))}\right]} \\
&\quad = \log \frac{p(A)}{\tilde{p}(A)} + \log \frac{p(\text{pa}_{\mathcal{G}}(A))}{\tilde{p}(\text{pa}_{\mathcal{G}}(A))} \\
&\quad \log \frac{OR_{num}}{OR_{den}} - \log \frac{\widetilde{OR}_{num}}{\widetilde{OR}_{den}} \\
&\quad + \log \sum_{A, \text{pa}_{\mathcal{G}}(A)} \tilde{p}(A) \frac{\widetilde{OR}_{num}}{\widetilde{OR}_{den}} \tilde{p}(\text{pa}_{\mathcal{G}}(A)) \\
&\quad - \log \sum_{A, \text{pa}_{\mathcal{G}}(A)} p(A) \frac{OR_{num}}{OR_{den}} p(\text{pa}_{\mathcal{G}}(A))
\end{aligned}$$

where we apply the Chen factorization for $p(A, \text{pa}_{\mathcal{G}}(A))$ and $\tilde{p}(A, \text{pa}_{\mathcal{G}}(A))$ and

$$\begin{aligned} OR_{num} &= OR(A, \text{pa}_{\mathcal{G}}(A))OR(A, \text{pa}_{\mathcal{G}}(A)^0)OR(A^0, \text{pa}_{\mathcal{G}}(A)) \\ OR_{den} &= E[OR(A, \text{pa}_{\mathcal{G}}(A)^0)|A^0]\times \\ &\quad E[OR(A^0, \text{pa}_{\mathcal{G}}(A))|\,\text{pa}_{\mathcal{G}}(A)^0] \end{aligned}$$

and analogously for the $\widetilde{OR}$'s.

Suppose we pick $\tilde{p}(A, \text{pa}_{\mathcal{G}}(A)) = p(A)p(\text{pa}_{\mathcal{G}}(A))$. Then $A \perp\!\!\!\perp \text{pa}_{\mathcal{G}}(A)$ in $\tilde{p}$ and so $\frac{\widetilde{OR}_{num}/\widetilde{OR}_{den}}{\sum_{A,\text{pa}_{\mathcal{G}}(A)} \tilde{p}(A)\frac{\widetilde{OR}_{num}}{\widetilde{OR}_{den}}\tilde{p}(\text{pa}_{\mathcal{G}}(A))} = 1$. Thus, the previous expression simplifies to:

$$\log \frac{OR_{num}}{OR_{den}} - \log \sum_{A,\text{pa}_{\mathcal{G}}(A)} p(A)\frac{OR_{num}}{OR_{den}}p(\text{pa}_{\mathcal{G}}(A)) \tag{F.1}$$

Suppose we instead picked some *other* $\tilde{p}(A, \text{pa}_{\mathcal{G}}(A)) = \tilde{p}(A)\tilde{p}(\text{pa}_{\mathcal{G}}(A))$ (i.e. one in which $A \perp\!\!\!\perp \text{pa}_{\mathcal{G}}(A)$). Then the above expression would have additional non-zero terms $\log \frac{p(A)}{\tilde{p}(A)} + \log \frac{p(\text{pa}_{\mathcal{G}}(A))}{\tilde{p}(\text{pa}_{\mathcal{G}}(A))}$. For this alternative $\tilde{p}$ to yield a lower KL divergence than that given by Eq. F.1, one of the terms, $\log \frac{p(A)}{\tilde{p}(A)}$ or $\log \frac{p(\text{pa}_{\mathcal{G}}(A))}{\tilde{p}(\text{pa}_{\mathcal{G}}(A))}$, must be less than 0 (since the other terms in Eq. F.1 remain the same under the independence of $A$ and $\text{pa}_{\mathcal{G}}(A)$). However, if $\log \frac{p(A)}{\tilde{p}(A)} < 0$ then the KL-divergence of $\tilde{p}(A)$ from $p(A)$ is negative, which violates Gibbs' inequality. The same holds for the distributions over $\text{pa}_{\mathcal{G}}(A)$. We therefore can conclude that $\tilde{p}(A, \text{pa}_{\mathcal{G}}(A)) = p(A)p(\text{pa}_{\mathcal{G}}(A))$ is the KL-closest distribution to $p(A, \text{pa}_{\mathcal{G}}(A))$ such that $A \perp\!\!\!\perp \text{pa}_{\mathcal{G}}(A)$.

Now, as a subclaim, we prove: if $\tilde{p}$ is KL-closest to $p$, then any conditional obtained from $\tilde{p}$ (by dividing by some (potentially conditional) distribution $p^*$ over a variable $V \in \mathbf{B}$), is KL-closest to the corresponding conditional obtained from $p$.

This is a simple consequence of the formula for KL-divergence:

$$D_{KL}(\frac{p}{p^*}||\frac{\tilde{p}}{p^*}) \propto \log \frac{\frac{p}{p^*}}{\frac{\tilde{p}}{p^*}} = \log \frac{p}{\tilde{p}}$$

The KL-divergence between the two distributions does not change by conditioning.

By the above two subclaims, the KL-closest distribution $\tilde{p}(A|\operatorname{pa}_{\mathcal{G}}(A))$ to $p(A|\operatorname{pa}_{\mathcal{G}}(A))$ is $p(A)$. By the local Markov property of DAGs, the claim holds, with $\tilde{p}(\mathbf{V} \setminus A) = p(\mathbf{V} \setminus A)$. □

**Theorem 12** *Let $\mathbf{V}$ be a set of random variables with $p(\mathbf{V})$ corresponding to a DAG $\mathcal{G}$. Let $A \in \mathbf{V}$ and $\mathbf{B} \subseteq \mathbf{V}$ such that $\mathbf{B} \subseteq \operatorname{pa}_{\mathcal{G}}(A)$. Let $\mathcal{P}(\mathbf{V})$ be the set of probability distributions that factorize according to $\mathcal{G}$. Then*

$$\begin{aligned}
&p(A|\operatorname{pa}_{\mathcal{G}}(A) \setminus \mathbf{B}) \prod_{V \in \mathbf{V} \setminus A} p(V|\operatorname{pa}_{\mathcal{G}}(V)) \\
&\quad = \operatorname*{arg\,min}_{\tilde{p} \in \mathcal{P}(\mathbf{V})} D_{KL}(p||\tilde{p}) \ \textit{s.t.}\ A \perp\!\!\!\perp \mathbf{B}|\operatorname{pa}_{\mathcal{G}}(A) \setminus \mathbf{B}
\end{aligned}$$

*Proof:* We adapt the argument from Thm. 11. We can express the KL-divergence of $\tilde{p}(A, \operatorname{pa}_{\mathcal{G}}(A))$ from $p(A, \operatorname{pa}_{\mathcal{G}})$ as proportional to:

$$\begin{aligned}
&\log \frac{p(A, \mathbf{B}|\operatorname{pa}_{\mathcal{G}}(A) \setminus \mathbf{B}) p(\operatorname{pa}_{\mathcal{G}}(A) \setminus \mathbf{B})}{\tilde{p}(A, \mathbf{B}|\operatorname{pa}_{\mathcal{G}}(A) \setminus \mathbf{B}) \tilde{p}(\operatorname{pa}_{\mathcal{G}}(A) \setminus \mathbf{B})} \\
&= \log \frac{\left[\frac{p(A|\operatorname{pa}_{\mathcal{G}}(A)\setminus\mathbf{B}) \frac{OR_{num}}{OR_{den}} p(\mathbf{B}|\operatorname{pa}_{\mathcal{G}}(A)\setminus\mathbf{B})}{\sum_{A,\mathbf{B}} p(A|\operatorname{pa}_{\mathcal{G}}(A)\setminus\mathbf{B}) \frac{OR_{num}}{OR_{den}} p(\mathbf{B}|\operatorname{pa}_{\mathcal{G}}(A)\setminus\mathbf{B})}\right]}{\left[\frac{\tilde{p}(A|\operatorname{pa}_{\mathcal{G}}(A)\setminus\mathbf{B}) \frac{\widetilde{OR}_{num}}{\widetilde{OR}_{den}} \tilde{p}(\mathbf{B}|\operatorname{pa}_{\mathcal{G}}(A)\setminus\mathbf{B})}{\sum_{A,\mathbf{B}} \tilde{p}(A|\operatorname{pa}_{\mathcal{G}}(A)\setminus\mathbf{B}) \frac{\widetilde{OR}_{num}}{\widetilde{OR}_{den}} \tilde{p}(\mathbf{B}|\operatorname{pa}_{\mathcal{G}}(A)\setminus\mathbf{B})}\right]} \\
&+ \log \frac{p(\operatorname{pa}_{\mathcal{G}}(A) \setminus \mathbf{B})}{\tilde{p}(\operatorname{pa}_{\mathcal{G}}(A) \setminus \mathbf{B})} \\
&= \log \frac{p(A|\operatorname{pa}_{\mathcal{G}}(A) \setminus \mathbf{B})}{\tilde{p}(A|\operatorname{pa}_{\mathcal{G}}(A) \setminus \mathbf{B})} + \log \frac{p(\mathbf{B}|\operatorname{pa}_{\mathcal{G}}(A) \setminus \mathbf{B})}{\tilde{p}(\mathbf{B}|\operatorname{pa}_{\mathcal{G}}(A) \setminus \mathbf{B})} \\
&+ \log \frac{OR_{num}}{OR_{den}} - \log \frac{\widetilde{OR}_{num}}{\widetilde{OR}_{den}} \\
&+ \log \sum_{A,\mathbf{B}} \tilde{p}(A|\operatorname{pa}_{\mathcal{G}}(A) \setminus \mathbf{B}) \frac{\widetilde{OR}_{num}}{\widetilde{OR}_{den}} \tilde{p}(\mathbf{B}|\operatorname{pa}_{\mathcal{G}}(A) \setminus \mathbf{B}) \\
&- \log \sum_{A,\mathbf{B}} p(A|\operatorname{pa}_{\mathcal{G}}(A) \setminus \mathbf{B}) \frac{OR_{num}}{OR_{den}} p(\mathbf{B}|\operatorname{pa}_{\mathcal{G}}(A) \setminus \mathbf{B}) \\
&+ \log \frac{p(\operatorname{pa}_{\mathcal{G}}(A) \setminus \mathbf{B})}{\tilde{p}(\operatorname{pa}_{\mathcal{G}}(A) \setminus \mathbf{B})}
\end{aligned}$$

where we apply the Chen factorization for $p(A, \mathrm{pa}_{\mathcal{G}}(A))$ and $\tilde{p}(A, \mathrm{pa}_{\mathcal{G}}(A))$ and

$$\begin{aligned} OR_{num} = {} & OR(A, \mathbf{B} | \mathrm{pa}_{\mathcal{G}}(A) \setminus \mathbf{B}) \\ & \times OR(A^0, \mathbf{B} | \mathrm{pa}_{\mathcal{G}}(A) \setminus \mathbf{B}) \\ & \times OR(A, \mathbf{B}^0 | \mathrm{pa}_{\mathcal{G}}(A) \setminus \mathbf{B}) \\ OR_{den} = {} & E[OR(A, \mathbf{B}^0 | \mathrm{pa}_{\mathcal{G}}(A) \setminus \mathbf{B}) | A^0, \mathrm{pa}_{\mathcal{G}}(A) \setminus \mathbf{B}] \\ & \times E[OR(A^0, \mathbf{B} | \mathrm{pa}_{\mathcal{G}}(A) \setminus \mathbf{B}) | \mathbf{B}^0, \mathrm{pa}_{\mathcal{G}}(A) \setminus \mathbf{B}] \end{aligned}$$

and analogously for the $\widetilde{OR}$'s.

Suppose we pick $\tilde{p}(A, \mathrm{pa}_{\mathcal{G}}(A)) = p(A | \mathrm{pa}_{\mathcal{G}}(A) \setminus \mathbf{B}) p(\mathbf{B} | \mathrm{pa}_{\mathcal{G}}(A) \setminus \mathbf{B}) p(\mathrm{pa}_{\mathcal{G}}(A) \setminus \mathbf{B})$. Then $A \perp\!\!\!\perp \mathbf{B} | \mathrm{pa}_{\mathcal{G}}(A) \setminus \mathbf{B}$ in $\tilde{p}$ and so $\frac{\widetilde{OR}_{num} / \widetilde{OR}_{den}}{\sum_{A,\mathbf{B}} \tilde{p}(A | \mathrm{pa}_{\mathcal{G}}(A) \setminus \mathbf{B}) \frac{\widetilde{OR}_{num}}{\widetilde{OR}_{den}} \tilde{p}(\mathbf{B} | \mathrm{pa}_{\mathcal{G}}(A) \setminus \mathbf{B})} = 1$. Thus, the previous expression simplifies to:

$$\begin{aligned} & \log \frac{OR_{num}}{OR_{den}} \\ & \quad - \log \sum_{A,\mathbf{B}} p(A | \mathrm{pa}_{\mathcal{G}}(A) \setminus \mathbf{B}) \frac{OR_{num}}{OR_{den}} p(\mathbf{B} | \mathrm{pa}_{\mathcal{G}}(A) \setminus \mathbf{B}) \end{aligned} \tag{F.2}$$

Suppose we instead picked some *other* $\tilde{p}(A, \mathrm{pa}_{\mathcal{G}}(A)) = \tilde{p}(A) \tilde{p}(\mathrm{pa}_{\mathcal{G}}(A))$ (i.e. one in which $A \perp\!\!\!\perp \mathbf{B} | \mathrm{pa}_{\mathcal{G}}(A) \setminus \mathbf{B}$). Then the above expression would have additional non-zero terms $\log \frac{p(A | \mathrm{pa}_{\mathcal{G}}(A) \setminus \mathbf{B})}{\tilde{p}(A | \mathrm{pa}_{\mathcal{G}}(A) \setminus \mathbf{B})} + \log \frac{p(\mathbf{B} | \mathrm{pa}_{\mathcal{G}}(A) \setminus \mathbf{B})}{\tilde{p}(\mathbf{B} | \mathrm{pa}_{\mathcal{G}}(A) \setminus \mathbf{B})} + \log \frac{p(\mathrm{pa}_{\mathcal{G}}(A) \setminus \mathbf{B})}{\tilde{p}(\mathrm{pa}_{\mathcal{G}}(A) \setminus \mathbf{B})}$. For this alternative $\tilde{p}$ to yield a lower KL divergence than that given by Eq. F.2, one of the terms in the above sum must be less than 0 (since the other terms in Eq. F.2 remain the same under the conditional independence of $A$ and $\mathbf{B}$). However, if $\log \frac{p(A | \mathrm{pa}_{\mathcal{G}}(A) \setminus \mathbf{B})}{\tilde{p}(A | \mathrm{pa}_{\mathcal{G}}(A) \setminus \mathbf{B})} < 0$ then the KL-divergence of $\tilde{p}(A | \mathrm{pa}_{\mathcal{G}}(A) \setminus \mathbf{B})$ from $p(A | \mathrm{pa}_{\mathcal{G}}(A) \setminus \mathbf{B})$ is negative, which violates Gibbs' inequality. The same holds for the distributions over $\mathrm{pa}_{\mathcal{G}}(A) \setminus \mathbf{B}$ and $\mathbf{B} | \mathrm{pa}_{\mathcal{G}}(A) \setminus \mathbf{B}$. We therefore can conclude that $\tilde{p}(A, \mathrm{pa}_{\mathcal{G}}(A)) = p(A | \mathrm{pa}_{\mathcal{G}}(A) \setminus \mathbf{B}) p(\mathbf{B} | \mathrm{pa}_{\mathcal{G}}(A) \setminus \mathbf{B}) p(\mathrm{pa}_{\mathcal{G}}(A) \setminus \mathbf{B})$ is the KL-closest distribution to $p(A, \mathrm{pa}_{\mathcal{G}}(A))$ such that $A \perp\!\!\!\perp \mathbf{B} | \mathrm{pa}_{\mathcal{G}}(A) \setminus \mathbf{B}$.

By the above argument and application of the conditioning argument in Thm. 11, the KL-closest distribution $\tilde{p}(A | \mathrm{pa}_{\mathcal{G}}(A))$ to $p(A | \mathrm{pa}_{\mathcal{G}}(A))$ satisfying the necessary independence constraint is $p(A|\,\mathrm{pa}_{\mathcal{G}}(A) \setminus \mathbf{B})$. By the local Markov property, the result is immediate since if chose $\tilde{p} = p$ for variables $\mathbf{V} \setminus \{A, \mathrm{pa}_{\mathcal{G}}(A)\}$.

□

**Theorem** 13 *Let* $\mathbf{V}$ *be a set of random variables with* $p(\mathbf{V})$ *corresponding to a DAG* $\mathcal{G}$*. Let* $\mathbf{A} \in \mathbf{V}$ *and for each* $A \in \mathbf{A}$ *define* $\mathrm{In}(A) \subseteq \mathrm{pa}_{\mathcal{G}}(A)$*, the set of parents of* $A$ *whose edges into* $A$ *we wish to remove. Let* $\mathcal{P}(\mathbf{V})$ *be the set of probability distributions that factorize according to* $\mathcal{G}$*. Then*

$$\begin{aligned}\prod_{A\in\mathbf{A}} p(A|\,\mathrm{pa}_{\mathcal{G}}(A) \setminus \mathrm{In}(A)) &\prod_{V\in\mathbf{V}\setminus\mathbf{A}} p(V|\,\mathrm{pa}_{\mathcal{G}}(V)) \\ &= \underset{\tilde{p}\in\mathcal{P}(\mathbf{V})}{\arg\min}\, D_{KL}(p||\tilde{p}) \\ &\quad s.t.\;\; A \perp\!\!\!\perp \mathrm{In}(A)|\,\mathrm{pa}_{\mathcal{G}}(A) \setminus \mathrm{In}(A) \;\; \forall A \in \mathbf{A}\end{aligned}$$

*Proof:* We prove the claim inductively. When $|\mathbf{A}| = 1$, the claim holds trivially by application of Thm 12.

Suppose $|\mathbf{A}| > 1$. Impose a reverse topological ordering $\prec$ on $\mathbf{V}$ (e.g. variables have higher indexes in the ordering than their parents). This ordering assumption is not necessary to prove the claim, however it helps simplify the argument.

Suppose that for some $\mathbf{A}' \subset \mathbf{A}$, where every $A \in \mathbf{A}'$ precedes every $A^* \in \mathbf{A} \setminus \mathbf{A}'$ in $\prec$, we know

$$\begin{aligned}\tilde{p}(\mathbf{V}) = &\prod_{A\in\mathbf{A}'} p(A|\,\mathrm{pa}_{\mathcal{G}}(A) \setminus \mathrm{In}(A)) \\ &\times \prod_{V\in\mathbf{V}\setminus\mathbf{A}'} p(V|\,\mathrm{pa}_{\mathcal{G}}(V))\end{aligned} \tag{F.3}$$

is the KL-closest distribution to $p(\mathbf{V})$ which satisfies $A \perp\!\!\!\perp \mathrm{In}(A)|\,\mathrm{pa}_{\mathcal{G}}(A) \setminus \mathrm{In}(A)$ for all $A \in \mathbf{A}'$. Then it suffices to show for some $A^* \in \mathbf{A} \setminus \mathbf{A}'$ that

$$\begin{aligned}\tilde{p}(\mathbf{V}) = &\prod_{A\in(\mathbf{A}'\cup A^*)} p(A|\,\mathrm{pa}_{\mathcal{G}}(A) \setminus \mathrm{In}(A)) \\ &\times \prod_{V\in\mathbf{V}\setminus(\mathbf{A}'\cup A^*)} p(V|\,\mathrm{pa}_{\mathcal{G}}(V))\end{aligned}$$

is the KL-closest distribution to $p(\mathbf{V})$ that satisfies $A \perp\!\!\!\perp \mathrm{In}(A)|\,\mathrm{pa}_{\mathcal{G}}(A) \setminus \mathrm{In}(A)$ for all $A \in \mathbf{A}' \cup A^*$.

We can factorize $p$ (and analogously $\tilde{p}$) by chain rule:

$$\begin{aligned} p(\mathbf{V}) = p(A^{*},\, \text{In}(A^{*})|\, \text{pa}_{\mathcal{G}}(A^{*}) \setminus \text{In}(A^{*})) \\ \times\, p(\text{pa}_{\mathcal{G}}(A^{*}) \setminus \text{In}(A^{*})) \\ \times\, p(\mathbf{V} \setminus (A^{*} \cup \text{pa}_{\mathcal{G}}(A^{*}))) \end{aligned}$$

By application of Cor. 6, we can re-write the first term as $\frac{X_p}{Y_p}$, where:

$$\begin{aligned} X_p =& p(A^{*}|\, \text{pa}_{\mathcal{G}}(A^{*}) \setminus \text{In}(A^{*})) \\ &\times \frac{OR_{num}}{OR_{den}} \\ &\times p(\text{In}(A^{*})|\, \text{pa}_{\mathcal{G}}(A^{*}) \setminus \text{In}(A^{*})) \end{aligned}$$

and $Y_p = \sum_{A^{*},\text{In}(A^{*})} X_p$, and analogously for $X_{\tilde{p}}$ and $Y_{\tilde{p}}$. Similar to previous arguments, for notational simplicity, we use the shorthands $OR_{num} =$

$$\begin{aligned} &OR(A^{\star}, \text{In}(A^{\star})|\, \text{pa}_{\mathcal{G}}(A^{\star} \setminus \text{In}(A^{\star}), \mathbf{V} \setminus (A^{\star} \cup \text{pa}_{\mathcal{G}}(A^{\star}))) \\ \times &OR(A^{\star 0}, \text{In}(A^{\star})|\, \text{pa}_{\mathcal{G}}(A^{\star} \setminus \text{In}(A^{\star}), \mathbf{V} \setminus (A^{\star} \cup \text{pa}_{\mathcal{G}}(A^{\star}))) \\ \times &OR(A^{\star}, \text{In}(A^{\star})^{0}|\, \text{pa}_{\mathcal{G}}(A^{\star} \setminus \text{In}(A^{\star}), \mathbf{V} \setminus (A^{\star} \cup \text{pa}_{\mathcal{G}}(A^{\star}))) \end{aligned}$$

and $OR_{den} =$

$$\begin{aligned} &E\Big[OR(A^{\star}, \text{In}(A^{\star})^{0})|A^{\star 0}, \\ &\qquad\qquad \text{pa}_{\mathcal{G}}(A^{\star} \setminus \text{In}(A^{\star}), \mathbf{V} \setminus (A^{\star} \cup \text{pa}_{\mathcal{G}}(A^{\star}))\Big] \\ \times &E\Big[OR(A^{\star 0}, \text{In}(A^{\star}))|\, \text{In}(A^{\star})^{0}, \\ &\qquad\qquad \text{pa}_{\mathcal{G}}(A^{\star} \setminus \text{In}(A^{\star}), \mathbf{V} \setminus (A^{\star} \cup \text{pa}_{\mathcal{G}}(A^{\star}))\Big] \end{aligned}$$

(analogously $\widetilde{OR}_{num}$ and $\widetilde{OR}_{den}$).

As before, we can express the KL-divergence from $p$ to $\tilde{p}$ as proportional to:

$$\begin{aligned}
&\log \frac{p(A^*, \operatorname{pa}_{\mathcal{G}}(A^*))p(\mathbf{V} \setminus (A^* \cup \operatorname{pa}_{\mathcal{G}}(A^*)))}{\tilde{p}(A^*, \operatorname{pa}_{\mathcal{G}}(A^*))\tilde{p}(\mathbf{V} \setminus (A^* \cup \operatorname{pa}_{\mathcal{G}}(A^*)))} \\
&= \Bigg[ \log \frac{p(A^* | \operatorname{pa}_{\mathcal{G}}(A^*) \setminus \operatorname{In}(A^*))}{\tilde{p}(A^* | \operatorname{pa}_{\mathcal{G}}(A^*) \setminus \operatorname{In}(A^*))} \\
&+ \log \frac{p(\operatorname{In}(A^*) | \operatorname{pa}_{\mathcal{G}}(A^*) \setminus \operatorname{In}(A^*))}{\tilde{p}(\operatorname{In}(A^*) | \operatorname{pa}_{\mathcal{G}}(A^*) \setminus \operatorname{In}(A^*))} \\
&+ \log \frac{OR_{num}}{OR_{den}} - \log \sum_{A^*, \operatorname{In}(A^*)} \Bigg[ p(A^* | \operatorname{pa}_{\mathcal{G}}(A^*) \setminus \operatorname{In}(A^*)) \\
&\times \frac{OR_{num}}{OR_{den}} p(\operatorname{In}(A^*) | \operatorname{pa}_{\mathcal{G}}(A^*) \setminus \operatorname{In}(A^*)) \Bigg] \\
&- \frac{\widetilde{OR}_{num}}{\widetilde{OR}_{den}} + \log \sum_{A^*, \operatorname{In}(A^*)} \Bigg[ \tilde{p}(A^* | \operatorname{pa}_{\mathcal{G}}(A^*) \setminus \operatorname{In}(A^*)) \\
&\times \frac{\widetilde{OR}_{num}}{\widetilde{OR}_{den}} \tilde{p}(\operatorname{In}(A^*) | \operatorname{pa}_{\mathcal{G}}(A^*) \setminus \operatorname{In}(A^*)) \Bigg] \\
&+ \log \frac{p(\operatorname{pa}_{\mathcal{G}}(A^*) \setminus \operatorname{In}(A^*))}{\tilde{p}(\operatorname{pa}_{\mathcal{G}}(A^*) \setminus \operatorname{In}(A^*))} \\
&+ \log \frac{p(\mathbf{V} \setminus (A^* \cup \operatorname{pa}_{\mathcal{G}}(A^*)))}{\tilde{p}(\mathbf{V} \setminus (A^* \cup \operatorname{pa}_{\mathcal{G}}(A^*)))} \Bigg]
\end{aligned}$$

Suppose we let $\tilde{p}(A^* | \operatorname{pa}_{\mathcal{G}}(A^*)) = p(A^* | \operatorname{pa}_{\mathcal{G}}(A^*) \setminus \operatorname{In}(A^*))$ and $\tilde{p}(\operatorname{In}(A^*) | \operatorname{pa}_{\mathcal{G}}(A^* \setminus \operatorname{In}(A^*))) = p(\operatorname{In}(A^*) | \operatorname{pa}_{\mathcal{G}}(A^* \setminus \operatorname{In}(A^*)))$. Then, similar to the previous theorems, we induce conditional independence between $A^\star$ and $\operatorname{In}(A^\star)$ given $A^\star$'s other parents $\operatorname{pa}_{\mathcal{G}}(A^\star) \setminus \operatorname{In}(A^\star)$. In turn, the above expression simplifies to the following:

$$\begin{aligned}
&= \log \frac{OR_{num}}{OR_{den}} - \log \sum_{A^*, \operatorname{In}(A^*)} \Bigg[ p(A^* | \operatorname{pa}_{\mathcal{G}}(A^*) \setminus \operatorname{In}(A^*)) \\
&\times \frac{OR_{num}}{OR_{den}} p(\operatorname{In}(A^*) | \operatorname{pa}_{\mathcal{G}}(A^*) \setminus \operatorname{In}(A^*)) \Bigg] \\
&+ \log \frac{p(\operatorname{pa}_{\mathcal{G}}(A^\star) \setminus \operatorname{In}(A^\star)}{\tilde{p}(\operatorname{pa}_{\mathcal{G}}(A^\star) \setminus \operatorname{In}(A^\star)} \\
&+ \log \frac{p(\mathbf{V} \setminus (A^* \cup \operatorname{pa}_{\mathcal{G}}(A^*)))}{\tilde{p}(\mathbf{V} \setminus (A^* \cup \operatorname{pa}_{\mathcal{G}}(A^*)))}
\end{aligned}$$

Under the assumption of a topological ordering $\prec$, choosing this choice of $\tilde{p}$ does not affect whether the constraint $A \perp\!\!\!\perp \operatorname{In}(A) | \operatorname{pa}_{\mathcal{G}}(A)$ for $A \in \mathbf{A}' \cup A^*$ holds. This is because of the assumption made in Eq. F.3. As a consequence, the ratio of terms with respect to $\text{pa}_{\mathcal{G}}(A^*) \setminus \text{In}(A^*)$ cancels in equality 1 above, leaving us with:

$$
\begin{aligned}
&= \log \frac{OR_{num}}{OR_{den}} - \log \sum_{A^*, \text{In}(A^*)} \Big[ p(A^* | \, \text{pa}_{\mathcal{G}}(A^*) \setminus \text{In}(A^*)) \\
&\times \frac{OR_{num}}{OR_{den}} p(\text{In}(A^*) | \, \text{pa}_{\mathcal{G}}(A^*) \setminus \text{In}(A^*)) \Big] \\
&+ \log \frac{p(\mathbf{V} \setminus (A^* \cup \text{pa}_{\mathcal{G}}(A^*)))}{\tilde{p}(\mathbf{V} \setminus (A^* \cup \text{pa}_{\mathcal{G}}(A^*)))}
\end{aligned}
$$

Now suppose we wish find a $p^*$ that yields a lower KL-divergence, corresponding to decrease the quantity in the above expression by changing $\tilde{p}$ for one or more of the terms. By application of the argument in Thm. 12, changing $\tilde{p}(A^* | \, \text{pa}_{\mathcal{G}}(A^*))$ to a function other than $p(A^* | \, \text{pa}_{\mathcal{G}}(A^*) \setminus \text{In}(A))$ would necessarily violate Gibbs' inequality. So, we must consider changing $\tilde{p}(V | \, \text{pa}_{\mathcal{G}}(V))$ for some $V \in \mathbf{V} \setminus (A^* \cup \text{pa}_{\mathcal{G}}(A^*))$.

If $V \in \text{nd}_{\mathcal{G}}(A^*)$ then, by the local Markov property of DAGs, $A^* \perp\!\!\!\perp V | \, \text{pa}_{\mathcal{G}}(A^*) \setminus \text{In}(A^*)$ and so choosing $\tilde{p}(V | \, \text{pa}_{\mathcal{G}}(V)) = p(V | \, \text{pa}_{\mathcal{G}}(V))$ will maintain the necessary independence constraints and ensure that the term for $V$ has 0 contribution to the KL quantity above. That is $\log p(V | \, \text{pa}_{\mathcal{G}}(V)) - \log \tilde{p}(V | \, \text{pa}_{\mathcal{G}}(V)) = 0$. Changing $\tilde{p}(V | \, \text{pa}_{\mathcal{G}}(V))$ will therefore not move $\tilde{p}$ closer to $p$.

If, on the other hand, $V \in \text{de}_{\mathcal{G}}(A^*)$, then $\text{In}(V)$ is either empty or non-empty. If $\text{In}(V) = \emptyset$ then by the same argument as for $V \in \text{nd}_{\mathcal{G}}(A^*)$, choosing $\tilde{p}(V | \, \text{pa}_{\mathcal{G}}(V)) = p(V | \, \text{pa}_{\mathcal{G}}(V))$ will ensure that the necessary constraints hold and that $V$'s contribution the KL-divergence expression will be 0. If $\text{In}(V) \neq \emptyset$, then $\tilde{p}(V | \, \text{pa}_{\mathcal{G}}(V))$ was already set to be $p(V | \, \text{pa}_{\mathcal{G}}(V) \setminus \text{In}(V))$ by the assumption in Eq. F.3. By the argument in Thm. 12, changing this setting of $\tilde{p}$ would violate Gibbs' inequality.

By the above argument, as well as the argument given in Thm. 11 that states that applying conditioning to two distributions doesn't affect their KL-divergence, we have shown that

$$
\prod_{A \in (\mathbf{A}' \cup A^*)} p(A | \, \text{pa}_{\mathcal{G}}(A) \setminus \text{In}(A)) \prod_{V \in \mathbf{V} \setminus (\mathbf{A}' \cup A^*)} p(V | \, \text{pa}_{\mathcal{G}}(V))
$$

is the KL-closest distribution to $p$ which satisfies $A \perp\!\!\!\perp \mathrm{In}(A) | \mathrm{pa}_{\mathcal{G}}(A) \setminus \mathrm{In}(A)$ for all $A \in \mathbf{A}' \cup A^*$. By induction, the claim of the theorem for $\mathbf{A}$ follows immediately. □

**Theorem 14** *Let $\mathbf{V}$ be a set of random variables with $p(\mathbf{V})$ corresponding to a DAG $\mathcal{G}$. Let $\mathbf{A} \subseteq \mathbf{V}$ and assume that for some $\mathbf{a}$ we have $p(\mathbf{A} = \mathbf{a}) > 0$. Let $\mathcal{P}(\mathbf{V})$ be the set of probability distributions that factorize according to $\mathcal{G}$. Then*

$$\prod_{V \in \mathbf{V} \setminus \mathbf{A}} p(V | \mathrm{pa}_{\mathcal{G}}(V))|_{\mathbf{A}=\mathbf{a}} = \arg\min_{\tilde{p} \in \mathcal{P}(\mathbf{V})} D_{KL}(p || \tilde{p})$$
$$s.t.\ \tilde{p}(A_i | \mathrm{nd}_{\mathcal{G}}(A_i)) = I(A_i = a_i)\, \forall i \in [|\mathbf{A}|]$$

*where* $[|\mathbf{A}|] = \{1, \ldots, |\mathbf{A}|\}$.

*Proof:* This is a simple consequence of Thm. 13. For each $A \in \mathbf{A}$, if we let $\mathrm{In}(A) = \mathrm{pa}_{\mathcal{G}}$, then by the local Markov property, we have

$$\prod_{A \in \mathbf{A}} p(A) \prod_{V \in \mathbf{V} \setminus \mathbf{A}} p(V | \mathrm{pa}_{\mathcal{G}}(V))$$

is the KL-closest distribution to $p(\mathbf{V})$ that satisfies $A \perp\!\!\!\perp \mathrm{nd}_{\mathcal{G}}(A)$ for all $A \in \mathbf{A}$. Now, by previous arguments, replacing each $p(A)$ with $I(A_i = a_i)$ maintains the KL-closeness of $\tilde{p}$ since $\tilde{p}(\mathbf{V} \setminus \mathbf{A}) = p(\mathbf{V} \setminus \mathbf{A})$ and now the required constraint holds. Since we are replacing each $p(A)$ with an indicator, this is equivalent to just evaluating $p(\mathbf{V} \setminus \mathbf{A})$ with $\mathbf{A} = \mathbf{a}$:

$$\prod_{V \in \mathbf{V} \setminus \mathbf{A}} p(V | \mathrm{pa}_{\mathcal{G}}(V))|_{\mathbf{A}=\mathbf{a}}$$

□

We extend the above result to the case of edge interventions. To simplify our argument, we formulate this theorem in terms of extended graphs which are inspired by Robins, Richardson, and Spirtes (2009) and requires the following additional background notation (Malinsky, Shpitser, and Richardson, 2019):

For a set of variables $\mathbf{A}$ and a set of edges $\alpha$ out of $\mathbf{A}$, define for each $A_i \in \mathbf{A}$, the synthetic nodes $A_i^{ch} = \{A_i^j | V_j \in \mathrm{ch}_{\mathcal{G}}(A_i)\}$. That is, for each $V_j \in \mathrm{ch}_{\mathcal{G}}(A_i)$, we create a synthetic node $A_i^j$. Let $\mathbf{A}^{ch} = \bigcup_{A_i \in \mathbf{A}}$.

Define the extended graph of $\mathcal{G}(\mathbf{V})$, denoted $\mathcal{G}^e(\mathbf{V} \cup \mathbf{A}^{ch})$, as the graph obtained by adding the synthetic $A_i^j$'s to $\mathcal{G}$ with edges $A_i \to A_i^j \to V_j$ if and only if $A_i \to V_j$ appears in $\mathcal{G}$. The relationship for each edge of type $A_i \to A_i^j$ as assumed to be deterministic. Following Malinsky, Shpitser, and Richardson (2019), $\mathcal{G}^e(\mathbf{V} \cup \mathbf{A}^{ch})$ is a valid DAG under the structural equation model assumption.

**Theorem 15** *Let $\mathbf{V}$ be a set of random variables with $p(\mathbf{V})$ corresponding to a DAG $\mathcal{G}$. Let $\alpha$ be a set of edges in $\mathcal{G}$ and let $\mathbf{A}_\alpha = \{A|(AB)_\to \in \alpha\} \subseteq \mathbf{V}$. For the corresponding $\mathbf{A}^{\text{ch}}$ and $\mathcal{G}^e(\mathbf{V} \cup \mathbf{A}^{\text{ch}})$, if we let $\mathcal{P}^e(\mathbf{V})$ be the set of probability distributions that factorize according to $\mathcal{G}^e$ and assume for some $\mathbf{a}^{\text{ch}}$, $p(\mathbf{A}^{\text{ch}} = \mathbf{a}^{\text{ch}}) > 0$ then,*

$$\prod_{V \in \mathbf{V}} p^e(V|\operatorname{pa}_{\mathcal{G}^e}(V)) = \arg\min_{\tilde{p}^e \in \mathcal{P}^e(\mathbf{V})} D_{KL}(p^e || \tilde{p}^e) \ s.t.$$
$$\tilde{p}^e(A_i|\operatorname{nd}_{\mathcal{G}^e}(A_i)) = I(A_i = a_i) \ for \ i = \{1, \ldots, |\mathbf{A}^{\text{ch}}|\}$$

*Proof:* This result follows directly from Thm. 14. By re-expressing $\mathcal{G}$ as $\mathcal{G}^e$, the intervention is no longer in terms of a set of edges $\alpha$ but rather a set of nodes $\mathbf{A}^{ch}$. We can simply apply the result of Thm. 14 where $\mathbf{A}^{ch}$ corresponds to the set of nodes $\mathbf{A}$ for which we are inducing independence with their non-descendants. □

## F.3 Experimental Setup

The models for $C, A$, and $Y$ are parametrized by $\tau_C$, $\tau_A = [\tau_{A_0}, \tau_{A_C}, \tau_{A_{C_\mathcal{N}}}]$, and $\tau_Y = [\tau_{Y_0}, \tau_{Y_A}, \tau_{Y_C}, \tau_{Y_{A_\mathcal{N}}}, \tau_{Y_{C_\mathcal{N}}}]$, specified in Table F-I. $C$ is a 3-dimensional vector, with each component $C^l$ drawn from a Bernoulli distribution with probability $\tau_{C^l}$; $A$

and $Y$ are generated using the following parametric models:

$$\begin{aligned}
&p(A_i = 1 | C_i, \{C_j | j \in \mathcal{N}_i\}; \tau_A) \\
&\qquad = expit\Big(\tau_{A_0} + \tau_{A_C} \cdot C_i + \sum_{j \in \mathcal{N}_i} \tau_{A_{C_{\mathcal{N}}}} \cdot C_j\Big) \\
&p(Y_i = 1 | C_i, A_i, \{C_j, A_j | j \in \mathcal{N}_i\}; \tau_Y) \\
&\qquad = expit\Big(\tau_{Y_0} + \tau_{Y_A} A_i + \tau_{Y_C} \cdot C_i \\
&\qquad\qquad + \sum_{j \in \mathcal{N}_i} \big(\tau_{Y_{C_{\mathcal{N}}}} \cdot C_j + \tau_{Y_{A_{\mathcal{N}}}} \cdot A_j\big)\Big)
\end{aligned}$$

| Parameter | Value |
|---|---|
| $\tau_C$ | [.7, .3, .5] |
| $\tau_A$ | [1, 3, .15, .2, .1, .15, .15] |
| $\tau_Y$ | [2.5, 1.2, -1, 1.2, .2, -.13, -1, -.2, -.3] |

Table F-I. Parameters for data generation in simulation studies

For the first experiment, we use the Erdős-Rényi (with attachment probability $p = .05$), Barabasi-Albert (with a preferential attachment edge count of 4), and Watts-Strogatz (with nearest neighbor attachment of 4 and an edge re-wiring probability of .25) network generators.

## Estimation Details

For homogeneous connections, estimating the post-intervention value of $Y_i$ is done by simply adding the connecting unit to $\mathcal{N}_i$ for the sake of forming covariate vectors on which we perform inference.

For known policy interventions, we consider adding a weight to the terms associated with the added neighbor. This corresponds, for instance, to joining one unit gaining an addition connection on a social media service and also algorithmically promoting the content of new neighbor unit. To estimate $Y_i$ here, we simply multiply the new neighbor's variables by the known weight and form covariate vectors as in the homogeneous case. In our simulations we use a weight of 1.2.

Finally, for unknown policy interventions, we repeat the process for known policies where a weight is added to the adjoined neighbors. Here, however, we choose the weight by maximizing a function $g(Y_i, Y_j) = min(\frac{Y_i+Y_j}{2}, .3)$. This corresponds to ensuring we satisfy a 'worst-case' scenario for the outcomes of the newly joined neighbors. We chose .3 as the floor for this function since the mean $Y$ in our data-generating process was .395 and we wanted to simulate not making one unit better off at the expense of making the other substantially worse off. We chose optimal parameters using standard optimization software (Jones, Oliphant, and Peterson, 2014).

Stochastic severance interventions are estimated analogously to homogeneous connections. We remove the terms relating to the severed connection from the feature vector for predicting $A_i$ and $Y_i$ and perform inference using our logistic regression models. Severance interventions performed with interventional values for the severed neighbor's $C$ and $A$ values are estimated by simply replacing the variables in the Monte Carlo sampling procedure with the interventional values according to the g-formula (Robins, 1986). For our simulations we chose the cross-unit interventional values for $C_j$ and $A_j$ to be 0 and 1 respectively. In either case, when a unit has no pre-intervention neighbors, the estimate of their outcome is the same in both the pre- and post-intervention worlds.

## Extended Results

| Network Size | Bias CI |
|---|---|
| 4 | (-.0082, .0051) |
| 8 | (-.0088, .0011) |
| 16 | (-.0057, .0006) |
| 32 | (-.0009, .0040) |
| 64 | (-.0009, .0016) |

Table F-II. 95% confidence intervals for the bias of estimates of stochastic severance task with varied network sizes.

| Attachment Prob. | Bias CI |
|---|---|
| .01 | (-.0008, .0017) |
| .05 | (-.0018, .0026) |
| .10 | (-.0043, .0003) |
| .15 | (-.0014, .0016) |
| .20 | (-.0011, .0004) |

Table F-III. 95% confidence intervals for the bias of estimates of stochastic severance task with varied Erdős-Rényi attachment probabilities.

| Sample Size | Bias CI |
|---|---|
| 10 | (-.0278, .0178) |
| 100 | (-.0075, .0057) |
| 1,000 | (-.0007, .0028) |
| 10,000 | (-.0001, .0006) |

Table F-IV. 95% confidence intervals for the bias of estimates of stochastic severance task with varied sample sizes.

# Bibliography

Snow, John (1849). *On the mode of communication of cholera.* John Churchill.

Robins, James M, Miguel Angel Hernan, and Babette Brumback (2000). *Marginal structural models and causal inference in epidemiology.*

Robins, James M et al. (1992). "G-estimation of the effect of prophylaxis therapy for Pneumocystis carinii pneumonia on the survival of AIDS patients." In: *Epidemiology*, pp. 319–336.

Papadogeorgou, Georgia, Christine Choirat, and Corwin M Zigler (2019). "Adjusting for unmeasured spatial confounding with distance adjusted propensity score matching." In: *Biostatistics* 20.2, pp. 256–272.

Nabi, Razieh et al. (2022). "Causal inference in the presence of interference in sponsored search advertising." In: *Frontiers in Big Data*, p. 54.

Pearl, Judea (2009). *Causality: Models, Reasoning, and Inference.* 2nd ed. Cambridge University Press.

Wright, Sewall (1934). "The method of path coefficients." In: *The annals of mathematical statistics* 5.3, pp. 161–215.

Fisher, Ronald Aylmer (1935). "Design of experiments." In: *British Medical Journal* 1.3923, p. 554.

Rubin, Donald B (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies." In: *Journal of educational Psychology* 66.5, p. 688.

Robins, James (1986). "A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect." In: *Mathematical modelling* 7.9-12, pp. 1393–1512.

Shpitser, Ilya and Judea Pearl (2006). "Identification of joint interventional distributions in recursive semi-Markovian causal models." In: *Proceedings of the National Conference on Artificial Intelligence.* Vol. 21. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, p. 1219.

Shpitser, Ilya (2013). "Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding." In: *Cognitive science* 37.6, pp. 1011–1035.

Bareinboim, Elias and Judea Pearl (2016). "Causal inference and the data-fusion problem." In: *Proceedings of the National Academy of Sciences* 113.27, pp. 7345–7352.

Pearl, Judea and Elias Bareinboim (2011). “Transportability of causal and statistical relations: A formal approach.” In: *Twenty-fifth AAAI conference on artificial intelligence*.

Bhattacharya, Rohit, Razieh Nabi, and Ilya Shpitser (2020). “Semiparametric inference for causal effects in graphical models with hidden variables.” In: *arXiv preprint arXiv:2003.12659*.

Rotnitzky, Andrea and Ezequiel Smucler (2019). “Efficient adjustment sets for population average treatment effect estimation in non-parametric causal graphical models.” In: *arXiv preprint arXiv:1912.00306*.

Chakraborty, Bibhas and EE Moodie (2013). *Statistical methods for dynamic treatment regimes*. Springer.

Bertsekas, Dimitri P. and John Tsitsiklis (1996). *Neuro-Dynamic Programming*. Athena Publishing.

Shalizi, Cosma Rohilla and Andrew C Thomas (2011). “Homophily and contagion are generically confounded in observational social network studies.” In: *Sociological methods & research* 40.2, pp. 211–239.

Mnih, Volodymyr et al. (2015). “Human-level control through deep reinforcement learning.” In: *Nature* 518.7540, p. 529.

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le (2014). “Sequence to sequence learning with neural networks.” In: *Advances in neural information processing systems*, pp. 3104–3112.

Davis, Gary A and Sylvia B Rimm (1989). *Education of the gifted and talented*. Prentice-Hall, Inc.

Hodges, Jaret et al. (2018). “A meta-analysis of gifted and talented identification practices.” In: *Gifted Child Quarterly* 62.2, pp. 147–174.

Cox, David R (1958). *Planning of experiments*. Vol. 20. Wiley New York.

Hudgens, Michael G and M Elizabeth Halloran (2008). “Toward causal inference with interference.” In: *Journal of the American Statistical Association* 103.482, pp. 832–842.

Ogburn, Elizabeth L, Tyler J VanderWeele, et al. (2014). “Causal diagrams for interference.” In: *Statistical science* 29.4, pp. 559–578.

Ogburn, Elizabeth L, Ilya Shpitser, and Youjin Lee (2018). “Causal inference, social networks, and chain graphs.” In: *arXiv preprint arXiv:1812.04990*.

Shpitser, Ilya (2015). “Segregated graphs and marginals of chain graph models.” In: *Advances in Neural Information Processing Systems*, pp. 1720–1728.

Pearl, Judea (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan and Kaufmann, San Mateo.

Richardson, Thomas S and James M Robins (2013). “Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality.”

In: *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper* 128.30, p. 2013.

Hiscock, Rosemary et al. (2012). “Socioeconomic status and smoking: a review.” In: *Annals of the New York Academy of Sciences* 1248.1, pp. 107–123.

Cohen, Aaron J and CA Pope 3rd (1995). “Lung cancer and air pollution.” In: *Environmental health perspectives* 103.suppl 8, pp. 219–224.

Verma, Thomas and Judea Pearl (1990). *Equivalence and synthesis of causal models.* UCLA, Computer Science Department.

Richardson, Thomas S et al. (2017). “Nested Markov properties for acyclic directed mixed graphs.” In: *arXiv preprint arXiv:1701.06686*.

Tian, Jin and Judea Pearl (2002). “A general identification condition for causal effects.” In: *Proceedings of the National Conference on Artificial Intelligence*, pp. 567–573.

Shpitser, Ilya and Eli Sherman (2018). “Identification of personalized effects associated with causal pathways.” In: *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence.*

Avin, Chen, Ilya Shpitser, and Judea Pearl (2005). “Identifiability of Path-Specific Effects.” In: *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI-05)*. Vol. 19. Morgan Kaufmann, San Francisco, pp. 357–363.

Pearl, Judea (2001). “Direct and Indirect Effects.” In: *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI-01)*. Morgan Kaufmann, San Francisco, pp. 411–420.

Robins, James M. and Sander Greenland (1992). “Identifiability and Exchangeability of Direct and Indirect Effects.” In: *Epidemiology* 3, pp. 143–155.

Nabi, Razieh and Ilya Shpitser (2018). “Estimation of Personalized Effects Associated With Causal Pathways.” In: *Proceedings of the Thirty Fourth Conference on Uncertainty in Artificial Intelligence (UAI).*

Hernan, Miguel A. et al. (2006). “Comparison of Dynamic Treatment Regimes via Inverse Probability Weighting.” In: *Basic and Clinical Pharmacology and Toxicology* 98 (3), pp. 237–242.

Miles, Caleb et al. (2017). “Quantifying an Adherence Path-Specific Effect of Antiretroviral Therapy in the Nigeria PEPFAR Program.” In: *Journal of the American Statistical Association.*

Shpitser, Ilya and Eric J. Tchetgen Tchetgen (2016). “Causal Inference with a Graphical Hierarchy of Interventions.” In: *Annals of Statistics* 44.6, pp. 2433–2466. arXiv: 1411.2127 [stat].

Tian, Jin (2008). “Identifying dynamic sequential plans.” In: *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence.*

Huang, Yimin and Marco Valtorta (2006). “Pearl’s Calculus of Interventions is Complete.” In: *Twenty Second Conference On Uncertainty in Artificial Intelligence.*

Tchetgen Tchetgen, Eric J. and Tyler J. VanderWeele (2012). "On Causal Inference in the Presence of Interference." In: *Statistical Methods in Medical Research* 21.1, pp. 55–75.

Peña, Jose M (2018). "Reasoning with alternative acyclic directed mixed graphs." In: *Behaviormetrika*, pp. 1–34.

— (2016). "Learning Acyclic Directed Mixed Graphs from Observations and Interventions." In: *Conference on Probabilistic Graphical Models*, pp. 392–402.

Maier, Marc, Katerina Marazopoulou, and David Jensen (2013). "Reasoning about independence in probabilistic models of relational data." In: *arXiv preprint arXiv:1302.4381*.

Arbour, David, Dan Garant, and David Jensen (2016). "Inferring network effects from observational data." In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 715–724.

Sherman, Eli and Ilya Shpitser (2018). "Identification and Estimation of Causal Effects from Dependent Data." In: *Advances in Neural Information Processing Systems*, pp. 9424–9435.

Lauritzen, Steffen L (1996). *Graphical models*. Vol. 17. Clarendon Press.

Lauritzen, Steffen L and Thomas S Richardson (2002). "Chain graph models and their causal interpretations." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.3, pp. 321–348.

Tchetgen, Eric J Tchetgen, Isabel Fulcher, and Ilya Shpitser (2017). "Auto-g-computation of causal effects on a network." In: *arXiv preprint arXiv:1709.01577*.

Richardson, Thomas (2003). "Markov properties for acyclic directed mixed graphs." In: *Scandinavian Journal of Statistics* 30.1, pp. 145–157.

Sobel, Michael E. (2006). "What do randomized studies of housing mobility demonstrate? Causal inference in the face of interference." In: *Journal of the American Statistical Association* 101.476, pp. 1398–1407.

Kenny, David A, Deborah A Kashy, and William L Cook (2020). *Dyadic data analysis*. Guilford Publications.

Drton, Mathias (2009). "Discrete chain graph models." In: *Bernoulli* 15.3, pp. 736–753.

Sherman, Eli, David Arbour, and Ilya Shpitser (2020). "General Identification of Dynamic Treatment Regimes Under Interference." In: *arXiv preprint arXiv:2004.01218*.

Viviano, Davide (2019). "Policy Targeting under Network Interference." In: *arXiv preprint arXiv:1906.10258*.

Blackwell, Matthew (2013). "A framework for dynamic causal inference in political science." In: *American Journal of Political Science* 57.2, pp. 504–520.

Sherman, Eli and Ilya Shpitser (2019). "Intervening on Network Ties." In: *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*. AUAI Press.

Erdős, Paul and Alfréd Rényi (1960). "On the evolution of random graphs." In: *Publ. Math. Inst. Hung. Acad. Sci* 5.1, pp. 17–60.

Watts, Duncan J and Steven H Strogatz (1998). "Collective dynamics of 'small-world' networks." In: *Nature* 393.6684, p. 440.

Albert, Réka and Albert László Barabási (2002). "Statistical mechanics of complex networks." In: *Reviews of modern physics* 74.1, p. 47.

Murphy, Kevin Patrick and Stuart Russell (2002). "Dynamic bayesian networks: representation, inference and learning." In.

Ogburn, Elizabeth L et al. (2017). "Causal inference for social network data." In: *arXiv preprint arXiv:1705.08527*.

Chetty, Raj, Nathaniel Hendren, and Lawrence F Katz (2016). "The effects of exposure to better neighborhoods on children: New evidence from the Moving to Opportunity experiment." In: *American Economic Review* 106.4, pp. 855–902.

Malinsky, Daniel (May 2018). "Intervening on structure." en. In: *Synthese* 195.5, pp. 2295–2312. DOI: 10.1007/s11229-017-1341-z. URL: http://link.springer.com/10.1007/s11229-017-1341-z (visited on 09/06/2018).

Koller, Daphne et al. (2007). *Introduction to statistical relational learning*. MIT press.

Friedman, Nir et al. (1999). "Learning probabilistic relational models." In: *IJCAI*. Vol. 99, pp. 1300–1309.

Neyman, Jersey (1923). "Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes." In: *Roczniki Nauk Rolniczych* 10, pp. 1–51.

Tian, Jin and Judea Pearl (2001). "Causal discovery from changes." In: *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., pp. 512–521.

Halpern, Joseph Y and Judea Pearl (2005). "Causes and explanations: A structural-model approach. Part I: Causes." In: *The British journal for the philosophy of science* 56.4, pp. 843–887.

Woodward, James (2001). "Causation and manipulability." In.

Korb, Kevin B et al. (2004). "Varieties of causal intervention." In: *Pacific Rim International Conference on Artificial Intelligence*. Springer, pp. 322–331.

Eberhardt, Frederick and Richard Scheines (2007). "Interventions and causal inference." In: *Philosophy of Science* 74.5, pp. 981–995.

Beck, Nathaniel and Jonathan N Katz (2011). "Modeling dynamics in time-series–cross-section political economy data." In: *Annual Review of Political Science* 14, pp. 331–352.

Robinson, Joshua W and Alexander J Hartemink (2009). "Non-stationary dynamic Bayesian networks." In: *Advances in neural information processing systems*, pp. 1369–1376.

Subbaswamy, Adarsh, Bryant Chen, and Suchi Saria (2019). "A universal hierarchy of shift-stable distributions and the tradeoff between stability and performance." In: *arXiv preprint arXiv:1905.11374*.

Witty, Sam et al. (2019). “Bayesian causal inference via probabilistic program synthesis.” In: *arXiv preprint arXiv:1910.14124*.

Galhotra, Sainyam et al. (2022). “HypeR: Hypothetical Reasoning With What-If and How-To Queries Using a Probabilistic Causal Approach.” In: *arXiv preprint arXiv:2203.14692*.

Spirtes, Peter et al. (2000). *Causation, prediction, and search.* MIT press.

Greenland, Sander and William D Finkle (1995). “A critical look at methods for handling missing covariates in epidemiologic regression analyses.” In: *American journal of epidemiology* 142.12, pp. 1255–1264.

Greenland, Sander (1996). “Basic methods for sensitivity analysis of biases.” In: *International journal of epidemiology* 25.6, pp. 1107–1116.

Haber, Noah A et al. (2022). “DAG With Omitted Objects Displayed (DAGWOOD): A framework for revealing causal assumptions in DAGs.” In: *Annals of Epidemiology* 68, pp. 64–71.

Angrist, Joshua D, Guido W Imbens, and Donald B Rubin (1996). “Identification of causal effects using instrumental variables.” In: *Journal of the American statistical Association* 91.434, pp. 444–455.

Shpitser, Ilya, Zach Wood-Doughty, and Eric J Tchetgen Tchetgen (2021). “The proximal id algorithm.” In: *arXiv preprint arXiv:2108.06818*.

Manski, Charles F (2003). *Partial identification of probability distributions.* Vol. 5. Springer.

Duarte, Guilherme et al. (2021). “An automated approach to causal inference in discrete settings.” In: *arXiv preprint arXiv:2109.13471*.

VanderWeele, Tyler J, Stijn Vansteelandt, and James M Robins (2014). “Effect decomposition in the presence of an exposure-induced mediator-outcome confounder.” In: *Epidemiology (Cambridge, Mass.)* 25.2, p. 300.

Brody, Debra J, Laura A Pratt, and Jeffery P Hughes (2018). “Prevalence of depression among adults aged 20 and over: United States, 2013-2016.” In.

Rush, A John (2007). “The varied clinical presentations of major depression disorder.” In: *The Journal of clinical psychiatry*.

Alboni, Paolo et al. (2008). “Is there an association between depression and cardiovascular mortality or sudden death?” In: *Journal of Cardiovascular Medicine* 9.4, pp. 356–362.

Richards, C Steven and Michael W O’Hara (2014). *The Oxford handbook of depression and comorbidity.* Oxford University Press.

De Choudhury, Munmun et al. (2013). “Predicting depression via social media.” In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 7. 1.

Coppersmith, Glen, Mark Dredze, and Craig Harman (2014). “Quantifying mental health signals in Twitter.” In: *Proceedings of the workshop on computational* *linguistics and clinical psychology: From linguistic signal to clinical reality*, pp. 51–60.

Elazar, Yanai and Yoav Goldberg (2018). "Adversarial removal of demographic attributes from text data." In: *arXiv preprint arXiv:1808.06640*.

De Choudhury, Munmun et al. (2016). "Discovering shifts to suicidal ideation from mental health content in social media." In: *Proceedings of the 2016 CHI conference on human factors in computing systems*, pp. 2098–2110.

Wood-Doughty, Zach et al. (2017). "How does twitter user behavior vary across demographic groups?" In: *Proceedings of the Second Workshop on NLP and Computational Social Science*, pp. 83–89.

Loveys, Kate et al. (2018). "Cross-cultural differences in language markers of depression online." In: *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pp. 78–87.

Amir, Silvio, Mark Dredze, and John W Ayers (2019). "Mental health surveillance over social media with digital cohorts." In: *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pp. 114–120.

Aguirre, Carlos, Keith Harrigian, and Mark Dredze (2021). "Gender and Racial Fairness in Depression Research using Social Media." In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Coppersmith, Glen et al. (2015). "CLPsych 2015 shared task: Depression and PTSD on Twitter." In: *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pp. 31–39.

Benton, Adrian, Margaret Mitchell, and Dirk Hovy (2017). "Multi-task learning for mental health using social media text." In: *arXiv preprint arXiv:1712.03538*.

Baumgartner, Jason et al. (2020). "The pushshift reddit dataset." In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 14, pp. 830–839.

Budge, Stephanie L, Jill L Adelson, and Kimberly AS Howard (2013). "Anxiety and depression in transgender individuals: the roles of transition status, loss, social support, and coping." In: *Journal of consulting and clinical psychology* 81.3, p. 545.

Wolohan, JT et al. (2018). "Detecting linguistic traces of depression in topic-restricted text: Attending to self-stigmatized depression with NLP." In: *Proceedings of the First International Workshop on Language Cognition and Computational Models*, pp. 11–21.

Larson, Brian N (2017). "Gender as a variable in natural-language processing: Ethical considerations." In.

Yates, Andrew, Arman Cohan, and Nazli Goharian (Sept. 2017). "Depression and Self-Harm Risk Assessment in Online Forums." In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 2968–2978. DOI: 10.18653/v1/D17-1322. URL: https://www.aclweb.org/anthology/D17-1322.

Pennebaker, James W, Roger J Booth, and Martha E Francis (2007). "LIWC2007: Linguistic inquiry and word count." In: *Austin, Texas: liwc. net.*

Jones, Karen Sparck (1972). "A statistical interpretation of term specificity and its application in retrieval." In: *Journal of documentation.*

Cohan, Arman et al. (2018). "SMHD: a Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions." In: *Proceedings of the 27th International Conference on Computational Linguistics.* Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 1485–1497. URL: https://www.aclweb.org/anthology/C18-1126.

Pedregosa, Fabian et al. (2011). "Scikit-learn: Machine learning in Python." In: *the Journal of machine Learning research* 12, pp. 2825–2830.

McCullough, Michael E and David B Larson (1999). "Religion and depression: A review of the literature." In: *Twin Research and Human Genetics* 2.2, pp. 126–136.

Hoffman, Bruce, Jacob Ware, and Ezra Shapiro (2020). "Assessing the threat of incel violence." In: *Studies in Conflict & Terrorism* 43.7, pp. 565–587.

Nabi, Razieh, Daniel Malinsky, and Ilya Shpitser (2019). "Learning optimal fair policies." In: *International Conference on Machine Learning.* PMLR, pp. 4674–4682.

Wang, Zijian and David Jurgens (2018). "It's going to be okay: Measuring Access to Support in Online Communities." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.* Brussels, Belgium: Association for Computational Linguistics, pp. 33–45. DOI: 10.18653/v1/D18-1004. URL: https://www.aclweb.org/anthology/D18-1004.

Bailey, Molly K. et al. (2019). "Characteristics of 30-Day All-Cause Hospital Readmissions, 2010–2016 (Statistical Brief #248)." In: *Rockville, MD: Agency for Healthcare Research and Quality.Retrieved from www.hcup-us.ahrq.gov/reports/statbriefs/sb248-Hospital-Readmissions-2010-2016.jsp.*

Commission, Medicare Payment Advisory (2007). "Report to the Congress: promoting greater efficiency in Medicare. June 2007." In.

Weiss, Audrey J., Anne Elixhauser, and Claudia Steiner (2006). "Readmissions to US hospitals by procedure, 2010: Statistical Brief #154." In: Healthcare Cost and Utilization Project (HCUP) Statistical Briefs. Agency for Health Care Policy and Research (US), Rockville, MD.

Kilic, Arman et al. (2017). "Development and validation of a score to predict the risk of readmission after adult cardiac operations." In: *The Annals of Thoracic Surgery* 103.1, pp. 66–73.

Henry, Katharine E. et al. (2015). "A targeted real-time early warning score (TREWScore) for septic shock." In: *Science translational medicine* 7.299. pmid:26246167, 299ra122.

Sherman, Eli et al. (2021). "Towards Understanding the Role of Gender in Deploying Social Media-Based Mental Health Surveillance Models." In: *Proceedings of the Sev-*

*enth Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pp. 217–223.

Wiens, Jenna, Eric Horvitz, and John Guttag (2012). “Patient risk stratification for hospital-associated c. diff as a time-series classification task.” In: *Advances in Neural Information Processing Systems* 25, pp. 467–475.

Lundervold, Alexander Selvikvåg and Arvid Lundervold (2019). “An overview of deep learning in medical imaging focusing on MRI.” In: *Zeitschrift für Medizinische Physik* 29.2, pp. 102–127.

Garavaglia, Susan and Asha Sharma (1998). “A smart guide to dummy variables: Four applications and a macro.” In: *Proceedings of the northeast SAS users group conference*. Vol. 43.

Breiman, Leo (2001). “Random forests.” In: *Machine Learning* 45.1, pp. 5–32.

Chen, Tianqi and Carlos Guestrin (2016). “Xgboost: A scalable tree boosting system.” In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794.

Boser, Bernhard E., Isabelle M. Guyon, and Vladimir N. Vapnik (1992). “A training algorithm for optimal margin classifiers.” In: *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 144–152.

Lundberg, Scott and Su-In Lee (2017). “A unified approach to interpreting model predictions.” In: *arXiv preprint arXiv:1705.07874*.

Hernández-Orallo, José, Peter Flach, and César Ferri Ramírez (2012). “A unified view of performance metrics: Translating threshold choice into expected classification loss.” In: *Journal of Machine Learning Research* 13, pp. 2813–2869.

Fanari, Zaher et al. (2017). “Predicting readmission risk following coronary artery bypass surgery at the time of admission.” In: *Cardiovascular Revascularization Medicine* 18.2, pp. 95–99.

Benuzillo, Jose et al. (2018). “Predicting readmission risk shortly after admission for CABG surgery.” In: *Journal of cardiac surgery* 33.4, pp. 163–170.

Tam, Derrick Y. et al. (2018). “A clinical risk scoring tool to predict readmission after cardiac surgery: an Ontario administrative and clinical population database study.” In: *Canadian Journal of Cardiology* 34.12, pp. 1655–1664.

Brown, Jeremiah R. et al. (2018). “Utility of biomarkers to improve prediction of readmission or mortality after cardiac surgery.” In: *The Annals of Thoracic Surgery* 106.5, pp. 1294–1301.

Liu, Yun et al. (2019). “How to read articles that use machine learning: users’ guides to the medical literature.” In: *Jama* 322.18, pp. 1806–1816.

Murphy, Barbara M. et al. (2008). “Living alone predicts 30-day hospital readmission after coronary artery bypass graft surgery.” In: *European Journal of Cardiovascular Prevention & Rehabilitation* 15.2, pp. 210–215.

Mazzeffi, Michael et al. (2020). "Racial disparity in cardiac surgery risk and outcome: report from a statewide quality initiative." In: *The Annals of Thoracic Surgery* 110.2, pp. 531–536.

Wadhera, Rishi K et al. (2018). "Association of the Hospital Readmissions Reduction Program with mortality among Medicare beneficiaries hospitalized for heart failure, acute myocardial infarction, and pneumonia." In: *Jama* 320.24, pp. 2542–2552.

Healy, Ryan et al. (2020). "Assessment of Patient Ambulation Profiles to Predict Hospital Readmission, Discharge Location, and Length of Stay in a Cardiac Surgery Progressive Care Unit." In: *JAMA network open* 3.3, e201074.

Besag, Julian (1975). "Statistical Analysis of Lattice Data." In: *The Statistician* 24.3, pp. 179–195.

Westreich, D. et al. (2012). "The parametric g-formula to estimate the effect of highly active antiretroviral therapy on incident AIDS or death." In: *Statistics in Medicine* 31.18.

Chen, Hua Yun (2007). "A Semiparametric Odds Ratio Model for Measuring Association." In: *biometrics* 63, pp. 413–421.

Robins, James, Thomas Richardson, and Peter Spirtes (2009). "On identification and inference for direct effects." In: *Epidemiology*.

Malinsky, Daniel, Ilya Shpitser, and Thomas Richardson (2019). "A Potential Outcomes Calculus for Identifying Conditional Path-Specific Effects." In: *Proceedings of Machine Learning Research*. Ed. by Kamalika Chaudhuri and Masashi Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, pp. 3080–3088. URL: http://proceedings.mlr.press/v89/malinsky19b.html.

Jones, Eric, Travis Oliphant, and Pearu Peterson (2014). "${$SciPy$}$: open source scientific tools for ${$Python$}$." In.

# Biographical Sketch

Eli Sherman is a doctoral candidate in the Computer Science Department at Johns Hopkins University. His research spans the theoretical-applied spectrum in causal inference and healthcare. On the theoretical end, he has made substantial progress on developing general theories for the identification of causal effects using graphical models with a focus on dependent data settings and policy evaluation. His applied contributions have centered on translational efforts at the intersection of applied machine learning and healthcare, including the development of clinical decision support models for predicting hospital readmission and hypokalemia. Mr. Sherman's broader interests center on the practical and societal implications of adopting machine learning and causal inference methodologies, including resource allocation, fairness, and the relationship between AI and the law.

Mr. Sherman's work has been recognized by awards at the AMIA Annual Symposium, a doctoral fellowship from the Johns Hopkins Mathematical Institute for Data Science, and a Google PhD Fellowship in machine learning. He has also been recognized for his teaching contributions, having been named a finalist for the Johns Hopkins University graduate student teaching award for his instruction in the course "Introduction to Causal Inference". During his doctoral studies, Mr. Sherman completed internships at IBM Research and Adobe Research. Eli received his BS in computer science, with honors, from the University of Michigan in 2017 and a Masters of Science in Engineering from Johns Hopkins University in 2019.