

**BEYOND CLASSICAL CAUSAL MODELS: PATH DEPENDENCE,  
ENTANGLED MISSINGNESS AND GENERALIZED COARSENING**

by  
Ranjani Srinivasan

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

Baltimore, Maryland  
January, 2023

© 2022 Ranjani Srinivasan  
All rights reserved

# Abstract

Classical causal models generally assume relatively simple settings like static observations, complete observability and independent and identically distributed (i.i.d.) data samples. For many systems of scientific interest, such assumptions are unrealistic. More recent work has explored models with complex properties including (time-invariant) temporal dynamics, data dependence, as well as missingness within the causal inference framework. Inspired by these advances, this dissertation goes beyond these classical causal inference models to explore the following complications that can arise in some causal systems– (i) *path dependence*, whereby systems exhibit state-specific causal relationships and a temporal evolution that could be counterfactually altered (ii) *entangled missingness*, where missingness occurs in data together with causal dependence and finally, (iii) *generalized coarsening*, where systems entail causal processes operating at multiple timescales, and estimands of interest lie at a timescale different from that in which data is observed. In particular, we use the language of *graphical causal models* and discuss an important component of the causal inference pipeline, namely *identification*, which links the counterfactual of interest to the observed data via a set of assumptions. In some cases, we also discuss *estimation*, which allows us to obtain identified parameters from finite samples of data. We illustrate the use of these novel models on observational data obtained from biomedical and clinical settings.

# Thesis Readers

Dr. Ilya Shpitser (Primary Advisor)  
John C. Malone Associate Professor  
Department of Computer Science  
Whiting School of Engineering, Johns Hopkins University

Dr. Archana Venkataraman  
John C. Malone Assistant Professor  
Department of Electrical and Computer Engineering  
Whiting School of Engineering, Johns Hopkins University

Dr. Trac Duy Tran  
Professor  
Department of Electrical and Computer Engineering  
Whiting School of Engineering, Johns Hopkins University

Dr. Daniel Malinsky  
Assistant Professor  
Department of Biostatistics  
Mailman School of Public Health, Columbia University

# Acknowledgements

First and foremost, I would like to thank my adviser Ilya, for taking the chance on me five years ago when I had little to no background in causal inference. He was incredibly patient with my learning curve, especially in the first few years. His enthusiasm for consistent (even if slow) progress has helped me learn a lot and be productive during graduate school, which can sometimes seem like an arduous and unending chapter in life. He has been very supportive of my endeavors outside of school too. I am immensely grateful to have had an adviser who treats his students as peers, and shares my opinion that graphs are cool!

Next, I would like to thank my committee members - Dr. Archana Venkataraman, Dr. Daniel Malinsky and Dr. Trac Tran, for consenting to guide me through this last phase of my Ph.D. I am confident that their thoughts on my work will improve its presentation substantially, and likely encourage novel ideas for future research.

I express my heartfelt gratitude to a slew of my collaborators at the Johns Hopkins University and the Johns Hopkins Hospital, without whom my journey as a researcher would have been impossible. Dr. Elizabeth (Betsy) Ogburn has collaborated with me through the better part of graduate school, and also graciously served on my GBO committee. She has always had plenty of constructive feedback that has improved my work, and I continue to be in awe of her scholarship. Dr. Glenn Whitman, Dr. Marc Sussman and Dr. Stefano Schena are remarkable physicians who literally save lives everyday and yet find the time to have regular conversations with me. Mondays will

remain etched in my mind as cardiac surgery meeting days forever, and I will terribly miss being able to interact so regularly with such accomplished (and yet unassuming) people. Diane Alejo, with JHH IT, has been at the center of this team, and I would like to thank her for being so generous with her time; she knows everything there is to know about EHR data. And a special thanks to Diane for arranging a cardiac valve replacement surgery for us to watch; it is a cherished memory. A quick thank you to Joe DiNatale, who has helped us a ton with data related questions as well. Also, I want to thank Dr. Narges Ahmidi for introducing me to this team, and setting up the entire database which was the foundation for two projects in my Ph.D. I also want to thank her, and Dr. Swaroop Vedula, for sharing the septoplasty dataset that we used in our study of path dependent systems. It was always my dream to do clinical translational work in graduate school, and I could not have asked for a better team to do it with. I hope we made a dent, and that we can improve patient care with analytical methods in the years to come!

Johns Hopkins was my cherished destination as an undergrad, and JHU ECE was everything I was told it would be. I enjoyed the classes I took and the company of friends and collaborators I made. The department events made me feel like I was part of a wonderful community. I would especially like to thank Belinda Blinkoff and Dana Walter-Shock for guiding me (and all other graduate students) through the milestones of graduate school, and making everything go by seamlessly. The student community at JHU is absolutely wonderful, and there is an eternal, friendly spirit in the air, one of cooperation to further everyone's potential, and I am so thankful I chose to spend a significant part of my 20's here.

I would like to thank my labmates (and ex-labmates) - Jaron, Rohit, Zach, Razieh, Eli, Numair, Noam and Dan - for creating the best, transparent and most collaborative group environment anyone could ask for. They are my one-stop-shop as friends,

colleagues, teachers and mentors. Sharing a space and white-boarding ideas with these folks almost singularly made my Ph.D worth it. Friends like family, they say, but as international students, friends are really the only family most of us have in the U.S. I am just incredibly lucky that I have built some wonderful friendships here, that will outlast my stay in Baltimore. All those friends I picked up as roommates, badminton and running buddies, dance buddies, affinity group buddies - you all know who you are. I thank you for keeping my life in Baltimore so vibrant. I hope for all of us to grow and flourish as we look to the future.

Of course, my acknowledgements would be incomplete without the mention of my family. It suffices to say that all I am today is thanks to my parents. My mom has taught me ambition, and my dad humility. They have always been proud of me and supported everything I wanted to do, and I hope this long document I wrote makes me their favorite child. My sister, the good (and current favorite) daughter, is the one I go to when I am frustrated, angry, annoyed or sad. I might not buy what she says but at least I am cracking up when the call ends. My brother-in-law cheers every achievement of mine like it were his own. My parents-in-law and (another) brother-in-law are family I picked up along my Ph.D. journey, and that means I have had twice the encouragement and support I previously had. And finally, the biggest and most special thank you to my partner Praveen who has seen me through it all, and thinks no crazy dream of mine is too big.

# Contents

<b>Abstract</b> . . . . .	<b>ii</b>
<b>Acknowledgements</b> . . . . .	<b>iv</b>
<b>Contents</b> . . . . .	<b>vii</b>
<b>List of Figures</b> . . . . .	<b>xi</b>
<b>Chapter 1 Introduction</b> . . . . .	<b>1</b>
1.1 Motivation . . . . .	4
<b>Chapter 2 Preliminaries</b> . . . . .	<b>8</b>
2.1 Fundamental Assumptions in Causal Inference . . . . .	8
2.2 Statistical and Causal DAG models . . . . .	10
2.3 Temporal Models . . . . .	12
2.4 Hidden Variable Causal Models . . . . .	15
2.4.1 The Nested Markov Factorization . . . . .	16
2.5 Interference . . . . .	18
2.6 Missing Data . . . . .	20
2.7 Chain Graph Models . . . . .	23
<b>Chapter 3 Path Dependent Models</b> . . . . .	<b>26</b>
3.1 Introduction . . . . .	26

3.1.1	Contributions . . . . .	27
3.2	Background . . . . .	28
3.2.1	Identification in Causal DBNs with Hidden Variables . . . . .	29
3.3	A Simple PDSEM . . . . .	31
3.4	Fully Observed PDSEMs . . . . .	34
3.5	PDSEMs with Hidden Variables . . . . .	37
3.6	Statistical Inference . . . . .	40
3.7	Experiments . . . . .	42
3.7.1	Simulation Study . . . . .	42
3.7.2	Septoplasty Application . . . . .	43
3.8	Conclusions . . . . .	45
<b>Chapter 4</b>	<b>Entangled Missingness . . . . .</b>	<b>46</b>
4.1	Introduction . . . . .	46
4.1.1	Contributions . . . . .	48
4.2	Background . . . . .	48
4.3	Motivating Example . . . . .	48
4.4	Graphical Models for Entangled Missingness . . . . .	50
4.4.1	Entanglements Without Missingness Interference . . . . .	51
4.4.2	Entanglements With Missingness Interference . . . . .	55
4.4.2.1	Beyond Dyads . . . . .	57
4.5	Identification Results . . . . .	61
4.5.1	Without Missingness Interference . . . . .	61
4.5.2	With Missingness Interference . . . . .	63
4.6	Experiments . . . . .	68
4.7	Conclusions . . . . .	69
<b>Chapter 5</b>	<b>Generalized Coarsening . . . . .</b>	<b>72</b>

5.1	Introduction . . . . .	72
5.1.1	Contributions . . . . .	73
5.2	Background . . . . .	74
5.3	Motivating Example: Cellular Differentiation . . . . .	74
5.4	Punctuated Causal Models . . . . .	77
5.4.1	Identification in Special Cases . . . . .	84
5.5	Inference . . . . .	87
5.5.1	Statistical Estimators . . . . .	87
5.5.2	Estimators for Specific Targets . . . . .	89
5.5.2.1	Target $\beta \equiv \mathbb{E}[Y_{t=1}^{1,2}(a^1, a^2)]$ . . . . .	89
5.5.2.2	Target $\psi \equiv \mathbb{E}[Y_{t=2}^{1,1}(a^1, a^2)]$ . . . . .	90
5.6	Experiments . . . . .	91
5.6.1	Simulation Study . . . . .	91
5.6.2	Cellular Reprogramming . . . . .	92
5.7	Conclusions . . . . .	95
<b>Chapter 6 Conclusions and Future Work . . . . .</b>		<b>96</b>
<b>Bibliography . . . . .</b>		<b>98</b>
<b>Appendix I Additional Background on Causal Graphical Models . . . . .</b>		<b>108</b>
A.	Graph Preliminaries . . . . .	108
B.	The Nested Markov Factorization . . . . .	108
C.	kth Order Markov Temporal Causal Models . . . . .	110
<b>Appendix II Path Dependent Models: Additional Material . . . . .</b>		<b>111</b>
A.	$k$ -th order Markov Temporal Causal Models . . . . .	111
B.	Representing a PDSEM as a DBN . . . . .	112
C.	Proofs . . . . .	113

<b>Appendix III</b>	<b>Entangled Missingness: Additional Material . . . . .</b>	<b>115</b>
A.	Binary Parameterization of Nested Markov Models . . . . .	115
A..1	Binary Parameterization of Missing Data Models . . . . .	115
B.	Proofs . . . . .	116
<b>Appendix IV</b>	<b>Generalized Coarsening: Additional Material . . . . .</b>	<b>122</b>
A.	Proofs . . . . .	122
B.	Cellular Reprogramming: Plots . . . . .	130

# List of Figures

<b>Figure 2.1</b>	(a) DAG representation of the conditionally ignorable model; (b) CDAG of the conditionally ignorable model, with additional background context $W$ . . . . .	11
<b>Figure 2.2</b>	(a) Prior network CDAG $\mathcal{G}_1$ , representing the state of the dynamic Bayesian network at time $t = 1$ . Here, $W = \emptyset$ (b) A CDAG $\mathcal{G}_{+1}$ representing the transitions in a dynamic Bayesian network. (c) A dynamic Bayesian network model unrolled to four time steps. . . . .	14
<b>Figure 2.3</b>	(a) A hidden variable DAG, and (b) its latent projection ADMG. . . . .	16
<b>Figure 2.4</b>	Example of a DAG illustration for interference between two units. . . . .	19
<b>Figure 2.5</b>	A missing data DAG example corresponding to (a) a MCAR model, (b) a MAR model, and (c) a MNAR model, all with $Z \equiv \{Y\}$ , and (d) A missing data MAR model with $Z \equiv \{A, Y\}$ . . . . .	22

<b>Figure 3.1</b>	(a),(d) Prior network hidden variable DAGs $\mathcal{G}_1$ , representing the state at time $t = 1$ . (b),(e) Conditional hidden variable DAGs $\mathcal{G}_{+1}$ representing the transitions in the network, with (e) leading to a first-order Markov model, and (b) leading to higher order dependences to unobserved hidden variables $U_t$ linking multiple time points. (c),(f) Latent projection ADMGs of the unrolled hidden variable DBNs to three time steps. . . . .	30
<b>Figure 3.2</b>	A simple PDSEM. (a) Causal structure of the initial state $S^1$ . (b) The state transition diagram. (c),(d),(e) Causal diagrams representing possible transitions and subsequent states. (f) Causal relationships in a system evolving according to the state transitions: $s^1 \rightarrow s^2 \rightarrow s^3$ . (g) A snapshot of a possible PDSEM trajectory that terminates in 3 timepoints is represented as an unrolled ADMG. . . . .	32
<b>Figure 3.3</b>	A hidden variable PDSEM. (a) Causal structure of the initial state $S^1$ . (b) The state transition diagram. (c),(d),(e) Latent projected causal diagrams representing possible transitions and subsequent states. (f) A snapshot of a possible PDSEM trajectory represented as an unrolled ADMG . . . . .	39
<b>Figure 3.4</b>	Histograms of the number of transitions in a surgery under two different interventions: when a more experienced surgeon performs the entire procedure, and when a less experienced trainee performs the entire procedure. . . . .	44

<b>Figure 3.5</b>	Histograms of hypothetical surgeries performed only by a junior trainee surgeon (blue) versus hypothetical surgeries performed only by a senior attending surgeon (orange). Surgeries performed by the attending are slightly longer ( $\mu = 244.3.91, \sigma = 139.9$ ) than those of the trainee ( $\mu = 233.5, \sigma = 125.9$ ). . . . .	45
<b>Figure 3.6</b>	The state transition diagram for the surgery data application. . . . .	45
<b>Figure 4.1</b>	Four scenarios representing all possible ways target law dependence and missingness process dependence may arise in a dyadic partial interference setting without missingness interference arising. . . . .	52
<b>Figure 4.2</b>	Four scenarios representing all possible ways target law dependence and missingness process dependence may arise in a dyadic partial interference setting when missingness interference is present. We assume that only outcomes $Y_i$ are missing, and $R_i$ is shorthand for $R_{Y_i}$ . . . . .	55
<b>Figure 4.3</b>	A simple example of a general network with $n$ units undergoing missingness interference. This graph represents an MCAR model, $\mathbf{R} \perp\!\!\!\perp \mathbf{O}, \mathbf{Z}_i^{(1, \mathbf{r}_{\text{aff}(i)})}$ . . . . .	58
<b>Figure 4.4</b>	A general network with $d$ units featuring missingness interference. (a) - (d) enumerate all possible ways target law dependence and missingness process dependence may arise, as a generalization of the dyad in Fig. 4.2. . . . .	59
<b>Figure 4.5</b>	Two illustrations of Theorem 4. ADMG in (a) does not have a e-colluding path. ADMG in (b) features the e-colluding path $Z_2^{(1, r_1=1)} \rightarrow R_3 \leftarrow R_1$ . . . . .	67
<b>Figure 4.6</b>	The model used to generate synthetic data for our experiments. . . . .	69

<b>Figure 4.7</b>	Bias recorded in bootstrapped estimates of targets (shown in x-axis), MAR case (left) and MNAR (right). We compare our adjusted IPW estimates (denoted by an asterisk (*)) on the x-axis) to the unadjusted estimate, which is obtained by ignoring the network missingness structure underpinning the data. Error bars represent quantiles $q_{0.05}$ and $q_{0.95}$ across 50 bootstrap samples. . . . .	70
<b>Figure 5.1</b>	Reprogramming requires a reversal (above) of a differentiated cell into a pluripotent stem cell (i.e. iPS), which next may undergo a differentiation process (below) into another differentiated cell. . . . .	75
<b>Figure 5.2</b>	(a) A simple PCM with four ECPs (grey dashed blocks), including one trivial ECP for treatment A, and three observable macrostates (milestones) of cell differentiation. (b) Unrolled ECP, which corresponds to the first milestone (endogenous Oct4 expression) in the PCM in (a). The ECP contains microstate variables that unroll to infinite time. The grey dashed edge from (a) to (b) is not part of the causal diagram, and is used only to indicate that (b) is an unrolled version of an ECP component of PCM in (a). (c) Microstate counterfactuals that result from an intervention $A = a$ : $V_{t=1}^{1,1}(a)$ is neural cell identity at time point $t = 1$ had the transcription factor been set to $a$ in the (sub)process when endogenous Oct4 expression is being attained. . . . .	79
<b>Figure 5.3</b>	The model that was used to generate data in the simulation study. (a) The PCM (b) The expanded equilibrium process in the ECP involving $Y^{1,1}$ and $Y^{1,2}$ . . . . .	89

<b>Figure 5.4</b>	Histograms showing how $\hat{\beta} - \hat{\beta}_{trial}$ varies across bootstrap trials for targets $\hat{\beta} = E[Y_{t=1}^{1,2}(a^1, a^2)]$ on the left and $E[Y_{t=2}^{1,1}(a^1, a^2)]$ on the right, for the intervention $a^1 = a^2 = 0$ . . . . .	93
<b>Figure 5.5</b>	Sample trajectories corresponding to 2 dimensions, namely MEF and epithelial identities. The blue trajectory is derived the Gibbs compatible PCM, and the orange one represents the trajectory from the original dataset, coarsely sampled every 12 hours. . . . .	94
<b>Figure II.1</b>	A causal DBN encoding the PDSEM in Fig. 3.2, via (a) the prior network, and (b) the complete transition network with context-specific independences. . . . .	112
<b>Figure III.1</b>	Examples where $P(\tilde{\mathbf{Z}}^{(r=1)}, \mathbf{R})$ is not identified. The proof is demonstrated using parameter counting. (a)-(c): Extended self-censoring (neighbor-censoring), (a) depicts the data generating process, (b) ADMG for $P(\tilde{\mathbf{Z}}^{(r=1)}, \mathbf{R})$ , (c) ADMG for observed law. (d)-(f): Extended collider, (d) depicts the data generating process, (e) ADMG for $P(\tilde{\mathbf{Z}}^{(r=1)}, \mathbf{R})$ , (f) ADMG for observed law. . . . .	118
<b>Figure III.2</b>	(a) e-Colluding paths between $Z_i^{(1, r_{\text{aff}(i)}=1)}$ and $R_j$ where $R_j \in \mathbf{R}_{\text{aff}(i)}$ ( $i$ and $j$ are neighbors) (b) Projecting out $Z_i^{(1, r_{\text{aff}(i)}=1)}$ . . . . .	119
<b>Figure IV.1</b>	Trajectories for the first 15 (of 32) dimensions for medium 2i. The blue trajectory is derived from Gibbs compatible PCM and the orange trajectory is plotted from the original dataset for comparison, sampled every 12 hours. See the next page for the other dimensions. . . . .	130

<b>Figure IV.2</b>	Trajectories for 17 (of the 32) dimensions for medium 2i, continued from Fig. IV.1. The blue trajectory is derived from Gibbs compatible PCM and the orange trajectory is plotted from the original dataset for comparison, sampled every 12 hours. . . . .	131
<b>Figure IV.3</b>	Trajectories for the first 15 (of 32) dimensions for medium serum. The blue trajectory is derived from Gibbs compatible PCM and the orange trajectory is plotted from the original dataset for comparison, sampled every 12 hours. See the next page for the other dimensions. . . . .	132
<b>Figure IV.4</b>	Trajectories for 17 (of the 32) dimensions for medium serum, continued from Fig. IV.3. The blue trajectory is derived from Gibbs compatible PCM and the orange trajectory is plotted from the original dataset for comparison, sampled every 12 hours. . . . .	133

# Chapter 1

## Introduction

Cause-effect relationships form an integral part of our intuitive understanding, and thus our description, of the world around us as humans. For example, we might have heard the words “Incessant rains caused the roads to flood yesterday”, or said, “My cough is the effect of a flu infection”. Unsurprisingly, much scientific inquiry tends to be causal in nature too, and various attempts to define and understand *causality* have been made, since the days of Democritus, back in 400 BC <sup>1</sup>. More recently, since Jerzy Neyman’s work in 1923 [Splawa-Neyman et al., 1990], *counterfactual* scenarios arising from hypothetical thought experiments have become a prominent tool to reason about causality. For example, in a surgery, one might propose answering the question, "What would the patient’s outcome *have been* if the surgery was performed by an attending, instead of a resident surgeon?", to assess the effect of experience on surgical outcome. Or, "What would the reduction in carbon emissions *have been* if this policy were to be enacted?", to assess the effect of a particular law on the environment. Methods in the field of *causal inference* [Pearl, 2000, Hernan and Robins, 2020] are developed to pose and evaluate such counterfactual queries to shed light on cause-effect relationships.

The presence and magnitude of cause-effect relationships are often investigated using

---

<sup>1</sup>“I would rather discover one causal law than be King of Persia”, quoted in [Pearl, 2000]

a randomized controlled trial (RCT). In an ideal RCT, the treatment randomization scheme ensures that all dependence between treatment and outcome due to spurious factors is eliminated, and observed dependence between treatment and outcome is due to a causal relationship. However, running an RCT to answer causal questions is not always feasible or ethical. For example, there may not be enough resources to run large trials, or it may not be ethical to randomize which patient receives a life-saving drug (or not) for a debilitating disease. The field of observational causal inference seeks to bridge this gap: observational causal inference seeks to answer counterfactual questions using *observational (non-randomized)* data. It does so by emulating a target RCT [Hernán and Robins, 2016] and obtaining an estimate of the causal effects an RCT, if conducted, would have yielded. Observational data from various fields like biomedicine and public health, socio-economics, climate science among others, is growing rapidly in size with the advancement of data collection technology and, so is the curiosity about what observational causal inference methods can bring to data analysis.

One can better understand the structure of a hypothetical target experiment using the simple illustration in Table 1-I, which records variables associated with a surgery that was either performed by a resident trainee or an experienced attending. Let us assume that investigators are interested in total surgery time as the outcome. For any patient who underwent surgery, only one of their two possible outcomes, is recorded in data; the outcome that is contrary to fact, i.e., the outcome of a surgery performed by the other surgeon (in grey), is not accessible. In general, causal effects are formulated as population level comparisons involving both counterfactuals. In this example, we might construe the (average) causal effect of surgeon experience as the difference in mean surgery time under the attending ( $=1.4$ ) versus resident ( $=1$ ), giving us a value of  $0.4$ . We are able to easily obtain this number if we have access to both counterfactuals for every individual.

In real settings, one has to investigate whether the causal effect is *identified* from actually available data. Parameters must be identified in order for parameter estimation to be well-defined; a parameter is said to be identified if it is a unique function of the observed data distribution. That is, one must establish a map between the observed data and counterfactuals via *assumptions* about the data generating process and the causal model, which can be formulated via a causal graph. In fact, this dissertation will describe causal models mainly using the language of graphical causal models [Pearl, 2000, Hernan and Robins, 2020]. We note, however, that substantial causal inference literature exists that does not rely on graphical model; see [Rubin, 1974] for example.

The observational causal inference pipeline using graphical models begins with the dataset and population of interest. Typically, the dataset would consist of one or more outcomes and possible causes (treatments) one would like to investigate, and in addition, other variables that might be relevant, like background information on demographics. Understanding the context in which data was collected and how good the measurement systems were, are critical to good data analysis. For example, there could be seemingly innocuous phenomena, like a clerical typing error rendering an arbitrary entry wrong or missing, or familial connections between study participants which renders the data samples not independent of each other. Or simply, the data might arise from a setting that has significant temporal structure and should not

Outcome under attending	Outcome under resident	Surgery by	Observed outcome
1.2	0.7	Resident	0.7
1.4	1.1	Attending	1.4
2.2	0.6	Attending	2.2
0.8	1.6	Resident	1.6

**Table 1-I.** The hypothetical truth table for a surgery: The first two columns record patient outcomes if each patient were operated on by an attending, and independently, by a resident. In real data, however, we have access to only the third and fourth columns, the surgeon who actually performed the surgery and the patient's outcome.

be treated as static observations. Considering the context from which data arises is crucial to using an appropriate causal model.

Following dataset curation, one posits a graphical causal model linking the variables in the data. Such a model might be posited directly on consultation with domain-specific literature or *discovered* using a variety of causal discovery methods [Spirtes et al., 2001, Ramsey et al., 2018] or obtained using a combination of both approaches. With the graph and data in hand, the next step involves investigating whether the parameter of interest is identified in the given model [Tian and Pearl, 2002, Shpitser, 2017], and a significant portion of this dissertation involves discussing identification theory for novel models and targets of interest. Finally, the last step typically involves *estimating* the parameter from observed data, if the parameter is indeed identified [Bickel and Doksum, 2015]. It is worth noting that there are other steps in the pipeline that might be considered, like sensitivity analysis of our identified estimates or obtaining bounds of estimates in the case of unidentified targets [Manski, 1990, Balke and Pearl, 1997, Richardson et al., 2014]. We shall not discuss these topics in this dissertation, however.

## 1.1 Motivation

Beginning with Sewell Wright’s path analysis of graphical models for animal husbandry applications back in 1934, observational causal inference has relied on certain assumptions to identify causal effects [Wright, 1934]. In Wright’s case, he assumed linear models. Thereafter, works in [Rubin, 1974, Pearl, 2009, Robins, 1986] have discussed more general causal models, with fewer and less restrictive assumptions, resulting in seminal results in model selection and general identification theories for non-parametric models [Tian and Pearl, 2002, Shpitser and Pearl, 2006, Richardson et al., 2017]. While these theories provide a strong foundation, many of these models still applied directly only to relatively simple settings from which data arose, like lack

of censoring or independent and identically distributed (i.i.d.) subjects in the data. For most scientific systems, such models tend to be fairly unrealistic. More recently, considerable work has been done to explore more realistic attributes in causal settings. For example, public health studies, signal processing systems or many biomedical signals often exhibit temporal dynamics, and models and identification theory for time-invariant dynamic causal systems were developed to capture them [Peters et al., 2013, Blondel et al., 2017]. Similarly, the recognition that ties between subjects in public health studies or social network analyses is unavoidable, resulted in the study of data dependence and *interference*: causal dependence of one experimental unit’s variables on another unit’s variables [Aronow and Samii, 2013, Hudgens and Halloran, 2008a, Basse et al., 2019, Bowers et al., 2013, Ogburn and VanderWeele, 2014]<sup>2</sup>. These works pushed the techniques of causal modeling forward, allowing for a more nuanced and richer description of certain causal systems.

The work in this dissertation is motivated by the recognition that good causal analysis of observational data relies on a sound understanding of the attributes of the system from which the data arises and the ecosystem in which it was recorded, and the need to develop models that enable a rich characterization of these systems. We aim to explore specific system attributes that go beyond what classical causal models have been able to capture and primarily study the question of identification in these models. In some cases, we will also discuss estimation of identified models. We briefly list below, the types of system attributes we are interested in, and motivate why we would be interested in them.

**Path dependence:** Causal analyses of longitudinal data have historically assumed that the qualitative causal structure relating variables remains invariant over time. In structured systems that transition between qualitatively different states in discrete time steps, such an approach is deficient. For example, consider the task of modeling surgical

---

<sup>2</sup>For a detailed set of references, please refer to pertinent material presented in Chapter 2

procedures. Surgeries are often divided into discrete stages, each with a distinct goal, variables and relationships among them, possibly not shared across stages: contrast a skill-intensive cartilage repair against a routine incision closure that can be automated, as an example. Existing time-invariant models cannot capture the attributes of *path dependence*: (i) time-varying variables with state-specific causal relationships, (ii) an intervention that results in a completely altered temporal evolution of the system, due to downstream state transitions distinct from these observed in data. We introduce the path dependent structural equation model (PDSEM) to describe such systems.

**Entangled missingness:** There is growing interest in causal and statistical inference for settings with data dependence, i.e. data samples are assumed to not be i.i.d. However, there are a surprisingly small number of methods overall, and *no* graphical causal methods, to account for missingness in dependent data settings. As a motivating example for such a setting, consider a mass public health surveillance effort, where demographic and family data are collected on a large population of individuals from which a smaller group is invited to participate in a substudy where more detailed data are collected. An individual’s choice to enroll in the substudy may depend both on their own characteristics, and on the characteristics of, or choices made by, other individuals in their social network. Not accounting for such factors (of dependence and missingness), will result in bias. For such systems, we develop a framework for causal inference in the presence of *entangled missingness*, defined as missingness with dependence.

**Generalized coarsening:** In many physical systems, causal processes operate at different levels of temporal granularity. For instance, measurables in a “macrostate” such as pressure and temperature, are a result of dynamics in “microstates” corresponding to particle motion leading to a thermal *equilibrium*. In general, detailed information about microstates cannot be obtained from *coarsened* data on macrostates

(at equilibrium) alone. For these types of causal systems, we pose the problem of inferring microstate information from data on macrostates as a formal problem of causal identifiability, by developing a specialized model, called the punctuated causal model (PCM) to describe these systems. Then, we are able to discuss identification in the models, and the assumptions required to identify the microstate parameters.

The dissertation is organized as follows. Chapter 2 is a detailed review of graphical causal models. It sets up the notation and introduces concepts that will provide a foundation for the rest of the dissertation. All of this content is prior work that will be cited appropriately. The following chapters constitute original work, and are as follows: chapter 3 discusses path dependent causal models, chapter 4 expounds on entangled missingness, and chapter 5 discusses generalized coarsening. Within each of these chapters, we dive into the motivation behind the model, prior work, our contributions including model formulation and identification results, followed by experiments and future directions for the topic at hand. Chapter 6 provides closing thoughts and concludes the dissertation. We would like to note that the research for chapters 3 and 4 was conducted mainly by the author, with valuable input from the co-authors of the corresponding research articles. The research for chapter 5 was conducted jointly by the author of the dissertation with Numair Sani, as first co-authors. As a consequence, (1) we present only a part of the results here and, (2) some of the material in his (future) dissertation might be similar to that in this chapter; our individual contributions remain distinct, however.

# Chapter 2

## Preliminaries

In this chapter, we set up the notation and terminology necessary to follow the rest of the dissertation. We point out that much of the content in this chapter (as well as the following chapters) has been reproduced from relevant parts of the published works of the author of this dissertation or their manuscripts in preparation.

### 2.1 Fundamental Assumptions in Causal Inference

A *counterfactual* or *potential* outcome  $Y_i(a)$  in causal inference is defined as the outcome that would have been observed if, possibly contrary to fact, a unit (or sample)  $i$  had received treatment  $A = a$ . Causal effects are typically defined as contrasts of counterfactual outcomes, e.g.  $\beta := \mathbb{E}[Y(a) - Y(a')]$  is a comparison of the expected value of counterfactual outcomes in a world in which every unit receives treatment  $A = a$  compared with a world in which every unit receives treatment  $A = a'$ . This parameter  $\beta$  is often referred to as the average causal effect (ACE). Since potential outcomes are not directly observed, assumptions are needed to link the counterfactual distributions and the observed data distribution, from which samples are actually drawn, to identify  $\beta$ .

A standard set of assumptions used to express  $\beta$  as a functional of observed data are the following:

1. *Consistency*: The observed outcome realization is equal to the counterfactual outcome realization had treatment been set to the observed value, or  $Y(a) = Y$  if  $A = a$ .
2. *Positivity*: All treatment assignments, possibly conditioned on a set of baseline covariates  $\mathbf{C}$ , have positive support. That is,  $p(a \mid \mathbf{C}) > 0$  for all  $a$  in the support of  $A$  and support of  $\mathbf{C}$  such that  $P(\mathbf{C} = c) > 0$ .
3. *Conditional Ignorability*: Potential outcomes are independent of the treatment assignment, possibly conditioned on a set of baseline covariates  $\mathbf{C}$ . That is,  $Y(a) \perp\!\!\!\perp A \mid \mathbf{C}$  for all  $a$ .

Under these assumptions, the parameter  $\beta$  is identified from the observed data distribution  $p(Y, A, \mathbf{C})$  via the *adjustment functional*:  $\mathbb{E}[\mathbb{E}[Y \mid A = a, \mathbf{C}] - \mathbb{E}[Y \mid A = a', \mathbf{C}]]$ .

Graphical models can be used to encode these types of conditional independence assumptions in a visually intuitive fashion, and facilitate causal inference. In the remaining sections of this chapter, we discuss prior work on graphical models that sets the background for the following chapters. We first discuss directed acyclic graph (DAG) models. DAG models provide a gentle and intuitive introduction to the graphical model world, and are quite popular in the statistics and machine learning literature. Following DAG models, we introduce the dynamic bayesian network (DBN) model for temporal data. While DAG models are generally used to depict static information, a simple extension has been formulated in the form of a DBN model to describe temporal relationships between variables. DBN models also tend to be commonly used in signal processing and machine learning domains. In this dissertation, DBNs are used in Chapter 3 and Chapter 5. In the presence of hidden variables, a different type of graphical model called an acyclic directed mixed graph, which is derived from a hidden variable DAG using a set of rules (known as the latent

projection operator), is often preferable to using the DAG directly. ADMGs can be used to represent identification theory for a class of hidden variable causal models. We describe ADMGs after DBNs and use them in Chapter 3 and Chapter 4. Next, we discuss how DAGs (and ADMGs) have been extended to encode interference and missing data settings, and use these formulations in Chapter 4. Finally, we also briefly introduce chain graph models for systems in equilibrium, which are only considered in our study of generalized coarsening in Chapter 5. Any background material that might be relevant but not absolutely necessary to this dissertation has been discussed in Appendix I.

## 2.2 Statistical and Causal DAG models

A directed acyclic graph (DAG)  $\mathcal{G}(\mathbf{V})$  is a graph  $\mathcal{G}$  with a vertex set  $\mathbf{V}$  representing variables  $\mathbf{V}$ , and consists only of directed edges ( $\rightarrow$ ) between any two vertices, with no directed cycles. A directed cycle is a path  $V \rightarrow \dots \rightarrow V$ , that begins and ends with the same vertex  $V$ . The *statistical* model of a DAG, also called a *Bayesian network*, is the set of distributions that Markov factorize with respect to the DAG as  $p(\mathbf{V}) = \prod_{V \in \mathbf{V}} p(V \mid \text{pa}_{\mathcal{G}}(V))$  where  $\text{pa}_{\mathcal{G}}(V)$  are parents of  $V$  in graph  $\mathcal{G}$ ; we say  $U$  is a parent of  $V$  if  $U \rightarrow V$ , for  $U, V \in \mathbf{V}$ . All conditional independence restrictions in  $p(\mathbf{V})$  are encoded in the DAG  $\mathcal{G}(\mathbf{V})$ , and can be read off the graph by applying a set of d-separation rules on the paths in  $\mathcal{G}(\mathbf{V})$  [Pearl, 1988].

*Causal* models of a DAG are also sets of distributions but on counterfactual random variables. They combine a generative model of  $p(\mathbf{V})$  with the theory of interventions to yield distributions over counterfactual random variables. Each variable  $V$  in a causal model is determined from values of its parents  $\text{pa}_{\mathcal{G}}(V)$  and an exogenous noise variable  $\epsilon_V$  via an invariant causal mechanism called a *structural equation*  $f_V(\text{pa}_{\mathcal{G}}(V), \epsilon_V)$ . Intervention operations in causal models replace each structural equation  $f_V(\text{pa}_{\mathcal{G}}(V), \epsilon_V)$  for  $V \in \mathbf{A} \subset \mathbf{V}$  by one that sets  $V$  to a constant value in  $\mathbf{a}$



**Figure 2.1.** (a) DAG representation of the conditionally ignorable model; (b) CDAG of the conditionally ignorable model, with additional background context  $\mathbf{W}$ .

corresponding to  $V$  [Pearl, 2009]. The joint distribution of variables in  $\mathbf{Y} \equiv \mathbf{V} \setminus \mathbf{A}$  after the intervention  $\text{do}(\mathbf{a})$  was performed is denoted by  $p(\mathbf{Y} \mid \text{do}(\mathbf{a}))$ , equivalently written as  $p(\{V(\mathbf{a}) : V \in \mathbf{Y}\})$ , or  $p(\mathbf{Y}(\mathbf{a}))$ , where  $V(\mathbf{a})$  is a counterfactual random variable or a potential outcome. We assume that the structural equations and noise terms are such that the resulting  $p(\mathbf{V})$  is a positive distribution.

The widely used *non-parametric structural equation model with independent errors (NPSEM-IE)* [Pearl, 2009], which is the model we will resort to, throughout this dissertation, assumes additionally that the joint distribution of all exogenous terms are marginally independent:  $p(\epsilon) = \prod_{V \in \mathbf{V}} p(\epsilon_V)$ . The NPSEM-IE implies the DAG factorization of  $p(\mathbf{V})$  with respect to  $\mathcal{G}(\mathbf{V})$ , and a truncated DAG factorization known as the *g-formula*:

$$p(\mathbf{Y}(\mathbf{a})) = \prod_{V \in \mathbf{Y}} p(V \mid \text{pa}_{\mathcal{G}}(V)) \big|_{\mathbf{A}=\mathbf{a}} \quad (2.1)$$

for every  $\mathbf{A} \subseteq \mathbf{V}$ , and  $\mathbf{Y} = \mathbf{V} \setminus \mathbf{A}$ .

As an example, the DAG representation of the conditionally ignorable model discussed in 2.1 is shown in Fig. 2.1 (a). The joint distribution factorizes as:  $p(\mathbf{V}) = p(C)p(A \mid C)p(Y \mid A, C)$ .

Graphical models can be extended to represent the dependence of a set of variables  $\mathbf{V}$  on a set of variables  $\mathbf{W}$  held fixed to values  $\mathbf{w}$ . Such models are defined using conditional versions of graphs. A *conditional DAG (CDAG)*  $\mathcal{G}(\mathbf{V}, \mathbf{W})$  is a graph where vertices  $\mathbf{V}$  represent random variables, and vertices  $\mathbf{W}$  provide fixed context, and denote random variables  $\mathbf{W}$  set to values  $\mathbf{w}$ . In a CDAG, for every  $W \in \mathbf{W}$ ,  $\text{pa}_{\mathcal{G}}(W) = \emptyset$ . A

statistical model of a CDAG  $\mathcal{G}(\mathbf{V}, \mathbf{W})$  is a set of conditional distributions that can be written as  $p(\mathbf{V} \mid \mathbf{W}) = \prod_{V \in \mathbf{V}} p(V \mid \text{pa}_{\mathcal{G}}(V))$ , where  $\text{pa}_{\mathcal{G}}(V)$  may include elements in  $\mathbf{W}$ . See Fig. 2.1(b) for an example of a CDAG, where  $W$  has no parents and is placed inside a square to indicate that it is fixed to a constant value.

A *conditional causal model*, associated with a CDAG  $\mathcal{G}(\mathbf{V}, \mathbf{W})$ , may be viewed as defined by a set of structural equations  $f_V(\text{pa}_{\mathcal{G}}(V), \epsilon_V)$ , where  $\text{pa}_{\mathcal{G}}(V)$  may include elements in  $\mathbf{W}$ . In such models, interventional distributions  $p(\mathbf{Y}(\mathbf{a}) \mid \mathbf{W})$  (for  $\mathbf{Y} = \mathbf{V} \setminus \mathbf{A}$ ) depend on  $\mathbf{W}$  in general, and their identification is obtained by a generalization of the g-formula:

$$p(\mathbf{Y}(\mathbf{a}) \mid \mathbf{W}) = \prod_{V \in \mathbf{Y}} p(V \mid \text{pa}_{\mathcal{G}}(V))|_{\mathbf{A}=\mathbf{a}} \quad (2.2)$$

where  $\text{pa}_{\mathcal{G}}(V)$  may include elements in  $\mathbf{W}$ .

Before we discuss temporal graphical models, we would like to state that, in this work, we are primarily interested in *non-parametric identification* of the parameters in a model, i.e., there are no restrictions on the structural equations of the causal model aside from what the graph asserts. Our use of the word identification is synonymous with non-parametric identification for the purposes of this dissertation.

## 2.3 Temporal Models

While Bayesian networks lend themselves well to modeling static data, data that changes over time requires more sophisticated models. An extension of the Bayesian network model for discrete time temporal systems is the popular *dynamic Bayesian network (DBN)* model [Murphy, 2012]. We introduce temporal models, and in specific DBN models, in this section, to serve as background material to later describe temporal relationships between variables in Chapter 3 and Chapter 5.

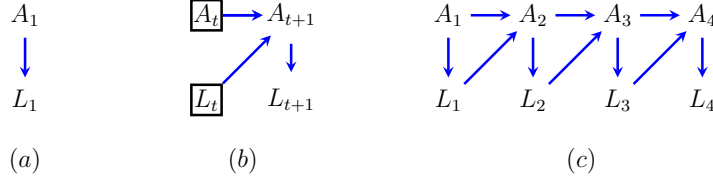
A conditional DBN (CDBN) generalizes a DBN given context  $\mathbf{W}$ , and is specified

by a pair of CDAGs, and a corresponding pair of factorized distributions. The *prior network* CDAG  $\mathcal{G}_1(\mathbf{V}_1, \mathbf{W})$ , or simply  $\mathcal{G}_1$ , containing random variables  $\mathbf{V}_1$  and fixed vertices  $\mathbf{W}$ , and the corresponding distribution  $p(\mathbf{V}_1 | \mathbf{W}) = \prod_{V \in \mathbf{V}_1} p(V | \text{pa}_{\mathcal{G}_1}(V))$  represent the state of the system at the first time step. The *transition network* CDAG  $\mathcal{G}_{+1}(\mathbf{V}_{+1}, \mathbf{V} \cup \mathbf{W})$ , or simply  $\mathcal{G}_{+1}$ , involves fixed vertices  $\mathbf{V}, \mathbf{W}$  and random variables  $\mathbf{V}_{+1}$  representing vertices at time point  $t + 1$ .  $\mathcal{G}_{+1}$  as well as its corresponding distribution  $p(\mathbf{V}_{+1} | \mathbf{V} \cup \mathbf{W}) = \prod_{V \in \mathbf{V}_{+1}} p(V | \text{pa}_{\mathcal{G}_{+1}}(V))$  represent the way variables  $\mathbf{V}_{+1}$  at any time point  $t + 1$  depend on each other, and on variables  $\mathbf{V}$  at the prior time point  $t$ , as well as possibly on values  $\mathbf{w}$  of  $\mathbf{W}$ . This kind of dependence leads to a *first-order Markov* CDBN, as variables at  $t + 1$  only depend on variables at  $t$  or  $\mathbf{W}$ , and not on other prior variables, such as those at timepoint  $t - 1$ .

A CDBN represents temporal dynamics up to any time point  $T$  via a distribution  $p(\mathbf{V}_1 \cup \mathbf{V}_2 \cup \dots \cup \mathbf{V}_T | \mathbf{W})$  where  $\mathbf{V}_1$  represents the variables in the prior network  $\mathcal{G}_1$ , and  $\mathbf{V}_t$  for  $t = 2, \dots, T$  represents reindexed copies of  $\mathbf{V}_{+1}$  corresponding to each subsequent time point. This distribution is represented in a tractable way by the following “unrolled” factorization, using  $\mathcal{G}_1$  and  $\mathcal{G}_{+1}$ :

$$\begin{aligned} p(\mathbf{V}_1 \cup \dots \cup \mathbf{V}_T | \mathbf{W}) &= p(\mathbf{V}_1 | \mathbf{W}) \prod_{t=1}^{T-1} p(\mathbf{V}_{t+1} | \mathbf{V}_t \cup \mathbf{W}) \\ &= \prod_{V \in \mathbf{V}_1} p(V | \text{pa}_{\mathcal{G}_1}(V)) \prod_{t=1}^{T-1} \prod_{V \in \mathbf{V}_{t+1}} p(V | \text{pa}_{\mathcal{G}_{+1}}(V)). \end{aligned}$$

A simple DBN, i.e., a CDBN where  $\mathbf{W} = \emptyset$ , is shown in Fig. 2.2, where the prior network (2.2(a)) contains two variables  $A$  and  $L$ , and the transition network (2.2(b)) shows connections among the state variables in the prior state at time  $t$  and the subsequent state at time  $t + 1$ . Fig. 2.2(c) shows the DBN implied by these prior and transition networks unrolled to 4 time steps.



**Figure 2.2.** (a) Prior network CDAG  $\mathcal{G}_1$ , representing the state of the dynamic Bayesian network at time  $t = 1$ . Here,  $W = \emptyset$  (b) A CDAG  $\mathcal{G}_{+1}$  representing the transitions in a dynamic Bayesian network. (c) A dynamic Bayesian network model unrolled to four time steps.

CDBNs can be naturally extended to represent causal models by assuming that both prior and transition networks are *causal* CDAGs. In other words, we assume values of every variable  $V$  in both the prior and the transition network is determined, via a structural equation  $f_V(\cdot)$ , in terms of its observed parents  $\text{pa}_{\mathcal{G}_1}(V)$  (or  $\text{pa}_{\mathcal{G}_{+1}}(V)$ ) and an exogenous noise term  $\epsilon_V$ . If we further assume that all exogenous noise variables are marginally independent, we arrive at a DBN version of the NPSEM-IE, where in addition to the  $g$ -formula (2.1) holding for the prior network, the *conditional  $g$ -formula* holds for the transition network:

$$p(\mathbf{Y}_{t+1}(\mathbf{a})|\mathbf{V}_t, \mathbf{W}) = \prod_{V \in \mathbf{Y}_{t+1}} p(V|\text{pa}_{\mathcal{G}_{+1}}(V))|_{\mathbf{A}=\mathbf{a}}, \quad (2.3)$$

for any  $\mathbf{A} \subseteq \mathbf{V}_{t+1}$ , and  $\mathbf{Y}_{t+1} = \mathbf{V}_{t+1} \setminus \mathbf{A}$ . Thus, a causal CDBN unrolled to a set of time points  $1, \dots, T$  yields a standard causal CDAG model with vertices  $\mathbf{V}_{1:T} \equiv \mathbf{V}_1 \cup \mathbf{V}_2 \cup \dots \cup \mathbf{V}_T$  and  $\mathbf{W}$ . For an intervention that sets  $\mathbf{A} \subseteq \mathbf{V}_{1:T}$  to constant values  $\mathbf{a}$ , the interventional distribution  $p(\mathbf{Y}_{1:T}(\mathbf{a}))$ , where  $\mathbf{Y}_{1:T} = \mathbf{V}_{1:T} \setminus \mathbf{A}$ , is identified by:

$$\prod_{V \in \mathbf{V}_1 \setminus \mathbf{A}} p(V|\text{pa}_{\mathcal{G}_1}(V)) \prod_{t=1}^{T-1} \prod_{V \in \mathbf{V}_{t+1} \setminus \mathbf{A}} p(V|\text{pa}_{\mathcal{G}_{+1}}(V)) \Big|_{\mathbf{A}=\mathbf{a}} \quad (2.4)$$

The first-order Markov assumption in CDBN models may be relaxed to a  $k$ th-order Markov assumption, where the model at any time step depends on variables in at

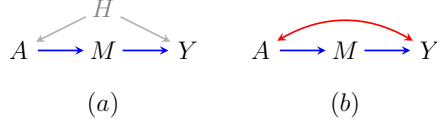
most  $k$  prior time steps, a generalization we describe in Appendix I.

## 2.4 Hidden Variable Causal Models

In causal models where all relevant variables are observed, the g-formula (2.1) provides an elegant link between observed data and counterfactual distributions. In practice, however, there are *hidden (latent)* variables: variables that are relevant to answering the scientific query being investigated, but not recorded in data. Representations for such data using a DAG  $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$ , or CDAG  $\mathcal{G}(\mathbf{V} \cup \mathbf{H}, \mathbf{W})$ , where  $\mathbf{V}$  and  $\mathbf{H}$  correspond to observed and hidden variables, respectively, and  $\mathbf{W}$  corresponds to fixed observed context, is not very helpful; applying (2.1) to  $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$  or  $\mathcal{G}(\mathbf{V} \cup \mathbf{H} \cup \mathbf{W})$  results in an expression that involves unobserved variables  $\mathbf{H}$ . Inferences made by marginalizing out  $\mathbf{H}$  may be sensitive to assumptions made about the state spaces for  $\mathbf{H}$  and the latent variable model may contain singularities at which asymptotics are irregular [Drton, 2009]. Additionally, such a model does not form a tractable search space: an arbitrary number of hidden variables and associated structures may be incorporated that are consistent with the observed data distribution.

A popular alternative is to represent a class of hidden variable DAGs  $\mathcal{G}_i(\mathbf{V} \cup \mathbf{H}_i)$  by a single *acyclic directed mixed graph* ADMG  $\mathcal{G}(\mathbf{V})$  that consists only of observed variables, and contains directed ( $\rightarrow$ ) and bidirected ( $\leftrightarrow$ ) edges and no directed cycles via the *latent projection* operation [Verma and Pearl, 1990]. In an ADMG, the bidirected edges represent unobserved confounding. A simple example is shown in Fig. 2.3: the hidden variable DAG in Fig. 2.3(a) can be viewed alternatively by its latent projection ADMG in Fig. 2.3(b). Conditional independence statements from a distribution  $p(\mathbf{V})$  may be read off an ADMG by a generalization of d-separation called *m-separation*.

In hidden variable models, not every causal effect is identified, and identification



**Figure 2.3.** (a) A hidden variable DAG, and (b) its latent projection ADMG.

theory may be expressed on the latent projection ADMG. The ADMG  $\mathcal{G}(\mathbf{V})$  captures relationships between observed variables  $\mathbf{V}$  implied by the factorization of  $p(\mathbf{V} \cup \mathbf{H})$  with respect to  $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$  via the *nested Markov factorization* of  $p(\mathbf{V})$  with respect to  $\mathcal{G}(\mathbf{V})$  [Richardson et al., 2017]. That is, the nested Markov factorization completely avoids modeling hidden variables but it captures all equality constraints a hidden variable DAG factorization imposes. Just as identification in DAGs may be viewed in terms of a modified DAG factorization (2.1), identification in a hidden variable DAG  $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$  may be viewed in terms of a modified nested factorization of  $\mathcal{G}(\mathbf{V})$ , as we describe next.

#### 2.4.1 The Nested Markov Factorization

The nested Markov factorization of  $p(\mathbf{V})$  with respect to  $\mathcal{G}(\mathbf{V})$  is defined in terms of *Markov kernels* of the form  $q_{\mathbf{D}}(\mathbf{D} \mid \text{pa}_{\mathcal{G}}(\mathbf{D}) \setminus \mathbf{D})$ , where set  $\mathbf{D} \subseteq \mathbf{V}$  is *intrinsic* in  $\mathcal{G}(\mathbf{V})$ . Kernels  $q_{\mathbf{D}}(\mathbf{D} \mid \text{pa}_{\mathcal{G}}(\mathbf{D}) \setminus \mathbf{D})$  are objects that resemble conditional densities  $p(V \mid \text{pa}_{\mathcal{G}}(V))$  that arise in the Markov factorization for a DAG; kernels are non-negative and normalize to 1 for every value of  $\text{pa}_{\mathcal{G}}(\mathbf{D}) \setminus \mathbf{D}$ . A set  $\mathbf{S}$  is intrinsic in  $\mathcal{G}(\mathbf{V})$  if  $p(\mathbf{S} \mid \text{do}(\text{pa}(\mathbf{S}) \setminus \mathbf{S}))$  is identified.

The nested Markov factorization asserts that the observed margin  $p(\mathbf{V})$  can be expressed as a product  $\prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}(\mathbf{V}))} q_{\mathbf{D}}(\mathbf{D} \mid \text{pa}_{\mathcal{G}}(\mathbf{D}) \setminus \mathbf{D})$  of kernels where  $\mathcal{D}(\mathcal{G}(\mathbf{V}))$  is the set of bidirected connected components, called *districts*, in  $\mathcal{G}(\mathbf{V})$ . The factorization implies certain other kernels associated with *reachable sets* may be expressed as similar products of intrinsic kernels. Finally, the modified form of the factorization may be used to express *any* interventional distribution identified from  $p(\mathbf{V})$ .

Given a latent projection ADMG  $\mathcal{G}(\mathbf{V})$  representing a hidden variable causal model, and any disjoint subsets  $\mathbf{Y}, \mathbf{A}$  of  $\mathbf{V}$ , let  $\mathbf{Y}^*$  be the set of ancestors of  $\mathbf{Y}$  in  $\mathcal{G}(\mathbf{V})$  via directed paths that do not pass through  $\mathbf{A}$ , and let  $\mathcal{G}_{\mathbf{Y}^*}$  be the *induced subgraph* of  $\mathcal{G}(\mathbf{V})$  containing only vertices in  $\mathbf{Y}^*$  and edges among these vertices. [Shpitser and Pearl, 2006, Richardson et al., 2017] showed that any interventional distribution  $p(\mathbf{Y}(\mathbf{a}))$  is identified from  $p(\mathbf{V})$  given  $\mathcal{G}(\mathbf{V})$  if and only if every bidirected connected component in  $\mathcal{G}_{\mathbf{Y}^*}$  is intrinsic. Moreover, if  $p(\mathbf{Y}(\mathbf{a}))$  is identified, it is given by the following margin of the modified nested Markov factorization, made up of the appropriate kernels:

$$p(\mathbf{Y}(\mathbf{a})) = \sum_{\mathbf{Y}^* \setminus (\mathbf{Y} \cup \mathbf{A})} \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}_{\mathbf{Y}^*})} q_{\mathbf{D}}(\mathbf{D} | \text{pa}_{\mathcal{G}}(\mathbf{D}) \setminus \mathbf{D}) |_{\mathbf{A}=\mathbf{a}}. \quad (2.5)$$

Reverting to the example in Fig. 2.3(b), the ADMG has intrinsic sets  $\{A\}$ ,  $\{M\}$ ,  $\{A, Y\}$ , and  $\{Y\}$ , with the corresponding kernels:  $q_A(A) \equiv p(A)$ ,  $q_M(M|A) \equiv p(M|A)$ ,  $q_{A,Y}(A, Y|M) \equiv p(Y|M, A)p(A)$ , and  $q_Y(Y|M) \equiv \sum_A p(Y|M, A)p(A)$ .

The observed margin  $p(A, M, Y)$  factorizes as  $q_{A,Y}(A, Y|M)q_M(M|A)$ , via the nested Markov factorization. Further,  $p(Y(a))$  is identified from  $p(A, M, Y)$  and equal to  $\sum_M q_Y(Y|M)q_M(M|a) = \sum_M (\sum_{A'} p(Y|M, A')p(A'))p(M|a)$ , which is the *front-door formula* [Pearl, 1995].

A conditional ADMG (CADMG)  $\mathcal{G}(\mathbf{V}, \mathbf{W})$  can be obtained as a latent projection of a CDAG  $\mathcal{G}(\mathbf{V} \cup \mathbf{H}, \mathbf{W})$ , and generalizes an ADMG with observed context  $\mathbf{W}$ , in the same way that a CDAG generalizes a DAG. A nested Markov factorization of the marginal distribution  $p(\mathbf{V}|\mathbf{W})$  can be defined directly on the latent projection CADMG  $\mathcal{G}(\mathbf{V}, \mathbf{W})$ . In addition,  $p(\mathbf{Y}(\mathbf{a})|\mathbf{W})$  identified in a hidden variable causal model represented by  $\mathcal{G}(\mathbf{V} \cup \mathbf{H}, \mathbf{W})$  is always equal to a modified version of a nested factorization [Richardson et al., 2017] associated with  $\mathcal{G}(\mathbf{V}, \mathbf{W})$ . See Appendix I for details on the nested Markov factorization, reachable and intrinsic sets, and

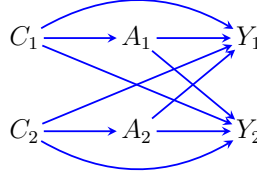
identification theory in ADMGs and CAMDGs.

Naturally, our next step would be to discuss temporal models with hidden variables and related identification results as we did with fully observed models, but we reserve this discussion for Chapter 3, specifically Section 3.2.1, as our novel work on PDSEMs is closely interlinked to a reformulation of existing results on this topic. Instead, we will proceed to discuss *interference*.

## 2.5 Interference

Historically, causal inference approaches have made the assumption that all units are independent and identically distributed (i.i.d.). This assumption might be reasonable where interactions between units are negligible. However, quite often, human beings influence one another through their social networks (online and offline), and we say that *interference* is present when one unit may causally affect other units [Aronow and Samii, 2013, Athey et al., 2018, Basse et al., 2019, Basse and Airolidi, 2018, Bowers et al., 2013, Cai et al., 2019, Eck et al., 2022, Eckles et al., 2017, Forastiere et al., 2021, Graham et al., 2010, Halloran and Struchiner, 1995, Halloran and Hudgens, 2012, Hong and Raudenbush, 2006, Hudgens and Halloran, 2008b, Jagadeesan et al., 2020, Toulis and Kao, 2013, Leung, 2020, Liu and Hudgens, 2014, Papadogeorgou et al., 2019, Puelz et al., 2019, Rosenbaum, 2007, Rubin, 1990, Sävje, 2021, Sävje et al., 2021, Sobel, 2006, Tchetgen and VanderWeele, 2012, Toulis et al., 2018, VanderWeele, 2010]. In some works, interference is defined to be the causal effect of one unit’s treatment on another’s outcome; but we use the term more generally, to mean interactions between any types of variables: covariates, treatments, or outcomes, and when applicable, variables to indicate the presence or absence of a potentially missing variable in the data records (commonly called *missingness indicators*; see Section 2.6).

In this dissertation, we will assume *partial interference*, i.e. interference occurs within



**Figure 2.4.** Example of a DAG illustration for interference between two units.

*blocks* of units of fixed finite size, but the blocks themselves are i.i.d. All of our identification results, however, can be extended to *full interference* settings where all units interact with one another (and the effective sample size is 1), or to non-iid blocks. In the setting of i.i.d. blocks of fixed (and identical) size, we can switch our lens to treat each *block* as a unit for estimation.

Let there be  $b$  blocks, each block consisting of  $m = n/b$  units, where  $n$  is the total number of units in the network. Let  $\mathbf{A} \equiv (A_1, \dots, A_m)$  be the vector of treatment assignments for units  $i = 1, \dots, m$ , and the  $m$ -dimensional vector  $\mathbf{a}$  be a realization in the support of  $\mathbf{A}$ . Similarly, let  $\mathbf{Y} \equiv (Y_1, \dots, Y_m)$  and  $\mathbf{C} \equiv (\mathbf{C}_1, \dots, \mathbf{C}_m)$  be the vector of outcomes and covariates respectively. We define  $Y_i(\mathbf{a})$  to be the counterfactual outcome of unit  $i$ , where the treatment vector  $\mathbf{A}$  is intervened on and set to  $\mathbf{a}$ ; we must index the counterfactual outcome of unit  $i$  by interventions not only performed on unit  $i$ , but also on other units that belong to the same block because of the interactions. Compare this to the counterfactual  $Y(a)$  in Eqn. (2.1) for a setting with no interference. Here  $A = a$  represents only the unit's own treatment. Parameters of interest in interference problems and related estimation strategies are described in detail in [Ogburn and VanderWeele, 2014, Hudgens and Halloran, 2008a, Tchetgen Tchetgen and VanderWeele, 2012, Tchetgen et al., 2017].

Ogburn and VanderWeele [2014] proposed causal diagrams for interference, an extension of causal DAGs to interference problems. Fig. 2.4 is a typical example of how DAGs are used to represent interference, in this case in a block of 2 units (commonly referred to as a *dyad*). In this DAG, the tuple  $(C_1, A_1, Y_1)$  corresponds to variables of

unit 1 and the tuple  $(C_2, A_2, Y_2)$  corresponds to variables of unit 2. The presence of edges between these two tuples encodes the causal influence of one unit’s variables on another unit’s variables. There are alternate graphical representations of interference, including those using *chain graph models* [Bhattacharya et al., 2019a, Tchetgen Tchetgen et al., 2017], but we will not discuss them in this work.

The principles of graphical causal models and identification generalize from the i.i.d. setting to settings with interference, with the only difference being that the graph now represents an entire block of units rather than a single unit representing an i.i.d. realization.

## 2.6 Missing Data

Missing data is a perennial problem in data analyses of all types, and may arise due to dropout from studies, loss to followup, imperfect data collection, survey non-response, among other reasons. Systematically missing data records can substantially bias subsequent analyses if not properly addressed. In this section, we discuss missing data models developed for i.i.d. settings.

A missing data model encodes assumptions about how missingness arises and how it relates to the underlying variables. The model is a set of distributions defined over a set of random variables  $\{\mathbf{O}, \mathbf{Z}^{(1)}, \mathbf{R}, \mathbf{Z}\}$ , where  $\mathbf{O}$  denotes the set of variables that are always observed,  $\mathbf{Z}^{(1)}$  denotes the set of variables that are potentially missing,  $\mathbf{R}$  denotes the set of missingness indicators of the variables in  $\mathbf{Z}^{(1)}$ , and  $\mathbf{Z}$  denotes the set of *observed proxies* of the variables in  $\mathbf{Z}^{(1)}$ . Given  $Z^{(1)} \in \mathbf{Z}^{(1)}$  and its corresponding missingness indicator  $R_Z \in \mathbf{R}$ , the observed proxy  $Z$  is deterministically defined by the following:

$$Z = \begin{cases} Z^{(1)} & \text{if } R_Z = 1, \\ ? & \text{if } R_Z = 0. \end{cases} \quad (2.6)$$

An interesting way to approach missing data is using the lens of causal models: each

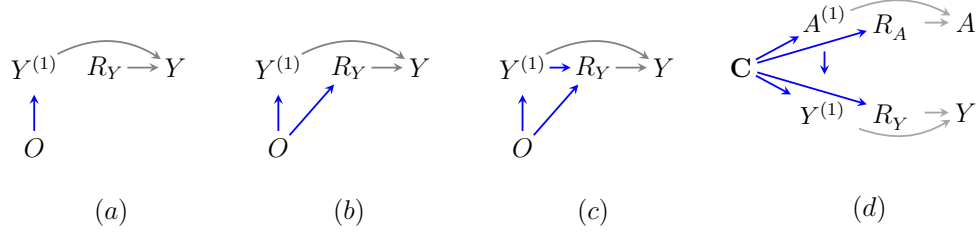
missingness indicator  $R \in \mathbf{R}$  may be viewed as a treatment variable that can be intervened on, and each  $Z \in \mathbf{Z}$  as an observed outcome. Thus,  $Z^{(1)}$  is a counterfactual random variable had we, possibly contrary to fact, intervened and set the corresponding missingness indicator  $R_Z$  to 1. This makes (2.6) the missing data equivalent of the consistency assumption in causal inference. Note, however, that the missingness indicator does not operate entirely identical to a treatment variable in that we do not see  $R_Z = 0$  as the treatment variable being set to 0. That is, in the missing data case, when  $R_Z = 0$ ,  $Z$  is not observed, and hence its value is set, deterministically, to “?”, unlike when the outcome for treatment being set to 0 typically has a well-defined, non-deterministic value in the support of  $Z$ .

The distribution  $p(\mathbf{O}, \mathbf{Z}^{(1)}, \mathbf{R})$  is called the full law, the distribution  $p(\mathbf{O}, \mathbf{Z}^{(1)})$  the target law, the distribution  $p(\mathbf{O}, \mathbf{Z}, \mathbf{R})$  the observed data law or simply the observed law, and the conditional distribution  $p(\mathbf{R} \mid \mathbf{Z}^{(1)}, \mathbf{O})$  the missingness process, or the missing data propensity score.

Typically we are interested in a functional of the target law, often of the form  $\mathbb{E}[h(\mathbf{O}, \mathbf{Z}^{(1)})]$ , and the goal of missing data methods is to identify such functionals in terms of the observed data law. Just like with causal inference, it is common to assume the missingness consistency assumption (2.6) and a missingness positivity assumption:

$$p(\mathbf{R} = 1 \mid \mathbf{Z}^{(1)}, \mathbf{O}) > 0. \quad (2.7)$$

For nonparametric identification of the full law and therefore of any functionals thereof, other assumptions, in addition to consistency and positivity, are necessary. Such assumptions can be encoded in a DAG  $\mathcal{G}(\mathbf{V})$ , where vertices  $\mathbf{V}$  correspond to random variables in  $\mathbf{O} \cup \mathbf{Z}^{(1)} \cup \mathbf{R} \cup \mathbf{Z}$ , and certain additional restrictions are placed on  $\mathcal{G}(\mathbf{V})$ : (i) each  $Z \in \mathbf{Z}$  has only two parents,  $R_Z$  and  $Z^{(1)}$ , and (ii) variables in  $\mathbf{R}$  cannot point to variables in  $\mathbf{O} \cup \mathbf{Z}^{(1)}$ . Restriction (i) is imposed by definition of  $Z$  via (2.6). In



**Figure 2.5.** A missing data DAG example corresponding to (a) a MCAR model, (b) a MAR model, and (c) a MNAR model, all with  $Z \equiv \{Y\}$ , and (d) A missing data MAR model with  $Z \equiv \{A, Y\}$ .

order to distinguish the deterministic relations implied by (2.6) from probabilistic relations, we draw edges pointing into  $Z \in \mathbf{Z}$  in gray. Restriction (ii) is imposed to ensure that, while changes in  $R_Z$  cause changes in the observed variables  $Z \in \mathbf{Z}$ , they do not result in changes to the underlying full data  $Z^{(1)} \in \mathbf{Z}^{(1)}$ .

A missing data DAG model is a set of distributions defined over variables in  $\mathbf{O} \cup \mathbf{Z}^{(1)} \cup \mathbf{R} \cup \mathbf{Z}$  that factorize with respect to a DAG obeying the above restrictions, as follows:

$$p(\mathbf{O}, \mathbf{Z}^{(1)}, \mathbf{R}, \mathbf{Z}) = \prod_{V \in \mathbf{O} \cup \mathbf{Z}^{(1)} \cup \mathbf{R}} p(V \mid \text{pa}_{\mathcal{G}}(V)) \times \prod_{Z \in \mathbf{Z}} p(Z \mid R_Z, Z^{(1)}). \quad (2.8)$$

Examples of missing data DAGs are shown in Fig. 2.5: Fig. 2.5(a) corresponds to a *missing completely at random* (MCAR) model where  $R_Y \perp\!\!\!\perp O, Y^{(1)}$ , Fig. 2.5(b) corresponds to a *missing at random* (MAR) model where  $R_Y \perp\!\!\!\perp Y^{(1)} \mid O$ , and Fig. 2.5(c) corresponds to a *missing not at random* (MNAR) model where neither independence holds.

If a missing data model contains hidden variables  $\mathbf{H}$ , it may be represented by the latent projection ADMG  $\mathcal{G}(\mathbf{O}, \mathbf{Z}, \mathbf{Z}^{(1)}, \mathbf{R})$  of the DAG  $\mathcal{G}(\mathbf{H}, \mathbf{O}, \mathbf{Z}, \mathbf{Z}^{(1)}, \mathbf{R})$ . In such models, the target of inference is some function of the margin of the full data law  $p(\mathbf{O}, \mathbf{Z}^{(1)}, \mathbf{R})$  where  $\mathbf{H}$  is marginalized out.

It is worth pointing out that there are many non-graphical approaches to dealing with missing data as well. A simple and popular approach is to use only fully observed

rows in a *complete-case* analysis, justifiable only if the underlying mechanism is MCAR [Rubin, 1976]. For MAR scenarios, many approaches including expectation maximization [Dempster et al., 1977, Horton and Laird, 1999, Little and Rubin, 2002], multiple imputation [Rubin, 1988, Schafer, 1999], inverse probability weighting [Robins et al., 1994a, Li et al., 2013] have been developed. For MNAR problems, often parametric or semiparametric restrictions have been imposed on the underlying data distribution and missingness selection model, such that they yield identification [Little and Rubin, 2002, Tchetgen Tchetgen et al., 2016, Wu and Carroll, 1988, Wang et al., 2014, Miao et al., 2016, Miao and Tchetgen Tchetgen, 2016, Sun et al., 2018]. In this dissertation, we restrict ourselves to graphical modeling approaches and non-parametric identification.

The full law in a missing data model is identified *if and only if* the missingness mechanism is identified by some functional of the observed data law  $g(p(\mathbf{R}, \mathbf{O}, \mathbf{Z}))$ , because

$$p(\mathbf{O}, \mathbf{Z}^{(1)}, \mathbf{R}) = p(\mathbf{O}, \mathbf{Z}^{(1)}) \times \underbrace{p(\mathbf{R} \mid \mathbf{O}, \mathbf{Z}^{(1)})}_{g(p(\mathbf{R}, \mathbf{O}, \mathbf{Z}))} = \underbrace{\frac{p(\mathbf{O}, \mathbf{Z}^{(1)}, \mathbf{R} = 1)}{p(\mathbf{R} = 1 \mid \mathbf{O}, \mathbf{Z}^{(1)})}}_{g(p(\mathbf{R}, \mathbf{O}, \mathbf{Z}))|_{\mathbf{R}=1}} \times \underbrace{p(\mathbf{R} \mid \mathbf{O}, \mathbf{Z}^{(1)})}_{g(p(\mathbf{R}, \mathbf{O}, \mathbf{Z}))}. \quad (2.9)$$

For instance, the full law in the MAR missing data DAG model example in Fig. 2.5(d) is identified because the missingness mechanism  $p(R_A, R_Y \mid \mathbf{C}, A^{(1)}, Y^{(1)}) = p(R_A, R_Y \mid \mathbf{C})$  is a function of observed data. The hierarchy of MCAR, MAR and MNAR mechanisms holds relevance in the context of identification simply due to the fact that if missingness is not at random, the full law may not be identified. We discuss non-parametric identification in missing data models at length in Chapter 4.

## 2.7 Chain Graph Models

*Chain graphs* (CGs) are a class of mixed graphs containing directed and undirected (–) edges with no partially directed cycles, i.e., it is not possible to create a directed

cycle by orienting any of the undirected edges [Lauritzen, 1996]. A chain graph with no undirected edges is simply a DAG. We denote a chain graph  $\mathcal{G}^c(\mathbf{V})$  with vertices  $\mathbf{V}$  by the shorthand  $\mathcal{G}^c$ . For  $\mathbf{B} \subseteq \mathbf{V}$ , we define a *block*<sup>1</sup>  $\mathbf{B}$  to be the maximal set of vertices, where every vertex pair in the subgraph  $\mathcal{G}_{\mathbf{B}}$  is connected by an undirected path; the subgraph  $\mathcal{G}_{\mathbf{B}}$  is defined as a graph with a vertex set  $\mathbf{B}$  with all edges in  $\mathcal{G}^c$  between elements in  $\mathbf{B}$ , and an undirected path is a path that contains only undirected edges. The set  $\mathcal{B}(\mathcal{G}^c)$  denotes the set of all blocks  $\mathbf{B}$  in  $\mathcal{G}^c$ .

A CG model is a set of distributions associated with a CG  $\mathcal{G}^c$  that can be written in terms of a two-level factorization, the first associated with all blocks in  $\mathcal{B}$ , and the second corresponding to elements in each block  $\mathbf{B}$ . For a more detailed account on chain graphs, their Markov properties and the factorization, see [Lauritzen and Richardson, 2002].

Work in [Lauritzen and Richardson, 2002] justifies the use of a chain graph (under the Lauritzen-Wermuth-Freydenburg (LWF) interpretation) for systems where the underlying data is obtained via equilibrium-generating dynamics. In this context, a CG is defined via a combination of structural equation semantics and Gibbs sampling, the latter enabling equilibrium dynamics. In particular, in a causal CG  $\mathcal{G}^c$ , the distribution  $p(\mathbf{B} \mid \text{pa}_{\mathcal{G}^c}(\mathbf{B}))$  for each block  $\mathbf{B}$  is determined via a Gibbs sampler on variables  $B \in \mathbf{B}$ , where the distribution  $p(B \mid \mathbf{B} \setminus \{B\}, \text{pa}_{\mathcal{G}^c}(\mathbf{B}))$  is determined via a structural equation  $f_B(\mathbf{B} \setminus \{B\}, \text{pa}_{\mathcal{G}^c}(\mathbf{B}), \epsilon_{\mathbf{B}})$ . The intervention  $\text{do}(b)$  that sets  $B$  to  $b$  replaces the structural equation for  $B$  by the assignment  $b$ . In a causal CG model, for any disjoint  $\mathbf{Y}, \mathbf{A}$ , the distribution  $p(\mathbf{Y}(\mathbf{a}))$  is identified by the CG version of the g-formula [Lauritzen and Richardson, 2002]:

---

<sup>1</sup>This usage of block is not to be confused with the usage of block in the context of interference. We have remained faithful to the definitions of these entities in existing literature, and will clarify the usage of the word as it appears in our text.

$$p(\mathbf{Y}(\mathbf{a})) = \prod_{\mathbf{B} \in \mathcal{B}(\mathcal{G}^c)} \mathbf{p}(\mathbf{B} \setminus \mathbf{A} \mid \text{pa}(\mathbf{B}, \mathbf{B} \cap \mathbf{A}))|_{\mathbf{A}=\mathbf{a}} \quad (2.10)$$

With this, we have completed an overview of the preliminaries required to follow our work documented in Chapters 3, 4 and 5.

# Chapter 3

## Path Dependent Models

### 3.1 Introduction

Causal dynamic Bayesian networks [Blondel et al., 2017], discussed in Section 2.3, are a discrete-time generalization of causal DAGs, and can model causal relationships in temporal processes evolving in discrete time. However, these models have generally been used in settings where the causal structure remains invariant over time. For example, analysis of the impact of anti-retroviral therapy on HIV infection progression assumed the same variables relevant for the patient health and the same causal relationships linking them at each time point in the study [Hernán et al., 2000]. Changes tracked over time (such as HIV developing resistance to the current drug) are thus *quantitative*, with the underlying causal structure remaining unchanged over time. However, many systems undergo *qualitative* changes as well, where observability, relevance, and causal relationships of variables vary over time. Furthermore, interventions in such systems might alter the downstream evolution of the system to be different from that observed in data. We call systems that exhibit these characteristics *path-dependent*, inspired by the economics literature [Liebowitz and Margolis, 2002].

A motivating example that we use in this work [Srinivasan et al., 2021] is that of a septoplasty surgery, a procedure performed on the nasal cartilage (or septum) to relieve

nasal obstruction. The surgery consists of (atleast) five distinct phases: (1) opening of the septum, (2) raising septal flaps, (3) removal of the deviated septal cartilage and bone, (4) reconstruction, and (5) closing of the septum. Each stage is associated with a distinct set of variables and relationships among them that may not be shared across stages: contrast a skill-intensive task like cartilage repair against a routine incision closure which could potentially be automated, as an example. Further, procedures performed at a particular stage can go wrong, forcing surgeons to “double back” to correct mistakes or deal with complications. Surgeon experience often determines how likely it is that previous stages of the surgery need to be revisited. These features make a (septoplasty) surgery a path-dependent system.

The goal of causal inference in this setting is to help assign surgeons to perform different stages of the surgery while navigating the tradeoff between the need to train resident surgeons on the one hand, and operating costs and patient safety on the other. Addressing this tradeoff entails using retrospective data to estimate outcomes of surgery trajectories that *differ from those actually observed* due to counterfactually different choices of surgeon assignment in past stages of the surgery. Other examples where path dependence may naturally arise include life course studies examining economic disparities in society or patient outcomes in hospitals using Electronic Health Record (EHR) data.

### 3.1.1 Contributions

Our contributions to the causal inference literature are as follows. We introduce the *path-dependent structural equation model (PDSEM)* for causal systems that exhibit qualitative changes over time, observed or unobserved confounding, and path-dependence on counterfactual choices in the past. PDSEMs generalize causal dynamic Bayesian networks by allowing complex and looping stage transitions between distinct yet tractable causal models, and generalize Markov decision processes used in reinforce-

ment learning [Sutton and Barto, 2018, Zhang and Bareinboim, 2016] by representing each state as a graphical causal model that allows confounding between actions and outcomes. We give a complete identification theory for our model. In particular, in the special case where the PDSEM is first order Markov, all identification queries may be decomposed into queries pertaining to observed transition probabilities between states, a generalization of results for causal DBNs in [Blondel et al., 2017]. Finally, we show how statistical inference may be performed by a combination of plug-in estimation and Monte Carlo sampling [Bickel and Doksum, 2015], generalizing similar schemes developed for longitudinal causal models [Westreich et al., 2012].

The roadmap to this chapter is as follows. In Section 3.2, we discuss causal DBNs in literature, and in Section 3.2.1 we reformulate some existing results to generalize them later. In Section 3.3, we present a simple example of a PDSEM for the reader to get a quick glimpse of what constitutes the model, followed by a more rigorous treatment of PDSEMs for fully observed data and then with hidden variables in Sections 3.4 and 3.5, respectively. Finally, we give a quick summary of our experimental results in Section 3.7. For a more thorough treatment of the content, please refer to the author’s work in [Srinivasan et al., 2021].

## 3.2 Background

Causal DBNs, the model we build on and generalize, have been considered in prior work. [Peters et al., 2013] illustrated how structural equations can be used in the context of time series data, addressing issues of identifiability. [Malinsky and Spirtes, 2018, 2019, Mogensen et al., 2018] presented structure learning algorithms for causal dynamic networks and applied them to macroeconomic data. [Blondel et al., 2017] developed an identification algorithm and transportability results for dynamic causal networks.

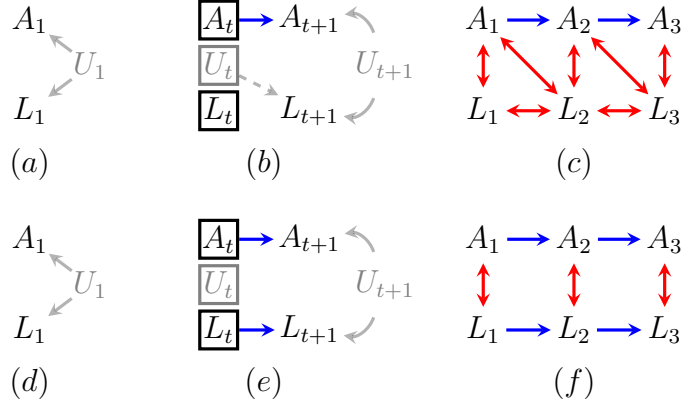
### 3.2.1 Identification in Causal DBNs with Hidden Variables

[Blondel et al., 2017] showed how identification in hidden variable causal DBNs may be decomposed into a set of independent problems, pertaining to conditional state transition distributions. We reformulate these results using the language of nested Markov models (see Section 2.4.1) to facilitate identification theory and statistical inference in PDSEMs. We start with an assumption that allows us to view the marginal version of a DBN, defined only on observed variables, as a first-order Markov DBN.

**Assumption 1.** *Transition network  $\mathcal{G}_{+1}$  between time points  $t$  and  $t + 1$  only depends on fixed variables in the previous time step  $t$  that are observed.*

If  $\mathcal{G}_{+1}$  depends on fixed variables that are hidden, the resulting DBN may result in observed variables in step  $t + 1$  depending on observed variables earlier than  $t$  even if observed variables in  $t$  are conditioned on, resulting in a model that is not first order Markov. For example, consider the DBN specified by prior and transition networks in Fig. 3.1 (a) and (b). Because  $L_{t+1}$  depends on unobserved  $U_t$ , and  $U_t$  influences  $L_t$ , “unrolling” this network, and taking the latent projection yields an ADMG shown in Fig. 3.1 (c), where  $L_3$  ends up being dependent on  $L_1$ , even after conditioning on  $L_2$  and  $A_2$  (due to the “explaining away” phenomenon arising when a shared effect  $L_2$  of two variables  $U_2$  and  $U_1$  is conditioned on). On the other hand, the DBN specified by prior and transition networks in Fig. 3.1 (d) and (e) does not suffer from this issue, as the transition network only depends on observed variables  $L_t, A_t$ , yielding a latent projection of the “unrolled” model shown in Fig. 3.1 (f), which factorizes into time step specific conditional distributions:  $p(A_1, L_1)p(A_2, L_2|A_1, L_1)p(A_3, L_3|A_2, L_2)$ .

In general, given a hidden variable prior network  $\mathcal{G}_1$  on  $\mathbf{V}_1, \mathbf{H}_1$ , and transition network  $\mathcal{G}_{+1}$  on  $\mathbf{V}_{t+1}, \mathbf{H}_{t+1}$  given  $\mathbf{V}_t$ , the hidden variable DBN may be represented by latent projections of the prior and transition networks: an ADMG  $\mathcal{G}_1$  on  $\mathbf{V}_1$ , and



**Figure 3.1.** (a),(d) Prior network hidden variable DAGs  $\mathcal{G}_1$ , representing the state at time  $t = 1$ . (b),(e) Conditional hidden variable DAGs  $\mathcal{G}_{+1}$  representing the transitions in the network, with (e) leading to a first-order Markov model, and (b) leading to higher order dependences to unobserved hidden variables  $U_t$  linking multiple time points. (c),(f) Latent projection ADMGs of the unrolled hidden variable DBNs to three time steps.

a CADMG  $\mathcal{G}_{+1}$  on  $\mathbf{V}_{t+1}$  given  $\mathbf{V}_t$ , and the corresponding marginal distributions  $p(\mathbf{V}_1)$  and  $p(\mathbf{V}_{t+1,t}|\mathbf{V}_t)$ . The “unrolled” version of the factorization of this model is:  $p(\mathbf{V}_1) \prod_{t=1}^T p(\mathbf{V}_{t+1,t}|\mathbf{V}_t)$ , where each term nested Markov factorizes with respect to either  $\mathcal{G}_1$  or  $\mathcal{G}_{+1}$  by results in [Richardson et al., 2017]. If the underlying DAGs correspond to causal models, the hidden variable DBN yields identification theory where modified nested factorization (2.5) is applied at every time point, just as (2.1) was applied at every point in a fully observed causal DBN to yield (2.4).

Given a fixed set of time points  $1, \dots, T$ , vertices  $\mathbf{V}_{1:T} \equiv \mathbf{V}_1 \cup \mathbf{V}_2 \cup \dots \cup \mathbf{V}_T$ , and disjoint subsets  $\mathbf{A}, \mathbf{Y} \subseteq \mathbf{V}_{1:T}$ , we have the following generalization of results in [Blondel et al., 2017]:

*Lemma 1.* Under Assumption 1,  $p(\mathbf{Y}(\mathbf{a}))$  is identified from a hidden variable causal DBN model represented by latent projections  $\mathcal{G}_1$  on  $\mathbf{V}_1$  and  $\mathcal{G}_{+1}$  on  $\mathbf{V}_{t+1}$  given  $\mathbf{V}_t$  if and only if every bidirected connected component in  $\mathcal{G}_{1\mathbf{Y}_1^*}$  (the induced subgraph of  $\mathcal{G}_1$ ) is intrinsic in  $\mathcal{G}_1$ , and every bidirected component in  $\mathcal{G}_{+1\mathbf{Y}_i^*}$  (the induced subgraph of  $\mathcal{G}_{+1}$ ) is intrinsic in  $\mathcal{G}_{+1}$ , where  $\mathbf{Y}_1^*$  is the set of ancestors of  $\mathbf{Y} \cap \mathbf{V}_1$  not through  $\mathbf{A} \cap \mathbf{V}_1$  in  $\mathcal{G}_1$ , and for every  $i \in 2, \dots, T$ ,  $\mathbf{Y}_i^*$  is the set of ancestors of  $\mathbf{Y} \cap \mathbf{V}_i$  not

through  $\mathbf{A} \cap \mathbf{V}_i$  in  $\mathcal{G}_{+1}$ . Moreover, if  $p(\mathbf{Y}(\mathbf{a}))$  is identified, we have

$$\left( \sum_{\mathbf{Y}_1^* \setminus ((\mathbf{Y} \cup \mathbf{A}) \cap \mathbf{V}_1)} \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}_1 \mathbf{Y}_1^*)} q_{\mathbf{D}}^1(\mathbf{D} | \text{pa}_{\mathcal{G}}(\mathbf{D}) \setminus \mathbf{D}) |_{\mathbf{A}=\mathbf{a}} \right) \times \prod_{i=2}^T \left( \sum_{\mathbf{Y}_i^* \setminus ((\mathbf{Y} \cup \mathbf{A}) \cap \mathbf{V}_i)} \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}_{+1} \mathbf{Y}_i^*)} q_{\mathbf{D}}^{+1}(\mathbf{D} | \text{pa}_{\mathcal{G}}(\mathbf{D}) \setminus \mathbf{D}) |_{\mathbf{A}=\mathbf{a}} \right),$$

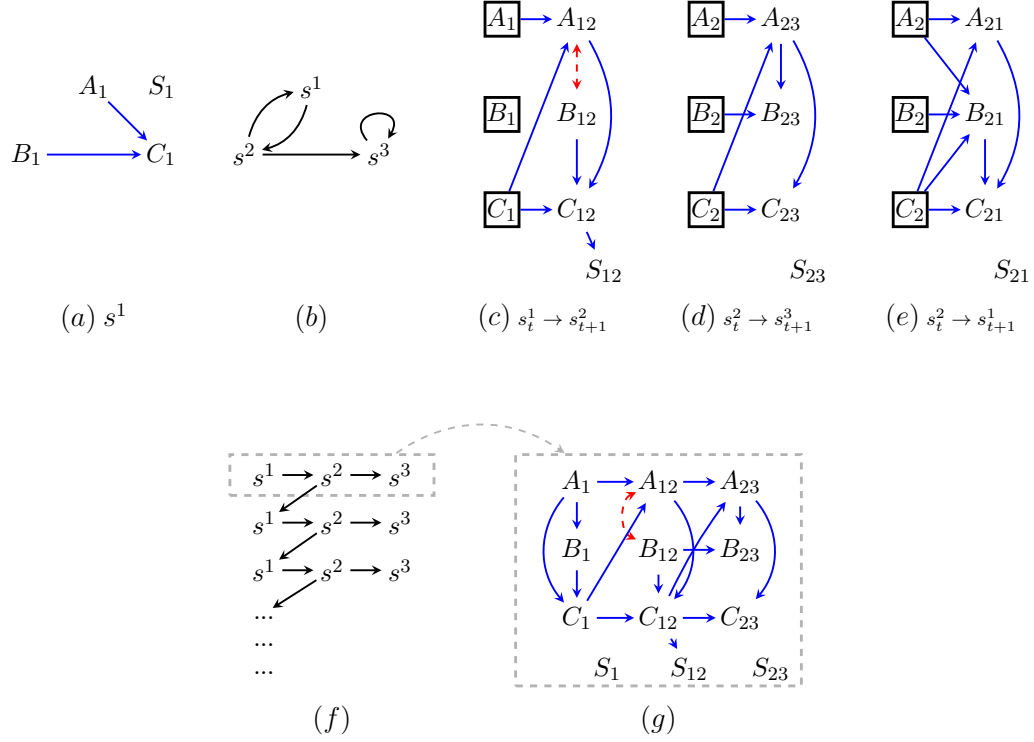
where  $q_{\mathbf{D}}^1$  and  $q_{\mathbf{D}}^{+1}$  are kernels corresponding to intrinsic sets that are districts in  $\mathcal{D}(\mathcal{G}_1 \mathbf{Y}_1^*)$  and  $\mathcal{D}(\mathcal{G}_{+1} \mathbf{Y}_1^*)$  in the nested Markov factorizations of  $\mathcal{G}_1$  and  $\mathcal{G}_{+1}$ , respectively.

This result, unlike in [Blondel et al., 2017], allows arbitrary sets of treatments in a DBN. The proof is in Appendix II. Results for systems with fixed context  $\mathbf{W}$  are discussed in [Srinivasan et al., 2021]. If Assumption 1 does not hold, causal effects in causal DBNs may still be identified for any finite  $T$  (see supplementary material of [Srinivasan et al., 2021]), Section 4.2 in [Blondel et al., 2017]. However, the resulting functional will likely be computationally intractable.

### 3.3 A Simple PDSEM

Let us use a simple illustrative example inspired by the surgery setting. We assume a surgery will consist of only three states:  $s^1$  (“incision”),  $s^2$  (“modification of tissue”), and  $s^3$  (“closing the incision”). Further, each state has the following variables:  $A$  (patient status prior to any procedures in the current stage),  $B$  (experience of surgeon performing the procedure in the current stage) and  $C$  (the observed patient outcome for the stage after procedure is performed), all observed. The surgery always starts at  $s^1$ , and concludes upon reaching  $s^3$ . Procedures performed in  $s^2$  may either succeed, leading to  $s^3$ , or fail with some probability, leading the surgeon to revisit  $s^1$ . The state transition diagram for this scenario is shown in Fig. 3.2 (b).

The causal diagram in Fig. 3.2 (a) shows relationships between variables in  $s^1$  and functions similar to the prior network in a causal DBN. In addition to variables



**Figure 3.2.** A simple PDSEM. (a) Causal structure of the initial state  $S^1$ . (b) The state transition diagram. (c),(d),(e) Causal diagrams representing possible transitions and subsequent states. (f) Causal relationships in a system evolving according to the state transitions:  $s^1 \rightarrow s^2 \rightarrow s^3$ . (g) A snapshot of a possible PDSEM trajectory that terminates in 3 timepoints is represented as an unrolled ADMG.

$A_1, B_1$  and  $C_1$ , it contains  $S_1$ , representing the state to transition to at time step 1. In our simple model, the state  $s^1$  transitions to  $s^2$  with probability 1, and so  $S_1$  represents a degenerate probability distribution and does not depend on any other variable. In general, however, the probability associated with  $S_1$  may depend on other variables in the current state. Transitions are specified by multiple causal CDAGs, one for every allowed state transition. These CDAGs are shown in Fig. 3.2(c),(d) and (e) (where dashed edges are ignored). These graphs include transition edges representing relationships between variables in the state at time  $t$  and variables in the state at time  $t + 1$ , and state-specific relationships among variables at time  $t + 1$ . We assume that the state spaces of variables associated with each state are the same across state transition and prior graphs. For example, the state spaces of  $A_1, B_1, C_1$  in

Fig. 3.2(a) and  $A_{21}, B_{21}, C_{21}$  in Fig. 3.2(c) are the same, but the variables themselves (and the causal graphs relating them) are not. This implies values may be indexed by state, e.g.  $a_1$  can refer without loss of generality to a value of  $A_1$  or  $A_{21}$ . Similarly, conditional distributions that depend on variables in a prior state are well-defined if those variables are indexed by the prior state only, e.g.  $p(A_{12}|A_1)$  is a shorthand for “a density over  $A_{12}$  in transition (1, 2) given any value  $a_1$  of any variable of the form  $A_{i1}$ .” Causal graphs in Fig. 3.2(a),(c),(d),(e), along with the state-transition diagram Fig. 3.2(b), *completely* describe the fully observed PDSEM. Complex state dynamics are captured by distinct state causal DAGs and path-dependence is a consequence of state transitions that may depend on variables in the current state, and not just the state itself.

This example describes a randomized controlled trial where the surgeon operating during state  $s^2$  is randomly assigned, hence  $B_{12}$  in the transition graph in Fig. 3.2(c) has no parents. Otherwise, we encode standard causal relationships we expect:  $C$  in the previous state influences  $A, C$  in the next, and  $A$  in the previous state influences  $A$  in the next. Surgeon assignment  $B_{12}$  in  $s^2$  influences assignments in subsequent stages, whether they are  $s^1$  or  $s^3$ . The state transition at  $s^2$  depends on the outcome  $C$  at that state. In  $s^3$ ,  $B$  does not influence  $C$ , since closing the incision is a task adequately performed independent of surgeon experience. The observed data factorization of a fully-observed PDSEM is not finite, but yields a well defined joint distribution  $p_\infty$  over possible trajectories shown schematically in Fig. 3.2(f):

$$\begin{aligned}
& p_1 \prod_{t=1}^{\infty} (p_{12})^{\mathbb{I}(s_t^1, s_{t+1}^2)} (p_{23})^{\mathbb{I}(s_t^2, s_{t+1}^3)} (p_{21})^{\mathbb{I}(s_t^2, s_{t+1}^1)} 1^{\mathbb{I}(s_t^3)} \\
& p_1 = p(A_1)p(B_1|A_1)p(C_1|A_1, B_1)\tilde{p}(S_1) \\
& p_{12} = p(A_{12}|A_1, C_1)p(B_{12})p(C_{12}|B_{12}, A_{12}, C_1)p(S_{12}|C_{12}) \\
& p_{23} = p(A_{23}|A_2, C_2)p(B_{23}|B_2, A_{23})p(C_{23}|A_{23}, C_2)\tilde{p}(S_{23}) \\
& p_{21} = p(A_{21}|A_2, C_2)p(B_{21}|B_2, A_2, C_2)p(C_{21}|C_2, B_{21}, A_{21})\tilde{p}(S_{21}),
\end{aligned}$$

where  $s_t^i$  is the event “the state at time  $t$  is  $s^i$ ”, and all  $\tilde{p}$  are deterministic by definition

of our model.

PDSEMs allow us to reason about counterfactual questions such as: “what would happen if all procedures are performed by the resident surgeon ( $B = b$ ), possibly contrary to fact?”. The counterfactual joint distribution  $p_\infty(b)$  is obtained by standard structural equation replacement semantics [Pearl, 2009], on the state-specific marginal and conditional counterfactual distributions:

$$p_1(b) \prod_{t=1}^{\infty} (p_{12}(b))^{\mathbb{I}(s_t^1, s_{t+1}^2)} (p_{23}(b))^{\mathbb{I}(s_t^2, s_{t+1}^3)} (p_{21}(b))^{\mathbb{I}(s_t^2, s_{t+1}^1)} \mathbb{1}^{\mathbb{I}(s_t^3)},$$

which is identified by using the g-formula for every component of the factorization, in a generalization of (2.4), yielding:

$$\begin{aligned} p_0^* & \prod_{t=1}^{\infty} (p_{12}^*)^{\mathbb{I}(s_t^1, s_{t+1}^2)} (p_{23}^*)^{\mathbb{I}(s_t^2, s_{t+1}^3)} (p_{21}^*)^{\mathbb{I}(s_t^2, s_{t+1}^1)} \mathbb{1}^{\mathbb{I}(s_t^3)} \\ p_1^* &= p(A_1)p(C_1|A_1, b)\tilde{p}(S_1) \\ p_{12}^* &= p(A_{12}|A_1, C_1)p(C_{12}|b, A_{12}, C_1)p(S_{12}|C_{12}) \\ p_{23}^* &= p(A_{23}|A_2, C_2)p(C_{23}|A_{23}, C_2)\tilde{p}(S_{23}) \\ p_{21} &= p(A_{21}|A_2, C_2)p(C_{21}|C_2, b, A_{21})\tilde{p}(S_{21}). \end{aligned}$$

While the distribution  $p(S_{12}|C_{12})$  remains the same, the probability that  $s^1$  is visited from  $s^2$  is likely higher in  $p_\infty(b)$  compared to  $p_\infty$ . This is because  $B_{12}$ , counterfactually set to  $b$ , causes  $C_{12}$ , and  $C_{12}$  causes  $S_{12}$ . Thus, PDSEMs encode counterfactually changing state transition probabilities from their observed values.

### 3.4 Fully Observed PDSEMs

An arbitrary PDSEM is defined using a set of states  $\mathbf{s}$ , with initial state  $s^1$ , an absorbing state  $s^*$ , a set  $\mathcal{T}$  of state index pairs of the form  $(i, j)$ , where  $s^i \neq s^*$  representing allowed state transitions, a DAG  $\mathcal{G}_1$  on  $\mathbf{V}_1$  for the initial state  $s^1$ , and for each  $(i, j) \in \mathcal{T}$ , a CDAG  $\mathcal{G}_{ij}$  on  $\mathbf{V}_{ij}$  given  $\mathbf{V}_i$ . Variables  $S_1 \in \mathbf{V}_1, \{S_{ij} \in \mathbf{V}_{ij} : (i, j) \in \mathcal{T}\}$  determine probabilities of transitioning from state to state. Just as in a causal DBN, the DAG  $\mathcal{G}_1$ , and CDAGs  $\mathcal{G}_{ij}$  represent structural equation models for the initial

state, and the appropriate state transitions, respectively. That is, in the initial state, each variable  $V \in \mathbf{V}_1$  is determined via  $f_V(\text{pa}_{\mathcal{G}}(V), \epsilon_V)$ . Similarly, for each variable  $V \in \mathbf{V}_{ij}$  in any state transition represented by  $\mathcal{G}_{ij}$ . We assume  $S_1, \{S_{ij} : (i, j) \in \mathcal{T}\}$  have no outgoing edges (this is without loss of generality, as structural equations are already state-specific in a PDSEM).

A first order Markov PDSEM obeys the following assumption that ensures that we need not condition on any context in the past except variables in the prior state.

**Assumption 2.** *For every state  $s^j$ , any CDAG  $\mathcal{G}_{ij}$  or DAG  $\mathcal{G}_j$  will have random variables that share state spaces.*

We thus denote the values of any  $\mathbf{V}_{ij}$  for any transition  $(i, j)$  into state  $j$  by  $\mathbf{v}_j$  (note the lack of dependence on  $i$ ). As in our example, we index conditional densities that depend on variables in a prior state by that state only, e.g.  $p(A_{12}|A_1)$ .

Define  $\mathbf{V} \equiv \mathbf{V}_1 \cup \left(\bigcup_{(i,j) \in \mathcal{T}} \mathbf{V}_{ij}\right)$ . A PDSEM yields an observed distribution  $p_\infty(\mathbf{V})$  with the factorization:

$$p_1(\mathbf{V}_1) \prod_{t=1}^{\infty} \left( \prod_{(i,j) \in \mathcal{T}} (p_{ij}(\mathbf{V}_{ij}|\mathbf{V}_i))^{\mathbb{I}(s_t^i, s_{t+1}^j)} \right) 1^{\mathbb{I}(s_t^*)} \quad (3.1)$$

where  $p_{ij}(\mathbf{V}_{ij}|\mathbf{V}_i) = \prod_{V \in \mathbf{V}_{ij}} p(V|\text{pa}_{\mathcal{G}_{ij}}(V))$  and  $p_1(\mathbf{V}_1) = \prod_{V \in \mathbf{V}_1} p(V|\text{pa}_{\mathcal{G}_1}(V))$ .

An intervention in a PDSEM is defined on a set of treatment variables  $\mathbf{A} \equiv \bigcup_{(i,j) \in \mathcal{T}} \mathbf{A}_{ij}$  and set to values  $\mathbf{a}$  with the property that for any  $(i, j), (k, j) \in \mathcal{T}$ , the same values  $\mathbf{a}_j$  are being set to  $\mathbf{A}_{ij}$  and  $\mathbf{A}_{kj}$ . Define  $\mathbf{Y}_{ij}$  in each transition graph  $\mathcal{G}_{ij}$  to be all variables in that state not in  $\mathbf{A}_{ij}$ , with their corresponding values being  $\mathbf{y}_j$ , their union being  $\mathbf{Y}$ , and the values of the union being  $\mathbf{y}$ .

A new counterfactual distribution  $p_\infty(\mathbf{Y}(\mathbf{a}))$  is obtained from the counterfactual initial state distribution  $p_1(\mathbf{Y}_1(\mathbf{a}_1))$ , and transition distributions  $p_{ij}(\mathbf{Y}_{ij}(\mathbf{a}_j)|\mathbf{Y}_i(\mathbf{a}_i))$  as:

$$p_1(\mathbf{Y}_1(\mathbf{a}_1)) \prod_{t=1}^{\infty} \left( \prod_{(i,j) \in \mathcal{T}} (p_{ij}(\mathbf{Y}_{ij}(\mathbf{a}_j) | \mathbf{Y}_i(\mathbf{a}_i)))^{\mathbb{I}(s_t^i, s_{t+1}^j)} \right) 1^{\mathbb{I}(s_t^*)}$$

Individual counterfactual distributions are obtained using standard structural equation replacement semantics. Since the initial state and transitions are defined using structural equations, we obtain the following identification result, which generalizes the DBN g-formula (2.4) to PDSEMs.

*Lemma 2.* Given a fully observed PDSEM, each factor of the distribution  $p_{\infty}(\mathbf{Y}(\mathbf{a}))$  is identified from  $p_{\infty}(\mathbf{V})$  as:

$$\begin{aligned} p_1(\mathbf{Y}_1(\mathbf{a}_1)) &\equiv \prod_{V \in \mathbf{Y}_1 \setminus \mathbf{A}_1} p_1(V | \text{pa}_{\mathcal{G}_1}(V)) \Big|_{\mathbf{A}_1 = \mathbf{a}_1} \\ p_{ij}(\mathbf{Y}_{ij}(\mathbf{a}_j) | \mathbf{Y}_i(\mathbf{a}_i)) &\equiv \prod_{V \in \mathbf{Y}_{ij} \setminus \mathbf{A}_j} p_{ij}(V | \text{pa}_{\mathcal{G}_{ij}}(V)) \Big|_{\substack{\mathbf{A}_i = \mathbf{a}_i, \\ \mathbf{A}_j = \mathbf{a}_j}} \end{aligned} \quad (3.2)$$

A PDSEM may be generalized from a first order to a  $k$ th-order Markov model, where variables in a particular state, can depend on variables in at most  $k$  prior states. This involves an appropriate generalization of Assumption 2, and specification of a larger set of transition networks. Details are in Appendix II.

It is worth noting that if all transition networks in a PDSEM obey a single consistent topological order, it is possible to encode a PDSEM simply by a causal DBN. Such an encoding will be inefficient and non-intuitive, however, since this causal DBN would represent restrictions of a PDSEM via context-specific independences in a large transition network representing a Cartesian product of possible transition networks of a PDSEM. If a consistent topological order on variables in transition networks does not exist, PDSEMs do not have a known causal DBN representation. For more details, refer to [Srinivasan et al., 2021] and Appendix II.

### 3.5 PDSEMs with Hidden Variables

To extend causal inference to latent variable PDSEMs, in addition to Assumption 1 and Assumption 2, we assume the probabilities of any state transition trajectories are observed.

**Assumption 3.** *The variables  $S_{ij}$  for any  $(i, j) \in \mathcal{T}$  governing state transition probabilities are observed.*

The latent variable PDSEMs then decompose into an initial state and a set of transitions such that causal inference results may be stated without loss of generality using latent projection ADMGs (and CADMGs) of appropriate DAGs and CDAGs. In addition, the fact that variables  $S_{ij}$  are observed implies we can evaluate counterfactual state transition probabilities, provided they are identified. Next we provide a formal definition of hidden variable PDSEMs.

Given the initial state DAG  $\mathcal{G}_1$  on  $\mathbf{V}_1, \mathbf{H}_1$  and the set of transition CDAGs  $\mathcal{G}_{ij}$  on  $\mathbf{V}_{ij}, \mathbf{H}_{ij}$  given  $\mathbf{V}_i$ , for all  $(i, j) \in \mathcal{T}$ , define a PDSEM such that:

1. the variables  $\mathbf{V} \equiv \{\mathbf{V}_1\} \cup \bigcup_{(i,j) \in \mathcal{T}} \mathbf{V}_{ij}$ , are observed and  $\mathbf{H} \equiv \{\mathbf{H}_1\} \cup \bigcup_{(i,j) \in \mathcal{T}} \mathbf{H}_{ij}$  are hidden,
2. all state transition variables are observed, that is,  $S_1 \in \mathbf{V}_1$ ,  $S_{ij} \in \mathbf{V}_{ij}$  for every  $(i, j) \in \mathcal{T}$ , and
3. every state has the same observed and hidden variables regardless of transition, or, for every  $j$  and all  $(i, j), (k, j) \in \mathcal{T}$ ,  $\mathbf{H}_{ij} = \mathbf{H}_{kj}$  and  $\mathbf{V}_{ij} = \mathbf{V}_{kj}$ .

Given this definition of a latent variable PDSEM, the distribution  $p_\infty(\mathbf{V})$  is obtained from applying the usual transition probabilities to the margin at the initial state  $p_1(\mathbf{V}_1) \equiv \sum_{\mathbf{H}_1} p_1(\mathbf{V}_1 \dot{\cup} \mathbf{H}_1)$ , and the margins of all transition probabilities  $p_{ij}(\mathbf{V}_{ij} | \mathbf{V}_i) \equiv \sum_{\mathbf{H}_{ij}} p_{ij}(\mathbf{V}_{ij} \dot{\cup} \mathbf{H}_{ij} | \mathbf{V}_i)$ . Here,  $\dot{\cup}$  refers to the disjoint union.

Fix a set of observed treatment variables  $\mathbf{A}$ , the union of  $\{\mathbf{A}_{ij} : (i, j) \in \mathcal{T}\}$ , such that  $\mathbf{a}_j$  are set to  $\mathbf{A}_{ij}, \mathbf{A}_{kj}$  for any  $(i, j), (k, j) \in \mathcal{T}$ , and the set of outcomes  $\mathbf{Y}_{ij} = \mathbf{V}_{ij} \setminus \mathbf{A}_{ij}$  for any  $(i, j) \in \mathcal{T}$ , with  $\mathbf{Y}$  the union of  $\{\mathbf{Y}_{ij} : (i, j) \in \mathcal{T}\}$ . Define the set of *strict parents* of  $\mathbf{V}$  as follows:  $\text{pa}_{\mathcal{G}}^s(\mathbf{V}) = \text{pa}_{\mathcal{G}}(\mathbf{V}) \setminus \mathbf{V}$ .

Identification for  $p_{\infty}(\mathbf{Y}(\mathbf{a}))$  in a latent variable PDSEM reduces to identification theory for  $p_1(\mathbf{Y}_1(\mathbf{a}_1))$  in the latent projection ADMG  $\mathcal{G}_1$  on  $\mathbf{V}_1$ , and  $p_{ij}(\mathbf{Y}_{ij}(\mathbf{a}_j)|\mathbf{V}_i(\mathbf{a}_i))$  in the latent projection CADMG  $\mathcal{G}_{ij}$  on  $\mathbf{V}_{ij}$  given  $\mathbf{V}_i$ , as follows:

*Lemma 3.* Under Assumptions 1, 2 and 3, given a latent variable PDSEM represented by  $\mathcal{G}_1$  and  $\{\mathcal{G}_{ij} : (i, j) \in \mathcal{T}\}$ ,  $p_{\infty}(\mathbf{Y}(\mathbf{a}))$  is identified from  $p_{\infty}(\mathbf{V})$  if and only if every bidirected component in  $\mathcal{G}_{1\mathbf{Y}_1^*}$  is intrinsic in  $\mathcal{G}_1$ , and every bidirected component in  $\mathcal{G}_{ij\mathbf{Y}_j^*}$  is intrinsic in  $\mathcal{G}_{ij}$  for every  $i$  and  $j$ . Moreover, if  $p_{\infty}(\mathbf{Y}(\mathbf{a}))$  is identified, it is equal to

$$p_1(\mathbf{Y}_1(\mathbf{a}_1)) \prod_{t=1}^{\infty} \left( \prod_{(i,j) \in \mathcal{T}} (p_{ij}(\mathbf{Y}_{ij}(\mathbf{a}_j)|\mathbf{Y}_i(\mathbf{a}_i)))^{\mathbb{I}(s_{t-1}^i, s_t^j)} \right) 1^{\mathbb{I}(s_{t-1}^*)} \quad (3.3)$$

where

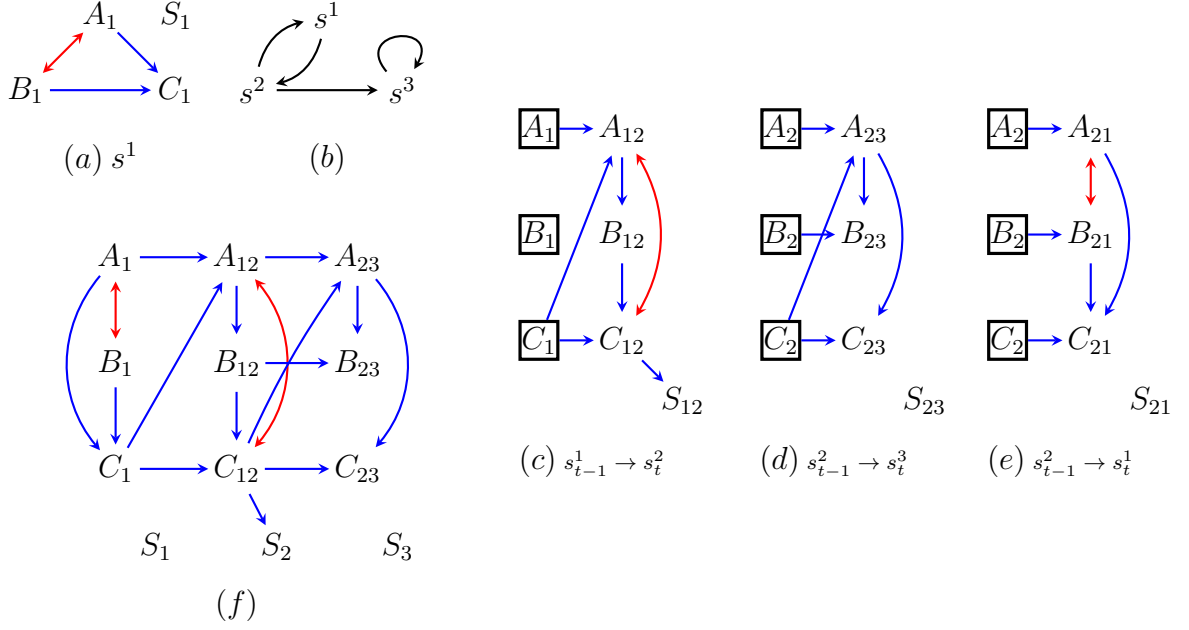
$$p_1(\mathbf{Y}_1(\mathbf{a}_1)) = \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}_{1\mathbf{Y}_1^*})} q_{\mathbf{D}}^1(\mathbf{D} | \text{pa}_{\mathcal{G}_1}^s(\mathbf{D})) \Big|_{\mathbf{A}_1 = \mathbf{a}_1}, \quad (3.4)$$

where each kernel  $q_{\mathbf{D}}^1(\mathbf{D} | \text{pa}_{\mathcal{G}_1}^s(\mathbf{D}))$  is in the nested Markov factorization of  $p_1(\mathbf{V}_1)$  with respect to  $\mathcal{G}_1$ , and

$$p_{ij}(\mathbf{Y}_{ij}(\mathbf{a}_j)|\mathbf{Y}_i(\mathbf{a}_i)) = \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}_{\mathbf{V}_{ij} \setminus \mathbf{A}_{ij}})} q_{\mathbf{D}}^{ij}(\mathbf{D} | \text{pa}_{\mathcal{G}_{ij}}^s(\mathbf{D})) \Big|_{\substack{\mathbf{A}_i = \mathbf{a}_i, \\ \mathbf{A}_j = \mathbf{a}_j}}, \quad (3.5)$$

where each kernel  $q_{\mathbf{D}}^{ij}(\mathbf{D} | \text{pa}_{\mathcal{G}_{ij}}^s(\mathbf{D}))$  is in the nested Markov factorization of  $p_{ij}(\mathbf{V}_{ij}|\mathbf{V}_i)$  with respect to  $\mathcal{G}_{ij}$ .

Before we discuss experimental results, let us consider the example in Fig. 3.3 to elucidate the above lemma. The figure shows a slightly modified version of the example in Fig. 3.2. Here, the first two states of the system involve hidden variables. Transition graphs are in Fig. 3.3(c)-(e).



**Figure 3.3.** A hidden variable PDSEM. (a) Causal structure of the initial state  $S^1$ . (b) The state transition diagram. (c),(d),(e) Latent projected causal diagrams representing possible transitions and subsequent states. (f) A snapshot of a possible PDSEM trajectory represented as an unrolled ADMG

The nested factorization for the initial graph in Fig. 3.3 (a) has intrinsic sets

$$(a) : \{A_1\}, \{B_1\}, \{C_1\}, \{A_1, B_1\}, \{S_1\}$$

with corresponding kernels

$$\begin{aligned} (a) : q_{A_1}(A_1) &\equiv p(A_1); q_{B_1}(B_1) = p(B_1); \\ q_{C_1}(C_1|A_1, B_1) &\equiv p(C_1|A_1, B_1); \\ q_{A_1, B_1}(A_1, B_1) &\equiv p(A_1, B_1); q_{S_1}(S_1) \equiv p(S_1). \end{aligned} \tag{3.6}$$

Similarly, the nested factorizations for the transition graphs in Fig. 3.3 (c),(d),(e) have intrinsic sets:

$$\begin{aligned} (c) : & \{A_{12}\}, \{B_{12}\}, \{C_{12}\}, \{A_{12}, C_{12}\}, \{S_{12}\} \\ (d) : & \{A_{23}\}, \{B_{23}\}, \{C_{23}\}, \{S_{23}\} \\ (e) : & \{A_{21}\}, \{B_{21}\}, \{A_{21}, B_{21}\}, \{C_{21}\}, \{S_{21}\}, \end{aligned}$$

with corresponding kernels

$$\begin{aligned}
(c) : q_{A_{12}}(A_{12}|C_1) &\equiv p(A_{12}|C_1); q_{B_{12}}(B_{12}|A_{12}) \equiv p(B_{12}|A_{12}); \\
q_{C_{12}}(C_{12}|C_1, B_{12}) &\equiv \sum_{A_{12}} p(C_{12}|B_{12}, A_{12}, C_1) p(A_{12}|C_1); \\
q_{A_{12}, C_{12}}(A_{12}, C_{12}|B_{12}, C_1) &\equiv p(C_{12}|B_{12}, A_{12}, C_1) p(A_{12}|C_1); q_{S_{12}}(S_{12}|C_{12}) \equiv p(S_{12}|C_{12}). \\
(d) : q_{A_{23}}(A_{23}|A_2, C_2) &\equiv p(A_{23}|A_2, C_2); q_{B_{23}}(B_{23}|B_2, A_{23}) \equiv p(B_{23}|B_2, A_{23}); \\
q_{C_{23}}(C_{23}|C_2, A_{23}) &\equiv p(C_{23}|C_2, A_{23}); q_{S_{23}}(S_{23}) \equiv p(S_{23}). \\
(e) : q_{A_{21}}(A_{21}|A_2) &\equiv p(A_{21}|A_2); q_{B_{21}}(B_{21}|B_2) \equiv p(B_{21}|B_2); q_{A_{21}, B_{21}}(A_{21}, B_{21}) \equiv p(A_{21}, B_{21}); \\
q_{C_{21}}(C_{21}|C_2, B_{21}, A_{21}) &\equiv p(C_{21}|C_2, B_{21}, A_{21}); q_{S_{21}}(S_{21}) \equiv p(S_{21}).
\end{aligned} \tag{3.7}$$

Applying the nested Markov factorization on the trajectory in Fig. 3.3 (f), we obtain the following factorization:

$$\begin{aligned}
& p(A_1, B_1, C_1) \cdot p(A_{12}, B_{12}, C_{12}|A_1, B_1, C_1) \cdot p(A_{23}, B_{23}, C_{23}|A_{12}, B_{12}, C_{12}) \\
&= \underbrace{\{q_{A_1, B_1}(A_1, B_1) q_{C_1}(C_1 | A_1, B_1)\}}_{(a)} \cdot \underbrace{\{q_{A_{12}, C_{12}}(A_{12}, C_{12}|B_{12}, A_1, C_1) q_{B_{12}}(B_{12}|A_{12})\}}_{(c)} \cdot \\
&\quad \underbrace{\{q_{A_{23}}(A_{23}|A_{12}, C_{12}) \cdot q_{A_{23}}(B_{23}|B_{12}, A_{23}) \cdot q_{C_{23}}(C_{23}|C_{12}, A_{23})\}}_{(d)},
\end{aligned}$$

where the kernels are given in (3.6) and (3.7) above.

## 3.6 Statistical Inference

Given counterfactual distributions identified via (3.2), (3.4), and (3.5), if a parametric likelihood may be specified in terms of components of these equations, statistical inference may be performed by plug-in estimation and Monte-Carlo simulation. In a fully observed PDSEM, the likelihood may be obtained from (3.1) by imposing parametric models for every Markov factor. In a hidden variable PDSEM, the likelihood may be obtained from the nested Markov factorization for the marginal distribution associated with the prior network ADMG, and the conditional distributions associated with transition network CADMGs. These likelihoods are available in multivariate normal assumption on the observed data, which we illustrate via the simulation study, and discrete state spaces, via the Möbius inversion formula parameterization discussed

in [Evans and Richardson, 2018] (see experiments in Section 3.7).

Given model parameters obtained by maximizing the PDSEM observed data likelihood, counterfactual distributions in (3.2), (3.4), and (3.5) may be obtained by simulating PDSEM trajectories using these modified factorizations, evaluated at MLE parameter values. Confidence intervals for any counterfactual parameter of interest may be obtained by parametric bootstrap.

However, an analogous approach is not straightforward for nested Markov parameterizations of the marginal PDSEM representing a PDSEM with hidden variables. In our simulations, we use a specific generative model for our continuous variables, i.e, the linear Gaussian Structural Equation model. Another choice based on work in [Evans and Richardson, 2014] is the Möbius parameterization for binary variables. However, this is ill-suited for drawing samples. Instead, existing approaches to sampling from a nested Markov discrete likelihood involve first converting the likelihood expressed in terms of the Möbius parameters to one expressed as a the joint distribution  $p(\mathbf{V})$  (from which it is easy to generate samples for a discrete sample space of  $\mathbf{V}$ ). Importantly, such a conversion leads to an intractable object that requires storage and running time exponential in size of  $\mathbf{V}$ . This holds *even if* the underlying model dimension of the nested Markov model is small. The situation is radically different from that of DAG models, where a small model dimension directly leads to a computationally efficient sampling scheme. For settings beyond Gaussian and discrete data, statistical inference strategies are significantly more complicated and have been discussed in [Bhattacharya et al., 2020]. While there exist promising approaches, based on the nested Markov generalization of the variable elimination algorithm [Shpitser et al., 2011], in general the problem remains open.

## 3.7 Experiments

### 3.7.1 Simulation Study

We simulate data and perform statistical inference using the PDSEM shown in Fig. 3.2. The system has states  $\{s^1, s^2, s^3\}$  and variables  $\{A, B, C\}$  in each state. Additionally,  $s^2$  has a hidden common cause of  $A$  and  $B$ . This is represented by the red (dotted) bidirected edge  $A \leftrightarrow B$  in the latent projected ADMG in Fig. 3.2(c). Patient health status  $A$ , surgeon experience  $B$ , and duration of the stage of surgery  $C$ , are all continuous variables. State and transition graphs are identical to those in Fig. 3.2. This PDSEM was used to consider the causal impact of surgeon experience (measured by total operating time in their career) on average surgery length. This outcome is easy to measure, and is known to serve as an informative proxy for other measures of surgery quality, such as follow-up assessments of quality of life [Rambachan et al., 2013, Jackson et al., 2011].

Parameters associated with the given generative model are  $p(S_{t+1} = s^j | S_t = s^i, \mathbf{V}_t)$ , where  $s_t^i \rightarrow s_{t+1}^j$  is a transition allowed by the model, and  $p(V_{t+1}^{ij} = v | S_{t+1} = s^j, S_t = s^i, \mathbf{V}_t)$ , where  $V^{ij} \in \{A^{ij}, B^{ij}, C^{ij}\}$ , where  $s_t^i \rightarrow s_{t+1}^j$  is an allowed transition. These are chosen to be reasonable for the surgery application, yielding a distribution Markov relative to appropriate graphs. We simulated  $N = 10000$  “surgeries,” with initial state  $s^1$ . Transition probabilities were generated using a logistic regression on variables in the current state, with transitions eventually terminating at the absorbing state. Each variable  $V_i$  is generated from a set of linear structural equations with correlated errors. Using generated data, state transition probabilities were estimated using maximum likelihood. Parameters for the structural equation model were estimated using the RICF algorithm [Drton et al., 2009], implemented in the Ananke package [Bhattacharya et al.].

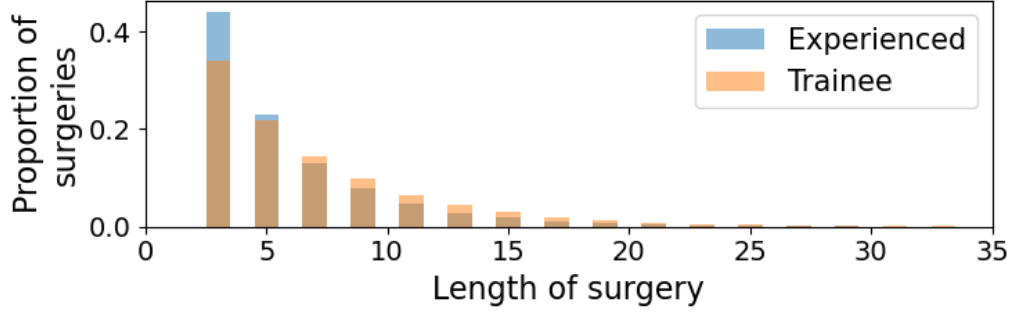
We assessed the causal impact of surgeon experience on operating time by generating

two sets of sampled surgery trajectories where, in each stage of the surgery, the surgeon was intervened to have higher (vs. lower) career operating time by one unit. These trajectories may be viewed as a Monte Carlo sampling scheme for evaluating the functional given by (3.3), (3.4) and (3.5). This approach generalizes similar schemes developed for longitudinal causal models [Westreich et al., 2012]. The comparison of these two sets of trajectories may be viewed as a generalization of the average causal effect (ACE) from classical longitudinal causal models to PDSEMs.

The results are shown in Fig. 3.4. Surgeries performed by experienced surgeons are shorter ( $\mu = 5.79$ ,  $\mathbf{q}_{0.05} = 3$ ,  $\mathbf{q}_{0.95} = 13$ ) than those performed by trainees ( $\mu = 7.02$ ,  $\mathbf{q}_{0.05} = 3$ ,  $\mathbf{q}_{0.95} = 17$ ) where  $\mathbf{q}_p$  denotes the  $p^{\text{th}}$  quantile. Surgeries performed by trainees have higher variance.

### 3.7.2 Septoplasty Application

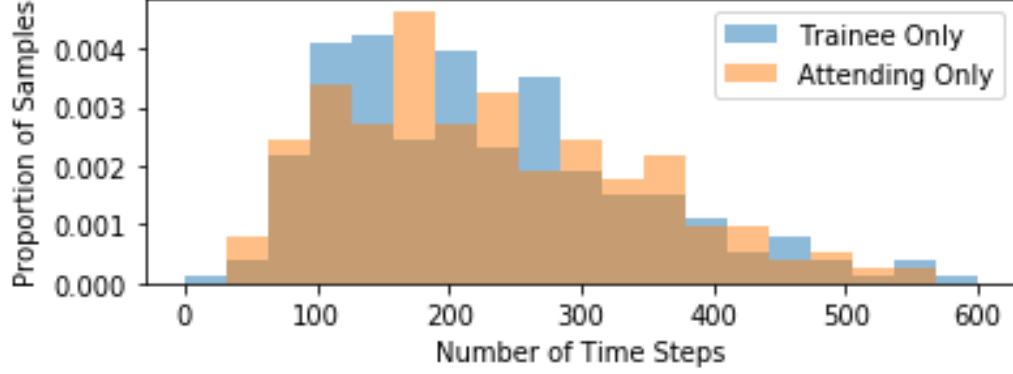
We are interested in the causal impact of surgeon experience on the average length of surgery, in the context of septoplasty. Our dataset consists of 236 septoplasty procedures conducted at our institution’s research hospital. A total of 57343 timestamped records were collected, including tool and personnel activity. Surgeries consist of six distinct phases:  $s^1$  (opening of the septum),  $s^2$  (raising septal flaps),  $s^3$  (removal of deviated septal cartilage and bone),  $s^4$  (reconstruction),  $s^5$  (closing of the incision), and  $s^6$  (other activity). An artificial absorbing state  $s^{\text{end}}$  represents the end of procedures. Procedures are often led by an attending, with a surgeon trainee assisting. Of the surgeries, 42.79% of them were performed fully by the leading attending; the others by a team. Also, attending surgeons perform for 64.98% of all operating time and trainees the rest. Twelve different surgical tools were tracked for use. The state transition diagram representing allowed state transitions is presented in Fig. 3.6. We discretized all variables into two categories, and fit model parameters by maximum likelihood.



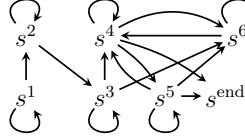
**Figure 3.4.** Histograms of the number of transitions in a surgery under two different interventions: when a more experienced surgeon performs the entire procedure, and when a less experienced trainee performs the entire procedure.

While there are certainly unobserved but relevant confounding variables in the problem we consider (such as underlying patient state), we assume these variables influence treatment variables (identity of the surgeon), as well as variables in the next stage only via relevant observed variables (such as duration of the stage, and tools currently in use). In addition to implying Assumption 1, this implies identifiability of the parameter of interest (a contrast of the average length of surgery had experienced vs inexperienced surgeon performed all stages) is given by Lemma 2, and statistical inference may be performed as if the prior network were a DAG, and every transition network were a CDAG, without loss of generality.

Estimation of  $p(s_t | s_{t-1}, \mathbf{v}_{t-1})$  at all levels of  $s_{t-1}, \mathbf{v}_{t-1}$  is not always possible due to finite sample limitations. To address this, we apply additive smoothing to  $p(s_t | s_{t-1}, \mathbf{v}_{t-1})$ , based on the empirical distribution  $p(s_t | s_{t-1})$ . Results are presented in Fig. 3.5. We have made considerable assumptions in modeling our PDSEM and have closely matched the generative model to the empirical distribution; see [Srinivasan et al., 2021] for more details. We observe that the causal effect of surgeon skill on surgery length, given our learned parameters, is close to zero. This indicates that policies that govern the trade-off between the need to train surgeons, and overall surgery quality (as quantified by our outcome) are effective at our institution.



**Figure 3.5.** Histograms of hypothetical surgeries performed only by a junior trainee surgeon (blue) versus hypothetical surgeries performed only by a senior attending surgeon (orange). Surgeries performed by the attending are slightly longer ( $\mu = 244.391, \sigma = 139.9$ ) than those of the trainee ( $\mu = 233.5, \sigma = 125.9$ ).



**Figure 3.6.** The state transition diagram for the surgery data application.

### 3.8 Conclusions

We developed the Path Dependent Structural Equation Model (PDSEM) for longitudinal data unifying complex state structure from DBNs and complex state transition dynamics from MDPs. We have described counterfactuals associated with these causal models that can alter the subsequent temporal evolution of the system, identification theory for such counterfactuals in terms of the observed data distribution, and estimation. We showed the utility of the model in clinical settings, in data from a septoplasty procedure. We also extended our results to  $k$ th order Markov systems, and compared how DBNs might fall short of PDSEMs in describing certain types of systems. Developing novel methods for efficient Monte Carlo sampling based statistical inference for hidden variable PDSEMs with the nested Markov model is an area worth exploring in the future.

# Chapter 4

## Entangled Missingness

### 4.1 Introduction

*Systematically* missing data records, including due to survey non-response, dropout or loss-to-followup, imperfect data collection and other reasons, can substantially bias analyses if not properly addressed. While plenty of methods have been developed to address missing data, much of it treats samples as independent and identically distributed (i.i.d.) [Rubin, 1976, Little, 2021, Little and Rubin, 2002, Glymour, 2006, Daniel et al., 2012, Martel García, 2013, Mohan et al., 2013, Thoemmes and Rose, 2014, Tian, 2015, Shpitser, 2016, Bhattacharya et al., 2019b, Nabi et al., 2020, Mohan and Pearl, 2021, Scharfstein et al., 2021, Nabi and Bhattacharya, 2022]. The i.i.d. assumption is only reasonable when interactions between units under investigation are negligible and can be ignored. In recent years there has been increasing recognition that many settings are subject to dependence among data samples and in particular, *interference*, where variables measured on one unit may have a causal effect on those measured on another unit through the varied ways in which humans influence one another, via offline or online network relationships.

Literature has largely lagged behind in recognizing that data dependence and missing data might occur simultaneously, with the exception of a few pieces of work - [Chang

et al., 2020] investigated multiple imputation techniques for missing data in health data networks; [Smith et al., 2017] characterized bias in networks empirically, where values are designed to be missing systematically; [Gile and Handcock, 2017] used a likelihood-based modeling approach for health studies with partially observed data; [Almquist and Butts, 2018] proposed estimation methods for network logistic regression models in the presence of missing data. But there are no graphical causal methods that address *entangled missingness*, defined as missingness with dependence.

Briefly, let us consider a few example settings that involve both data dependence and missingness. First, consider a study of vaccine effectiveness in which one unit’s vaccination status may help to protect their friends or family members from contracting an infectious disease and some records end up missing due to unknown clerical errors at the hospital. Or, consider instead, a mass public health surveillance effort, where demographic and family data are collected on a large population of individuals from which a smaller group is invited to participate in substudy where more detailed data are collected; an individual’s choice to enroll in the substudy may depend both on their own characteristics and on the characteristics of, or choices made by, other individuals in their social network. And finally, consider a course evaluation survey answered by students. Answers to survey questions in this case would certainly be unobserved for any individual who did not attend or who dropped out before the end of the course. However, because classes have social and collaborative components, the underlying values of responses of a particular individual, *had they attended*, will potentially differ depending on the attendance of that individual’s peers. As we will illustrate, these different examples describe different types of *entanglements* and must be treated accordingly to reduce bias.

### 4.1.1 Contributions

This work represents the first graphical causal method to study systems that exhibit both data dependence and missingness. First, we define a hierarchy of entanglements that can arise in systems. *Target law dependence* occurs when the full data distribution exhibits either statistical or causal dependence. *Missingness process dependence* occurs when the missingness indicator of one unit may depend on (variables of) other units. *Missingness interference* occurs when the underlying variables of interest in a missing data problem may be indexed by the missingness indicators of multiple units. For this last setting, we propose modeling multiple versions of underlying variables in a way that is structurally similar to counterfactual variables or potential outcomes in causal inference. We show that existing missing data models may be extended to describe entanglements arising from target law dependence and missingness process dependence, while those arising from missingness interference require a novel framework. We present sound and complete non-parametric identification results under this framework. Finally, we demonstrate the use of this novel modeling approach on synthetic data.

## 4.2 Background

We have already outlined the main concepts required, including graphical causal models of DAGs and ADMGs (Section 2.2 and Section 2.4), interference (Section 2.5) and missing data models for i.i.d data (Section 2.6) in Chapter 2. Notation will remain consistent with that developed in earlier chapters.

## 4.3 Motivating Example

Consider the following example, where investigators are interested in evaluating outcomes of a health and wellness program that involves a diet and workout regimen of participants, based on their baseline health indicators. Participation in the study is through voluntary registration. Bob and Anne are neighbors, and see the advertisement

for enrollment. Imagine two different scenarios: (1) Both Bob and Anne feel strongly about the program and enroll together. Since they participate together, they are able to easily motivate each other and adhere to their regimen. As a result, Anne records a successful outcome, improved metabolism and better BMI. Alternatively, (2) Bob enrolls and his baseline variables are collected but he does not actually participate. Anne goes through the program alone but finds it challenging to follow effectively, and her results are not as good. Bob's final outcome (BMI) is not measured as he does not participate. But Anne has two possible hypothetical program results under the two different scenarios: one where Bob participates, and the other where he does not. To put it another way, the observed results for Anne depend on the missingness status of Bob's results in the study. Similar arguments might be made for Bob.

Formally, for Anne (the unit indexed by 1), the outcome BMI has two possible values:  $Y_1^{(r_{Y_1}=1, r_{Y_2}=0)}$  for when Bob does not participate and his BMI is, hence, not recorded ( $r_{Y_2} = 0$ ) and  $Y_1^{(r_{Y_1}=1, r_{Y_2}=1)}$  for when Bob's BMI is recorded ( $r_{Y_2} = 1$ ). For the purposes of this example, we make two simplifications here. One, we assume that the only missing variable for any unit is the outcome, and hence denote  $r_{Y_1}$  by  $r_1$  and  $r_{Y_2}$  by  $r_2$ ; this assumption will be relaxed in the more general case (see Section 4.4.2.1). Second, we drop  $r_1$  from the superscript and denote the counterfactuals as  $Y_1^{(1, r_2=0)}$  and  $Y_1^{(1, r_2=1)}$ . This latter simplification can be done without ambiguity because Anne's outcomes are defined only when  $r_1 = 1$ . Anne's missingness status  $R_1$  and observable BMI  $Y_1$  are recorded in the study data. This observed outcome is not just a function of her counterfactuals  $Y_1^{(1, r_2=0)}$ ,  $Y_1^{(1, r_2=1)}$  and her missingness status  $R_1$ , but also the missingness status of Bob's outcome,  $R_2$ . One can think of  $R_2$  as a switch, selecting which counterfactual,  $Y_1^{(1, r_2=0)}$  or  $Y_1^{(1, r_2=1)}$ , is realized in the observation. By symmetry in this particular example, we assume the same is true for Bob. Thus, we have the following relationships between observed values  $Y_1, Y_2$  of neighbors Bob and Anne and

the corresponding counterfactuals:

$$Y_1 \leftarrow \begin{cases} (1 - r_2)Y_1^{(1,r_2=0)} + r_2Y_1^{(1,r_2=1)} & \text{if } r_1 = 1 \\ ? & \text{if } r_1 = 0 \end{cases}$$

$$Y_2 \leftarrow \begin{cases} (1 - r_1)Y_2^{(1,r_1=0)} + r_1Y_2^{(1,r_1=1)} & \text{if } r_2 = 1 \\ ? & \text{if } r_2 = 0 \end{cases}$$

Since this setting involves multiple counterfactual versions of  $Y_1$ , it cannot be captured by existing (i.i.d.) missing data models described in Section 2.6, which are restricted to counterfactuals of the form  $Y_1^{(r_1=1)}$  or  $Y_1^{(1)}$  for short, one per (each variable of a) unit. This example represents a special type of entanglement called missingness interference, and represents a gap in current modeling approaches. We will discuss it in more detail in Section 4.4.2.

## 4.4 Graphical Models for Entangled Missingness

We discuss three types of *entanglements* by which missing data and data dependence, can occur together. The following definitions pertain to the distribution  $p(\mathbf{R}, \mathbf{O}, \mathbf{Z}^{(1)}, \mathbf{Z})$  with all hidden variables  $\mathbf{H}$ , if any, marginalized out <sup>1</sup>.

1. Target Law Dependence ( $\mathcal{E}_1$ ): Counterfactuals  $\mathbf{Z}_i^{(1)}$  and  $\mathbf{Z}_j^{(1)}$  of units  $i, j$  in a block depend on each other. This means the target law  $p(\mathbf{O}, \mathbf{Z}^{(1)})$  does not factorize into unit-specific marginal distributions. Such a situation arises in problems where the underlying distribution, had there been no missing data, exhibits data dependence or interference.
2. Missingness Process Dependence ( $\mathcal{E}_2$ ): Missingness indicators  $\mathbf{R}_i$  for unit  $i$  depend on variables corresponding to unit  $j$ , and thus the missingness process  $p(\mathbf{R} \mid \mathbf{O}, \mathbf{Z}^{(1)})$  does not factorize into unit-specific factors.

---

<sup>1</sup>This distinction is necessary since in the presence of hidden variables, the same underlying data generating process might appear to be coming from different types of entanglements in the two different laws  $p(\mathbf{R}, \mathbf{O}, \mathbf{Z}^{(1)}, \mathbf{Z}, \mathbf{H})$ , pertaining to the hidden variable DAG and  $p(\mathbf{R}, \mathbf{O}, \mathbf{Z}^{(1)}, \mathbf{Z})$ , the latent projection ADMG. To remain consistent, we define entanglements on the latter.

3. Missingness Interference ( $\mathcal{E}_3$ ): Observed variables  $\mathbf{Z}_i$  for unit  $i$  depend on missingness indicators  $\mathbf{R}_j$  for unit  $j$ .<sup>2</sup> Under  $\mathcal{E}_3$ , the counterfactual variable corresponding to the variable  $Z_i$  must be indexed not only by  $R_{Z_i}$  being set to 1, but also by the missingness status of (variables of) other units. For instance the counterfactual  $Z_i^{(R_{Z_i}=1, R_{Z_j}=0)}$  would correspond to variable  $Z$  of unit  $i$ , had it been observed ( $R_{Z_i} = 1$ ) and had variable  $Z$  of unit  $j$  been missing.

In Section 4.4.1, we describe illustrative examples and corresponding graphs for entanglements *without* missingness interference; these settings can be described adequately with a simple reinterpretation of existing missing data models and graphs for i.i.d. settings. In Section 4.4.2, we describe scenarios with missingness interference, extend the new notation we briefly introduced in Section 4.3 and present graphical representations for such settings.

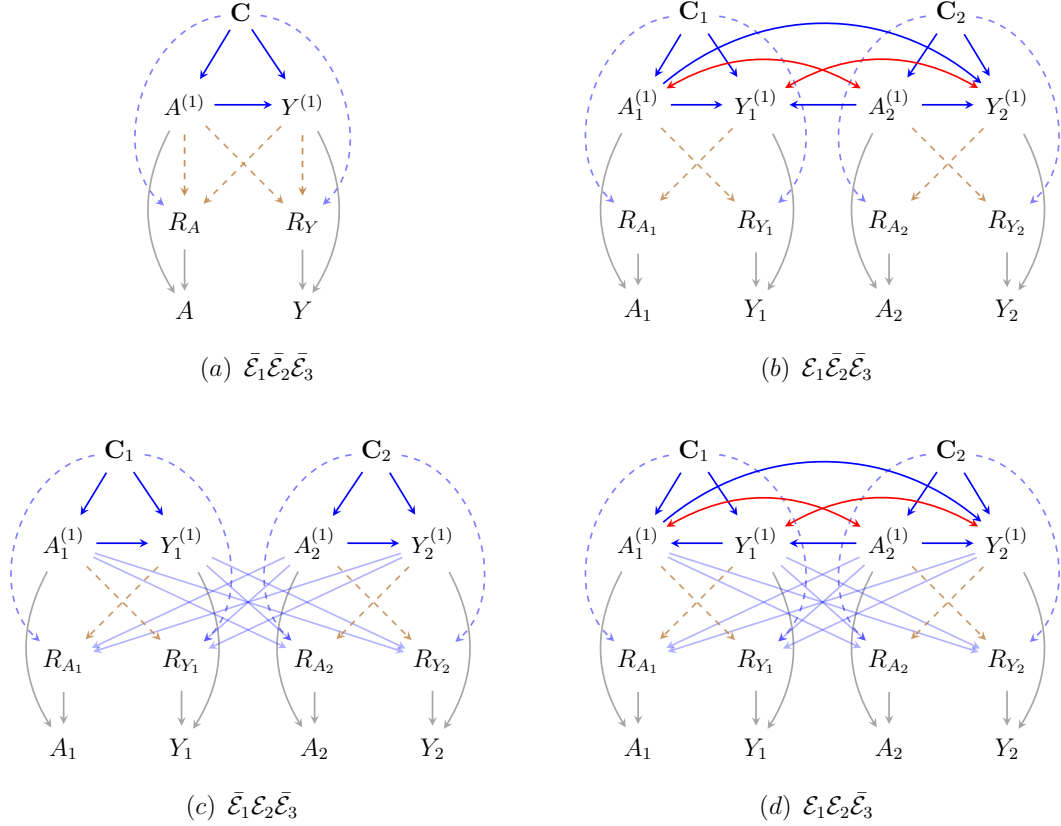
#### 4.4.1 Entanglements Without Missingness Interference

**Scenario 1:** Consider studying the effect of drug  $A$  on disease status  $Y$  from a community hospital database. Assume hospital records are incomplete and missingness indicators  $R_{A_i}$ ,  $R_{Y_i}$  denote whether the treatment and disease status values were recorded for patient  $i$ , with counterfactuals  $Y_i^{(1)}$ ,  $A_i^{(1)}$  denoting the true (but possibly unrecorded) values of these variables. Baseline covariates, denoted by  $\mathbf{C}_i$  for unit  $i$ , are always observed and include age and alcohol consumption. In this example,  $\mathbf{Z} \equiv \{A, Y\}$  and  $\mathbf{O} \equiv \{\mathbf{C}\}$ . For simplicity of presentation, we will work with dyads.

In order to provide a contrast between settings with i.i.d. missing data models and those with entanglement, we first start with a scenario which involves no kind of entanglement ( $\bar{\mathcal{E}}_1 \bar{\mathcal{E}}_2 \bar{\mathcal{E}}_3$ ) within this setup and build the examples further to illustrate how entanglements might arise. Within each example, we also point out the hierarchy of

---

<sup>2</sup>This can be interpreted as an extension of classical interference, where one unit's treatment affects another unit's outcome, except that the treatment has been replaced by the missingness indicator. Hence the name missingness interference.



**Figure 4.1.** Four scenarios representing all possible ways target law dependence and missingness process dependence may arise in a dyadic partial interference setting without missingness interference arising.

MCAR, MAR and MNAR mechanisms as this will be of use in reconciling identification results presented in Section 4.5.

**Scenario 1.1** ( $\bar{\mathcal{E}}_1 \bar{\mathcal{E}}_2 \bar{\mathcal{E}}_3$ ). When the disease being investigated is not contagious, and the missingness process, pertaining to non-response, deficiencies in data collection, and so on, does not exhibit dependence across patients, we might be able to assume that a patient's record does not influence the record of any other patient, i.e., data is i.i.d.

The graph in Fig. 4.1(a) depicts this situation. Including different set of edges in Fig. 4.1(a) yields either MCAR where  $R_A, R_Y \perp\!\!\!\perp A^{(1)}, Y^{(1)}, \mathbf{C}$  (if only solid edges are included), MAR where  $R_A, R_Y \perp\!\!\!\perp A^{(1)}, Y^{(1)} \mid \mathbf{C}$  (if in addition to the solid lines, blue

dashed edges are also included)<sup>3</sup>, or MNAR (if at least one brown dashed edge is included). Edges  $A \rightarrow R_A$  and  $Y \rightarrow R_Y$  are called *self-censoring* or *self-masking* edges [Brown, 1990, Mohan et al., 2018, Nabi et al., 2020].

**Scenario 1.2** ( $\mathcal{E}_1 \bar{\mathcal{E}}_2 \bar{\mathcal{E}}_3$ ). Suppose the outcome  $Y$  is an infectious disease, like Covid-19. Since patients can infect one another, data on patients residing in the same household or geographic area are likely to exhibit dependence in outcomes. In addition, surging infections may lead to treatment shortages, which will lead to dependence among treatments due to *allocational interference* [Ogburn and VanderWeele, 2014]; geographically localized treatment shortages occurred at several points early in the Covid-19 pandemic. Finally, successful prevention or treatment of an infectious outcome for unit  $i$  may influence outcomes for other units  $j$  by preventing potential transmission from  $i$  to  $j$ . Missingness mechanisms remain independent across units, however.

Fig. 4.1(b) depicts this setting. Bidirected arrows encode dependence (due to shared hidden causes) between  $A_1^{(1)}$  and  $A_2^{(1)}$  and between  $Y_1^{(1)}$  and  $Y_2^{(1)}$ , respectively. For simplicity of visualization, we omit connections among baseline variables  $\mathbf{C}_1, \mathbf{C}_2$  and other variables in other units.

The edge subgraph of Fig. 4.1(b) consisting only of solid edges corresponds to a MCAR model, since  $\mathbf{R} \perp\!\!\!\perp \mathbf{Y}^{(1)}, \mathbf{A}^{(1)}, \mathbf{C}$  by the m-separation criterion [Richardson, 2003]. Similarly, the edge subgraph of Fig. 4.1(b) consisting of all edges other than dashed brown corresponds to a MAR model, since  $\mathbf{R} \perp\!\!\!\perp \mathbf{Y}^{(1)}, \mathbf{A}^{(1)} \mid \mathbf{C}$ . Finally, including dashed brown edges in the MAR model yields an MNAR model.

**Scenario 1.3** ( $\bar{\mathcal{E}}_1 \mathcal{E}_2 \bar{\mathcal{E}}_3$ ). Assume that we are again dealing with a non-contagious disease, and that treatment allocation for one unit is not influenced by any features

---

<sup>3</sup>Note that the MAR version is identical to that in Fig. 2.5(d) as all of Scenario 1.1 pertains to i.i.d. settings.

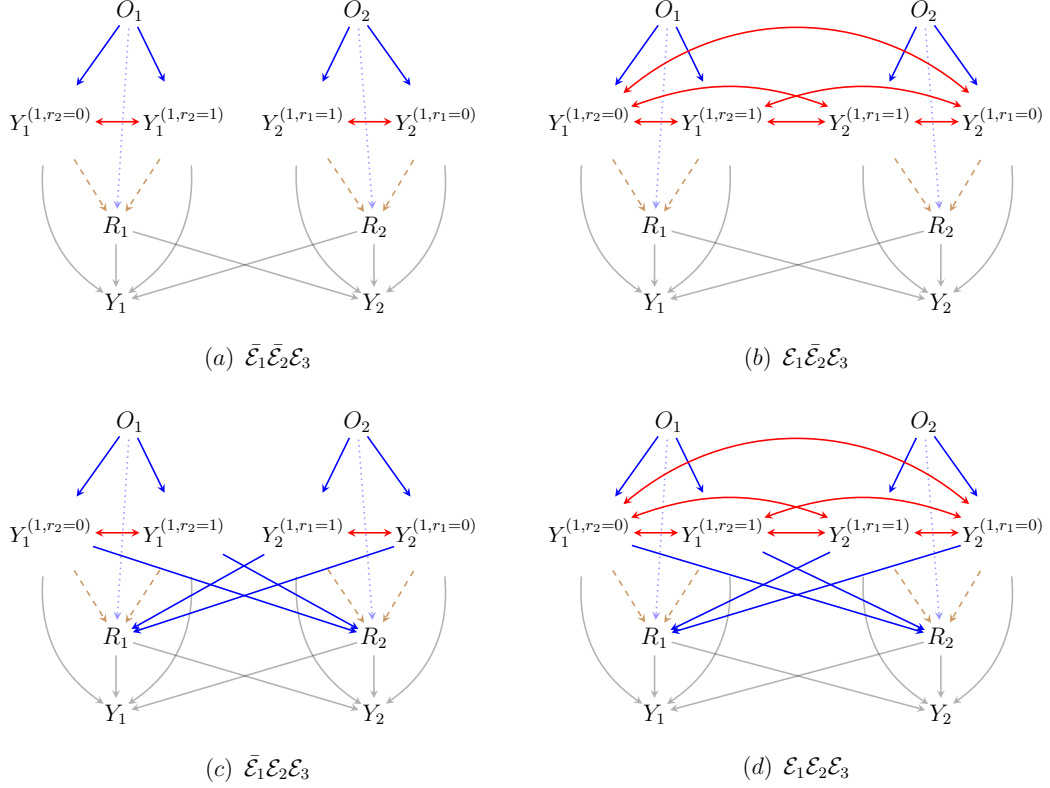
of other units. However, the missingness process for one unit does depend on other units. For example, data includes members of a family who may jointly decide to change healthcare providers, resulting in missing treatment and outcome data for all members of the family.

Fig. 4.1(c) depicts this setting, and the light blue edges between the counterfactuals  $A_1^{(1)}, Y_1^{(1)}$  and missingness indicators  $R_{A_2}, R_{Y_2}$ , as well as  $Y_2^{(1)}, Y_2^{(1)}$  and  $R_{A_1}, R_{Y_1}$  represent missingness process dependence. Edges from  $\mathbf{C}_2$  to variables of unit 1, and from  $\mathbf{C}_1$  to variables of unit 2 are omitted for simplicity.

An MCAR version of this scenario simply reduces to scenario  $\bar{\mathcal{E}}_1\bar{\mathcal{E}}_2\bar{\mathcal{E}}_3$ , representing i.i.d. data. An example of a MAR mechanism in this scenario corresponds to the absence of light solid blue edges in Fig. 4.1(c) and instead having edges  $\mathbf{C}_1 \rightarrow R_{A_2}$ ,  $\mathbf{C}_1 \rightarrow R_{Y_2}$ ,  $\mathbf{C}_2 \rightarrow R_{A_1}$  and  $\mathbf{C}_2 \rightarrow R_{Y_1}$  (not shown).

**Scenario 1.4** ( $\mathcal{E}_1\mathcal{E}_2\bar{\mathcal{E}}_3$ ). This scenario simply combines complications in scenarios 1.2 and 1.3, such that both target law dependence and missingness process dependence are present. The edges can be interpreted exactly as before, with scenarios 1.2 and 1.3. Specifically, the light blue solid edges between counterfactuals and missingness indicators make the missingness process dependence MNAR. Instead, if the dependence was on observed covariates, the mechanism would be MAR. Dashed edges, both brown and blue operate within a unit and simply induce dependence that is MNAR or MAR, respectively.

Finally, we note that the graphs in Fig. 4.1 can be extended in a straightforward fashion to a network where blocks are not just dyads, but are of arbitrary size  $m$ , to illustrate the different types of entanglement discussed here. Identification results corresponding to the settings described until now are discussed in Section 4.5.1.



**Figure 4.2.** Four scenarios representing all possible ways target law dependence and missingness process dependence may arise in a dyadic partial interference setting when missingness interference is present. We assume that only outcomes  $Y_i$  are missing, and  $R_i$  is shorthand for  $R_{Y_i}$ .

#### 4.4.2 Entanglements With Missingness Interference

In some situations where missingness and data dependence occur together, like the example of Bob and Anne in Section 4.3, observed realizations of underlying variables are influenced by missingness indicators of multiple units, leading to a missing data analogue of *interference problems* in causal inference. As aforementioned, this type of situation cannot be captured by missing data models described so far, nor described by the graphs in Fig. 4.1. Instead, we use the notation developed in Section 4.3.

We call, for every unit  $i$ , the set of units whose missingness indicators index the counterfactuals of unit  $i$ , as *affectors* of  $i$ , or  $\text{aff}(i)$ , in the network, inspired by the use of the word in neuroscience [Ebeling and Feistel, 2011]. This relationship need not be

symmetric, and one unit can be the affector of another, without the reverse being true. With these considerations in mind, we now describe the four scenarios that arise when missingness interference may occur, extending the example of Anne and Bob.

**Scenario 2.1** ( $\bar{\mathcal{E}}_1\bar{\mathcal{E}}_2\mathcal{E}_3$ ). First, consider a setting where there is neither target law dependence, nor missingness process dependence, but only missingness interference. Let us assume that diets and exercise regimens are recommended to participants, based solely on their measured baseline covariates. As an approximation, we assume that a person’s BMI is affected only by their actions, and hence the target laws are not directly influenced by each other. As described, Anne is more likely to workout if her friend Bob does ( $R_j \rightarrow Y_i$ ), affecting her BMI. We assume that  $Y_i^{(1,r_j=0)}$  and  $Y_i^{(1,r_j=1)}$  are associated for any unit  $i$  via a hidden common cause  $H_i$ , resulting in a bidirected ( $\leftrightarrow$ ) edge between the counterfactuals in the latent projection graphs. See Fig. 4.2(a). Brown edges are self-censoring edges.

**Scenario 2.2** ( $\mathcal{E}_1\bar{\mathcal{E}}_2\mathcal{E}_3$ ). Now, assume instead that Bob and Anne are siblings. One might imagine that the effect of diet and exercise on a human being are mediated by their genetics, which siblings share, giving rise to target law dependence. Additionally, assume that those conducting the study randomize who gets to participate, from those who register (and hence, whether someone is included in the study is independent of their sibling). This can give rise to a structure as shown in Fig. 4.2(b). Adherence to prescribed regimens is dependent on a person’s sibling being included in the study, giving rise to two underlying counterfactual BMIs for each unit. Here too, the dependence between counterfactuals is assumed to be due to a shared common cause  $H$ , resulting in target law dependence being represented by bidirected edges.

**Scenario 2.3** ( $\bar{\mathcal{E}}_1\mathcal{E}_2\mathcal{E}_3$ ). Assume Bob and Anne are not genetically related. Further, assume that knowledge of Bob’s potential BMI if he enrolled in the study (or not) makes Anne more likely to enroll herself, and the investigators include everyone who

registers, without randomly selecting a subset of the entries, giving rise to missingness process dependence. This would result in a graph like in Fig. 4.2(c).

The blue edge  $Y_2^{(1, r_1=1)} \rightarrow R_1$  introduces a more general type of self-censoring, which we call *affector-censoring*. This corresponds to the situation where a counterfactual variable of a unit (which is indexed by its affector's missingness status) censors that very same missingness status.

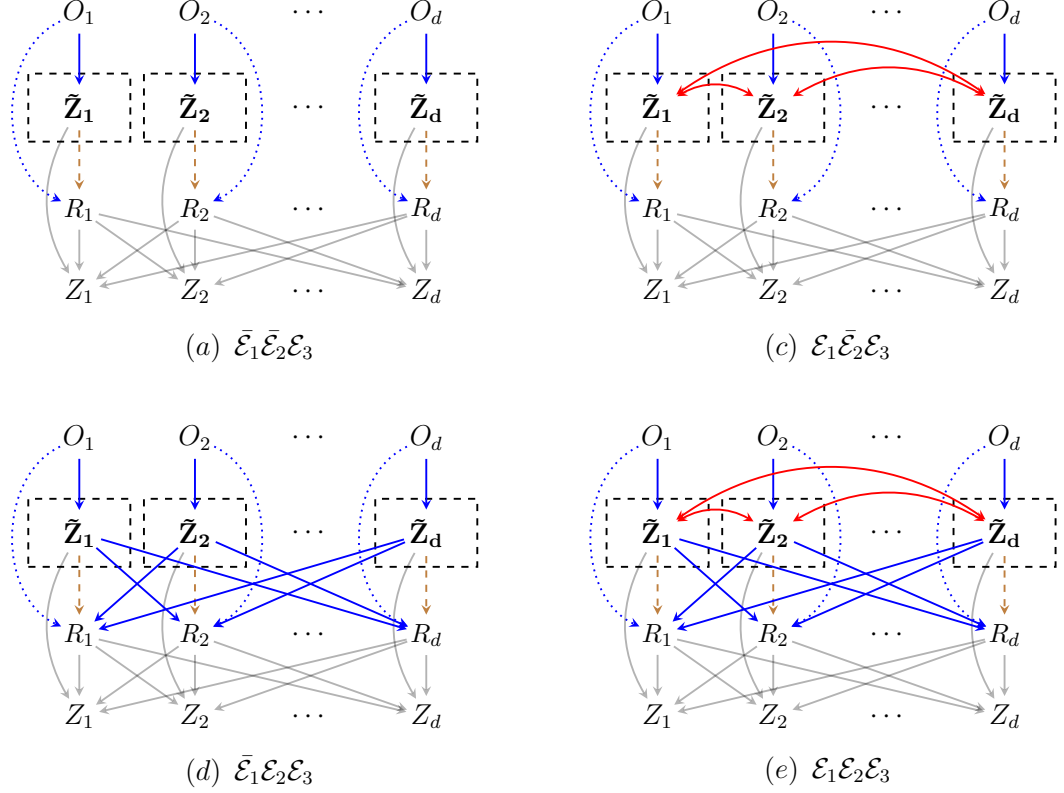
**Scenario 2.4** ( $\mathcal{E}_1\mathcal{E}_2\mathcal{E}_3$ ). Finally assume a situation where Bob and Anne share genetics and also enroll based on their knowledge of how fit the other person is likely to become as part of the study, combining all three possible types of entanglements in one setup. This scenario is illustrated via the graph in Fig. 4.2(d).

#### 4.4.2.1 Beyond Dyads

Interactions in many realistic settings are not restricted to dyads and hence we describe counterfactuals (and graphs) for networks with missingness interference. For simplicity of presentation, we will restrict ourselves to cases where, whenever  $j \in \text{aff}(i)$ , if the edge  $R_{Z_j} \rightarrow Z_i$  exists, then  $R_{Z'_j} \rightarrow Z_i$  for all  $Z_j, Z'_j \in \mathbf{Z}_j$ . That is, whenever  $j \in \text{aff}(i)$ , we are able to use a single missingness indicator  $\mathbf{R}_j$  which determines whether  $\mathbf{Z}_j$  is missing, instead of individual variables  $R_{Z_j}$  in the index of counterfactuals corresponding to unit  $i$ . And that, within a unit, the edge  $R_{Z_i} \rightarrow Z'_i$  is absent, for distinct  $Z, Z'$ . This simply allows us to replace  $R_{Z_i} = 1$  by 1, in the index of counterfactuals of unit  $i$ . To understand the implications of removing these restrictions, see the note at the end of this section.

We assume a network of  $n$  individuals. The full data is formed by i.i.d. realizations of the full law  $p(\{\mathbf{O}_i, \tilde{\mathbf{Z}}_i, \mathbf{R}_{\mathbf{Z}_i}, \mathbf{H}_i : i = 1, \dots, n\})$ , where  $\tilde{\mathbf{Z}}_i \equiv \{\mathbf{Z}_i^{(1, \mathbf{r}_{\text{aff}(i)})} : \mathbf{r}_{\text{aff}(i)} \text{ a value of } \mathbf{R}_{\text{aff}(i)}\}$  represents counterfactual versions of variables  $\mathbf{Z}_i$  for unit  $i$ , the set  $\text{aff}(i)$  refers to the units that are affectors of unit  $i$ , and  $\mathbf{R}_{\text{aff}(i)} \equiv \{\mathbf{R}_{\mathbf{Z}_j} : j \in \text{aff}(i)\}$  where  $\mathbf{R}_{\mathbf{Z}_j}$  refers to the set of all missingness indicators  $R_{Z_j}$  for all variables  $Z_j \in \mathbf{Z}_j$





**Figure 4.4.** A general network with  $d$  units featuring missingness interference. (a) - (d) enumerate all possible ways target law dependence and missingness process dependence may arise, as a generalization of the dyad in Fig. 4.2.

and (ii) every element  $Z_i$  in  $\mathbf{Z}$  has all its corresponding counterfactuals and  $R_{Z_i}$  as parents. Restriction (ii) is a consequence of the consistency property in (4.1).

A latent projection ADMG, obtained by projecting out fully hidden variables  $\cup_{i=1}^n \mathbf{H}_i$  from this DAG, will be our preferred choice of graphical representation, just as in Fig. 4.1 and Fig. 4.2. A simple example of such an ADMG for general networks is shown in Fig. 4.3, where each unit  $i$  has only one missing variable  $Z_i$ . Dependence among counterfactuals  $\mathbf{Z}_i^{(1, \mathbf{r}_{\text{aff}(i)})}$  is represented by bidirected edges, obtained by latent projecting  $H_i$  out. In this example, every unit is an affector of every other unit.

When the incoming and outgoing edges of vertices  $Z_i^{(1, \mathbf{r}_{\text{aff}(i)})}$  and  $Z_i^{(1, \mathbf{r}'_{\text{aff}(i)})}$  are identical for all  $\mathbf{r}, \mathbf{r}'$  in the graph, we can make the graphical representation even more compact by introducing the vertex  $\tilde{\mathbf{Z}}_i$  corresponding to the set  $\tilde{\mathbf{Z}}_i$  in place of all its elements

$Z_i^{(1, \mathbf{r}_{\text{aff}(i)})}$ , for all  $i$ . Such graphs, which we will denote by  $\tilde{\mathcal{G}}$ , are shown in Fig. 4.4. In particular, the graph composed of solid edges in Fig. 4.4 (a) is a condensed representation of the graph in Fig. 4.3, and any vertex  $\tilde{\mathbf{Z}}_i$  (shown in a dashed black box) compactly represents the full set of vertices  $Z_i^{(1, \mathbf{r}_{\text{aff}(i)}=0, \dots, 0)}$ ,  $Z_i^{(1, \mathbf{r}_{\text{aff}(i)}=1, \dots, 0)}$ ,  $\dots$  and  $Z_i^{(1, \mathbf{r}_{\text{aff}(i)}=1, \dots, 1)}$ , all connected by bidirected edges. In  $\tilde{\mathcal{G}}$ , we interpret the edge  $V \rightarrow \tilde{\mathbf{Z}}_i$ , for any node  $V$ , to mean that  $V$  influences all the counterfactuals  $\mathbf{Z}_i^{(1, \mathbf{r}_{\text{aff}(i)})}$  in  $\tilde{\mathbf{Z}}_i$ . Similarly  $V \leftarrow \tilde{\mathbf{Z}}_i$  is interpreted to mean that all counterfactuals  $\mathbf{Z}_i^{(1, \mathbf{r}_{\text{aff}(i)})}$  in  $\tilde{\mathbf{Z}}_i$ , influence  $V$ .

Further, it is helpful to view  $\tilde{\mathcal{G}}$  in Fig. 4.4 (a) as the generalized version of Fig. 4.2 (a), where there is no target law dependence or missingness process dependence but only missingness interference. The graph can encode MCAR (only solid edges), MAR (with blue dotted edges) and MNAR (with brown dashed edges) processes, just as before. Graphs in Fig. 4.4 (b)-(d) are generalized versions of graphs in Fig. 4.2 (b)-(d). Graphs  $\tilde{\mathcal{G}}$  are useful for illustrative purposes as they are compact, but as we show in Section 4.5.2, identification is more straightforward in models where counterfactuals do *not* share identical edges. Hence, we will use graphs  $\mathcal{G}$ , like those in Fig. 4.3 (and not  $\tilde{\mathcal{G}}$  in Fig. 4.4), in the remainder of the paper.

We have looked at examples where all potentially missing variables of a unit are either all observed or all missing, meaning only a single missingness indicator per unit is needed. In general, this may not be true. This gives rise to two different complications: (i) there could be edges from  $R_{Z_j}$  to  $Z'_i$  for some units  $i, j$  and variables  $Z, Z' \in \mathbf{Z}$ , and (ii) edges from a missingness indicator of a unit to another variable in that same unit (e.g.  $R_{Z_i} \rightarrow Z'_i$ ) might be present. If (i) occurs, we cannot index counterfactuals simply by the missingness indicators of other units as we had done before, but of specific variables of those units. To be consistent with our prior definition, we call a unit the affector of another if *any* missingness indicator of the first unit indexes the counterfactual of the latter unit. If (ii) occurs, the missingness interference is within the

unit, and not across units, making the data i.i.d. in the absence of other interactions. In both of these cases, the graphical structure and mathematical framework remain the same as what we have discussed so far. What changes is the interpretation of relationships between units and their recorded variables.

## 4.5 Identification Results

Having set up the notation and graphical framework for entangled missingness, we next discuss identification in these models.

### 4.5.1 Without Missingness Interference

We recognized in Sec. 4.4.1 that the (graphical) models in the absence of missingness interference are essentially identical to existing i.i.d. models for missing data, with the only difference arising from how we view the smallest unit of investigation - whether it is one individual or a set of individuals interacting within a block.

The parameter of interest in a missing data model is generally of the form  $\beta = \mathbb{E}[h(\mathbf{Z}^{(1)}, \mathbf{O})]$  for some known function  $h$  of  $\mathbf{Z}^{(1)}$ , but identification results will be for the full law  $p(\mathbf{Z}^{(1)}, \mathbf{O}, \mathbf{R})$ . If this full law is nonparametrically identified from the observed data then so is any functional of it.

A *sound* and *complete* algorithm for full law identification in i.i.d. missing data DAG models with fully observed variables has been proposed in [Nabi et al., 2020].<sup>4</sup> The authors also provide full law identification in missing data models with hidden variables in the same work. For both types of graphs, DAGs and ADMGs, their identification criterion relies on the notion of the Markov blanket [Pearl, 1988, Richardson et al., 2017]: in an ADMG  $\mathcal{G}$ , the Markov blanket of a vertex  $V$ , denoted by  $\text{mb}_{\mathcal{G}}(V)$ , consists of all variables sharing an edge with  $V$  or with a collider path to  $V$ ; a collider path is

---

<sup>4</sup>A sound and complete algorithm corresponds to necessary and sufficient identification assumptions.

path where all vertices on the path are colliders of the form  $\rightarrow o \leftarrow, \leftrightarrow o \leftarrow, \leftrightarrow o \leftrightarrow$ . In a DAG, which is an ADMG with no bidirected edges, the Markov blanket reduces to the set of variables with an edge in common with  $V$ , and variables that share a child with  $V$ . We provide a brief description of their results in Theorem 1 and Theorem 2 below, as they apply directly to entangled missingness settings without missingness interference.

*Theorem 1.* [Full law identification in DAGs [Nabi et al., 2020]]

In a missing data model, represented by DAG  $\mathcal{G}$ , the full law  $p(\mathbf{R}, \mathbf{Z}^{(1)}, \mathbf{O})$  is identified if and only if  $Z^{(1)} \notin \text{mb}_{\mathcal{G}}(R_Z), \forall Z^{(1)} \in \mathbf{Z}^{(1)}$ . Thus, for the full law to be identified, no edge of the form  $Z^{(1)} \rightarrow R_Z$  can be present (no *self-censoring*) and no structure of the form  $Z^{(1)} \rightarrow R_Z \leftarrow R_Z$  can be present (no *colluders*). The identifying functional is given by Eq. 2.9, where the missingness mechanism  $p(\mathbf{R} \mid \mathbf{Z}^{(1)}, \mathbf{O})$  is given by an odds ratio parameterization [Chen, 2007]:

$$\frac{1}{\sigma(\mathbf{Z}^{(1)}, \mathbf{O})} \times \prod_{k=1}^K p(R_k \mid \mathbf{R}_{-k} = 1, \mathbf{Z}^{(1)}, \mathbf{O}) \times \prod_{k=2}^K \text{OR}(R_k, R_{\prec k} \mid R_{\succ k} = 1, \mathbf{Z}^{(1)}, \mathbf{O}), \quad (4.2)$$

where  $\mathbf{R}_{-k} = \mathbf{R} \setminus R_k, R_{\prec k} = \{R_1, \dots, R_{k-1}\}, R_{\succ k} = \{R_{k+1}, \dots, R_K\}$ ,

$$\begin{aligned} \text{OR}(R_k, R_{\prec k} \mid R_{\succ k} = 1, \mathbf{Z}^{(1)}, \mathbf{O}) \\ = \frac{p(R_k \mid R_{\succ k} = 1, R_{\prec k}, \mathbf{Z}^{(1)}, \mathbf{O})}{p(R_k = 1 \mid R_{\succ k} = 1, R_{\prec k}, \mathbf{Z}^{(1)}, \mathbf{O})} \times \frac{p(R_k = 1 \mid R_{-k} = 1, \mathbf{Z}^{(1)}, \mathbf{O})}{p(R_k \mid R_{-k} = 1, \mathbf{Z}^{(1)}, \mathbf{O})}, \end{aligned}$$

and  $\sigma(\mathbf{Z}^{(1)}, \mathbf{O}) = \sum_r \{\prod_{k=1}^K p(r_k \mid R_{-k} = 1, \mathbf{Z}^{(1)}, \mathbf{O}) \times \prod_{k=2}^K \text{OR}(r_k, r_{\prec k} \mid R_{\succ k} = 1, \mathbf{Z}^{(1)}, \mathbf{O})\}$  is the normalizing function.

Consider the i.i.d. graph in Fig. 2.5(d). The odds ratio parameterization of the score  $p(R_A, R_Y \mid \mathbf{C}, A^{(1)}, Y^{(1)})$  is identical to its regular DAG parameterization since  $p(R_A \mid R_Y = 1, \mathbf{C}, A^{(1)}, Y^{(1)}) = p(R_A \mid \mathbf{C}), p(R_Y \mid R_A = 1, \mathbf{C}, A^{(1)}, Y^{(1)}) = p(R_Y \mid \mathbf{C})$ , and  $\text{OR}(R_A, R_Y \mid Y^{(1)}, A^{(1)}, \mathbf{C}) = \sum_{R_Y, R_A} p(R_A \mid \mathbf{C}) \times p(R_Y \mid \mathbf{C}) = 1$ , and the normalizing term is one. Self-censoring edges  $A^{(1)} \rightarrow R_A$  and  $Y^{(1)} \rightarrow R_Y$  prevent identification of the full law in Fig. 4.1 (a).

The following theorem summarizes the full law identification results in [Nabi et al., 2020] for ADMGs obtained as latent projections of hidden variable missing data DAGs.

*Theorem 2.* [Full law identification in ADMGs [Nabi et al., 2020]]

In a missing data model represented by a hidden variable DAG  $\mathcal{G}(\mathbf{R}, \mathbf{Z}^{(1)}, \mathbf{O}, \mathbf{H})$  and its latent projection ADMG  $\mathcal{G}(\mathbf{R}, \mathbf{Z}^{(1)}, \mathbf{O})$ , the full law  $p(\mathbf{R}, \mathbf{Z}^{(1)}, \mathbf{O})$  is identified if and only if  $Z^{(1)} \notin \text{mb}_{\mathcal{G}}(R_Z), \forall Z^{(1)} \in \mathbf{Z}^{(1)}$ . Thus, for the full law to be identified, no pair  $(Z^{(1)}, R_Z)$  should be connected directly (a.k.a. no *self-censoring*) or through a collider path (a.k.a. no *colluding paths*). Moreover, the identification of the missingness mechanism is given by the odds ratio parameterization, as stated in Theorem 1.

None of the graphs in Fig. 4.1 have a colluding path. However, as examples, if the path  $A_1^{(1)} \rightarrow R_{Y_1} \leftarrow R_{A_1}$  or the path  $A_1^{(1)} \leftrightarrow A_2^{(1)} \leftrightarrow R_{A_1}$  were to exist in any of the graphs, then the full law would not be identified.

Settings with entanglements but no missingness interference are amenable to these (existing) identification results, but valid identification requires careful consideration of the dependence engendered by the entanglements, and whether they introduce self-censoring or colluding paths in the corresponding graph.

## 4.5.2 With Missingness Interference

In settings with missingness interference, the joint distribution over all counterfactuals, i.e. *the full law*, involves counterfactual variables from multiple *worlds*, like  $Y_1^{(1, r_2=0)}$  and  $Y_1^{(1, r_2=1)}$  in the case of Anne in Section 4.4.2. This implies that the full law is not identified without very strong assumptions<sup>5</sup>. Instead, we will consider identification of *single-world* objects.

---

<sup>5</sup>See discussion on rank preservation [Hernan and Robins, 2020] for an example of a type of strong assumption relating counterfactuals across worlds.

Single-world objects in general missing data networks are denoted by  $h(\tilde{\mathbf{Z}}; \mathbf{r})$ , where  $\tilde{\mathbf{Z}} \equiv \{\tilde{\mathbf{Z}}_i : i \in \{1, 2, \dots, n\}\}$  is the set of all counterfactuals. Single-world objects are composed of counterfactuals  $\{Z_i^{(1, \mathbf{r}_{\text{aff}(i)})} : i \in \{1, 2, \dots, n\} \text{ and } \mathbf{R} = \mathbf{r} \text{ such that } r_i = 1 \text{ for any } i \text{ in this set}\}$ . For example, in a 3-unit network, where  $\mathbf{r}_{\text{aff}(1)} = (r_2, r_3)$ ,  $\mathbf{r}_{\text{aff}(2)} = (r_1, r_3)$  and  $\mathbf{r}_{\text{aff}(3)} = (r_1, r_2)$ , we might choose  $\mathbf{r} = (r_1, r_2, r_3) = (1, 0, 1)$ . Then,  $p(Z_1^{(1, r_2=0, r_3=1)}, Z_3^{(1, r_1=1, r_2=0)})$  is a valid single-world object. No valid single-world object can include  $Z_2^{(1, \mathbf{r}_{\text{aff}(2)})}$  as  $r_2 = 0$ , and  $Z_2^{(1, \mathbf{r}_{\text{aff}(2)})}$  are not consistent with  $\mathbf{r}$  for any  $\mathbf{r}_{\text{aff}(2)}$ .

The first identification result we outline shows that, analogues of MCAR and MAR models when missingness interference is present, yield non-parametric identification for any single-world object  $h(\tilde{\mathbf{Z}}; \mathbf{r})$ .

*Theorem 3.* In a missing data ADMG  $\mathcal{G}$  with missingness interference, valid single-world objects  $h(\tilde{\mathbf{Z}}; \mathbf{r})$  consisting of a set of counterfactuals  $\mathbf{Z}' \equiv \bigcup_i \{\mathbf{Z}_i^{(1, \mathbf{r}_{\text{aff}(i)})}\}$ ,  $i \in \{1, \dots, n\}$  are identified when either of these two conditions is satisfied: (1)  $\mathbf{R}' \perp\!\!\!\perp \mathbf{O}, \tilde{\mathbf{Z}}$  (MCAR), or (2)  $\mathbf{R}' \perp\!\!\!\perp \tilde{\mathbf{Z}} | \mathbf{O}$  (MAR), where  $\mathbf{R}'$  refers to the set of all missingness indicators  $R$  that index counterfactuals in  $h(\tilde{\mathbf{Z}}; \mathbf{r})$ . The object  $h(\tilde{\mathbf{Z}}; \mathbf{r})$  is a function of  $p(\mathbf{Z}', \mathbf{R}, \mathbf{O})$ , and the identifying functional is given by:

$$p(\mathbf{Z}', \mathbf{R}, \mathbf{O}) = p(\mathbf{Z}', \mathbf{O}) \times p(\mathbf{R} | \mathbf{O}, \mathbf{Z}') = \frac{p(\mathbf{Z}', \mathbf{R} = \mathbf{r}, \mathbf{O})}{p(\mathbf{R} = \mathbf{r} | \mathbf{O})} \times p(\mathbf{R} | \mathbf{O}) \quad (4.3)$$

where propensity scores are obtained by simple m-separation or (d-separation) rules on ADMG (or DAG) factorization.

*Proof:* See Appendix for proof. □

If the missing data model corresponds to MNAR, single-world objects are not always identified. But a special single-world object called the *full-observability law*,  $P(\tilde{\mathbf{Z}}^{(\mathbf{r}=\mathbf{1})}, \mathbf{R})$ , where  $\tilde{\mathbf{Z}}^{(\mathbf{r}=\mathbf{1})} \equiv \{\mathbf{Z}_i^{(1, \mathbf{r}_{\text{aff}(i)}=\mathbf{1})} : i \in \{1, 2, \dots, n\}\}$  is the set of all counterfactuals where missingness status is set to full observability (i.e, missingness pattern is

1), can in fact be identified under certain assumptions, as we will soon show. The full observability law might be thought of as a close analogue to the full law in i.i.d settings. It corresponds to the distribution  $p(Y_1^{(1,r_2=1)}, Y_2^{(1,r_1=1)}, R_1, R_2)$  in our example with Anne and Bob. We outline our assumptions and definitions before presenting our result for identifying the full observability law in models for entangled missingness corresponding to MNAR.

Let the set of all counterfactuals corresponding to patterns with at least one zero be denoted by  $\tilde{\mathbf{Z}}^{(\mathbf{r} \neq \mathbf{1})} = \tilde{\mathbf{Z}} \setminus \tilde{\mathbf{Z}}^{(\mathbf{r} = \mathbf{1})}$ .

**Assumption 4.** *In a missing data ADMG  $\mathcal{G}$  with missingness interference,  $\text{ch}_{\mathcal{G}}(\tilde{\mathbf{Z}}^{(\mathbf{r} \neq \mathbf{1})}) \cap \mathbf{R} = \emptyset$ . Here, the definition  $\text{ch}_{\mathcal{G}}(V)$  applies disjunctively to a set  $\mathbf{V}$ , i.e.,  $\text{ch}_{\mathcal{G}}(\mathbf{V}) = \bigcup_{V \in \mathbf{V}} \text{ch}_{\mathcal{G}}(V)$ .*

In other words, the only type of counterfactual that can be a parent of any  $R_i \in \mathbf{R}$  has the form  $Z_j^{(1, \mathbf{r}_{\text{aff}(j)} = \mathbf{1})}$ , for not necessarily distinct  $i, j$ .

We also define the following new entities to be used in our results:

1. *e-colluding path*: The pair  $(Z_i^{(1, \mathbf{r}_{\text{aff}(i)} = \mathbf{1})}, R_k)$  have an (extended- or) e-colluding path if  $Z_i^{(1, \mathbf{r}_{\text{aff}(i)} = \mathbf{1})}$  and  $R_k$  are connected through at least one colliding path<sup>6</sup> that does not go through an observed proxy, and  $R_k \in \mathbf{R}_{\text{aff}(i)}$
2. *e-colluder*:  $Z_i^{(1, \mathbf{r}_{\text{aff}(i)} = \mathbf{1})} \rightarrow R_j \leftarrow R_k$ , where  $R_j \notin \mathbf{R}_{\text{aff}(k)}$  and  $R_k \in \mathbf{R}_{\text{aff}(i)}$
3. *e-self-censoring* or *affector-censoring*:  $Z_i^{(1, \mathbf{r}_{\text{aff}(i)} = \mathbf{1})} \rightarrow R_j \leftarrow R_k$  where  $R_k \in \mathbf{R}_{\text{aff}(i)}$

**Theorem 4.** In a missing data ADMG  $\mathcal{G}$  with missingness interference, under Assumption 4, the full-observability law  $P(\tilde{\mathbf{Z}}^{(\mathbf{r} = \mathbf{1})}, \mathbf{R})$  is identified if and only if there is no e-colluding path. Further, if  $\mathcal{G}$  is a missing data DAG, the full-observability law is identified if and only if there is no e-colluder and no e-self-censoring. The identifying

---

<sup>6</sup>A path between vertices  $V_i$  and  $V_j$  is a colliding path if every vertex  $V_k$  in the path is a collider, i.e., bears ones of these forms: (1)  $\rightarrow V_k \leftarrow$ , (2)  $\leftrightarrow V_k \leftrightarrow$ , (3)  $\rightarrow V_k \leftrightarrow$ , (4)  $\leftrightarrow V_k \leftarrow$

functional is given by

$$p(\tilde{\mathbf{Z}}^{(r=1)}, \mathbf{R}) = \mathbf{p}(\tilde{\mathbf{Z}}^{(r=1)}) \times \underbrace{p(\mathbf{R} \mid \tilde{\mathbf{Z}}^{(r=1)})}_{\mathbf{g}(\mathbf{p}(\mathbf{R}, \mathbf{Z}))} = \frac{\mathbf{p}(\tilde{\mathbf{Z}}^{(r=1)}, \mathbf{R} = \mathbf{1})}{\underbrace{p(\mathbf{R} = \mathbf{1} \mid \tilde{\mathbf{Z}}^{(r=1)})}_{\mathbf{g}(\mathbf{p}(\mathbf{R}, \mathbf{Z}))|_{\mathbf{R}=\mathbf{1}}}} \times \underbrace{p(\mathbf{R} \mid \tilde{\mathbf{Z}}^{(r=1)})}_{\mathbf{g}(\mathbf{p}(\mathbf{R}, \mathbf{Z}))}. \quad (4.4)$$

and missingness mechanism  $p(\mathbf{R} \mid \tilde{\mathbf{Z}}^{(r=1)})$  is identified using the OR parameterization given below:

$$p(\mathbf{R} \mid \tilde{\mathbf{Z}}^{(r=1)}) = \frac{1}{\sigma} \times \prod_{k=1}^K p(R_k \mid R_{-k} = 1, \tilde{\mathbf{Z}}^{(r=1)}) \times \prod_{k=2}^K \text{OR}(R_k, R_{\prec k} \mid R_{\succ k} = 1, \tilde{\mathbf{Z}}^{(r=1)})$$

where notation and OR is consistent with Section 4.5.1.

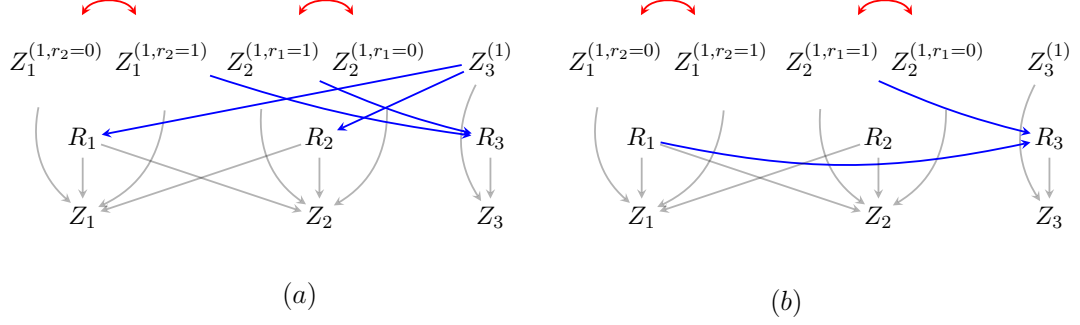
*Proof:* The proof extends the proof for sound and complete identification of the missing data full law in ADMGs in the absence of colluding paths in [Nabi et al., 2020]. Proof is in the Appendix.  $\square$

We illustrate Theorem 4 using two different examples, one where the full observability law is identified, and one where the graphical criterion is violated, and hence the law is not identified.

In the example in Fig. 4.5(a), there are no e-colluding paths. So, by Theorem 4, the full observability law  $P(Z_1^{(1, r_2=2)}, Z_2^{(1, r_1=1)}, Z_3^{(1)}, R_1, R_2, R_3)$  should be identified, and it is indeed identified as shown:

$$\begin{aligned} & P(Z_1^{(1, r_2=1)}, Z_2^{(1, r_1=1)}, Z_3^{(1)}, R_1, R_2, R_3) \\ &= \frac{p(Z_1^{(1, r_2=1)}, Z_2^{(1, r_1=1)}, Z_3^{(1)}, R_1 = 1, R_2 = 1, R_3 = 1)}{p(R_1 = 1, R_2 = 1, R_3 = 1 \mid Z_1^{(1, r_2=1)}, Z_2^{(1, r_1=1)}, Z_3^{(1)})} \\ & \quad \times p(R_1 \mid Z_3^{(1)}) \times p(R_2 \mid Z_3^{(1)}) \times p(R_3 \mid Z_1^{(1, r_2=1)}, Z_2^{(1, r_1=1)}) \\ &= \frac{p(Z_1^{(1, r_2=1)}, Z_2^{(1, r_1=1)}, Z_3^{(1)}, R_1 = 1, R_2 = 1, R_3 = 1)}{p(R_1 = 1, R_2 = 1, R_3 = 1 \mid Z_1^{(1, r_2=1)}, Z_2^{(1, r_1=1)}, Z_3^{(1)})} \\ & \quad \times p(R_1 \mid Z_3, R_3 = 1) \times p(R_2 \mid Z_3, R_3 = 1) \times p(R_3 \mid Z_1, Z_2, R_1 = 1, R_2 = 1), \end{aligned}$$

since  $R_1 \perp\!\!\!\perp R_3 \mid Z_3^{(1)}$ ,  $R_2 \perp\!\!\!\perp R_3 \mid Z_3^{(1)}$ ,  $R_3 \perp\!\!\!\perp R_1, R_2 \mid Z_1^{(1, r_2=1)}, Z_2^{(1, r_1=1)}$ , and rules of consistency allow us to replace counterfactuals by their observed proxies.



**Figure 4.5.** Two illustrations of Theorem 4. ADMG in (a) does not have a e-colluding path. ADMG in (b) features the e-colluding path  $Z_2^{(1,r_1=1)} \rightarrow R_3 \leftarrow R_1$

The full observability law is however not identified in Fig. 4.5(b) because of the colluding path  $Z_2^{(1,r_1=1)} \rightarrow R_3 \leftarrow R_1$ . We can try to understand why, using the following intuition - for identification, the missingness mechanism  $p(\mathbf{R}|\text{pa}(\mathbf{R}))$  should be identified for all levels of  $\mathbf{R}$ . Writing the missingness mechanism as  $p(R_1)p(R_2)p(R_3|R_1, Z_2^{(1,r_1=1)})$ , we realize that we cannot identify the last term unless we are able to set *both*  $R_1$  and  $R_2$  to the value 1, and by consistency, replace the counterfactual  $Z_2^{(1,r_1=1)}$  by observed proxy  $Z_2$ . Doing so in the joint would mean that we cannot identify this quantity for levels of  $\mathbf{R}$  that set  $R_1$  or  $R_2$  to 0.

The rigorous argument for non-identification involves counting the parameters required to characterize the full observability law against the observed law in a binary model, and showing that the observed law has fewer parameters and hence it is not possible to uniquely map back to the full observability law. A detailed account of parameter counting in examples with e-colluding paths has been deferred to the Appendix, under the completeness section of the proof of Theorem 4, as knowledge of the Möbius parameterization of the nested Markov model for binary variable ADMGs is required to follow the procedure for parameter counting, which is also discussed in the Appendix.

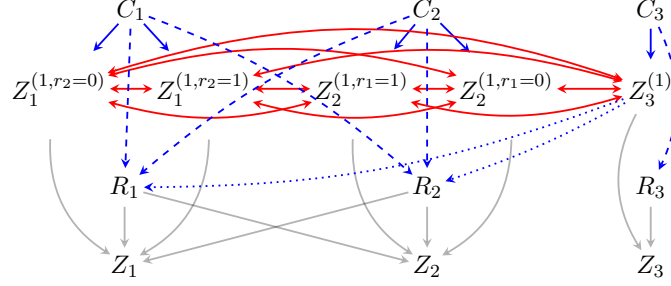
## 4.6 Experiments

We generate synthetic data and attempt to recover the ground truth in entangled missingness settings where the targets we are interested in, are identified. In particular, we reserve our interest to the cases where missingness interference ( $\mathcal{E}_3$ ) is present, as results for settings where missingness interference is absent have been discussed in literature before [Nabi et al., 2020].

Data was simulated based on the graph in Fig. 4.6. There are three units in the network, and units 1 and 2 are neighbors. When the graph consists only the solid edges, we have an MCAR scenario, and including the dashed blue edges ( $C_1 \rightarrow R_1$ ,  $C_1 \rightarrow R_2$ ,  $C_2 \rightarrow R_1$ ,  $C_2 \rightarrow R_2$ ,  $C_3 \rightarrow R_3$ ), we get a MAR scenario and finally with all the edges (adding  $Z_3^{(1)} \rightarrow R_2$ ,  $Z_3^{(1)} \rightarrow R_1$ ), we get an MNAR case.

Parameters associated with the generative model are as follows:  $P(C_1), P(C_2), P(C_3)$ , all chosen to be univariate normal distributions, counterfactuals  $Z_1^{(1,r_2=0)}, Z_1^{(1,r_2=1)}$ ,  $Z_3^{(1,r_1=0)}, Z_3^{(1,r_1=1)}$  and  $Z_3^{(1)}$  are drawn from a multivariate normal distribution. The means of the multivariate normal distribution are given by linear functions of the corresponding parents  $C$  in the graph, and covariance matrix has no zero entries; this ensures that all the counterfactuals are associated with each other. Missingness indicators  $R_i$  are generated according to the scenario, using a binomial distribution where the binomial probability is chosen completely at random (MCAR) or is a logistic sigmoid of a linear function of the parents of  $R_i$  (MAR, MNAR). Observed proxies  $Z_i$  are a deterministic function of the counterfactuals and missingness indicators, as given by the consistency assumption.

We are interested in identifying the following functionals:  $\mathbb{E}[Z_1^{(1,r_2=0)}]$ ,  $\mathbb{E}[Z_1^{(1,r_2=1)}]$ ,  $\mathbb{E}[Z_2^{(1,r_1=0)}]$ ,  $\mathbb{E}[Z_2^{(1,r_1=1)}]$ , under MCAR, MAR and MNAR conditions. In the given graph, since there are no e-colluders, we should be able to identify the full observability



**Figure 4.6.** The model used to generate synthetic data for our experiments.

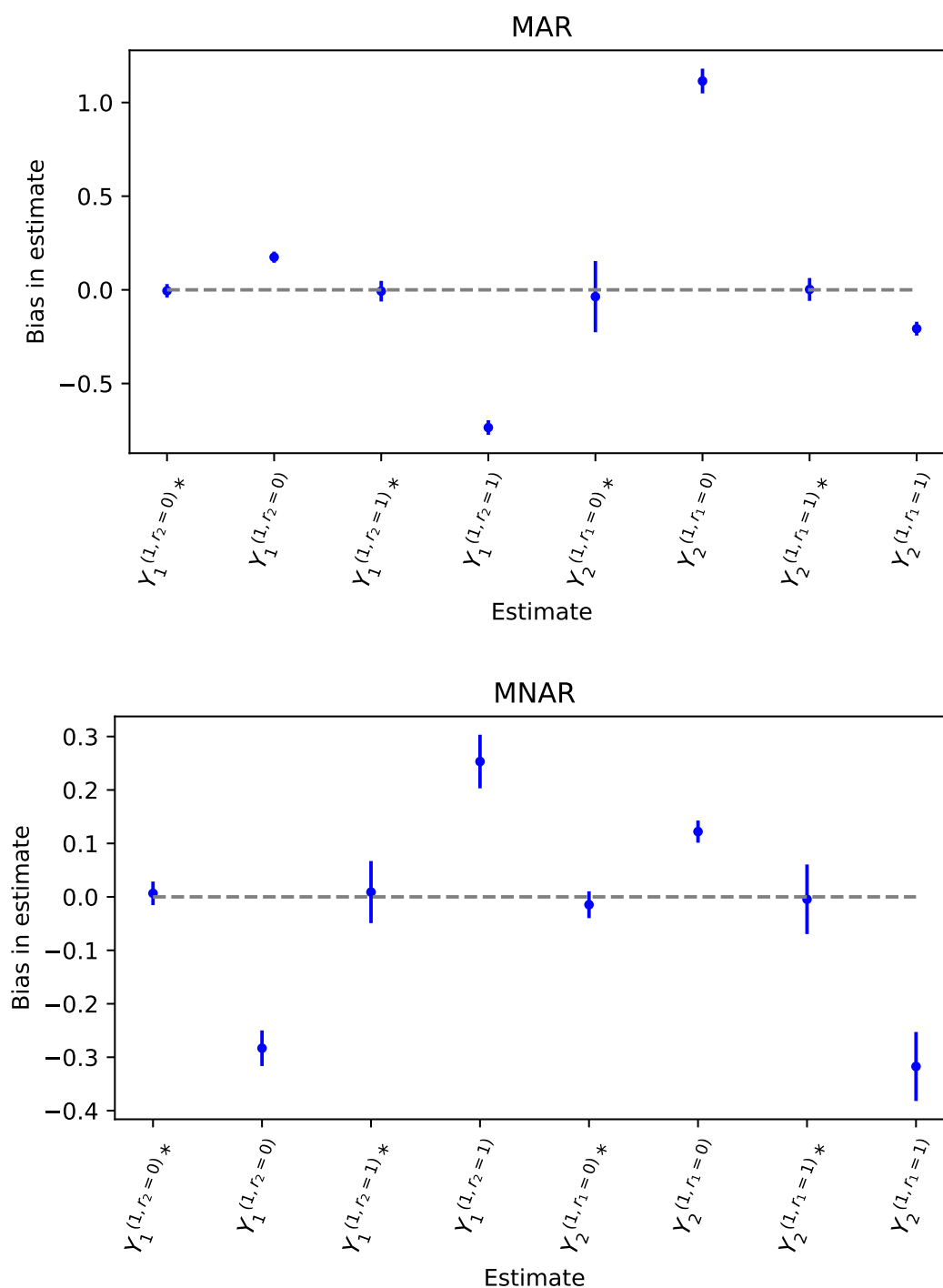
law (or its marginals) even in the MNAR case<sup>7</sup>. We use an estimating equation of the form  $\mathbb{E}[\frac{\mathbb{I}(\mathbf{R}=\mathbf{1})}{\pi(\mathbf{C}, \tilde{\mathbf{Z}})} h(\tilde{\mathbf{Z}})]$  where  $\pi(\mathbf{C}, \tilde{\mathbf{Z}})$  is the inverse weight function composed of propensity scores and  $h(\cdot)$  is the functional we are interested in. The estimator uses regression models for the weights, and expectations are done empirically.

A total of 50000 samples were generated from the network, and estimation was done over a bootstrap of 50 samples. The results from these experiments are shown in Fig. 4.7. We compare the bias of an estimate against the ground truth (which we have from model parameters), in two different scenarios - when we adjust for the network structure and the entanglement appropriately, and one where we do not, for MCAR, MAR and MNAR. In MCAR, the bias would be the same for the two different approaches since we can treat the complete rows in the dataset as an unbiased, representative dataset. However, for MAR and MNAR situations, estimates based on complete rows without adjustment would be biased, as shown in the figure. Estimates obtained after adjustment are labeled with an asterisk (\*) in the x-axis.

## 4.7 Conclusions

In this work, we developed a comprehensive hierarchy of entangled missingness to understand settings with dependent data and missingness, using the language of causal graphical models. We recognized that the (conventional) full law distribution in the

<sup>7</sup>In this particular example,  $\mathbb{E}[Z_1^{(1,r_2=0)}]$  and  $\mathbb{E}[Z_2^{(1,r_1=0)}]$  are also identified by simple graph factorization and m-separation



**Figure 4.7.** Bias recorded in bootstrapped estimates of targets (shown in x-axis), MAR case (left) and MNAR (right). We compare our adjusted IPW estimates (denoted by an asterisk (\*) on the x-axis) to the unadjusted estimate, which is obtained by ignoring the network missingness structure underpinning the data. Error bars represent quantiles  $q_{0.05}$  and  $q_{0.95}$  across 50 bootstrap samples.

presence of missingness interference is in general not identified because of the presence of cross-world counterfactuals. Hence, we developed sound and complete identification results via a graphical criterion for the full observability law, a marginal of the full law. We demonstrate, using synthetic experiments, that if we do not account for the right kind of entanglement, estimates are biased. Exploring nonparametric identification for other margins (including joints that have counterfactuals for when at least one unit is missing) is a promising area of future work.

# Chapter 5

## Generalized Coarsening

### 5.1 Introduction

Causal inference methods generally involve making inferences about a parameter in a larger full data distribution on counterfactual random variables from a smaller observed data distribution, as we noted right in the beginning, in Section 1. Typically, the latter is obtained from the former by a combination of marginalization (applied to unobserved variables), and *coarsening* (which relates observed and counterfactual variables). However, many causal systems are not well-described via a coarsened relationship between the full and the observed data distribution. Instead, we might have a more *generalized* form of coarsening in them. In particular, these systems might be composed of “macrostates” consisting of measurable variables on which data is available, and “microstates” with dynamics operating at a finer temporal granularity, that *lead* to observables in macrostates. This is similar to descriptions in statistical physics literature, of microstates involving particle motion leading to observables such as pressure or temperature [Touchette, 2015]. Such a description can also be used to model processes such as viral infection dynamics where microstates correspond to individual disease status that evolves according to transmission dynamics, and macrostates correspond to population-level disease incidence and contagion levels measured by metrics such as  $r_0$ , the average number of infections an infected person is

expected to cause among susceptible individuals [Ridenhour et al., 2014, Coburn et al., 2009]. In this work, we will take the illustrative example of cellular differentiation and reprogramming, where microscopic cellular processes are investigated using macroscopic measurements involving RNA sequencing (see Section 5.3).

It is well known in the statistical physics literature [Reif, 1965] that in a system described by a relationship of microstates and macrostates, information about the former is not, in general, possible to obtain using data on the latter. This is simply due to the fact that many possible configurations of microstates can lead to the same observable macrostate. In statistical terms, microstate parameters are *not identified* from data on macrostates. We shall use the language of graphical causal models to pose the problem of inferring causal microstate parameters from data on macrostates as a causal identification problem, and show that under some conditions, interesting microstate information may be obtained from macrostate data.

### 5.1.1 Contributions

In this work, we define the *punctuated causal model (PCM)* that links causal processes corresponding to microstates with observables in macrostates, that are a result of the microstates reaching equilibria via discrete time dynamics. While in general it is not possible to reason about microstates, we show that under some assumptions, we are able to reason about (and identify) microstate counterfactuals from the data on macrostates alone, generalizing the g-formula [Pearl, 2009] in fully observed causal models. We discuss maximum likelihood and semi-parametric estimation for identified parameters, and illustrate our proposed framework with experiments on synthetic data as well as data from cellular reprogramming experiments.

The author would like to note that they share credit for this work with collaborator Numair Sani, who will present some of the results (not discussed in detail here) in his own dissertation. This will include expounding on identification results and proofs,

PCMs for hidden variable settings, as well as a recursive formulation for estimation functionals.

## 5.2 Background

We will use temporal causal models for discrete time systems, described in Section 2.3, as the foundation for developing the punctuated causal model. We include a few additional details that are relevant to the discussion in this chapter, below.

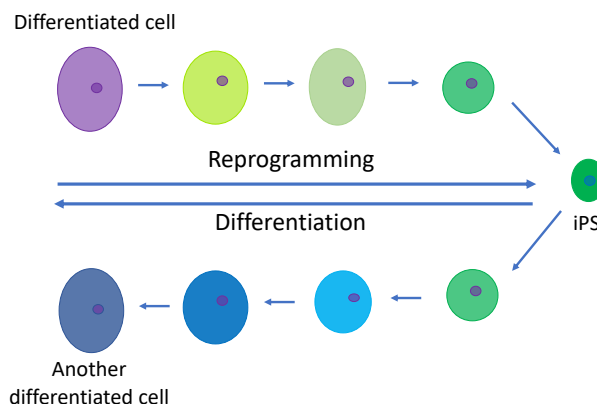
A causal CDBN allows interventions on any subset of variables in the model. But, in this work we restrict ourselves to interventions on the same set of variables in each time point. That is, given  $\mathbf{A} \subseteq \mathbf{V}$ , we denote  $\mathbf{A}_1, \dots, \mathbf{A}_T$  as copies of  $\mathbf{A}$  indexed by time, and  $\bar{\mathbf{A}}_t = (\mathbf{A}_1, \dots, \mathbf{A}_t)$  for any  $t = 1, \dots, T$ . We define  $\bar{\mathbf{V}}_t$  and  $\bar{\mathbf{Y}}_t$  similarly for  $\mathbf{V}_t$  and  $\mathbf{Y}_t = \mathbf{V}_t \setminus \mathbf{A}_t$ , respectively, for any  $t = 1, \dots, T$ . The resulting counterfactual distribution  $p(\bar{\mathbf{Y}}_T(\bar{\mathbf{a}}_T)|\mathbf{W})$  factorizes into a product of counterfactual distributions associated with the causal model corresponding to  $\mathcal{G}_1$  representing the hypothetical outcomes at the first time point, and a set of “unrolled” causal models corresponding to  $\mathcal{G}_{+1}$ , representing hypothetical outcomes at all subsequent time points. Specifically, we have the following factorization:

$$p(\bar{\mathbf{Y}}_T(\bar{\mathbf{a}}_T)|\mathbf{W}) = p(\mathbf{Y}_1(\bar{\mathbf{a}}_1)|\mathbf{W}) \prod_{t=1}^{T-1} p(\mathbf{Y}_{t+1}(\bar{\mathbf{a}}_{t+1})|\mathbf{Y}_t(\bar{\mathbf{a}}_t) \cup \mathbf{W}),$$

with this (unrolled) counterfactual distribution identified by the appropriate generalization of the g-formula.

## 5.3 Motivating Example: Cellular Differentiation

To ground our discussion, we will use the illustrative example of cellular differentiation. Biological cells undergo *differentiation*, a process in which undeveloped cells multiply and transform into highly specialized cells that perform functions in multicellular organisms [Alvarado and Yamanaka, 2014]. It involves changes in cell size, shape,



**Figure 5.1.** Reprogramming requires a reversal (above) of a differentiated cell into a pluripotent stem cell (i.e. iPS), which next may undergo a differentiation process (below) into another differentiated cell.

membrane potential, metabolic activity, and responsiveness to signals, which are largely driven by controlled modifications in gene expression. Conversely, modern epigenetics allows scientists to manipulate and reverse the differentiation process [Cieřlar-Pobuda et al., 2017]. By using forced gene expression, highly differentiated and specialized somatic cells (e.g. fibroblasts) can be *reprogrammed* into induced pluripotent stem cells (iPSCs) which resemble embryonic cells, which can then become any kind of specialized cell. These processes are widely used in therapeutics for disease modeling, regenerative medicine, and drug discovery. Fig. 5.1 shows the processes of differentiation and reprogramming, and is reproduced from [Cieřlar-Pobuda et al., 2017].

Reprogramming is a stochastic, heterogeneous process, with substantial variation in the speed and path that cells take to convert to their ultimate state, commonly referred to as cell “fate” [Schiebinger et al., 2019], at *equilibrium*.<sup>1</sup> The efficiency of the process depends on the starting cell population, the bio-chemical environment of the cells and other factors that scientists are still studying today [Francesconi et al., 2019, Schiebinger et al., 2019]. Many questions about cellular differentiation remain

<sup>1</sup>We use equilibrium as in statistical mechanics, when the observable macrostate of a system is time-invariant [Reif, 1965].

unresolved. Do cells convert via homogenous or heterogenous pathways? Do all cells convert with the same speed? What are the determinants of variation in the speed and path of conversion?

Mechanisms of cellular differentiation are often studied by experiments, where somatic cells (from model animals) are subjected to an array of transcription factors (Oct4, Sox2, c-Myc, and Klf4) in vitro, which have been shown to induce cell conversion, over many days [Takahashi and Yamanaka, 2006]. These factors might be added according to specific protocols to study the effect of such manipulations on cell conversion. *Macroscopic* cell states are measured using gene expression data, collected at intervals of hours or days. However, relevant cellular processes leading to these observed macrostates occur at much faster timescales, due to *microscopic* intracellular interactions. Further, differentiation and reprogramming are documented to involve “milestone” events along the way to ultimate cell fates, each of which might be approximated as an *equilibrium* attained by the system, locally in time.<sup>2</sup> Thus, data from such cellular differentiation experiments can be naturally understood as being composed of milestone measurements from macrostates after interventions on microstate cellular processes have been performed [MacArthur and Lemischka, 2013].

Modelled in this way, questions about cellular differentiation can be investigated using tools from causal inference, provided a causal model is formulated that can link microstate parameters, representing cellular responses to transcription factors introduced via an intervention, with macrostate measurements, representing recorded and observed data. *Associative* temporal graphical models, including Markov chain models have been used extensively to study cellular differentiation [Yates et al., 2017,

---

<sup>2</sup>It is approximate as the system does not attain total time-invariance, and does undergo further cellular differentiation. But, we believe this approximation can be justified due to the relative timescales of the observed macrostates against those of intracellular processes, and is useful in modeling system evolution.

Hu et al., 2011]. Our model is explicitly causal, allowing us to take advantage of available experimental data by considering responses to *hypothetical* interventions, and thus allowing useful hypotheses to be generated for subsequent experimentation.

## 5.4 Punctuated Causal Models

To presage what is to come, note that we will build our punctuated causal model (PCM) using building blocks called *equilibrium causal processes* (ECPs), each of which represents a process that runs to equilibrium with a set of observable macrostate variables, and causally influences downstream ECPs only via these variables. In particular, we use a causal DAG to represent causal connections among observable macrostate variables. But these connections do not necessarily correspond to parent-child relationships arising from structural equation mechanisms, as in a typical causal model. Instead, they are mediated by discrete time dynamics of microstates that reach an equilibrium within an ECP, resulting in observable macrostate variables. Microstate dynamics of an ECP may potentially be influenced by macrostate variables that occur prior to that ECP. We call our model *punctuated* because the causal connections involve multiple equilibria proceeding in “fits and starts,” by analogy with the concept of punctuated equilibrium in evolutionary biology [Gould and Gould, 2009]. First, we define all these entities associated with a general PCM in a mathematically rigorous fashion, followed by a discussion on the corresponding counterfactuals of interest in a causal analysis of such systems. Then, in Section 5.4.1, we discuss a special case of equilibrium dynamics that yields identification.

We describe discrete time microstate dynamics by means of a special type of causal Bayesian network: the *equilibrium causal process* (ECP).

*Definition 1* (equilibrium causal process (ECP)). An equilibrium causal process is a causal CDBN associated with CDAGs  $\mathcal{G}_{+1}(\mathbf{V}_{+1}, \mathbf{V}_1 \cup \mathbf{W})$  and  $\mathcal{G}_1(\mathbf{V}_1, \mathbf{W})$  such that

for any  $\mathbf{A} \subseteq \mathbf{V}$  and any value  $\mathbf{w}$  of  $\mathbf{W}$ , the distribution  $p(\mathbf{Y}_T(\bar{\mathbf{a}}_T)|\mathbf{W} = \mathbf{w})$  with probabilities  $p(\mathbf{Y}_T(\bar{\mathbf{a}}_T) = \mathbf{y}|\mathbf{W} = \mathbf{w})$  defined as

$$\sum_{\mathbf{y}_1(\bar{\mathbf{a}}_1), \mathbf{y}_2(\bar{\mathbf{a}}_2), \dots, \mathbf{y}_{T-1}(\bar{\mathbf{a}}_{T-1})} p(\mathbf{y}_1(\bar{\mathbf{a}}_1)|\mathbf{w}) \times \left( \prod_{t=1}^{T-1} p(\mathbf{y}_{t+1}(\bar{\mathbf{a}}_{t+1})|\mathbf{y}_t(\bar{\mathbf{a}}_t) \cup \mathbf{w}) \right)$$

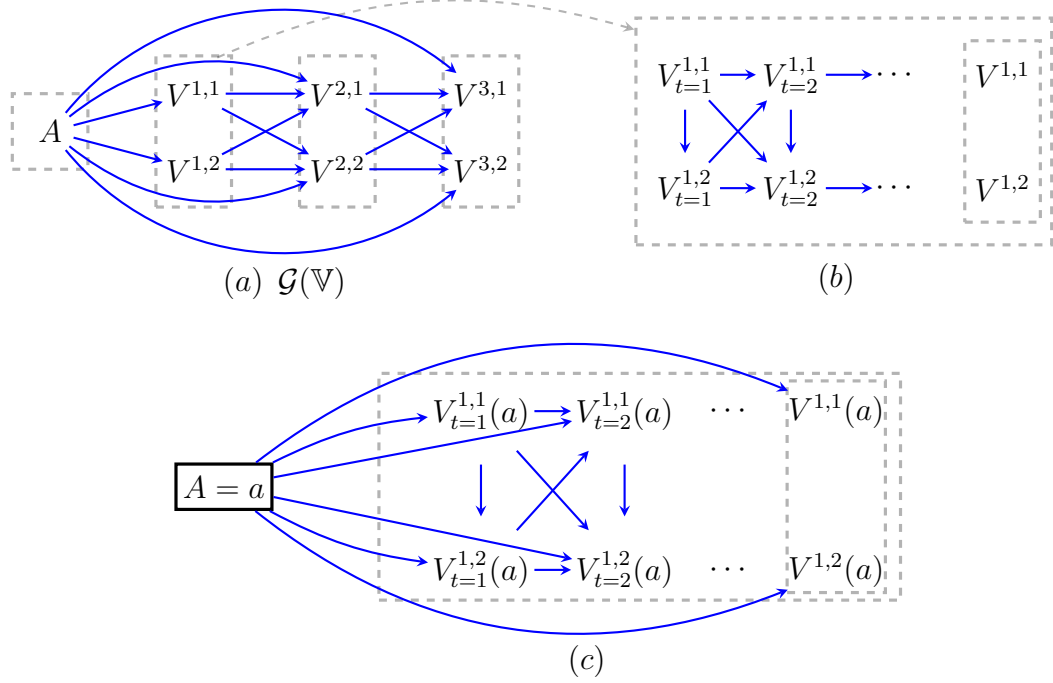
converges to a unique equilibrium distribution  $p(\mathbf{Y}(\mathbf{a})|\mathbf{W} = \mathbf{w})$  as  $T \rightarrow \infty$ . Here the set of random variables  $\mathbf{Y}(\mathbf{a})$  is interpreted to mean the set of outcomes  $\mathbf{Y}$  at equilibrium, had the intervention  $\mathbf{A} = \mathbf{a}$  taken place at every time point.

In words, an ECP is a causal CDBN such that every possible intervention (including the empty intervention) leads to a unique intervention-specific equilibrium distribution.

*Definition 2* (Microstate And Macrostate Variables). A microstate variable is a random variable indexed by time  $t$ , generated according to the structural equations associated with the CDAGs  $\mathcal{G}_{+1}(\mathbf{V}_{+1}, \mathbf{V}_1 \cup \mathbf{W})$  and  $\mathcal{G}_1(\mathbf{V}_1, \mathbf{W})$ . A macrostate variable is a random variable distributed according to the equilibrium distribution of its corresponding microstate variables.

To define a punctuated causal model (PCM), fix a set of macrostate variables  $\mathbf{V}$ , a partition  $\mathbb{V} = \{\mathbf{V}^1, \dots, \mathbf{V}^K\}$  of  $\mathbf{V}$ , and a DAG  $\mathcal{G}(\mathbb{V})$  representing causal connections among macrostates  $\{\mathbf{V}^1, \dots, \mathbf{V}^K\}$ . These connections will be mediated by dynamics represented by ECPs, modeled by CDAGs  $\mathcal{G}_{+1}^k(\mathbf{V}_{+1}^k, \mathbf{V}_1^k \cup \mathbf{W}^k)$ ,  $\mathcal{G}_1^k(\mathbf{V}_1^k, \mathbf{W}^k)$  for every  $k = 1, \dots, K$ , where each  $\mathbf{W}^k$  is  $\bigcup_{\mathbf{V}^j \in \text{pa}_{\mathcal{G}(\mathbb{V})}(\mathbf{V}^k)} \mathbf{V}^j$ . In other words,  $\mathbf{W}^k$  represents dependence of  $\mathbf{V}^k$  on a subset of prior macrostate variables, as encoded by  $\mathcal{G}(\mathbb{V})$ .

*Definition 3* (punctuated causal model (PCM)). A punctuated causal model (PCM) associated with  $\mathcal{G}(\mathbb{V})$  and  $\{\mathcal{G}_{+1}^k(\mathbf{V}_{+1}^k, \mathbf{V}_1^k \cup \mathbf{W}^k), \mathcal{G}_1^k(\mathbf{V}_1^k, \mathbf{W}^k) : k = 1, \dots, K\}$  associates an ECP with each pair of graphs  $\mathcal{G}_{+1}^k(\mathbf{V}_{+1}^k, \mathbf{V}_1^k \cup \mathbf{W}^k)$ ,  $\mathcal{G}_1^k(\mathbf{V}_1^k, \mathbf{W}^k)$ . In other words, macrostate variables in  $\mathbf{V}^k$  represent variables in the corresponding ECP reaching equilibrium.



**Figure 5.2.** (a) A simple PCM with four ECPs (grey dashed blocks), including one trivial ECP for treatment  $A$ , and three observable macrostates (milestones) of cell differentiation. (b) Unrolled ECP, which corresponds to the first milestone (endogenous Oct4 expression) in the PCM in (a). The ECP contains microstate variables that unroll to infinite time. The grey dashed edge from (a) to (b) is not part of the causal diagram, and is used only to indicate that (b) is an unrolled version of an ECP component of PCM in (a). (c) Microstate counterfactuals that result from an intervention  $A = a$ :  $V_{t=1}^{1,1}(a)$  is neural cell identity at time point  $t = 1$  had the transcription factor been set to  $a$  in the (sub)process when endogenous Oct4 expression is being attained.

We also allow a special “trivial” case where modeling temporal dynamics on a set of macrostate variables  $\mathbf{V}^k$  is not necessary, in which case the ECP may be replaced by an ordinary (conditional) causal model associated with a CDAG  $\mathcal{G}^k(\mathbf{V}^k, \mathbf{W}^k)$ .

We illustrate these definitions by an example shown in Fig. 5.2, which represents an idealized experiment where cellular reprogramming is affected by externally manipulating gene expression using transcription factors. Prior work in [Schiebinger et al., 2019] shows that three significant milestones are observed in reprogramming: (1) endogenous Oct4 expression marked by emergence of Oct4-EGFP+ cells, (2) initial signs of pluripotency via expression of marker genes such as Nanog, Zfp42, Dppa4

and Esrrb, (3) conversion to final iPSC fate, in that order. The treatment variable  $A$  in Fig. 5.2 corresponds to a set of transcription factors, representing one of a set of alternative experimental conditions, and (trivially) the first macrostate. The cellular response to this treatment, recorded using gene-expression profiles, is represented by three macrostates, shown as three blocks of variables:  $\mathbf{V}^1 \equiv \{V^{1,1}, V^{1,2}\}$ ,  $\mathbf{V}^2 \equiv \{V^{2,1}, V^{2,2}\}$  and  $\mathbf{V}^3 \equiv \{V^{3,1}, V^{3,2}\}$ . These blocks pertain to the three observed reprogramming milestones in temporal order. We assume that variable pairs  $V^{j,i}$ ,  $j \in \{1, 2\}$  are a simplified two-variable representation of cell-identity (e.g. neural-identity and pluripotent-identity), derived from gene-expression profiles. Cell-identity is a common concept used in computational biology to represent how *similar* a given cell is to a certain type of cell using correlational measures on gene-expression. In our example, we assume that in all three blocks  $V^{i,1}$  represents neural identity, and  $V^{i,2}$ , pluripotent for  $i \in 1, 2, 3$ .

As shown in Fig. 5.2, there are four macrostates:  $\mathbf{A}$  representing the transcription factors, and  $\mathbf{V}^1$ ,  $\mathbf{V}^2$ ,  $\mathbf{V}^3$ , representing reprogramming milestones. Microstate variables  $V_{t=1}^{1,1}$ ,  $V_{t=2}^{1,1}$ ,  $\dots$  represent the neural identity of a measured cell over (infinite) time in the first ECP corresponding to the first milestone (endogenous Oct4 expression), and microstate variables  $V_{t=1}^{1,2}$ ,  $V_{t=2}^{1,2}$ ,  $\dots$  represent pluripotent identity during the same process and timepoints. The experiment involves two more such ECPs that follow in succession, giving rise to the PCM shown. Macrostates are causally affected by temporally prior macrostates, as well as the treatment variable  $A$ . These causal connections among macrostates and the treatment are shown by the DAG  $\mathcal{G}(\mathbb{V})$  in Fig. 5.2(a), and are mediated by ECPs, with the ECP corresponding to the first macrostate displayed in its unrolled state. The precise graph topology of this unrolled ECP is, in turn, determined by its constituent CDAGs (see Fig. 5.2(b)). Note that macrostate variables  $V^{1,1}$  and  $V^{2,1}$  are obtained after the ECP converges to an equilibrium.

The behavior of the PCM, given that no interventions are performed, yields the observed data distribution where all macrostate variables  $\mathbf{V}$  in the PCM are obtained after each ECP, left in its natural state, reaches equilibrium. The following lemma gives the factorization of the observed data distribution associated with a PCM into the equilibrium distributions of its component ECPs, conditioned on the values of the mediating ECPs.

*Lemma 4.* For a value assignment  $\mathbf{v}$  to  $\mathbf{V}$ , the observed data distribution associated with a PCM is:

$$p(\mathbf{V} = \mathbf{v}) = \prod_{k=1}^K p(\mathbf{v}^k \mid \mathbf{w}^k), \quad (5.1)$$

where  $\mathbf{w}^k$  and  $\mathbf{v}^k$  are values of  $\mathbf{W}^k$  and  $\mathbf{V}^k$  consistent with  $\mathbf{v}$ .

This resembles the factorization of the observed data distribution associated with a DAG, but the difference is that, here, each factor represents the equilibrium distribution of an ECP. For example, we can factorize the observed data distribution for the PCM in Fig. 5.2 according to this lemma as:

$$\begin{aligned} & p(A = a, V^{1,1} = v^{1,1}, V^{1,2} = v^{1,2}, V^{2,1} = v^{2,1}, V^{2,2} = v^{2,2}, V^{3,1} = v^{3,1}, V_2^3 = v^{3,2}) \\ &= p(A = a) p(V^{1,1} = v^{1,1}, V^{2,1} = v^{2,1} \mid A = a) \\ & \quad \times p(V^{2,1} = v^{2,1}, V^{2,2} = v^{2,2} \mid V^{1,1} = v^{1,1}, V^{2,1} = v^{2,1}, A = a) \\ & \quad \times p(V^{3,1} = v^{3,1}, V^{3,2} = v^{3,2} \mid V^{2,1} = v^{2,1}, V^{2,2} = v^{2,2}, A = a) \end{aligned}$$

Next, we define counterfactual quantities associated with interventions on PCMs.

*Definition 4* (Macrostate Counterfactual Distributions). Fix a set of treatment variables  $\mathbf{A} \subseteq \mathbf{V}$ , and let  $\mathbf{Y} = \mathbf{V} \setminus \mathbf{A}$ . We assume that an intervention on  $\mathbf{A}$  is an operation that sets the values of the corresponding microstate variables at every time point of every relevant ECP to  $\mathbf{a}$ . A macrostate counterfactual distribution corresponding to

$\mathbf{Y}$ , had we intervened on the variables in  $\mathbf{A}$  and set them to  $\mathbf{a}$ , is defined as:

$$p(\mathbf{Y}(\mathbf{a}) = \mathbf{y}) \equiv \prod_{k=1}^K p(\mathbf{Y}^k(\mathbf{a}^k) = \mathbf{y}^k \mid \mathbf{w}^k), \quad (5.2)$$

where  $\mathbf{Y}^k = \mathbf{V}^k \setminus \mathbf{A}$ ,  $\mathbf{a}^k$  are a subset of  $\mathbf{a}$  pertaining to  $\mathbf{V}^k \cap \mathbf{A}$ ,  $\mathbf{y}^k, \mathbf{w}^k$  are values of  $\mathbf{Y}^k, \mathbf{W}^k$  consistent with  $\mathbf{y}$  and  $\mathbf{a}$ , and  $p(\mathbf{Y}^k(\mathbf{a}^k) = \mathbf{y}^k \mid \mathbf{w}^k)$  is the equilibrium distribution obtained from the  $k^{\text{th}}$  ECP, corresponding to  $\mathcal{G}_{+1}^k(\mathbf{V}_{+1}^k, \mathbf{V}^k \cup \mathbf{W}^k)$  and  $\mathcal{G}_1^k(\mathbf{V}^k, \mathbf{W}^k)$ .

In Fig. 5.2, the macrostate counterfactual distribution when  $\mathbf{A} = \{V^{1,1}\}$  is set to value  $v^{1,1}$  can be factorized as

$$\begin{aligned} p(A = a, V^{1,2}(v^{1,1}) = v^{1,2}, V^{2,1}(v^{1,1}) = v^{2,1}, V^{2,2}(v^{1,1}) = v^{2,2}, V^{3,1}(v^{1,1}) = v^{3,1}, V^{3,2}(v^{1,1}) = v^{3,2}) \\ = p(A = a)p(V^{1,2}(v^{1,1}) = v^{1,2} \mid A = a) \\ \times p(V^{2,1}(v^{1,1}) = v^{2,1}, V^{2,2}(v^{1,1}) = v^{2,2} \mid V^{1,2}(v^{1,1}) = v^{1,2}, A = a) \\ \times p(V^{3,1}(v^{1,1}) = v^{3,1}, V^{3,2}(v^{1,1}) = v^{3,2} \mid V^{2,1}(v^{1,1}) = v^{2,1}, V^{2,2}(v^{1,1}) = v^{2,2}, A = a) \end{aligned}$$

Each factor in this factorization, except the first factor corresponding to  $A$ , is obtained after the corresponding ECP reaches equilibrium given the corresponding intervention. Next, we define microstate counterfactuals.

*Definition 5* (Microstate Counterfactuals). Fix a set of treatment variables  $\mathbf{A} \subseteq \mathbf{V}$ , and fix a subset  $\mathbf{Y} \subseteq \mathbf{V} \setminus \mathbf{A}$  of outcome variables, which can be partitioned into block-specific subsets  $\mathbf{Y}^k \equiv \mathbf{V}^k \cap \mathbf{Y}$ . Fix a set of finite horizon time points  $T^k$  for each non-empty  $\mathbf{Y}^k$ , with the microstate outcomes of interest being  $\mathbf{Y}_{T^k}^k \equiv \{Y_{t=T^k}^{k,i} : Y^{k,i} \in \mathbf{Y}^k\}$ . Then a microstate counterfactual corresponds to the result of intervening on  $\mathbf{A}$ , where  $\mathbf{A}$  is set to the same value at every time point.

In order for the microstate counterfactual distribution for each  $\mathbf{Y}_{T^k}^k$  to be well-defined, we assume that for every  $\mathbf{Y}^k$ ,  $\{\mathbf{V}_j : \mathbf{Y}_j \subseteq \mathbf{Y} \setminus \mathbf{Y}^k\} \cap \text{an}_{\mathcal{G}(\mathbb{V})}(\mathbf{V}^k) = \emptyset$ , i.e., the microstate behavior of a variable that determines any macrostate variable causally relevant for

$\mathbf{Y}^k$  is rendered irrelevant. In other words, for every variable in our model, we are either interested in its microstate or macrostate behavior, but never both at once. A valid set of microstate counterfactuals could be  $V_{t=2}^{1,1}(a)$  and  $V_{t=2}^{1,2}(a)$ , with  $k = 1, T_k = 2, \mathbf{A} = A$ , if  $A$  were set to  $a$ . Before we can define the microstate counterfactual distribution, we must first define the *macrostate context*.

*Definition 6* (Macrostate Context). The macrostate context  $\mathbf{C}_{\mathbf{Y}}$  is equal to  $\text{an}_{\mathcal{G}(\mathbb{V})}(\{\mathbf{V}^k : \mathbf{Y}^k \neq \emptyset\}) \setminus \bigcup_k \{\mathbf{V}^k : \mathbf{Y}^k \neq \emptyset\}$ .

In words, the macrostate context is all variables whose macrostate behavior is needed to determine the microstate values of each element in  $\mathbf{Y}_{T^k}^k$  at the corresponding time horizon  $T^k$ . The macrostate context in Fig. 5.2 for  $\mathbf{Y} \equiv \{V^{2,1}, V^{2,2}\}$  is  $\{\{A\}, \{V^{1,1}, V^{2,1}\}\}$ .

*Definition 7* (Microstate Counterfactual Distribution). For an intervention that sets  $\mathbf{A}$  to the same value at every time point, the distribution over microstate counterfactuals  $p(\{\mathbf{Y}_{T^k}^k(\mathbf{a}) : \mathbf{Y}^k \neq \emptyset\})$  is known as the microstate counterfactual distribution. The microstate counterfactual distribution decomposes into a product of distributions corresponding to ECPs in  $\mathbf{C}_{\mathbf{Y}}$  and any  $\mathbf{V}^k$  if  $\mathbf{Y}^k \neq \emptyset$  as:

$$\begin{aligned} p(\{\mathbf{Y}_{T^k}^k(\mathbf{a}) : \mathbf{Y}^k \neq \emptyset\}) &\equiv \sum_{\mathbf{C}_{\mathbf{Y}}(\mathbf{a})} \prod_{\mathbf{V}^k \in \mathbf{C}_{\mathbf{Y}}} p(\mathbf{Y}^k(\mathbf{a}^k) = \mathbf{y}^k \mid \mathbf{w}^k) \\ &\times \prod_{\mathbf{V}^k : \mathbf{Y}^k \neq \emptyset} \left[ \sum_{\mathbf{y}_1^k(\bar{\mathbf{a}}_1^k), \mathbf{y}_2^k(\bar{\mathbf{a}}_2^k), \dots, \mathbf{y}_{T-1}^k(\bar{\mathbf{a}}_{T-1}^k)} p(\mathbf{y}_1^k(\mathbf{a}_1^k) \mid \mathbf{w}^k) \cdot \left( \prod_{t=1}^{T^k-1} p(\mathbf{y}_{t+1}^k(\bar{\mathbf{a}}_{t+1}^k) \mid \mathbf{y}_t^k(\bar{\mathbf{a}}_t^k) \cup \mathbf{w}^k) \right) \right]. \end{aligned} \quad (5.3)$$

As an illustration, the microstate counterfactual distribution for the condition  $\mathbf{A} = \{A\}$  is set to value  $a$ ,  $k = 1, T_k = 2, \mathbf{Y}^k = \{V^{1,1}, V^{1,2}\}$  is given by

$$\begin{aligned} p(A = a, V_{t=1}^{1,1}(a) = v_{t=1}^{1,1}, V_{t=1}^{1,2}(a) = v_{t=1}^{1,2}, V_{t=2}^{1,1}(a) = v_{t=2}^{1,1}, V_{t=2}^{1,2}(a) = v_{t=2}^{1,2}) \\ = 1 \times p(V_{t=1}^{1,1}(a) = v_{t=1}^{1,1}) \times p(V_{t=1}^{1,2}(a) = v_{t=1}^{1,2} \mid V_{t=1}^{1,1}(a) = v_{t=1}^{1,1}) \\ \times p(V_{t=2}^{1,1}(a) = v_{t=2}^{1,1} \mid V_{t=1}^{1,1}(a) = v_{t=1}^{1,1}), V_{t=1}^{1,2}(a) = v_{t=1}^{1,2}) \end{aligned}$$

$$\times p(V_{t=2}^{1,2}(a) = v_{t=2}^{1,2} \mid V_{t=1}^{1,1}(a) = v_{t=1}^{1,1}), V_{t=1}^{1,2}(a) = v_{t=1}^{1,2}, V_{t=2}^{1,1}(a) = v_{t=2}^{1,1})$$

In standard causal inference or missing data problems, identification aims to recover the target parameter of the full data distribution from the observed data distribution, where the latter is obtained by a *coarsening process* from the former, generally using some form of the consistency assumption [Hernan and Robins, 2020, Splawa-Neyman et al., 1990]. Indeed, had the observed data distribution consisted of conditional distributions in the conditional causal model of each ECP, identification of the pre-equilibrium and equilibrium counterfactual distributions in (5.2) and (5.3) would be easily obtained by the standard application of the g-formula.

However, we are interested in settings where data on microstate dynamics of the constituent ECPs is *not* available directly, and only summarized via the macrostate distribution in (5.1). In other words, the identification problem involving PCMs that we consider is as follows: the target parameters are functionals of (5.2) and (5.3), while the observed data distribution corresponds to (5.1).

#### 5.4.1 Identification in Special Cases

We discuss in [Sani\* et al., 2022] that, while microstate and macrostate counterfactuals are not identified from observed data on macrostates in general, identification may be obtained under additional structural assumptions on the model. We have four main results, pertaining to identification of macrostate and microstate counterfactuals in ECPs and corresponding PCMs respectively, under these assumptions. We merely state these identification results here in order to discuss estimation of identified parameters, later in Section 5.5. For a detailed account of identification and proofs of Theorems 5-8, please refer to [Sani\* et al., 2022]<sup>3</sup>.

*Definition 8* (Gibbs Compatible ECP). An ECP associated with CDAGs  $\mathcal{G}_1(\mathbf{V}_1 \cup \mathbf{W})$

---

<sup>3</sup>Co-author Numair Sani will also discuss these results in detail in his dissertation.

and  $\mathcal{G}_{+1}(\mathbf{V}_{+1}, \mathbf{V}_1 \cup \mathbf{W})$  is said to be Gibbs compatible if it satisfies the following:

$$\forall i \quad p(\mathbf{V}_{+1}^i \mid \text{pa}_{\mathcal{G}_{+1}}(\mathbf{V}_{+1}^i)) = p(\mathbf{V}^i \mid \mathbf{S}_{\mathbf{V}^i})$$

where  $\mathbf{V}^i$  is the macrostate variable corresponding to  $\mathbf{V}_{+1}^i$ , and  $\mathbf{S}_{\mathbf{V}^i}$  is the set of macrostate variables corresponding to  $\text{pa}_{\mathcal{G}_{+1}}(\mathbf{V}_{+1}^i)$ .

In Fig. 5.2, for the ECP given by  $\mathbf{V}^1 = \{V^{1,1}, V^{1,2}\}$ ,  $\mathbf{W}^1 = \{A\}$  to be Gibbs compatible, the following conditions must hold:

$$\begin{aligned} p(V_{t+1}^{1,1}, V_{t+1}^{1,2} \mid V_t^{1,1}, V_t^{1,2}, A) &= p(V_{t+1}^{1,1} \mid V_t^{1,2}, A) p(V_{t+1}^{1,2} \mid V_{t+1}^{1,1}, A) \\ p(V_{t+1}^{1,1} \mid V_t^{1,2}, A) &= p(V^{1,1} \mid V^{1,2}, A) \\ p(V_{t+1}^{1,2} \mid V_{t+1}^{1,1}, A) &= p(V^{1,2} \mid V^{1,1}, A) \end{aligned}$$

Given a Gibbs compatible ECP, macrostate counterfactuals are identified from the observed macrostate distribution  $p(\mathbf{V})$  by the following theorem:

*Theorem 5.* Fix  $\mathbf{A} \subseteq \mathbf{V}$  in a Gibbs compatible ECP represented by a pair of CDAGs  $\mathcal{G}_1(\mathbf{V}_1 \cup \mathbf{W})$  and  $\mathcal{G}(\mathbf{V}_{+1}, \mathbf{V}_1 \cup \mathbf{W})$ . The macrostate counterfactual  $\mathbf{Y}(\mathbf{a})$  where  $\mathbf{Y} := \mathbf{V} \setminus \mathbf{A}$  is identified from the observed data distribution on macrostates  $p(\mathbf{V})$  as

$$p(\mathbf{Y}(a) = \mathbf{y} \mid \mathbf{W}) = p(\mathbf{Y}^k = \mathbf{y} \mid \mathbf{A}^k = \mathbf{a}^k, \mathbf{W})$$

Since multiple microstates can correspond to the same observed macrostate, microstate counterfactuals associated with a Gibbs compatible ECP are not identified from the observed macrostate distribution alone. But, if  $p(\mathbf{V}_1 \mid \text{pa}_{\mathcal{G}_1}(\mathbf{V}_1))$  is known, the following result holds:

*Theorem 6.* Fix  $\mathbf{A} \subseteq \mathbf{V}$  in a Gibbs compatible ECP represented by a pair of CDAGs  $\mathcal{G}_1(\mathbf{V}_1 \cup \mathbf{W})$  and  $\mathcal{G}(\mathbf{V}_{+1}, \mathbf{V}_1 \cup \mathbf{W})$ . The microstate counterfactual for  $\mathbf{Y} = \mathbf{V} \setminus \mathbf{A}$  at time  $T$  is identified as

$$p(\mathbf{Y}_T(\mathbf{a}) = \mathbf{y} \mid \mathbf{W}) = \sum_{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_{T-1}} p(\mathbf{y}_1 \mid \text{pa}_{\mathcal{G}_1}(\mathbf{y}_1)) \cdot \left( \prod_{t=1}^{T-1} p(\mathbf{y}_{t+1} \mid \text{pa}_{\mathcal{G}_{+1}}(\mathbf{y}_{t+1})) \right) \Big|_{A=a}$$

from the observed data law  $p(\mathbf{V})$  on macrostates and the starting distribution  $p(\mathbf{V}_1 \mid \text{pa}_{\mathcal{G}_1}(\mathbf{V}_1))$ , which is assumed to be known.

Identification of macrostate counterfactuals in a PCM is given by the following theorem:

*Theorem 7.* Given a PCM associated with DAG  $\mathcal{G}(\mathbb{V})$  and CDAGs  $\{\mathcal{G}_{+1}^k(\mathbf{V}_{+1}^k, \mathbf{V}_1^k \cup \mathbf{W}^k), \mathcal{G}_1^k(\mathbf{V}_1^k, \mathbf{W}^k) : k = 1, \dots, K\}$  corresponding to Gibbs compatible ECPs, fix a treatment set  $\mathbf{A} \subseteq \mathbf{V}$  and let  $\mathbf{Y} = \mathbf{V} \setminus \mathbf{A}$ . Then the macrostate counterfactual is identified from the observed data law on macrostates  $p(\mathbf{V})$  as

$$p(\mathbf{Y}(\mathbf{a}) = \mathbf{y}) = \prod_{k=1}^K p(\mathbf{Y}^k = \mathbf{y}^k \mid \mathbf{W}^k \setminus \mathbf{A} = \mathbf{y}^{\mathbf{W}^k \setminus \mathbf{A}}, \mathbf{a}^{\mathbf{A} \cap \mathbf{V}^k}, \mathbf{a}^{\mathbf{W}^k \cap \mathbf{A}})$$

The macrostate distribution of a PCM consisting of Gibbs compatible ECPs will obey the Markov properties of an LWF chain graph since they share equilibrium-generating Gibbs dynamics (see Section 2.7), and identification of macrostate counterfactuals associated with such a PCM will be identical to the chain graph g-formula (2.10).

Microstate counterfactuals in PCMs are identified as stated below:

*Theorem 8.* Given a PCM associated with DAG  $\mathcal{G}(\mathbb{V})$  and  $\{\mathcal{G}_{+1}^k(\mathbf{V}_{+1}^k, \mathbf{V}_1^k \cup \mathbf{W}^k), \mathcal{G}_1^k(\mathbf{V}_1^k, \mathbf{W}^k) : k = 1, \dots, K\}$ , fix a treatment set  $\mathbf{A} \subseteq \mathbf{V}$  and let  $\mathbf{Y} \subseteq \mathbf{V} \setminus \mathbf{A}$ . Given timepoints of interest  $T^k$  for each  $\mathbf{Y}^k$ , the microstate counterfactual  $\{Y_{T^k}^k(\mathbf{a}) : \mathbf{Y}^k \neq \emptyset\}$  is identified from the observed data distribution on macrostates  $p(\mathbf{V})$  and the starting microstate distributions  $p(\mathbf{V}_1^k \mid \text{pa}_{\mathcal{G}_1^k}(\mathbf{V}_1^k))$  as

$$\begin{aligned} p(\mathbf{C}_{\mathbf{Y}}(\mathbf{a}), \{\mathbf{Y}_{T^k}^k(\mathbf{a}) : \mathbf{Y}^k \neq \emptyset\}) &\equiv \prod_{\mathbf{V}^k \in \mathbf{C}} p(\mathbf{V}^k \mid \text{pa}_{\mathcal{G}}(\mathbf{V}^k)) \\ &\times \prod_{\mathbf{V}^k : \mathbf{Y}^k \neq \emptyset} \left[ \sum_{\mathbf{y}_1^k, \mathbf{y}_2^k, \dots, \mathbf{y}_{T^k-1}^k} p(\mathbf{y}_1^k \mid \text{pa}_{\mathcal{G}_1^k}(\mathbf{y}_1^k)) \cdot \left( \prod_{t=1}^{T^k-1} p(\mathbf{y}_{t+1}^k \mid \text{pa}_{\mathcal{G}_{+1}^k}(\mathbf{y}_{t+1}^k)) \right) \right] \mid \bar{A} = \bar{a} \end{aligned}$$

For examples illustrating these theorems, please refer to [Sani\* et al., 2022].

In PCMs that are not obtained from Gibbs compatible ECPs, identification of counterfactual distributions is challenging and we leave treatment of other types of equilibrium-generating dynamics for future work.

## 5.5 Inference

We now discuss estimation of identified parameters in PCMs. But first, we provide some preliminaries on statistical estimators. Readers who are acquainted with influence-function based estimators may skip the background in Section 5.5.1 and go straight to novel results in Section 5.5.2.

### 5.5.1 Statistical Estimators

Properties typically desired from statistical estimators are consistency, asymptotical normality and  $\sqrt{n}$ -rates of estimation, since these give us guarantees on the correctness of our estimator as well as quantification of uncertainty of our estimate [Bickel and Doksum, 2015].

If a correct parametric likelihood can be assumed, the plug-in principle [Bickel and Doksum, 2015] can be used to derive estimators. Conditional on the assumptions being true, plug-in estimators are  $\sqrt{n}$ -consistent and asymptotically normal. For example, the average causal effect (ACE),  $\mathbb{E}[Y(a) - Y(a')]$ , given outcome  $Y$ , treatment  $A$  and observed covariates  $C$ , is identified from the observed data distribution  $p(Y, A, C)$  under the assumption  $Y(a) \perp\!\!\!\perp A \mid C$  as:

$$ACE = \mathbb{E}[\mathbb{E}[Y \mid A = a, C]] - \mathbb{E}[\mathbb{E}[Y \mid A = a', C]]$$

If we specify a parametric form for  $\mathbb{E}[Y \mid A = a, C]$ , written as  $\mu(A, C; \eta)$ , a plug-in estimator can be obtained using the following:  $\frac{1}{n} \sum_{i=1}^n \mu(a, C_i; \hat{\eta}) - \mu(a', C_i; \hat{\eta})$ , where  $\hat{\eta}$  is the maximum likelihood estimate of  $\eta$ .

However, we may not be able to correctly specify the likelihood of the model in

practice. In such situations, we would like to make minimal assumptions on the model while still obtaining estimators that are  $\sqrt{n}$ -consistent and asymptotically normal. A class of *regular and asymptotically linear estimators* (RAL) [Tsiatis, 2006, Robins et al., 1994b] provide us with a principled approach to do so.

Given a model indexed by an infinite dimensional parameter  $\theta$ , a scalar parameter of interest  $\psi$  can be viewed as a function of  $\theta$ , i.e. as  $\psi(\theta)$ . An estimator  $\hat{\psi}_n$  of this parameter, based on  $n$  i.i.d. samples of data  $Z \sim p(Z; \theta)$ , is a RAL estimator if there exists a function  $U_\psi(Z)$ , with zero mean and finite variance such that

$$\sqrt{n} \times (\hat{\psi}_n - \psi) = \frac{1}{\sqrt{n}} \times \sum_{i=1}^n U_\psi(Z_i) + o_p(1)$$

Here  $o_p(1)$  is a term that converges to zero in probability as  $n$  goes to infinity, and  $U_\psi(Z)$  is called the *influence function* of the estimator [Tsiatis, 2006].

Influence functions provide estimators for  $\psi$  by defining estimating equations of the form  $\mathbb{P}_n[U_\psi(Z)] = 0$ . The resulting estimator obtained by solving the estimating equations will be *consistent and asymptotically normal* (CAN) with an asymptotical variance equal to the variance of the influence function<sup>4</sup>, i.e.

$$\sqrt{n}(\hat{\psi}_n - \psi) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \text{var}(U_\psi(Z)))$$

To obtain the influence function for a parameter  $\psi(\theta)$ , for a model indexed by infinite dimensional parameter  $\theta$ , we define the *score vector* for an observation  $Z \sim p(Z; \theta)$ , as  $S_\theta(Z; \theta_0) = \left. \frac{\partial \log p(Z; \theta)}{\partial \theta} \right|_{\theta=\theta_0}$ . The influence function for  $\hat{\psi}_n$ , the estimator of parameter  $\psi(\theta)$ , can be obtained using the integral equation:

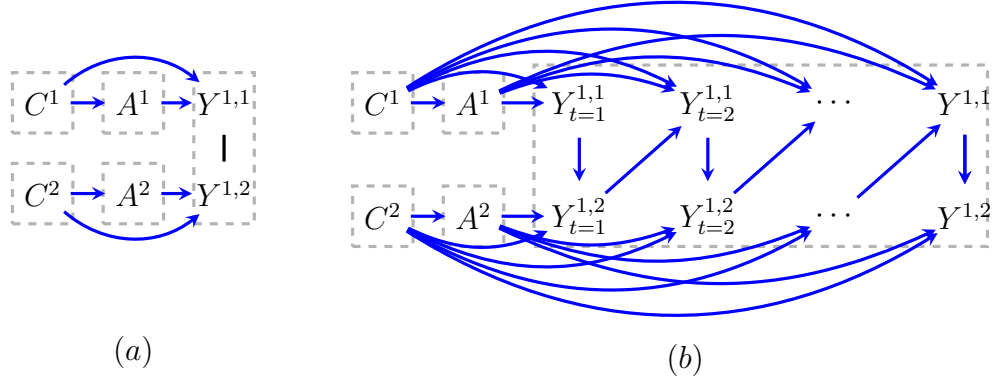
$$\left. \frac{\partial \psi(\theta)}{\partial \theta} \right|_{\theta=\theta_0} = \mathbb{E}[U_\psi(Z) S_\theta(Z; \theta_0)]$$

Estimators based on influence functions often also have robustness properties, i.e. certain nuisance<sup>5</sup> functions can be misspecified while still obtaining a consistent

---

<sup>4</sup>We use  $\xrightarrow{\mathcal{D}}$  to denote convergence in distribution.

<sup>5</sup>A nuisance parameter or function is an entity whose value we are not directly interested in, but whose value determines or modifies the quantities we are interested in.



**Figure 5.3.** The model that was used to generate data in the simulation study. (a) The PCM (b) The expanded equilibrium process in the ECP involving  $Y^{1,1}$  and  $Y^{1,2}$

estimate for the target parameter. Additionally, influence functions allow for slower than  $\sqrt{n}$  convergence rates for nuisance functions, allowing for the use of highly flexible machine learning models to fit nuisance functions [Bickel et al., 1993, Chernozhukov et al., 2018] while still obtaining desirable statistical guarantees.

## 5.5.2 Estimators for Specific Targets

We derive influence function based estimators for two microstate counterfactuals in Fig. 5.3, namely  $\beta \equiv \mathbb{E}[Y_{t=1}^{1,2}(a^1, a^2)]$  and  $\psi \equiv \mathbb{E}[Y_{t=2}^{1,1}(a^1, a^2)]$ . In this example, we use  $C$ ,  $A$  and  $Y$  in place of  $V$ : nodes  $Y$  denote the outcome,  $A$ , the treatment and  $C$ , the observed covariates. While we only provide influence functions (and associated robustness properties) for two microstate counterfactuals here, influence functions can be derived for any arbitrary microstate counterfactual that is identified, and a recursive formulation is discussed in [Sani\* et al., 2022].

### 5.5.2.1 Target $\beta \equiv \mathbb{E}[Y_{t=1}^{1,2}(a^1, a^2)]$

Target  $\beta \equiv \mathbb{E}[Y_{t=1}^{1,2}(a^1, a^2)]$  is identified from the observed data distribution as  $\sum_{Y^{1,1}, Y^{1,2}, C^1, C^2} \mathbb{E}[Y^{1,2} \mid Y^{1,1}, a^1, a^2, C^1, C^2] p(Y_{t=1}^{1,1} \mid a^1, a^2, C^1, C^2) p(C^1, C^2)$ . We give the influence function and robustness properties of the estimator in Theorems 9 and 10 respectively. Proofs are deferred to Appendix IV.

*Theorem 9.* The influence function for  $\beta$  is given as

$$\begin{aligned} U_\beta(Z) = & \frac{I(A^1 = a^1)I(A^2 = a^2)}{p(Y^{1,1}, A^1, A^2 \mid C^1, C^2)} p(Y_{t=1}^{1,1} \mid A^1, A^2, C^1, C^2) \times \\ & \{Y^{1,2} - \mathbb{E}[Y^{1,2} \mid Y^{1,1}, A^1, A^2, C^1, C^2]\} \\ & + \sum_{Y^{1,1}} \mathbb{E}[Y^{1,2} \mid Y^{1,1}, a^1, a^2, C^1, C^2] p(Y_{t=1}^{1,1} \mid a^1, a^2, C^1, C^2) - \beta \end{aligned}$$

*Theorem 10.* The estimator obtained by solving the influence function for  $\beta$  is doubly robust as long as one of  $p(Y^{1,1}, A^1, A^2 \mid C^1, C^2)$  or  $\mathbb{E}[Y^{1,2} \mid Y^{1,1}, a^1, a^2, C^1, C^2]$  is specified correctly.

### 5.5.2.2 Target $\psi \equiv \mathbb{E}[Y_{t=2}^{1,1}(a^1, a^2)]$

We present the influence function for  $\psi$ , and its robustness properties, but these are slightly different than those exhibited by  $U(\beta)$ . We use the short-hand  $\pi \equiv p(Y_{t=1}^{1,1} \mid A^1, A^2, C^1, C^2)$  and  $\phi_{a^1, a^2}(Y^{1,1} \mid Y^{1,2}) \equiv \mathbb{E}[Y^{1,1} \mid Y^{1,2}, A^1, A^2, C^1, C^2]$ .

Target  $\psi$  is identified from the observed data distribution as  $\sum_{Y^{1,1}, Y^{1,2}, C^1, C^2} \mathbb{E}[Y^{1,1} \mid Y^{1,2}, a^1, a^2, C^1, C^2] p(Y^{1,2} \mid Y^{1,1}, a^1, a^2, C^1, C^2) \pi p(C^1, C^2)$  and its corresponding influence function is given below.

*Theorem 11.* The influence function for  $\psi$  is given as

$$\begin{aligned} U(\psi) = & \left\{ \sum_{Y^{1,1}} \frac{I(A^1 = a^1)I(A^2 = a^2)\pi p(Y^{1,2} \mid Y^{1,1}, A^1, A^2, C^1, C^2)}{p(Y^{1,2}, A^1, A^2 \mid C^1, C^2)} \right\} \\ & \times \left\{ Y^{1,1} - \mathbb{E}[Y^{1,1} \mid Y^{1,2}, A^1, A^2, C^1, C^2] \right\} \\ & + \left\{ \frac{I(A^1 = a^1)I(A^2 = a^2)\pi}{p(Y^{1,1}, A^1, A^2 \mid C^1, C^2)} \right\} \\ & \times \left\{ \phi_{a^1, a^2}(Y^{1,1} \mid Y^{1,2}) - \mathbb{E}[\phi_{a^1, a^2}(Y^{1,1} \mid Y^{1,2}) \mid Y^{1,1}, A^1, A^2, C^1, C^2] \right\} \\ & + \sum_{Y^{1,1}, Y^{1,2}} \mathbb{E}[Y^{1,1} \mid Y^{1,2}, a^1, a^2, C^1, C^2] p(Y^{1,2} \mid Y^{1,1}, a^1, a^2, C^1, C^2) \pi - \psi \end{aligned}$$

Since the model has more nuisance parameters, it will exhibit different robustness properties.

*Theorem 12.* The influence function based estimator for  $\psi$  above exhibits 2 out of 4 robustness, where as long as the following pairs of models are specified correctly, we obtain a consistent and unbiased estimator:

$$\begin{aligned} & \left( \mathbb{E}[Y^{1,1} \mid Y^{1,2}, a^1, a^2, C^1, C^2], \quad p(Y^{1,2} \mid Y^{1,1}, a^1, a^2, C^1, C^2) \right) \\ & \left( p(Y^{1,2}, A^1, A^2 \mid C^1, C^2), \quad p(Y^{1,2} \mid Y^{1,1}, A^1, A^2, C^1, C^2) \right) \\ & \left( p(Y^{1,1}, A^1, A^2 \mid C^1, C^2), \quad \mathbb{E}[Y^{1,1} \mid Y^{1,2}, A^1, A^2, C^1, C^2] \right) \end{aligned}$$

All proofs and derivations are in Appendix IV.

## 5.6 Experiments

Next, we demonstrate the utility of our model through simulations (Section 5.6.1) and a data application (Section 5.6.2) involving cellular reprogramming.

### 5.6.1 Simulation Study

Data is simulated according to the PCM in Fig. 5.3(a), an extension of the simplified model of cell reprogramming in Fig. 5.2, but which allows for incremental complexity in the form of covariates  $C^1, C^2$ , that can be thought of as macro-environmental factors. Treatment variables  $A^1$  and  $A^2$  correspond to transcription factors; these variables are assumed to be binary for simplicity. As before,  $Y^{1,1}$  and  $Y^{1,2}$  correspond to the macrostate of an infinite ECP, and pertain to neural and pluripotent identity, and are also binary. Additionally, we assume that the ECP is Gibbs compatible [Lauritzen and Richardson, 2002], giving rise to the undirected edge  $Y^{1,1} - Y^{1,2}$  as interpreted in [Lauritzen and Richardson, 2002]. While the experiment measures macrostate variables  $Y^{1,1}$  and  $Y^{1,2}$ , we are interested in how microstate variables  $Y_{t=1}^{1,1}, Y_{t=1}^{1,2}, Y_{t=2}^{1,1} \dots$  respond to interventions on  $A^1$  and  $A^2$  on average, i.e., counterfactuals  $\mathbb{E}[Y_{t=1}^{1,2}(a^1, a^2)]$  and  $E[Y_{t=2}^{1,2}(a^1, a^2)]$ . These quantities inform us “what the expected pluripotent or neural identity would have been at the first and second timepoint, respectively (as measured

Target	Analytical	Plug-in	IF-based ( $q_{0.05}, q_{0.95}$ )
$\mathbb{E}[Y_{t=1}^{1,2}(a^1, a^2)]$	0.619	0.618	0.624 (0.617, 0.630)
$\mathbb{E}[Y_{t=2}^{1,1}(a^1, a^2)]$	0.370	0.371	0.372 (0.365, 0.378)

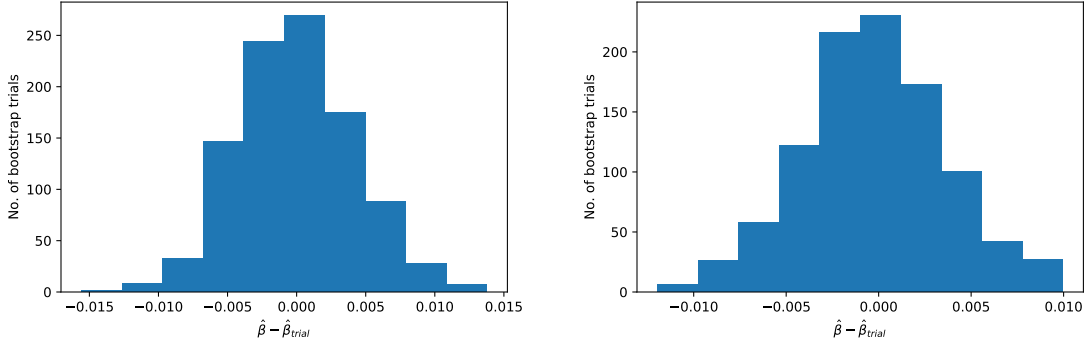
**Table 5-I.** Bootstrapped estimates of targets for a chosen intervention  $a^1 = 0, a^2 = 0$  in a scale appropriate for the experiment), had the transcription factors been set to  $a^1, a^2$ ?. We assume that the distribution  $\pi(Y_{t=1}^{1,1}) = p(Y_{t=1}^{1,1} | A^1, A^2, C^1, C^2)$  is known in order to identify these targets.

We use the following distributions to generate data:  $P(C^1), P(C^2), P(A^1|C^1), P(A^2|C^2)$  are all Bernoulli with the parameter chosen from  $\mathcal{U}[0, 1]$ ; the joint distribution  $P(Y^{1,1}, Y^{1,2}|A^1, A^2, C^1, C^2)$  is also set as a table of Bernoulli probabilities for all levels of  $Y^{1,1}, Y^{1,2}, A^1, A^2, C^1$  and  $C^2$ . Gibbs factors  $P(Y^{1,1} = 1|A^1, A^2, C^1, C^2, Y^{1,2})$  and  $P(Y^2 = 1|A^1, A^2, C^1, C^2, Y^1)$  can be determined uniquely from the joint distributions and starting distribution for  $\pi(Y_{t=1}^{1,1})$  is also set to Bernoulli with a known parameter. Gibbs sampling was done over 5000 initiations, for 2000 iterations. We use the running mean, a popular convergence diagnostic metric, to determine that equilibrium is always attained within the first 2000 iterations.

We estimate the targets  $\mathbb{E}[Y_{t=1}^{1,2}(a^1, a^2)]$  and  $E[Y_{t=2}^{1,1}(a^1, a^2)]$  using two methods: (1) plug-in using MLE estimates of the identification formulae in Theorem 8 and (2) influence-function based estimator of the targets as given in Theorems 9 and 11, over a bootstrap of 1000 samples. And each of these targets is compared to the analytical estimate obtained from the underlying generative model. Table 5-I records the results for a particular intervention  $a^1 = 0, a^2 = 0$ . Variation across bootstrap trials is documented in Fig. 5.4.

### 5.6.2 Cellular Reprogramming

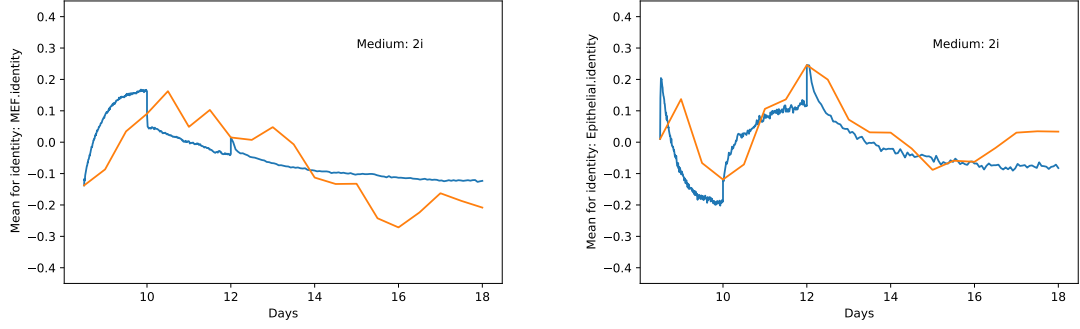
We are interested in estimating the (counterfactual) trajectories of cells in reprogramming, modeled using a PCM, similar to the one in Fig. 5.2. Our data, central to the



**Figure 5.4.** Histograms showing how  $\hat{\beta} - \hat{\beta}_{trial}$  varies across bootstrap trials for targets  $\hat{\beta} = E[Y_{t=1}^{1,2}(a^1, a^2)]$  on the left and  $\mathbb{E}[Y_{t=2}^{1,1}(a^1, a^2)]$  on the right, for the intervention  $a^1 = a^2 = 0$ .

work in [Schiebinger et al., 2019], is obtained from the NCBI Gene Expression Omnibus (ID: GSE122662). It consists of single-cell RNA-sequencing data obtained from 251203 cells over 18 days. As part of our analysis, we assume that data is collected only when the transcription factors are added and when the previously cited milestones, namely (1) endogenous Oct4 expression marked by emergence of Oct4-EGFP+ cells, (2) initial signs of pluripotency via expression of marker genes such as Nanog, Zfp42, Dppa24 and Esrrb, and (3) conversion to final iPSC fate, occur. These happen on Day 10, Day 12 and Day 18 of the culture, respectively. The reprogramming process initiates on the day marked 8.5, when the transcription factors are added, which we will consider as the starting point of the trajectory for our purposes. In the experiment, there are two arms, one consisting of cells in medium 2i, and the other in serum; for details on these media, please refer to [Schiebinger et al., 2019]. Investigators in [Schiebinger et al., 2019] were interested in tracing the trajectory of these cell cultures over time; see [Schiebinger et al., 2019] for more details about the data and experiments. In our analysis, we will restrict our attention to mapping out the trajectory of the microstate variables over time in either medium, as we describe below.

We assume that the dynamics of the cellular reprogramming process is captured well by a PCM. In particular, just as described in Section 5.3 and Section 5.4, there are three



**Figure 5.5.** Sample trajectories corresponding to 2 dimensions, namely MEF and epithelial identities. The blue trajectory is derived the Gibbs compatible PCM, and the orange one represents the trajectory from the original dataset, coarsely sampled every 12 hours.

ECPs that make up the reprogramming trajectory, each of them Gibbs compatible, and there is a trivial block corresponding to the treatment that precedes them (see Fig. 5.2). One notable difference between the figure and the data is that real data from our experiment is high-dimensional and consists of 32 dimensions of cell-identity, in contrast to the two dimensions shown in our simplified illustration. These 32 dimensions are obtained from gene-expression data using the dimensionality-reduction method explained in [Schiebinger et al., 2019], an approach standard in computational biology. Intuitively, each of these dimensions corresponds to a normalized correlation obtained between the cell’s expression and a known gene-expression for a specific kind of cell, like a neural cell, for example. Each of the 32 dimensions and their interpretations are discussed in [Sani\* et al., 2022].

Since the three ECPs are Gibbs compatible, we can use the starting distribution (which we have from data) as well as the equilibrium distribution at each of the milestone points to deduce the microstate trajectory. For each of the ECPs, we assume that the Gibbs factors  $P(Y^i|A, \mathbf{Y}^{-i})^6$  are given by kernel ridge regressions. The goodness of fit of these regressions is shown in the supplementary material of [Sani\* et al., 2022]. The Gibbs trajectories are generated using these regressions and the starting distribution,

<sup>6</sup> $\mathbf{Y}^{-i} \equiv \mathbf{Y} \setminus Y^i$  where  $\mathbf{Y}$  is the 32-dimensional cell-identity vector, and  $A$  is binary, corresponding to the medium, serum or 2i.

for each of the cells originally recorded in the data. The equilibrium was considered to be attained when the difference in the running mean at a sample distance of 10 was under the threshold 0.005. The mean pre-equilibrium trajectory for the three ECPs, across each of the dimensions was appended together, maintaining relative temporal widths, for each medium. See Fig. 5.5 for some examples. For the rest of the plots, please refer to Appendix IV. We infer from these plots that Gibbs compatible PCMs capture the overall trend observed in the cellular reprogramming process, along each of the cell identity dimensions. Some of the regressions and trajectories deviate from the ground truth, which suggests that there might be room to explore other conditional distributions and equilibrium dynamics.

Unlike in the example with synthetic data, we do not attempt to recover the microstate counterfactuals at each timepoint since the data is extremely high-dimensional. The identifying functional is algebraically intensive to derive, but is technically identified. We reserve extensions to higher dimensions for future work.

## 5.7 Conclusions

We developed a framework to reason about interventions performed on stochastic processes reaching equilibrium. We define microstate and macrostate counterfactuals, and provide sound and complete identification algorithms for these. We also derived flexible semiparametric estimators for the identified counterfactuals. Finally, we demonstrate the utility of our model through experiments on synthetic data and a data application involving cellular reprogramming. Exploring different types of equilibrium generating dynamics, as well as extensions for high-dimensional data are promising avenues for future work.

# Chapter 6

## Conclusions and Future Work

In this dissertation, we introduced graphical causal models for settings with various complications in data, including path dependence and entangled missingness. We also extended causal modeling approaches to systems where observed data is not merely a coarsened version of the full data distribution as in conventional causal inference settings; instead the observed data is obtained as a consequence of equilibrium generating dynamics and they occur at different timescales.

Within each of these studies, we motivated the need for these models with various clinical and healthcare applications and explored various parts of the causal inference pipeline, with a special emphasis on identification and in some cases, estimation. We also pointed out that, while we have uncovered some new ideas and established novel results, many unanswered theoretical questions remain in each of these topics. For example, in hidden variable path dependent systems, inference strategies remain limited to relatively simple model assumptions; efficient Monte Carlo sampling procedures, based on the nested Markov model would be an interesting avenue to pursue. In settings of entangled missingness, there is plenty to be explored, like valid targets of interest that do not assume full observability, associated identification theories and estimation procedures.

The author of this dissertation has worked on a few other topics (not in this document) during their graduate study, including applications of causal inference to electronic health record (EHR) data, and recognizes that bridging the gap between theory and practice in causal inference is what they would be most excited to pursue in the future. The models developed here, while they have been used in specific datasets like a surgery or cellular reprogramming, directly apply to various settings in the clinic. For example, entangled missingness is ubiquitous in hospital settings, for e.g. in allocational matters of hospital resources. Path dependence is crucial to how a patient moves (and their condition improves or deteriorates) within the hospital system. However, while causal inference has provided us with rich tools to model and investigate EHR data, application of these methods to high-dimensional data, remains tremendously challenging, due to several issues not limited to dataset sample sizes, mixed data-types, coding errors, existing estimation strategies among others. There appears to be growing recognition within clinical medicine that observational causal inference tools can be valuable when RCTs cannot be conducted [Hoffman et al., 2022, Lee et al., 2022] and the author would like to make the best of this opportunity to improve patient care using causal inference methods in the years to come.

# Bibliography

- Zack W Almquist and Carter T Butts. Dynamic network analysis with missing data: theory and methods. *Statistica Sinica*, 28(3):1245–1264, 2018.
- Alejandro Sánchez Alvarado and Shinya Yamanaka. Rethinking differentiation: stem cells, regeneration, and plasticity. *Cell*, 157(1):110–119, 2014.
- P. M. Aronow and C. Samii. Estimating average causal effects under general interference. *Technical Report*, 2013.
- S. Athey, D. Eckles, and G.W. Imbens. Exact p-values for network interference. *Journal of the American Statistical Association*, 113 (521):230–240, 2018.
- Alexander Balke and Judea Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176, 1997.
- G. Basse, A. Feller, and P. Toulis. Randomization tests of causal effects under interference. *Biometrika*, 106 (2):487–494, 2019.
- G. W. Basse and E. M. Airolidi. Model-assisted design of experiments in the presence of network-correlated outcomes. *Biometrika*, 105 (4):849–858, 2018.
- Rohit Bhattacharya, Jaron J. R. Lee, Razieh Nabi, and Ilya Shpitser. *Ananke*: A python package for causal inference with graphical models. URL <https://ananke.readthedocs.io/en/latest/index.html>.
- Rohit Bhattacharya, Daniel Malinsky, and Ilya Shpitser. Causal inference under interference and network uncertainty. In *Proceedings of the Thirty Fifth Conference on Uncertainty in Artificial Intelligence (UAI-35th)*. AUAI Press, 2019a.
- Rohit Bhattacharya, Razieh Nabi, Ilya Shpitser, and James Robins. Identification in missing data models represented by directed acyclic graphs. In *Proceedings of the Thirty Fifth Conference on Uncertainty in Artificial Intelligence (UAI-35th)*. AUAI Press, 2019b.
- Rohit Bhattacharya, Razieh Nabi, and Ilya Shpitser. Semiparametric inference for causal effects in graphical models with hidden variables. *arXiv preprint arXiv:2003.12659*, 2020.
- Peter J Bickel and Kjell A Doksum. *Mathematical statistics: basic ideas and selected topics, volumes I-II package*. Chapman and Hall/CRC, 2015.
- Peter J Bickel, Chris AJ Klaassen, Peter J Bickel, Ya’acov Ritov, J Klaassen, Jon A Wellner, and YA’Acov Ritov. *Efficient and adaptive estimation for semiparametric models*, volume 4. Springer, 1993.

- Gilles Blondel, Marta Arias, and Ricard Gavaldà. Identifiability and transportability in dynamic causal networks. *International Journal of Data Science and Analytics*, 3(2): 131–147, 2017.
- J. Bowers, F. M. M., and P. C. Reasoning about interference between units: A general framework. *Political Analysis*, 21:97–124, 2013.
- C Hendricks Brown. Protecting against nonrandomly missing data in longitudinal studies. *Biometrics*, pages 143–155, 1990.
- X. Cai, W. W. Loh, and F. W. Crawford. Identification of causal intervention effects under contagion. *arXiv preprint arXiv:1912.04151*, 2019.
- Changge Chang, Yi Deng, Xiaoqian Jiang, and Qi Long. Multiple imputation for analysis of incomplete data in distributed health data networks. *Nature communications*, 11(1): 1–11, 2020.
- Hua Yun Chen. A semiparametric odds ratio model for measuring association. *biometrics*, 63:413–421, 2007.
- Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.
- Artur Cieřlar-Pobuda, Viktoria Knoflach, Mikael V Ringh, Joachim Stark, Wirginia Likus, Krzysztof Siemianowicz, Saeid Ghavami, Andrzej Hudecki, Jason L Green, and Marek J Łos. Transdifferentiation and reprogramming: overview of the processes, their similarities and differences. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1864(7): 1359–1369, 2017.
- Brian J Coburn, Bradley G Wagner, and Sally Blower. Modeling influenza epidemics and pandemics: insights into the future of swine flu (h1n1). *BMC medicine*, 7(1):1–8, 2009.
- Rhian M. Daniel, Michael G. Kenward, Simon N. Cousens, and Bianca L. De Stavola. Using causal diagrams to guide analysis in missing data problems. *Statistical Methods in Medical Research*, 21(3):243–256, 2012.
- A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.
- Mathias Drton. Likelihood ratio tests and singularities. *The Annals of Statistics*, 37(2): 979–1012, 2009.
- Mathias Drton, Michael Eichler, and Thomas S. Richardson. Computing maximum likelihood estimates in recursive linear models with correlated errors. *Journal of Machine Learning Research*, 10(10), 2009.
- Werner Ebeling and Rainer Feistel. *Physics of Self-organization and Evolution*. John Wiley & Sons, 2011.
- Daniel J Eck, Olga Morozova, and Forrest W Crawford. Randomization for the susceptibility effect of an infectious disease intervention. *Journal of Mathematical Biology*, 85(4):1–22, 2022.

- Dean Eckles, Brian Karrer, and Johan Ugander. Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference*, 5(1), 2017.
- Robin J. Evans and Thomas S. Richardson. Markovian acyclic directed mixed graphs for discrete data. *Annals of Statistics*, pages 1–30, 2014.
- Robin J. Evans and Thomas S. Richardson. Smooth, identifiable supermodels of discrete DAG models with latent variables. *Bernoulli*, 2018. (to appear).
- Laura Forastiere, Edoardo M Airoidi, and Fabrizia Mealli. Identification and estimation of treatment and interference effects in observational studies on networks. *Journal of the American Statistical Association*, 116(534):901–918, 2021.
- Mirko Francesconi, Bruno Di Stefano, Clara Berenguer, Luisa de Andrés-Aguayo, Marcos Plana-Carmona, Maria Mendez-Lago, Amy Guillaumet-Adkins, Gustavo Rodriguez-Esteban, Marta Gut, Ivo G Gut, et al. Single cell rna-seq identifies the origins of heterogeneity in efficient cell transdifferentiation and reprogramming. *Elife*, 8:e41627, 2019.
- Krista J Gile and Mark S Handcock. Analysis of networks with missing data with application to the national longitudinal study of adolescent health. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(3):501–519, 2017.
- M Maria Glymour. Using causal diagrams to understand common problems in social epidemiology. *Methods in social epidemiology*, pages 393–428, 2006.
- Stephen Jay Gould and Stephen Jay Gould. *Punctuated equilibrium*. Harvard University Press, 2009.
- Bryan S. Graham, Guido W. Imbens, and Geert Ridder. Measuring the effects of segregation in the presence of social spillovers: a nonparametric approach. Technical report, National Bureau of Economic Research, 2010.
- M Elizabeth Halloran and Michael G Hudgens. Causal inference for vaccine effects on infectiousness. *The international journal of biostatistics*, 8(2):1–40, 2012.
- M. Elizabeth Halloran and Claudio J. Struchiner. Causal inference in infectious diseases. *Epidemiology*, pages 142–151, 1995.
- Miguel A Hernán and James M Robins. Using big data to emulate a target trial when a randomized trial is not available. *American journal of epidemiology*, 183(8):758–764, 2016.
- Miguel A Hernan and James M Robins. *Causal Inference: What If*. CRC Press: Taylor and Francis Group, 2020.
- Miguel A. Hernán, Babette Brumback, and James M. Robins. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*, pages 561–570, 2000.
- Katherine L Hoffman, Edward J Schenck, Michael J Satlin, William Whalen, Di Pan, Nicholas Williams, and Iván Díaz. Comparison of a target trial emulation framework vs cox regression to estimate the association of corticosteroids with covid-19 mortality. *JAMA Network Open*, 5(10):e2234425–e2234425, 2022.

- Guanglei Hong and Stephen W. Raudenbush. Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association*, 101(475):901–910, 2006.
- Nicholas J Horton and Nan M Laird. Maximum likelihood analysis of generalized linear models with missing covariates. *Statistical Methods in Medical Research*, 8(1):37–50, 1999.
- Zhirui Hu, Minping Qian, and Michael Q Zhang. Novel markov model of induced pluripotency predicts gene expression changes in reprogramming. In *BMC systems biology*, volume 5, pages 1–12. BioMed Central, 2011.
- M.G. Hudgens and M.E. Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842, 2008a.
- Michael G Hudgens and M Elizabeth Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842, 2008b.
- Timothy D. Jackson, Jeffrey J. Wannares, Todd R. Lancaster, David W. Rattner, and Matthew M. Hutter. Does speed matter? The impact of operative time on outcome in laparoscopic surgery. *Surgical Endoscopy*, 25(7):2288–2295, 2011.
- Ravi Jagadeesan, Natesh S Pillai, and Alexander Volfovsky. Designs for estimating the treatment effect in networks with interference. *The Annals of Statistics*, 48(2):679–712, 2020.
- Steffan L. Lauritzen. *Graphical Models*. Oxford, U.K.: Clarendon, 1996.
- Steffen L. Lauritzen and Thomas S. Richardson. Chain graph models and their causal interpretations (with discussion). *Journal of the Royal Statistical Society: Series B*, 64: 321–361, 2002.
- Jaron JR Lee, Ranjani Srinivasan, Chin Siang Ong, Diane Alejo, Stefano Schena, Ilya Shpitser, Marc Sussman, Glenn JR Whitman, and Daniel Malinsky. Causal determinants of postoperative length of stay in cardiac surgery using causal graphical learning. *The Journal of Thoracic and Cardiovascular Surgery*, 2022.
- Michael P Leung. Treatment and spillover effects under network interference. *Review of Economics and Statistics*, 102(2):368–380, 2020.
- Lingling Li, Changyu Shen, Xiaochun Li, and James M Robins. On weighting approaches for missing data. *Statistical methods in medical research*, 22(1):14–30, 2013.
- Stan Liebowitz and Stephen Margolis. Path dependence. *Encyclopedia of Law and Economics*, 2002.
- Roderick J Little. Missing data assumptions. *Annual Review of Statistics and Its Application*, 8:89–107, 2021.
- Roderick J.A. Little and Donald B. Rubin. *Statistical analysis with missing data*. Wiley Series in Probability and Statistics. Wiley, 2002. ISBN 9780471183860. URL <https://books.google.ca/books?id=aYPwAAAAAAAJ>.
- Lan Liu and Michael G Hudgens. Large sample randomization inference of causal effects

- in the presence of interference. *Journal of the american statistical association*, 109(505): 288–301, 2014.
- Ben D MacArthur and Ihor R Lemischka. Statistical mechanics of pluripotency. *Cell*, 154(3):484–489, 2013.
- Daniel Malinsky and Peter Spirtes. Causal structure learning from multivariate time series in settings with unmeasured confounding. In *Proceedings of 2018 ACM SIGKDD Workshop on Causal Discovery*, pages 23–47, August 2018.
- Daniel Malinsky and Peter Spirtes. Learning the structure of a nonstationary vector autoregression. *Proceedings of Machine Learning Research*, 89:2986–2994, April 2019. ISSN 2640-3498.
- Daniel Malinsky, Ilya Shpitser, and Eric J Tchetgen Tchetgen. Semiparametric inference for nonmonotone missing-not-at-random data: the no self-censoring model. *Journal of the American Statistical Association*, pages 1–9, 2021.
- Charles F. Manski. Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323, 1990. ISSN 00028282. URL <http://www.jstor.org/stable/2006592>.
- Fernando Martel García. Definition and diagnosis of problematic attrition in randomized controlled experiments. *Available at SSRN 2302735*, 2013.
- Wang Miao and Eric J Tchetgen Tchetgen. On varieties of doubly robust estimators under missingness not at random with a shadow variable. *Biometrika*, 103(2):475–482, 2016.
- Wang Miao, Peng Ding, and Zhi Geng. Identifiability of normal and normal mixture models with nonignorable missing data. *Journal of the American Statistical Association*, 111(516): 1673–1683, 2016.
- Soren W. Mogensen, Daniel Malinsky, and Niels R. Hansen. Causal learning for partially observed stochastic dynamical systems. In *UAI*, pages 350–360, 2018.
- Karthika Mohan and Judea Pearl. Graphical models for processing missing data. *Journal of the American Statistical Association*, pages 1–16, 2021.
- Karthika Mohan, Judea Pearl, and Jin Tian. Graphical models for inference with missing data. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1277–1285. Curran Associates, Inc., 2013. URL <http://papers.nips.cc/paper/4899-graphical-models-for-inference-with-missing-data.pdf>.
- Karthika Mohan, Felix Thoemmes, and Judea Pearl. Estimation with incomplete data: The linear case. In *Proceedings of the International Joint Conferences on Artificial Intelligence Organization*, 2018.
- Kevin P. Murphy. *Machine learning: A probabilistic perspective*. MIT Press, September 2012. ISBN 978-0-262-30432-0.
- Razieh Nabi and Rohit Bhattacharya. On testability and goodness of fit tests in missing data models. *arXiv preprint arXiv:2203.00132*, 2022.

- Razieh Nabi, Rohit Bhattacharya, and Ilya Shpitser. Full law identification in graphical models of missing data: Completeness results. In *International Conference on Machine Learning*, pages 7153–7163. PMLR, 2020.
- Elizabeth L. Ogburn and Tyler J. VanderWeele. Causal diagrams for interference. *Statistical Science*, 29(4):559–578, 2014.
- Georgia Papadogeorgou, Fabrizia Mealli, and Corwin M Zigler. Causal inference with interfering units for cluster and population level treatment allocation programs. *Biometrics*, 75(3):778–787, 2019.
- Judea Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan and Kaufmann, San Mateo, 1988.
- Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–709, 1995. URL [citeseer.ist.psu.edu/55450.html](http://citeseer.ist.psu.edu/55450.html).
- Judea Pearl. *Causality: Models, reasoning, and inference*. Cambridge University Press, 2000. ISBN 0-521-77362-8.
- Judea Pearl. *Causality: models, reasoning, and inference*. Cambridge University Press, 2 edition, 2009. ISBN 978-0521895606.
- Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on time series using restricted structural equation models. *Advances in Neural Information Processing Systems 26*, pages 154–162, 2013.
- David Puelz, Guillaume Basse, Avi Feller, and Panos Toulis. A graph-theoretic approach to randomization tests of causal effects under general interference. *arXiv preprint arXiv:1910.10862*, 2019.
- Aksharananda Rambachan, Lauren M. Mioton, Sujata Saha, Neil Fine, and John Y. S. Kim. The impact of surgical duration on plastic surgery outcomes. *European Journal of Plastic Surgery*, 36(11):707–714, 2013.
- Joseph D Ramsey, Kun Zhang, Madelyn Glymour, Ruben Sanchez Romero, Biwei Huang, Imme Ebert-Uphoff, Savini Samarasinghe, Elizabeth A Barnes, and Clark Glymour. Tetrad—a toolbox for causal discovery. In *8th International Workshop on Climate Informatics*, 2018.
- Frederick Reif. *Fundamentals of Statistical and Thermal Physics*. McGraw-Hill, 1965.
- Amy Richardson, Michael G Hudgens, Peter B Gilbert, and Jason P Fine. Nonparametric bounds and sensitivity analysis of treatment effects. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 29(4):596, 2014.
- Thomas S. Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30(1):145–157, 2003.
- Thomas S. Richardson, James M. Robins, and Ilya Shpitser. Nested Markov properties for acyclic directed mixed graphs. In *Twenty Eighth Conference on Uncertainty in Artificial Intelligence (UAI-12)*, 2012.

- Thomas S. Richardson, Robin J. Evans, James M. Robins, and Ilya Shpitser. Nested Markov properties for acyclic directed mixed graphs, 2017. Working paper.
- Benjamin Ridenhour, Jessica M Kowalik, and David K Shay. Unraveling r 0: Considerations for public health applications. *American journal of public health*, 104(2):e32–e41, 2014.
- James M. Robins. A new approach to causal inference in mortality studies with sustained exposure periods – application to control of the healthy worker survivor effect. *Mathematical Modeling*, 7:1393–1512, 1986.
- James M. Robins, Andrea Rotnitzky, and Lue P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89:846–866, 1994a.
- James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994b.
- Paul R. Rosenbaum. Interference between units in randomized experiments. *Journal of the American Statistical Association*, 102(477):191–200, 2007.
- Donald B. Rubin. Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.
- Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. ISSN 00063444. URL <http://www.jstor.org/stable/2335739>.
- Donald B Rubin. An overview of multiple imputation. In *Proceedings of the survey research methods section of the American statistical association*, pages 79–84. Citeseer Princeton, NJ, USA, 1988.
- Donald B Rubin. Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5(4):472–480, 1990.
- Numair Sani\*, Ranjani Srinivasan\*, Prashanthi Ravichandran, and Ilya Shpitser. Causal inference across timescales. Draft in preparation, 2022.
- Fredrik Sävje. Causal inference with misspecified exposure mappings. *arXiv preprint arXiv:2103.06471*, 2021.
- Fredrik Sävje, Peter Aronow, and Michael Hudgens. Average treatment effects in the presence of unknown interference. *Annals of statistics*, 49(2):673, 2021.
- Joseph L Schafer. Multiple imputation: a primer. *Statistical methods in medical research*, 8(1):3–15, 1999.
- Daniel O Scharfstein, Jaron JR Lee, Aidan McDermott, Aimee Campbell, Edward Nunes, Abigail G Matthews, and Ilya Shpitser. Markov-restricted analysis of randomized trials with non-monotone missing binary outcomes: Sensitivity analysis and identification results. *arXiv preprint arXiv:2105.08868*, 2021.
- Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, et al. Optimal-transport anal-

- ysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.
- Eli Sherman and Ilya Shpitser. Identification and estimation of causal effects from dependent data. In *Advances in neural information processing systems*, pages 9424–9435, 2018.
- Ilya Shpitser. Consistent estimation of functions of data missing non-monotonically and not at random. *Advances in Neural Information Processing Systems*, 29:3144–3152, 2016.
- Ilya Shpitser. Identification in graphical causal models. In *Handbook of Graphical Models*, 2017.
- Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06)*. AAAI Press, Palo Alto, 2006.
- Ilya Shpitser and Eli Sherman. Identification of personalized effects associated with causal pathways. In *Proceedings of the 34th Annual Conference on Uncertainty in Artificial Intelligence (UAI-18)*, 2018.
- Ilya Shpitser, Thomas S. Richardson, and James M. Robins. An efficient algorithm for computing interventional distributions in latent variable causal models. In *Uncertainty in Artificial Intelligence*, volume 27. AUAI Press, 2011.
- Jeffrey A Smith, James Moody, and Jonathan H Morgan. Network sampling coverage ii: The effect of non-random missing data on network measurement. *Social networks*, 48: 78–99, 2017.
- Michael E Sobel. What do randomized studies of housing mobility demonstrate? causal inference in the face of interference. *Journal of the American Statistical Association*, 101 (476):1398–1407, 2006.
- Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. Springer Verlag, New York, 2 edition, 2001. ISBN 978-0262194402.
- Jerzy Splawa-Neyman, Dorota M Dabrowska, and TP Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pages 465–472, 1990.
- Ranjani Srinivasan, Jaron JR Lee, Rohit Bhattacharya, and Ilya Shpitser. Path dependent structural equation models. In *Uncertainty in Artificial Intelligence*, pages 161–171. PMLR, 2021.
- BaoLuo Sun, Lan Liu, Wang Miao, Kathleen Wirth, James Robins, and Eric J Tchetgen Tchetgen. Semiparametric estimation with data missing not at random using an instrumental variable. *Statistica Sinica*, 28(4):1965, 2018.
- Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, October 2018. ISBN 978-0-262-35270-3.
- Kazutoshi Takahashi and Shinya Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *cell*, 126(4):663–676, 2006.

- Eric J Tchetgen Tchetgen and Tyler J VanderWeele. On causal inference in the presence of interference. *Statistical methods in medical research*, 21(1):55–75, 2012.
- Eric J. Tchetgen Tchetgen and Tyler J. VanderWeele. On causal inference in the presence of interference. *Statistical Methods in Medical Research*, 21(1):55–75, 2012.
- Eric J. Tchetgen Tchetgen, Linbo Wang, and BaoLuo Sun. Discrete choice models for nonmonotone nonignorable missing data: Identification and inference. <https://arxiv.org/abs/1607.02631>, 2016. Working paper.
- Eric J. Tchetgen Tchetgen, Isabel Fulcher, and Ilya Shpitser. Auto-g-computation of causal effects on a network. <https://arxiv.org/abs/1709.01577>, 2017. Working paper.
- Felix Thoemmes and Norman Rose. A cautious note on auxiliary variables that can increase bias in missing data problems. *Multivariate Behavioral Research*, 49(5):443–459, 2014.
- Jin Tian. Missing at random in graphical models. In *Artificial Intelligence and Statistics*, pages 977–985. PMLR, 2015.
- Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Eighteenth National Conference on Artificial Intelligence*, pages 567–573, 2002. ISBN 0-262-51129-0.
- Hugo Touchette. Equivalence and nonequivalence of ensembles: thermodynamic, macrostate, and measure levels. *Journal of Statistical Physics*, 159(5):987–1016, 2015.
- Panos Toulis and Edward Kao. Estimation of causal peer influence effects. In *International conference on machine learning*, pages 1489–1497. PMLR, 2013.
- Panos Toulis, Alexander Volfovsky, and Edoardo M Airoidi. Propensity score methodology in the presence of network entanglement between treatments. *arXiv preprint arXiv:1801.07310*, 2018.
- Anastasios Tsiatis. *Semiparametric Theory and Missing Data*. Springer-Verlag New York, 1st edition edition, 2006.
- Tyler J VanderWeele. Direct and indirect effects for neighborhood-based clustered and longitudinal data. *Sociological methods & research*, 38(4):515–544, 2010.
- Thomas S. Verma and Judea Pearl. Equivalence and synthesis of causal models. Technical Report R-150, Department of Computer Science, University of California, Los Angeles, 1990.
- Sheng Wang, Jun Shao, and Jae Kwang Kim. An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica*, pages 1097–1116, 2014.
- Daniel Westreich, Stephen R. Cole, Jessica G. Young, Frank Palella, Phyllis C. Tien, Lawrence Kingsley, Stephen J. Gange, and Miguel A. Hernán. The parametric g-formula to estimate the effect of highly active antiretroviral therapy on incident aids or death. *Statistics in Medicine*, 31(18):2000–2009, 2012.
- Sewall Wright. The method of path coefficients. *The annals of mathematical statistics*, 5(3):161–215, 1934.

- Margaret C Wu and R. J. Carroll. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, 175-188, 1988.
- Christian A Yates, Matthew J Ford, and Richard L Mort. A multi-stage representation of cell proliferation as a markov process. *Bulletin of mathematical biology*, 79(12):2905–2928, 2017.
- Junzhe Zhang and Elias Bareinboim. Markov decision processes with unobserved confounders: A causal approach. Technical report, Technical Report R-23, Purdue AI Lab, 2016.

# Appendix I

## Additional Background on Causal Graphical Models

### A. Graph Preliminaries

Let capital letters  $X$  denote random variables, and let lower case letters  $x$  values of  $X$ . Sets of random variables are denoted  $\mathbf{V}$ , and sets of values  $\mathbf{v}$ . For a subset  $\mathbf{A} \subseteq \mathbf{V}$ ,  $\mathbf{v}_{\mathbf{A}}$  denotes the subset of values in  $\mathbf{v}$  of variables in  $\mathbf{A}$ . Domains of  $X$  and  $\mathbf{X}$  are denoted by  $\mathfrak{X}_X$  and  $\mathfrak{X}_{\mathbf{X}}$ , respectively.

Standard genealogic relations on graphs are as follows: parents, children, descendants, siblings and ancestors of  $X$  in a graph  $\mathcal{G}$  are denoted by  $\text{pa}_{\mathcal{G}}(X)$ ,  $\text{ch}_{\mathcal{G}}(X)$ ,  $\text{de}_{\mathcal{G}}(X)$ ,  $\text{si}_{\mathcal{G}}(X)$ ,  $\text{an}_{\mathcal{G}}(X)$ , respectively [Lauritzen, 1996]. These relations are defined disjunctively for sets, e.g.  $\text{pa}_{\mathcal{G}}(\mathbf{X}) \equiv \bigcup_{X \in \mathbf{X}} \text{pa}_{\mathcal{G}}(X)$ . By convention, for any  $X$ ,  $\text{an}_{\mathcal{G}}(X) \cap \text{de}_{\mathcal{G}}(X) \cap \text{dis}_{\mathcal{G}}(X) = \{X\}$ .

We will also define the set of *strict parents* as follows:  $\text{pa}_{\mathcal{G}}^s(\mathbf{X}) = \text{pa}_{\mathcal{G}}(\mathbf{X}) \setminus \mathbf{X}$ . Given any vertex  $V$  in an ADMG  $\mathcal{G}$ , define the *ordered Markov blanket* of  $V$  as  $\text{omb}_{\mathcal{G}}(V) \equiv (\text{dis}_{\mathcal{G}}(V) \cup \text{pa}_{\mathcal{G}}(\text{dis}_{\mathcal{G}}(V))) \setminus V$ . Given a graph  $\mathcal{G}$  with vertex set  $\mathbf{V}$ , and  $\mathbf{S} \subseteq \mathbf{V}$ , define the *induced subgraph*  $\mathcal{G}_{\mathbf{S}}$  to be a graph containing the vertex set  $\mathbf{S}$  and all edges in  $\mathcal{G}$  among elements in  $\mathbf{S}$ .

### B. The Nested Markov Factorization

The nested Markov factorization has been explained succinctly in the main text of this dissertation. Here, we provide further details and restate some of the results we have already discussed, in context of new background provided below.

The nested Markov factorization of  $p(\mathbf{V}|\mathbf{W})$  with respect to a CADMG  $\mathcal{G}(\mathbf{V}, \mathbf{W})$  links *kernels*, mappings derived from  $p(\mathbf{V}|\mathbf{W})$  and CADMGs derived from  $\mathcal{G}(\mathbf{V}, \mathbf{W})$  via a *fixing* operation.

**Kernel:** A kernel  $q_{\mathbf{V}}(\mathbf{V}|\mathbf{W})$  is a mapping from values in  $\mathbf{W}$  to normalized densities over  $\mathbf{V}$  [Lauritzen, 1996]. A conditional distribution is a familiar example of a kernel, in that  $\sum_{\mathbf{v} \in \mathbf{V}} q_{\mathbf{V}}(\mathbf{v}|\mathbf{w}) = 1$ . Conditioning and marginalization are defined in kernels in the usual way: For  $\mathbf{A} \subseteq \mathbf{V}$ ,  $q_{\mathbf{V}}(\mathbf{A}|\mathbf{W}) \equiv \sum_{\mathbf{V} \setminus \mathbf{A}} q_{\mathbf{V}}(\mathbf{V}|\mathbf{W})$  and  $q_{\mathbf{V}}(\mathbf{V} \setminus \mathbf{A}|\mathbf{A} \cup \mathbf{W}) \equiv \frac{q_{\mathbf{V}}(\mathbf{V}|\mathbf{W})}{q_{\mathbf{V}}(\mathbf{A}|\mathbf{W})}$ .

**Fixability and the fixing operator:** A variable  $V \in \mathbf{V}$  in a CADMG  $\mathcal{G}$  is fixable if  $\text{de}_{\mathcal{G}}(V) \cap \text{des}_{\mathcal{G}}(V) = \emptyset$ . In other words,  $V$  is fixable if paths  $V \leftrightarrow \dots \leftrightarrow B$  and  $V \rightarrow \dots \rightarrow B$  do not both exist

in  $\mathcal{G}$  for any  $B \in \mathbf{V} \setminus \{V\}$ .

We define a fixing operator  $\phi_V(\mathcal{G})$  for graphs, and a fixing operator  $\phi_V(q; \mathcal{G})$  for kernels. Given a CADMG  $\mathcal{G}(\mathbf{V}, \mathbf{W})$ , with a fixable  $V \in \mathbf{V}$ ,  $\phi_V(\mathcal{G}(\mathbf{V}, \mathbf{W}))$  yields a new CADMG  $\mathcal{G}(\mathbf{V} \setminus \{V\}, \mathbf{W} \cup \{V\})$  obtained from  $\mathcal{G}(\mathbf{V}, \mathbf{W})$  by moving  $V$  from  $\mathbf{V}$  to  $\mathbf{W}$ , and removing all edges with arrowheads into  $V$ . Given a kernel  $q_{\mathbf{V}}(\mathbf{V}|\mathbf{W})$ , and a CADMG  $\mathcal{G}(\mathbf{V}, \mathbf{W})$ , the operator  $\phi_V(q_{\mathbf{V}}(\mathbf{V}|\mathbf{W}), \mathcal{G}(\mathbf{V}, \mathbf{W}))$  yields a new kernel:

$$q_{\mathbf{V} \setminus \{V\}}(\mathbf{V} \setminus \{V\}|\mathbf{W} \cup \{V\}) \equiv \frac{q_{\mathbf{V}}(\mathbf{V}|\mathbf{W})}{q_{\mathbf{V}}(V|\text{omb}_{\mathcal{G}}(V))}$$

**Fixing sequences:** A sequence  $\langle V_1, \dots, V_k \rangle$  is said to be *valid* in  $\mathcal{G}(\mathbf{V}, \mathbf{W})$  if  $V_1$  fixable in  $\mathcal{G}(\mathbf{V}, \mathbf{W})$ ,  $V_2$  is fixable in  $\phi_{V_1}(\mathcal{G}(\mathbf{V}, \mathbf{W}))$ , and so on. If any two sequences  $\sigma_1, \sigma_2$  for the same set  $\mathbf{S} \subseteq \mathbf{V}$  are fixable in  $\mathcal{G}$ , they lead to the same CADMG. The graph fixing operator can be extended to a set  $\mathbf{S}$ :  $\phi_{\mathbf{S}}(\mathcal{G})$ . This operator is defined as applying the vertex fixing operation in any valid sequence  $\sigma$  for set  $\mathbf{S}$ .

Given a sequence  $\sigma_{\mathbf{S}}$ , define  $\eta(\sigma_{\mathbf{S}})$  to be the first element in  $\sigma_{\mathbf{S}}$ , and  $\tau(\sigma_{\mathbf{S}})$  to be the subsequence of  $\sigma_{\mathbf{S}}$  containing all elements but the first. Given a sequence  $\sigma_{\mathbf{S}}$  on elements in  $\mathbf{S}$  valid in  $\mathcal{G}(\mathbf{V}, \mathbf{W})$ , the kernel fixing operator  $\phi_{\sigma_{\mathbf{S}}}(q_{\mathbf{V}}(\mathbf{V}|\mathbf{W}), \mathcal{G}(\mathbf{V}, \mathbf{W}))$  is defined to be equal to  $q_{\mathbf{V}}(\mathbf{V}|\mathbf{W})$  if  $\sigma_{\mathbf{S}}$  is the empty sequence, and  $\phi_{\tau(\sigma_{\mathbf{S}})}(\phi_{\eta(\sigma_{\mathbf{S}})}(q_{\mathbf{V}}(\mathbf{V}|\mathbf{W}); \mathcal{G}(\mathbf{V}, \mathbf{W})), \phi_{\eta(\sigma_{\mathbf{S}})}(\mathcal{G}(\mathbf{V}, \mathbf{W})))$  otherwise.

**Reachability:** Given a CADMG  $\mathcal{G}(\mathbf{V}, \mathbf{W})$ , a set  $\mathbf{R} \subseteq \mathbf{V}$  is called *reachable* if there exists a sequence for  $\mathbf{V} \setminus \mathbf{R}$  valid in  $\mathcal{G}(\mathbf{V}, \mathbf{W})$ . In other words, if  $\mathbf{S}$  is fixable in  $\mathcal{G}$ ,  $\mathbf{V} \setminus \mathbf{S}$  is reachable.

**Intrinsic sets:** A set  $\mathbf{R}$  reachable in  $\mathcal{G}(\mathbf{V}, \mathbf{W})$  is *intrinsic* in  $\mathcal{G}(\mathbf{V}, \mathbf{W})$  if  $\phi_{\mathbf{V} \setminus \mathbf{R}}(\mathcal{G})$  contains a single district,  $\mathbf{R}$  itself. The set of intrinsic sets in a CADMG  $\mathcal{G}$  is denoted by  $\mathcal{I}(\mathcal{G})$ .

**Nested Markov factorization:** A distribution  $p(\mathbf{V}|\mathbf{W})$  is said to obey the *nested Markov factorization* with respect to the CADMG  $\mathcal{G}(\mathbf{V}, \mathbf{W})$  if there exists a set of kernels of the form  $\{q_{\mathbf{S}}(\mathbf{S}|\text{pa}_{\mathcal{G}}(\mathbf{S})) : \mathbf{S} \in \mathcal{I}(\mathcal{G})\}$  such that for every valid sequence  $\sigma_{\mathbf{R}}$  for a reachable set  $\mathbf{R}$  in  $\mathcal{G}$ , we have:

$$\phi_{\sigma_{\mathbf{R}}}(p(\mathbf{V}|\mathbf{W}); \mathcal{G}(\mathbf{V}, \mathbf{W})) = \prod_{\mathbf{D} \in \mathcal{D}(\phi_{\mathbf{R}}(\mathcal{G}(\mathbf{V}, \mathbf{W})))} q_{\mathbf{D}}(\mathbf{D}|\text{pa}_{\mathcal{G}}^{\mathbf{S}}(\mathbf{D}))$$

If a distribution obeys this factorization, then for any reachable  $\mathbf{R}$ , any two valid sequences on  $\mathbf{R}$  applied to  $p(\mathbf{V}|\mathbf{W})$  yield the same kernel  $q_{\mathbf{R}}(\mathbf{R}|\mathbf{V} \setminus \mathbf{R})$ . Hence, kernel fixing may be defined on sets, just as graph fixing. In this case, for every  $\mathbf{D} \in \mathcal{I}(\mathcal{G})$ ,  $q_{\mathbf{D}}(\mathbf{D}|\text{pa}_{\mathcal{G}}^{\mathbf{S}}(\mathbf{D})) \equiv \phi_{\mathbf{V} \setminus \mathbf{D}}(p(\mathbf{V}|\mathbf{W}); \mathcal{G}(\mathbf{V}, \mathbf{W}))$ .

The *district factorization* or *Tian factorization* of  $p(\mathbf{V}|\mathbf{W})$  results from the nested factorization:

$$\begin{aligned} p(\mathbf{V}|\mathbf{W}) &= \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}(\mathbf{V}, \mathbf{W}))} q_{\mathbf{D}}(\mathbf{D}|\text{pa}_{\mathcal{G}}^{\mathbf{S}}(\mathbf{D})) \\ &= \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}(\mathbf{V}, \mathbf{W}))} \left( \prod_{D \in \mathbf{D}} p(D|\text{pre}_{\prec}(D)) \right), \end{aligned}$$

where  $\text{pre}_{\prec}(D)$  is the set of predecessors of  $D$  according to a topological total ordering  $\prec$ . Each factor  $\prod_{D \in \mathbf{D}} p(D|\text{pre}_{\prec}(D))$  is only a function of  $\mathbf{D} \cup \text{pa}_{\mathcal{G}}(\mathbf{D})$  under the nested factorization.

An important result in [Richardson et al., 2017] states that if  $p(\mathbf{V} \cup \mathbf{H}|\mathbf{W})$  obeys the factorization for a CDAG  $\mathcal{G}(\mathbf{V} \cup \mathbf{H}, \mathbf{W})$ , then  $p(\mathbf{V}|\mathbf{W})$  obeys the nested factorization for the latent projection CADMG  $\mathcal{G}(\mathbf{V}, \mathbf{W})$ .

**Identification:** Not every interventional distribution  $p(\mathbf{Y}(\mathbf{a}))$  is identified in a hidden variable causal model. However, *every*  $p(\mathbf{Y}(\mathbf{a})|\mathbf{W})$  identified from  $p(\mathbf{V}|\mathbf{W})$  can be expressed as a modified nested factorization as follows:

$$\begin{aligned} p(\mathbf{Y}(\mathbf{a})|\mathbf{W}) &= \sum_{\mathbf{Y}^* \setminus \mathbf{Y}} \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}_{\mathbf{Y}^*})} p(\mathbf{D} | \text{do}(\text{pa}_{\mathcal{G}}^s(\mathbf{D})))|_{\mathbf{A}=\mathbf{a}} \\ &= \sum_{\mathbf{Y}^* \setminus \mathbf{Y}} \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}_{\mathbf{Y}^*})} \phi_{\mathbf{V} \setminus \mathbf{D}}(p(\mathbf{V}|\mathbf{W}); \mathcal{G}(\mathbf{V}, \mathbf{W}))|_{\mathbf{A}=\mathbf{a}} \end{aligned}$$

where  $\mathbf{Y}^* \equiv \text{an}_{\mathcal{G}(\mathbf{V}(\mathbf{a}), \mathbf{W})}(\mathbf{Y}) \setminus \mathbf{a}$ . That is,  $p(\mathbf{Y}(\mathbf{a})|\mathbf{W})$  is only identified if it can be expressed as a factorization, where every piece corresponds to a kernel associated with a set intrinsic in  $\mathcal{G}(\mathbf{V}, \mathbf{W})$ . Moreover, no piece in this factorization contains elements of  $\mathbf{A}$  as random variables.

## C. $k$ th Order Markov Temporal Causal Models

Causal DBNs may be generalized to  $k$ th-order Markov models, where variables in a particular time step depend on variables in at most  $k$  prior states.

A  $k$ th-order Markov DBN consists of a single *prior network*  $\mathcal{G}_1$ , which is a DAG with vertices  $\mathbf{V}_1$ , a set of  $k-1$  initial transition networks  $\mathcal{G}_2, \dots, \mathcal{G}_k$ , where each  $\mathcal{G}_i$  is a CDAG with random vertices  $\mathbf{V}_i$  and fixed vertices  $\bigcup_{j=1}^{i-1} \vec{V}_j$ , and a transition network  $\mathcal{G}_t$  with random vertices  $\mathbf{V}_t$  and fixed vertices  $\bigcup_{j=t-k}^{t-1} \vec{V}_j$ . Each DAG and CDAG in a  $k$ th-order DBN is associated with a factorization of the corresponding joint or conditional distribution. The “unrolled” factorization of the DBN makes use of the prior distribution  $p_1(\mathbf{V}_1)$  and initial transition network distributions  $p_i(\mathbf{V}_i|\mathbf{V}_1, \dots, \mathbf{V}_{i-1})$  for the first  $k-1$  steps, and then uses a repeated version of the transition network distribution  $p_t(\mathbf{V}_t|\mathbf{V}_{t-k}, \dots, \mathbf{V}_{t-1})$ :

$$\prod_{V \in \mathbf{V}_1} p_1(V | \text{pa}_{\mathcal{G}_1}(V)) \prod_{\tilde{t}=1}^{k-1} \prod_{V \in \mathbf{V}_{\tilde{t}}} p_{\tilde{t}}(V | \text{pa}_{\mathcal{G}_{\tilde{t}}}(V)) \cdot \prod_{\tilde{t}=k}^{T-1} \prod_{V \in \mathbf{V}_{\tilde{t}}} p_{\tilde{t}}(V | \text{pa}_{\mathcal{G}_{\tilde{t}}}(V)). \quad (\text{I.1})$$

The causal version of a  $k$ th-order Markov DBN is obtained in the natural way by endowing each DAG and CDAG with structural equation model semantics, and obtaining standard identification results, via the g-formula, and the ID algorithm in cases hidden variables are present.

The relaxation of the first-order Markov assumption in these models does not come without a cost: additional transition networks must be specified, and all transition networks may potentially depend on a larger set of variables, resulting in a more difficult statistical inference problem on model parameters.

# Appendix II

## Path Dependent Models: Additional Material

### A. $k$ -th order Markov Temporal Causal Models

PDSEMs may also be relaxed to a  $k$ th-order Markov model, similar to DBNs (shown in Appendix I). For example, given a model with 3 states, if we wish all transitions to depend on two rather than one prior state, we would need to specify a prior network (with a corresponding causal model), a set of 3 single-step transition networks (corresponding to steps from the initial state to any of the 3 possible states), and then finally a set of 9 transition networks, representing variables in one of three states that depend on any two prior states (which may involve states repeating). Such a model would have a separate transition network  $\mathcal{G}_{\langle 1,2,3 \rangle}$  for variables in state 3 at time  $t$ , where state 2 was visited at time  $t - 1$ , and state 1 was visited at time  $t - 2$ , and a transition network  $\mathcal{G}_{\langle 2,1,3 \rangle}$  for variables in state 3 at time  $t$ , where state 1 was visited at time  $t - 1$ , and state 2 was visited at time  $t - 2$ .

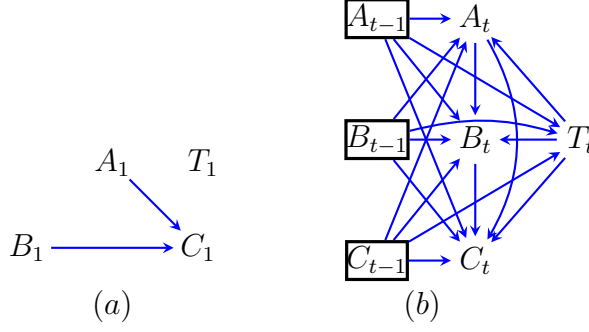
In general, a  $k$ th-order Markov PDSEM with  $S$  states will have a single prior network DAG corresponding to the initial state,  $S^i$  transition networks CDAGs that depend on  $i$  prior states (for  $i = 1, \dots, k - 1$ ), indexed by sequences of states visited (starting with the initial state and ending in one of the states  $s \in S$ ), and  $S^{k+1}$  transition network CDAGs that depend on  $k$  prior states, indexed by sequences of states visited, and ending in one of the states  $s \in S$ . Note that the initial transition networks all assume that the starting state is the initial state, while the transition network does not.

In addition, a  $k$ th-order Markov PDSEM makes the following assumption, that generalizes Assumption 2 in Chapter 3 of the main text:

**Assumption 5.** *For every state  $s^j$ , any CDAG  $\mathcal{G}_{\langle \dots j \dots \rangle}$  or DAG  $\mathcal{G}_j$  that mentions variables in state  $j$  will have corresponding random variables that share state spaces.*

As was the case with DBNs, each DAG or CDAG in a PDSEM is associated with a causal model, which induces an appropriate DAG or CDAG factorization and g-formula for identification of interventional distributions. These, in turn, yield PDSEM factorizations that naturally generalize those in Section 3.4.

Let  $\mathcal{T}_{\tilde{k}}$  be a set of all valid state transition sequences  $\sigma_{\tilde{k}}$  of size  $\tilde{k} = 1, \dots, k - 1$  that start with the initial state, and  $\mathcal{T}$  be a set of all valid state transition sequences  $\sigma_k$  of size  $k$ . Further, let  $\mathbf{V}_{\sigma}$  be random variables in the final state in a state transition sequence  $\sigma$ , while  $\mathbf{W}_{\sigma}$  be fixed variables in states prior to the final state in  $\sigma$ . Finally, let  $\mathbb{I}(\sigma)$  be the indicator that the current state is the final



**Figure II.1.** A causal DBN encoding the PDSEM in Fig. 3.2, via (a) the prior network, and (b) the complete transition network with context-specific independences.

state in  $\sigma$ , and the  $|\sigma| - 1$  prior states were the states prior to the last state in  $\sigma$ . We then obtain the following observed data factorization of the  $k$ th-order Markov PDSEM:

$$\begin{aligned}
& p_1(\mathbf{V}_1) \prod_{t=1}^{k-1} \left( \prod_{\sigma_{\tilde{k}} \in \mathcal{T}_{\tilde{k}}} \left( p(\mathbf{V}_{\sigma_{\tilde{k}}} | \mathbf{W}_{\sigma_{\tilde{k}}}) \right)^{\mathbb{I}(\sigma_{\tilde{k}})} \right) \prod_{t=k}^{\infty} \left( \prod_{\sigma_k \in \mathcal{T}} \left( p(\mathbf{V}_{\sigma_k} | \mathbf{W}_{\sigma_k}) \right)^{\mathbb{I}(\sigma_k)} \right) 1^{\mathbb{I}(s_t^*)} \\
& p_1(\mathbf{V}_1) = \prod_{V \in \mathbf{V}_1} p(V | \text{pa}_{\mathcal{G}_1}(V)); \quad p_{\sigma_{\tilde{k}}}(\mathbf{V}_{\sigma_{\tilde{k}}} | \mathbf{W}_{\sigma_{\tilde{k}}}) = \prod_{V \in \mathbf{V}_{\sigma_{\tilde{k}}}} p(V | \text{pa}_{\mathcal{G}_{\sigma_{\tilde{k}}}}(V)); \quad p_{\sigma_k}(\mathbf{V}_{\sigma_k} | \mathbf{W}_{\tilde{k}}) \\
& = \prod_{V \in \mathbf{V}_{\sigma_k}} p(V | \text{pa}_{\mathcal{G}_{\sigma_k}}(V));
\end{aligned}$$

Extensions to truncated factorizations representing interventional distributions, and hidden variable versions of these models are straightforward generalizations of the  $k = 1$  case, described in the main body.

As was the case with DBNs, relaxation of the first-order Markov assumption to a  $k$ th-order Markov assumption comes at a cost – many additional transition networks must be specified, and the resulting statistical inference is more likely to suffer from the curse of dimensionality.

## B. Representing a PDSEM as a DBN

If variables in all transition networks in a PDSEM obey a single consistent topological order, one may encode a PDSEM by a causal DBN as follows. First, define a transition variable  $T$  with values representing all possible state transition pairs  $(s_i, s_j)$  in a PDSEM. Then, use this variable as a parent of every variable in the single transition network allowed by a DBN, and use it to select a subset of all possible parents to implement transition specific networks of a PDSEM via context-specific independence.

In the example shown in Fig. 3.2, one topological order on variables that is consistent for the prior network and all transition networks is  $A \prec B \prec C$ . Thus, a causal DBN representing the example PDSEM would have a prior network shown in Fig. II.1 (a), and a complete conditional DAG as a transition network shown in Fig. II.1 (b), with a factorization:  $p(C_t | B_t, A_t, T_t, V_{t-1}) \cdot p(B_t | A_t, T_t, V_{t-1}) \cdot p(A_t | T_t, V_{t-1}) \cdot p(T_t | V_{t-1})$ , where  $V_{t-1} \equiv C_{t-1}, B_{t-1}, A_{t-1}, T_{t-1}$ . Note that in this transition network, every state variable has the transition variable as a parent, and this parent is used to implement state transition independences in a PDSEM via context-specific independence. For example, the Markov factor  $p(B_t | A_t, T_t, C_{t-1}, B_{t-1}, A_{t-1}, T_{t-1})$  will not depend on  $A_t$  unless  $T_t$  has value  $(s_2, s_3)$ .

Note that this representation, is in some sense, isomorphic to PDSEMs. The causal DBN factorization exhibits no independences, and all interesting probabilistic and causal structure is obtained via context-specific independences, which would be represented explicitly in transition networks of a PDSEM.

In addition, if no consistent topological order on variables in all transition networks in a PDSEM exists, then there is no known representation scheme for such a PDSEM using causal DBNs.

## C. Proofs

**Lemma 1** *Under Assumption 1,  $p(\mathbf{Y}(\mathbf{a}))$  is identified from a hidden variable causal DBN model represented by latent projections  $\mathcal{G}_1$  on  $\mathbf{V}_1$  and  $\mathcal{G}_{+1}$  on  $\mathbf{V}_{t+1}$  given  $\mathbf{V}_t$  if and only if every bidirected connected component in  $\mathcal{G}_{1,\mathbf{Y}_1^*}$  (the induced subgraph of  $\mathcal{G}_1$ ) is intrinsic in  $\mathcal{G}_1$ , and every bidirected component in  $\mathcal{G}_{+1,\mathbf{Y}_i^*}$  (the induced subgraph of  $\mathcal{G}_{+1}$ ) is intrinsic in  $\mathcal{G}_{+1}$ , where  $\mathbf{Y}_1^*$  is the set of ancestors of  $\mathbf{Y} \cap \mathbf{V}_1$  not through  $\mathbf{A} \cap \mathbf{V}_1$  in  $\mathcal{G}_1$ , and for every  $i \in 2, \dots, T$ ,  $\mathbf{Y}_i^*$  is the set of ancestors of  $\mathbf{Y} \cap \mathbf{V}_i$  not through  $\mathbf{A} \cap \mathbf{V}_i$  in  $\mathcal{G}_{+1}$ . Moreover, if  $p(\mathbf{Y}(\mathbf{a}))$  is identified, we have*

$$\left( \sum_{\mathbf{Y}_1^* \setminus ((\mathbf{Y} \cup \mathbf{A}) \cap \mathbf{V}_1)} \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}_{1,\mathbf{Y}_1^*})} q_{\mathbf{D}}^1(\mathbf{D} | \text{pa}_{\mathcal{G}}(\mathbf{D}) \setminus \mathbf{D}) |_{\mathbf{A}=\mathbf{a}} \right) \times \prod_{i=2}^T \left( \sum_{\mathbf{Y}_i^* \setminus ((\mathbf{Y} \cup \mathbf{A}) \cap \mathbf{V}_i)} \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}_{+1,\mathbf{Y}_i^*})} q_{\mathbf{D}}^{+1}(\mathbf{D} | \text{pa}_{\mathcal{G}}(\mathbf{D}) \setminus \mathbf{D}) |_{\mathbf{A}=\mathbf{a}} \right),$$

where  $q_{\mathbf{D}}^1$  and  $q_{\mathbf{D}}^{+1}$  are kernels corresponding to intrinsic sets representing elements of  $\mathcal{D}(\mathcal{G}_{1,\mathbf{Y}_1^*})$  and  $\mathcal{D}(\mathcal{G}_{+1,\mathbf{Y}_i^*})$  in the nested Markov factorizations of  $\mathcal{G}_1$  and  $\mathcal{G}_{+1}$ , respectively.

*Proof:* We want to obtain  $p(\mathbf{Y}(\mathbf{a}))$  from the observed joint  $p(\mathbf{V}_{1:T})$ . Using identification result 2.5 on the unrolled ADMG gives  $\sum_{\mathbf{Y}^* \setminus \mathbf{Y}} p(\mathbf{Y}^*(\mathbf{a})) = \sum_{\mathbf{Y}^* \setminus \mathbf{Y}} \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}_{\text{unrolled}\mathbf{Y}^*})} p(\mathbf{D}(\text{pa}(\mathbf{D}) \setminus \mathbf{D})) |_{\mathbf{A}=\mathbf{a}}$ . Assumption 1 ensures that no district  $\mathbf{D}$  spans time points, and parents  $\text{pa}(\mathbf{D})$  at time  $t+1$  lie either at  $t$  or  $t+1$ . This allows us to write  $\sum_{\mathbf{Y}^* \setminus \mathbf{Y}} p(\mathbf{Y}^*(\mathbf{a})) = \sum_{\mathbf{Y}^* \setminus \mathbf{Y}} \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}_{1,\mathbf{Y}^*})} p(\mathbf{D}(\text{pa}(\mathbf{D}) \setminus \mathbf{D})) |_{\mathbf{A}=\mathbf{a}} \times \prod_{t=1}^{T-1} \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}_{+1,\mathbf{Y}^*})} p(\mathbf{D}(\text{pa}(\mathbf{D}) \setminus \mathbf{D})) |_{\mathbf{A}=\mathbf{a}}$ . Applying the identification results in [Richardson et al., 2012] to the prior network ADMG  $\mathcal{G}_1$  and extensions of these results in [Sherman and Shpitser, 2018] to the transition network CADMGs  $\mathcal{G}_{+1}$ , these counterfactual conditionals can be replaced by given modified nested factorizations, provided every appropriate bidirected connected set in the prior or transition graph is intrinsic in that graph.

Note that completeness of our procedure does not immediately follow from the completeness argument in [Shpitser and Pearl, 2006]. This is because a completeness argument entails constructing in any ADMG  $\mathcal{G}(\mathbf{V})$  where identification fails two causal models which agree on the observed data distribution  $p(\mathbf{V})$ , but disagree on  $p(\mathbf{Y}(\mathbf{a}))$ . Furthermore, the construction employed in [Shpitser and Pearl, 2006] relied on an unrestricted causal model inducing a given latent projection ADMG  $\mathcal{G}(\mathbf{V})$ . However, in the case of causal DBNs, the model is *not* unrestricted – indeed there is a very strong restriction that all transition networks at any time point share all structural equations.

Nevertheless, it is possible to extend the completeness proof in [Shpitser and Pearl, 2006] to yield completeness of the procedure in this lemma by employing an extended construction modeled after one in [Shpitser and Sherman, 2018].

From this point on, we will refer to  $\mathcal{G}_{1:T}(\mathbf{V}_{1:T})$  by  $\mathcal{G}(\mathbf{V})$  for simplicity. Assume  $p(\mathbf{Y}(\mathbf{a}))$  is not identified in  $\mathcal{G}(\mathbf{V})$ , and assume there exists a hedge structure ancestral of  $\mathbf{Y}'$ . Note that by first order Markov assumption, the hedge structure must lie entirely in a transition network in a single time step. Fix a subgraph  $\tilde{\mathcal{G}}$  of  $\mathcal{G}(\mathbf{V})$  containing the hedge, the set  $\mathbf{Y}$ , a set of vertices  $\mathbf{S}$  making up

directed paths from every element of the root set  $\mathbf{R}$  to some element of  $\mathbf{Y}'$  (without loss of generality we assume these vertices do not have more than one child).

We extend  $\tilde{\mathcal{G}}$  with a new set of vertices  $\mathbf{S}^*$  that are copies of  $\mathbf{S}$  with the property that if  $S \in \mathbf{S}$  has a parent in  $\mathbf{R}$ , so does the corresponding  $S^* \in \mathbf{S}^*$ , and if  $T \in \mathbf{S}$  is a parent of  $S \in \mathbf{S}$ , the corresponding  $T^* \in \mathbf{S}^*$  is a parent of  $S^* \in \mathbf{S}^*$ . We then apply the counterexample construction connecting the hedge structure to  $\mathbf{Y}'$  appearing in [Shpitser and Pearl, 2006] to elements of  $\mathbf{S}^*$ . In particular, we make sure that  $\sum_{\mathbf{S}^*} p(\mathbf{Y}'|\mathbf{S}^*)p(\mathbf{S}^*|\mathbf{R})$  is a one-to-one map. This implies  $p(\mathbf{Y}(\mathbf{a}))$  is not identified in an extended model containing vertices  $\mathbf{V}$  and  $\mathbf{S}^*$ . Lemma 1 in [Shpitser and Sherman, 2018] then implies  $p(\mathbf{Y}(\mathbf{a}))$  is also not identified in  $\mathcal{G}(\mathbf{V})$ , establishing our result.  $\square$

**Lemma 2** *Given a fully observed PDSEM, each factor of the distribution  $p_\infty(\mathbf{Y}(\mathbf{a}))$  is identified from  $p_\infty(\mathbf{V})$  as:*

$$p_1(\mathbf{Y}_1(\mathbf{a}_1)) \equiv \prod_{V \in \mathbf{Y}_1 \setminus \mathbf{A}_1} p_1(V | \text{pa}_{\mathcal{G}_1}(V)) \Big|_{\mathbf{A}_1 = \mathbf{a}_1}$$

$$p_{ij}(\mathbf{Y}_{ij}(\mathbf{a}_j) | \mathbf{Y}_i(\mathbf{a}_i)) \equiv \prod_{V \in \mathbf{Y}_{ij} \setminus \mathbf{A}_j} p_{ij}(V | \text{pa}_{\mathcal{G}_{ij}}(V)) \Big|_{\substack{\mathbf{A}_i = \mathbf{a}_i, \\ \mathbf{A}_j = \mathbf{a}_j}}$$

*Proof:* This follows from the factorization of  $p_\infty(\mathbf{V}(\mathbf{a}))$  into elements of the form  $p_1(\mathbf{Y}_1(\mathbf{a}_1))$ , and  $p_{ij}(\mathbf{Y}_j(\mathbf{a}_j) | \mathbf{Y}_i(\mathbf{a}_i))$ , the fact that  $\mathcal{G}_1, \{\mathcal{G}_{ij} : (i, j) \in \mathcal{T}\}$  define causal models under standard structural equation semantics, and equation 2.1.  $\square$

**Lemma 3** *Under Assumptions 1, 2 and 3, given a latent variable PDSEM represented by  $\mathcal{G}_1$  and  $\{\mathcal{G}_{ij} : (i, j) \in \mathcal{T}\}$ ,  $p_\infty(\mathbf{Y}(\mathbf{a}))$  is identified from  $p_\infty(\mathbf{V})$  if and only if every bidirected component in  $\mathcal{G}_{1\mathbf{Y}_1^*}$  is intrinsic in  $\mathcal{G}_1$ , and every bidirected component in  $\mathcal{G}_{ij\mathbf{Y}_j^*}$  is intrinsic in  $\mathcal{G}_{ij}$  for every  $i$  and  $j$ . Moreover, if  $p_\infty(\mathbf{Y}(\mathbf{a}))$  is identified, it is equal to*

$$p_1(\mathbf{Y}_1(\mathbf{a}_1)) \prod_{t=1}^{\infty} \left( \prod_{(i,j) \in \mathcal{T}} (p_{ij}(\mathbf{Y}_{ij}(\mathbf{a}_j) | \mathbf{Y}_i(\mathbf{a}_i)))^{\mathbb{I}(s_{t-1}^i, s_t^j)} \right) 1^{\mathbb{I}(s_{t-1}^*)} \quad (\text{II.1})$$

where

$$p_1(\mathbf{Y}_1(\mathbf{a}_1)) = \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}_{1\mathbf{Y}_1^*})} q_{\mathbf{D}}^1(\mathbf{D} | \text{pa}_{\mathcal{G}_1}^s(\mathbf{D})) \Big|_{\mathbf{A}_1 = \mathbf{a}_1}, \quad (\text{II.2})$$

where each kernel  $q_{\mathbf{D}}^1(\mathbf{D} | \text{pa}_{\mathcal{G}_1}^s(\mathbf{D}))$  is in the nested Markov factorization of  $p_1(\mathbf{V}_1)$  with respect to  $\mathcal{G}_1$ , and

$$p_{ij}(\mathbf{Y}_{ij}(\mathbf{a}_j) | \mathbf{Y}_i(\mathbf{a}_i)) = \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}_{\mathbf{V}_{ij} \setminus \mathbf{A}_{ij}})} q_{\mathbf{D}}^{ij}(\mathbf{D} | \text{pa}_{\mathcal{G}_{ij}}^s(\mathbf{D})) \Big|_{\substack{\mathbf{A}_i = \mathbf{a}_i, \\ \mathbf{A}_j = \mathbf{a}_j}} \quad (\text{II.3})$$

where each kernel  $q_{\mathbf{D}}^{ij}(\mathbf{D} | \text{pa}_{\mathcal{G}_{ij}}^s(\mathbf{D}))$  is in the nested Markov factorization of  $p_{ij}(\mathbf{V}_{ij} | \mathbf{V}_i)$  with respect to  $\mathcal{G}_{ij}$ .

*Proof:* Assumption 3 implies all state transitions are known, and thus allows us to proceed by induction on any sequence of state transitions with positive probability after  $t$  steps.

Unrolling the prior network, and appropriate transition networks for such a sequence yields an ADMG representing the observed data distribution had that transition taken place, with Assumption 1 implying that districts in this ADMG do not span multiple time steps. This immediately implies the conclusion by the same argument used in the proof of Lemma 1.

In fact, this argument works for any transition sequence of any size.  $\square$

# Appendix III

## Entangled Missingness: Additional Material

### A. Binary Parameterization of Nested Markov Models

From the nested factorization (discussed in Section B. of Appendix I), intrinsic sets given their parents form the atomic units of the nested Markov model. Using this observation, a smooth parameterization of discrete nested Markov models was provided by [Evans and Richardson, 2014]. We provide a brief description of how to derive the Möbius parameters of a binary nested Markov model.

For each district  $\mathbf{D} \in \mathcal{D}(\mathcal{G})$ , consider all possible subsets  $\mathbf{S} \subseteq \mathbf{S}$ . If  $\mathbf{S}$  is intrinsic in  $\phi_{\mathbf{V} \setminus \mathbf{S}}(\mathcal{G})$ , define the head  $\mathbf{H}$  of the intrinsic set to be all vertices in  $\mathbf{S}$  that are childless in  $\phi_{\mathbf{V} \setminus \mathbf{S}}(\mathcal{G})$  and the tail  $\mathbf{T}$  to be all parents of the head in the CADMG  $\phi_{\mathbf{V} \setminus \mathbf{S}}(\mathcal{G})$ , excluding the head itself. Formally,  $\mathbf{H} \equiv \{V \in \mathbf{S} \mid \text{ch}_{\phi_{\mathbf{V} \setminus \mathbf{S}}(\mathcal{G})}(V) = \emptyset\}$  and  $\mathbf{T} \equiv \text{pa}_{\phi_{\mathbf{V} \setminus \mathbf{S}}(\mathcal{G})}(\mathbf{H}) \setminus \mathbf{H}$ . The corresponding set of Möbius parameters for this intrinsic head and tail pair parameterizes the kernel  $q_{\mathbf{S}}(\mathbf{H} = \mathbf{0} \mid \mathbf{T})$ ; i.e., the kernel where all variables outside the intrinsic set  $\mathbf{S}$  are fixed, and all elements of the head are set to zero given the tail. Note that these parameters are, in general, variationally dependent (in contrast to variationally independent in the case of an ordinary DAG model) as the heads and tails in these parameter sets may overlap. The joint density for any query  $p(\mathbf{V} = \mathbf{v})$  can be obtained through the Möbius inversion formula; see [Lauritzen, 1996, Evans and Richardson, 2014] for details. We will denote  $q_{\mathbf{S}}(\mathbf{H} = \mathbf{0} \mid \mathbf{T})$  simply as  $q(\mathbf{H} = \mathbf{0} \mid \mathbf{T})$  as it is generally clear what variables are still random in the kernel corresponding to a given intrinsic set.

#### A..1 Binary Parameterization of Missing Data Models

We use the parameterization described earlier to count the number of parameters required to parameterize the full observability law of a missing data ADMG and its corresponding observed law. We then use this to reason that if the number of parameters in the full observability law exceeds those in the observed law, it is impossible to establish a map from the observed law to the full law. This in turn implies that such a full observability law is not identified. In the full observability, the deterministic factors, i.e., proxies given parents can be ignored as the probability of those events is always 1. However, while counting the observed law, we are careful to treat counterfactuals as unobserved and obtain the corresponding ADMG. The Möbius parameters are then derived in a similar manner as before, but with additional constraint that if  $Z_i$  appears in the head of a parameter, and missingness indicators  $R_i$  or  $\mathbf{R}_{\text{aff}(i)}$  appear in the tail, then the kernel must be restricted to

cases where  $R_i = 1$  and  $\mathbf{R}_{\text{aff}(i)} = \mathbf{1}$ . This is because, (1) when  $R_i = 0$ , the probability of the head taking any value aside from those where  $Z_i = ?$  is deterministically 0, and (2) cases where  $\mathbf{R}_{\text{aff}(i)}$  are set to values different from 1 are irrelevant to the identification of the full observability law. The consideration that  $R_i = 1$  always holds, but that  $\mathbf{R}_{\text{aff}(i)} = \mathbf{1}$  is only for identifying the full observability law. For other cases, considerations vary.

## B. Proofs

We restate the theorems and outline their proofs here.

**Theorem 3** *In a missing data ADMG  $\mathcal{G}$  with missingness interference, valid single-world objects  $h(\tilde{\mathbf{Z}}; \mathbf{r})$  consisting of a set of counterfactuals  $\mathbf{Z}' \equiv \bigcup_i \{\mathbf{Z}_i^{(1, \mathbf{r}_{\text{aff}(i)})}\}$ ,  $i \in \{1, \dots, n\}$  are identified when either of these two conditions is satisfied: (1)  $\mathbf{R}' \perp\!\!\!\perp \mathbf{O}, \tilde{\mathbf{Z}}$  (MCAR), or (2)  $\mathbf{R}' \perp\!\!\!\perp \tilde{\mathbf{Z}} | \mathbf{O}$  (MAR), where  $\mathbf{R}'$  refers to the set of all missingness indicators  $R$  that index counterfactuals in  $h(\tilde{\mathbf{Z}}; \mathbf{r})$ . The object  $h(\tilde{\mathbf{Z}}; \mathbf{r})$  is a function of  $p(\mathbf{Z}', \mathbf{R}, \mathbf{O})$ , and the identifying functional is given by:*

$$p(\mathbf{Z}', \mathbf{R}, \mathbf{O}) = p(\mathbf{Z}', \mathbf{O}) \times p(\mathbf{R} | \mathbf{O}, \mathbf{Z}') = \frac{p(\mathbf{Z}', \mathbf{R} = \mathbf{r}, \mathbf{O})}{p(\mathbf{R} = \mathbf{r} | \mathbf{O})} \times p(\mathbf{R} | \mathbf{O}) \quad (\text{III.1})$$

where propensity scores are obtained by simple  $m$ -separation or ( $d$ -separation) rules on ADMG (or DAG) factorization.

*Proof:*  $\mathbf{Z}'$  is the set of counterfactuals in  $h(\tilde{\mathbf{Z}}; \mathbf{r})$  such that they construct a valid single-world counterfactual in the world  $\mathbf{R} = \mathbf{r}$ . Further, let the set of all missingness indicators that index  $\mathbf{Z}'$  be  $\mathbf{R}'$ , and that  $\mathbf{R}' = \mathbf{r}'$  when  $\mathbf{R} = \mathbf{r}$ . We are interested in identifying a distribution  $P(\mathbf{Z}')$  or a function  $h(\tilde{\mathbf{Z}}; \mathbf{r})$  thereof. The crucial factor here is that no variable, whose missingness indicator is 0, is present in  $h(\cdot)$ .

In case (1), we can write  $P(\mathbf{Z}') = P(\mathbf{Z}' | \mathbf{R}' = \mathbf{r}')$  since  $\mathbf{R}' \perp\!\!\!\perp \tilde{\mathbf{Z}}$  and  $\mathbf{Z}' \subset \tilde{\mathbf{Z}}$ . And by consistency, we can replace all counterfactuals by their corresponding proxies and the object is identified.

In case (2), we can write  $P(\mathbf{Z}') = P(\mathbf{Z}' | \mathbf{O}) \times P(\mathbf{O}) = P(\mathbf{Z}' | \mathbf{O}, \mathbf{R}' = \mathbf{r}') \times P(\mathbf{O})$  since  $\mathbf{R}' \perp\!\!\!\perp \tilde{\mathbf{Z}} | \mathbf{O}$  and  $\mathbf{Z}' \subset \tilde{\mathbf{Z}}$ . And by consistency, we can identify the distribution as we can replace all counterfactuals by their corresponding proxies.

□

**Theorem 4** *In a missing data ADMG  $\mathcal{G}$  with missingness interference, under Assumption 4, the full-observability law  $P(\tilde{\mathbf{Z}}^{(r=1)}, \mathbf{R})$  is identified if and only if there is no  $e$ -colluding path. Further, if  $\mathcal{G}$  is a missing data DAG, the full-observability law is identified if and only if there is no  $e$ -colluder and no  $e$ -self-censoring. The identifying functional is given by*

$$p(\tilde{\mathbf{Z}}^{(r=1)}, \mathbf{R}) = p(\tilde{\mathbf{Z}}^{(r=1)}) \times \underbrace{p(\mathbf{R} | \tilde{\mathbf{Z}}^{(r=1)})}_{g(\mathbf{p}(\mathbf{R}, \mathbf{Z}))} = \frac{p(\tilde{\mathbf{Z}}^{(r=1)}, \mathbf{R} = \mathbf{1})}{\underbrace{p(\mathbf{R} = \mathbf{1} | \tilde{\mathbf{Z}}^{(r=1)})}_{g(\mathbf{p}(\mathbf{R}, \mathbf{Z}))|_{\mathbf{R}=\mathbf{1}}}} \times \underbrace{p(\mathbf{R} | \tilde{\mathbf{Z}}^{(r=1)})}_{g(\mathbf{p}(\mathbf{R}, \mathbf{Z}))}.$$

and missingness mechanism  $p(\mathbf{R} | \tilde{\mathbf{Z}}^{(r=1)})$  is identified using the OR parameterization given below:

$$p(\mathbf{R} | \tilde{\mathbf{Z}}^{(r=1)}) = \frac{1}{\sigma} \times \prod_{k=1}^K p(R_k | R_{-k} = 1, \tilde{\mathbf{Z}}^{(r=1)}) \times \prod_{k=2}^K \text{OR}(R_k, R_{\prec k} | R_{\succ k} = 1, \tilde{\mathbf{Z}}^{(r=1)})$$

**Soundness:** The absence of  $e$ -colluding paths results in identification.

The absence of a e-colluding path between  $Z_i^{(1, \mathbf{r}_{\text{aff}(i)})}$  and  $R_k \in \mathbf{R}_{\text{aff}(i)}$  implies that  $Z_i^{(1, \mathbf{r}_{\text{aff}(i)})} \notin \text{mb}_{\mathcal{G}}(R_k)$  where

$$\text{mb}_{\mathcal{G}}(V) = \{\text{pa}_{\mathcal{G}}(V), \text{dis}_{\mathcal{G}}(V), \text{pa}_{\mathcal{G}}(\text{dis}_{\mathcal{G}}(V)), \text{ch}_{\mathcal{G}}(V), \text{pa}_{\mathcal{G}}(\text{ch}_{\mathcal{G}}(V)), \\ \text{dis}_{\mathcal{G}}(\text{ch}_{\mathcal{G}}(V)), \text{pa}_{\mathcal{G}}(\text{dis}_{\mathcal{G}}(\text{ch}_{\mathcal{G}}(V)))\}$$

Let  $\tilde{\mathbf{Z}}^{(\mathbf{r}_{\mathbf{k}}=1)} = \{\mathbf{Z}_{\mathbf{i}}^{(1, \mathbf{r}_{\text{aff}(\mathbf{i})}=1)} : \mathbf{R}_{\mathbf{k}} \in \mathbf{R}_{\text{aff}(\mathbf{i})}\} \cup \mathbf{Z}_{\mathbf{k}}^{(1, \mathbf{r}_{\text{aff}(\mathbf{k})})}$ . In words,  $\tilde{\mathbf{Z}}^{(\mathbf{r}_{\mathbf{k}}=1)}$  is the set of all counterfactuals that correspond to full observability, and are indexed (and influenced) by  $R_k$ .

By Markov property, we have that  $V \perp\!\!\!\perp \mathbf{V} \setminus \text{mb}_{\mathcal{G}}(V) | \text{mb}_{\mathcal{G}}(V)$ . Therefore, the absence of e-colluding paths implies the following assumptions:

$$R_k \perp\!\!\!\perp \{Z_i^{(1, \mathbf{r}_{\text{aff}(i)})} \in \tilde{\mathbf{Z}}^{(\mathbf{r}_{\mathbf{k}}=1)}\} \mid \{\mathbf{R} \setminus \mathbf{R}_{\mathbf{k}}, \tilde{\mathbf{Z}} \setminus \mathbf{Z}_{\mathbf{i}}^{(1, \mathbf{r}_{\text{aff}(i)})}\} \quad \text{for } \mathbf{R}_{\mathbf{k}} \in \mathbf{R}$$

Given these assumptions, we can identify  $p(\mathbf{R} | \tilde{\mathbf{Z}}^{(\mathbf{r}=1)})$  using the OR parameterization. The proof is similar to the identification proof of the no self-censoring model given in [Malinsky et al., 2021] and the representation of it in [Nabi et al., 2020]. Since we are assuming  $\text{ch}_{\mathcal{G}}(\tilde{\mathbf{Z}}^{(\mathbf{r} \neq 1)}) \cap \mathbf{R} = \emptyset$ , it suffices to only ID  $p(\mathbf{R} | \tilde{\mathbf{Z}}^{(\mathbf{r}=1)})$ .

$$p(\mathbf{R} | \tilde{\mathbf{Z}}^{(\mathbf{r}=1)}) = \frac{1}{\sigma} \times \prod_{k=1}^K p(R_k | \mathbf{R}_{-k} = 1, \tilde{\mathbf{Z}}^{(\mathbf{r}=1)}) \times \prod_{k=2}^K \text{OR}(R_k, \mathbf{R}_{\prec k} | \mathbf{R}_{\succ k} = 1, \tilde{\mathbf{Z}}^{(\mathbf{r}=1)})$$

where notation and OR is consistent with Section 2.6.

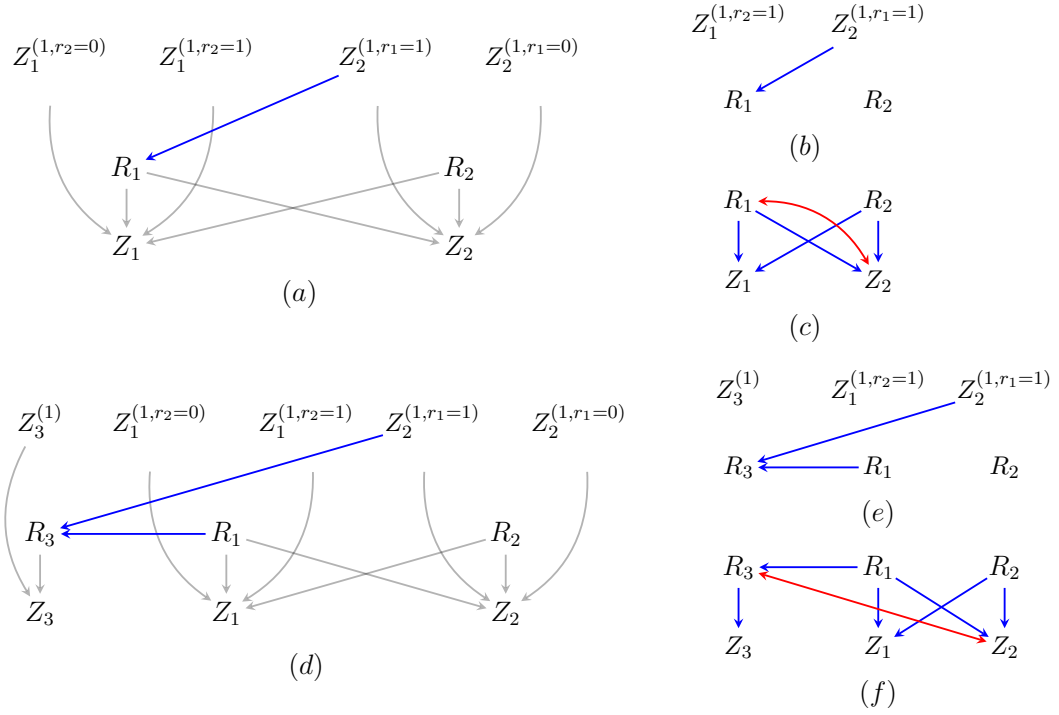
We have  $p(R_k | \mathbf{R}_{-k} = 1, \tilde{\mathbf{Z}}^{(\mathbf{r}=1)}) = p(R_k | \mathbf{R}_{-k} = 1, \tilde{\mathbf{Z}}^{(\mathbf{r}=1)} \setminus \tilde{\mathbf{Z}}^{(\mathbf{r}_{\mathbf{k}}=1)})$ . By consistency, every counterfactual past the conditioning bar is equal to the observed proxy. Then, pairwise OR terms are identified because  $\text{OR}(R_k, R_i | \mathbf{R}_{-(k,i)} = 1, \tilde{\mathbf{Z}}^{(\mathbf{r}=1)})$  is not a function of  $\tilde{\mathbf{Z}}^{(\mathbf{r}_i=1)}$  or  $\tilde{\mathbf{Z}}^{(\mathbf{r}_{\mathbf{k}}=1)}$ . Finally higher order terms are ID in similar ways [Nabi et al., 2020, Malinsky et al., 2021].

**Completeness:** One approach to demonstrate completeness is to count the number of parameters required to parameterize the full-observability law of a missing data graph and its corresponding observed law and reason that if the former requires more parameters than the latter, it is impossible to identify all the parameters of the full-observability law uniquely. In order to do so, we assume all variables are binary, and adopt the binary parameterization of Nested Markov Models (see Appendix I).

First, we present two simple e-colluding path examples to show how the parameter counting argument proceeds. Thereafter, we provide the general argument for completeness for all graphs.

Consider the simple graph in Fig. III.1(a) which has e-self-censoring  $Z_2^{(1, r_1=1)} \rightarrow R_1$ . To count the parameters required for the full-observability law, we follow the procedure in [Nabi et al., 2020]. The binary parameterization of the full law of a missing data ADMG is the same as counting in an ordinary ADMG, with all the irrelevant counterfactuals projected out, except that deterministic factors  $P(Z_i | R_j, Z_i^{(1, r_{\text{aff}(i)})})$  can be ignored. This gives us the graph in Fig. III.1(b). The 5 parameters associated with this graph are  $q(Z_1^{(1, r_2=1)} = 0)$ ,  $q(Z_2^{(1, r_1=1)} = 0)$ ,  $q(R_2 = 0)$  and  $q(R_1 = 0 | Z_2^{(1, r_1=1)} = 1)$ ,  $q(R_1 = 0 | Z_2^{(1, r_1=1)} = 0)$ .

Next, we obtain the graph in (c) for the observed law, by projecting out all the counterfactuals in (a). Counting the observed law under full observability in entangled missingness settings has a special consideration: if  $Z_i$  appears in the head of a parameter, and any of the corresponding missingness indicators  $R_{\text{aff}(i)}$  appear in the tail, the kernel must be restricted to cases where  $R_i = 1$  and  $\mathbf{R}_{\text{aff}(i)} = \mathbf{1}$ . This is because, (1) when  $R_i = 0$ , the probability of the head taking any value aside from those where  $Z_i = ?$  is deterministically 0, and (2) cases where  $R_{\text{aff}(i)}$  are set to values



**Figure III.1.** Examples where  $P(\tilde{\mathbf{Z}}^{(r=1)}, \mathbf{R})$  is not identified. The proof is demonstrated using parameter counting. (a)-(c): Extended self-censoring (neighbor-censoring), (a) depicts the data generating process, (b) ADMG for  $P(\tilde{\mathbf{Z}}^{(r=1)}, \mathbf{R})$ , (c) ADMG for observed law. (d)-(f): Extended colluder, (d) depicts the data generating process, (e) ADMG for  $P(\tilde{\mathbf{Z}}^{(r=1)}, \mathbf{R})$ , (f) ADMG for observed law.

different from 1 are irrelevant to the identification of the full observability law<sup>1</sup>. The observed law in (c) can be parameterized using only 4 parameters, one each corresponding to  $q(R_1 = 0)$ ,  $q(R_2 = 0)$ ,  $q(R_1 = 0, Z_2 = 0 | R_2 = 1)$  and  $q(Z_1 = 0 | R_1 = 1, R_2 = 1)$ . That is one less than the full observability law and hence the latter may not be uniquely determined from data.

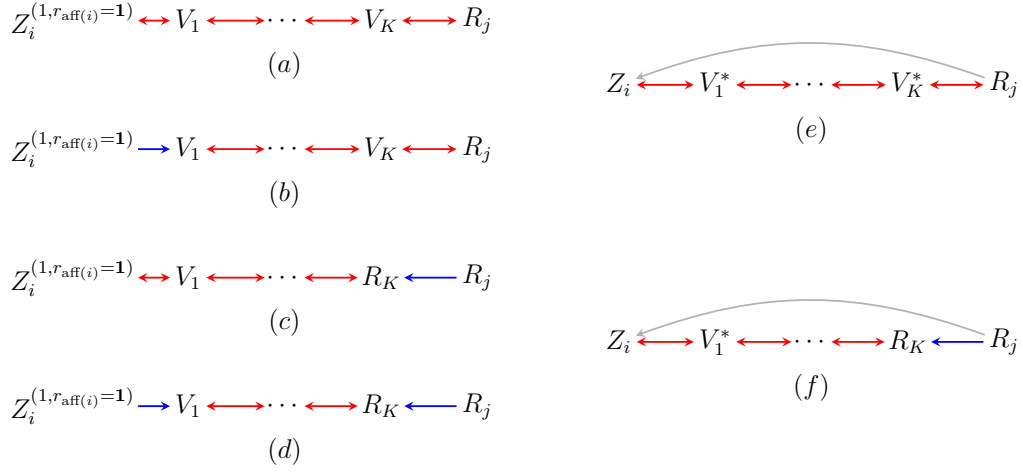
Second, consider the graph in Fig. III.1(d), which has an e-colluder  $Z_2^{(1,r_1=1)} \rightarrow R_3 \leftarrow R_1$ . The full observability law is shown in (e) and observed law in (f). Counting for (e) yields a total of 9 parameters:  $q(Z_3^{(1)} = 0)$ ,  $q(Z_1^{(1,r_2=1)} = 0)$ ,  $q(Z_2^{(1,r_1=1)} = 0)$ ,  $q(R_1 = 0)$ ,  $q(R_2 = 0)$  and finally,  $q(R_3 = 0 | R_1, Z_2^{(1,r_1=1)})$ , which accounts for 4 parameters. The observed law, on the other hand, needs only 8 parameters:  $q(R_1 = 0)$ ,  $q(R_2 = 0)$ ,  $q(R_3 = 0 | R_1 = 1)$ ,  $q(R_3 = 0 | R_1 = 0)$ ,  $q(Z_3 = 0 | R_3 = 1)$ ,  $q(Z_1 = 0 | R_1 = 1, R_2 = 1)$ ,  $q(Z_2 = 0 | R_1 = 1, R_2 = 1)$  and  $q(Z_2 = 0, R_3 = 0 | R_1 = 1, R_2 = 1)$ . Hence, it is not possible to uniquely map back to the full observability law from observed data.

Finally, we present the general argument for arbitrary graphs.

Assume that there are  $n$  variables in  $\tilde{\mathbf{Z}}$ . For simplicity, assume that all counterfactuals are independent of each other, i.e.,  $Z_i^{(1, \mathbf{r}_{\text{aff}(i)} = \mathbf{r})} \perp\!\!\!\perp Z_i^{(1, \mathbf{r}'_{\text{aff}(i)} = \mathbf{r}')}$ , when  $\mathbf{r} \neq \mathbf{r}'$  and  $Z_i^{(1, \mathbf{r}_{\text{aff}(i)} = \mathbf{r})} \perp\!\!\!\perp Z_j^{(1, \mathbf{r}_{\text{aff}(j)} = \mathbf{r}')}$ , when  $\mathbf{r}$  may or may not be equal to  $\mathbf{r}'$ . Graphically, there are no edges between counterfactuals<sup>2</sup>. And by

<sup>1</sup>The consideration that  $R_i = 1$  always holds, but that  $R_{\text{aff}(i)} = 1$  is only for identifying the full observability law. For other cases, considerations vary.

<sup>2</sup>This assumption does not affect the generality of completeness for other types of models because



**Figure III.2.** (a) e-Colluding paths between  $Z_i^{(1, r_{\text{aff}(i)}=1)}$  and  $R_j$  where  $R_j \in \mathbf{R}_{\text{aff}(i)}$  ( $i$  and  $j$  are neighbors) (b) Projecting out  $Z_i^{(1, r_{\text{aff}(i)}=1)}$

Assumption 4,  $\tilde{\mathbf{Z}}^{(r \neq 1)}$  is not a parent of any  $R \in \mathbf{R}$ , and can be projected out (adding no new edges) without affecting the Markov blanket of any variable in  $\tilde{\mathbf{Z}}^{(r=1)}$ ,  $\mathbf{R}$  or proxies  $\mathbf{Z}$ .

Shown in Fig. III.2 (a)-(d) are all possible e-colluding paths between  $Z_i^{(1, r_{\text{aff}(i)}=1)}$  and  $R_j$ . Here,  $R_j \in \mathbf{R}_{\text{aff}(i)}$ . Assume there are  $K$  variables,  $V_1 \dots V_K$ , that lie on the smallest collider path between  $Z_i^{(1, r_{\text{aff}(i)}=1)}$  and  $R_j$ . For this to be the smallest collider path, we require that  $Z_l^{(1, r_{\text{aff}(l)}=1)}$  or  $R_l$  are not in  $V_1, \dots, V_K$  when  $i$  and  $l$  share the same neighborhood, i.e., when  $R_l \in \mathbf{R}_{\text{aff}(i)}$  and  $R_i \in \mathbf{R}_{\text{aff}(l)}$ . If not, we could truncate the path to have the smallest such path that goes between  $Z_i^{(1, r_{\text{aff}(i)}=1)}$  and  $R_j$  such that  $i$  and  $j$  are neighbors. Fig III.2 (e) shows the projection of (a) and (b), and (f) shows the projection of (c) and (d).  $V^* \in \mathbf{Z} \setminus Z_l$ ,  $\mathbf{R} \setminus R_l$ ,  $i$  and  $l$  are neighbors.

We now consider each of these paths (a)-(d) and their corresponding latent projections (e) and (f), as if they appear in a larger graph that is otherwise completely disconnected. Akin to what we did with the examples earlier, we count the number of Möbius parameters (as a function of  $K$ ), and show that the full observability law always has more parameters than the observed law. When we place these colluding paths in a larger graph with arbitrary connectivity, the full observability law is still not identified because of the discrepancy arising from the g-colluding path alone. That is, any edge super graph (super model) is also not identified.

The following fact will be used towards counting parameters in a binary model: Given a bidirected chain  $V_1 \leftrightarrow \dots \leftrightarrow V_{K'}$  of length  $K'$ , the number of parameters required to parameterize this chain is  $\frac{K'(K'+1)}{2}$ , corresponding to parameters given by the following:

$$\begin{aligned}
 & q(V_1 = 0) q(V_1 = V_2 = 0) \dots q(V_1 = \dots V_{K'} = 0) : K' \text{ params} \\
 & q(V_2 = 0) q(V_2 = V_3 = 0) \dots q(V_2 = \dots V_{K'} = 0) : K' - 1 \text{ params} \\
 & \dots \\
 & q(V_{K'} = 0) : 1 \text{ param}
 \end{aligned}$$

#### Parameter counting for Fig. III.2(a), (b), (e)

this model is a submodel of others that involve dependence between counterfactuals, and completeness in this model guarantees completeness in others.

1. Number of Möbius parameters in Fig. III.2(a) is  $\frac{(K+2)(K+3)}{2}$ 
  - It is a bidirected chain  $Z_i^{(1, \mathbf{r}_{\text{aff}(i)}=1)} \leftrightarrow \dots \leftrightarrow R_j$  of length  $K' = K + 2$ .
2. Number of Möbius parameters in Fig. III.2(b) is  $\frac{(K+2)(K+3)}{2}$ 
  - $q(Z_i^{(1, \mathbf{r}_{\text{aff}(i)}=1)} = 0)$  i.e. 1 parameter.
  - A bidirected chain  $V_2 \leftrightarrow \dots \leftrightarrow V_K \leftrightarrow R_j$  of length  $K' = K$ , i.e.  $\frac{K(K+1)}{2}$  parameters.
  - Intrinsic sets involving  $V_1$ , i.e.,  $q(V_1 = 0 | Z_i^{(1, \mathbf{r}_{\text{aff}(i)}=1)})$ ,  $q(V_1 = V_2 = 0 | Z_i^{(1, \mathbf{r}_{\text{aff}(i)}=1)})$ ,  $\dots$ ,  $q(V_1 = V_2 = \dots R_j = 0 | Z_i^{(1, \mathbf{r}_{\text{aff}(i)}=1)})$  which have 2 parameters each, leading to  $2 * (K + 1)$  parameters totally.
3. Number of Möbius parameters in Fig. III.2(e) is  $\frac{(K+2)(K+3)}{2} - 1$ 
  - The bidirected chain  $V_1^* \leftrightarrow \dots \leftrightarrow R_j$  of length  $K' = K + 1$  has  $\frac{(K+1)(K+2)}{2}$  parameters. It should be noted that for any  $Z_l \in \{V_1 \dots V_K\}$ , in the head of a Möbius parameter, if  $R \in R_l \cup \mathbf{R}_{\text{aff}(l)}$  is the parameter's tail,  $R$  is always set to 1 deterministically. Hence, it reduces to counting the simple bidirected chain.
  - The number of intrinsic sets involving  $Z_i$  is  $K + 1$  (and not  $K + 2$ ) since  $R_i$  is not fixable, and the set  $\{Z_i, V_1^*, \dots, V_K^*\}$  is not intrinsic. Each of these intrinsic sets corresponds to two parameters, so  $2 * (K + 1)$  parameters.

#### Parameter counting for Fig. III.2(c), (d), (f)

1. Number of Möbius parameters in Fig. III.2(c) is  $\frac{(K+2)(K+3)}{2}$ 
  - $q(R_j = 0)$  i.e. 1 parameter.
  - A bidirected chain  $Z_i^{(1, \mathbf{r}_{\text{aff}(i)}=1)} \leftrightarrow \dots V_{K-1}$  of length  $K' = K$ , so  $\frac{K(K+1)}{2}$  parameters.
  - Intrinsic sets involving  $R_K$ , i.e.,  $q(R_K = 0 | R_j)$ ,  $q(R_K = V_{K-1} = 0 | R_j)$ ,  $\dots$ ,  $q(R_K = V_{K-1} = \dots Z_i^{(1, \mathbf{r}_{\text{aff}(i)}=1)} = 0 | R_j)$ , i.e.,  $2 * (K + 1)$  parameters.
2. Number of Möbius parameters in Fig. III.2(d) is  $\frac{(K+2)(K+3)}{2}$ 
  - $q(Z_i^{(1, \mathbf{r}_{\text{aff}(i)}=1)} = 0)$  i.e. 1 parameter.
  - $q(R_j = 0)$ , i.e., 1 parameter.
  - A bidirected chain  $V_2 \leftrightarrow \dots \leftrightarrow V_{K-1}$  of length  $K' = K - 2$ , i.e.  $\frac{(K-2)(K-1)}{2}$  parameters.
  - Intrinsic sets involving  $V_1$  and not  $R_K$ , i.e.,  $q(V_1 = 0 | Z_i^{(1, \mathbf{r}_{\text{aff}(i)}=1)})$ ,  $q(V_1 = V_2 = 0 | Z_i^{(1, \mathbf{r}_{\text{aff}(i)}=1)})$ ,  $\dots$ ,  $q(V_1 = V_2 = \dots V_{K-1} = 0 | Z_i^{(1, \mathbf{r}_{\text{aff}(i)}=1)})$  which give 2 parameters each, leading to  $2 * (K - 1)$  parameters.
  - Intrinsic sets involving  $R_K$  and not  $V_1$ , i.e.,  $q(R_K = 0 | Z_i^{(1, \mathbf{r}_{\text{aff}(i)}=1)})$ ,  $q(R_K = V_2 = 0 | Z_i^{(1, \mathbf{r}_{\text{aff}(i)}=1)})$ ,  $\dots$ ,  $q(R_K = V_2 = \dots V_{K-1} = 0 | Z_i^{(1, \mathbf{r}_{\text{aff}(i)}=1)})$  which give 2 parameters each, leading to  $2 * (K - 1)$  parameters.
  - The one intrinsic set involving both  $V_1$  and  $R_K$ , i.e.,  $q(V_1 = V_2 = \dots R_K = 0 | R_j, Z_i^{(1, \mathbf{r}_{\text{aff}(i)}=1)})$  corresponding to 4 parameters.
3. Number of Möbius parameters in Fig. III.2(f) is  $\frac{(K+2)(K+3)}{2} - 1$ 
  - $q(R_j = 0)$ , i.e., 1 parameter.
  - The bidirected chain  $Z_i \leftrightarrow V_1^* \leftrightarrow \dots \leftrightarrow V_{K-1}$  of length  $K' = K$  has  $\frac{K(K+1)}{2}$  parameters. As before, for any  $Z_l \in \{V_1 \dots V_{K-1}\}$ , in the head of a Möbius parameter, if  $R \in R_l \cup \mathbf{R}_{\text{aff}(l)}$  is the parameter's tail,  $R$  is always set to 1 deterministically.

- Intrinsic sets involving  $R_K$ , i.e.,  $q(R_K = 0|R_j)$ ,  $q(R_K, V_{K-1} = 0|R_j), \dots, q(R_K, V_{K-1}, \dots, V_1 = 0|R_j)$  corresponding to  $2 * K$  parameters, and the intrinsic set  $q(R_K, V_{K-1}, \dots, V_1 = 0, Z_i = 0|R_j = 1)$  which only corresponds to 1 parameter (instead of 2) since  $R_j \in \mathbf{R}_{\text{aff}(i)}$  and has to be set to 1.

# Appendix IV

## Generalized Coarsening: Additional Material

### A. Proofs

**Theorem 9** *The influence function for  $\beta$  is given as*

$$U_\beta(Z) = \frac{I(A^1 = a^1)I(A^2 = a^2)}{p(Y^{1,1}, A^1, A^2 \mid C^1, C^2)} p(Y_{t=1}^{1,1} \mid A^1, A^2, C^1, C^2) \times \\ \{Y^{1,2} - \mathbb{E}[Y^{1,2} \mid Y^{1,1}, A^1, A^2, C^1, C^2]\} \\ + \sum_{Y^{1,1}} \mathbb{E}[Y^{1,2} \mid Y^{1,1}, a^1, a^2, C^1, C^2] p(Y_{t=1}^{1,1} \mid a^1, a^2, C^1, C^2) - \beta$$

*Proof.* The parameter of interest  $\beta$  is identified as:

$$\sum_{Y^{1,1}, C^1, C^2} \mathbb{E}[Y^{1,2} \mid Y^{1,1}, A^1, A^2, C^1, C^2] \pi(Y_{t=1}^{1,1}, A^1, A^2, C^1, C^2) p(C^1, C^2)$$

The influence function  $U_\beta(Z)$  is derived using the integral equation as

$$\left. \frac{\partial \beta}{\partial \theta} \right|_{\theta=\theta_0} = \mathbb{E}[U_\beta(Z) S_\theta(Z; \theta_0)]$$

$$\left. \frac{\partial \beta}{\partial \theta} \right|_{\theta=\theta_0} = \sum_{Y^{1,2}, Y^{1,1}, C^1, C^2} Y^{1,2} \frac{\partial p(Y^{1,2} \mid Y^{1,1}, A^1, A^2, C^1, C^2)}{\partial \theta} p(Y_{t=1}^{1,1}, A^1, A^2, C^1, C^2) p(C^1, C^2) \\ + \sum_{Y^{1,1}, C^1, C^2} \mathbb{E}[Y^{1,2} \mid Y^{1,1}, A^1, A^2, C^1, C^2] p(Y_{t=1}^{1,1}, A^1, A^2, C^1, C^2) \frac{\partial p(C^1, C^2)}{\partial \theta}$$

From here, each term is modified until it looks like the integral equation and the influence function is obtained by visual inspection. Starting with the second term in the above equation

$$\sum_{Y^{1,1}, C^1, C^2} \mathbb{E}[Y^{1,2} \mid Y^{1,1}, A^1, A^2, C^1, C^2] p(Y_{t=1}^{1,1}, A^1, A^2, C^1, C^2) \frac{\partial p(C^1, C^2)}{\partial \theta}$$

Since  $S(C^1, C^2; \theta) = \frac{\partial \log p(C^1, C^2)}{\partial \theta} = \frac{1}{p(C^1, C^2)} \frac{\partial p(C^1, C^2)}{\partial \theta}$ ,  $\frac{\partial p(C^1, C^2)}{\partial \theta} = S_\theta(C^1, C^2; \theta_0) p(C^1, C^2)$ , the above can be rewritten as

$$= \sum_{Y^{1,1}} \mathbb{E}_{C^1, C^2} \left[ \mathbb{E}[Y^{1,2} | Y^{1,1}, A^1, A^2, C^1, C^2] p(Y_{t=1}^{1,1}, A^1, A^2, C^1, C^2) S_\theta(C^1, C^2; \theta_0) \right]$$

From the properties of the score, we have

$$\mathbb{E}_{C^1, C^2} \left[ \mathbb{E}_{C^1, C^2} \left[ \sum_{Y^{1,1}} \mathbb{E}[Y^{1,2} | Y^{1,1}, A^1, A^2, C^1, C^2] p(Y_{t=1}^{1,1}, A^1, A^2, C^1, C^2) \right] S_\theta(C^1, C^2; \theta_0) \right] = 0$$

Denoting  $\sum_{Y^{1,1}} \mathbb{E}[Y^{1,2} | Y^{1,1}, A^1, A^2, C^1, C^2] p(Y_{t=1}^{1,1}, A^1, A^2, C^1, C^2) = q_{C^1, C^2}$ , the above equation can be rewritten as

$$\mathbb{E}_{C^1, C^2} [(q_{C^1, C^2} - \mathbb{E}[q_{C^1, C^2}]) S_\theta(C^1, C^2; \theta_0)]$$

Since  $\mathbb{E}[S(Y^{1,1}, Y^{1,2}, A^1, A^2 | C^1, C^2)] = 0$ , the above equation can be rewritten as

$$= \mathbb{E} [\mathbb{E} [\{q_{C^1, C^2} - \mathbb{E}[q_{C^1, C^2}]\} \{S_\theta(C^1, C^2; \theta_0) + S_\theta(Y^{1,1}, Y^{1,2}, A^1, A^2 | C^1, C^2; \theta_0)\} | C^1, C^2]]$$

Since  $\mathbb{E}_{C^1, C^2} [q_{C^1, C^2}] = \psi$ , the contribution of the second term to the influence function can be written as:

$$U_2(\psi) = \sum_{Y^{1,1}} \mathbb{E}[Y^{1,2} | Y^{1,1}, A^1, A^2, C^1, C^2] p(Y_{t=1}^{1,1}, A^1, A^2, C^1, C^2) - \psi$$

Similarly, the first term, is rewritten using the score function as:

$$\begin{aligned} &= \sum_{Y^{1,2}, Y^{1,1}, C^1, C^2} Y^{1,2} S_\theta(Y^{1,2} | Y^{1,1}, A^1, A^2, C^1, C^2; \theta_0) \\ &\times p(Y^{1,2} | Y^{1,1}, A^1, A^2, C^1, C^2) p(Y_{t=1}^{1,1}, A^1, A^2, C^1, C^2) p(C^1, C^2) \end{aligned}$$

Introducing two indicator functions  $\mathbb{I}(A^1 = a^1) \mathbb{I}(A^2 = a^2)$ , and summing over  $A^1, A^2$ , this term can be rewritten as

$$= \mathbb{E} \left[ \frac{I(A^1 = a^1) I(A^2 = a^2)}{p(Y^{1,1}, A^1, A^2, C^1, C^2)} Y^{1,2} p(Y_{t=1}^{1,1}, A^1, A^2, C^1, C^2) S_\theta(Y^{1,2} | Y^{1,1}, A^1, A^2, C^1, C^2; \theta_0) \right]$$

Next, from iterated expectations and the properties of score functions, note

$$\begin{aligned} &\mathbb{E} \left[ \frac{I(A^1 = a^1) I(A^2 = a^2)}{p(Y^{1,1}, A^1, A^2, C^1, C^2)} p(Y_{t=1}^{1,1}, A^1, A^2, C^1, C^2) \mathbb{E}[Y^{1,2} | Y^{1,1}, A^1, A^2, C^1, C^2] \right. \\ &\quad \left. \times S_\theta(Y^{1,2} | Y^{1,1}, A^1, A^2, C^1, C^2; \theta_0) \right] = 0 \end{aligned}$$

So, we can rewrite the partial derivative of the first term as

$$= \mathbb{E} \left[ \frac{I(A^1 = a^1) I(A^2 = a^2)}{p(Y^{1,1}, A^1, A^2, C^1, C^2)} p(Y_{t=1}^{1,1}, A^1, A^2, C^1, C^2) \{Y^{1,2} - \mathbb{E}[Y^{1,2} | Y^{1,1}, A^1, A^2, C^1, C^2]\} \right]$$

$$\times S_{\theta}(Y^{1,2} \mid Y^{1,1}, A^1, A^2, C^1, C^2; \theta_0) \Big]$$

Using the property of score functions, we can add in the missing score piece to get:

$$= \mathbb{E} \left[ \frac{I(A^1 = a^1)I(A^2 = a^2)}{p(Y^{1,1}, A^1, A^2 \mid C^1, C^2)} p(Y_{t=1}^{1,1}, A^1, A^2, C^1, C^2) \{Y^{1,2} - \mathbb{E}[Y^{1,2} \mid Y^{1,1}, A^1, A^2, C^1, C^2]\} \right. \\ \left. S_{\theta}(Y^{1,2}, Y^{1,1}, A^1, A^2, C^1, C^2; \theta_0) \right]$$

So, the influence function is

$$U_{\beta}(Z) = \frac{I(A^1 = a^1)I(A^2 = a^2)}{p(Y^{1,1}, A^1, A^2 \mid C^1, C^2)} p(Y_{t=1}^{1,1}, A^1, A^2, C^1, C^2) \{Y^{1,2} - \mathbb{E}[Y^{1,2} \mid Y^{1,1}, A^1, A^2, C^1, C^2]\} \\ + \sum_{Y^{1,1}} \mathbb{E}[Y^{1,2} \mid Y^{1,1}, A^1, A^2, C^1, C^2] p(Y_{t=1}^{1,1}, A^1, A^2, C^1, C^2) - \beta$$

□

**Theorem 10** *The estimator obtained by solving the influence function for  $\beta$  is doubly robust as long as one of  $p(Y^{1,1}, A^1, A^2 \mid C^1, C^2)$  or  $\mathbb{E}[Y^{1,2} \mid Y^{1,1}, A^1, A^2, C^1, C^2]$  is specified correctly.*

*Proof.* The influence function provides an unbiased and consistent estimator of  $\beta$  as long as at least one of  $p(Y_1, A^1, A^2 \mid C^1, C^2)$  or  $\mathbb{E}[Y_2 \mid Y_1, A^1, A^2, C^1, C^2]$  is specified correctly. To prove this, we first misspecify each of the nuisance models individually and show the influence function is still mean zero.

First, we incorrectly specify  $p(Y^{1,1}, A^1, A^2 \mid C^1, C^2)$  is mis-specified as  $p^*(Y^{1,1}, A^1, A^2 \mid C^1, C^2)$ , but correctly specify  $\mathbb{E}[Y^{1,2} \mid Y^{1,1}, A^1, A^2, C^1, C^2]$ . The expectation of the influence function can be written as:

$$\mathbb{E}[U_{\beta}(Z)] = \mathbb{E} \left[ \frac{I(A^1 = a^1)I(A^2 = a^2)}{p^*(Y^{1,1}, A^1, A^2 \mid C^1, C^2)} p(Y_{t=1}^{1,1}, A^1, A^2, C^1, C^2) \{Y^{1,2} - \mathbb{E}[Y^{1,2} \mid Y^{1,1}, A^1, A^2, C^1, C^2]\} \right] \\ + \mathbb{E} \left[ \sum_{Y^{1,1}} \mathbb{E}[Y^{1,2} \mid Y^{1,1}, A^1, A^2, C^1, C^2] p(Y_{t=1}^{1,1}, A^1, A^2, C^1, C^2) \right]$$

Since  $\mathbb{E}[Y^{1,2} \mid Y^{1,1}, A^1, A^2, C^1, C^2]$  is specified correctly,  $\sum_{Y^{1,1}} \mathbb{E}[\mathbb{E}[Y^{1,2} \mid Y^{1,1}, A^1, A^2, C^1, C^2] p(Y_{t=1}^{1,1}, A^1, A^2, C^1, C^2)] = \beta$ . To prove the expectation of the influence function is mean 0, it remains to show is that the first term is mean 0. This can be seen by

$$\mathbb{E} \left[ \frac{I(A^1 = a^1)I(A^2 = a^2)}{p^*(Y^{1,1}, A^1, A^2 \mid C^1, C^2)} p(Y_{t=1}^{1,1}, A^1, A^2, C^1, C^2) \{Y^{1,2} - \mathbb{E}[Y^{1,2} \mid Y^{1,1}, A^1, A^2, C^1, C^2]\} \right] \\ = \mathbb{E} \left[ \mathbb{E} \left[ \frac{I(A^1 = a^1)I(A^2 = a^2)}{p^*(Y^{1,1}, A^1, A^2 \mid C^1, C^2)} p(Y_{t=1}^{1,1}, A^1, A^2, C^1, C^2) \right. \right. \\ \left. \left. \times \{Y^{1,2} - \mathbb{E}[Y^{1,2} \mid Y^{1,1}, A^1, A^2, C^1, C^2]\} \mid Y^{1,1}, A^1, A^2, C^1, C^2 \right] \right] \\ = \mathbb{E} \left[ \mathbb{E} \left[ \frac{I(A^1 = a^1)I(A^2 = a^2)}{p^*(Y^{1,1}, A^1, A^2 \mid C^1, C^2)} p(Y_{t=1}^{1,1}, A^1, A^2, C^1, C^2) Y^{1,2} \mid Y^{1,1}, A^1, A^2, C^1, C^2 \right] \right] \\ - \mathbb{E} \left[ \frac{I(A^1 = a^1)I(A^2 = a^2)}{p^*(Y^{1,1}, A^1, A^2 \mid C^1, C^2)} p(Y_{t=1}^{1,1}, A^1, A^2, C^1, C^2) \mathbb{E}[Y^{1,2} \mid Y^{1,1}, A^1, A^2, C^1, C^2] \right]$$

$$= 0$$

Similar steps can be followed when  $\mathbb{E}[Y^{1,2} \mid Y^{1,1}, A^1, A^2, C^1, C^2]$  is incorrectly specified as  $\mathbb{E}^*[Y^{1,2} \mid Y^{1,1}, A^1, A^2, C^1, C^2]$ . Grouping the influence function into two terms and examining:

$$\begin{aligned} \mathbb{E}[U_\beta(Z)] &= \mathbb{E} \left[ \frac{I(A^1 = A^1)I(A^2 = A^2)}{p(Y_1, A^1, A^2 \mid C^1, C^2)} p(Y_{1,t=1}, A^1, A^2, C^1, C^2) Y_2 \right] \\ &\quad - \mathbb{E} \left[ \frac{I(A^1 = A^1)I(A^2 = A^2)}{p(Y_1, A^1, A^2 \mid C^1, C^2)} p(Y_{1,t=1}, A^1, A^2, C^1, C^2) \mathbb{E}^*[Y_2 \mid Y_1, A^1, A^2, C^1, C^2] \right] \\ &\quad + \mathbb{E} \left[ \sum_{Y^{1,1}} \mathbb{E}^*[Y^{1,2} \mid Y_1, A^1, A^2, C^1, C^2] \pi(Y_{1,t=1}, A^1, A^2, C^1, C^2) \right] - \beta \end{aligned}$$

As long as the propensity score is specified correctly,  $\mathbb{E} \left[ \frac{I(A^1 = a^1)I(A^2 = a^2)}{p(Y^{1,1}, A^1, A^2 \mid C^1, C^2)} \pi(Y^{1,1}, A^1, A^2, C^1, C^2) Y^{1,2} \right]$  will evaluate to the target parameter. And the remaining two expectation terms will cancel out, and the mean of the influence function will be zero. Consequently, the double robustness of this influence function is proved.  $\square$

**Theorem 11** *The influence function for  $\psi$  is given as*

$$\begin{aligned} U(\psi) &= \left\{ \sum_{Y^{1,1}} \frac{I(A^1 = a^1)I(A^2 = a^2)\pi p(Y^{1,2} \mid Y^{1,1}, A^1, A^2, C^1, C^2)}{p(Y^{1,2}, A^1, A^2 \mid C^1, C^2)} \right\} \times \\ &\quad \{Y^{1,1} - \mathbb{E}[Y^{1,1} \mid Y^{1,2}, A^1, A^2, C^1, C^2]\} \\ &\quad + \left\{ \frac{I(A^1 = a^1)I(A^2 = a^2)\pi}{p(Y^{1,1}, A^1, A^2 \mid C^1, C^2)} \right\} \times \\ &\quad \{q_{a^1, a^2}(Y^{1,1} \mid Y^{1,2}) - \mathbb{E}[\phi_{a^1, a^2}(Y^{1,1} \mid Y^{1,2}) \mid Y^{1,1}, A^1, A^2, C^1, C^2]\} \\ &\quad + \sum_{Y^{1,1}, Y^{1,2}} \mathbb{E}[Y^{1,1} \mid Y^{1,2}, a^1, a^2, C^1, C^2] p(Y^{1,2} \mid Y^{1,1}, a^1, a^2, C^1, C^2) \pi - \psi \end{aligned}$$

*Proof.* Following a similar procedure as Theorem 9 and taking the path wise derivative of the identifying functional for  $\psi$  we get

$$\begin{aligned} \frac{\partial \psi}{\partial \theta} \Big|_{\theta=\theta_0} &= \\ &\sum_{Y^{1,1}, Y^{1,2}, C^1, C^2} \left( \sum_{Y^{1,1}} \frac{\partial p(Y^{1,1} \mid Y^{1,2}, A^1, A^2, C^1, C^2)}{\partial \theta} \right) \times p(Y^{1,2} \mid Y^{1,1}, A^1, A^2, C^1, C^2) \times \\ &\quad p(Y_{t=1}^{1,1}, A^1, A^2, C^1, C^2) p(C^1, C^2) \\ &+ \sum_{Y^{1,1}, Y^{1,2}, C^1, C^2} \mathbb{E}[Y^{1,1} \mid Y^{1,2}, A^1, A^2, C^1, C^2] \times \frac{\partial p(Y^{1,2} \mid Y^{1,1}, A^1, A^2, C^1, C^2)}{\partial \theta} \times \\ &\quad p(Y_{t=1}^{1,1}, A^1, A^2, C^1, C^2) p(C^1, C^2) \\ &+ \sum_{Y^{1,1}, Y^{1,2}, C^1, C^2} \mathbb{E}[Y^{1,1} \mid Y^{1,2}, A^1, A^2, C^1, C^2] p(Y^{1,2} \mid Y^{1,1}, A^1, A^2, C^1, C^2) \times \frac{\partial p(C^1, C^2)}{\partial \theta} \times \\ &\quad p(Y_{t=1}^{1,1}, A^1, A^2, C^1, C^2) \end{aligned}$$

Each term will be examined individually and put into the form from the integral equation. Starting with the third term, we rewrite it as:

$$\begin{aligned}
&= \sum_{Y^{1,1}, Y^{1,2}, C^1, C^2} \mathbb{E}[Y^{1,1} \mid Y^{1,2}, A^1, A^2, C^1, C^2] p(Y^{1,2} \mid Y^{1,1}, A^1, A^2, C^1, C^2) \times \\
&\quad \pi(Y^{1,1}, A^1, A^2, C^1, C^2) S(C^1, C^2) p(C^1, C^2) \\
&= \mathbb{E}_{C^1, C^2} \left[ \sum_{Y^{1,1}, Y^{1,2}} \mathbb{E}[Y^{1,1} \mid Y^{1,2}, A^1, A^2, C^1, C^2] p(Y^{1,2} \mid Y^{1,1}, A^1, A^2, C^1, C^2) \times \right. \\
&\quad \left. \pi(Y^{1,1}, A^1, A^2, C^1, C^2) S(C^1, C^2) \right]
\end{aligned}$$

Denote  $\sum_{Y^{1,1}, Y^{1,2}} \mathbb{E}[Y^{1,1} \mid Y^{1,2}, A^1, A^2, C^1, C^2] p(Y^{1,2} \mid Y^{1,1}, A^1, A^2, C^1, C^2) \pi(Y^{1,1}, A^1, A^2, C^1, C^2) = q(C^1, C^2)$ . Then

$$\mathbb{E}_{C^1, C^2} [\mathbb{E}_{C^1, C^2} [q(C^1, C^2)] S(C^1, C^2)] = 0$$

Noting  $\mathbb{E}_{C^1, C^2} [q(C^1, C^2)] = \psi$ , the third term can be rewritten as:

$$\begin{aligned}
&= \mathbb{E}_{C^1, C^2} [(q(C^1, C^2) - \psi) S(C^1, C^2)] \\
&= \mathbb{E}_{C^1, C^2} [\mathbb{E}_{Y^{1,1}, Y^{1,2}, A^1, A^2 \mid C^1, C^2} [(q(C^1, C^2) - \psi) S(C^1, C^2)]] \\
&= \mathbb{E}_{C^1, C^2} [\mathbb{E}_{Y^{1,1}, Y^{1,2}, A^1, A^2 \mid C^1, C^2} [(q(C^1, C^2) - \psi) S(Y^{1,1}, Y^{1,2}, A^1, A^2, C^1, C^2)]] \\
&= \mathbb{E}_{Y^{1,1}, Y^{1,2}, A^1, A^2, C^1, C^2} [(q(C^1, C^2) - \psi) S(Y^{1,1}, Y^{1,2}, A^1, A^2, C^1, C^2)]
\end{aligned}$$

Therefore, the contribution of the third term to the influence function is:

$$U_3(\psi) = \sum_{Y^{1,1}, Y^{1,2}} \mathbb{E}[Y^{1,1} \mid Y^{1,2}, A^1, A^2, C^1, C^2] p(Y^{1,2} \mid Y^{1,1}, A^1, A^2, C^1, C^2) \pi(Y^{1,1}, A^1, A^2, C^1, C^2) - \psi$$

Now, examining the second term and introducing the score function:

$$= \mathbb{E} \left\{ \frac{\mathbb{I}(A^1 = a^1) \mathbb{I}(A^2 = a^2) \pi(Y^{1,1}, A^1, A^2, C^1, C^2)}{p(Y^{1,1}, A^1, A^2 \mid C^1, C^2)} \right\} \mathbb{E}[Y^{1,1} \mid Y^{1,2}, A^1, A^2, C^1, C^2] \times \\
S(Y^{1,2} \mid Y^{1,1}, A^1, A^2, C^1, C^2)$$

Now, denoting  $\left\{ \frac{\mathbb{I}(A^1 = a^1) \mathbb{I}(A^2 = a^2) \pi(Y_{i=1}^{1,1}, a^1, a^2, C^1, C^2)}{p(Y^{1,1}, A^1, A^2 \mid C^1, C^2)} \right\} \mathbb{E}[Y^{1,1} \mid Y^{1,2}, A^1, A^2, C^1, C^2]$  as  $q(Y^{1,1}, Y^{1,2}, A^1, A^2, C^2, C^1)$ , and noting

$$\mathbb{E} [\mathbb{E} [q(Y^{1,1}, Y^{1,2}, a^1, a^2, C^1, C^2) \mid Y^{1,1}, a^1, a^2, C^1, C^2] S(Y^{1,2} \mid Y^{1,1}, A^1, A^2, C^1, C^2)] = 0$$

Using the above, the second term is rewritten as:

$$\begin{aligned}
&= \mathbb{E} [(q(Y^{1,1}, Y^{1,2}, A^1, A^2, C^1, C^2) \\
&\quad - \mathbb{E}[q(Y^{1,1}, Y^{1,2}, A^1, A^2, C^1, C^2) \mid Y^{1,1}, A^1, A^2, C^1, C^2]) S(Y^{1,2} \mid Y^{1,1}, A^1, A^2, C^1, C^2)]
\end{aligned}$$

From the properties of score function,

$$\begin{aligned}
&\mathbb{E} [(q(Y^{1,1}, Y^{1,2}, A^1, A^2, C^1, C^2) \\
&\quad - \mathbb{E}[q(Y^{1,1}, Y^{1,2}, A^1, A^2, C^1, C^2) \mid Y^{1,1}, A^1, A^2, C^1, C^2]) S(Y^{1,1}, A^1, A^2, C^1, C^2)] = 0
\end{aligned}$$

And so we get

$$= \mathbb{E}[(q(Y^{1,1}, Y^{1,2}, A^1, A^2, C^1, C^2) - \mathbb{E}[q(Y^{1,1}, Y^{1,2}, A^1, A^2, C^1, C^2) | Y^{1,1}, A^1, A^2, C^1, C^2])S(Y^{1,2}, Y^{1,1}, A^1, A^2, C^1, C^2)]$$

The contribution of the second term to the influence function can be written as

$$U_2(\psi) = q(Y^{1,1}, Y^{1,2}, A^1, A^2, C^1, C^2) - \mathbb{E}[q(Y^{1,1}, Y^{1,2}, A^1, A^2, C^1, C^2) | Y^{1,1}, A^1, A^2, C^1, C^2]$$

Now, repeating the procedure for the first term:

$$\begin{aligned} &= \sum_{Y^{1,1}, Y^{1,2}, C^1, C^2} \left( \sum_{Y^{1,1}} Y^{1,1} p(Y^{1,1} | Y^{1,2}, A^1, A^2, C^1, C^2) S(Y^{1,1} | Y^{1,2}, A^1, A^2, C^1, C^2) \right) \\ & p(Y^{1,2} | Y^{1,1}, A^1, A^2, C^1, C^2) \pi(Y^{1,1}, A^1, A^2, C^1, C^2) p(C^1, C^2) \\ &= \sum_{Y^{1,2}, C^1, C^2} \mathbb{E}[Y^{1,1} S(Y^{1,1} | Y^{1,2}, A^1, A^2, C^1, C^2) | Y^{1,2}, A^1, A^2, C^1, C^2] \\ & \times \left\{ \sum_{Y^{1,1}} p(Y^{1,2} | Y^{1,1}, A^1, A^2, C^1, C^2) \pi(Y^{1,1}, A^1, A^2, C^1, C^2) p(C^1, C^2) \right\} \\ &= \sum_{Y^{1,2}, A^1, A^2, C^1, C^2} \mathbb{I}(A^1 = a^1) \mathbb{I}(A^2 = a^2) \mathbb{E}[Y^{1,1} S(Y^{1,1} | Y^{1,2}, A^1, A^2, C^1, C^2) | Y^{1,2}, A^1, A^2, C^1, C^2] \\ & \left\{ \sum_{Y^{1,1}} p(Y^{1,2} | Y^{1,1}, A^1, A^2, C^1, C^2) \pi(Y^{1,1}, A^1, A^2, C^1, C^2) p(C^1, C^2) \right\} \end{aligned}$$

Multiplying and dividing by  $p(Y^{1,2}, A^1, A^2, C^1, C^2)$ , shifting summations around to get a full expectation:

$$\begin{aligned} &= \left\{ \sum_{Y^{1,1}} \mathbb{E}_{Y^{1,2}, A^1, A^2, C^1, C^2} \left[ \frac{I(A^1 = a^1) I(A^2 = a^2) \pi(Y^{1,1}, A^1, A^2, C^1, C^2) p(Y^{1,2} | Y^{1,1}, A^1, A^2, C^1, C^2)}{p(Y^{1,2}, A^1, A^2, C^1, C^2)} \right] \right\} \\ & \mathbb{E}[Y^{1,1} S(Y^{1,1} | Y^{1,2}, A^1, A^2, C^1, C^2) | Y^{1,2}, A^1, A^2, C^1, C^2] \\ &= \mathbb{E} \left[ \frac{I(A^1 = a^1) I(A^2 = a^2) \{ \sum_{Y^{1,1}} \pi(Y^{1,1}, A^1, A^2, C^1, C^2) p(Y^{1,2} | Y^{1,1}, A^1, A^2, C^1, C^2) \}}{p(Y^{1,2}, A^1, A^2 | C^1, C^2)} Y^{1,1} \times \right. \\ & \quad \left. S(Y^{1,1} | Y^{1,2}, A^1, A^2, C^1, C^2) \right] \end{aligned}$$

So, the first term of the influence function can be written as

$$\begin{aligned} U_1(\psi) &= \frac{I(A^1 = a^1) I(A^2 = a^2) \{ \sum_{Y^{1,1}} \pi(Y^{1,1}, A^1, A^2, C^1, C^2) p(Y^{1,2} | Y^{1,1}, A^1, A^2, C^1, C^2) \}}{p(Y^{1,2}, A^1, A^2 | C^1, C^2)} \\ & \left\{ Y^{1,1} - \mathbb{E} \left[ \frac{I(A^1 = a^1) I(A^2 = a^2)}{p(Y^{1,2}, A^1, A^2 | C^1, C^2)} \right] \right. \\ & \quad \left. \left\{ \sum_{Y^{1,1}} \pi(Y^{1,1}, A^1, A^2, C^1, C^2) p(Y^{1,2} | Y^{1,1}, A^1, A^2, C^1, C^2) \right\} Y^{1,1} | Y^{1,2}, A^1, A^2, C^1, C^2 \right\} \right\} \end{aligned}$$

Putting all three of these pieces together gives:

$$U(\psi) = \sum_{Y'_1} \frac{I(A^1 = a^1) I(A^2 = a^2) \pi(Y'_1, A^1, A^2, C^1, C^2) p(Y^{1,2} | Y'_1, A^1, A^2, C^1, C^2)}{p(Y^{1,2}, A^1, A^2 | C^1, C^2)} Y^{1,1}$$

$$\begin{aligned}
& - \sum_{Y'_1} \frac{I(A^1 = a^1)I(A^2 = a^2)\pi(Y'_1, A^1, A^2, C^1, C^2)p(Y^{1,2} | Y'_1, A^1, A^2, C^1, C^2)}{p(Y^{1,2}, A^1, A^2 | C^1, C^2)} \times \\
& \quad \mathbb{E}[Y^{1,1} | Y^{1,2}, A^1, A^2, C^1, C^2] \\
& + \left\{ \frac{I(A^1 = a^1)I(A^2 = a^2)\pi(Y^{1,1}, A^1, A^2, C^1, C^2)}{p(Y^{1,1}, A^1, A^2 | C^1, C^2)} \right\} \mathbb{E}[Y^{1,1} | Y^{1,2}, A^1, A^2, C^1, C^2] \\
& - \left\{ \frac{I(A^1 = a^1)I(A^2 = a^2)\pi(Y^{1,1}, A^1, A^2, C^1, C^2)}{p(Y^{1,1}, A^1, A^2 | C^1, C^2)} \right\} \times \\
& \quad \mathbb{E}[\mathbb{E}[Y^{1,1} | Y^{1,2}, A^1, A^2, C^1, C^2] | Y^{1,1}, A^1, A^2, C^1, C^2] \\
& + \sum_{Y^{1,1}, Y^{1,2}} \mathbb{E}[Y^{1,1} | Y^{1,2}, A^1, A^2, C^1, C^2] p(Y^{1,2} | Y^{1,1}, A^1, A^2, C^1, C^2) \pi(Y^{1,1}, A^1, A^2, C^1, C^2) \\
& - \psi
\end{aligned}$$

□

**Theorem 12** *The influence function based estimator for  $\psi$  above exhibits 2 out of 4 robustness, where as long as the following pairs of models are specified correctly, we obtain a consistent and unbiased estimator:*

$$\begin{aligned}
& (\mathbb{E}[Y^{1,1} | Y^{1,2}, a^1, a^2, C^1, C^2], \quad p(Y^{1,2} | Y^{1,1}, a^1, a^2, C^1, C^2)) \\
& (p(Y^{1,2}, A^1, A^2 | C^1, C^2), \quad p(Y^{1,2} | Y^{1,1}, A^1, A^2, C^1, C^2)) \\
& (p(Y^{1,1}, A^1, A^2 | C^1, C^2), \quad \mathbb{E}[Y^{1,1} | Y^{1,2}, A^1, A^2, C^1, C^2])
\end{aligned}$$

*Proof.* The statement of the theorem says that as long as any of the pairs of models are correctly specified, we will obtain a consistent estimator for  $\psi$

$$\mathbb{E}[Y^{1,1} | Y^{1,2}, A^1, A^2, C^1, C^2] \& p(Y^{1,2} | Y^{1,1}, A^1, A^2, C^1, C^2) \quad (\text{IV.1})$$

$$p(Y^{1,2}, A^1, A^2 | C^1, C^2) \& p(Y^{1,2} | Y^{1,1}, A^1, A^2, C^1, C^2) \quad (\text{IV.2})$$

$$p(Y^{1,1}, A^1, A^2 | C^1, C^2) \& \mathbb{E}[Y^{1,1} | Y^{1,2}, A^1, A^2, C^1, C^2] \quad (\text{IV.3})$$

Starting with the proof assuming only  $p(Y^{1,2}, A^1, A^2 | C^1, C^2) \& p(Y^{1,2} | Y^{1,1}, A^1, A^2, C^1, C^2)$  are specified correctly, and denote the rest of the models that are misspecified as  $p^*, \mathbb{E}^*$ . The first term of the influence function can then be written as:

$$\begin{aligned}
& \sum_{Y^{1,1}, Y^{1,2}, A^1, A^2, C^1, C^2} \left\{ \sum_{Y_{t=1}^{1,1}} \frac{I(A^1 = a^1)I(A^2 = a^2)\pi(Y_{t=1}^{1,1}, A^1, A^2, C^1, C^2)p(Y^{1,2} | Y_{t=1}^{1,1}, A^1, A^2, C^1, C^2)}{p(Y^{1,2}, A^1, A^2 | C^1, C^2)} Y_{t=1}^{1,1} \right\} \\
& \quad p(Y^{1,1}, Y^{1,2}, A^1, A^2, C^1, C^2) \\
& = \sum_{Y^{1,1}, Y^{1,2}, A^1, A^2, C^1, C^2} \left\{ \sum_{Y_{t=1}^{1,1}} \frac{\pi(Y_{t=1}^{1,1}, A^1, A^2, C^1, C^2)p(Y^{1,2} | Y_{t=1}^{1,1}, A^1, A^2, C^1, C^2)}{p(Y^{1,2}, A^1, A^2 | C^1, C^2)} Y_{t=1}^{1,1} \right\} \times \\
& \quad p(Y^{1,1}, Y^{1,2}, A^1, A^2, C^1, C^2) \\
& = \sum_{Y_{t=1}^{1,1}, Y^{1,2}, A^1, A^2, C^1, C^2} \left\{ \sum_{Y_{t=1}^{1,1}} \frac{\pi(Y_{t=1}^{1,1}, A^1, A^2, C^1, C^2)p(Y^{1,2} | Y_{t=1}^{1,1}, A^1, A^2, C^1, C^2)}{p(Y^{1,2}, A^1, A^2 | C^1, C^2)} Y_{t=1}^{1,1} \times \right. \\
& \quad \left. p(Y^{1,1} | Y^{1,2}, A^1, A^2, C^1, C^2) \right\} \times p(Y^{1,2}, A^1, A^2, C^1, C^2)
\end{aligned}$$

$$\begin{aligned}
&= \sum_{Y_{t=1}^{1,1}, Y^{1,2}, C^1, C^2} \left\{ \frac{\pi(Y_{t=1}^{1,1}, A^1, A^2, C^1, C^2) p(Y^{1,2} | Y_{t=1}^{1,1}, A^1, A^2, C^1, C^2)}{p(Y^{1,2}, A^1, A^2 | C^1, C^2)} \mathbb{E}[Y^{1,1} | Y^{1,2}, A^1, A^2, C^1, C^2] \right\} \\
&\quad p(Y^{1,2}, A^1, A^2, C^1, C^2) \\
&= \sum_{Y_{t=1}^{1,1}, Y^{1,2}, C^1, C^2} \left\{ \pi(Y_{t=1}^{1,1}, A^1, A^2, C^1, C^2) p(Y^{1,2} | Y_{t=1}^{1,1}, A^1, A^2, C^1, C^2) \times \right. \\
&\quad \left. \mathbb{E}[Y^{1,1} | Y^{1,2}, A^1, A^2, C^1, C^2] \right\} \times p(C^1, C^2) = \psi
\end{aligned}$$

The rest of the terms are mean 0 by the properties of iterated expectations.

Next, we consider the case where only  $\mathbb{E}[Y^{1,1} | Y^{1,2}, A^1, A^2, C^1, C^2]$  and  $p(Y^{1,2} | Y^{1,1}, A^1, A^2, C^1, C^2)$  are specified correctly, and denote the rest of the misspecified models with an \*. Now, consider the following term:

$$\begin{aligned}
&\mathbb{E} \left[ \sum_{Y^{1,1}, Y^{1,2}} \mathbb{E}[Y^{1,1} | Y^{1,2}, A^1, A^2, C^1, C^2] p(Y^{1,2} | Y^{1,1}, A^1, A^2, C^1, C^2) \pi(Y_{t=1}^{1,1}, A^1, A^2, C^1, C^2) \right] \\
&= \sum_{Y^{1,1}, Y^{1,2}, C^1, C^2} \mathbb{E}[Y^{1,1} | Y^{1,2}, A^1, A^2, C^1, C^2] p(Y^{1,2} | Y^{1,1}, A^1, A^2, C^1, C^2) \times \\
&\quad \pi(Y^{1,1}, A^1, A^2, C^1, C^2) p(C^1, C^2) = \psi
\end{aligned}$$

The rest of the terms are mean 0, which can be seen by applying iterated expectation once.

Finally, Consider the case where  $p(Y^{1,1}, A^1, A^2 | C^1, C^2)$  and  $\mathbb{E}[Y^{1,1} | Y^{1,2}, A^1, A^2, C^1, C^2]$  are correctly specified. Examining the term:

$$\begin{aligned}
&\mathbb{E} \left[ \left\{ \frac{I(A^1 = a^1) I(A^2 = a^2) \pi(Y_{t=1}^{1,1}, A^1, A^2, C^1, C^2)}{p(Y^{1,1}, A^1, A^2 | C^1, C^2)} \right\} \mathbb{E}[Y^{1,1} | Y^{1,2}, A^1, A^2, C^1, C^2] \right] \\
&= \sum_{Y^{1,1}, Y^{1,2}, A^1, A^2, C^1, C^2} \left\{ \frac{I(A^1 = a^1) I(A^2 = a^2) \pi(Y^{1,1}, A^1, A^2, C^1, C^2)}{p(Y^{1,1}, A^1, A^2 | C^1, C^2)} \right\} \times \\
&\quad \mathbb{E}[Y^{1,1} | Y^{1,2}, A^1, A^2, C^1, C^2] p(Y^{1,1}, Y^{1,2}, A^1, A^2, C^1, C^2) \\
&= \sum_{Y^{1,1}, Y^{1,2}, C^1, C^2} \left\{ \frac{\pi(Y^{1,1}, A^1, A^2, C^1, C^2)}{p(Y^{1,1}, A^1, A^2 | C^1, C^2)} \right\} \mathbb{E}[Y^{1,1} | Y^{1,2}, A^1, A^2, C^1, C^2] p(Y^{1,1}, Y^{1,2}, A^1, A^2, C^1, C^2) \\
&= \sum_{Y^{1,1}, Y^{1,2}, C^1, C^2} \pi(Y^{1,1}, A^1, A^2, C^1, C^2) \mathbb{E}[Y^{1,1} | Y^{1,2}, A^1, A^2, C^1, C^2] \times \\
&\quad p(Y^{1,2} | Y^{1,1}, A^1, A^2, C^1, C^2) p(C^1, C^2) \\
&= \psi
\end{aligned}$$

The rest of the terms are mean 0 by an application of iterated expectations.  $\square$

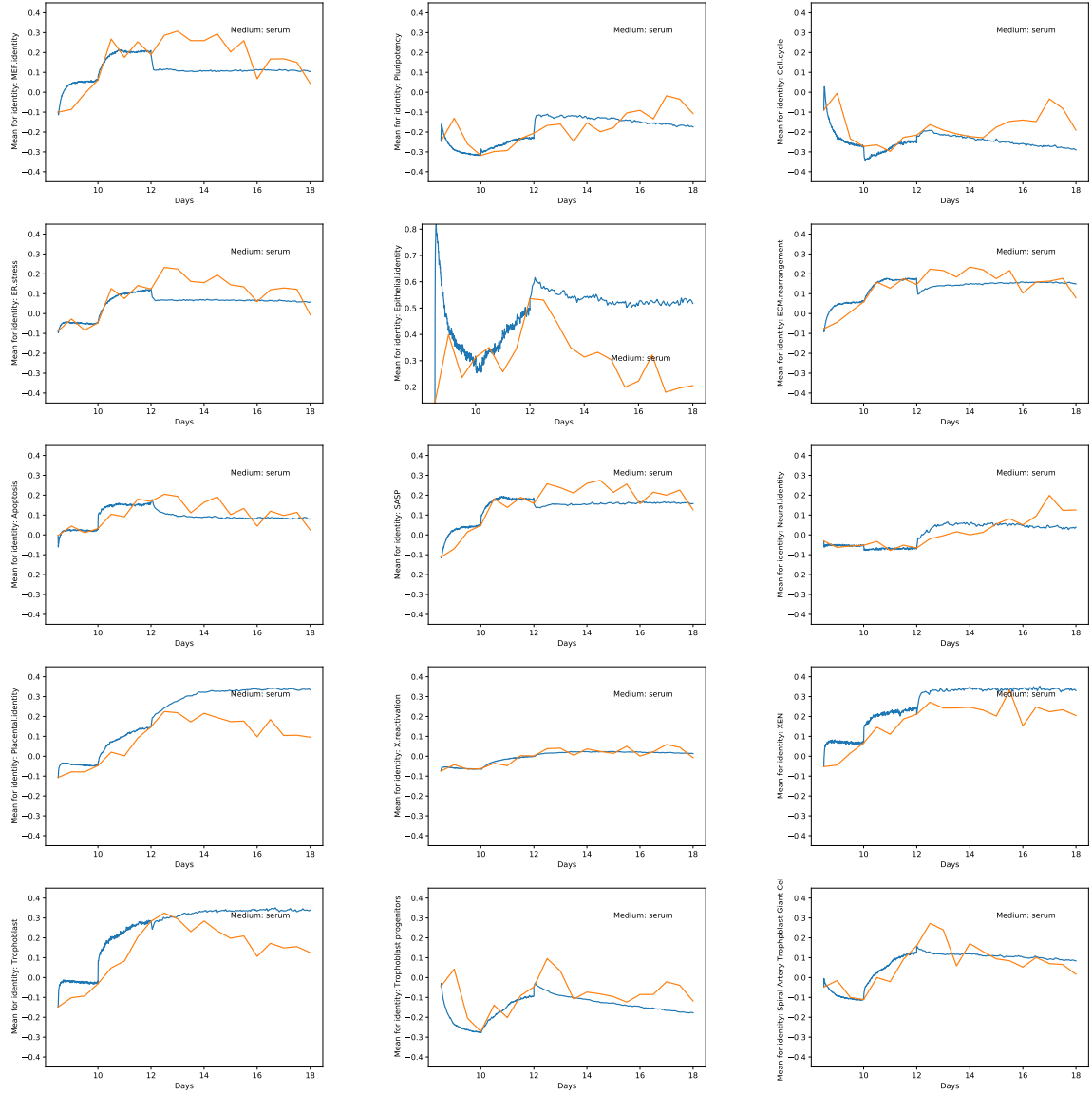
## B. Cellular Reprogramming: Plots



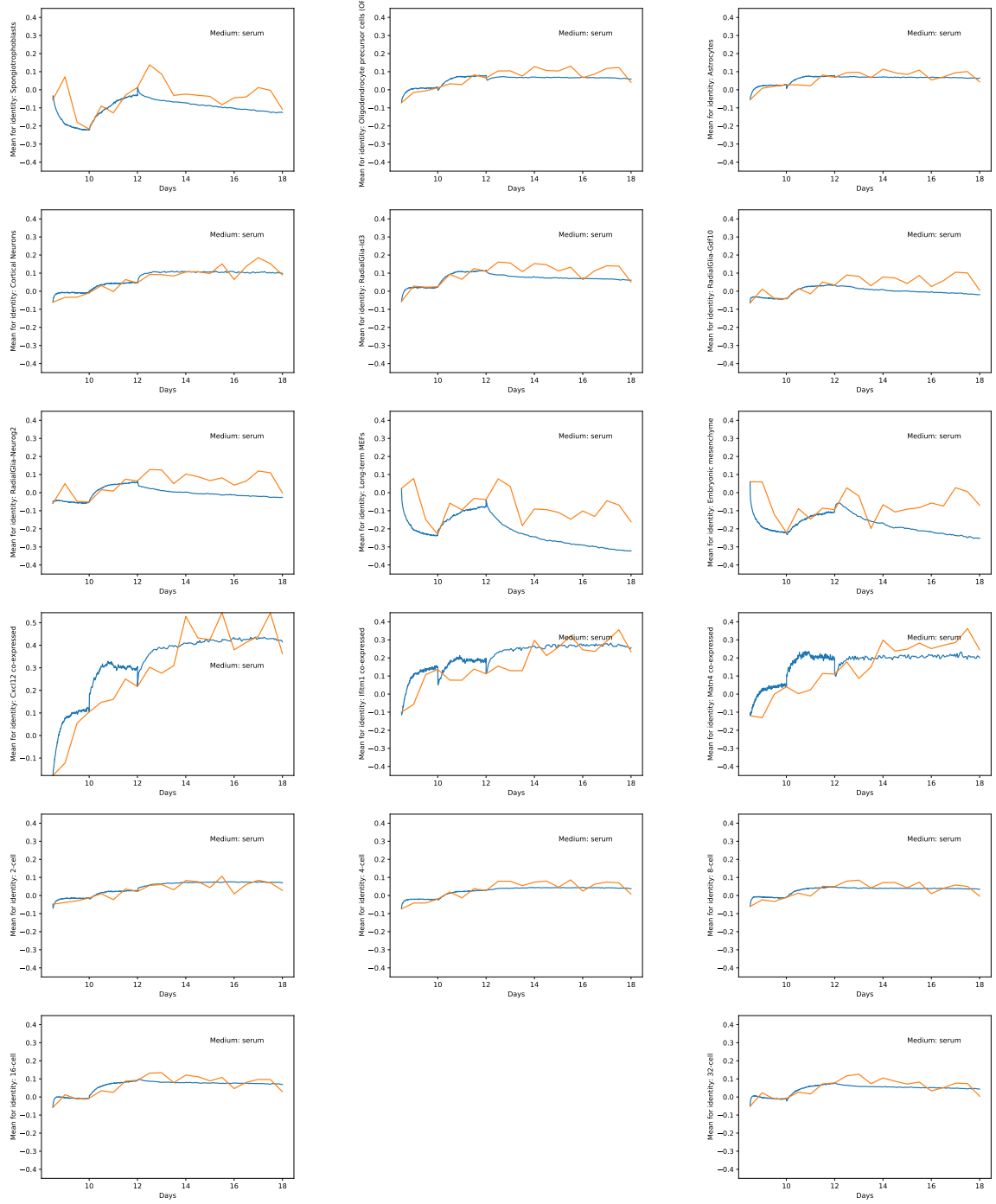
**Figure IV.1.** Trajectories for the first 15 (of 32) dimensions for medium 2i. The blue trajectory is derived from Gibbs compatible PCM and the orange trajectory is plotted from the original dataset for comparison, sampled every 12 hours. See the next page for the other dimensions.



**Figure IV.2.** Trajectories for 17 (of the 32) dimensions for medium 2i, continued from Fig. IV.1. The blue trajectory is derived from Gibbs compatible PCM and the orange trajectory is plotted from the original dataset for comparison, sampled every 12 hours.



**Figure IV.3.** Trajectories for the first 15 (of 32) dimensions for medium serum. The blue trajectory is derived from Gibbs compatible PCM and the orange trajectory is plotted from the original dataset for comparison, sampled every 12 hours. See the next page for the other dimensions.



**Figure IV.4.** Trajectories for 17 (of the 32) dimensions for medium serum, continued from Fig. IV.3. The blue trajectory is derived from Gibbs compatible PCM and the orange trajectory is plotted from the original dataset for comparison, sampled every 12 hours.