Problems in Identification and Estimation: Algorithms for Pathogen, Ancestral, and Rashomon Analysis

Aparajithan Venkateswaran

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2024

Reading Committee:
Tyler McCormick, Chair
Emilija Perković, Chair
Alex Luedtke

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

Problems in Identification and Estimation:
Algorithms for Pathogen, Ancestral, and Rashomon Analysis

Aparajithan Venkateswaran

Co-Chairs of the Supervisory Committee:
Professor Tyler McCormick
Department of Statistics

Professor Emilija Perković
Department of Statistics

This dissertation answers three questions on identifiability and estimability that arise in policy-making and causal discovery. First, we study contact tracing as a tool to prevent the spread of infectious diseases. We show how to substantially improve the efficiency of contact tracing using multi-armed bandits that leverage heterogeneity in how infectious a sick person is. We propose to test contacts of infected persons to ascertain whether they are likely to be a "high infector" and to find additional infections only if it is likely to be highly fruitful. Using administrative COVID-19 contact tracing datasets, we show that an easily implementable strategy in the field performs at nearly optimal levels.

Second, we robustly estimate heterogeneities in the outcome of interest with respect to a factorial feature space. We partition this factorial space into "pools" of feature combinations where the outcome differs only across the pools. We fully enumerate the Rashomon Partition Set (RPS), a collection of all partitions with sufficiently high posterior density. Using the ℓ_0 prior, which we show is minimax optimal, we calculate approximation error relative to the entire posterior and bound the size of the RPS. In three empirical settings (charitable giving, chromosomal structure, and microfinance), we highlight robust conclusions, including affirmations and reversals of extant literature findings.

Third, we restrict Markov equivalence classes of causal maximal ancestral graphs (MAGs) that agree with expert knowledge in the form of edge orientations. We can uniquely represent this equivalence class using its essential graph. We revise two previously described graphical orientation rules and present a novel rule to add expert knowledge. We provide an algorithm for adding expert knowledge and show that it is complete for edge marks in the circle component of the essential graph. We also provide an algorithm for verifying completeness in the general case.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my dissertation committee, Kevin Jamieson, Alex Luedtke, Tyler McCormick, and Ema Perković, for their invaluable support and guidance. I am deeply indebted and forever grateful to my advisors, Tyler and Ema, for countless hours of advising and innumerable opportunities they have given me over the years. I am thankful to Tyler for helping me identify exciting problems in the social sciences and answer them rigorously without losing a grasp of the thread. I am thankful to Ema for introducing me to the world of graphical models and teaching me to argue technical details behind intuitions. Working with them has been immensely rewarding and equally enjoyable. Their constant feedback and encouragement were a catalyst for this dissertation.

I want to thank Tyler's and Ema's research groups for encouraging me to be a better scientist. Thanks to Jess Kunke, Sara LaPlante, Shane Lubold, Vydhourie Thiyageswaran, and Steve Wilkins-Reeves for many math and peer advising sessions. I also want to thank Yen-Chi Chen, Carlos Cinelli, Alex, and Thomas Richardson for kindling my passion for statistics through their courses. I am grateful to Arun Chandrasekhar, Jishnu Das, and Anirudh Sankar for being excellent collaborators. I would also like to thank Dan Larremore for showing me that science can be fun and inspiring me to go on this journey.

I also want to thank my wonderful friends and family, near and far, for cheering me on. Many thanks to Medha Agarwal, James Buenfil, Ellen Considine, Jillian Fisher, Kayla Irish, Alex Kokot, Ronak Mehta, Shreya Prakash, Sandra Sajeev, and Suyog Soti for food, music, movies, games, and late-night conversations. Lastly, I am especially thankful to my parents and sister for their unwavering support through this and every endeavor; I would not be where I am without them.

DEDICATION

To my parents and sister.

"If nobody asked questions, then we would never learn anything."

— Shallan, *Oathbringer*

Chapter 1
INTRODUCTION

"The part that stories leave out is everything that comes before."

— Tress, *Tress of the Emerald Sea*

Here is a cartoon of statistical learning: (1) construct a data-generating model based on our view of the world, (2) collect data, and (3) estimate the model parameters. The goal is to use insights from this model to improve our understanding of the world and make decisions.

If only science was as simple as that. There is uncertainty at every step along the way. There is model uncertainty – if we knew with absolute certainty how the world works, there would be no need for this exercise in the first place. There is uncertainty in data collection – sampling errors, measurement errors, finiteness, etc. Together, these are reflected in our estimated models and quantities of interest. Then, the role of the scientist is to learn how the world *truly* works by minimizing the uncertainty in their *post hoc* belief.

Naturally, several questions arise. What if we are unwilling (or unable) to make strong assumptions about our world? What if our assumptions are underspecified, overspecified, or misspecified? What if our data is biased? What if it has a low rank, i.e., highly correlated or high dimensional? These can be classified as either an *identifiability* problem or an *estimability* problem.

A data-generating model (or a quantity of interest) is identifiable if, given access to an infinite number of observations, it is possible to learn the true underlying parameters. Formally, for a model P_θ where θ is the unknown parameter of interest, we say that θ is identifiable if $P_{\theta_1} \neq P_{\theta_2}$ for every $\theta_1 \neq \theta_2$ (cf. Chapter 1, Definition 5.2 of Lehmann and

Casella (2006), and; p. 62, Equation 5.34 of Van der Vaart (2000)). If the quantity of interest is indeed identifiable, then estimating it from finite data is a well-defined task, i.e., it exists and is unique. However, with a finite sample, the estimates may not be stable, i.e., the solution may not be sufficiently smooth given the data. This is referred to as an estimability problem. When identifiability is guaranteed, regularization is a commonly used smoothing technique to achieve stability (Tikhonov et al., 1943; Bickel and Li, 2006). Existence, uniqueness, and smoothness are the Hadamard conditions of a well-posed problem (Hadamard, 1902). See Maclaren and Nicholson (2019) for a unifying review of identification and estimation problems.

The three works contained in this dissertation revolve around identifiability and estimability. In the first problem, we wish to estimate the underlying infectiousness of each sick person during an epidemic to uncover more infections through contact tracing. Here, practical challenges and finite lifetimes of infections make it infeasible to estimate the true infectiousness. In the second problem, we seek to robustly estimate heterogeneities in factorial feature space by enumerating near-optimal models that explain the world differently, i.e., the estimated model is unstable. Contrary to traditional statistical and machine learning wisdom, in these two tasks, we highlight that learning the *sub-optimal* can be more useful than finding the *optimal* – "good-enough is better." In the final problem, we turn to causal discovery where multiple equally-optimal models could have generated the data, i.e., the true model is not identified. We seek to obtain the true model by pruning this set of equally optimal models using expert knowledge about the system.

The remainder of this dissertation is organized as follows:

Chapter 2 In this chapter, we study the problem of contact tracing of an infectious disease epidemic. Infectious diseases spread when infected people pass the infection to their contacts. We know that there is heterogeneity in individual infectiousness (Hagenaars et al., 2004; Bolzoni et al., 2007; Arinaminpathy et al., 2020). In other words, some people are more infectious than others. The public health agency's goal in contact tracing is to identify the

infected individuals, quarantine or treat them, and control the spread of the disease.

Contact tracing can be framed as a multi-armed bandit (Lattimore and Szepesvári, 2020). In a multi-armed bandit, a player is provided with a set of arms to play. When the player pulls an arm, they are rewarded. The reward of each arm is unknown, and the player's goal is to identify a sequence of arms to pull to maximize their reward. In the context of contact tracing, the *arms* are infected people. *Pulling an arm* corresponds to testing a contact of that infected person. The public health agency, the *player*, aims to uncover as many infections as possible utilizing the least number of tests. Contact tracing differs from the standard multi-armed bandit in that each person only has a finite number of contacts (a setting termed as a *mortal bandit* by Chakrabarti et al. (2008)).

At the heart of this bandit strategy is a focus on learning. In the typical conceptualization of contact tracing, contacts of an infected person are tested to find more infections. Under a learning-first framework, however, contacts of infected persons are tested to ascertain whether the infected person is likely to be a "high infector" and to find additional infections only if it is likely to be highly fruitful. We establish the optimality of two bandit algorithms, Adaptive Greedy sampling and Pilot sampling. Both of these algorithms take a "good-enough is better" approach to identify *everyone* who is more infectious than the average person (rather than just the *most* infectious person).

Using three administrative contact tracing datasets from India and Pakistan during the COVID-19 pandemic, we demonstrate that this approach, perhaps surprisingly, improves efficiency. We find 80% of infections with just 40% of contacts, while current approaches test twice as many contacts to identify the same number of infections. We further show that a simple strategy easily implemented in the field performs at nearly optimal levels, allowing for feasible contact tracing. These results are immediately transferable to contact tracing in any epidemic.

This is joint work with Jishnu Das and Tyler McCormick (Venkateswaran et al., 2023).

Chapter 3 Zooming out to the broader scientific inquiry, we often wish to learn how some outcomes vary with features of interest. For example, how do various drug combinations affect health outcomes, or how does technology adoption depend on incentives and demographics? Suppose we are in a factorial space, i.e., our covariates are discrete and ordered. A natural way to answer this question is to partition the feature space into "pools" of feature combinations where the outcome differs across the pools (but not within a pool). Existing approaches for identifying such partitions either (i) search for a single *optimal* partition under some assumptions about the association between covariates (for example, decision trees (Breiman et al., 1984) and Lasso (Tibshirani, 1996)) or (ii) attempt to *sample* from the *entire* set of possible partitions (for example, random forests (Breiman, 2001a), Bayesian Lasso (Park and Casella, 2008), and BARTs (Chipman et al., 2010)). These ignore the reality that, especially with correlation structure in covariates, many ways to partition the covariate space may be indistinguishable from a statistical perspective despite very different implications for policy or science. This is an estimability problem (here, we assume that the true model is identified if we have infinite data and treat non-identifiability separately in Chapter 4). Such model multiplicity is called the *Rashomon effect* (Breiman, 2001b).[1]

We develop an alternative perspective, called *Rashomon Partition Sets* (RPSs). Each item in the RPS partitions the factorial space of covariates using a tree-like geometry. RPSs incorporate *all* partitions with posterior values near that of the *maximum a posteriori* partition, even if they offer substantively very different explanations. We do so using a prior that makes no assumptions about the associations between covariates, the ℓ_0 prior, which we show is minimax optimal. Conditional on being in the RPS, we characterize the approximation error relative to the full posterior of any measurable function of the vector of feature combination effects on outcomes. We also show that the RPS is much smaller than the space of all partitions and provide an algorithm to enumerate the RPS.

[1]The name can be traced back to the 1950 Japanese film *Rashomon*, directed by Akira Kurosawa, where a tragic event is re-told from four wildly different perspectives while the facts remain the same. This movie itself was based on two short stories by Ryunosuke Akutagawa, *Rashōmon* (1915) and *In a Grove* (1922).

We apply our method to three empirical settings: price match effects on charitable donations, heterogeneity in chromosomal structure (telomere length), and the introduction of microfinance to small businesses in India. We highlight robust conclusions learned from the RPS in each of these settings. These include affirmations and reversals of extant literature's findings in each setting.

This is joint work with Anirudh Sankar, Arun Chandrasekhar, and Tyler McCormick and is currently under review (Venkateswaran et al., 2024).

Chapter 4 Next, we turn to causal discovery from observational data using graphical models. The commonly used directed acyclic graph cannot correctly represent causal relationships in the presence of unmeasured confounders. Instead, we use graphical models called *ancestral graphs*. Mathematically, one can construct the so-called *maximal ancestral graph* (MAG) by beginning with the true data-generating model that includes all unmeasured confounders and marginalizing over the latent variables (Richardson and Spirtes, 2002). The true data-generating MAG usually cannot be identified from observational data alone because more than one MAG may encode the same conditional independence relationships in the observational distribution. We call such MAGs Markov equivalent and uniquely represent this Markov equivalence class by an *essential graph*, otherwise known as a *partial ancestral graph* (Ali et al., 2009).

In this chapter, we study the problem of restricting Markov equivalence classes of MAGs that agree with knowledge obtained beyond the observational data, i.e., expert knowledge. For example, if the causal relationship between two variables is not identified in the Markov equivalence class but domain knowledge exactly identifies it, then we refer to this as expert knowledge. In practice, this takes the form of edge marks in the causal graph.

We seek to learn the essential graph *restricted* to expert knowledge. This is analogous to Meek (1995)'s work on DAGs. Our contributions in the latent variable setting are three-fold. First, we use various Markov equivalence characterizations to prove important properties of these restricted equivalence classes. Second, we present three sound graphical orientation

rules, two of which generalize previously known rules, for adding expert knowledge to an essential graph. Third, we provide an algorithm for including this expert knowledge and show that our algorithm is complete in certain settings i.e., in these settings, the output of our algorithm is a *restricted essential ancestral graph*. Outside of our specified settings, we provide an algorithm for checking whether a graph is a restricted essential graph and discuss its runtime.

This is joint work with Emilija Perković.

Chapter 5 In the final chapter, I provide concluding remarks and highlight several important questions that the methods introduced in this dissertation now allow us to think about.

I provide commentary at the beginning of each chapter on how it relates to the other works discussed here. However, each chapter is self-contained and can be read independently without sacrificing continuity. Technical details, proofs, additional simulations, and other supplementary materials are presented at the end of the dissertation in their respective appendices.

Chapter 2

FEASIBLE CONTACT TRACING

> "As any problem to overcome is merely a set of smaller problems to overcome in a sequence, he divided his goal of becoming a dragon into three steps."
>
> — Wit, *Rhythm of War*

This chapter is largely adapted from Venkateswaran et al. (2023) (to be submitted and available on arXiv). This is joint work with Jishnu Das and Tyler McCormick. We acknowledge Sanmay Das and Anja Sautmann for providing valuable feedback and Eva Tourangeau for research assistance.

The key question in this chapter is to identify whose contacts should be tested to maximize the number of infections uncovered during contact tracing of an infectious disease. Assuming we have a list of contacts, answering this question boils down to finding the most infectious people and testing their contacts. Finding such a person can be hard. Fortunately, and perhaps surprisingly, we show that we don't need to find the most infectious person; testing contacts of someone who is more infectious than the average person is sufficient. This is the idea that "good enough is better" – it is better to find the near-optimal than the optimal. We will revisit and fully explore this concept in Chapter 3 through the Rashomon effect.

2.1 Introduction

Contact tracing, quarantines, and other pharmaceutical and non-pharmaceutical interventions are key in the fight against infectious diseases. It proceeds as a branching process on the network of contacts (Huerta and Tsimring, 2002; Lloyd-Smith et al., 2005). Assuming that transmission events are independent of each other, we begin with an arbitrary infected

person and test each of their contacts for the infection. If the test returns positive, indicating the presence of an infection, we start the process anew with a new set of potentially infected people.

Contact tracing has been deployed widely and effectively in infectious diseases that are geographically localized or spread slowly, such as Ebola or HIV (for example Hyman et al., 2003; Saurabh and Prateek, 2017). However, contact tracing poses substantial logistical and financial challenges for infections that spread more quickly and easily. During the COVID-19 pandemic, for example, contact tracing proved challenging, and in countries like the United Kingdom, many contacts could not be reached or tested in time. This is unsurprising – in typical networks, while most people have very few contacts, some may have thousands, especially in public-facing jobs. Given the wide variation in the number of contacts, we ask whether there are ways to reduce the number of contacts who need to be tested without a commensurate decline in the number of new infections uncovered. Interestingly, while attempts to improve the efficacy of contact tracing have focused on innovations that allow more contacts to be reached through the use of cell phones and other passive data (see Danquah et al., 2019; Zhao et al., 2020), there is little research on whose contacts should be traced. Feasible contact tracing remains an elusive goal in the face of a rapidly moving pandemic like COVID-19.

Here, we develop the insight that the testing of any contact provides new information about the *infector*, which can then be leveraged to improve contact tracing. We show that substantial gains are possible if there is heterogeneity in the likelihood that someone passes on the infection to others, a term we label per-contact infectivity (PCI), which could arise from biological or behavioral factors. To see how information about PCI interacts with contact tracing, suppose there are only two types of people in the population: those who pass on the infection with probability 1 to every person they meet and those who never pass the infection, i.e., pass on the infection with probability 0. In such a world, a simple algorithm that tests exactly one contact of every infected person, to begin with, and tests further contacts if and only if the tested contact is positive can result in substantial cost

savings because additional tests do not provide more information about the infection status of untested contacts.

We expand on this intuition in this chapter. The core message is that when there is heterogeneity in PCI, there are massive gains to learning about PCI quickly. We further contend that strategies that prioritize learning this way are feasible in practice, for instance, by testing a subset of people living with the person. We demonstrate that the kind of heterogeneity in PCI that we need to realize significant gains in the efficiency of contact tracing is consistent with the data from three different South Asian locations during COVID-19. Our feasible algorithms would have allowed 80% of infections to have been detected by testing only 40% of the contacts of infected persons in two of the datasets (from Punjab, India, and southern India) and 60% of contacts in the third (from Punjab, Pakistan). In contrast, currently employed strategies that test all contacts of an infected person need to test 80% of the population to uncover the same number of infections.

This chapter is structured as follows. We first pose contact tracing as a multi-armed bandit and review related works in Section 2.2. In Section 2.3, we establish a theoretical framework and provide new results on the asymptotic optimality of different contact tracing algorithms. Turning from asymptotic results, in Section 2.4 we then consider finite samples and show that the performance of the algorithms is sensitive to the specific distribution of infections. For instance, depending on the distributional parameters, one algorithm can outperform another, allowing us to outline specific policy actionable guidelines for the appropriate choice of an algorithm. In Section 2.5, we take these insights to administrative contact tracing data from Punjab (Pakistan), Punjab (India), and southern India during the COVID-19 pandemic. We estimate the distribution of PCI for each dataset and show a large variation in PCI among infected persons in our settings. Then, we show that the bandit algorithms we outline are far more efficient than a naive sampling strategy and reconcile our findings with those from the empirical simulations. Finally, we conclude our discussion and provide directions for future research in Section 2.6.

2.2 Contact Tracing as Bandits

Bandit algorithms are a staple tool for sequential decision-making under uncertainty. The general setup is as follows. A decision-maker faces a choice between several options. For each option, there is a reward for choosing that option that comes from a probability distribution. The reward distributions are not known in advance, so the decision-maker faces a trade-off. Continuing to choose the same option provides a reliable reward, though moving to a different option might provide an even higher reward. Under this uncertainty, the goal is to construct a sequence of decisions amongst the set of options that maximizes the reward. As we discuss in the next section, there is a deep and active literature that develops and evaluates bandit algorithms in a wide range of contacts. This section focuses on the connection between these algorithms and contact tracing. In contact tracing, the decisions represent people whose contacts could be tested, and the reward refers to the number of new infections the decision-maker discovers.

The key insight in this chapter is that each person infects their contacts at a different rate, called per-contact infectivity (PCI). In bandit language, this heterogeneity means that the reward distribution across people varies widely, with some infected people likely to infect many of their contacts while others (or most, as we see in our empirical examples) infect few or none. This leads us to consider two things for contact tracing. First, we want to identify highly infectious individuals. Second, we do not want to spend a lot of effort trying to determine who is the most infectious, i.e., it is sufficient to find someone who is more infectious than the average person. This is called the *explore-exploit* trade-off in the context of bandits. We want to *exploit* the infectious people to uncover more positive infections while also *exploring* to identify those who may be more infectious.

Earlier, we considered a scenario where the PCI is binary, 0 or 1. When the distribution of PCI is continuous, this becomes analogous to what's known as a multi-armed bandit problem. The possible *arms* are infected individuals. *Pulling an arm* refers to testing a contact of an infected person. The *payoff* is measured through the number of infections identified, and

the *time horizon* is a function of the number of people tested. In the bandit setting, there is a trade-off between *exploiting* infectious people available in a current state versus *exploring* for more infectious people with greater payoffs. The optimal strategy determines whether to continue testing the contacts of someone with a known PCI or to test the contacts of someone whose PCI is unknown.

Although there are clear parallels, contact tracing differs from the standard bandit in two ways. First, unlike a standard bandit where the arms are fixed, for an infectious disease, the arms appear and disappear rapidly – new people get sick, and people who were previously sick either recover or die within a fairly short period relative to the length of the epidemic. Thus, it serves little to know the precise PCI of an arm if it exists only for a limited period. Second, since each person has a fixed number of contacts, there are only so many times each *arm* can be *pulled*, and the potential rewards differ based on the number of contacts. This variant of the standard bandit is called the *mortal multi-armed bandit* (Chakrabarti et al., 2008).[1]

The mortal multi-armed bandit changes the goal of the learning algorithm away from finding the *best* arm that can be exploited indefinitely to an approach where it is sufficient to find a *good-enough* arm by placing emphasis on arms that live longer, i.e., people with more contacts. For this problem, two algorithms called Adaptive Greedy and Stochastic Sampling (which we call Pilot Sampling) have been shown to be able to theoretically identify the maximum number of infections per test in the long run by Chakrabarti et al. (2008). We take these algorithms one step further by additionally incorporating the reward distribution (the PCI distribution) to show that they remain optimal. In doing so, we also demonstrate the value of information about the PCI distribution in determining the optimal approach

[1]It is worth noting that mortal multi-armed bandits are similar to another class of bandit problems called *rotting bandits* (Levine et al., 2017). In rotting bandits, the mean reward distributions are not stationary. In particular, the expected mean reward of each arm decays as a function of the number of times the arm has been pulled. In that sense, mortal bandits can roughly be seen as a discretization of rotting bandits where the decaying function is defined step-wise. Seznec et al. (2019) show that when the number of arms is fixed, rotting bandits are no harder than the standard stochastic bandit. However, when there are infinitely many arms, the problem becomes significantly harder (Kim et al., 2022). In our analysis, we will assume that there are only finitely many arms. And we leave the scenario with infinite arms for future research.

through empirical simulations. One crucial advantage in our setting is that our goal is to define a strategy that is *better* than current approaches to contact tracing, which generally involves testing all contacts. In that sense, our problem is easier than finding the *best*, i.e., the optimal strategy that most other works tackle.

2.2.1 Related Literature

There is a long history of research on bandits. Although the current version of the problem was formalized only in Robbins (1952), such sequential optimization problems were considered as early as Thompson (1933). Optimal solutions to these problems using index-based rules have been discussed by Gittins (1979). Another class of bandit algorithms that came to be known as the upper confidence bounds (UCB) was first put forth by Lai and Robbins (1985). There are other strategies, such as greedy algorithms and probability matching. Sutton and Barto (2018) and Bubeck et al. (2012) provide an extensive literature review of bandit algorithms.[2]

Although bandits are predominantly associated with machine learning and computer science, a growing body of literature uses bandits to model human decisions. Cohen et al. (2007) suggest that humans make decisions in exploration-exploitation problems using index rules, reminiscent of Gittins (1979). UCB-type algorithms have been used in decision-making by Reverdy et al. (2014) and Wu et al. (2018). These have found applications in medical decision-making. For instance, Frank and Zeckhauser (2007) viewed the treatment of depression as a bandit problem, and Currie and MacLeod (2020) show that more skilled doctors tend to favor a strategy with greater experimentation when searching for treatments for depression. Perhaps a more direct application of bandits is in designing experiments as noted by Athey and Imbens (2019). In fact, Thompson sampling, one of the earliest bandit

[2]There are several variants of the classical bandit problem besides the mortal bandit such as *contextual bandits* where we observe covariates as well as rewards (Langford and Zhang, 2007), *adversarial bandits* where an adversary changes the reward structure (Auer and Cesa-Bianchi, 1998), *infinite-armed bandits* where there is an infinite number of arms to play (Agrawal, 1995), and *non-stationary bandits* where there is a drift in the mean rewards (Besbes et al., 2014).

algorithms put forth by Thompson (1933) was developed to guide data collection by identifying treatment arms that units should be assigned to. We add to this growing literature in economics by finding optimal contact tracing strategies using bandits.

Since 2020, the COVID-19 pandemic has sparked a body of work that explores the connection between contact tracing and bandits. For example, Grushka-Cohen et al. (2020) use a bandit-style framework by assigning risk scores and ranking individuals based on expressed symptoms and characteristics to identify people who need to be tested. However, they do not explicitly leverage heterogeneity in infectivity, which is where we show we can derive substantial gains. Wang et al. (2020) construct an agent-based model to simulate infectious disease dynamics and use bandits to trace contacts. Bastani et al. (2021) use batched bandits to identify groups of people to test at nation borders and found 1.85 times as many asymptomatic travelers as random surveillance testing. Meister and Kleinberg (2021) model the spread of infection in two distinct phases, an infection phase and a contact tracing phase, and derive optimal strategies. Chugg and Ho (2021) study a related problem of estimating the prevalence of the disease at a given time step using bandits.

Additionally, some work has identified the heterogeneity in infectiousness that we exploit here. Hagenaars et al. (2004) explore this in a spatial context. Bolzoni et al. (2007) caution that disease control strategies should account for heterogeneity and Miller (2007) found that epidemics are more likely when variance in infectivity is large. More recently, Arinaminpathy et al. (2020) quantified heterogeneity in infectivity in the transmission of COVID-19.

The key distinction in our work is that we exploit this heterogeneity to motivate a learning-first perspective for contact tracing. In the presence of heterogeneity, the priority is to learn about the PCI of the *infector*. The infector's PCI gives the decision-maker critical information about the reward distribution, resulting in more efficient decisions about who to test next. We make three contributions to the literature based on that key shift in perspective.

First, we consider active learning algorithms arising from heterogeneity in PCI in the context of contact tracing. Although previous contributions have proposed active learning

algorithms and identified heterogeneity in infectivity for multiple infectious diseases, they have not been addressed jointly in the literature. Second, in merging the two and being the first to draw the connection to mortal bandits, we also provide novel theoretical results on the asymptotic properties of two bandit algorithms. Third, ours is also the first work to use contact tracing data to show that bandit algorithms can lead to marked declines in the fraction of contacts needing to be sampled without a commensurate loss in the number of infected individuals identified. Taken together, our results provide policy-actionable and feasible methods for contact tracing in the field.

2.3 Methodology

In this section, we present new results on the optimality of different strategies. We formally introduce the mortal bandit problem and define loss functions to quantify performance called regret and Bayesian regret. Then, we present a lower bound on the Bayesian regret. Next, we introduce two commonly used algorithms for the mortal bandit setting: Adaptive Greedy and Pilot Sampling. Then, we provide novel bounds on the Bayesian regret. In deriving these bounds, we assume that no new arms appear i.e., arms can only die. In practice, this means that a policymaker has a fixed group of infected persons and must decide how to allocate tests amongst the contacts of those individuals. Finally, we describe the intuition behind a variant of these algorithms where we sample arms by lifetime, which in the context of contact tracing corresponds to sampling based on the total number of contacts. We show that this strategy does not change our asymptotic optimality results.

2.3.1 Problem Setup and Notation

Consider the bandit problem with N arms (infected individuals). Pulling an arm i (testing a contact of i) rewards the decision-maker with a reward X_i ($X_i = 1$ if contact is positive and 0 otherwise). This reward comes from a distribution P_{μ_i} with unknown mean $\mu_i \in [0, 1]$ (the PCI). Each arm can only be pulled for a maximum of L_i times where L_i is known. We call L_i the lifetime of the arm (the number of contacts of i) and say that arm i is alive at time

t if we haven't already pulled it L_i times at time t. In other words, the infected person still has more contacts left to test. An arm is playable (more contacts can be tested) if and only if it is alive. This is in contrast to the standard bandit problem, where there is no limit to the number of times an arm can be pulled. In a setup where arms have limitless numbers of pulls, the goal of the bandit algorithm is to find the arm with maximum payoff (explore) and then play the maximally rewarding arm indefinitely (exploit). In contact tracing, this trade-off is less straightforward since even the most productive arm will eventually die (i.e., all contacts will be tested). We will assume that $\{(\mu_i, L_i)\}_{i=1}^N$ are independently and identically distributed (i.i.d.) from some joint prior Γ. We will denote the marginal distribution of μ as Γ_μ.

The agent sequentially pulls arms in order to maximize cumulative reward over T turns (the number of tests available). Let $H_t = \{(a_\tau, X_{a_\tau,\tau})\}_{\tau=1}^{t-1}$ denote the history of the decision maker's actions a_τ and the corresponding rewards $X_{a_\tau,\tau}$ up to time $t-1$. Then, we define the decision maker's policy π as a mapping from the history H_t to the next action $a_t \in \{1, \ldots, N\}$.

Define $\mu_t^\star = \max_i \mu_i \times \mathbb{I}\{\text{arm } i \text{ is alive at } t\}$. For a fixed set of mean rewards $(\mu_1, \ldots, \mu_N)$ and some history of actions according to a policy π, we define three different kinds of regret for the decision maker as follows:

$$R_T(\pi \mid \mu) = \sum_{t=1}^{T} \mu_t^\star - X_{a_t,t} \qquad \text{Realized regret}$$

$$\mathbb{E}[R_T(\pi \mid \mu)] = \sum_{t=1}^{T} \mu_t^\star - \mu_{a_t} \qquad \text{Mean regret}$$

$$BR_{T,N}(\pi) = \mathbb{E}_\Gamma \mathbb{E}[R_T(\pi \mid \mu)] \qquad \text{Bayesian regret}$$

There are two sources of randomness. The first source is reward realization, as we assume that the reward generation process is stochastic. This is reasonable even though the lifetime is finite because we do not know anything about the arm: we have a list of contacts, but we have no information about how many contacts are infected. The second source of randomness is the distribution of μ_i itself. This comes from the fact that there is heterogeneity in PCI,

i.e., some people are more infectious than other people, as shown by Arinaminpathy et al. (2020) for COVID-19.

Given the stochastic nature of the problem, instead of analyzing realized regret, we *average* out the randomness and analyze the resulting Bayesian regret. The goal of the policymaker is then to reduce the Bayesian regret i.e., their cumulative regret from pulling a sub-optimal arm over T turns averaged across all possible mean rewards, lifetimes, and realizations of data.[3]

We use $\mathcal{O}, \Theta$ to denote the usual order asymptotics, and $\tilde{\mathcal{O}}$ to denote $\mathcal{O}$ ignoring logarithmic factors. Formally, we say $f(x) = \mathcal{O}(g(x))$ if there is a $M > 0$ and $x_0 > 0$ such that for all $x \geq x_0$, $|f(x)| \leq Mg(x)$. In other words, asymptotically, $f(x)$ does not grow at a faster rate than $g(x)$. We say that $f(x) = \Theta(g(x))$ if there are constants $m, M > 0$ and $x_0 > 0$ such that for all $x \geq x_0$, $mg(x) \leq f(x) \leq Mg(x)$. In other words, asymptotically, $f(x)$ and $g(x)$ grow at the same rate. Finally, we say $f(x) = \tilde{\mathcal{O}}(g(x))$ if there is a $k > 0$ such that $f(x) = \mathcal{O}(g(x) \log^k x)$. In other words, asymptotically, $f(x)$ does not grow at a faster rate than $g(x)$ up to some logarithmic factors.

2.3.2 A Lower Bound on Bayesian Regret

The following definition, commonly used in the analysis of many-armed bandits, enables us to study the behavior of the mean rewards (see Wang et al., 2008; Carpentier and Valko, 2015; Bayati et al., 2020, for examples).

Definition 2.3.1 (γ-regular distribution)**.** *For $\gamma > 0$, a distribution Q with support $[0, 1]$ is called γ-regular if $\mathbb{P}_Q(\mu > 1 - \epsilon) = \Theta(\epsilon^\gamma)$ when $\epsilon \to 0$.*

Commonly used distributions including the Beta distribution, which we will see throughout this chapter, are γ-regular. In particular, for Beta(α, β), $\gamma = \alpha + \beta - 1$ whenever $\beta > 1$

[3]Of course, one may imagine an adversarial scenario where the mean rewards of all arms are highly concentrated near the maximum mean reward. Then, regret, as defined here, may not be the right objective to minimize as it will remain small. Perhaps maximizing the number of infections identified is a better objective. We leave this adversarial setting as an open problem for future researchers.

and $\gamma = \alpha$ otherwise.

γ controls the tail behavior of μ allowing us to bound regret in the worst-case scenario. Intuitively, when

1. $\gamma < 1$, the density is concentrated towards 1,

2. $\gamma = 1$, the density is (roughly) uniform near 1, and

3. $\gamma > 1$, the density is concentrated away from 1.

When the density is concentrated towards 1, we expect a lot of arms to be highly rewarding as many people are highly infectious. So as $\gamma \ll 1$, the regret in the best case (the lower bound) will be smaller. The lower bound will be larger as γ becomes large i.e., the density is concentrated away from 1. This is reflected in the theoretical lower bound in Theorem 2.3.2 which was originally shown for the standard bandit by Bayati et al. (2020) and we state without proof for the moral bandit case.

Theorem 2.3.2 (Lower bound (Theorem 3.1 of Bayati et al. (2020))). *Consider the mortal bandit setting. Suppose that mean rewards μ are drawn from a γ-regular distribution and that there is a constant c such that $T, N \geq c$. Then, there is a constant an absolute constant C such that for any policy π,*

$$BR_{T,N}(\pi) \geq C \min(N, T^{\gamma/(\gamma+1)}) \tag{2.1}$$

This theorem is important because it immediately places a benchmark against which we can measure the performance of our algorithms. If an algorithm approaches the lower bound, we know that its performance compares favorably to any other policy that may be considered. In fact, we will go one step better, by showing that both the algorithms we discuss for contact tracing asymptotically achieve the lower bound of Theorem 2.3.2 up to constant or logarithmic factors.

Our analysis throughout the rest of this chapter will depend on (some subset of) the following assumptions:

(A1) The joint distribution of the mean reward and lifetime is $(\mu, L) \sim \Gamma$ where the marginal Γ_μ is γ-regular.

(A2) At any time during the game, we have at least $N_m \geq 1$ arms to play from.

(A3) The reward distribution P_μ is 1-subgaussian.

(A4) The lifetime of an arm is independent of its mean reward i.e., $\mu_i \perp L_i$ for all arms i.

(A5) The reward distribution is $P_\mu \equiv \text{Bern}(\mu)$

(A6) The mean rewards for arms $\{\mu_i\}_{i=1}^N$ are i.i.d. $\text{Beta}(\alpha, \beta)$.

(A7) The average lifetime $L \geq K$ for some $K > 0$.

We now analyze the asymptotic properties of two algorithms that are used in this context.

2.3.3 Adaptive Greedy Sampling

The first algorithm we discuss is called Adaptive Greedy sampling and it is based on greedy sampling (Chakrabarti et al., 2008). In greedy sampling, at time t, we pull an arm that has the largest sample mean at that time. Instead, Adaptive Greedy chooses to explore a different arm with probability $1 - \max_i \widehat{\mu}_i^t$. Here $\widehat{\mu}_i^t$ is the sample mean of arm i at time t and the max is taken over all arms that are alive. So if $\max_i \widehat{\mu}_i^t = 1$, then the algorithm decides to be greedy and pulls arm $\text{argmax}_i \widehat{\mu}_i^t$ but if $\max_i \widehat{\mu}_i^t = 0$, the algorithm randomly pulls an arm that is available to play. This method is described in Algorithm 1.

Example 2.3.3 (How does Adaptive Greedy work?)**.** *Suppose that mean rewards are* $\mu \in [0, 1]$ *and* $\mu \sim \Gamma$*. Suppose that there are only two arms with unknown rewards* $\mu_1, \mu_2 \in [0, 1]$ *and known lifetimes,* L_1, L_2*.*

Suppose that $\mu_1 = 0.1$ and $\mu_2 = 0.8$. Let $\widehat{\mu}_i$ be the current estimate of μ_i. In Adaptive Greedy, we start with $\widehat{\mu}_i = 0$. Since $\max \widehat{\mu}_i = 0$, we will for sure explore the space of arms. Let's say, we pick arm 2 and pull it. Suppose we are rewarded. Then, we update $\widehat{\mu}_2 = 0.5$. Now, we explore with probability $1 - \max \widehat{\mu}_i = 0.5$ and exploit arm 2 (current highest mean reward) otherwise. Suppose, we exploit arm 2 and are rewarded again. Then, $\widehat{\mu}_2 = 0.67$. Now, we exploit arm 2 with a larger probability, 0.67. Maybe next time, we choose to explore arm 1 out of randomness and are not rewarded so $\widehat{\mu}_1 = 0$ still. As we can see, this algorithm can, in the long run, settle upon the arm with the largest mean reward. However, it is not strictly greedy as it does allow some random exploration based on how good our current best arm is. □

We now present a new result in Theorem 2.3.4 showing that Adaptive Greedy shares the same asymptotic behavior as Greedy sampling. While this is a new bound for the Bayesian regret, we note that Tracà et al. (2020) state a novel bound for the mean regret. We provide a step-by-step technical discussion in Appendix A.1.

Theorem 2.3.4 (Bayesian Regret for Adaptive Greedy)**.** *Under the assumptions (A1)-(A5), for* $\epsilon \in (0, 1/3)$ *and* $N > \log T$*,*

$$BR_{T,N}(AG) = \begin{cases} \tilde{\mathcal{O}}\left(TN^{-1/\gamma} + N \min(\sqrt{T}, N^{1/\gamma})^{1-\gamma}\right), & \gamma < 1 \\ \tilde{\mathcal{O}}\left(TN^{-1/\gamma} + N\right), & \gamma \geq 1 \end{cases}$$

Observe that when $N \leq T^{\gamma/(\gamma+1)}$, $BR_{T,N}(\text{AG}) = \tilde{\mathcal{O}}(T^{\gamma/(\gamma+1)})$ and otherwise, $BR_{T,N}(\text{AG}) = \tilde{\mathcal{O}}(N)$. To match the lower bound described in Theorem 2.3.2, we subsample arms as in Bayati et al. (2020). Therefore, when $N > T^{\gamma/(\gamma+1)}$ we perform Adaptive Greedy on a subset of $m = \Theta(T^{\gamma/(\gamma+1)})$ arms to obtain $BR_{T,N}(\text{AG}) = \tilde{\mathcal{O}}(T^{\gamma/(\gamma+1)})$. Thus, asymptotically, the Bayesian regret of adaptive greedy (and the subsampled version) matches the optimal lower bound up to some log factors.

Understanding Theorem 2.3.4 Studying the behavior of Adaptive Greedy is difficult as both sources of randomness (that generate the mean reward and the reward realization) come into play immediately. So we can think of two complementary events. If we have bad luck, then no matter whose contacts we test (infectious person or not), we will never uncover new infections. In other words, the realized rewards are much much smaller than the true mean rewards (this can be quantified by invoking the sub-gaussian assumption). The first term in the asymptotic bound of Theorem 2.3.4 defines the likelihood of having bad luck. If we have good luck, then there will be at least one infectious person who infects people at a rate close to their PCI. In other words, there is at least one arm whose realized rewards are close to the true mean reward. The second term in the asymptotic bound defines the regret in this situation.

To illustrate the role of γ, consider the case when $\gamma \ll 1$. Here, the PCI density is concentrated near 1. Since nearly everyone is highly infectious, the likelihood of bad luck is negligible. Therefore, the second term in the asymptotic bound plays a more important role.

2.3.4 Pilot Sampling

The second sampling method we consider is Stochastic sampling, which we call this *Pilot Sampling* (Chakrabarti et al., 2008). The idea is to pull an arm a finite number of times that is smaller than its lifetime. If the sample mean of the arm based on this *pilot* meets a predetermined threshold, then we deem this arm to be highly rewarding and pull it until it dies.

In the context of contact tracing, pilot sampling means testing a pre-defined number of contacts, then only testing the remaining contacts if enough infections are found in the initial set (the *pilot* set). This approach would allow health officials to, for example, test a subset of the people living with an infected person and only test remaining contacts if enough family members test positive. This approach also illustrates the gains of knowing the distribution of infectiousness since it prioritizes finding the most infectious individuals.

Algorithm 1 Adaptive Greedy

Input: T budget, N arms

Set $t = 0$
while $t < T$ **do**
  $t = t + 1$
  if $t \leq N$ **then**
    Pull arm t and record reward
  else
    $X \sim \text{Bern}(\max_{i \text{ is alive}} \widehat{\mu}_i^t)$
    if $X = 1$ **then**
      $i^* = \text{argmax}_{i \text{ is alive}} \widehat{\mu}_i^t$
    else
      Sample i^* uniformly from all available arms
    Pull arm i^* and update sample mean

Algorithm 2 Pilot Sampling

Input: T budget, N arms, K pilot size

Set $t = 0$
while $t < T$ **do**
  Sample i uniformly from all available arms
  Pull arm i $\min\{T - t, K_i, L_i\}$ times
  $t = t + \min\{T, K_i, L_i\}$
  if Reward > 0 and $t < T$ **then**
    Pull arm i $\min\{T - t, L_i - K_i\}$ times
    $t = t + \min\{T - t, L_i - K_i\}$
  else
    Discard arm i
  Add all new infections as new arms

This method is described in Algorithm 2.[4] We discuss how to optimally identify the pilot group size and threshold in Appendix A.2.4.

Example 2.3.5 (How does Pilot Sampling work?)**.** *Suppose that mean rewards are* $\mu \in [0,1]$ *and* $\mu \sim \Gamma$*. Suppose that there are only two arms with unknown rewards* $\mu_1, \mu_2 \in [0,1]$ *and known lifetimes,* L_1, L_2*.*

Suppose, we set the parameters $K = 3$ and, as before, suppose that $\mu_1 = 0.1$ and $\mu_2 = 0.8$. Maybe we choose arm 1 first. We pull it $K = 3$ times and receive no reward. So, we move on to arm 2. Maybe we pull it $K = 3$ times and observe 2 rewards. Since we found at least one reward, we pull it until it dies. Then, we go to the next arm. If we are left with no more new arms, we can start playing arms from the discard pile. □

We now present a Bayesian regret bound for Pilot sampling in Theorem 2.3.6 with the technical discussion and proofs presented in Appendix A.2.

Theorem 2.3.6 (Bayesian regret of pilot sampling)**.** *Suppose that assumption (A1) and (A7) hold. Then,* $BR_{T,N}(Pilot) = \mathcal{O}(\min\{N, T\})$.

For $N \leq T^{\gamma/(\gamma+1)}$, $BR_{T,N}(\text{Pilot}) = \mathcal{O}(N)$. For $N > T^{\gamma/(\gamma+1)}$, running pilot sampling on a subset of $m = \Theta(T^{\gamma/(\gamma+1)})$ arms gives $BR_{T,N}(\text{Pilot}) = \mathcal{O}(T^{\gamma/(\gamma+1)})$. Thus, asymptotically, Pilot sampling achieves the same order as the lower bounds in Theorem 2.3.2. In particular observe that this does not have any log factors, unlike Adaptive Greedy.

Understanding Theorem 2.3.6 Pilot sampling is easier to analyze. In the worst case, we go through all N arms or we exhaust our budget but are not rewarded at all. It is important to emphasize that, besides logarithmic terms, both the Adaptive Greedy and the Pilot Sampling algorithms achieve the optimal lower bound in big-O asymptotics. That is, asymptotically, the difference between the Bayesian regret of Pilot Sampling and the optimal lower bound does not grow with N. This does not imply that it necessarily attains

[4] Chakrabarti et al. (2008) also describe a variant where an arm deemed to be highly rewarding can be discarded if its cumulative reward becomes too small. We do not consider it in our study.

the optimal lower bound – a point that we return to below. The exact asymptotic behavior has some dependence on the distribution of the lifetimes. Since we make no distributional assumption on the lifetimes, we don't see it in the big-O bound.

While these are highly favorable results for the algorithms we propose, in Appendix A.3, we also investigate when the asymptotic behavior described in Theorems 2.3.4 and 2.3.6 are achieved. We find that this generally occurs in the range of N between 10^3 and 10^6 depending on the distribution of mean rewards. The exception is when $\alpha < \beta < 1$ we need $N > 10^{12}$ before Adaptive Greedy achieves the asymptotic behavior. For these parameter values, the Beta distribution is mostly uniform, but rising in both tails (α controls the behavior on the left tail and β on the right tail). Fortunately, as we will discuss below, the size of the population N does not preclude real-world applicability in terms of dramatic efficiency gains.

2.3.5 Sampling Arms by Lifetime

An issue that we have not introduced thus far is that we expect and see in empirical data, that the number of contacts varies substantially between individuals. In particular, we often see a right-skewed distribution of contacts where a small fraction of individuals have a very large number and most have substantially fewer. We see this pattern in our empirical examples and it has also been documented extensively in work on measuring weak ties networks (for example McCormick et al., 2010; DiPrete et al., 2011). Through a minimal example, we motivate a variant of the algorithms where we sample arms proportional to their lifetimes instead of uniformly.

Example 2.3.7 (Sampling arms by lifetime)**.** *Suppose that mean rewards are $\mu \in \{0, 1\}$ and that $\mathbb{P}(\mu = 1) = p$. Then, pulling an arm just once will tell us what the mean reward is. If we pull arm i and obtain $X_i = 0$, then we know that the mean reward of that arm is $\mu_i = 0$ and if we obtain $X_i = 1$, then we know $\mu_i = 1$. Note that in this scenario, Adaptive Greedy sampling and Pilot sampling are identical. Suppose that there are only two possible lifetimes L_A and L_B with $L_A > L_B$ where $\mathbb{P}(L = L_A) = q$ and $\mathbb{P}(L = L_B) = 1 - q$. Suppose*

that there are only two arms with unknown rewards $\mu_1, \mu_2 \in \{0, 1\}$ *and known lifetimes,* $L_1, L_2 \in \{L_A, L_B\}$*. Let the budget be* $T < L_1 + L_2$ *(if* $T \geq L_1 + L_2$*, we get to pull both arms until their death and all policies are optimal).*

Suppose we pull arm i first. Then, we can show that the expected reward $\mathcal{R}_i$ (with respect to μ_1, μ_2) is

$$\mathcal{R}_i = p^2 T + p(1-p)\left(\min\{T, L_i\} + \min\{T-1, L_{1+|i-2|}\}\right)$$

Consider four policies π_1, π_2, π_3, and π_4. In π_1, we pull arm 1 first. In π_2, we pull arm 2 first. In π_3, arms are chosen uniformly at random i.e., with probability 0.5. And in π_4, we choose arms with probability proportional to the lifetime i.e., with probability $L_i/(L_1 + L_2)$. The expected rewards (with respect to μ_i) of these policies are

$$\begin{aligned}
\mathcal{R}(\pi_1) &= p^2 T + p(1-p)\left(\min\{T, L_1\} + \min\{T-1, L_2\}\right) \\
\mathcal{R}(\pi_2) &= p^2 T + p(1-p)\left(\min\{T, L_2\} + \min\{T-1, L_1\}\right) \\
\mathcal{R}(\pi_3) &= p^2 T + \frac{p(1-p)}{2}\left(\min\{T, L_1\} + \min\{T, L_2\} + \min\{T-1, L_1\} + \min\{T-1, L_2\}\right) \\
\mathcal{R}(\pi_4) &= p^2 T + \frac{p(1-p)}{L_1 + L_2}\left(L_1\left(\min\{T, L_1\} + \min\{T-1, L_2\}\right) + L_2\left(\min\{T, L_2\} + \min\{T-1, L_1\}\right)\right).
\end{aligned}$$

Without loss of generality, assume $L_1 > L_2$. It is easy to see that $\mathcal{R}(\pi_1) \geq \mathcal{R}(\pi_4) \geq \mathcal{R}(\pi_3) \geq \mathcal{R}(\pi_2)$. Therefore, sampling arms based on lifetime matters. The intuition is that it is more rewarding to learn the mean reward of an arm that lives longer because we can exploit it for a longer time. If we define $\kappa = L_2/L_1 < 1$, then κ represents the heterogeneity in lifetime – a smaller κ denotes a larger heterogeneity. As $\kappa \to 1$, $\mathcal{R}(\pi_4) \to \mathcal{R}(\pi_3)$ and as $\kappa \to 0$, $\mathcal{R}(\pi_4) \to \mathcal{R}(\pi_1)$. Essentially, π_4 represents the continuum between a policy that is agnostic to lifetimes (i.e., π_3) and a policy that greedily chooses arms with longer lifetimes (i.e., π_1). Thus, a larger heterogeneity results in a larger average reward when we sample by lifetime.

Now, we take the expectation of $\mathcal{R}(\cdot)$ with respect to L_i. Since $\mathbb{E}\mathcal{R}(\pi_1) = \mathbb{E}\mathcal{R}(\pi_2)$, we

will formulate π_1 as π_1', a policy that chooses the arm with the longer lifetime first, and remove π_2 from consideration. Simple calculations reveal that $\mathbb{E}\mathcal{R}(\pi_1') \geq \mathbb{E}\mathcal{R}(\pi_4) \geq \mathbb{E}\mathcal{R}(\pi_4)$. And a similar argument regarding heterogeneity as measured by L_B/L_A can be made: π_4 represents a continuum between π_3 and π_1'. □

Although the analysis, as in Example 2.3.7, becomes complicated when we have more arms or allow μ to be continuous, the underlying intuition remains the same: estimating the mean reward for an arm with a longer lifetime is more rewarding than estimating the mean reward for an arm with shorter lifetime with the same precision. This is because we can play the arm with the longer lifetime for a longer time. Further, the additional reward we gain from sampling based on lifetimes is more pronounced when the heterogeneity in lifetime is larger. This variant is obtained by modifying line 11 of Algorithm 1 and line 3 of Algorithm 2 to sample by lifetime (or degree) instead of uniformly i.e., $\mathbb{P}(\text{arm } i) \propto L_i$.

We've shown in the example above that the *expected* reward from an arm is higher if the arm's lifetime is longer (i.e. a person has more contacts). In our previous theoretical results, though, we've been concerned not with the expected performance but with the asymptotic big-O bound, which quantifies the order of the worst-case performance. In what might seem like a paradoxical result, Corollaries 2.3.8 and 2.3.9 say that the previously established asymptotics are unaffected. This insight can be very useful in settings where we do not have access to the lifetimes of arms as it says that, asymptotically, one method does not outperform the other in the worst case. We provide a technical discussion of this argument in Appendix A.1 and A.2.

Corollary 2.3.8 (Adaptive Greedy sampling by lifetime)**.** *Under the assumptions (A1)-(A5), for any $\epsilon \in (0, 1/3)$, the Bayesian regret of the adaptive greedy algorithm where we sample by lifetime obeys the same asymptotics in Theorem 2.3.4.*

Corollary 2.3.9 (Pilot sampling by lifetime)**.** *Suppose that assumptions (A2), and (A7) hold. Then the Bayesian regret of the pilot sampling algorithm when choosing arms by lifetime obeys the same asymptotics in Theorem 2.3.6.*

Corollaries 2.3.8 and 2.3.9 say that in the worst case, the asymptotic behavior of the Bayesian regret does not depend on whether we sample arms uniformly or by lifetime. This is because, in the worst case, we choose arms with the worst mean reward. That is, the worst-case scenario is one where we happen to choose people with very low infectivity (and thus few infections). The result is still finding few infections, regardless of whether the person has many contacts or few. Rather than focusing on the number of contacts (which could be logistically challenging to obtain in practice anyway), the crucial idea is to learn the mean reward of an arm quickly, regardless of how it is chosen, to help us decide whether to exploit that arm or explore other arms.[5]

2.3.6 Summary of Theoretical Results

We first showed that the theoretical lower bound on the Bayesian regret of standard bandits, established by Bayati et al. (2020), extends to mortal bandits in Theorem 2.3.2. Theorem 2.3.4 then showed that the upper bound of Adaptive Greedy sampling achieves this lower bound up to some logarithmic factors in big $\mathcal{O}$. Theorem 2.3.6 showed that Pilot sampling achieves the lower bound in big $\mathcal{O}$. In Example 2.3.7, we demonstrated that sampling by lifetime can lead to a higher expected reward than when sampling arms uniformly. However, Corollaries 2.3.8 and 2.3.9 show that sampling based on lifetime does not change the asymptotic behavior.

Next, we focus on empirical simulations of algorithmic performance for different distributions of mean rewards and lifetimes. This addresses three issues. First, as is well understood, matching orders in big $\mathcal{O}$ does not mean that one algorithm is equally efficient as the other. This is evident from Example 2.3.7. Second, finite sample behavior (in terms of N) can be very different from asymptotic behavior (see Appendix A.3). Third, from a perspective based

[5]It is worth noting that focusing search on arms that live longer has been brought up in earlier research. Tracà et al. (2020) describe a similar idea in mortal bandits by exploring arms that live longer. They restrict the exploration phase in Adaptive Greedy to arms that are in the top k% of the distribution of the remaining lifetimes and they tune k using a subset of available data. In contrast, our approach is free of hyperparameters and also works for Pilot sampling.

purely on asymptotic behavior, it is unclear if there is value in learning the distribution of the mean rewards i.e., γ, and the lifetimes.

2.4 Simulations

We perform a series of numerical experiments to evaluate the performance of the algorithms described in Section 2.3. We also compare these algorithms with a baseline naive sampler that picks an arm uniformly at random and pulls it until it dies, effectively not performing any learning, and with the widely used Thompson sampling algorithm, which we discuss in Appendix A.4. In particular, Thompson sampling has been shown to enjoy the best Bayesian regret bounds for a variety of model classes in the stochastic (non-mortal) setting (Russo and Van Roy, 2014).

2.4.1 Simulation Setup

For generating synthetic data, we fix N arms with prior parameters for mean reward (α, β). For each arm i, we draw mean reward $\mu_i \sim \text{Beta}(\alpha, \beta)$ and lifetime $L_i \sim F$, where the choice of F is described below. We randomly choose $X_i \sim \text{Binomial}(L_i, \mu_i)$ pulls as rewards. We set a total budget of T pulls.[6] For prior distribution on the mean reward, we choose $(\alpha, \beta) \in \{(0.09, 0.6), (1, 3), (1, 1), (10, 10), (3, 1), (0.6, 0.09\}$. This allows us to vary the skew of the mean rewards, which range from distributions that have some highly infective people (first and fifth) to those with a uniform or normal PCI distribution (third and fourth) and those where most people are not infective (the last). In Figure A.6, we visualize the densities of these distributions. Regardless of the choice of (α, β), the priors for all Thompson sampling simulations were initialized at $\text{Beta}(1, 1)$, a uniform prior. For the average lifetime, we chose two different families of distribution, zero-truncated Poisson and Pareto. We chose Poisson because it is a relatively homogeneous distribution with the same mean and variance. We chose the mean to be 500. We chose Pareto as the other distribution as it is a heavy-tailed

[6] If we have $\sum_{i=1}^{N} L_i < T$, then we add new arms to the data until $\sum_{i=1}^{N} L_i > T$.

distribution. For Pareto, we fixed the location as 1 and chose the shape = 0.6. Based on the literature on social networks, we expect that the degree distribution across the population will be heavily right-skewed (see Newman and Park, 2003; McCormick et al., 2010; DiPrete et al., 2011).

The key result we will show is that when the degree distribution is heavy-tailed or when the PCI distribution is right-skewed, Pilot Sampling yields significant advantages over either the Adaptive Greedy or Thompson sampling algorithms.

Simulation Results

First, consider the case where lifetimes are Poisson distributed. We set the initial number of arms $N = 10$, the average lifetime to $\lambda = 500$, and the budget to $T = 50,000$. The results from these simulations are shown in Figure 2.1. For a heavily right-skewed distribution of the mean reward (Beta$(0.09, 0.6)$, Thompson and pilot sampling perform the best. For a distribution with low variance (Beta$(10, 10)$), Thompson and adaptive greedy perform the best. For a distribution that is not heavily skewed or does not have a small variance, all three methods perform roughly the same. Finally, for a left-skewed distribution (Beta$(3, 1)$, Beta$(0.6, 0.09)$), Thompson and adaptive greedy perform the best. Overall, it seems like Thompson is consistently doing well in all situations while the other two algorithms fail in some extreme cases. While some of these conclusions agree with the bounds presented in Figure A.1, it is clear that the bounds do not tell the full story. We perform similar simulations with $\lambda = 10$ and $T = 1000$ in Appendix A.5.4.

Next, we used a heavy-tailed lifetime distribution. The lifetime was drawn from a Pareto distribution with shape 0.6 and location 1 (so the expectation diverges). Here, we fix the number of arms $N = 5000$. The results are shown in Figure 2.2. Note that pilot sampling consistently dominates the other algorithms. This is because pilot sampling commits to exhaustive testing when it finds a positive in the pilot group. In heavy-tailed scenarios, the regret in committing to a bad arm is small: (i) we choose an arm with an average lifetime and are done quickly or (ii) we choose an arm with a very high lifetime and are proportionally

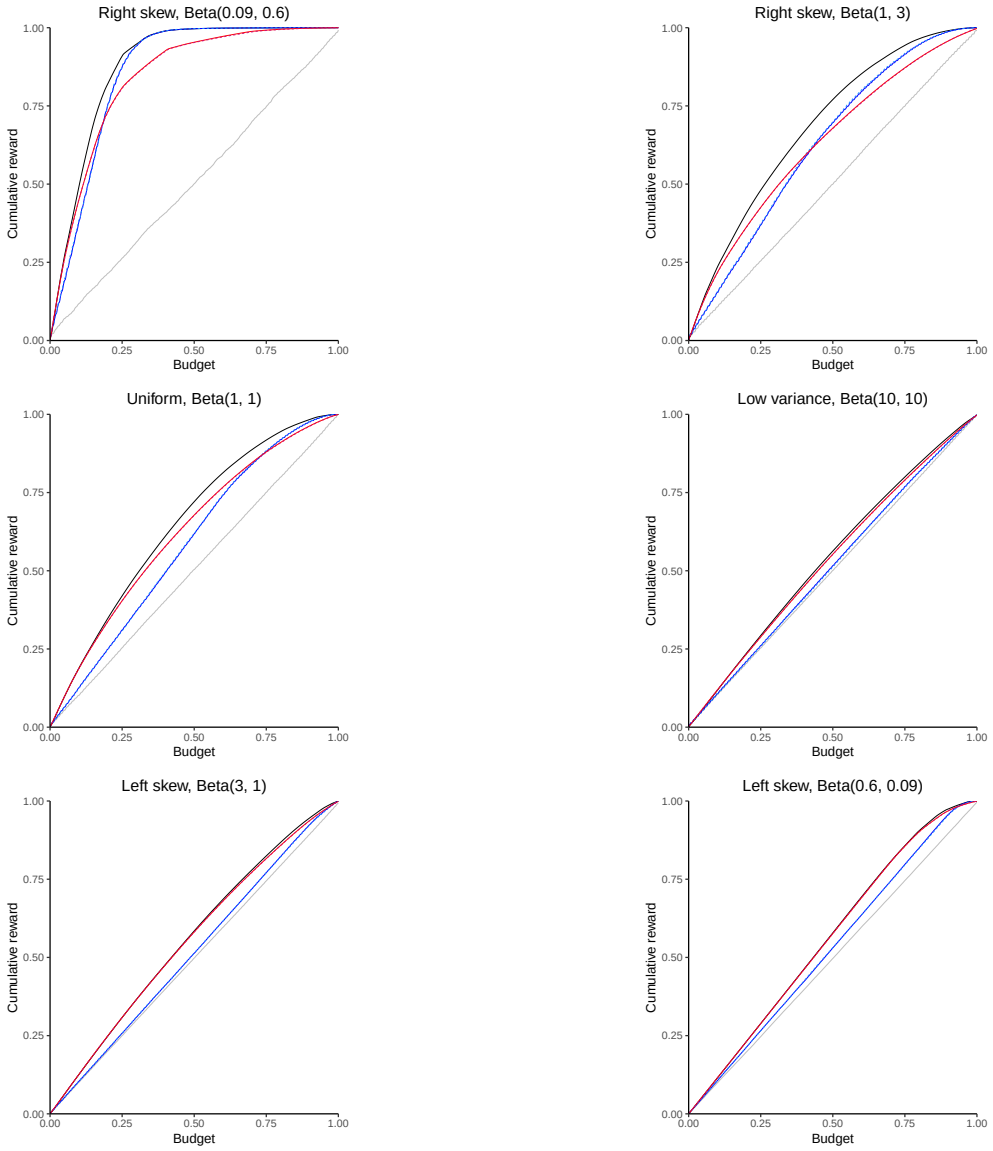


Figure 2.1: Cumulative reward over the total time horizon for different policies and various reward distributions based on data simulated without branching and Poisson(500) lifetime. The axes are normalized to facilitate visual comparison. Thompson sampling is in black; pilot sampling with uniform sampling and lifetime sampling are in dark blue and light blue, respectively; adaptive greedy with uniform sampling and sampling by lifetime are in red and pink respectively, and; naive sampling is in grey. Pilot sampling performs better when the rewards are heavily right skewed while adaptive greedy performs better in other scenarios. Thompson sampling appears to perform consistently well in all scenarios. In this setup, sampling by degree seems identical to sampling uniformly.

given a larger reward. In the Poisson case with no heavy tail, the loss in committing to a bad arm is relatively the same for all arms. Meanwhile, other algorithms are afraid of committing to exploit unless they are very sure. So, they continue searching for a good arm for a long time. This is also the same reason why we see Naive sampling beat other methods in some cases. But since it does not intelligently choose to commit arms, Pilot sampling performs better. Also, observe that the sampling-by-lifetime variant of Adaptive Greedy dominates the lifetime-agnostic variant.

We see two main patterns in these simulations. First, as the mean rewards become more and more right-skewed, there is little difference in how all the methods compare against each other. This is due to the fact for a right-skewed distribution, the probability that we choose an arm with a high reward is larger than in the left-skewed case. So all methods perform relatively similarly in this case.

Second, Adaptive Greedy performs better when the mean rewards have low variability. For example, compare Beta (1, 1) and Beta (10, 10) in Figures 2.1 and 2.2. The reason is that Adaptive Greedy learns the mean reward of an arm by pulling it once. Thus, it takes more pulls in a high-variance setting than in a low-variance setting to learn the mean reward of the arm with the same precision. Compare this to Pilot sampling where we pull an arm multiple times to get a more precise estimate of its mean reward. In a high-variance setting, Pilot sampling has an edge over Adaptive Greedy which can spend too much time exploring. In a low variance setting, Pilot sampling's gain in this precision by pulling an arm more than once is not substantial. Therefore, Pilot sampling can waste resources by making these unnecessary additional pulls that Adaptive Greedy does not. This difference does not seem to matter in cases where the lifetimes have a heavy tail but does become important when the distribution of the lifetime does not have a heavy tail.

At first blush, these results appear to contradict our regret bounds in Theorems 2.3.4 and 2.3.6. Based on our discussion in Section 2.3, most of our simulations, we appear to be in the regime where asymptotic behavior holds (also see Appendix A.3) and the asymptotic behavior says that Adaptive Greedy should perform worse than Pilot sampling which, based

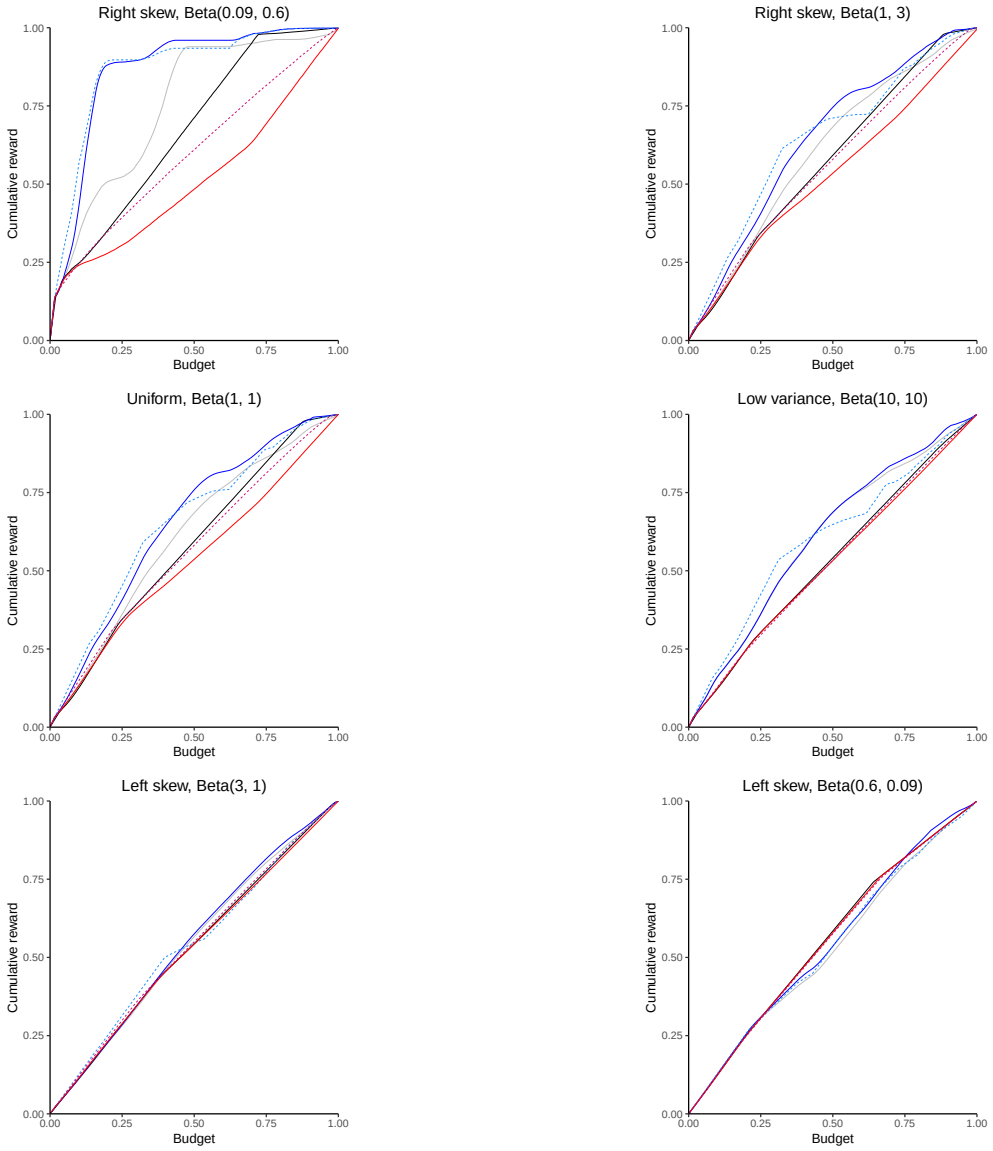


Figure 2.2: Cumulative reward over the total time horizon for different policies and various reward distributions based on data simulated without branching and Pareto(1, 0.6) lifetime. The axes are normalized to facilitate visual comparison. Thompson sampling is in black; pilot sampling with uniform sampling and lifetime sampling are in dark blue and light blue respectively; adaptive greedy with uniform sampling and sampling by lifetime are in red and pink respectively, and; naive sampling is in grey. Pilot sampling outperforms all other methods in all scenarios except when rewards are heavily left-skewed. In pilot sampling, it is unclear whether sampling uniformly or by lifetime is better. For adaptive greedy, it is clear that sampling by lifetime dominates sampling arms uniformly.

on simulations, is generally not true. In fact, this is only true in the heavy-tailed or heavily right-skewed cases. The reason this is not contradictory is that the asymptotic behavior only reflects the upper bounds of the Bayesian regret. This is a scenario where upper bounds are not indicative of the average-case behavior. We note that (Bayati et al., 2020) also report very similar results for the standard bandit. They show that even though a greedy algorithm does not achieve universal rate optimality, it performs extremely well in practice.

Guidelines for choosing a sampler The insights from these simulations are summarized in these guidelines: If the degree distribution is heavy-tailed, we should use Pilot sampling. If the degree distribution is not heavy-tailed, then we should use Pilot Sampling if the rewards are heavily right-skewed; use Adaptive Greedy or Thompson Sampling if rewards have a small variance or are heavily left-skewed, and; the choice does not matter in other cases. If the degree distribution is not heavy-tailed and we do not know how the rewards are distributed, we should default to Thompson Sampling as it consistently performs well. In Appendix A.5.2, we show that these findings hold true in scenarios where new arms can appear. In Appendix A.3, we provide additional discussion of the asymptotic behavior and empirical simulations.

2.5 Results on COVID-19 Contact Tracing

A fundamental insight from the previous sections is that the algorithms we have proposed have similar asymptotic bounds, but the mean performance depends critically on the shape of the degree and PCI distribution, which are both empirical quantities. In the last section of the chapter, we therefore turn to the data to estimate these distributions. Using these data, we implemented the different sampling policies described in Section 2.3. These datasets were collected as a part of administrative contact tracing efforts during the outbreak of COVID-19 in India and Pakistan. The first dataset was collected from Punjab (Pakistan), the second dataset was collected from Punjab (India), and the third dataset was collected from southern India (from parts of Andhra Pradesh and Tamil Nadu). In Appendix A.6, we summarize the

properties of these datasets.

Prior to discussing our estimations, we must acknowledge that our datasets are not perfect, and there are a number of data gaps that we cannot address. For instance, the data collected from Punjab, Pakistan exhibits low infectivity because at the time of collection, 485,853 people were still awaiting test results and, given low infectivity, we treat these individuals as healthy in our simulations. Similarly, the data collected from southern India contains only summary-level information i.e., total counts of traced and infected people, rather than a 'line-listing' of each individual contact and whether they were infected. Therefore, it was not possible to trace the infected individuals i.e., contact tracing does not proceed as a branching process. Finally, none of these data come from prospective studies with careful lab studies that can help establish the progeny of an infection. This implies, for instance, that if B is a contact of A and B is now infected, we will follow the contact-tracing line and assume that the causal infection link went from A to B, and not because A and B both were infected from a different source, or because B infected A. This is a strong assumption, but it is likely to hold in the dataset from Punjab, India, that was carried out through the period of a stringent lockdown with very limited outside contact.

Our idea then is not necessarily to demonstrate the value of our methods in a carefully prospective study, but rather to assess whether, in very different datasets from different settings and policies, we obtain similar results in terms of the shape of the degree and PCI distribution. To the extent that we do, it increases our confidence in the underlying estimates.

2.5.1 Estimating Individual Heterogeneity in Infectivity

We estimated the parameters of the Beta model for the infectivity of the population using the full dataset with a Bayesian shrinkage estimator as in Arinaminpathy et al. (2020).

An immediate estimator of PCI for person i, denoted by μ_i, would be the ratio of infected, z_i, to the total number of people, d_i, that i came into contact with i.e., $\widehat{\theta}_i = z_i/d_i$. However, since the distribution of degree is skewed, this naive estimator would have a different variance

for each individual as the total number of contacts changes. Instead, we will use a Bayesian shrinkage estimator following Arinaminpathy et al. (2020). Therefore, the individual PCI estimates for high-contact individuals will remain mostly unchanged while those of low-contact individuals will be shrunken towards the overall mean.

In particular, we model the log odds of the individual PCI as following a normal distribution with a common mean and variance,

$$\mu_i = \text{logit}^{-1}(\theta_i) = \frac{1}{1+e^{-\theta_i}}$$
$$\theta_i \sim N(\overline{\theta}, \sigma_\theta^2)$$

where θ_i is the log-odds of μ_i and $\overline{\theta}$ is the overall mean. The variance σ_θ^2 is inversely proportional to the shrinkage. As $\sigma_\theta^2 \to 0$, $\mu_i \to \overline{\theta}$ and as $\sigma_\theta^2 \to \infty$, $\mu_i \to z_i/d_i$. These hyperparameters are estimated using Monte Carlo Markov Chain methods with diffuse priors. Since $\overline{\theta}, \sigma_\theta^2 > 0$, μ_i is guaranteed to be between 0 and 1. We refer the reader to the Arinaminpathy et al. (2020, Supplementary) for more details.

We visualize the distribution of infectivity as estimated using Bayes shrinkage in Figure 2.3.

2.5.2 Results on Contact Tracing

We used the PCI distribution estimated previously to initialize Thompson sampling and determine the pilot group size for mortal bandits. Although we used the full dataset to estimate the pilot group size, we note that the group size is not very sensitive to the prior distribution, especially in the right-skewed cases. Based on these parameters and sizes of the datasets, we are in the regime where the asymptotic behavior described in Theorems 2.3.4 and 2.3.6 hold (see Appendix A.3).

The results are shown in Figure 2.4. The results were averaged over 100 simulations to account for randomly drawing infected people and their contacts. Here, Naive sampling is a straight line because it is equivalent to testing every single person and the rate of uncovering

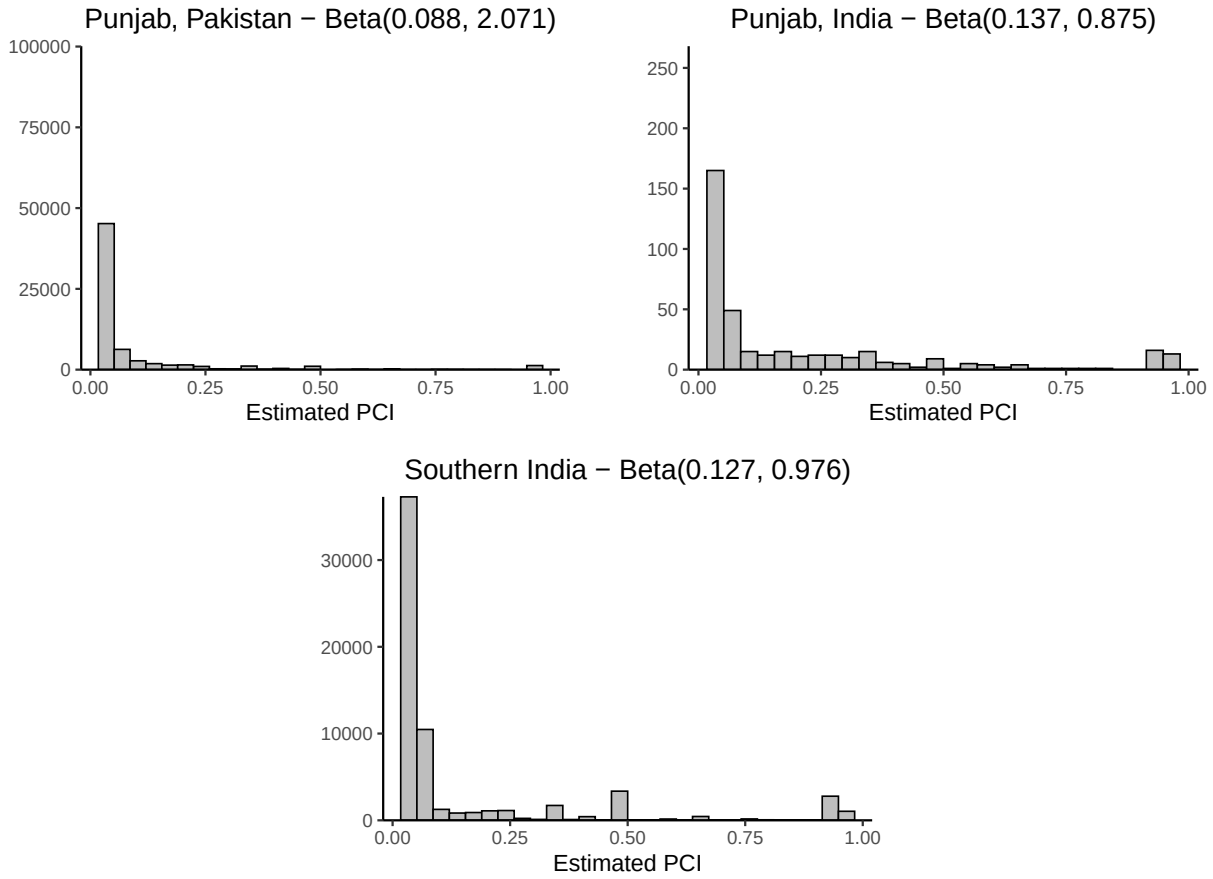


Figure 2.3: Distribution of estimated PCI from the three different datasets. PCI was estimated using a Bayes shrinkage estimator. The histogram of the PCI is shown in grey. Using the estimated PCI, parameters of a Beta distribution were fit using the method of moments. As we can see, the Beta distribution is heavily right-skewed.

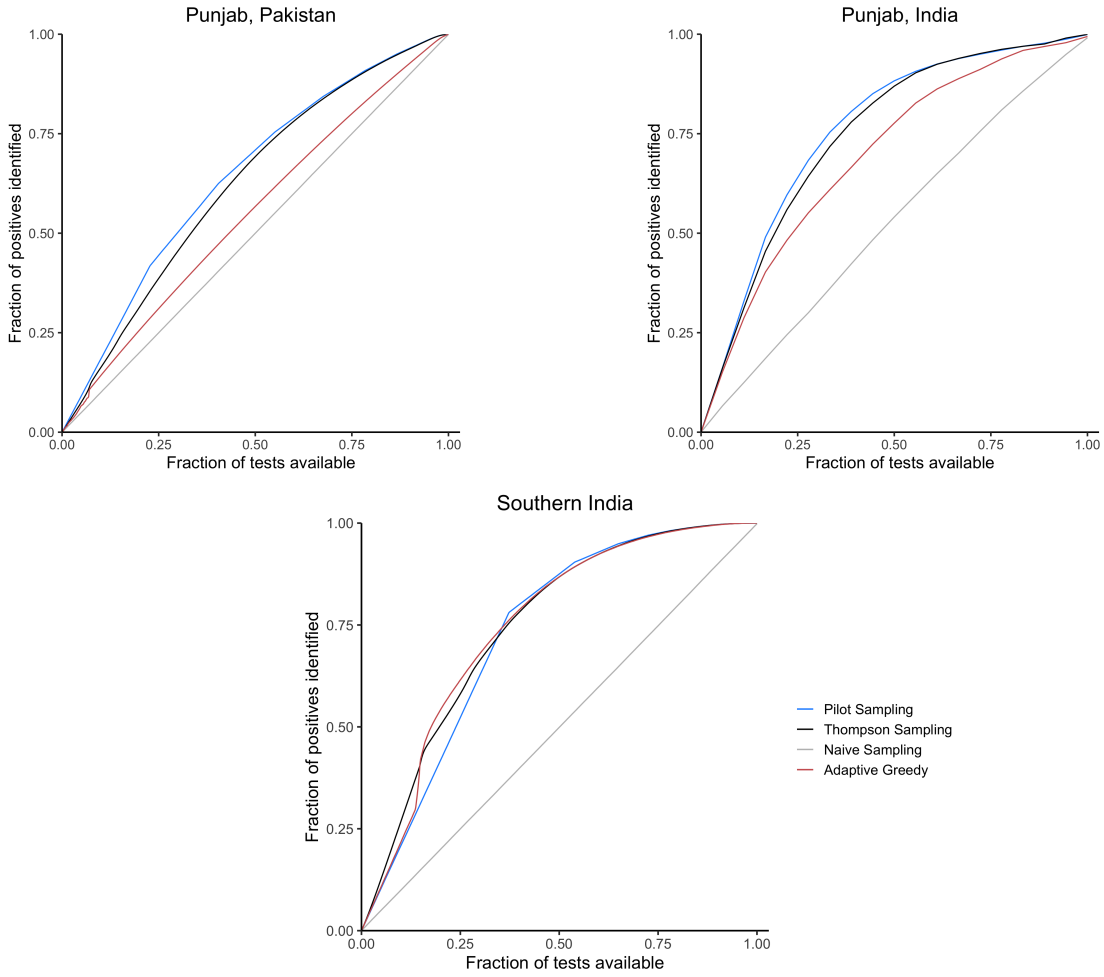


Figure 2.4: The top left figure shows results on Punjab, Pakistan dataset. The top right figure shows results on the Punjab, India dataset. The bottom figure shows results on the southern India dataset. Pilot sampling and Thompson sampling are clearly doing better when we have branching data (both Punjab datasets). However, Thompson sampling may be logistically difficult making pilot sampling favorable when implementing contact tracing. Figure A.12 shows the same plot with absolute numbers of tests and infections.

Algorithm	Punjab, Pakistan	Punjab, India	Southern India
Naive Sampling	80%	78%	80%
Thompson Sampling	**62%**	**39%**	42%
Adaptive Greedy	75%	50%	41%
Pilot Sampling	**62%**	**39%**	**40%**

Table 2.1: Fraction of tests performed in order to identify 80% of all infections.

infections reflects the proportion of infections in the full population. As expected, Pilot sampling, Thompson sampling, and Adaptive Greedy vastly outperform Naive sampling. Pilot sampling and Thompson sampling perform at similar levels with Pilot sampling dominating both the Punjab datasets. Adaptive Greedy does not perform well here. On the other hand, in the dataset from southern India, all three algorithms perform similarly. In Table 2.1, we look at how many people need to be tested to identify 80% of all infections. The advantages are quite dramatic – the use of Pilot sampling would have allowed teams to pick up 80% of the infected individuals in the samples with only 40% of the tests in Pakistan and South India and 62% of the tests in Punjab, India compared to 80% of tests with Naive sampling.

This demonstrates that the conclusions we drew from controlled simulations do in fact hold up when confronted with data collected during the COVID-19 pandemic. Clearly, Thompson sampling and Pilot sampling have an edge over Adaptive Greedy. These findings along with simulations in Section 2.4 strongly suggest that we should abandon Naive sampling in favor of more efficient methods.

One reason why governments may be reluctant to try active learning algorithms is because of their complexity and other logical constraints. For instance, door-to-door testing makes Thompson sampling and Adaptive Greedy approaches hard to implement in the field but may be more amenable to phone call testing. What makes our results particularly appealing is that in all three datasets, the distribution of PCI and degree is such that pilot sampling either performs better or very similarly to other more complex algorithms. Pilot sampling can be easily implemented in the field and offers an opportunity for governments to engage

in contact tracing that they can actually do.

2.6 Discussion

In this chapter, we addressed the problem of efficient contact tracing by framing it as a mortal bandit problem. We showed that the lower bound for the Bayesian regret in standard and mortal bandits are identical, and presented new Bayesian regret bounds for the Adaptive Greedy and Pilot Sampling algorithms. Through empirical simulations, we provide guidelines for choosing appropriate policies. If the distribution of the lifetime of arms is heavy-tailed, then we should use Pilot Sampling. If the distribution of lifetime is not heavy-tailed, then we should use Pilot Sampling if the rewards are heavily right-skewed; use Adaptive Greedy or Thompson Sampling if rewards have a small variance or are heavily left-skewed, and; the choice does not matter in other cases. If the distribution of lifetimes is not heavy-tailed and we do not know how the rewards are distributed, we should default to Thompson Sampling. We use our theoretical results and findings from empirical simulations on data from COVID-19 contact tracing in three different regions – Punjab (Pakistan), Punjab (India), and South India. We show that Pilot Sampling outperforms Adaptive Greedy in both of the Punjab datasets and performs similarly to Adaptive Greedy in the dataset from southern India. These results are in line with the results from the empirical simulations.

We outline three possible extensions along with their relevance for contact tracing. First, we assumed that the distribution of mean rewards (PCI) is known. However, the distribution of PCI is unlikely to be known at the beginning of the epidemic. Therefore, contact tracing serves the dual purpose of controlling the epidemic through contact tracing and estimating the distribution of PCI in the population. Thus, the value of information regarding the distribution of PCI is very high in the initial stages of the epidemic – performance may be higher if all contacts are tested until the PCI distribution is known with low uncertainty.

Relatedly, we claimed that Pilot Sampling and Adaptive Greedy are able to match the order lower bounds when we run the policies on an appropriately sized subset of arms. The size of the arms depends on the γ parameter. In practice, this may be unavailable. For

example, during a new outbreak, we may not know the behavior of the disease. There are a variety of estimators with desirable convergence properties available from extreme value theory that help estimate β (for example, see Hill, 1975; Pickands III, 1975; Haan and Ferreira, 2006). The situation with bandits is slightly different as we never really observe the true mean reward of each arm. Carpentier and Valko (2015) tackle this case for the standard bandit. They propose a modification to the estimator of Carpentier and Kim (2015) and describe a two-phase algorithm that first estimates γ and then calls a different policy.

Second, we have disregarded any potential correlation between the number of contacts and PCI. If PCI is purely biological, this may not be a poor assumption, but if PCI is, in part behavioral, both positive and negative correlations may occur. Estimating this correlation with any precision requires considerable data. It will be useful to compute the sample size requirements in order to estimate this with sufficiently high power. After this correlation is estimated (or is known a priori), a natural question is the effect of this correlation on choosing an optimal contact tracing policy.

Third, we have abstracted from the full network structure thus far, in part because we just don't see how such a structure can be learned in the throes of an epidemic. However, it will clearly matter. Suppose one person's PCI based on data accumulated thus far is 20% and another's is 100%. Each has 5 remaining contacts to trace. However, the person whose PCI is 20% has one contact who regularly meets thousands of people, while the contacts of the person whose PCI is 100% each have no further contacts. Clearly, tracing and quarantining the contacts of the 20% person will still be the most effective action in this case. It will be beneficial to assess the extent to which a lack of knowledge of the full structure of the network graph affects our results.

Chapter 3

ROBUSTLY ESTIMATING HETEROGENEITY WITH RASHOMON PARTITIONS

> "That's the trouble with science. It's never done. Always upending itself. Ruining perfect systems for the little inconvenience of them being wrong."
>
> — Zahel, *Rhythm of War*

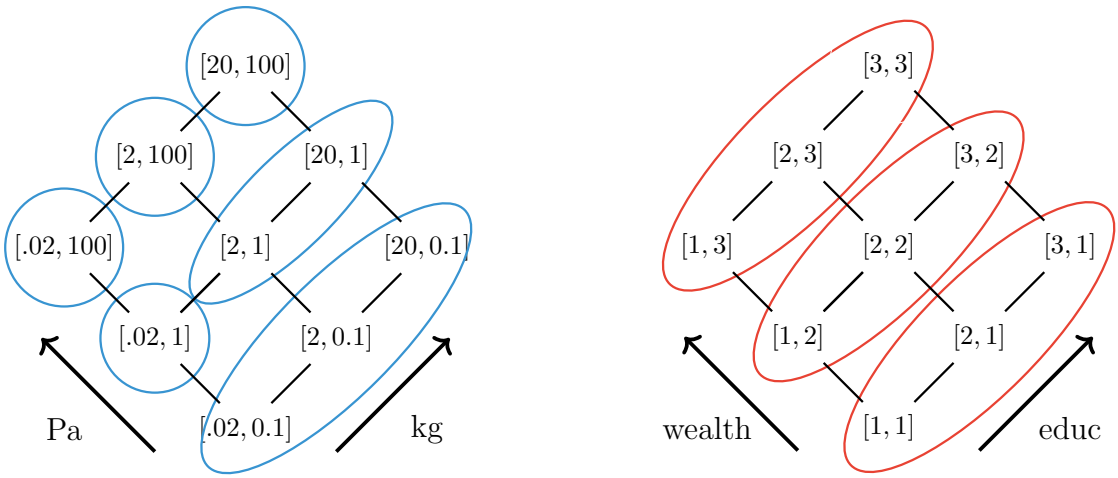
This chapter is adapted mainly from Venkateswaran et al. (2024) (currently under review and available on arXiv). It is joint work with Anirudh Sankar, Arun Chandrasekhar, and Tyler McCormick. We gratefully acknowledge Alberto Abadie, Isaiah Andrews, Abhijit Banerjee, Emily Breza, Kevin Chen, Paul Goldsmith-Pinkham, Rachel Heath, Muriel Niederle, Ashesh Rambachan, Cynthia Rudin, Rahul Singh, Davide Viviano, and Bo Zhang for their helpful discussions. We thank Garrett Allen and Jessica Kunke for their feedback on earlier versions of this chapter. We thank Brian Xu for exceptional research assistance.

In Chapter 2, we saw that working with the near-optimal instead of the optimal was practically valuable. In this chapter, we wrestle with the idea of optimality head-on. What does it mean for a model to be the "best fit" if the second best fit is statistically indistinguishable, i.e., the difference in loss or posterior probability is negligible? This is a problem when working with finite data even when the true model is identified, and it is called the Rashomon effect (Breiman, 2001b). Then, how does one reliably draw conclusions from data? How does the scientist choose the true model from a collection of near-optimal ones? We argue that, in light of the Rashomon effect, selecting a single model is a futile task. Instead, we should use all such high-quality models together to learn about the world and make robust decisions. We provide the machinery to do so. In three different contexts, we show how robust the existing findings are to the Rashomon effect. (Spoiler alert: not all are robust.)

3.1 Introduction

Researchers and policymakers often study settings where an outcome of interest varies with combinations of features or covariates (e.g., characteristics, treatment assignments) of a given unit. Examples include (1) learning what combination of drugs, at what frequency and dosages, and for what sub-groups reduce a given illness (e.g., in the cases of HIV and non-small cell lung cancer, Hammer et al. (1997); Cascorbi (2012); Nair et al. (2023)); (2) studying how an individual's wage is associated with combinations of age, education, parental wealth, race/ethnicity, and gender (Mincer, 1958; Aakvik et al., 2010; Forster et al., 2021); (3) analyzing vaccination campaigns that leverage incentives, reminders, network strategies across the wealth distribution (Chernozhukov et al., 2018; Banerjee et al., 2021), and; (4) determining when and why microfinance is more effective for certain sub-populations and markets (Banerjee et al., 2019).

Fundamentally, the researcher wants to learn a response function that describes how the outcome changes (or doesn't) when moving between levels of a feature, but doesn't know how sensitive the response function is to changes in covariate combinations. Both modern and classical statistical tools, either implicitly or explicitly, address this problem using *partitions*. They partition observations into "pools" where outcomes are similar within the pool but differ across pools, then compute a summary (or fit a model) to pool. Some models, such as Bayesian or frequentist tree models, are explicit about these partitions. Others, however, do so implicitly (e.g. a regression with a single binary covariate posits heterogeneity between people who in one group versus the other, and homogeneity otherwise). Banerjee et al. (2021) introduced Hasse diagrams as a geometric representation of partitions, an idea we substantially extend here. Figure 3.1 gives an example of two partitions. The Hasse defines partitions by removing (splitting on) edges to form disjoint connected components, which guarantees that all sets in a partition contain only "connected" feature combinations. Distinguishing meaningful from spurious heterogeneity then amounts to evaluating partitions. Estimation strategies and algorithms privilege different partitions in search of the partition



(a) e.g., $\mathbb{E}[y_i|(2,1)] = \mathbb{E}[y_i|(20,1)]$

(b) e.g., $\mathbb{E}[y_i|(2,3)] \neq \mathbb{E}[y_i|(2,2)]$

Figure 3.1: Two partitions, each representing a distinct model for heterogeneity in the outcome, y_i. The left panel shows heterogeneity in acceleration of a cube measured after dropping it a uniform gravitational field with drag as the mass of the cube and external pressure changes. The right panel shows Banerjee and Duflo (2010)'s model for interest rates as a function of borrower's wealth and education when there are high administrative costs relative to loan amount.

that captures meaningful complexity without sacrificing power, but the partitions capture the root of the heterogeneity directly.

In this paper, we propose *Rashomon Partition Sets (RPSs).* The RPS consists of all partitions that are close to the *maximum a posteriori (MAP)* partition in terms of posterior density. We can bound the difference between posterior quantities computed using the entire posterior and using only the RPS. We use an ℓ_0 prior, which we show is minimax optimal but does not impose additional restrictions on the association between covariates. Restrictions on the universe of partitions ensure that each partition corresponds to a scientifically plausible explanation. When combined, our prior and restrictions on the space of partitions, both of which are motivated scientifically, substantially improve computational efficiency, making it possible to enumerate the entire RPS. Conclusions using RPS, then, incorporate all partitions that are near the MAP, even if they offer substantively very different explanations, without

specifying the nature of associations between features. For experiments, policymakers can then weigh additional considerations (e.g., cost, equity, privacy) in choosing which policies from the RPS to implement. RPSs also yield insights to generate new scientific theories. Looking across models in the RPS, one can build an archetype of feature combinations that appear consistently and have consistent effects on the outcome, regardless of the structure imposed on other covariates by other high posterior partitions.

Rashomon alludes to Breiman (2001b)'s "Statistical Modeling: The Two Cultures" paper that describes "a multitude of different descriptions [equations $f(x)$] in a class of functions giving about the same minimum error rate" as the *Rashomon Effect*. In situating the RPS within existing literature, we consider two features: (i) building the RPS and (ii) the construction of partitions eligible for inclusion in the RPS, known as *permissible* partitions. We provide a high-level overview here and then a detailed literature review in Section 3.9.

First, we consider the construction of the RPS, taking as given for now that we have a set of partitions. Our goal is to represent uncertainty amongst scientifically plausible explanations of heterogeneity in an outcome. Scientifically plausible explanations should be supported by the observed data. RPSs, therefore, are *enumerative*, meaning that they consist of a list of *all* partitions with posterior density close to the MAP partition. In their seminal paper developing an *Occam's Window* approach to Bayesian Model Averaging (BMA) for graphical models, Madigan and Raftery (1994) express a similar philosophy:

> [standard BMA] does not accurately represent model uncertainty. Science is an iterative process in which competing models of reality are compared on the basis of how well they predict what is observed; models that predict much less well than their competitors are discarded. Most of the models in [standard BMA] have been discredited [...] so they should be discarded.

Madigan and Raftery (1994) use this logic to justify an approach that favors high posterior, simple models but constructs this set by sampling from the posterior. Sampling includes both models that are consistent with observed data and others that are highly unlikely, with

the latter comprising most of the posterior mass (Moulton, 1991). Our approach breaks from the literature that relies on sampling, instead using the MAP as an anchor and listing models with similar posterior density. Further, sampling explores the space of partitions/models, whereas the domain of science is explanations. Without restrictions on the space, there could be multiple partitions that correspond to a single explanation, and the number of partitions corresponding to each explanation will depend on the complexity of the explanation. This is particularly true for models like decision trees that split hierarchically amongst variables with no ordering (e.g., splitting a tree on education or income first is arbitrary and scientifically unimportant, though it generates multiple distinct partitions to sample). Sampling in the space of partitions, then, weights scientific explanations differentially based on the number of partitions corresponding to that explanation. To avoid this issue, we propose a set of rules for constructing partitions that ensures that each partition corresponds to a scientific explanation.

Second, we describe our principles for creating partitions. Ex-ante, with many variables and many feature combinations, the number of partitions is both enormous and unstructured (pooling the usage of 200mg of amoxicillin alone with 100mg ibuprofen alone, for example, is scientifically incoherent since the drugs act in completely different ways). We restrict ourselves to partitions that are scientifically interpretable using rules that we call *permissibility conditions*. We specify one such permissibility framework, but the details may be context dependent. In constructing our permissibility rules, we think of the world in increments: how changing a level in one variable marginally affects the outcome. The overall effect of some feature combination, then, comes from adding these marginal effects. This perspective leads us to define variants which are otherwise identical feature combinations that only differ by one feature by a single intensity value, taking one step up the Hasse. We complement marginal effects with *conditional independence* of pools to capture higher order dependence. For example, as seen in Figure 3.1, the structure of heterogeneity in acceleration under 100 Pa external pressure is different from 0.02 Pa. Conditional independence of pooling gives flexibility to capture this structure.

The general principle is that enforcing permissibility allows us to use the geometry of the factorial space to substantially reduce computation, while also ensuring that all partitions in the RPS are actually viable explanations for heterogeneity. This contrasts with other approaches that use tree-based partition structures. Xin et al. (2022), for example, explore the Rashomon set over decision trees as multiple partitions could correspond to the same explanation. This multiplicity is especially pronounced when using decision trees since they impose a false hierarchy over variables that are actually partially ordered. The remainder of the paper expands on these two tasks – defining partitions and constructing the RPS. In Section 3.2, we define the RPS formally. Then, we give statistical properties for an arbitrary set of partitions and introduce our minimax optimal ℓ_0 penalty in Section 3.3. We then give a formal definition of our robust partition structure in Section 3.4. We show that this combination allows us to bound the size of the RPS in Section 3.5 and enumerate it entirely in Section 3.6. Section 3.7 provides simulation evidence and Section 3.8 gives three empirical examples, highlighting robust archetypes in each setting. Finally, Section 3.9 and 3.10 provide a discussion of related work and future directions, respectively. All of our code is available at https://github.com/AparaV/rashomon-partition-sets.

3.2 Rashomon Partition sets

Suppose that there are n units (or individuals) and each has M features. The feature matrix is given by $\mathbf{X}_{1:n,1:M}$ and outcomes are $\mathbf{y} \in \mathbb{R}^n$. Every feature has R possible values, partially ordered. Let $\mathcal{K}$ be the set of all $K = R^M$ unique feature combinations. Depending on the context, we let k denote the vector of feature values or its index in $\mathcal{K}$. Each combination of features $k \in \mathcal{K}$ can be represented in a dummy binary matrix $\mathbf{D}$ with entries $D_{ik} = 1$ if and only if observation i has feature combination k. The dataset is $\mathbf{Z} := (\mathbf{y}, \mathbf{X})$. So, the researcher studies

$$\mathbf{y} = \mathbf{D}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{3.1}$$

where $\beta_k = \mathbb{E}[Y_i \mid D_{ik} = 1]$ is the expected outcome in the population given the feature combination and ϵ_i is some idiosyncratic mean-zero residual. A *partition*, Π, in the space of all partitioning models, $\mathcal{P}$, is a model of heterogeneity such that for every pool $\pi \in \Pi$, possibly a singleton, if feature combinations $k, k' \in \pi$, then $\beta_k = \beta_{k'}$. The posterior given the data $\mathbf{Z}$ is $\mathbb{P}(\Pi \mid \mathbf{Z})$. Let $\mathcal{P}^\star \subseteq \mathcal{P}$ be the set of permissible partitions that obey some permissibility rules (to be defined in Section 3.4). The RPS, then, is defined as follows.

Definition 3.2.1 (Rashomon Partition Set (RPS))**.** *For some posterior probability threshold* $\tau \in (0, 1)$*, we define the* Rashomon Partition Set *relative to a reference partition* Π_0*,* $\mathcal{P}_\tau(\Pi_0)$*, as*

$$\mathcal{P}_\tau(\Pi_0) = \{\Pi \in \mathcal{P}^\star : \ \mathbb{P}(\Pi \mid \mathbf{Z}) \geq (1 - \tau) \cdot \mathbb{P}(\Pi_0 \mid \mathbf{Z})\}. \tag{3.2}$$

The RPS relative to Π_0 is the set of partitions that have a similar or higher posterior value than the reference. In our analysis, we are interested in Π^{MAP}—the *maximum a posteriori (MAP)* partition, so we will focus on $\mathcal{P}_\tau(\Pi^{\text{MAP}})$. That is, in our setting the RPS is the set of partitions that are sufficiently close to the posterior of the MAP partition. We write this as $\mathcal{P}_\tau$, dropping the reference argument unless explicitly needed. We could, more generally, define an RPS based on any posterior threshold, θ, such that $\mathcal{P}_\theta = \{\Pi \in \mathcal{P}^\star : \ \mathbb{P}(\Pi \mid \mathbf{Z}) \geq \theta\}$. Defining the RPS in comparison to a reference partition, however, avoids the need to define "high" posterior values (which requires knowing at least the scale of the entire posterior since they are, tautologically, defined in contrast to "low" density values). To construct the RPS, we begin with an initialization partition, then enumerate all models with posteriors at least as high as this initialization partition (which by definition includes Π^{MAP}). We then construct the RPS by moving down the list of partitions, ordered by posterior value, until we reach $(1 - \tau)\mathbb{P}(\Pi^{\text{MAP}}|\mathbf{Z})$.

Given a posterior over the partition models, we have a posterior over the effects of various feature combinations, possibly pooled, on the outcome of interest conditional on the partition

models in this set. So

$$\mathbb{P}(\boldsymbol{\beta} \mid \mathbf{Z}, \mathcal{P}_\tau) = \sum_{\Pi \in \mathcal{P}_\tau} \mathbb{P}(\boldsymbol{\beta} \mid \mathbf{Z}, \Pi)\mathbb{P}(\Pi \mid \mathbf{Z}, \mathcal{P}_\tau), \tag{3.3}$$

and analogously for measurable functions of $\boldsymbol{\beta}$. Using Hasse diagrams and our permissibility criteria, defined below, avoids imposing an artificial hierarchy on the partially ordered set of covariates, which allows us to interpret the partitions in the RPS.[1] The posterior for $\boldsymbol{\beta}$ restricted to the RPS, Equation (3.3), is, of course not the same as the distribution over all possible partitions, $P_{\boldsymbol{\beta}|\mathbf{Z}}(\boldsymbol{\beta})$. We can, however, characterize the uniform approximation error of the posterior distribution of $\boldsymbol{\beta}$, and measurable functions of it, restricting to the RPS (Theorem 3.3.1 in Section 3.3). This result does not depend on a specific prior over partitions, but we do need to specify a prior to find RPS in practice. We propose an ℓ_0 prior that assumes only sparsity in heterogeneity and is robust to any potential correlation structure between covariates. We show that the ℓ_0 prior is minimax optimal (Theorem 3.3.3). When combined with our permissible partition rules (Section 3.4), we can calculate bounds on the size of the RPS (Theorem 3.5.3) and enumerate the entire RPS (Algorithm 3 and Theorem 3.6.3) in Section 3.6.

3.3 Statistical properties of Rashomon Partition Sets

We elaborate on the statistical framework underlying the RPS. We give results for the general definition of the RPS with respect to some threshold, θ, though, in practice we define the RPS relative to the MAP, defining $\theta = (1 - \tau) \cdot \mathbb{P}(\Pi^{\text{MAP}}|\mathbf{Z})$ i.e., $\mathcal{P}_\tau(\Pi^{\text{MAP}}) = \mathcal{P}_\theta$. We first obtain posteriors over the (functions of) effects of feature combinations. Given a unique partition Π with some probability $\mathbb{P}(\Pi \mid \mathbf{Z})$, it may be useful to know the likely effects of using that specific feature $k \in \pi \in \Pi$. This could be because, for instance, there may be scientific reasons to otherwise prefer one versus the other, heterogeneity in costs, logistical

[1] We discuss this distinction in more detail in Section B.8 in the context of frequentist tree-based methods that use the same geometry.

considerations, etc. There is also a statistical reason: the posterior may not be concentrated on just a few pools for some k but maybe for others. Therefore, we may be interested in the posterior over the entire set of permissible pools

$$P_{\boldsymbol{\beta}|\mathbf{Z}}(\boldsymbol{\beta}) = \sum_{\Pi \in \mathcal{P}^\star} P_{\boldsymbol{\beta}|\mathbf{Z}}(\boldsymbol{\beta} \mid \Pi) \cdot \mathbb{P}(\Pi \mid \mathbf{Z}),$$

where $P_{\boldsymbol{\beta}|\mathbf{Z}}$ is the distribution function of $\boldsymbol{\beta} \mid \mathbf{Z}$. Throughout our analysis, we will assume that $P_{\boldsymbol{\beta}|\mathbf{Z}}$ is a proper distribution i.e., it satisfies the Kolmogorov axioms. Our goal is to approximate functions of $P_{\boldsymbol{\beta}|\mathbf{Z}}$ using only the RPS. That is,

$$P_{\boldsymbol{\beta}|\mathbf{Z},\mathcal{P}_\theta}(\boldsymbol{\beta}) = \sum_{\Pi \in \mathcal{P}_\theta} P_{\boldsymbol{\beta}|\mathbf{Z},\mathcal{P}_\theta}(\boldsymbol{\beta} \mid \Pi) \cdot \mathbb{P}(\Pi \mid \mathbf{Z}, \mathcal{P}_\theta), \qquad \mathbb{P}(\Pi \mid \mathbf{Z}, \mathcal{P}_\theta) = \frac{\mathbb{P}(\Pi \mid \mathbf{Z})}{\sum_{\Pi' \in \mathcal{P}_\theta} \mathbb{P}(\Pi' \mid \mathbf{Z})},$$

meaning that the approximation only evaluates models in the RPS but is also normalized by the RPS. The quality of this approximation, of course, depends on both the shape of the posterior (i.e. how concentrated is the posterior around the highest probability models) and the structure of the RPS. Our first goal, then, is to describe how well we can approximate $P_{\boldsymbol{\beta}|\mathbf{Z}}$ using the RPS. And then, we discuss how to construct the posterior over partitions, $\mathbb{P}(\Pi \mid \mathbf{Z})$, using generalized Bayesian inference. Technical details for results discussed here are deferred to Appendix B.3.

3.3.1 Posterior over effects

Consider the RPS, $\mathcal{P}_\theta$. The Rashomon partitions allow for uniform approximation of the posterior over the effects vector $\boldsymbol{\beta}$.

Theorem 3.3.1 (Rashomon approximation of posterior effects)**.** *Let $f : \mathbb{R}^K \to \mathbb{R}^m$ be a measurable function of the effects $\boldsymbol{\beta}$, where K is the number of unique feature combinations and $m \geq 1$. Then, the posterior distribution of $f(\boldsymbol{\beta})$ over the Rashomon Partition Set*

uniformly approximates the entire posterior of $f(\boldsymbol{\beta})$ *in the sense that*

$$\sup_{\boldsymbol{t}} \left| F_{f(\boldsymbol{\beta})|\mathbf{Z},\mathcal{P}_\theta}(\boldsymbol{t}) - F_{f(\boldsymbol{\beta})|\mathbf{Z}}(\boldsymbol{t}) \right| \leq \frac{1}{|\mathcal{P}_\theta|\,\theta} - |\mathcal{P}_\theta|\,\theta,$$

where $F_{f(\boldsymbol{\beta})|\mathbf{Z}}$ *is the distribution function of the transformation* $f(\boldsymbol{\beta}) \mid \mathbf{Z}$ *and* $F_{f(\boldsymbol{\beta})|\mathbf{Z},\mathcal{P}_\theta}$ *is the same but conditioned on the RPS.*

With small θ or large $\mathcal{P}_\theta$, this tends to 0, meaning that the posterior approximation can be quite close to that calculated over the full support. Essentially this is saying that if we have enough models of high enough posterior probability, then the error is low. This could arise as a result of having a few very highly likely models or having many models that are only fairly likely. We visualize the behavior of the $1/|\mathcal{P}_\theta|\,\theta$ term in our empirical data analyses in Section 3.8.

Notice that setting $f(\boldsymbol{\beta}) = \boldsymbol{\beta}$ recovers the posterior of $\boldsymbol{\beta}$. f also covers other useful quantities derived from the vector $\boldsymbol{\beta}$. An obvious example is $f(\boldsymbol{\beta}) = \max_k \beta_k$ since conditional on a given variant k being estimated as the one with the maximum effect. There is a winner's curse since the selection of the maximum is positively biased, so the bias needs to be corrected (the posterior needs to be adjusted to have a lower mean) to undo this effect (Andrews et al., 2019). Other examples include the variability over outcomes across the feature combinations $\|\boldsymbol{\beta} - \sum_K \beta_k/K\|_2^2$ and quantiles of the expected outcome distribution.

We now focus specifically on estimating the full posterior mean, $\mathbb{E}_\Pi \boldsymbol{\beta} = \sum_{\Pi \in \mathcal{P}^\star} \boldsymbol{\beta}_\Pi \mathbb{P}(\Pi \mid \mathbf{Z})$, using the RPS. If we simply restricted ourselves to only models in the RPS, we would have $\mathbb{E}_{\Pi,\mathcal{P}_\theta} \boldsymbol{\beta} = \sum_{\Pi \in \mathcal{P}_\theta} \boldsymbol{\beta}_\Pi \mathbb{P}(\Pi \mid \mathbf{Z})$. For some priors on $\boldsymbol{\beta}$, we could approximate $\mathbb{P}(\Pi \mid \mathbf{Z})$ but this requires specifying a prior on $\boldsymbol{\beta}$ and a corresponding approximation with adequate accuracy (Appendix B.2.1 gives an example using Gaussian priors and a Laplace approximation). More generally, the easiest quantity to compute the expectation is by taking the mean of the effects weighted by the self-normalized posterior probabilities as in Equation (3.4). Of course, if the RPS captures most of the posterior density, then this method can validly approximate the posterior mean but that is not the goal of this estimator. This estimator simply tells us

what the effects are across the RPS.

$$\mathbb{E}_{\Pi|\mathcal{P}_\theta}\boldsymbol{\beta} = \sum_{\Pi\in\mathcal{P}_\theta} \boldsymbol{\beta}_\Pi \frac{\mathbb{P}(\Pi \mid \mathbf{Z}, \mathcal{P}_\theta)}{\sum_{\Pi'\in\mathcal{P}_\theta} \mathbb{P}(\Pi' \mid \mathbf{Z}, \mathcal{P}_\theta)}. \tag{3.4}$$

This approach contrasts with Bayesian and frequentist methods based on resampling trees. When resampling trees, under conditions where the mean is well-separated, we generally see the average to have appealing asymptotic properties. In any finite sample, though, we also expect that there will be several highly unappealing trees mixed in by chance. In our approach, in contrast, we look explicitly for partitions with the highest posterior probability, forgoing exploring the entire space to instead focus on the partitions with the highest posterior probability. We can then characterize the quality of this approximation for a given RPS construction.

Corollary 3.3.2. *The mean conditional effect in Equation (3.4) approximates the posterior mean effect restricted to the Rashomon set, $\mathbb{E}_{\Pi,\mathcal{P}_\theta}\boldsymbol{\beta}$, as*

$$\frac{\left\|\mathbb{E}_{\Pi|\mathcal{P}_\theta}\boldsymbol{\beta} - \mathbb{E}_{\Pi,\mathcal{P}_\theta}\boldsymbol{\beta}\right\|}{\left\|\mathbb{E}_{\Pi,\mathcal{P}_\theta}\boldsymbol{\beta}\right\|} = \mathcal{O}\left(\frac{1}{|\mathcal{P}_\theta|\,\theta} - 1\right).$$

If we further have that the effects are bounded like $\|\boldsymbol{\beta}_\Pi\| < \infty$ for all $\Pi \in \mathcal{P}^\star$, then mean conditional effect in Equation (3.4) approximates the posterior mean effect, $\mathbb{E}_\Pi\boldsymbol{\beta}$, as

$$\left\|\mathbb{E}_{\Pi|\mathcal{P}_\theta}\boldsymbol{\beta} - \mathbb{E}_\Pi\boldsymbol{\beta}\right\| = \mathcal{O}\left(\frac{1}{|\mathcal{P}_\theta|\,\theta} - |\mathcal{P}_\theta|\,\theta\right).$$

Corollary 3.3.2 says that our approximation depends on both the Rashomon threshold θ and the distribution of the models in the posterior space,$|\mathcal{P}_\theta|$. Of course, this result extends to functions of $\boldsymbol{\beta}$ as well. To better understand Corollary 3.3.2, first assume that models are uniformly distributed in the posterior probability space, i.e., $|\mathcal{P}_\theta| \propto 1$ is independent of θ. As θ gets closer to 0, the RPS collects more permissible partitions, so $|\mathcal{P}_\theta|\,\theta$ behaves like 1 and we get a better approximation using Equation (3.4). As θ gets closer to 1, the RPS is

sparser. Therefore, $|\mathcal{P}_\theta|\,\theta$ behaves like 0 blowing up the error. Now, consider a more complex model space where the models do not have a uniform posterior. If our models are clustered near the *maximum a posteriori* (MAP) model, then a large θ will, in fact, give a better approximation. Conversely, if models are clustered near a very small posterior probability, then a large θ will blow up our approximation error.

3.3.2 Posterior over partitions

Now, we turn to the proverbial elephant of constructing a tractable posterior for Π, $\mathbb{P}(\Pi \mid \mathbf{Z}) \propto \mathbb{P}(\mathbf{y} \mid \mathbf{D}, \Pi) \cdot \mathbb{P}(\Pi)$. We need to model the likelihood component and the prior. Nested within $\mathbb{P}(\mathbf{y} \mid \mathbf{D}, \Pi)$ is a prior over $\boldsymbol{\beta}$. In work on Bayesian tree models (e.g. Chipman et al. (2010)), the typical strategy is to define a prior over partitions and then define conjugate priors on $\boldsymbol{\beta}$ and related hyperparameters so that it is easy to evaluate each draw from the distribution over trees. We take a different approach based on generalized Bayesian inference (Bissiri et al., 2016), which requires specifying fewer distributions explicitly. However, in Appendix B.2.2, we also give an example of a fully specified Bayesian model where maximizing the likelihood $\mathbb{P}(\mathbf{y} \mid \mathbf{D}, \Pi)$ is equivalent to minimizing $\mathcal{L}(\Pi; \mathbf{Z})$, drawing a parallel with the previous work in the Bayesian literature. Let $\exp\{-\mathcal{L}(\Pi; \mathbf{Z})\}$ be the likelihood of $\mathbf{Z}$ where $\mathcal{L}(\Pi; \mathbf{Z})$ is the loss incurred by the partition Π. Further, let $\exp\{-\lambda H(\Pi)\}$ be the prior over $\mathcal{P}^\star$. Then, we have

$$\mathbb{P}(\Pi \mid \mathbf{Z}) \propto \exp\{-\mathcal{L}(\Pi; \mathbf{Z})\} \cdot \exp\{-\lambda H(\Pi)\} =: \exp\{-Q(\Pi)\}. \tag{3.5}$$

Specifically, we use the mean-squared error for the loss function,

$$\mathcal{L}(\Pi; \mathbf{Z}) = \frac{1}{n} \sum_{\pi \in \Pi} \sum_{k(i) \in \pi} (y_i - \widehat{\mu}_\pi)^2, \qquad \widehat{\mu}_\pi = \frac{\sum_{k(i) \in \pi} y_i}{\sum_{k(i) \in \pi} 1}. \tag{3.6}$$

For the prior, we take the view that, unless directed otherwise by the specifics of the science in the context of the study, the researcher does not know the correlation structure

between the various possible pools. That is they do not have a strong view on whether $\beta_k = \beta_{k'}$ is correlated with whether $\beta_{k''} = \beta_{k'}$. Therefore, we define the prior over the *number of distinct pools* i.e., $H(\Pi) = |\Pi|$, the size of the partition. The prior plays a regularizing role, putting more weight on less granular aggregations. It corresponds to the ℓ_0 penalty: conditional on the number of pools in a partition, all permissible partitions are equally likely.

The RPS, taken together with the ℓ_0 penalty, is similar in spirit to the Occam's Window approach used in the context of Bayesian Model Averaging by Madigan and Raftery (1994) and Madigan et al. (1996). These papers use a stochastic search over the discrete space of models that ultimately results in a set of high posterior models and discards more complicated models if simpler models are found to have higher posterior probability. Our approach, which does not do discrete model averaging, formalizes this notion by including a prior with an ℓ_0 penalty as part of the model, rather than using it to guide the search. In Theorem 3.3.3 we show that this choice of prior is minimax optimal.

Let $\mathcal{Q}$ be a family of priors for some expected outcome $\boldsymbol{\beta}$. For any prior $Q \in \mathcal{Q}$, denote the posterior over $\boldsymbol{\beta}$ given some data $\mathbf{Z}$ as $\mathrm{P}_{Q,\mathbf{Z}}$, i.e.,

$$\mathrm{P}_{Q,\mathbf{Z}}(\boldsymbol{\beta}) = \mathbb{P}(\boldsymbol{\beta} \mid \mathbf{Z}, \boldsymbol{\beta} \sim Q) = \frac{\mathbb{P}(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}) Q(\boldsymbol{\beta})}{\mathbb{P}(\mathbf{y} \mid \mathbf{X})}.$$

Let us fix the sparsity at h: there are h distinct pools in the partition. Define the restricted space of partitions as $\mathcal{P}_{|h} = \{\Pi \in \mathcal{P}^\star : H(\Pi) = h\}$. Let $N(h) = \left|\mathcal{P}_{|h}\right|$. The ℓ_0 penalty imposes a sparsity restriction on the number of pools. Therefore, at a fixed sparsity h, the ℓ_0 penalty corresponds to a uniform prior over $\mathcal{P}_{|h}$. Denote the ℓ_0 prior as P_{ℓ_0}. So for any $\Pi \in \mathcal{P}_{|h}$, $P_{\ell_0}(\Pi) = 1/N(h)$.

For any given $\boldsymbol{\beta}$, there is a corresponding permissible partition $\Pi_{\boldsymbol{\beta}} \in \mathcal{P}^\star$. Then we can define $\mathcal{Q}_{|h}$ be the family of priors for the restricted space of $\boldsymbol{\beta}$ such that there is some $\Pi_{\boldsymbol{\beta}} \in \mathcal{P}_{|h}$. Let $\mathcal{Q}_{\mathcal{P}|h}$ denote the family of priors, derived from $\mathcal{Q}_{|h}$, over partitions in $\mathcal{P}_{|h}$. We can traverse from $\mathcal{Q}_{|h}$ to $\mathcal{Q}_{\mathcal{P}|h}$ by noticing that for a given $\boldsymbol{\beta}$, there is a corresponding

permissible partition $\Pi_{\boldsymbol{\beta}} \in \mathcal{P}$ i,e, for any prior $Q \in \mathcal{Q}_{|h}$, we can define a prior over $\mathcal{P}_{|h}$,

$$Q_{\mathcal{P}|h}(\Pi) = \int_{\boldsymbol{\beta}} \mathbb{I}(\Pi_{\boldsymbol{\beta}} = \Pi) Q(\boldsymbol{\beta}) d\boldsymbol{\beta}, \quad \Pi \in \mathcal{P}_{|h}.$$

For reference, we define the supports for various priors in Table B.1.

For two priors $P, Q \in \mathcal{Q}_{\mathcal{P}|h}$, define the total variation distance as

$$\delta(P, Q) = \sup_{\Pi \in \mathcal{P}_{|h}} \left| \mathrm{P}_{P,\mathbf{Z}}(\Pi) - \mathrm{P}_{Q,\mathbf{Z}}(\Pi) \right|.$$

Theorem 3.3.3. *For a given sparsity h, the ℓ_0 penalty is minimax optimal in the sense that*

$$\sup_{Q \in \mathcal{Q}_{\mathcal{P}|h}} \delta(P_{P_{\ell_0},\mathbf{Z}}, P_{Q,\mathbf{Z}}) = \inf_{P \in \mathcal{Q}_{\mathcal{P}|h}} \sup_{Q \in \mathcal{Q}_{\mathcal{P}|h}} \delta(P_{P,\mathbf{Z}}, P_{Q,\mathbf{Z}}).$$

In other words, if one is unwilling to commit to any correlation structure for the model coefficients, the ℓ_0 penalty, which puts a prior on the *number* of selected features, is optimal for model selection. RPS can be built using other priors, but we advocate for using the robust one. We do not want to impose false independence or unwarranted assumptions on correlations on the relationship between the βs and instead want to be robust in an environment with possibly a complex and unknown correlational structure. We show how this choice lends to computational tractability in Section 3.5.

3.3.3 Loss

It is also useful to characterize the Rashomon threshold in the loss space. Specifically, we define $\theta := \theta(q_0, \epsilon) = e^{-q_0(1+\epsilon)}/c$ where $c := c(\mathbf{Z})$ is some scaling constant depending on the observed data, $q_0 = Q(\Pi_0)$ is the loss incurred of some good model Π_0, and ϵ is largest acceptable deviation from q_0. Then, the (q_0, ϵ)-*Rashomon Partition Set* (RPS), $\mathcal{P}_{q_0,\epsilon}$ is defined as

$$\mathcal{P}_{q_0,\epsilon} = \{\Pi \in \mathcal{P}^\star : \ \mathbb{P}(\Pi \mid \mathbf{Z}) \geq e^{-q_0(1+\epsilon)}/c\}.$$

This allows us to interpret Rashomon partitions with respect to a reference partition or pooling. Without any context, it is difficult to choose a threshold θ. However, if we have some good reference model Π_0, then we can pick a threshold that is not much worse than the performance of Π_0. In particular, using a reference Π_0, we can define a measure of performance of model Π with respect to Π_0 as

$$\xi(\Pi, \Pi_0) = \frac{\log \mathbb{P}(\Pi \mid \mathbf{Z}) - \log \mathbb{P}(\Pi_0 \mid \mathbf{Z})}{\log \mathbb{P}(\Pi_0 \mid \mathbf{Z}) + \log \mathbb{P}(\mathbf{y} \mid \mathbf{X})} = \frac{Q(\Pi) - Q(\Pi_0)}{Q(\Pi_0)} = \frac{Q(\Pi) - q_0}{q_0}. \tag{3.7}$$

ξ is essentially the log-likelihood ratio of the two models weighted by the log-likelihood of the reference model. For data that has a considerably higher likelihood (weighted by the likelihood of the model), the measure goes to 1. So when Π is a better fit than the reference, $\xi(\Pi, \Pi_0) < 0$. Conversely, when Π is a poorer fit than the reference, $\xi(\Pi, \Pi_0) > 0$. Note that if the two posteriors are identical, then $\xi(\Pi, \Pi_0) = 0$.

Suppose we know that Π_0 is a good model such that $\mathbb{P}(\Pi_0 \mid \mathbf{Z}) = e^{-q_0}/c$. It makes sense to only look at partitions Π such that $\xi(\Pi, \Pi_0) \leq \epsilon$ for some $\epsilon > 0$. We show how to recover the RPS using (q_0, ϵ) in Proposition 3.3.4.[2]

Proposition 3.3.4. *Fix $\Pi_0 \in \mathcal{P}^\star$ and let $q_0 = Q(\Pi_0)$. Then $\mathcal{P}_{q_0,\epsilon} = \{\Pi \in \mathcal{P}^\star : \ \xi(\Pi, \Pi_0) \leq \epsilon\}$.*

By construction, this result is almost trivial. However, Proposition 3.3.4 is very powerful because it circumvents needing to calculate the normalizing constant $c(\mathbf{Z})$ which can be intractable. The loss tolerance, ϵ, is specified by the researcher. If the goal is to approximate the full posterior, then ϵ should be chosen based on computational constraints, since adding more models to the RPS will improve the approximation. However, if the goal is scientific interpretation, we want to choose ϵ such that the models in the RPS are all very high quality since adding more models with little evidence in the data would dilute the results. We give

[2]This is related to the candidate models for Bayesian model selection used by Madigan and Raftery (1994). Their set of models is $\mathcal{A} = \{\Pi : \mathbb{P}(\Pi_0 \mid \mathbf{Z})/\mathbb{P}(\Pi \mid \mathbf{Z}) \leq \tilde{c}\}$ where Π_0 is the *maximum a posteriori* estimate. In the language of Rashomon sets, $\tilde{c} = (\mathbb{P}(\Pi_0 \mid \mathbf{Z})\mathbb{P}(\mathbf{Z}))^{-\epsilon}$.

an example of how to choose ϵ in practice in Section 3.8.

In practice, we estimate a good reference model Π_0 using an off-the-shelf technique. Then, we use its loss, $q_0 = Q(\Pi_0)$ and a loss tolerance ϵ to trace out the RPS $\mathcal{P}_{q_0,\epsilon}$ using Algorithm 3 described in Section 3.6. By definition, this will include the MAP, Π^{MAP}. Then, we can use the approximation error in Theorem 3.3.1 to prune $\mathcal{P}_{q_0,\epsilon}$. We find an appropriate cutoff τ (which is equivalent to the loss tolerance ϵ if the reference model is Π^{MAP}) by identifying where the error begins to drop steeply. This gives us $\mathcal{P}_\tau(\Pi^{\text{MAP}})$. We demonstrate this procedure in our empirical examples of Section 3.8.

3.4 Permissible partitions

Recall from Section 3.2 that we have M features taking on R discrete and ordered values and $\mathcal{K}$ is the set of all $K = R^M$ unique feature combinations. Therefore, we can partially order the feature combinations. For a feature combination $k \in \mathcal{K}$, let k_m denote the value that the m-th feature takes. We say $k \geq k'$ if and only if $k_m \geq k'_m$ for all $m = 1, \ldots, M$. We say $k > k'$ if $k \geq k'$ but $k \neq k'$, and say that k and k' are incomparable if there are two features m_1 and m_2 such that $k_{m_1} > k'_{m_1}$ and $k_{m_2} < k'_{m_2}$. We denote the expected outcome of feature combination k by β_k. We will motivate the rest of our discussion using the following running example.

Example 3.4.1. Consider an example where a physician is offering a treatment consisting of two drugs amoxicillin and ibuprofen. Amoxicillin is available at three dosages, $\{0 \text{ mg}, 250 \text{ mg}, 500 \text{ mg}\}$. Ibuprofen is available at three dosages, $\{0 \text{ mg}, 200 \text{ mg}, 400 \text{ mg}\}$. Let (a, b) denote dosages of amoxicillin and ibuprofen respectively. □

We now formalize the notions of partitions and pools.

Definition 3.4.2 (Pool)**.** *A* pool π *is a set of feature combinations having identical expected outcomes.*

For a given pool π, two feature combinations $k_1, k_2 \in \pi$ only if $\beta_{k_1} = \beta_{k_2}$. Note that the

converse is not true. That is, we could have $k_1 \in \pi_1$ and $k_2 \in \pi_2$ for $\pi_1 \neq \pi_2$ even though $\beta_{k_1} = \beta_{k_2}$.

Definition 3.4.3 (Partition)**.** *Given M features taking on R partially ordered values each, a* partition Π *is a partitioning of this feature space into pools.*

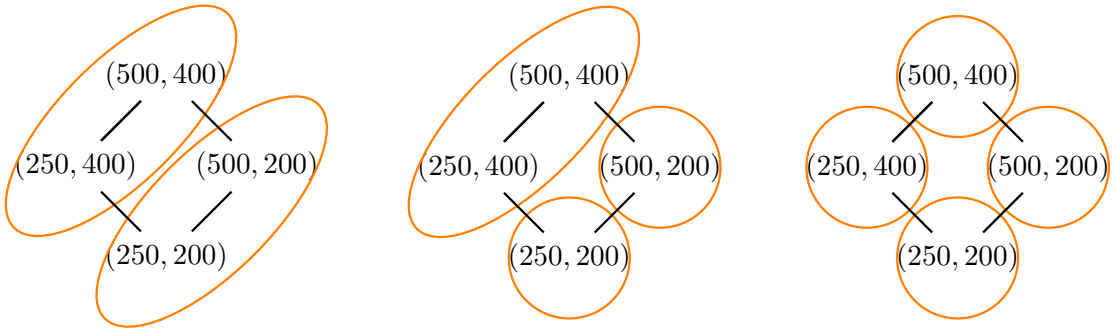
Many possible pools are not scientifically coherent. In Example 3.4.1, it is not scientifically coherent to say that for an ibuprofen only treatment should be pooled with an amoxicillin only treatment even if both have identical, non-zero effects relative to control – they are distinct drugs with distinct mechanisms. It makes sense to ask whether ibuprofen helps when it is prescribed alongside amoxicillin during a bacterial infection or, when both are administered simultaneously, what happens when dosages of both are increased.

Our goal is to learn scientifically permissible partitions of the feature space by heterogeneity in the expected outcomes. To do that we use *variants*, or feature combinations that differ by only one level, as finest grain building blocks for partitions.

Definition 3.4.4 (Variants)**.** *Two feature combinations $k < k'$ are* variants *if they have the same value of features for all but one and they vary by exactly one intensity value i.e., $k'_{-m} = k_{-m}$ and $k'_m = k_m + 1$ for some feature m.*

While variants reflect marginal changes, to characterize permissibility we also need a broader notion of (un)relatedness of feature combinations. To do this, we define a profile, which captures which features are "on" relative to base values. This allows us to establish a vocabulary for why, in the amoxicillin-ibuprofen treatment example, pooling different single-drug-only treatments is not sensible.

Definition 3.4.5 (Profile)**.** *A* profile $\rho(k)$ *is a binary vector indicating which of the M features have non-zero values, with the understanding that in a factorial design experiment, this indicates assignment to pure control and in a setting with heterogeneity we take (one of) the lowest value(s) as the base of 0.*



(a) Admissible partition (b) Inadmissible partition (c) Admissible partition

Figure 3.2: Hasse diagrams for amoxicillin and ibuprofen example. To see why Figure 3.2b fails Definition B.1.1, consider $\pi_i = \{(250\text{ mg}, 400\text{ mg}), (500\text{ mg}, 400\text{ mg})\}$ and $\pi_j = \{(500\text{ mg}, 200\text{ mg})\}$ with incomparable minima $(250\text{ mg}, 400\text{ mg}) \not\leqslant (500\text{ mg}, 200\text{ mg})$. This satisfies the antecedent but not the consequent of (3) (a).

For each profile ρ, let $\mathcal{K}_\rho = \{k \in \mathcal{K} : \rho(k) = \rho\}$ be the set of feature combinations in that profile. Within each profile, the partial ordering of feature combinations corresponds to the Hasse graph.

Definition 3.4.6 (Hasse)**.** *The Hasse for profile ρ, $\mathcal{H} = (\mathcal{K}_\rho, \mathcal{E})$, is a graph with nodes $\mathcal{K}_\rho$ and edges $\mathcal{E}$ relating the feature combinations through the partial ordering. Specifically, for two features $k, k' \in \mathcal{K}_\rho$, the edge $\langle k, k' \rangle \in \mathcal{E}$ if and only if k and k' are variants. We will also denote the edge $\langle k, k' \rangle$ as $e_{k,k'}$.*

Example 3.4.1 (continued). Drug combinations such as $k_1 = (500\text{ mg}, 200\text{ mg})$ correspond to the profile $\rho(k_1) = (1, 1)$, whereas single drug treatments such as $k_2 = (500\text{ mg}, 0\text{ mg})$ and $k_3 = (0\text{ mg}, 200\text{ mg})$ correspond to $\rho(k_2) = (1, 0)$ and $\rho(k_3) = (0, 1)$ respectively. The Hasse diagram for this setup, as well as example partitions for the $(1, 1)$ profile, are shown in Figure 3.2. □

In some scientific settings, profiles, as defined in Definition 3.4.5, may be too restrictive or unclear. In Example 3.4.1, domain knowledge said that amoxicillin and ibuprofen have

distinct drugs. When such regimes are unknown, they need to be learned. Therefore, we generalize the idea of profiles to be a subset of the feature space within which the partitions are permissible. When the regimes of scientific action are unknown, we will need to search for profiles and partitions.

Definition 3.4.7 (Generalized profile)**.** Generalized profiles *are formed by decomposing the entire feature space into the least number of subsets such that each subset corresponds to a regime of distinct scientific mechanisms.*

In our empirical examples, we will assume that scientific profiles are known and described as in Definition 3.4.5 but present algorithms that search for generalized profiles.

We imagine that the feature combination represents the physical world mechanistically capturing marginal changes: how changing a dosage in one drug marginally affects the outcome. Then one can think of the overall effect of some feature combination as summing up through these marginal effects. This amounts to considering variants, or taking one step up the Hasse.

Formally, we can re-parameterize the overall outcome of each feature k, β_k, into a sum of the marginals, which we denote by the vector $\boldsymbol{\alpha}$. Then, we have,

$$\beta_k = \sum_{k' \leq k; \rho(k) = \rho(k')} \alpha_{k'}. \tag{3.8}$$

This says that an expected outcome of a feature combination is the sum of expected marginal values leading up to it. This is identical to climbing up the Hasse. For instance, the treatment $(500 \text{ mg}, 400 \text{ mg})$ has value $\alpha_{500,400} = (\beta_{500,400} - \beta_{500,200}) - (\beta_{250,500} - \beta_{250,200})$. These will either capture a main effect of increasing a dosage (as on the sides of the Hasse) or an interaction effect between multiple dosage increases (as in the interior of the Hasse).

Of course, in this particular parameterization of $\boldsymbol{\beta}$, we chose to climb *up* the Hasse. We could have alternatively chosen to climb *down* the Hasse as $\beta_k = \sum_{k' \geq k; \rho(k) = \rho(k')} \gamma_{k'}$ for a marginal vector $\boldsymbol{\gamma}$. When the goal of the problem is to identify heterogeneity in $\boldsymbol{\beta}$,

there is no reason to prefer one parameterization of climbing the Hasse over the other.[3] Allowing for both parameterizations is coherent and not a contradiction – if the researcher is interested in the marginals, they can fix their interpretation to a single parameterization. Being agnostic to the direction of Hasse traversal has very important practical implications for computational feasibility, which we explore in Section 3.5.

Motivated by traversing the partial ordering while discarding all measure zero partitions, we present the formal definition of permissibility within a profile in Definition 3.4.8. We provide additional technical details with examples in Appendix B.1.

Definition 3.4.8 (Permissible partition of a profile)**.** *A partition* Π_0 *of a profile* ρ_0 *is* permissible *if*

(1) every $\pi \in \Pi_0$ *is a* pool *(cf. Definition 3.4.2), and*

(2) for every $\boldsymbol{\beta}$ *that generates* Π_0*, with respect to the Lebesgue measure,*

(a) the support of $\boldsymbol{\alpha}(\boldsymbol{\beta})$*,* $S_{\boldsymbol{\alpha}(\boldsymbol{\beta})} = \{\alpha_k \neq 0 \mid \alpha_k \in \boldsymbol{\alpha}(\boldsymbol{\beta})\}$*, is measurable, and*

(b) the support of $\boldsymbol{\gamma}(\boldsymbol{\beta})$*,* $S_{\boldsymbol{\gamma}(\boldsymbol{\beta})} = \{\gamma_k \neq 0 \mid \gamma_k \in \boldsymbol{\gamma}(\boldsymbol{\beta})\}$*, is measurable.*

We denote the set of all permissible partitions by $\mathcal{P}^\star$*.*

Definition 3.4.8 disregards partitions that are "measure zero," which are non-robust in the sense that the only rationalization for these splits rely on exact marginal effects that offset in a specific way. Such partitions require tremendous coincidence. Since partitions are discrete objects, it makes sense to ignore any partition that may arise from such measure zero events.[4] We believe that it is scientifically appropriate in a myriad of settings.

[3] Banerjee et al. (2021) used the climbing up parameterization with $\boldsymbol{\alpha}$. Here we take the view that this choice should not matter.

[4] Other techniques such as Lasso or decision trees may falsely estimate a non-zero posterior probability for such unrealistic partitions. We give examples of this in Appendix B.1.

Example 3.4.1 (continued). Permissibility interprets a pool π as the (lack of) marginal changes when climbing up or down an ordering. Consider the $(1, 1)$ treatment profile. Beginning from the lowest dosages of both (250 mg, 200 mg), we consider what happens when increasing the dosage of amoxicillin or ibuprofen. It could be that increasing amoxicillin has no effect (e.g. because the bacterial infection was highly localized) but increasing the dose of ibuprofen makes an appreciable difference (e.g., the patient feels much more relief from the pain). This is captured in Figure 3.2a.

We also want to avoid brittle pools relying on measure zero events. Starting from the lowest dosage of an amoxicillin-ibuprofen treatment, (250 mg, 200 mg), suppose that either increasing the dosage of *just* amoxicillin or the dosage of *just* ibuprofen has an appreciable effect. This rules out Figure 3.2a. When amoxicillin and ibuprofen are both raised to their largest dosages simultaneously, the effect would be a combination of each drug's dosage increase as well as an interaction effect between the two drugs. In a very special case, this interaction effect can exactly offset the effect of one drug's increase in dosage. For example, from a high 400 mg dose of ibuprofen, the benefits of a 250 mg dosage increase in amoxicillin (from 250 mg) can be *exactly* offset by an equal amount of stomach irritation that it causes so that 250 mg and 500 mg of amoxicillin have the exact same efficacy as 400 mg ibuprofen. Figure 3.2b captures this partition. However, this is a measure zero event. Almost surely, a stomach irritation does not *exactly* wash out the effect of a 250 mg dosage increase in amoxicillin.

Any amount of estimation noise, which is inevitable, makes brittle pools like Figure 3.2b unreliable. A permissible partition needs to be robust to estimation noise in the marginal increments, so, for example, the dosage combination (500 mg, 400 mg) has to be in a distinct pool as in Figure 3.2c. □

Definition 3.4.8 captures permissibility within a single profile, but we may also want to consider pooling across profiles. For example, it does not speak to the question of pooling treatments that add ibuprofen, as a temporary pain reliever, to a prescription of amoxicillin

against a bacterial infection – does introducing ibuprofen make an appreciable difference or not? In order to reason about this, we can partially order the profiles. Then, the ideas developed here naturally extend, which we describe in Appendix B.1.

Permissibility corresponds to topological restrictions on the Hasse. We can leverage this in representing the search problems for high-quality partitions to improve storage and computational performance. These benefits are byproducts of modeling permissibility natively. We can define partitions on the Hasse by removing (splitting on) edges in $\mathcal{E}$ to form disjoint connected components in $\mathcal{K}_\rho$. The removed edges correspond to non-zero marginal changes. This guarantees that all sets in a partition on the Hasse contain only "connected" feature combinations and generates a convex structure in the permissible partitions.

At a high level, permissibility restrictions generate equivalence classes amongst the edges in the Hasse, $\mathcal{E}$. The equivalence classes are those edges that can only be removed together to generate a partition. Suppose we decompose $\mathcal{E}$ into n mutually disjoint and exhaustive sets of edges $E_1, \ldots, E_n$, where each set E_j is an equivalence class. A partition Π induced by these equivalence classes says, Π removes edge e if and only if Π removes e' for every e' such that $e, e' \in E_i$ for some $i = 1, \ldots, n$. The equivalence classes will correspond to partitions where we pool along one of these edges if and only if we pool along the other. Let $\mathcal{E}'$ represent the set of edges that remain after the partition. Then the pruned graph $(\mathcal{K}, \mathcal{E}')$ specifies the corresponding partition.

Specifically, permissible partitions can be generated by identifying a unique decomposition of $\mathcal{E}$ into equivalent edges. The decomposition is given by,

$$\mathcal{E} = \bigcup_{m=1}^{M} \bigcup_{r=1}^{R} E_{m,r},$$

where for some feature m taking on value r, the equivalence class is $E_{m,r} = \{e_{k,k'} \in \mathcal{E} \mid k_m = r, k'_m = r+1\}$. In other words, edges between all pairs of variants k and k' that differ on the m-th feature i.e., $k_m = r$ and $k'_m = r+1$ for some arbitrary r belong to $E_{m,r}$.

This decomposition of $\mathcal{E}$ into equivalent edges corresponds to all "parallel" edges on the Hasse (see Figure 3.2 for example). We formalize this equivalent geometric interpretation of Definition 3.4.8 in Definition B.1.1.

The equivalence between the edges allows us to store partitions efficiently. If there are n equivalence classes, then there are 2^n possible partitions – we either split or pool across the edges in the i-th equivalence class. Rather than storing the partition as a set of pools or through a tree data structure, we can reduce the storage and calculation by a logarithmic factor by just keeping track of the hyperplanes induced by splits. That is, we can simply store a binary vector of length n. For interpretability purposes, we can reshape this vector into the $\mathbf{\Sigma} \in \{0,1\}^{M\times(R-2)}$ matrix. Here $\Sigma_{mr} = 1$ if and only if feature combinations with level r is pooled with feature combinations with factor $r+1$ in feature m. We walk through detailed examples in Appendix B.1.

Any estimation strategy implicitly takes some stand on partition structure. In Appendix B.1, we discuss permissibility restrictions in the context of several other commonly used models. We sketch examples using long/short regression, Lasso, decision trees, causal trees, and treatment variant aggregation. These alternative approaches impose permissibility restrictions that are naturally captured by our Hasse topology, although they are not presented formally as such. These permisssibility restrictions are generated by the choice of technique, rather than through any specific scientific consideration.

In the remainder of this paper, we will consider only permissible partitions and may refer to them only as partitions (dropping the "permissible" quantifier) unless we need to distinguish them. Our paper is modular: one could construct Rashomon Partitions with a different notion of permissibility.

3.5 Size of the Rashomon Partition Set

Given that we would like to enumerate $\mathcal{P}_\theta$ it is useful to calculate bounds on both its size and also $\mathcal{P}^\star$. Since any permissible partition requires each profile to respect Definition 3.4.8, it is sufficient to consider each profile independently. To develop an upper bound, we will

use m to denote the number of features with non-zero values in the profile we are focusing on, so $m \in \{1, \ldots, M\}$. All technical details are deferred to Appendix B.4.

The first observation is that $\mathcal{P}^\star$ is small relative to the total number of potential partitions.

Proposition 3.5.1. *In each profile, the total number of all possible partitions is* $\mathcal{O}\left(2^{2^{(R-1)^m}}\right)$, *and permissible partitions is* $\mathcal{O}(2^{m(R-2)})$.

Next, we will show that the size of the RPS is only polynomial in M and R. In Lemma 3.5.2, we crucially observe that the ℓ_0 prior bounds the number of pools in any Rashomon partition.

Lemma 3.5.2. *For a given Rashomon threshold* θ *and regularization parameter* λ, *any partition in the RPS,* $\mathcal{P}_\theta$, *can have at most* $H_\theta(\lambda)$ *pools,*

$$H_\theta(\lambda) = \left\lfloor -\frac{\ln(c\theta)}{\lambda} \right\rfloor,$$

where $c := c(\mathbf{Z})$ *is a normalization constant that depends only on* $\mathbf{Z}$ *and* $\lfloor \cdot \rfloor$ *is the floor function.*

Lemma 3.5.2 allows us to further reduce the number of the partitions in Proposition 3.5.1 by considering only partitions that meet this requirement. Even when the regimes of scientific action i.e., profiles, are unknown, we show that the size of the RPS is bounded polynomially in Theorem 3.5.3. In Lemma B.4.5 we bound the size of the RPS when the profiles are known apriori. Such a relationship between regularization and size of the model class was previously hypothesized and shown for empirical data by Semenova et al. (2022).

Theorem 3.5.3. *For a given Rashomon probability* θ *and regularization parameter* λ, *the size of the Rashomon Partition Set is bounded by*

$$|\mathcal{P}_\theta| = \begin{cases} \mathcal{O}\left(M^{2H-1} R^{H-1}\right), & R > M^{c_{\text{crit}}} \\ \mathcal{O}\left((MR)^{H-1} (\log_2(MR))^{-1}\right), & \text{else} \end{cases},$$

where $H := H_\theta(\lambda)$ *and* $c_{\text{crit}} = (\log_2 3 - 1)/(2 - \log_2 3)$.

Observe that $c_{\text{crit}} \approx 1.41$. In most settings, the number of factors R is fixed and the number of features M grows. Therefore, $R \leq M^{c_{\text{crit}}}$ as M grows, and we will fall under the second case which is smaller by a factor of M^H. In our empirical examples, we see that with just hundreds of partitions in our RPS we get close to the full posterior.

3.6 Enumerating Rashomon Partitions

Since we do not pool across profiles, we can enumerate the Rashomon Partition for each profile independently and then finally combine them in the end. We will first develop intuition to present an algorithm to enumerate the RPS for a single profile. Then, we will discuss how to combine these profile-specific RPS to get our RPS across all profiles. The intuition behind our enumeration is that any split we make introduces a new set of pools. If for some reason this split is very bad, then no matter what other split we make, we can never recover. Theorems 3.6.1 and 3.6.2 help us identify those poor splits. They rely on the fact that equivalent points having the exact same feature values will always belong to the same pool. However, equivalent units may not have the same outcome. Therefore, we will always incur some loss from these equivalent units (also see Angelino et al. (2017); Xin et al. (2022)). We defer technical details of the results to Appendix B.5.

Consider some partition matrix $\mathbf{\Sigma}$ for profile ρ, where the partition is given by $\Pi := \Pi(\mathbf{\Sigma})$. Given some data $\mathbf{Z} = (\mathbf{X}, \mathbf{y})$, we will use the mean squared error and the average outcome in pool $\pi \in \Pi$, $\widehat{\mu}_\pi$, as defined in Equation (3.6). However, the results generalize to any non-negative loss as we will see in Appendix B.7 with weighted mean-squared error.

Suppose we fix some indices $\mathcal{M}$ in $\mathbf{\Sigma}$. Define a new matrix $\mathbf{\Sigma}_{\mathfrak{f}}$,

$$\Sigma_{\mathfrak{f},(i,j)} = \begin{cases} \Sigma_{(i,j)}, & (i,j) \in \mathcal{M} \\ 0, & \text{else} \end{cases}.$$

In other words, $\mathbf{\Sigma}_{\mathfrak{f}}$ is a partition where all heterogeneity splits made by $\mathbf{\Sigma}$ corresponding

to indices in $\mathcal{M}$ are obeyed and we maximally split at all other places. Let $\Pi_{\mathfrak{f}} := \Pi(\mathbf{\Sigma}_{\mathfrak{f}})$ correspond to this maximal partition respecting $\mathbf{\Sigma}$ at indices $\mathcal{M}$. Next, define

$$\pi_{\mathfrak{f}} = \{k \in \mathcal{K} \mid k_i \leq j+1 \iff (i,j) \in \mathcal{M}\}$$

to be the set of all feature combinations covered by indices in $\mathcal{M}$. And we define the complement $\pi_{\mathfrak{f}}^c = \mathcal{K} \setminus \pi_{\mathfrak{f}}$. Finally, define $H(\Pi, \mathcal{M}) = \sum_{\pi \in \Pi} \mathbb{I}\{\pi \cap \pi_{\mathfrak{f}} \neq \varnothing\}$ to be the number of pools in Π consisting of feature combinations corresponding to indices $\mathcal{M}$.

Consider a procedure where we keep $\mathbf{\Sigma}$ constant at $\mathcal{M}$ and make further splits (not already implied by $\mathcal{M}$) at other indices only. Define child$(\mathbf{\Sigma}, \mathcal{M})$ to be all such $\mathbf{\Sigma}'$. Using the intuition we built earlier, our search algorithm in Algorithm 9 starts at some partition and fixes some heterogeneity splits. Theorem 3.6.1 says that if the loss incurred by these fixed heterogeneity splits is already too high, then we should discard this partition and its children.

Theorem 3.6.1. *Let θ_ϵ be the Rashomon threshold in the loss space i.e., $\Pi \in \mathcal{P}_{q,\epsilon}$ if and only if $Q(\Pi) < \theta_\epsilon$. Given a partition $\Pi := \Pi(\mathbf{\Sigma})$ for partition matrix $\mathbf{\Sigma}$, a set of fixed indices $\mathcal{M}$, and data $\mathbf{Z}$ consisting of n observations, define*

$$b(\mathbf{\Sigma}, \mathcal{M}; \mathbf{Z}) = \frac{1}{n} \sum_{\pi \in \Pi_{\mathfrak{f}}} \sum_{k(i) \in \pi} \mathbb{I}\left\{k(i) \in \pi_{\mathfrak{f}}\right\} (y_i - \widehat{\mu}_\pi)^2 + \lambda H(\Pi, \mathcal{M}). \tag{3.9}$$

If $b(\mathbf{\Sigma}, \mathcal{M}; \mathbf{Z}) > \theta_\epsilon$, then $\mathbf{\Sigma}$ and all $\mathbf{\Sigma}' \in$ child$(\mathbf{\Sigma}, \mathcal{M})$ are not in the Rashomon set $\mathcal{P}_{q,\epsilon}$.

Building on the same intuition, Theorem 3.6.2 "looks ahead" to see if this partition is of poor quality. If the loss incurred by feature combinations yet to be split is too high, then we should abandon this partition.

Theorem 3.6.2. *Consider the same setting as Theorem 3.6.1. Define*

$$b_{eq}(\mathbf{\Sigma}, \mathcal{M}; \mathbf{Z}) = \frac{1}{n} \sum_{\pi \in \Pi_{\mathfrak{f}}} \sum_{k(i) \in \pi} \mathbb{I}\left\{k(i) \in \pi_{\mathfrak{f}}^c\right\} (y_i - \widehat{\mu}_\pi)^2, \tag{3.10}$$

$$B(\mathbf{\Sigma}, \mathcal{M}; \mathbf{Z}) = b(\mathbf{\Sigma}, \mathcal{M}; \mathbf{Z}) + b_{eq}(\mathbf{\Sigma}, \mathcal{M}; \mathbf{Z}). \tag{3.11}$$

If $B(\mathbf{\Sigma}, \mathcal{M}; \mathbf{Z}) > \theta_\epsilon$*, then* $\mathbf{\Sigma}$ *and all* $\mathbf{\Sigma}' \in \text{child}(\mathbf{\Sigma}, \mathcal{M})$ *are not in the Rashomon set* $\mathcal{P}_{q,\epsilon}$*.*

Theorems 3.6.1 and 3.6.2 help aggressively cut down the search space by combining the lowest penalty on the splits already made and the lowest mean-squared error on the splits yet to be made. If this is already too high, then we abandon our search. We illustrate this in Algorithm 9. Here, we start with all feature combinations pooled together. We begin our search at the first feature trying to split the two variants with the lowest dosages into separate pools. We keep a queue of possible splits to consider. Whenever we remove a possible split from the queue, we check its viability using Lemma 3.5.2, and Theorems 3.6.1 and 3.6.2. If this is a bad split, we go to the next split in the queue. And if this is a good split (so far), we check if it already meets the Rashomon threshold and recursively add other possible splits to this queue. We also maintain a cache of splits that have been added to the queue at some point to avoid doubling back on old splits.

As noted before, we can solve each profile independently. In Algorithm 3, we explicitly show how to do this. Note that in line 5, we once again leverage Theorem 3.6.2 by noting that each profile will always incur some loss. Once we solve each profile independently, Algorithm 12 describes how to pool across profiles as defined in Definition B.1.8. Appendix B.5 describes this in more detail. Given our setup, it is easy to see how Algorithm 3 can be parallelized by delegating calls to Algorithm 9 and profile to separate threads.

Theorem 3.6.3. *Algorithm 3 correctly enumerates the Rashomon partition set.*

3.7 Simulations

In this section, we present two simulation studies that illustrate the Rashomon effect and highlight the importance of enumeration (as compared to sampling). In Appendix B.6, we provide an additional simulation study that compares the RPS with the Lasso, which relies on independence and thus will not identify the best policy. In contrast, the RPS tends to

Algorithm 3 `EnumerateRPS`$(M, R, H, \mathbf{Z}, q, \epsilon)$

Input: M features, R factors per feature, max pools H, data $\mathbf{Z}$, reference loss q, threshold ϵ

Output: Rashomon set $\mathcal{P}_{q,\epsilon}$

$\mathcal{P}_{q,\epsilon} = \varnothing$
$\mathcal{R}$ all sets of candidate profiles
for set of profiles $\rho \in \mathcal{R}$ **do**
    $H' = H - |\rho| + 1$
    $\mathcal{E} = [b_{eq}$of profile ρ_i for $\rho_i \in \rho]$ ▷ b_{eq} in Theorem 3.6.2 with zero matrix
    $\mathcal{P} =$ `dict()`
    for $\rho_i \in \rho$ **do**
        $q_i = q(1+\epsilon) -$ `sum`$(\mathcal{E}) + \mathcal{E}_{\rho_i}$
        $M_i =$ active features in ρ_i
        $R_i = R[M_i]$
        $\mathcal{P}[\rho_i] =$ `EnumerateRPS_profile`$(M_i, R_i, H', \mathbf{Z}, q_i)$ ▷ See Algorithm 9
        Sort partition matrices in $\mathcal{P}[\rho_i]$ on loss
    $\mathcal{P}' = \bigtimes_{\rho_i \in \rho} \mathcal{P}[\rho_i]$ ▷ Obtain candidate partitions with Cartesian product
    $\mathcal{P}_{q,\epsilon} = \mathcal{P}_{q,\epsilon} \cup$ `PoolProfiles`$(\mathcal{P}', \rho_0, \mathbf{Z}, q(1+\epsilon))$ ▷ See Algorithm 12
return $\mathcal{P}_{q,\epsilon}$

include the best policy (with others, of course), but also converges to the true best profile as the amount of data increases.

First, take a setting with four features. Each feature takes on four ordered factors including the control (which corresponds to zero, when the feature is inactive), $\{0, 1, 2, 3\}$. There are sixteen different feature profiles: $2^4 = 16$ possible combinations of active and inactive features. The control corresponds to the profile where all features are inactive. Our data-generating process groups all feature combinations in a given profile into the same pool. We will assume that the following profiles have a non-zero outcome:

$$\beta_{(0,0,0,1)} = 4.4,\ \sigma^2_{(0,0,0,1)} = 1,\ \beta_{(0,1,0,0)} = 4.3,\ \sigma^2_{(0,1,0,0)} = 1,\ \beta_{(0,1,0,1)} = 4.45,\ \sigma^2_{(0,1,0,1)} = 1,$$
$$\beta_{(1,0,1,0)} = 4.5,\ \sigma^2_{(1,0,1,0)} = 1.5,\ \beta_{(1,1,1,1)} = 4.35,\ \sigma^2_{(1,1,1,1)} = 1.$$

All other feature profiles have outcome $\beta = 0$ and variance $\sigma^2 = 1$. The feature profile

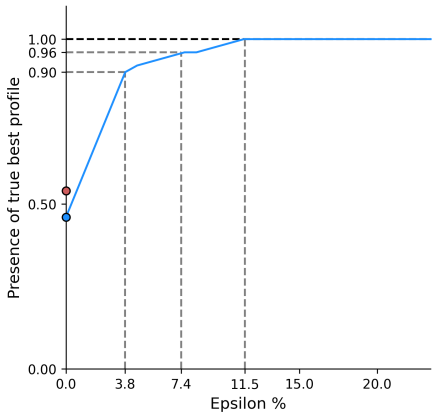


Figure 3.3: Simulation results. The plot shows often the true best profile is discovered as we increase the Rashomon threshold in the blue curve. With just $\epsilon \approx 0.038$, we recover the true best profile in the Rashomon set about 90% of the time. The red dot corresponds to how often Lasso recovers the true best profile.

$(1, 0, 1, 0)$ is the best, however, the other four profiles listed above are very close. We generated data with $n_k = 30$ data points per feature combination. Each profile has a different number of feature combinations depending on how many features are active. Therefore, each profile will have a different number of data points. The outcomes were drawn from $\mathcal{N}(\beta_i, \sigma_i^2)$. We averaged the results over $r = 100$ simulations. Figure 3.3 tells us how often the true best feature profile is present in RPS as a function of the threshold ϵ. Lasso selects the data generating model only about half the time, though as we increase ϵ we see that the data generating models is nearly always included in the RPS.

In the next simulation, we look at treatment effect in a binary treatment setting. We look at Wager and Athey (2018)'s Causal Random Forests (CRFs), which we discuss more in Section 3.9 and Appendix B.8. CRFs sample over the space of decision trees, so we ask how many causal trees are permissible and are in the RPS? Causal trees partition the features to find heterogeneity in the treatment effect directly. This is in contrast to partitions in the RPS that find heterogeneity in the outcome. To make the comparison fair, we set the outcome of the control group to a constant, 0, without any noise.

We simulate data with four features, the first feature being a binary treatment variable.

Table 3.1: Results for second simulation study. We compare how often CRFs find permissible partitions and how often they are present in the RPS. We vary both the number of trees in the CRF and the Rashomon threshold. Each cell shows the fraction of CRF trees inside the RPS (within parentheses are absolute counts). The numbers are averaged over 100 simulations.

	# trees = 20 (# permissible = 1.29)	# trees = 50 (# permissible = 2.67)	# trees = 100 (# permissible = 10.32)
$\epsilon = 0.1$ ($\lvert\mathcal{P}_\theta\rvert = 7.46$)	0% (0)	0% (0)	0% (0)
$\epsilon = 0.2$ ($\lvert\mathcal{P}_\theta\rvert = 46.6$)	0% (0)	0% (0)	0% (0)
$\epsilon = 0.3$ ($\lvert\mathcal{P}_\theta\rvert = 126.54$)	0.41% (0.52)	0.91% (1.15)	3.35% (4.24)
$\epsilon = 0.5$ ($\lvert\mathcal{P}_\theta\rvert = 823.81$)	0.16% (1.29)	0.32% (2.67)	1.25% (10.32)

The second feature takes on 3 ordered levels and the last two features take on 4 ordered levels. The following are the outcomes for the treatment group:

$$\beta_{(1,1,1:2,1:3)} = 2,\ \beta_{(1,1,1:2,4)} = 4,\ \beta_{(1,1,3:4,1:3)} = 2,\ \beta_{(1,1,3:4,4)} = 0,$$
$$\beta_{(1,2,1:2,1:3)} = 3,\ \beta_{(1,2,1:2,4)} = 5,\ \beta_{(1,2,3:4,1:3)} = 7,\ \beta_{(1,2,3:4,4)} = 1,$$
$$\beta_{(1,3,1:2,1:3)} = 1,\ \beta_{(1,3,1:2,4)} = -1,\ \beta_{(1,3,3:4,1:3)} = -1,\ \beta_{(1,3,3:4,4)} = -2.$$

We generate $n_k = 10$ data points per feature combination. In the treatment group, we drew outcomes from a $\mathcal{N}(\beta_k, 1)$ distribution. We averaged simulations over 100 iterations. The results are presented in Table 3.1. The vast majority of partitions sampled by CRFs are notscientifically coherent (permissible) partitions and thus cannot be interpreted as plausible explanations. This result is not specific to CRFs and would hold for any algorithm using unrestricted trees. Consequently, the number of trees that are in the RPS is also very small, meaning that, although averaging over trees has appealing asymptotic properties, the trees included in particular sample are unlikely to be high-quality explanations. This result is, again, not specific to CRFs but highlights the value of exploring uncertainty by enumerating high-quality explanations compared to sampling.

3.8 Empirical data examples

In three distinct environments, we emphasize robust conclusions, both affirming and challenging the established literature's findings in each context through the use of RPS.

3.8.1 Does price matter in charitable giving?

Karlan and List (2007) conducted a field experiment to better understand the anatomy of charitable giving. Most mechanisms for charitable contributions utilize high matching ratios (e.g., \$3 matched for each \$1 contributed by an individual) which is assumed to motivate giving. But this could be wasteful if an individual is just as likely to give with no match (or a low match) because, for instance, they have "warm glow" from giving. It is similarly possible that individuals care about matching caps from a public goods perspective: the returns to donating are lower with a lower cap. Whether a contribution policy is cost effective or wasteful crucially relies on how the features of the design line up with individuals' utility functions.

They used mail solicitations to prior donors of a non-profit political organization to study the effect of price on charitable donations. The data contains 50,083 individuals in the United States who had previously donated to the organization. All individuals received a letter soliciting donations. Those in the treatment group (33,396 people) included an additional paragraph describing that their donation will be matched. The letters were identical otherwise. Three treatment arms were cross-randomized: (i) the maximum size of the matching gift across all donations (\$25k, \$50k, \$100k, unspecified), (ii) the ratio of price match (1:1, 2:1, 3:1), and (iii) an example suggested donation amount ($1\times$, $1.25\times$, $1.5\times$ the subject's highest previous contribution). Additionally, they classify states as red or blue depending on whether they voted for George Bush or John Kerry in the 2004 U.S. presidential election. Together with the treatment assignments, our feature space is a $4 \times 3 \times 3 \times 2$ Hasse.

Karlan and List (2007) find that (1) match ratio does not matter; (2) gift maximum does not matter; (3) political leanings only matter at the extensive margin—having a match mo-

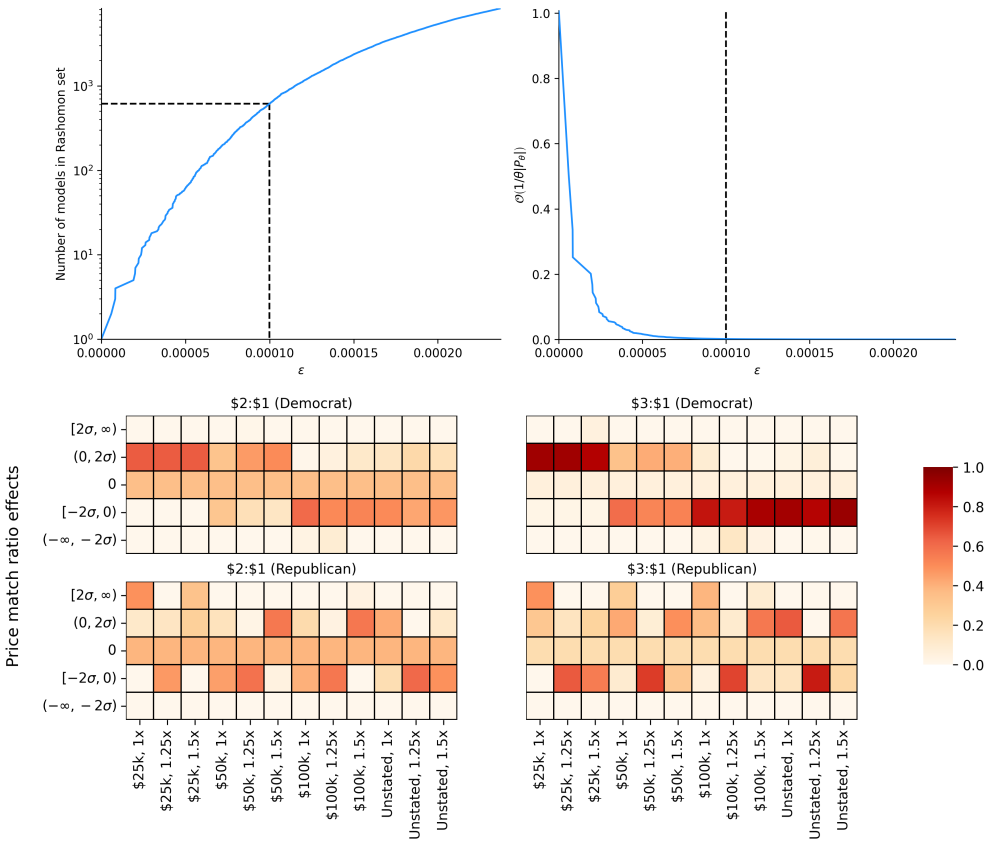


Figure 3.4: Results for the Karlan and List (2007) dataset. The top two panels show the size of the RPS and error term in Theorem 3.3.1 as a function of ϵ. Our choice of $\epsilon = 10^{-4}$ is highlighted by the black dashed line. The bottom panel shows the effect of price match of \$2:\$1 and \$3:\$1 relative to \$1:\$1 are stratified by political leaning and other treatments in the RPS.

tivates Republicans and not Democrats, though again ratio does not matter. Their findings falsify certain theoretical economic models and suggest that typical design which emphasizes matching and maxima are wasteful.

We revisit their findings through the RPS. Figure 3.4 shows how the set size and error bound change with ϵ. Using Figure 3.4 as a guide, we chose ϵ so that adding additional models to the RPS does not dramatically increase the approximation quality (akin to choosing the number of components in principal components using a scree plot). We choose $\epsilon = 10^{-4}$. A larger ϵ would dilute the RPS by adding more models that have little support in the data.

Figure 3.4 also shows the effect of price match of \$2:\$1 and \$3:\$1 relative to \$1:\$1 for all feature combinations.

The RPS strongly rejects all three of their conclusions. First, match ratio has a robust impact (positive and negative) for Democrats and positive for Republican, relative to control. Second, the gift maximum clearly matters, and in interesting ways. Particularly, Democrats are robustly encouraged by lower gift maxima and deterred by the very high/unrestricted ones, requiring a nuanced explanation. The Karlan and List (2007) regression analysis effectively averages out these cells. Third, political leaning matters for essentially all margins. Over 99% of the RPS split on political leaning. And there are interesting robust subtleties: the previously mentioned non-monotone effects of the match ratio for Democrats is not a fluke and merits investigation.

3.8.2 Heterogeneity in telomere length

Telomeres are regions of repeated nucleotide sequences near the end of the chromosome that protect the chromosome from damage. They reduce in length every time a cell divides eventually becoming so short that the cell can no longer divide. A recent literature has begun to examine what features are associated with (or possibly cause) changes in telomere length. For example, telomere shortening has been associated with cellular senescence and aging and has been thought to hold biomarkers as targets for genetic predispositions and anti-cancer therapies (Rossiello et al., 2022; Srinivas et al., 2020). Recent research suggests that there may only be a narrow range of healthy telomere lengths; anything extremal is at increased risk of immune system problems or cancer (Alder et al., 2018; Protsenko et al., 2020). Research has found heterogeneity by race, ethnicity, age, and even stress (Chae et al., 2014; Geronimus et al., 2015; Hamad et al., 2016; Vyas et al., 2021). Nonetheless, these studies are large-scale associative statistical analyses rather than those derived from micro-experimental data.

We use the National Health and Nutrition Examination Survey (NHANES) collected in 1999 and 2002. The survey included blood draws and from the samples DNA analyses were

performed and telomere length was estimated. Specifically, the dataset reports the mean T/S ratio (telomere length relative to standard reference DNA).[5] The dataset also contains socio-economic variables. To speak to the emerging literature on telomere heterogeneity, we focus on hours worked (a proxy for stress), age, gender, race, and education. Our goal is to study the RPS of this heterogeneity on T/S.[6]

We show our choice of ϵ for the Rashomon threshold in the top two panels of Figure 3.5. In the RPS, we found robust heterogeneity in race – specifically, we found no partition that pools features across different races. So the remainder of the analysis will stratify based on race. We found robust evidence of heterogeneity in gender only in White and Black races. All partitions for these races split males and females into separate pools. This was absent in Other races – only 23% of the partitions in the RPS split on gender.

For each race r, we find the length of telomeres stratified by each feature $m \in \{$Hours worked, Gender, Age, Education$\}$ relative to the lowest level of that feature, $\mathbb{E}[Y_i(x_m, \mathbf{x}_{-m}, r)] - \mathbb{E}[Y_i(1, \mathbf{x}_{-m}, r)]$.[7] We estimate this using the mean, $\widehat{y}(\mathbf{x}, r)$, as $\widehat{y}(x_m, \mathbf{x}_{-m}, r) - \widehat{y}(1, \mathbf{x}_{-m}, r)$. As in the previous case, we categorize this into five bins using the standard deviation of the difference in lengths across all models in the RPS. We average the counts in each bin and report them in the bottom panel of Figure 3.5. Again, the bin "0" corresponds to the case where the features were pooled together thereby having no difference in telomere lengths. We find very few robust patterns. As discussed earlier, we find robust differences in telomere lengths across males and females in the Black population and a robust non-difference in Other races. Similarly, we find a robust non-difference in Black and Other races in telomere lengths for people who work fewer than 40 hours.

---

[5] See https://wwwn.cdc.gov/nchs/nhanes/1999-2000/TELO_A.htm for details. Website last accessed on 2024-01-29.

[6] To operationalize, we removed all individuals who were missing data for our relevant covariates. We binned the number of hours worked: ≤ 20 hours, $21 - 40$ hours, and ≥ 40 hours. Gender was unordered: Male or Female. Age was categorized into five ordered discrete factors – ≤ 18 years, $19 - 30$ years, $31 - 50$ years, $51 - 70$ years, and > 70 years. Education was categorized into 3 ordered discrete factors – did not complete GED, finished GED but did not finish college, and received some college degree. Finally, race was categorized into three unordered factors – Black, White, and Other.

[7] For gender, we found telomere lengths of *Male* ($x_m = 2$) relative to *Female* ($x_m = 1$).

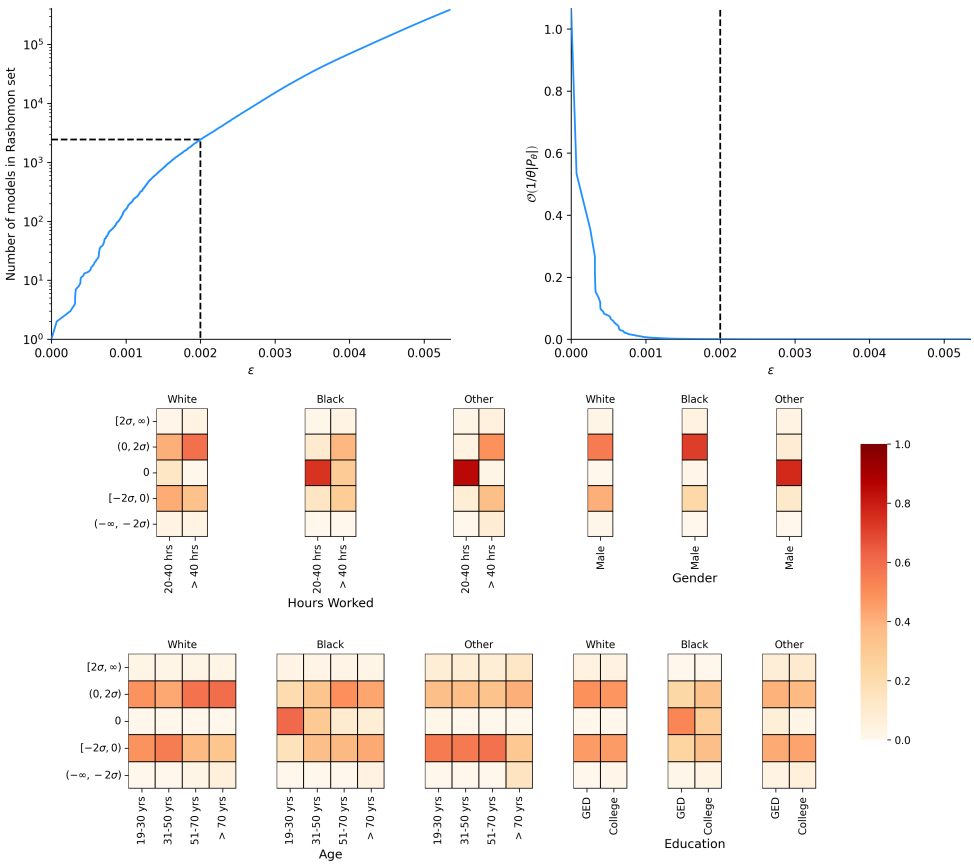


Figure 3.5: The top two panels show what happens as we increase ϵ in the NHANES dataset highlighting our choice of ϵ. In the bottom panel, we highlight heterogeneity in telomere length across the four features (hours worked, gender, age, and education) relative to the lowest level of that feature, sorted into race.

Our findings reveal an absence of robust evidence supporting the patterns highlighted in existing literature. Moreover, of the few robust patterns we do identify, several findings contradict prior research. We find Black males have longer telomeres than females. Among White people, we find older people have longer telomeres, which also contradicts existing research. This underscores the necessity for further exploration in this field using comprehensive data and appropriate statistical methods.

3.8.3 Heterogeneity in the impact of microcredit access

A large literature has looked at the impact of microfinance on several outcomes, ranging from private consumption to business outcomes to social outcomes (e.g., female empowerment). Mostly, the literature has found little beyond basic consumption effects (Angelucci et al., 2015; Attanasio et al., 2015; Augsburg et al., 2015; Banerjee et al., 2015; Crépon et al., 2015; Tarozzi et al., 2015; Meager, 2019), though there is suggestive evidence of some potential heterogeneity. One specific heterogeneity of interest concerns entrepreneurs: those with pre-existing businesses may be particularly benefited by the access to microfinance loans (Banerjee et al., 2019). Another concerns family size (Baland et al., 2008): the returns to credit access may vary by whether the household has more children.

We study the RPS in the Banerjee et al. (2015) data. The data is generated from a randomized controlled trial in which 102 neighborhoods in Hyderabad, India were randomly assigned to treatment or control, each with equal probability, where treatment meant that a partner microfinance organization, Spandana, entered. At baseline a number of characteristics of sampled individuals were collected, including the gender of the head of the household, the education status of the head of the household, the number of businesses previously owned by the household, and the number of children in the household. Additionally, at the neighborhood level, information about the share of households with debt, the share of households with businesses, total expenditure per capita in the region, and average literacy rates in the region were also collected at baseline. Amongst these regional characteristics, motivated by the literature we only look at the regional debt and business variables.

We look at outcomes from the second (longer term) endline, focusing on four spheres: (i) loans, (ii) household response (total expenditure, durables, temptation goods, labor supply), (iii) business (revenue, size, assets, profits), (iv) female empowerment (female business participation, education of daughters). We discretized the regional characteristics and the number of businesses previously owned into four levels based using quartiles. We set the first quartile as the "base control" i.e., we consider that characteristic to be active if it is one of

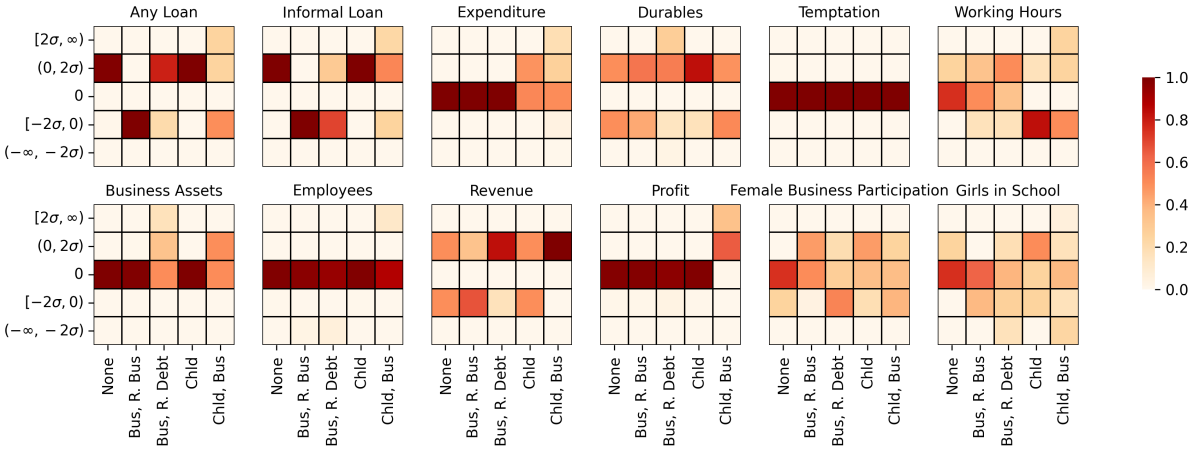


Figure 3.6: This plot visualizes the number of features with a positive, zero, or negative effect, averaged across partitions in the RPS. Each column corresponds to one of the five robust feature profiles described here where the label denotes which features are active (i.e., do not take the lowest level). "None" means that all features are taking these lowest values.

the higher three levels.

To study the impact of access to microcredit, we allow features across treatment and control profiles to be pooled together (see Definition B.1.8). Then, we measure the heterogeneous impact as the conditional average treatment effect, $\text{CATE}(\mathbf{x}) = \mathbb{E}[Y_i(1, \mathbf{x})] - \mathbb{E}[Y_i(0, \mathbf{x})]$. We find the sample mean estimate $\widehat{\text{CATE}}(\mathbf{x}) = \widehat{y}(1, \mathbf{x}) - \widehat{y}(0, \mathbf{x})$ where $\widehat{y}(1, \mathbf{x})$ is the estimated potential outcome for a household assigned to treatment with feature combination $\mathbf{x}$, and $\widehat{y}(0, \mathbf{z})$ is the estimated potential outcome for a household assigned to control with feature combination $\mathbf{z}$. If feature $\mathbf{x}$ is pooled across the treatment and control profiles, then $\widehat{\text{CATE}}(\mathbf{x}) = 0$ indicating no treatment effect heterogeneity in feature $\mathbf{x}$. Otherwise, $\widehat{\text{CATE}}(\mathbf{x}) \neq 0$ indicating treatment effect heterogeneity. Here we present $\widehat{\text{CATE}}(\mathbf{x})$ categorized into five bins based on effect sizes across the RPS, which captures robust (or non-robust) qualitative patterns. We sort $\mathbf{x}$ into various profiles and count the number of features in each profile that have a large positive, small positive, zero, small negative, and large negative effect, averaging the counts over all partitions in the RPS.

There are no robust conclusions to be had for most profiles. Appendix B.9, Figure B.13

shows the full set of profiles and outcomes. We also look at the treatment effect heterogeneity by gender, which has been a point of interest in the literature. For the most part, we find no robust heterogeneity. Here, out of the 16 profiles, we highlight the 5 most robust ones, although arguably only the first is particularly robust. In each of these (and across all 16 profiles), profits and employees are robustly unaffected by microcredit.

1. **Most robust archetype:** Large households, with no previous businesses, in a region with low baseline debt and business presence: (a) take more loans (including informal), (b) consume more, particularly durables but not temptation goods, and supply less labor, (c) increase revenue, but nothing robust to note about business asset accumulation, (d) see no changes in female empowerment.

2. **Other archetypes exhibiting some robustness:** (a) small households, with entrepreneurial experience, in a region with low baseline debt but high business presence, (b) small households, with no entrepreneurial experience, in a region with low baseline debt and business presence, (c) small households, with entrepreneurial experience, in a region with high baseline debt but low business presence, (d) large households, with entrepreneurial experience, in a region with low baseline debt and business presence.

The RPS provides an avenue for the researcher to identify archetypes: profiles where the treatment effects are robust across many outcomes of interest. This provides an avenue for theory-building. It also clearly demonstrates when, for numerous profiles, there is little robustness to be said: the data, without strong priors, cannot really speak to the impacts of microcredit in most cases. The RPS also gives policymakers guidance on robust interventions. For example, if the policymaker considered regions with high baseline debt, since robustly there are no positive profits and half the RPS suggests negative profits, they may not wish for the microcredit firm to enter this market. But in contrast, in other markets, e.g., low debt and business presence, for large non-entrepreneurial families since there are robustly no effects on profits and robustly positive effects on consumption and leisure, they can proceed with confidence.

3.9 Related work

Our work contributes based on ideas that are present in several vibrant literatures. In this section, we contextualize our work in reference to three lines of existing work. We also provide a more thorough discussion of four specific alternative approaches in Appendix B.8.

3.9.1 Related Work on the Rashomon Effect

Our work is, of course, related to literature prior work grappling with the Rashomon effect (Chatfield, 1995; Breiman, 2001b; McAllister, 2007; Tulabandhula and Rudin, 2014; D'Amour et al., 2022; Zhong et al., 2023). One line, reminiscent of dealing with p-hacking, identifies sets of estimands that generate similar objective function values (Marx et al., 2020; Coker et al., 2021; Watson-Daniels et al., 2023) and has been explored in the context of variable importance (Fisher et al., 2019; Dong and Rudin, 2020). Model multiplicity is now being recognized as an important problem in fields such as fairness and causal inference (Black et al., 2022; Pawelczyk et al., 2020; Kobylińska et al., 2023). The most related is Xin et al. (2022), who identify ϵ-Rashomon sets and a decision tree algorithm to enumerate the set of estimands (trees) that have squared loss smaller than a threshold slightly higher than that of a reference model.

Our work considers the Rashomon effect when addressing pressing questions in statistical inference and decision-making. We develop a Bayesian framework and define the RPS as the set of models with high posterior probability. This framework allows for unified inference across partitions and for specific effects. We can also provide a bound on the error in estimating the posterior using only the RPS. We show how to practically estimate the effects of different policies or feature combinations using the RPS and show that the error vanishes quickly in our empirical simulations. We show how to enumerate the entire RPS in the regression setting using scientifically sensible restrictions to cut down our search space. Finally, we formalize the notion of simple models using the ℓ_0 penalty as a sparsity constraint. We show in Theorem 3.3.3 that, in the absence of information about the correlation structure

of the parameters, the ℓ_0 penalty is minimax optimal. Semenova et al. (2022) hypothesized and showed using empirical simulations that regularization changes the size of the RPS. We establish and prove this relationship for the ℓ_0 penalty in Theorem 3.5.3.

3.9.2 Related Work on Bayesian Model Uncertainty

In our work, we address uncertainty across plausible models of heterogeneity. Our goal is to identify cases where distinct elements of the factorial have (nearly) indistinguishable outcomes, which we accomplish by creating partitions of the space of covariate interactions (though we generalize our approach to smoothing covariance matrices in Section B.7). In this sense, our setup is reminiscent of other work on Bayesian tree models that leverage priors over partitions, or trees (e.g., Chipman et al. (1998), Denison et al. (1998), Wu et al. (2007) or Bayesian Additive Regression Trees (BART) (Chipman et al., 2010)). Like this work, we put priors over complexity in the space of trees. In contrast to many subsequent papers in this line of work, our goal is not solely or even principally prediction, but the identification of sets of combinations of characteristics that are heterogeneous with respect to the outcome. We use Hasse diagrams that obey admissibility criteria and, critically, our computational approach does not involve sampling from the posterior, but rather identifying partitions that make up the RPS. Enumerating the RPS allows researchers to focus on a set of the highest posterior explanations for heterogeneity while avoiding the computational issues associated with sampling the extremely large space of trees. We also demonstrate in Section B.7 how to extend our framework to functions across pools (see, for example Chipman et al. (2002)). Our approach is also related to Bayesian Model Averaging (BMA), where each element of the model space is inherently meaningful (Raftery et al., 1997; Clyde, 2003). The notion of using a small set of simple models with high posterior probability models arises in Madigan and Raftery (1994) in the context of BMA for graphical models and more generally in Madigan et al. (1996). Unlike BMA, though, the dimension of $\boldsymbol{\beta}$ stays fixed throughout, though there are restrictions on $\boldsymbol{\beta}$ given a particular partition. This feature avoids the need for the computational issues associated with searching the extremely large

space of highly correlated models of different dimensions (Raftery et al., 1997; Hans et al., 2007; Onorante and Raftery, 2016) while preserving interpretability and a unified Bayesian inference framework. Analogously, Tian and He (2009) and Chen and Tian (2014) use this for causal discovery by finding high posterior equivalence classes of causal Bayesian networks.

3.9.3 Related Work on Learning Treatment Heterogeneity with Machine Learning Tools

A rapidly growing literature leverages ideas from machine learning to estimate treatment effect heterogeneity. Our approach is most closely related to Banerjee et al. (2021) (prior work in part by Chandrasekhar and Sankar), which developed the Hasse representation for treatment variant aggregation for a factorial randomized controlled trial. Their technique employs ℓ_1 regularization (Lasso) to pool treatment combinations. To grapple with the correlations in the design matrix from the factorial data, they employ a Puffer transformation (Jia and Rohe, 2015) to satisfy irrepresentability. They only prove that this transformation can be used for fully crossed RCTs since it is not obvious that the conditions are satisfied for generically factorialized covariate data. Our approach differs fundamentally in several ways, beyond taking a Bayesian rather than frequentist approach. First, we focus on robustness and allow for uncertainty in the selected model through the RPS. There is no such approach in their work. Second, regularizing using an ℓ_1 penalty imposes independence across adjacent models. This is exactly the opposite of what we would expect in practice (two treatment conditions with the same drugs at slightly different levels should have related outcomes). Third, their approach to robustness is to perturb the ℓ_1 penalization parameter. But this traces out a limited family of models for two reasons. To see this, notice that fundamentally the Rashomon Effect is about multimodality: the ℓ_1 approach privileges modes that are more consistent with independent priors irrespective of the penalization parameter. So this approach does not address the Rashomon Effect. Further, the Lasso approach in some sense is limited by roughly being a greedy algorithm: misleading local minima can severely affect which model is selected as optimal since it cannot explore robust alternatives. Our robust prior approach immediately takes us to a decision tree strategy that can explore beyond

local minima. Fourth, their regression approach requires recovering irrepresentability from a correlated design matrix, and the techniques are shown to work only in the crossed RCT setting. Our approach with the decision tree strategy and ℓ_0 prior applies to arbitrary factorial data structures since it is agnostic to the correlational structure.

Our work is also related to existing tree-based methods (e.g. the seminal work of Wager and Athey (2018) on causal forests in the context of treatment heterogeneity). Wager and Athey (2018) construct regression trees (every tree corresponding to some partition Π in our language) to describe heterogeneity in the space of covariates and then sample from the distribution over trees to (honestly) estimate conditional average treatment effects. Honesty here refers to an iterative sample splitting strategy to alleviate issues with estimating the tree and then doing inference using the same data. Both their thought experiments and goals are distinct. Beyond being Bayesian (which philosophically addresses "honesty"), our approach departs in several ways. First, in our setting, data are partially ordered, a restriction that is not captured in the regression trees. Second, we have a coherency requirement: without admissibility restrictions, unrestricted decision trees will put considerable mass on and arrive at partitions that are not real-world meaningful and yet be incorporated into their estimator. This is not desirable and we rule this out. Third, the sampling over trees means that there is no guarantee that the selected trees (or Πs) will be high quality. We provide guarantees, by definition of the RPS, on the quality of the selected partitions and we enumerate all of them. Fourth, to understand whether two adjacent feature combinations should be pooled, in their approach each of these queries must be tested individually which quickly runs into multiple hypothesis issues when done in mass. But our approach natively returns a posterior over all partitions and therefore jointly over all poolings. This delivers output amenable to theory-building: "archetypes" of pools that robustly exhibit certain effects. Fifth, the manner in which tree depth is controlled involves ensuring leafs (pools) have enough observations. This is sensible from a certain perspective, but for our purposes corresponds to the statistician having a prior that their data collection process stratifies observations against *unknown* true partition structure. This is both not a reasonable prior on first principles and also one that

changes as one samples more data, since its shape is defined by the data collection itself. Taken together, these features make it impossible to explore multiple potential models for heterogeneity, which is a fundamental goal of our work.

Finally, we contrast our work with recent work in econometrics that uses machine learning "proxies" to study heterogeneity in treatment effects. The logic is that in settings with a moderate-to-high dimensional covariate structure and little information about the relationship between covariates, machine learning tools can effectively capture patterns of how covariates are associated with heterogeneity in treatment effects rather than the exact covariate-based effects. A novel method developed in Chernozhukov et al. (2018), for example, constructs a framework for inference using proxies constructed by an arbitrary machine learning model. After constructing proxies (predictions of the outcome using features flexibly), Chernozhukov et al. (2018) cluster respondents into groups with the highest and lowest treatment efficacy using treatment outcomes predicted based on the proxies. These clusters, which are derived from amalgamating covariates through "black box" machine learning algorithms, can then be related back to observable covariates. We are interested in a different set of goals: rather than finding what features are associated with heterogeneity (and more extreme effects), we want to identify robust poolings.[8] The outputs of the proxy techniques do not lend themselves to addressing these questions nor do they readily provide any comment on robustness for the same reasons as those faced by the causal forests techniques previously discussed. Our work shows that admissibility restrictions and the geometry of the underlying problem provide enough structure to make search in the space of covariates possible and interpretable, alleviating the need to use black box algorithms to summarize relationships between covariates through proxies.

[8]Further, robustness is an exact finite sample n calculation, so in some sense, we are not worried about the high dimensional case of the number of parameters exploding relative to the number of observations.

3.10 Discussion

In this paper, we present an approach for estimating heterogeneity in outcomes based on a set of discrete covariates. We derive a fully Bayesian framework and an algorithm to identify *all* possible pooling across feature combinations with the highest posterior density: the Rashomon Partition Set. We provide bounds on the portion of the posterior captured by models for heterogeneity in this set, allowing researchers to compute posteriors for marginal effects and evaluate specific treatment combinations. Appealing to a Bayesian framework addresses the issue of multiple testing/selection that leaves practitioners with the unappealing choice between invalid inference and procedures such as data splitting that have implications for power, which are particularly problematic in settings where we expect the cost of data collection to be high. Meanwhile, by identifying a set of high posterior models rather than sampling from the entire posterior, we avoid the inefficiency and impracticality of existing Bayesian approaches to model uncertainty. By only considering scientifically plausible pools in a geometry that allows for partial ordering (Hasse diagrams), we substantially reduce the number of possible explanations for heterogeneity without sacrificing flexibility. Additionally, and critically, these choices mean that the resulting high posterior partitions are *interpretable* and useful for researchers and policymakers when designing future interventions or generalizations.

We now highlight two additional philosophical points about our approach. First, our approach is fundamentally generative in the sense that it produces insights that are directly interpretable. As we highlight in our empirical examples, we expect that Rashomon partitions themselves will be of interest for researchers or policymakers. They allow for the identification of the most robust conclusions, settings where policymakers can intervene without worrying about likely negative consequences, and defining "archetypes" for theory-building. In this way, our work contributes to a growing literature in artificial intelligence and machine learning that pushes back on the use of black box algorithms to make high-stakes decisions (see e.g., Rudin (2019)). While machine learning models may be effective

at estimating complex relationships between covariates, they also often do so in ways that obfuscate the influence of particular features (or combinations of features). Our approach presents an alternative that generates insights about sources of treatment effect heterogeneity based on combinations of observed covariates. Our work shows that admissibility restrictions and the geometry of the underlying problem provide enough structure to make search in the space of covariates possible, alleviating the need to use black box algorithms to summarize relationships between covariates through proxies.

Second, our work highlights the aperture that exists between statistical and practical decision-making. We take as given that in many moderate to high dimensional settings there will be interactions between features. With finite data, the result is a set of possible models for heterogeneity whose statistical performance is indistinguishable. Said another way, our work posits that the quest for the "best" statistical model is Sisyphean in essentially any scientifically interesting setting. While this may seem dire, it actually presents an opportunity to involve additional factors beyond model performance that are often critical in practice for making decisions. Amongst models in the RPS, a policymaker could choose based on, for example, implementation cost, equity considerations, or preserving privacy without sacrificing statistical performance.

There are many promising areas for future work in extending the framework we present here. First, we present results in terms of a posterior in a Bayesian framework. We could, however, also construct a similar structure under a frequentist paradigm. In such a setup, we would need to explore a re-splitting strategy (see Wager and Athey (2018), for example) to construct an "honest" set of Hasse diagrams. Furthermore, we could use our approach to identify groups that are systematically underrepresented in randomized trials (see Parikh et al. (2024), for example) and, as a further generalization, to compare results across trials (see for example Meager (2019)). Finally, our computational approach could be more generally valuable in a wide range of settings, in model selection for graphical models or in for discrete model averaging more generally.

Chapter 4

TOWARDS COMPLETE CAUSAL EXPLANATION WITH EXPERT KNOWLEDGE

"They were possible, they simply weren't known. That's the nature of science."

— Navani, *Rhythm of War*

This chapter contains work that is largely adapted from Venkateswaran and Perković (2024). It is joint work with Emilija Perković. This is based upon work supported by the National Science Foundation under Grant No. 2210210. We want to thank Chris Meek and Thomas Richardson for helpful discussions and valuable feedback.

In previous chapters, we implicitly assumed that we can estimate a unique true optimal model when given access to infinite data, i.e., the true model is identified. In this chapter, we ask, what if the true model cannot be identified from observational data alone? Then, one must rely on knowledge obtained otherwise – for example, through scientific properties or interventional studies. We examine how one can prune the set of equivalent models (based on observational distribution) using domain knowledge coherently. The goal is once again to learn how the world *truly* works. This time, we take a causal graphical model perspective.

4.1 Introduction

In the presence of latent variables, we use maximal ancestral graphs (MAGs) to represent causal relationships between the observed variables (Richardson and Spirtes, 2002). More than one MAG may encode the same conditional independence constraints in the observational distribution through the graphical m-separation criterion. Such MAGs are referred to as Markov equivalent. Therefore, from conditional independence constraints present in the observational distribution, we can learn a MAG up to its Markov equivalence class. Any

Markov equivalence class of MAGs can be uniquely represented by a partial ancestral graph (PAG), which we refer to as an essential ancestral graph. Generally, given an essential ancestral graph, some causal relationships are not identified. However, we may know some of these relationships from domain knowledge or by performing interventions. We refer to such information learned without observational data as expert, or background, knowledge.

In this paper, we are concerned with the properties of the Markov equivalence class of MAGs, and proper restrictions of such an equivalence class using expert knowledge. In particular, we consider restricting the Markov equivalence class to all MAGs that contain certain edge marks, similar to the work of Meek (1995) on restricting the Markov equivalence class of DAGs. We seek to recover a partial mixed graph $\mathcal{G}$ that uniquely represents such a restriction of the Markov equivalence class. We refer to this graph as a restricted essential ancestral graph or a PAG with background knowledge.

Previous work on this topic has focused on specific kinds of background knowledge. In particular, one line of work considers local background knowledge (Mooij et al., 2020; Wang et al., 2022, 2023) where all edge marks incident to a node A are oriented at once. This type of background knowledge represents knowledge that can be obtained from an intervention that sets the variable corresponding to A to a fixed value. Another line of work considers so-called tiered background knowledge (Andrews et al., 2020), which is background knowledge where a set of nodes $\mathbf{V}'$ is partitioned into disjoint sets, or tiers, $\mathbf{V}_{\mathbf{1}}^{\mathbf{prime}}, \ldots, \mathbf{V}_{\mathbf{k}}^{\mathbf{prime}}$, $k > 1$ such that the orientations of edges between the nodes in different partitions is known. This setting furthermore, does not allow for confounding to occur between the different tiers. Our aim in this work is to consider a more general class of background knowledge, where we make no restrictions outside of the assumption that we do not consider systems with selection bias Zhang (2008a).

Before considering background knowledge, we first review various Markov equivalence characterizations of MAGs that exist in the literature. We reconcile these characterizations and prove certain properties of restricted and unrestricted Markov equivalence classes. We show that colliders discriminated by some path in a MAG are present in every Markov

equivalent MAG, which was previously conjectured by Ali et al. (2009). Using this property, we also provide a new algorithm for constructing the essential ancestral graph corresponding to a given MAG in Algorithm 4. Then, we provide three new orientation rules that apply in our setting: (i) R4 in Theorem 4.5.6 generalizes existing Rule 4 used in the construction of the essential ancestral graph, (ii) R11 in Lemma 4.5.1 generalizes Rule 4 of Meek (1995) used for adding expert knowledge in directed acyclic graphs (DAGs), and (iii) R12 in Theorem 4.5.2 which is an entirely novel rule. Using these rules, Algorithm 5 shows how to add expert knowledge to essential ancestral graphs. We show that this algorithm is complete in specific settings (Theorems 4.6.5 and 4.6.7). In the absence of a formal proof for completeness of our rules, we provide an algorithm that can verify completeness for a given graph (Algorithm 6). Our simulation results lend evidence to completeness (see Section 4.6.2).

The remainder of this paper is organized as follows: In Section 4.2, we review basic graphical definitions and discuss various kinds of graphs. We discuss various Markov equivalence characterizations in Section 4.3 and formally define background knowledge in Section 4.4. In Section 4.5, we describe our new graphical orientation rules and state our main completeness results in Section 4.6. We provide concluding remarks in Section 4.7. All technical details are provided in the supplementary materials.

4.2 Preliminaries

Throughout the paper we denote sets of nodes in bold (for example $\mathbf{X}$), graphs in calligraphic font (for example $\mathcal{G}$) and nodes in a graph in uppercase letters (for example X).

Nodes and edges. A graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ consists of a set of nodes (variables) $\mathbf{V} = \{X_1, \ldots, X_p\}$ and a set of edges $\mathbf{E}$. We consider simple graphs, meaning that there is at most one edge between any pair of nodes. Two nodes are called *adjacent* if they are connected by an edge. Every edge has two edge marks that are one of arrowheads, tails, or circles. Edges can be *directed* $\rightarrow$, *bi-directed* $\leftrightarrow$, *non-directed* $\circ\!\!-\!\!\circ$, or *partially directed* $\circ\!\!\rightarrow$. We do not allow for undirected edges, $-$. We use $\bullet$ as a stand-in for any of the allowed edge marks. An edge is *into* (*out of*) a node X if the edge has an arrowhead (tail) at X. An arrowhead or tail

edge marks are called *invariant* and circle edge marks are called *variant*.

4.2.1 Paths and Path Properties

Paths. A *path* p from X to Y in $\mathcal{G}$ is a sequence of distinct nodes $\langle X, \ldots, Y \rangle$ in which every pair of successive nodes are adjacent in $\mathcal{G}$. A node V *lies on a path* p if V occurs in the sequence of nodes. If V lies on a path p, then p *contains* V. If $p = \langle X_1, X_2, \ldots, X_k, \rangle, k \geq 2$, then X_1 and X_k are *endpoints* of p, and any other node $X_i, 1 < i < k$, is a *non-endpoint* node on p. The *length* of a path equals the number of edges on the path.

Directed paths, possibly directed paths, and cycles. Let p be a path $p = \langle V_1, \ldots, V_k \rangle, k > 1$ in graph $\mathcal{G}$. Then p is called a *directed path* from V_1 to V_k, if $V_i \to V_{i+1}$ is on p for all $i \in \{1, \ldots k-1\}$, that is p is of the form $V_1 \to \cdots \to V_k$. Similarly, p is called a *possibly directed path* from V_1 to V_k if there is no edge $V_i \leftarrow\!\bullet V_j$, for $1 \leq i < j \leq k$ in $\mathcal{G}$. Note that, compared to previous definitions in the literature, we also disallow $V_i \leftarrow\!\circ V_j$ edges in possibly directed paths. A directed path from V_1 to V_k together with $V_k \to V_1$ forms a *directed cycle* of *length* k. A directed path from V_1 to V_k together with $V_k \bullet\!\!\to V_1$ forms an *almost directed cycle* of *length* k. Note that this is also different from previous definitions in the literature as we also call directed paths with $V_k \circ\!\!\to V_1$ edges as an almost directed cycle.

Ancestral relationships. If $X \to Y$, then X is a *parent* of Y, and Y is a child of X. If there is a directed path from X to Y, then X is an *ancestor* of Y, and Y is a *descendant* of X. We also use the convention that every node is a descendant and ancestor of itself. The sets of parents, descendants and ancestors of X in $\mathcal{G}$ are denoted by $\mathrm{Pa}(X, \mathcal{G})$, $\mathrm{De}(X, \mathcal{G})$ and $\mathrm{An}(X, \mathcal{G})$ respectively. If there is a possibly directed path from X to Y in $\mathcal{G}$, then we say that X is a *possible ancestor* of Y and that Y is a *possible descendant* of X, The sets of possible descendants and possible ancestors of X in $\mathcal{G}$ are denoted by $\mathrm{PossDe}(X, \mathcal{G})$ and $\mathrm{PossAn}(X, \mathcal{G})$ respectively. For a set of nodes $\mathbf{X} \subseteq \mathbf{V}$, we let $\mathrm{Pa}(\mathbf{X}, \mathcal{G}) = \cup_{X \in \mathbf{X}} \mathrm{Pa}(X, \mathcal{G})$, with analogous definitions for $\mathrm{De}(\mathbf{X}, \mathcal{G})$, $\mathrm{An}(\mathbf{X}, \mathcal{G})$, $\mathrm{PossDe}(\mathbf{X}, \mathcal{G})$ and $\mathrm{PossAn}(\mathbf{X}, \mathcal{G})$.

Subsequences and subpaths. A *subsequence* of a path p is a sequence of nodes ob-

tained by deleting some nodes from p without changing the order of the remaining nodes. A subsequence of a path is not necessarily a path. For a path $p = \langle X_1, X_2, \ldots, X_m \rangle$, the *subpath* from X_i to X_k $(1 \leq i \leq k \leq m)$ is the path $p(X_i, X_k) = \langle X_i, X_{i+1}, \ldots, X_k \rangle$.

Colliders, shields and definite non-colliders; Definite status paths. If a path p contains $X_i \bullet\!\!\rightarrow X_j \leftarrow\!\!\bullet X_k$ as a subpath, then X_j is a *collider* on p. A path $\langle X_i, X_j, X_k \rangle$ is an *(un)shielded triple* if X_i and X_k are (not) adjacent. A path is *unshielded* if all successive triples on the path are unshielded. A node X_j is a *definite non-collider* on a path p if there is at least one edge out of X_j on p, or if $X_i \bullet\!\!-\!\!\circ X_j \circ\!\!-\!\!\bullet X_k$ is a subpath of p and $\langle X_i, X_j, X_k \rangle$ is an unshielded triple (Zhang, 2008a). In a mixed graph, we refer to definite non-colliders as *non-colliders*. A node is of *definite status* on a path if it is a collider or a definite non-collider on the path. A path p is of definite status if every non-endpoint node on p is of definite status.

Collider paths and minimal collider paths. A *collider path* is a path consisting of at least three nodes, on which every non-endpoint node is a collider. A collider path p is *minimal* if no subsequence of p is also a collider path.

Discriminating and inducing paths. A path $p = \langle X, Q_1, \ldots Q_k, Y \rangle, k > 1$ from X to Y is a *discriminating path* (Zhang, 2008b) for Q_k in $\mathcal{G}$ if

(i) $p(X, Q_k)$ is a collider path in $\mathcal{G}$, and

(ii) $X \notin \mathrm{Adj}(Y, \mathcal{G})$, and

(iii) $Q_i \in \mathrm{Pa}(Y, \mathcal{G})$ for all $i \in \{1, \ldots, k-1\}$.

A path $p = \langle X, Q_1, \ldots, Q_k, Y \rangle, k > 1$ is an *inducing path* from X to Y in a graph $\mathcal{G}$ if

(i) $X \notin \mathrm{Adj}(Y, \mathcal{G})$, and

(ii) p is a collider path in $\mathcal{G}$, and

(iii) $Q_i \in \text{An}(\{X, Y\}, \mathcal{G})$, for all $i \in \{1, \ldots, k\}$.

Remark. Our definition of an inducing path differs from that of Ali et al. (2009) in that we exclude the case when the inducing path is an edge, called a *primitive inducing path* by Ali et al. (2009).

Blocking and m-separation. A definite status path p between nodes X and Y is *m-connecting*, or *open* given a set of nodes $\mathbf{Z}$ ($X, Y \notin \mathbf{Z}$) if every definite non-collider on p is not in $\mathbf{Z}$, and every collider on p has a descendant in $\mathbf{Z}$ (Richardson and Spirtes, 2002; Zhang, 2008a). Otherwise $\mathbf{Z}$ *blocks* p. If $\mathbf{Z}$ blocks all definite status paths between X and Y, we say that X and Y are m-separated given $\mathbf{Z}$ in $\mathcal{G}$ (Richardson and Spirtes, 2002). Otherwise, X and Y are m-connected given $\mathbf{Z}$ in $\mathcal{G}$. For pairwise disjoint subsets $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ of $\mathbf{V}$ in $\mathcal{G}$, we say that $\mathbf{X}$ and $\mathbf{Y}$ are m-separated given $\mathbf{Z}$ in $\mathcal{G}$ if X and Y are m-separated given $\mathbf{Z}$ in $\mathcal{G}$ for any $X \in \mathbf{X}$ and $Y \in \mathbf{Y}$. Otherwise, $\mathbf{X}$ and $\mathbf{Y}$ are m-connected given $\mathbf{Z}$ in $\mathcal{G}$. If $\mathbf{X}$ and $\mathbf{Y}$ are m-separated by $\mathbf{Z}$ in $\mathcal{G}$, we write $\mathbf{X} \perp_m \mathbf{Y} | \mathbf{Z}$. Otherwise, we write $\mathbf{X} \not\perp_m \mathbf{Y} | \mathbf{Z}$.

4.2.2 Types of Graphs

Mixed and partial mixed graphs. A graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ is a *mixed graph* if it only contains directed and bidirected edges. A graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ is a *partial mixed graph* if it only contains non-directed, partially directed, directed, and bidirected edges.

Induced subgraph, skeleton Let $\mathbf{X} \subseteq \mathbf{V}$ be a node set in a graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$. Then the $\mathbf{X}$ *induced subgraph* of $\mathcal{G}$, labeled $\mathcal{G}_{\mathbf{X}}$ is a graph that consists of vertices $\mathbf{X}$ and edges $\mathbf{E}_{\mathbf{X}}$, which are all edges in $\mathbf{E}$ for which both endpoints are in $\mathbf{X}$. A *skeleton* of a graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ is graph $\mathcal{G}_{\text{skel}} = (\mathbf{V}, \mathbf{E}')$, where $\mathbf{E}'$ is constructed from $\mathbf{E}$ by replacing each edge with a non-directed edge $\circ\!\!-\!\!\circ$. For a partial mixed graph $\mathcal{G}$, the subgraph of $\mathcal{G}$ consisting of all $\circ\!\!-\!\!\circ$ edges is called the *circle component* of $\mathcal{G}$ and labeled as $\mathcal{G}_C$.

Ancestral and maximal graphs, MAGs. A graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ is *ancestral* if it does not contain directed or almost directed cycles. An ancestral mixed graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ is *maximal* if for any pair of non-adjacent nodes $V_1, V_2 \in \mathbf{V}$, there exists a node set $\mathbf{S}$, $V_1, V_2 \notin \mathbf{S}$ such that $V_1 \perp_m V_2 \mid \mathbf{S}$ in $\mathcal{G}$. Equivalently, an ancestral mixed graph is maximal if it does not contain an inducing path $p = \langle X, Q_1, \ldots, Q_k, Y \rangle$, $k > 1$, such that X and Y are not adjacent (Theorem 4.2 of Richardson and Spirtes, 2002). A maximal ancestral mixed graph will be abbreviated as a MAG. Throughout our work, we assume there is no selection bias i.e., there are no undirected "−" edges.

Markov equivalence class of MAGs, essential ancestral graphs. Several MAGs can encode the same m-separation relationships. Such MAGs form a *Markov equivalence class* of graphs. The Markov equivalence class of MAGs can be uniquely represented by a partial mixed graph which we refer to as the *essential ancestral graph* (cf. Andersson et al. (1997)). Other works have also referred to this essential ancestral graph as a *partial ancestral graph (PAG)* (Richardson and Spirtes, 2002; Ali et al., 2009). Any non-circle edge-mark in an essential ancestral graph $\mathcal{G}$ corresponds to that same non-circle edge-mark in every MAG in the Markov equivalence class described by $\mathcal{G}$. Additionally, for every circle mark $X \circ\!\!-\!\!\bullet Y$ in an essential ancestral graph $\mathcal{G}$, the Markov equivalence class described by $\mathcal{G}$ contains a MAG with $X \leftarrow\!\!\bullet Y$ and a MAG with $X -\!\!\bullet Y$ (Zhang, 2008b). The circle component of an essential ancestral graph $\mathcal{G}$, is an induced subgraph of $\mathcal{G}$ (Zhang, 2008b).

4.3 Characterizing the Markov Equivalence Class of Maximal Ancestral Graphs

Beyond the m-separation criterion, there are three additional ways to characterize Markov equivalent MAGs. Spirtes and Richardson (1996) describe a characterization of Markov equivalence through discriminating paths: MAG $\mathcal{M}_1$ is Markov equivalent to MAG $\mathcal{M}_2$ if $\mathcal{M}_1$, and $\mathcal{M}_2$ share the same adjacencies and unshielded colliders, and if a path $\langle V_1, \ldots, V_{k-1}, V_k \rangle, k > 3$ is a discriminating path from V_1 to V_k for V_{k-1} in both $\mathcal{M}_1$ and $\mathcal{M}_2$, then the V_{k-1} is either a collider on both of these paths or a non-collider on both of these paths. Second, Zhao et al.

(2005) state that all Markov equivalent MAGs share the same skeleton and minimal collider paths. This minimal collider path characterization can be seen as a natural generalization of Markov equivalence classes of DAGs, which states that DAGS with the same skeleton and unshielded colliders are Markov equivalent. This characterization is key to several important results in this paper. Finally, Hu and Evans (2020) provide a non-parametric characterization through arrowheads and tails that arise when parameterizing discrete models.

In this section, we bridge the Spirtes and Richardson (1996) and Zhao et al. (2005) interpretations through two results in Theorem 4.3.1 and Theorem 4.3.2.

First, in Theorem 4.3.1, we show that *all* colliders discriminated by some path are invariant in the Markov equivalence class regardless of whether the discriminating path is invariant. This was previously conjectured by Ali et al. (2009).

Theorem 4.3.1 (Conjecture of Ali et al., 2009)**.** *Suppose that* $p = \langle A, Q_1, \ldots, Q_k, B, C\rangle$ *forms a discriminating path for* B *from* A *to* C *in MAG* $\mathcal{M}$*, and that* $\langle Q_k, B, C\rangle$ *is a collider. Then,* $\langle Q_k, B, C\rangle$ *is a collider in every* $\mathcal{M}^*$ *that is Markov equivalent to* $\mathcal{M}$*.*

Next, we turn to transforming a MAG into its essential ancestral graph. One can obtain an essential ancestral graph of a MAG $\mathcal{M}$ by taking the skeleton of $\mathcal{M}$ called $\mathcal{G}$, orienting those edge-marks in $\mathcal{G}$ that make up the non-endpoints of an unshielded collider in the same way as $\mathcal{M}$ and making all other edge-marks circles and then completing the following orientation rules (Spirtes et al., 2000; Zhang, 2008b):

R1 If $A \bullet\!\!\rightarrow B \circ\!\!-\!\!\bullet C$ is in $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ for some nodes $A, B, C \in \mathbf{V}$, and $A \notin \mathrm{Adj}(C, \mathcal{G})$ then orient $B \rightarrow C$.

R2 If $A \rightarrow B \bullet\!\!\rightarrow C$ or $A \bullet\!\!\rightarrow B \rightarrow C$ and $A \bullet\!\!-\!\!\circ C$, then orient $A \bullet\!\!\rightarrow C$.

R3 If $A \bullet\!\!\rightarrow B \leftarrow\!\!\bullet C$, $A \bullet\!\!-\!\!\circ D \circ\!\!-\!\!\bullet C$, $A \notin \mathrm{Adj}(C, \mathcal{G})$ and $D \bullet\!\!-\!\!\circ B$ is in $\mathcal{G}$, then orient $D \bullet\!\!\rightarrow B$.

Zhang-R4 If $p = \langle A, Q_1, \ldots, Q_{k-1}, Q_k, B\rangle$ is a discriminating path for Q_k in $\mathcal{G}$, and if $Q_k \circ\!\!-\!\!\bullet B$ is in $\mathcal{G}$; then if Q_k is in any m-separating set for A and B in $\mathcal{M}$, orient $Q_{k-1} \leftrightarrow Q_k \leftrightarrow B$;

Algorithm 4 MAGtoEssentialAncestralGraph

Input: MAG $\mathcal{M} = (\mathbf{V}, \mathbf{E})$.
Output: Partial mixed graph $\mathcal{G} = (\mathbf{V}, \mathbf{E}')$.
1: Let $\mathcal{G}_{skel}$ denote the skeleton of $\mathcal{M}$
2: Let $\mathcal{G} = \mathcal{G}_{skel}$
3: In $\mathcal{G}$, orient as arrowheads those edge marks that correspond to colliders on minimal collider paths in $\mathcal{M}$
4: Complete orientations according to R1-R3, Zhao-R4, R8-R10 in $\mathcal{G}$
5: **return** $\mathcal{G}$

otherwise, orient $Q_k \to B$.

R8 If $A \to B \to C$ and $A \circ\!\!\to C$ is in $\mathcal{G}$ then orient $A \to C$.

R9 If $A \circ\!\!\to C$ is in $\mathcal{G}$ and $p = \langle A, B, D, \ldots, C \rangle$ is an unshielded possibly directed path in $\mathcal{G}$ such that $B \notin \text{Adj}(C, \mathcal{G})$, then orient $A \to C$.

R10 If $A \circ\!\!\to C$ and $B \to C \leftarrow D$ are in $\mathcal{G}$, and if there are unshielded possibly directed paths $p_1 = \langle A, M_{11}, \ldots, M_{1l} = B \rangle, l \geq 1$ and $p_2 = \langle A, M_{21}, \ldots, M_{2r} = D \rangle, r \geq 1$ and if $M_{11} \neq M_{21}$ and $M_{11} \notin \text{Adj}(M_{21}, \mathcal{G})$, then orient $A \to C$.

Motivated by Zhao et al. (2005) characterization of the Markov equivalence class of MAGs, we introduce a reformulation of orientation rule Zhang-R4 that was independently identified by Wang et al. (2022):

Zhao-R4 If $p = \langle A, Q_1, \ldots, Q_{k-1}, Q_k, B \rangle$ is a discriminating path for Q_k in $\mathcal{G}$, and if $Q_k \circ\!\!\bullet B$ is in $\mathcal{G}$; then orient $Q_k \to B$.

This rephrasing of orientation rule Zhang-R4 lets us introduce Algorithm 4. We show in Theorem 4.3.2 that the output of Algorithm 4 will be the essential ancestral graph of the input MAG.

Theorem 4.3.2. *Let $\mathcal{M} = (\mathbf{V}, \mathbf{E})$ be a MAG and let $\mathcal{G}$ be the output of Algorithm 4 applied to $\mathcal{M}$, that is, $\mathcal{G} =$ `MAGtoEssentialAncestralGraph`$(\mathcal{M})$. Then $\mathcal{G}$ is the essential ancestral graph of $\mathcal{M}$.*

One may worry about algorithmic implementations of Algorithm 4 as the process of computing minimal collider paths seems cumbersome. We now note a result in Lemma 4.3.3 which in combination with Theorem 4.3.1 makes finding edge marks that make up minimal collider paths simpler.

Lemma 4.3.3. *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be an ancestral partial mixed graph. Furthermore, suppose edge orientations in $\mathcal{G}$ are closed under R1, R2, Zhao-R4. Let $p = \langle P_1, P_2, \ldots, P_k \rangle, k \geq 3$ be a minimal collider path in $\mathcal{G}$. Then for every $i \in \{2, \ldots, k-1\}$, one of the following holds:*

(i) $P_{i-1} \bullet\!\!\rightarrow P_i \leftarrow\!\!\bullet P_{i+1}$ *and* $P_{i-1} \notin \mathrm{Adj}(P_{i+1}, \mathcal{G})$*, or*

(ii) $\exists\ l \in \{1, \ldots, i-2\}$*, such that* $P_l \bullet\!\!\rightarrow P_{l+1} \leftrightarrow \cdots \leftrightarrow P_i \leftarrow\!\!\bullet P_{i+1}$ *is a discriminating collider path from P_l to P_{i+1} for P_i, or*

(iii) $\exists\ r \in \{i+2, \ldots k\}$ *such that* $P_r \bullet\!\!\rightarrow P_{r-1} \leftrightarrow \cdots \leftrightarrow P_{i+1} \leftrightarrow P_i \leftarrow\!\!\bullet P_{i-1}$ *is a discriminating collider path from P_r to P_{i-1} for P_i.*

According to Lemma 4.3.3 every non-endpoint node on a minimal collider path p is either a middle node of an unshielded collider on p, or the last collider on a discriminating collider path which is a subpath of p. Theorem 4.3.1 also tells us that all last colliders on discriminating collider paths in a MAG $\mathcal{M} = (\mathbf{V}, \mathbf{E})$ are invariant across the Markov equivalence class. Theorem 4.2 and Corollary 4.4 of Wienöbst et al. (2022) let us find such discriminating collider paths in a MAG in $O(|\mathbf{V}|^3)$ worst-case runtime. Hence, one may implement Step 3 of Algorithm 5 by finding unshielded colliders and discriminating collider paths in a MAG $\mathcal{M}$, and keeping the edgemarks corresponding to the last collider on every such path in $\mathcal{M}$.

4.4 Consistent Background Knowledge and Restricted Essential Ancestral Graphs

In the remainder of this paper, we study the problem restricting a Markov equivalence class of MAGs with background knowledge in the form of edge mark orientations. In this section, we formally introduce the notion of restricting an equivalence class and define the kind of background knowledge we consider. Then, we show some properties of this restricted equivalence class.

Definition 4.4.1 (Representing Graphs)**.** *A MAG $\mathcal{M} = (\mathbf{V}, \mathbf{E})$ is represented by a partial mixed graph $\mathcal{P} = (\mathbf{V}, \mathbf{E}')$, or $\mathcal{P}$ represents $\mathcal{M}$ if*

(i) $\mathcal{P}$ and $\mathcal{M}$ have the same skeleton and the same minimal collider paths, and

(ii) every invariant edge mark in $\mathcal{P}$ is identical in $\mathcal{M}$.

We use $[\mathcal{P}]$ to denote the set of MAGs represented by $\mathcal{P}$.

If $\mathcal{G}$ is an essential ancestral graph, $[\mathcal{G}]$ is the Markov equivalence class of MAGs represented by this essential ancestral graph. Further, by Theorem 2.1 of Zhao et al. (2005) (cf. C.1.2), for any partial mixed graph $\mathcal{G}$, $[\mathcal{G}]$ is necessarily a (possibly empty) subset of some Markov equivalence class.

Definition 4.4.2 (Edge mark orientations and background knowledge)**.** *$\langle\langle A, B\rangle\rangle$ is an edge mark orientation at B if $\langle\langle A, B\rangle\rangle$ denotes one of the following edge marks: $A \bullet\!\!\rightarrow B$, or $A \bullet\!\!- B$. A set of edge mark orientations $\mathcal{K}$ is called background knowledge.*

In this work, we do not allow our partial mixed graphs to contain edge marks of the form $-$, $\circ\!\!-$, or $-\!\!\circ$. Hence, we only consider sets of background knowledge $\mathcal{K}$ such that the following holds: $\langle\langle A, B\rangle\rangle$ of the form $A \bullet\!\!- B$ is in $\mathcal{K}$ if and only if $\langle\langle B, A\rangle\rangle$ of the form $B \bullet\!\!\rightarrow A$ is also in $\mathcal{K}$. The edge mark orientation $A \bullet\!\!\rightarrow B$ implies that the edge mark at B on $\langle A, B\rangle$ needs to be an arrowhead and does not imply anything about the edge mark at A on the

same edge. A bidirected edge $A \leftrightarrow B$ would be represented using two edge mark orientations $A \bullet\!\!\rightarrow B$ and $B \bullet\!\!\rightarrow A$, that is, with the following background knowledge $\mathcal{K} = \{A \bullet\!\!\rightarrow B, B \bullet\!\!\rightarrow A\}$.

Definition 4.4.3 (Consistent Background Knowledge)**.** *Background knowledge $\mathcal{K}$ is consistent with a partial mixed graph $\mathcal{P} = (\mathbf{V}, \mathbf{E})$ if there is a MAG $\mathcal{M}$ represented by $\mathcal{P}$ such that the edge $\langle A, B\rangle$ in $\mathcal{M}$ is oriented as $\langle\langle A, B\rangle\rangle$ for every $\langle\langle A, B\rangle\rangle \in \mathcal{K}$.*

Background knowledge can be used to restrict the Markov equivalence class to better represent the causal constraints we know to be true. Next, we define consistent background knowledge similar to Meek (1995) and background knowledge-restricted equivalence classes.

Definition 4.4.4 (Restricted Markov equivalence class)**.** *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be an essential ancestral graph and $\mathcal{K}$ be some background knowledge consistent with $\mathcal{G}$. Then $[\mathcal{G}]_{\mathcal{K}}$ is a restriction of the Markov equivalence class of MAGs represented by $\mathcal{G}$ to exactly those MAGs for which $\mathcal{K}$ is a set of consistent background knowledge. We call $[\mathcal{G}]_{\mathcal{K}}$ a restricted Markov equivalence class, or more precisely an $\mathcal{K}$-restricted Markov equivalence class.*

Definition 4.4.5 (Restricted essential ancestral graph)**.** *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be an essential ancestral graph and let $\mathcal{K}$ be background knowledge consistent with $\mathcal{G}$. Additionally, let $[\mathcal{G}]_{\mathcal{K}}$ be a restricted Markov equivalence class. Then $\mathcal{G}'$ is a restricted essential ancestral graph, or more precisely an $\mathcal{K}$-restricted essential ancestral graph if*

(i) $\mathcal{G}'$ has the same skeleton and the same minimal collider paths as $\mathcal{G}$,

(ii) a non-circle edge mark in $\mathcal{G}'$ is invariant across the $[\mathcal{G}]_{\mathcal{K}}$, and

(iii) for any circle edge mark in $\mathcal{G}'$ there is at least one MAG in $[\mathcal{G}]_{\mathcal{K}}$ such that this circle is replaced by a tail, and one MAG in $[\mathcal{G}]_{\mathcal{K}}$ where this circle is replaced by an arrowhead.

If $\mathcal{G}'$ is a restricted essential ancestral graph for $[\mathcal{G}]_{\mathcal{K}}$, then by construction of $\mathcal{G}'$, $[\mathcal{G}'] = [\mathcal{G}]_{\mathcal{K}}$. An essential ancestral graph $\mathcal{G}$ is an $\varnothing$-restricted essential ancestral graph.

We note that the below corollary follows from the soundness of orientation rules R1-Zhao-R4 and R8-R10 as shown by Zhang (2008b) and our Theorem 4.3.2 (see also Theorem 1 of Andrews et al. (2020), Theorem 20 of Mooij et al. (2020) and Theorem 2 of Wang et al. (2022)).

Corollary 4.4.6. *Let $\mathcal{G}$ be a restricted essential ancestral graph. Then orientations of $\mathcal{G}$ are closed under R1-Zhao-R4 and R8-R10.*

For certain types of tiered background knowledge $\mathcal{K}$ consistent with an essential ancestral graph $\mathcal{G}$, Andrews et al. (2020) show how to obtain an $\mathcal{K}$-restricted essential ancestral graph. Mooij et al. (2020) describe a causal discovery framework that pools data about the same system from multiple contexts and incorporates background knowledge $\mathcal{K}$ in the form of causal relationships between context variables and system variables. In particular, the edge mark orientations in their $\mathcal{K}$ take the form of $C_i \to X_j$ for certain pairs context and system variables (C_i, X_j). More recently, Wang et al. (2022) consider background knowledge $\mathcal{K}$ that consist of edge mark orientations $\{\langle\langle V_i, X_k\rangle\rangle : \forall\ V_i \in \text{Adj}(X_k)\}$ for some node X_k i.e., all edge mark orientations local for some nodes $\{X_k\}_{k=1}^m$ are known. Using Zhao-R4 along with other orientation rules of Zhang (2008b), they show how to obtain the $\mathcal{K}$-restricted essential ancestral graph.

In contrast to previous work, we provide a more general treatment by allowing any kind of (consistent) edge mark orientations as background knowledge. Before we describe new graphical orientation rules to add such background knowledge, we draw some connections between restricted essential ancestral graphs and existing orientation rules.

4.4.1 Maximality of Partial Mixed Graphs

For any MAG to be represented by a partial mixed graph $\mathcal{P}$, $\mathcal{P}$ must be ancestral and contain no inducing paths. If an ancestral mixed graph contains no inducing paths, then it is maximal (see Corollary 4.4 of Richardson and Spirtes (2002)). We expand the definition of maximality of partial mixed graphs using inducing paths.

Definition 4.4.7 (Possible inducing path)**.** *Let $\mathcal{P} = (\mathbf{V}, \mathbf{E})$ be a partial mixed graph and X and Y be distinct nodes in $\mathcal{P}$. A path $p = \langle X, Q_1, \ldots, Q_k, Y \rangle$, $k > 1$, is a possible inducing path in $\mathcal{P}$ if p is a collider path in $\mathcal{P}$, $X \notin \text{Adj}(Y, \mathcal{P})$, and $Q_i \in \text{PossAn}(\{X, Y\}, \mathcal{P})$ for all $i \in \{1, \ldots, k\}$.*

Definition 4.4.8 (Maximal partial mixed graph)**.** *Let $\mathcal{P} = (\mathbf{V}, \mathbf{E})$ be a partial mixed graph. We say that $\mathcal{P}$ is maximal if $\mathcal{P}$ contains no possible inducing paths.*

In Lemma 4.4.9, we make an important connection between maximality of partial mixed graphs and the orientation rules R1-R3 and Zhao-R4.

Lemma 4.4.9. *Let $\mathcal{P} = (\mathbf{V}, \mathbf{E})$ be an ancestral partial mixed graph such that edge orientations in $\mathcal{P}$ are closed under R1, R2, R3, Zhao-R4. If $\mathcal{P}$ does not contain an inducing path, then $\mathcal{P}$ also does not contain possible inducing paths i.e., $\mathcal{P}$ is maximal.*

This immediately allows us to show that any restricted essential ancestral graph is also both maximal and ancestral.

Corollary 4.4.10. *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a restricted essential ancestral graph. Then $\mathcal{G}$ is a maximal and ancestral partial mixed graph.*

Furthermore, we also have that any ancestral partial mixed graph that is Markov equivalent to some essential ancestral graph is both maximal and represents a (possibly, smaller) class of Markov equivalent MAGs.

Lemma 4.4.11. *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be an essential ancestral graph and let $\mathcal{G}' = (\mathbf{V}, \mathbf{E}')$ be an ancestral partial mixed graph, such that $\mathcal{G}$ and $\mathcal{G}'$ have the same skeleton and minimal collider paths, and every invariant edge mark in $\mathcal{G}$ is identical in $\mathcal{G}'$. Then $\mathcal{G}'$ is maximal and $[\mathcal{G}'] \subseteq [\mathcal{G}]$.*

4.5 Additional Orientation Rules

In this section, we present two graphical orientation rules that are distinct from R1-R3, Zhao-R4, and R8-R10. Using examples, we also demonstrate that both of these rules are

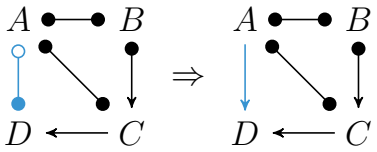


Figure 4.1: A representation of R11.

necessary when we have background knowledge.

4.5.1 Rule 11

We consider a generalization of R4 of Meek (1995), which we call R11 (Lemma 4.5.1).

Lemma 4.5.1. *Let A, B, C, D be distinct nodes in a partial mixed graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$.*

R11 Suppose that the partial mixed graph on the left side of Figure 4.1 is an induced subgraph of $\mathcal{G}$. Then in all MAGs represented by $\mathcal{G}$, the edge $A \circ\!\!-\!\!\bullet D$ is oriented as $A \to D$.

4.5.2 Rule 12

Another new rule, Rule 12 in Theorem 4.5.2, can be seen as a consequence of the ancestral restriction. Rule 12 can be seen as being related to Lemma 7.5 of Maathuis and Colombo (2015) (see Lemma C.1.8 in the Supplement).

Theorem 4.5.2. *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a partial mixed graph.*

R12 Suppose that there is an unshielded path of the form $V_1 \circ\!\!-\!\!\circ V_2 \circ\!\!-\!\!\circ \dots V_{i-1} \circ\!\!-\!\!\bullet V_i$, $i > 2$ in $\mathcal{G}$, as well as a path $V_i \to V_{i+1} \leftrightarrow V_1$ in $\mathcal{G}$. Then all MAGs represented by $\mathcal{G}$ contain $V_1 \leftarrow\!\!\bullet V_2$.

Example 4.5.3. Figure 4.2 shows an example where R12 is applied. The graph on the left-hand side of Figure 4.2 is an essential ancestral graph. Suppose we have the background knowledge that D is a cause of C and B is not a cause of C. This background knowledge is

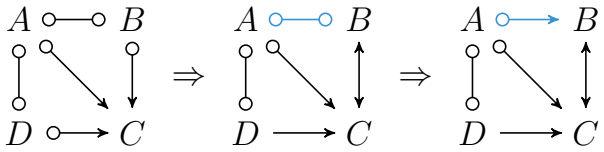


Figure 4.2: Example of orientation R12 in Theorem 4.5.2

added in the form of edge orientations $D \to C$ and $B \leftrightarrow C$ in the center graph of Figure 4.2. However, this center graph does not provide us with a complete causal explanation of the system as the orientations in it are not completed according to R12. To see this, consider paths $p = \langle B, A, D \rangle$ and $q = \langle D, C, B \rangle$. They satisfy R12, so we orient $A \circ\!\!\to B$. □

4.5.3 Revising Rule 4

We now move on to the most complicated of the new rules, which will in fact be a revision of Zhao-R4 (Theorem 4.5.6). To do this, we first define an almost collider path and an almost discriminating path (Definitions 4.5.4 and 4.5.5).

Definition 4.5.4 (Almost collider path)**.** *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a partial mixed graph. Let $p = \langle X = Q_0, Q_1, \ldots Q_k \rangle, k \geq 2$ be a path in $\mathcal{G}$. Then p is an almost collider path if*

(i) *(a)* Q_1 *is a collider on* p*, or*

(b) $Q_0 \bullet\!\!\to Q_1 \circ\!\!\to Q_2$*, and* $Q_0 \bullet\!\!\circ Q_2$ *are in* $\mathcal{G}$*, or*

(c) $Q_0 \bullet\!\!\circ Q_1 \leftarrow\!\!\bullet Q_2$ *and* $Q_0 \circ\!\!\to Q_2$ *are in* $\mathcal{G}$*,*

(ii) *for* $i \in \{2, \ldots, k-2\}$

(a) Q_i *is a collider on* p*, or*

(b) $Q_{i-1} \bullet\!\!\to Q_i \circ\!\!\to Q_{i+1}$*, and* $Q_{i-1} \leftarrow\!\!\circ Q_{i+2}$ *are in* $\mathcal{G}$*, or*

(c) $Q_{i-1} \leftarrow\!\!\circ Q_i \leftarrow\!\!\bullet Q_{i+1}$ *and* $Q_{i-1} \circ\!\!\to Q_{i+1}$ *are in* $\mathcal{G}$*,*

(iii) *(a)* Q_{k-1} *is a collider on* p*, or*

(b) $Q_{k-2} \bullet\!\!\rightarrow Q_{k-1} \circ\!\!-\!\!\bullet Q_k$, *and* $Q_{k-2} \leftarrow\!\!\circ Q_k$ *are in* $\mathcal{G}$, *or*

(c) $Q_{k-2} \leftarrow\!\!\circ Q_{k-1} \leftarrow\!\!\bullet Q_k$ *and* $Q_{k-2} \circ\!\!-\!\!\bullet Q_k$ *are in* $\mathcal{G}$.

Definition 4.5.5 (Almost discriminating path)**.** *Let* $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ *be a partial mixed graph. Let* $p = \langle X = Q_0, Q_1, \dots Q_k, Q_{k+1} = Y \rangle, k \geq 3$ *be a path in* $\mathcal{G}$. *Then* p *is an almost discriminating path for* Q_k *if*

(i) $X \notin \mathrm{Adj}(Y, \mathcal{G})$, *and*

(ii) for all $i \in \{1, \dots, k-1\}$, $Q_i \rightarrow Y$ *is in* $\mathcal{G}$, *and*

(iii) $p(X, Q_k)$ *is an almost collider path.*

Naturally, the definition above subsumes the definition of a discriminating path. This leads us to define a new orientation rule, which can be seen as a generalization of Zhao-R4.

Theorem 4.5.6. *Let* $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ *be a partial mixed graph.*

R4 If path $p = \langle X = Q_0, Q_1, \dots Q_k, Q_{k+1} = Y \rangle, k \geq 3$ *is an almost discriminating path for node* Q_k *in* $\mathcal{G}$ *and if* $Q_k \circ\!\!-\!\!\bullet Y$ *is in* $\mathcal{G}$, *then* $Q_k \rightarrow Y$ *is present in all MAGs represented by* $\mathcal{G}$.

The proof of Theorem 4.5.6 is in Section C.5. We remark that there is a connection between R11 and R4. Namely, R11 can be seen as a special case of R4 if we allow $k = 2$, as $B \notin \mathrm{Adj}(D, \mathcal{G})$, $B \in \mathrm{Adj}(A, \mathcal{G})$, $C \rightarrow D$, and $B \bullet\!\!\rightarrow C \bullet\!\!-\!\!\bullet A \circ\!\!-\!\!\bullet D$ are in $\mathcal{G}$. Hence, if edge $C \bullet\!\!-\!\!\bullet A$ is indeed of the form $C \leftarrow\!\!\bullet A$, then this is $\langle B, C, A \rangle$ is a collider path. Otherwise, if $C \bullet\!\!-\!\!\bullet A$ is of the form $C \circ\!\!-\!\!\bullet A$, then it $\langle B, C, A \rangle$ is an almost collider path.

Example 4.5.7. Figure 4.3 shows a more complex example of the application of R4. The graph on the left-hand side of Figure 4.3(a) is an essential ancestral graph $\mathcal{P}$. Suppose that we want to include background knowledge that A is not a parent of D. Since $\langle A, D \rangle$ is an

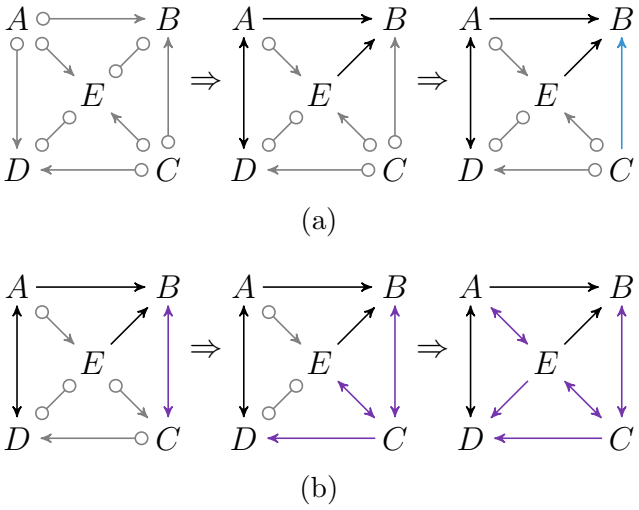


Figure 4.3: Graphs used in Example 4.5.7.

edge in $\mathcal{P}$, we can represent this knowledge as edge orientation $A \leftarrow\!\circ D$. This edge orientation is agreeable with $\mathcal{P}$, so we can include it.

Since $A \circ\!\!\rightarrow D$ is already in $\mathcal{P}$, adding our background knowledge results in $A \leftrightarrow D$. See graph $\mathcal{G}$ in the middle of Figure 4.3(a). Furthermore, since $D \notin \text{Adj}(B, \mathcal{P})$ and since $D \leftrightarrow A \circ\!\!\rightarrow B$, R1 implies $A \rightarrow B$. Now, R11 implies $E \rightarrow B$.

However, orientations in $\mathcal{G}$ are still not completed according to R4 due to path $p = \langle D, A, E, C, B \rangle$ in $\mathcal{G}$, which is an almost discriminating path in $\mathcal{G}$. To see this, consider that $D \notin \text{Adj}(B, \mathcal{G})$ and that $A \rightarrow B$, $E \rightarrow B$ are in $\mathcal{G}$. Furthermore, path $D \leftrightarrow A \circ\!\!\rightarrow E \leftarrow\!\circ C$ is an almost collider path in $\mathcal{G}$ due to the presence of the edge $D \circ\!\!-\!\!\circ E$.

Now, since $p = \langle D, A, E, C, B \rangle$ is an almost discriminating path in $\mathcal{G}$ R4 (Theorem 4.5.6) implies that $C \rightarrow B$ should be oriented. We include this orientation to obtain a partial mixed graph $\mathcal{G}'$ on the right-hand side of Figure 4.3(a). Orientations in $\mathcal{G}'$ are complete according to R1-R3, R4, R8-R10.

To explore why orienting $B \leftrightarrow C$ would lead to an issue, consider Figure 4.3(b). The left-hand-side graph in Figure 4.3(b) contains a graph derived from $\mathcal{G}$ by orienting $B \leftrightarrow C$.

The edge orientation $B \leftrightarrow C$ now implies a few more orientations. For instance, $C \leftrightarrow B$, R2 and $E \rightarrow B \leftrightarrow C$ imply $E \leftrightarrow B$. Furthermore, R1, and $B \leftrightarrow C \circ\!\!\rightarrow D$ imply $C \rightarrow D$. These two orientations are represented in the graph in the middle of Figure 4.3(b).

Next, R11 and thus, R4, imply $E \rightarrow D$. Lastly, R2 and $E \rightarrow D \leftrightarrow A$ imply $E \leftrightarrow A$. These two additional edge orientations are given in the mixed graph $\mathcal{G}^*$ on the right-hand side of Figure 4.3(b).

Graph $\mathcal{G}^*$ is ancestral. However, $\mathcal{G}^*$ contains path q of the form $D \leftrightarrow A \leftrightarrow E \leftrightarrow C \leftrightarrow B$, and $D \notin \text{Adj}(B, \mathcal{P})$ meaning that q is a new minimal collider path in $\mathcal{G}^*$ compared to $\mathcal{G}$. Moreover, edges $A \rightarrow B$, $E \rightarrow B$ $C \rightarrow D$ are in $\mathcal{G}^*$ implying that q is not only a new minimal collider path but also an inducing path in $\mathcal{G}^*$. Hence, $\mathcal{G}^*$ is not even a maximal ancestral graph. □

4.6 Incorporating Background Knowledge

We now introduce our algorithm for adding background knowledge to a partial mixed graph (Algorithm 5). To introduce this algorithm, we must first define admissible edge mark orientations.

Definition 4.6.1 (Admissible edge mark orientation)**.** *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a partial mixed graph and let $\langle\langle A, B\rangle\rangle$ be an edge mark orientation. Then $\langle\langle A, B\rangle\rangle$ is an admissible edge mark orientation for $\mathcal{G}$ if the following hold*

(i) edge $\langle A, B\rangle$ is in $\mathcal{G}$, and

(ii) $\langle\langle A, B\rangle\rangle$ is of the form $A \bullet\!\!\rightarrow B$, $\mathcal{G}$ contain $A \bullet\!\!-\!\!\circ B$ or $A \bullet\!\!\rightarrow B$, or

(iii) $\langle\langle A, B\rangle\rangle$ is of the form $A \bullet\!\!- B$ and $\mathcal{G}$ contains $A \bullet\!\!-\!\!\circ B$, or $A \leftarrow B$.

Remark. Note that all sets of background knowledge $\mathcal{K}$ considered in this manuscript satisfy the following:

Algorithm 5 addBgKnowledge

Input: Partial mixed graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, and a set of admissible background knowledge $\mathcal{K}$.
Output: Partial mixed graph $\mathcal{G}' = (\mathbf{V}, \mathbf{E}')$, or FAIL.

$\mathcal{G}' \leftarrow \mathcal{G}$
for edge mark orientation $\langle\langle A, B\rangle\rangle \in \mathcal{K}$ **do**
    if $\langle\langle A, B\rangle\rangle$ is admissible with $\mathcal{G}$ **then**
        Orient $\langle\langle A, B\rangle\rangle$ in $\mathcal{G}'$
        Complete orientations in $\mathcal{G}'$ according to R1,R2, R4, R8,R10, R11, and R12
    else
        return FAIL
return $\mathcal{G}'$

- for every $\langle\langle A, B\rangle\rangle \in \mathcal{K}$, such that $\langle\langle A, B\rangle\rangle$ is of the form $A \bullet\!\!- B$, we have that $\langle\langle B, A\rangle\rangle \in \mathcal{K}$, where $\langle\langle B, A\rangle\rangle$ is of the form $B \bullet\!\!\rightarrow A$.

Lemma 4.6.2. *Let $\mathcal{G}$ be an essential ancestral graph and $\mathcal{K}$ be a set of background knowledge edge marks consistent with $\mathcal{G}$. Then* `addBgKnowledge`$(\mathcal{G}, \mathcal{K})$ *(Algorithm 5) will not output FAIL.*

Proof. This follows from the definition of consistent background knowledge and soundness of R1, R2, R4, and R8, R10-R12. □

One may be surprised that our Algorithm 5 does not use orientation rules R3 and R9. We show in the next result that these rules are indeed not needed when adding background knowledge to an essential ancestral graph.

Lemma 4.6.3. *Let $\mathcal{G}$ be an essential ancestral graph and $\mathcal{K}$ be a set of background knowledge edge marks consistent with $\mathcal{G}$. Let $\mathcal{G}' =$* `addBgKnowledge`$(\mathcal{G}, \mathcal{K})$*. Then orientations in $\mathcal{G}'$ are complete with respect to R3 and R9.*

4.6.1 Completeness of Orientations Rules with Background Knowledge in Certain Scenarios

Next, we embark on proving the main results of this manuscript. As such, we first define the notion of completeness of edge marks, or completeness of edge orientations in a partial

mixed graph (Definition 4.6.4). We prove the completeness of edge marks for partial mixed graphs with certain kinds of background knowledge. For instance, background knowledge that completely orients all edges of the form $\circ\!\!\rightarrow$ in an essential ancestral graph (Theorem 4.6.5 and Corollary 4.6.6). We also show that if the background knowledge only adds tails to $\circ\!\!\rightarrow$ edges in an essential ancestral graph, this implies tail-completeness for the remaining $\circ\!\!\rightarrow$ edges from the essential ancestral graph, as well as the completeness of all other edge marks (Theorem 4.6.7).

Definition 4.6.4 (Completeness)**.** *Let $\mathcal{G}$ be a partial mixed graph and $\mathcal{P}$ be an essential ancestral graph such that $\mathcal{G}$ and $\mathcal{P}$ have the same skeleton, minimal collider paths, and the set of invariant edge marks in $\mathcal{P}$ is a subset of the set of invariant edge marks in $\mathcal{G}$. We say that the edge marks in $\mathcal{G}$ are complete if for every $A\circ\!\!-\!\!\bullet B$ edge in $\mathcal{G}$, there are two MAGs $\mathcal{M}_1$ and $\mathcal{M}_2$ represented by $\mathcal{G}$ containing the edges $A \rightarrow B$ and $A \leftarrow\!\!\bullet B$ respectively such that $\mathcal{M}_1, \mathcal{M}_2 \in [\mathcal{P}]$.*

Theorem 4.6.5. *Let $\mathcal{G}$ be an essential ancestral graph, and let $\mathcal{G}'$ be an ancestral partial mixed graph and such that $\mathcal{G}$ and $\mathcal{G}'$ have the same skeleton, the same set of minimal collider paths, and all invariant edge marks in $\mathcal{G}$ exist and are identical in $\mathcal{G}'$. Suppose furthermore, that edge mark orientations in $\mathcal{G}'$ are closed under R1-R4, R8-R12 and that every edge $A\circ\!\!\rightarrow B$ in $\mathcal{G}$ corresponds either to $A \rightarrow B$ or $A \leftrightarrow B$ in $\mathcal{G}'$. Then $\mathcal{G}'$ is a restricted essential ancestral graph.*

Corollary 4.6.6. *Let $\mathcal{G}$ be an essential ancestral graph and $\mathcal{K}$ be a set of background knowledge edge marks consistent with $\mathcal{G}$. Let $\mathcal{G}' = \texttt{addBgKnowledge}(\mathcal{G}, \mathcal{K})$. If every edge of the form $A\circ\!\!\rightarrow B$ in $\mathcal{G}$ corresponds to $A \rightarrow B$ or $A \leftrightarrow B$ in $\mathcal{G}'$, then $\mathcal{G}'$ is a $\mathcal{K}$-restricted essential ancestral graph.*

Theorem 4.6.7. *Let $\mathcal{G}$ be an essential ancestral graph and $\mathcal{K}$ be a set of background knowledge edge marks consistent with $\mathcal{G}$. Let $\mathcal{G}' = \texttt{addBgKnowledge}(\mathcal{G}, \mathcal{K})$. Furthermore, let $\mathcal{G}_C$ denote the circle component of $\mathcal{G}$, and let $\mathcal{G}'_C$ denote the subgraph of $\mathcal{G}'$ corresponding to the*

circle component of $\mathcal{G}$. *If there are no* $A \leftrightarrow B$ *edges in* $\mathcal{G}'$ *that are* $A \circ\!\!\rightarrow B$ *in* $\mathcal{G}$, *then the following hold:*

(i) For any edge $A \circ\!\!-\!\!\circ B$ *in* $\mathcal{G}'_C$, *there are three MAGs* $\mathcal{M}_1$, $\mathcal{M}_2$ *and* $\mathcal{M}_3$ *represented by* $\mathcal{G}'$ *such that* $A \rightarrow B$ *is in* $\mathcal{M}_1$, $A \leftarrow B$ *is in* $\mathcal{M}_2$, *and* $A \leftrightarrow B$ *is in* $\mathcal{M}_3$.

(ii) For any edge $A \circ\!\!\rightarrow B$ *in* $\mathcal{G}'_C$, *there are two MAGs* $\mathcal{M}_1$ *and* $\mathcal{M}_2$ *represented by* $\mathcal{G}'$ *such that* $A \rightarrow B$ *is in* $\mathcal{M}_1$, *and* $A \leftrightarrow B$ *is in* $\mathcal{M}_2$.

(iii) For any edge $A \circ\!\!\rightarrow B$ *in* $\mathcal{G}'$ *that is not in* $\mathcal{G}'_C$, *there is a MAG* $\mathcal{M}_1$ *represented by* $\mathcal{G}'$ *such that* $A \rightarrow B$ *is in* $\mathcal{M}_1$.

4.6.2 General Completeness of Orientation Rules with Background Knowledge

Unfortunately, we are unable to prove completeness of Algorithm 5 in a general setting.

We devise an algorithm for checking whether a partial mixed graph is a restricted essential ancestral graph called `verifyCompleteness` (Algorithm 6). This algorithm works better than a brute force method which would enumerate over all possible combinations invariant edge marks for the remaining $\circ$ edge marks by relying on a few key insights.

First, we know that it is sufficient for our enumeration procedure to only consider combinations of edge marks on the remaining $\circ\!\!\rightarrow$ edges from the original essential ancestral graph (Theorem 4.6.5). While exhaustively enumerating over this limited set of edge marks we need to ensure that, we do not introduce directed or almost directed cycles, or new minimal collider paths, and also that there is no other invariant edge mark that appears across the enumeration procedure.

To check that no directed or almost directed cycles are introduced, we show below (Lemma 4.6.8) that it is enough to check that no directed or almost directed cycle of length 3 is introduced. Once we have checked that the ancestral property holds, we move onto the maximal property (Lemma 4.4.11), that is checking for no new minimal collider paths.

To perform this check, we rely on the insight of Lemma 4.3.3 (see also Corollary C.6.3), as well as Theorem 4.3.1 and a remarkable result of Wienöbst et al. (2022) which we generalize to our setting. Namely, we show in Lemma 4.6.9 that no new minimal collider paths are introduced as long as no new unshielded colliders or discriminating collider paths are added in the enumeration process.

Algorithm 6 verifyCompleteness

Input: Essential ancestral graph $\mathcal{G}$, background knowledge $\mathcal{K}$, and a partial mixed graph $\mathcal{G}'$, such that $\mathcal{G}' = \texttt{addBgKnowledge}(\mathcal{G}, \mathcal{K})$

Output: TRUE or FALSE.

$\mathcal{A}_{\mathcal{G}'} \leftarrow \{\langle X, Y\rangle \mid X \circ\!\!\rightarrow Y \in \mathcal{G}' \cap X \circ\!\!\rightarrow Y \in \mathcal{G}\}$
$\mathcal{O} \leftarrow \{X \rightarrow Y \mid \langle X, Y\rangle \in \mathcal{A}_{\mathcal{G}'}\} \cup \{X \leftrightarrow Y \mid \langle X, Y\rangle \in \mathcal{A}_{\mathcal{G}'}\}$
$\mathcal{G}'_C \leftarrow$ subgraph of $\mathcal{G}'$ corresponding to circle component of $\mathcal{G}$
$\mathcal{C} \leftarrow \{\langle X, Y\rangle \mid X \bullet\!\!\circ Y \in \mathcal{G}'_C\}$
$\mathcal{C}_{\bullet\rightarrow} \leftarrow \varnothing$, $\mathcal{C}_{\bullet-} \leftarrow \varnothing$
while $\mathcal{O} \neq \varnothing$ **do**
    $\langle X, Y\rangle \leftarrow \mathcal{O}[1]$
    $\mathcal{G}'' \leftarrow \texttt{addBgKnowledge}(\mathcal{G}', \langle X, Y\rangle)$
    $\mathcal{A}_{\mathcal{G}''} \leftarrow \{\langle U, V\rangle \mid U \circ\!\!\rightarrow V \in \mathcal{G}'' \cap U \circ\!\!\rightarrow V \in \mathcal{G}\}$
    while $\mathcal{A}_{\mathcal{G}''} \neq \varnothing$ **do**
        $\langle U, V\rangle \leftarrow \mathcal{A}_{\mathcal{G}''}[1]$
        if some orientation of $\langle U, V\rangle$ exists in $\mathcal{O}$ **then**
            $\langle U', V'\rangle \leftarrow$ one orientation of $\langle U, V\rangle$ in $\mathcal{O}$
        else
            $\langle U', V'\rangle \leftarrow U \leftrightarrow V$
        $\mathcal{G}'' \leftarrow \texttt{addBgKnowledge}(\mathcal{G}'', \langle U', V'\rangle)$
        $\mathcal{A}_{\mathcal{G}''} \leftarrow \{\langle U, V\rangle \mid U \circ\!\!\rightarrow V \in \mathcal{G}'' \cap U \circ\!\!\rightarrow V \in \mathcal{G}\}$
    if there is a directed or almost directed cycle of length 3 in $\mathcal{G}''$ (Lemma 4.6.8), or if conditions (i) or (ii) of Lemma 4.6.9 fail between $\mathcal{G}'$ and $\mathcal{G}''$ **then**
        return FALSE
    $\mathcal{O} \leftarrow \mathcal{O} \setminus \{X \bullet\!\!\bullet Y \in \mathcal{G}'' \mid \langle X, Y\rangle \in \mathcal{A}_{g'}\}$
    $\mathcal{C}_{\bullet\rightarrow} \leftarrow \mathcal{C}_{\bullet\rightarrow} \cup \{X \bullet\!\!\rightarrow Y \mid \langle X, Y\rangle \in \mathcal{G}'_C\}$
    $\mathcal{C}_{\bullet-} \leftarrow \mathcal{C}_{\bullet-} \cup \{X \bullet\!\!- Y \mid \langle X, Y\rangle \in \mathcal{G}'_C\}$
if $\mathcal{C}_{\bullet\rightarrow} \setminus \{\mathcal{C}_{\bullet-} \cap \mathcal{C}\} \neq \varnothing$ **then**
    return FALSE
if $\mathcal{C}_{\bullet-} \setminus \{\mathcal{C}_{\bullet\rightarrow} \cap \mathcal{C}\} \neq \varnothing$ **then**
    return FALSE
return TRUE

Lemma 4.6.8. *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a partial mixed graph and $\mathcal{P}$ be an essential ancestral graph such that they have the same skeleton, and the set of all invariant edge marks in $\mathcal{P}$ is a subset of the invariant edge marks in $\mathcal{G}$. Furthermore, suppose that edge orientations in $\mathcal{G}$ are closed under R2, R8. If there is a directed or almost directed cycle in $\mathcal{G}$, then there is a directed or almost directed cycle of length 3 in $\mathcal{G}$.*

Lemma 4.6.9. *Let $\mathcal{G}' = (\mathbf{V}, \mathbf{E})$ be an ancestral partial mixed graph with edge mark orientations completed under R1, R2, and R4 and $\mathcal{G}$ be an essential ancestral graph such that they have the same skeleton and the set of all invariant edge marks in $\mathcal{G}$ is a subset of the invariant edge marks in $\mathcal{G}'$. Every minimal collider path in $\mathcal{G}$ is also a minimal collider path in $\mathcal{G}$ if and only if:*

(i) All unshielded colliders in $\mathcal{G}'$ are also unshielded colliders in $\mathcal{G}$, and

(ii) for every discriminating collider path $\langle X, Q_1, \ldots, Q_k, B, Y \rangle$, $k \geq 1$ in $\mathcal{G}'$, $B \leftarrow\!\bullet Y$ is in $\mathcal{G}$.

Completeness Verification Algorithm. In Algorithm 6, we present a method to verify completeness of our orientation rules. As input, we obtain the essential ancestral graph $\mathcal{G}$, and a partial mixed graph $\mathcal{G}'$ such that $\mathcal{G}' = \texttt{addBgKnowledge}(\mathcal{G}, \mathcal{K})$ for some background knowledge $\mathcal{K}$ consistent with $\mathcal{G}$. The algorithm returns TRUE if $\mathcal{G}'$ is a $\mathcal{K}$-restricted essential ancestral graph and FALSE otherwise. To verify this, we need to verify that for all $X \circ\!\!\rightarrow Y$ edges outside the subgraph $\mathcal{G}'_C$ (corresponding to the circle component of $\mathcal{G}$), there are two MAGs represented by $\mathcal{G}'$ with edges $X \rightarrow Y$ and $X \leftrightarrow Y$ respectively. These edges and orientations are given by $\mathcal{A}_{\mathcal{G}'}$ and $\mathcal{O}$ respectively. We pick one such orientation of edge $\langle X, Y \rangle$ and add it to $\mathcal{G}'$ using `addBgKnowledge` to obtain $\mathcal{G}''$. We iteratively add orientations to $\mathcal{G}''$ until there are no more $U \circ\!\!\rightarrow V$ edges in $\mathcal{G}''$ that correspond to $U \circ\!\!\rightarrow V$ edges in $\mathcal{G}$. We use Lemma 4.6.8 and Lemma 4.6.9 to verify wither $\mathcal{G}''$ and $\mathcal{G}'$ are Markov equivalent to each other in line 18. If they are not equivalent, we return FAIL.

We repeat this process to the original graph $\mathcal{G}'$ until there are no more orientations left in $\mathcal{O}$. During this process, we keep track of which edge marks in the circle component $\mathcal{G}'_C$ have been oriented into tails and arrowheads. If in the end, we find that any such edgemark has always been oriented as an arrowhead (or a tail) in line 23 (line 25), then we conclude that the $\mathcal{G}'$ is not a $\mathcal{K}$-restricted essential ancestral graph and return FAIL. Otherwise, we return TRUE indicating that the rules are complete for this graph.

Simulation Results. We perform simulations to lend evidence to completeness of rules and demonstrate the runtime of Algorithm 6. In these simulations, we randomly generate an DAG using the Erdoš-Renyi model, $G(n, p)$ where n is the number of nodes and p is the probability of an edge existing between two nodes using the `randomDAG` function from the `pcalg` package (the ordering is given by the indexing of the nodes). We select 10% of source and confounders in this DAG as latent variables and construct the corresponding MAG $\mathcal{M}$ and essential ancestral graph $\mathcal{G}$. We select k% of the non-identified edge marks in $\mathcal{G}$ as background knowledge and add them to $\mathcal{G}$ using Algorithm 5. Then we verify completeness using Algorithm 6 and compute its runtime. For each set of parameters, we simulated 1000 graphs. We found no violations of completeness in our simulations. The runtime for various parameter configurations is shown in Figure 4.4 and we report the size of the non-circle component and total number of non-identified circle edge marks in Tables 4.1 and 4.2 respectively. Our simulations were run using `R` v4.3.0 and `pcalg` v2.7-8 on a CPU with 4 cores and 30 GB RAM limit.

4.7 Discussion

In this paper, we looked at how to incorporate expert knowledge to refine the Markov equivalence class learned from observational data. We considered expert knowledge in the form of edge marks from the true MAG. As opposed to other existing methods in the literature that consider tiered or local knowledge (Andrews et al., 2020; Mooij et al., 2020; Wang et al., 2022, 2023), our approach is more general.

n \ p	**0.05**	**0.10**	**0.25**	**0.50**
10	0.42 (0)	1.33 (0)	4.76 (5)	8.54 (8)
12	0.74 (0)	2.41 (2)	6.74 (6)	10.79 (10)
15	1.47 (2)	4.36 (4)	9.13 (9)	13.90 (12)
20	3.49 (3)	8.71 (8)	12.50 (12)	19.02 (13)
25	6.43 (6)	13.41 (13)	13.82 (13)	27.32 (15)
30	10.26 (10)	18.16 (18)	14.86 (14)	63.65 (56)
35	14.39 (14)	22.83 (23)	15.69 (14)	158.73 (158)
40	19.11 (19)	26.57 (26)	15.94 (15)	298.90 (304)

Table 4.1: Average number of partially directed "$\circ\!\!\rightarrow$" edges for each (n, p) parameter combination. The medians are shown in parentheses.

n \ p	**0.05**	**0.10**	**0.25**	**0.50**
10	3.67 (4)	1.33 (6)	11.29 (11)	17.10 (16)
12	5.37 (0)	8.94 (8)	14.28 (13)	18.90 (17)
15	8.03 (2)	13.02 (13)	16.68 (16)	22.85 (20)
20	13.65 (3)	20.01 (20)	20.31 (20)	27.68 (21)
25	19.69 (20)	26.41 (26)	21.54 (21)	36.38 (26)
30	26.71 (26.5)	32.29 (32)	23.00 (22)	75.25 (65)
35	33.43 (33)	37.63 (37)	23.72 (22.5)	185.74 (181)
40	40.32 (40)	41.70 (42)	24.03 (23)	380.58 (377)

Table 4.2: Average number of all circle "$\circ$" edgemarks for each (n, p) parameter combination. The medians are shown in parentheses.

We provided three graphical orientation rules. Two of which revise previously established rules for causal graphs with and without latent confounders – R11 generalizes R4 of Meek (1995) and R4 generalizes R4 of FCI. The third rule in R12 is novel and we show that it is necessary in Example 4.5.3. Using these, we construct an algorithm to incorporate background knowledge. When all $\circ$ edge marks on $\circ\!\!\rightarrow$ edges within the essential ancestral are determined by background knowledge, we show our algorithm and thus, our set of orientation rules are complete (Theorem 4.6.5 and Corollary 4.6.6). Furthermore, when the background

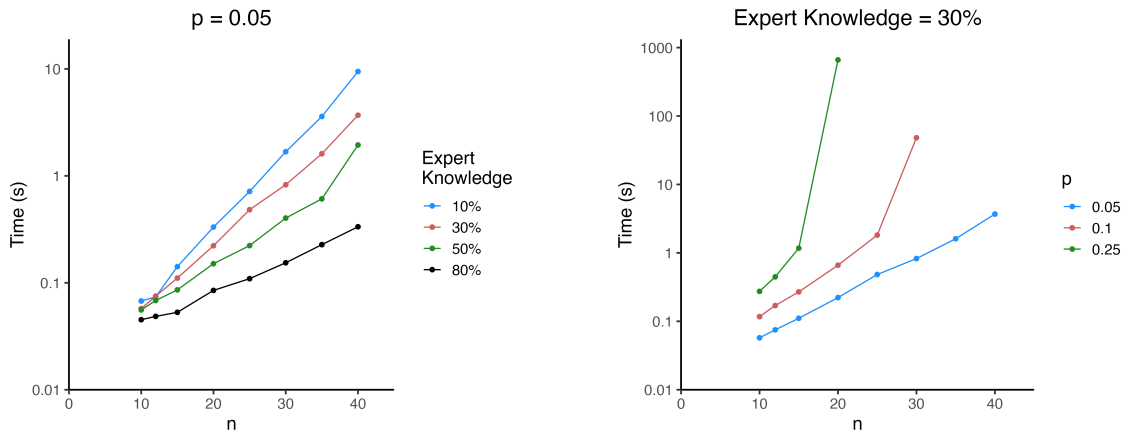


Figure 4.4: Runtime of Algorithm 6 under different simulation configurations. The left panel shows, for a fixed sparsity p, how the runtime decreases as we add more background knowledge. The right panel shows, for a fixed amount of background knowledge, how the runtime increases as the graph becomes more dense i.e., p increases.

knowledge only adds tails to $\circ\!\!\rightarrow$ edges within an essential ancestral graph, we show that the determined set of orientation rules and our Algorithm 5 are complete for $\circ$ edge marks within the circle component of the essential ancestral graph and also tail-complete for edge marks outside the circle component (Theorem 4.6.7).

We believe that our algorithm is complete in general and conduct simulation studies that do not find a disagreement with our orientation rules. Still, completeness remains unproven. One reason why is challenging in this setting is that established strategies, such as those of Zhang (2008b) fail. For instance, it is in general not possible to orient all $\circ\!\!\rightarrow$ edges outside of the circle component of the essential ancestral graph into $\rightarrow$ edges or into $\leftrightarrow$ edges, without incurring issues. Moreover, edges of the form $\circ\!\!\rightarrow$ that exist before and after adding background knowledge to an essential ancestral graph, cannot in general be oriented independently of each other or the rest of the graph. We now provide one specific example where any orientation of a partially directed edge leads to additional orientations in the partial mixed graph.

Hence, consider the partial mixed graph $\mathcal{G}'$ in Figure 4.5(a). The essential ancestral

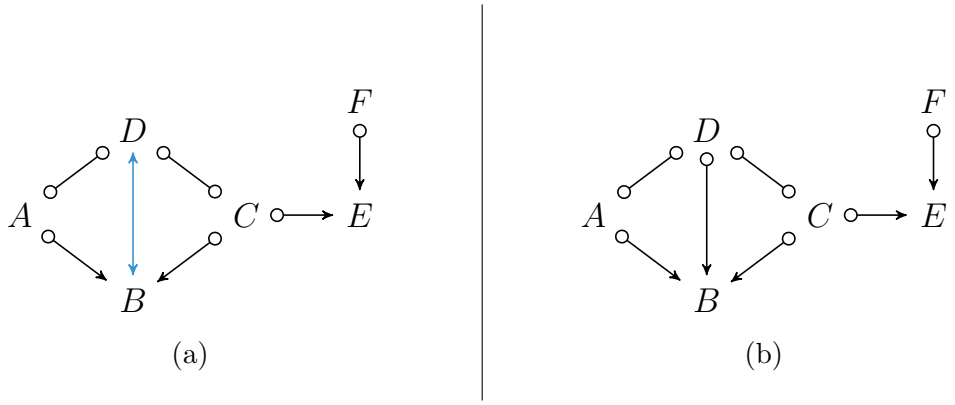


Figure 4.5: (a) A restricted essential ancestral graph $\mathcal{G}'$, (b) essential Ancestral graph $\mathcal{G}$.

graph $\mathcal{G}$ corresponding to $\mathcal{G}'$ is given in Figure 4.5(b). Hence, $\mathcal{G}'$ is obtained from $\mathcal{G}$ by adding background knowledge $\langle\langle B, D\rangle\rangle$ of the form $B \bullet\!\!\rightarrow D$ and completing orientation rules R1-R4, R8-R12. Furthermore, `verifyCompleteness`$(\mathcal{G}, \langle\langle B, D\rangle\rangle, \mathcal{G}'$ returns TRUE, thereby verifying that $\mathcal{G}'$ is a restricted essential ancestral graph.

Consider now edge mark at C on edge $\langle B, C\rangle$ in $\mathcal{G}'$. Orienting $B \leftrightarrow C$ in $\mathcal{G}'$, implies the orientation $C \rightarrow E$ by R1. However, orienting $B \leftarrow C$ in $\mathcal{G}'$ additionally implies $A \leftrightarrow B$ by R12. Hence, in $A \rightarrow B, B \rightarrow C, C \rightarrow E$ do not make up a set of consistent background knowledge for $\mathcal{G}'$ and neither do $A \leftrightarrow B, B \leftrightarrow C, C \leftrightarrow A$.

Our work also contains other results that researchers will find interesting and useful. In Theorem 4.3.1, we show that colliders discriminated by some path in a MAG will be present in every Markov equivalent MAG. This was previously conjectured by Ali et al. (2009). We reconcile different Markov equivalence characterizations to present an algorithm to obtain an essential ancestral graph from a MAG in Algorithm 4. From a pedagogical standpoint, we also provide an alternate proof to Meek (1995)'s completeness result in the case of DAGs.

Chapter 5

CONCLUSION

> "I should thank you not to demean me by insisting my art must be trying to accomplish something. In fact, you shouldn't enjoy art. You should simply admit that it exists, then move on. Anything else is patronizing."
>
> — Wit, *Rhythm of War*

In this dissertation, we looked at three problems relating to identification and estimation.

In Chapter 2, we saw how to perform contact tracing using mortal multi-armed bandits efficiently. We showed a novel lower bound for the Bayesian regret of the mortal bandit, which is identical to that of a standard bandit. We also showed that Pilot sampling and Adaptive Greedy sampling algorithms for the mortal bandit can asymptotically match this lower bound order, and we outlined practical guidelines. Using simulations and empirical data from administrative contact tracing of COVID-19, we showed that Pilot sampling can achieve near-optimal performance and be easily implemented in the field, allowing for feasible contact tracing.

In Chapter 3, we proposed a solution to the so-called Rashomon effect where multiple models are statistically indistinguishable but explain the data very differently. In a factorial feature space, we used a fully Bayesian framework to derive an algorithm that enumerates all partitions with the highest posterior density, i.e., the Rashomon Partition Set (RPS). We obtained approximation errors for calculating the full posterior and showed that the RPS grows only polynomially. Using three empirical examples, we saw how the Rashomon partitions allow researchers to identify and estimate robust conclusions for theory-building and policy-making.

In Chapter 4, we looked at a fundamental identifiability problem that arises in causal discovery from observational data in the presence of latent variables. Many causal relationships are unidentified since we can only learn Markov equivalence classes of maximal ancestral graphs (MAGs). We represented this equivalence class using a uniquely defined essential graph and properly restricted it using expert knowledge in the form of graphical edge marks. To add such expert knowledge, we developed new graphical orientation rules and generalized existing rules. We proved the completeness of these rules for edge marks inside the circle component and provided an algorithm to verify completeness in the general case.

While the works described here are primarily self-contained, they also provide the appropriate tools and frameworks to answer other important questions. We present some of them here, including ongoing work.

5.1 Ongoing and future work

5.1.1 Rashomon sets in high dimensions

In the worst case, the enumeration algorithm described in Chapter 3 will need to check every possible partition for RPS membership if we specify a loose threshold. Such adversarial settings can arise even if the RPS is a strict subset of the space of admissible partitions (Zhang, 1996; Morrison et al., 2016). This raises concerns regarding computational tractability in high-dimensional settings. One way out is dimension reduction. See Van Der Maaten et al. (2009); Ray et al. (2021) for reviews of dimension reduction techniques. However, using a lower dimensional embedding can impact interpretability. Then, how do we guarantee interpretability along with computational tractability? Perhaps we can select features in the lower dimensional embedding that best represent fewer features in the higher dimension. This ensures the researcher can easily interpret the selected low-dimensional features in the higher dimension. Of course, now we need to be mindful of post-selection inference.

5.1.2 Experiment design

In Chapter 3, we saw how to approximate the full posterior density using just the RPS. Suppose we find that credible intervals for the effects of some policies of interest are too wide. To improve our beliefs, we need to collect more data. One obvious way to do this is by collecting data for just those policies of interest. However, we know that, from partitions in the RPS, some policies are equivalent to others. This leads us to consider two things. First, collecting data directly about the policy of interest might be more expensive than from an equivalent policy. Second, we can improve our statistical power by collecting data from equivalent policies. Therefore, the RPS fundamentally changes how we approach experiment design. This idea is reminiscent of sequential experiment design in clinical trials (Mozgunov and Jaki, 2020) and batched bandits (Perchet et al., 2016). One can also use prior information about the effects of each policy by encoding it into the loss function (in Appendix B.2.2 of Chapter 3, we use diffuse priors).

5.1.3 Transportability

A related question to experiment design is transportability or generalizability – how can we transfer conclusions learned from one sub-population to another? This is also referred to as external validity. See Degtiar and Rose (2023) for a review of methods relating to transportability. The RPS presents an immediate framework for transfer learning. A naive approach would be to learn the RPS for several populations $X_1, \ldots, X_n$, and compare them to see for common patterns across populations. This is also related to our goal in Chapter 3 of learning "archetypes" for theory-building.

5.1.4 Rashomon sets of essential graphs

In Chapter 4, we identified causal relationships given an essential graph. We did not describe how this essential graph was learned to begin with. Given the connections we previously established between RPS and Bayesian model averaging in graphical models (Madigan and

Raftery, 1994), it is natural to ask, can we learn a set of essential graphs corresponding to the highest posterior? Chen and Tian (2014) previously demonstrated this for directed acyclic graphs (DAGs), i.e., in the absence of latent confounders. Using a score-based (using likelihoods) method for search for equivalence classes of MAGs (see, for example, Triantafillou and Tsamardinos (2016), Tsirlis et al. (2018), Rantanen et al. (2021)) to identify the Rashomon set of essential graphs is an obvious step forward (even if challenging) on this front.

5.1.5 Sequential causal discovery

Fast causal inference (FCI) is a constraint-based (using independence tests) algorithm for learning the essential graph (Spirtes et al., 1999). This algorithm is NP-Hard. If we have background knowledge, we should be able to improve the computational tractability of this algorithm by incorporating this knowledge, a divide-and-conquer approach, and our orientation rules. For example, Xie and Geng (2008) and Cai et al. (2013) use a divide-and-conquer approach for learning equivalence classes of DAGs. Such a sequential approach can also be framed in terms of missing data. For example, in the missing data setting, Strobl et al. (2018) look at FCI. Of course, previous work in this area does not assume we have background knowledge and orientation rules, which can considerably speed up the algorithm.

5.1.6 Verifying completeness of orientation rules

In Chapter 4, we showed that the graphical orientation rules are complete for all edges in the circle component of the essential graph. Specifically, Theorem 4.6.7 says that edges outside the circle component are tail-complete. Therefore, the only remaining piece is the completion of arrowheads for edges outside the circle component. Unsurprisingly, this corresponds to the same task that Zhang (2008b) remarks as "the most difficult to fulfill" when proving the completeness of the constraint-based MAG discovery algorithm. Zhang (2008b) presents a concept of arrowhead relevance used to identify all circle edge marks that could be oriented as arrowheads simultaneously. I believe that this is the crucial idea for proving completeness.

5.1.7 Causal effect identification

We can begin answering inference questions once we have a restricted essential graph, as described in Chapter 4. For example, which causal effects can be identified using the essential graph? This amounts to developing necessary and sufficient causal identification criteria. Perković (2020) answers an analogous question in the case without latent confounders.

This directly relates to an experiment design question. Suppose a causal effect of interest is unidentified. The identification criteria tell us which causal edges in the restricted essential graph must be known for identifiability. One might want to perform an interventional experiment to learn that edge (recall that this edge cannot be identified from observational data). Or, we can reverse engineer the set of all possible background knowledge that can identify this edge using our graphical orientation rules. This can be useful because it might be easier to intervene on some of the (reverse-engineered) background knowledge edges rather than directly on the edge of interest.

BIBLIOGRAPHY

A. Aakvik, K. G. Salvanes, and K. Vaage. Measuring heterogeneity in the returns to education using an education reform. *European Economic Review*, 54(4):483–500, 2010.

D. Agarwal. Computational advertising: the linkedin way. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 1585–1586, 2013.

D. Agarwal, B. Long, J. Traupman, D. Xin, and L. Zhang. Laser: A scalable response prediction platform for online advertising. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 173–182, 2014.

D. Agrawal, Y. Pote, and K. S. Meel. Partition function estimation: A quantitative study. *arXiv preprint arXiv:2105.11132*, 2021.

R. Agrawal. The continuum-armed bandit problem. *SIAM journal on control and optimization*, 33(6):1926–1951, 1995.

S. Agrawal and N. Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1. JMLR Workshop and Conference Proceedings, 2012.

J. K. Alder, V. S. Hanumanthu, M. A. Strong, A. E. DeZern, S. E. Stanley, C. M. Takemoto, L. Danilova, C. D. Applegate, S. G. Bolton, D. W. Mohr, et al. Diagnostic utility of telomere length testing in a hospital-based setting. *Proceedings of the National Academy of Sciences*, 115(10):E2358–E2365, 2018.

W. P. Alexander and S. D. Grimshaw. Treed regression. *Journal of Computational and Graphical Statistics*, 5(2):156–175, 1996.

A. R. Ali, T. S. Richardson, P. L. Spirtes, and J. Zhang. Towards characterizing Markov equivalence classes for directed acyclic graphs with latent variables. *Conference on Uncertainty in Artificial Intelligence*, 2005.

R. A. Ali, T. S. Richardson, and P. Spirtes. Markov equivalence for ancestral graphs. *The Annals of Statistics*, 37(5B):2808–2837, 2009.

S. A. Andersson, D. Madigan, and M. D. Perlman. A characterization of markov equivalence classes for acyclic digraphs. *The Annals of Statistics*, 25(2):505–541, 1997.

B. Andrews, P. Spirtes, and G. F. Cooper. On the completeness of causal discovery in the presence of latent confounding with tiered background knowledge. In *International Conference on Artificial Intelligence and Statistics*, pages 4002–4011, 2020.

I. Andrews, T. Kitagawa, and A. McCloskey. Inference on winners. Technical report, National Bureau of Economic Research, 2019.

E. Angelino, N. Larus-Stone, D. Alabi, M. Seltzer, and C. Rudin. Learning certifiably optimal rule lists. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 35–44, 2017.

M. Angelucci, D. Karlan, and J. Zinman. Microcredit impacts: Evidence from a randomized microcredit program placement experiment by compartamos banco. *American Economic Journal: Applied Economics*, 7(1):151–182, 2015.

N. Arinaminpathy, J. Das, T. McCormick, P. Mukhopadhyay, and N. Sircar. Quantifying heterogeneity in SARS-CoV-2 transmission during the lockdown in India. *Medrxiv*, 2020.

S. Athey and G. Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

S. Athey and G. W. Imbens. Machine learning methods that economists should know about. *Annual Review of Economics*, 11:685–725, 2019.

O. Attanasio, B. Augsburg, R. De Haas, E. Fitzsimons, and H. Harmgart. The impacts of microfinance: Evidence from joint-liability lending in mongolia. *American Economic Journal: Applied Economics*, 7(1):90–122, 2015.

P. Auer and N. Cesa-Bianchi. On-line learning with malicious noise and the closure algorithm. *Annals of mathematics and artificial intelligence*, 23:83–99, 1998.

B. Augsburg, R. De Haas, H. Harmgart, and C. Meghir. The impacts of microcredit: Evidence from bosnia and herzegovina. *American Economic Journal: Applied Economics*, 7 (1):183–203, 2015.

J.-M. Baland, R. Somanathan, and L. Vandewalle. Microfinance Lifespans: A Study of Attrition and Exclusion in Self-Help Groups in India. In *India Policy Forum*, volume 4(1), pages 159–210. National Council of Applied Economic Research, 2008.

A. Banerjee, E. Duflo, R. Glennerster, and C. Kinnan. The miracle of microfinance? evidence from a randomized evaluation. *American economic journal: Applied economics*, 7(1):22–53, 2015.

A. Banerjee, A. G. Chandrasekhar, E. Duflo, and M. O. Jackson. Using gossips to spread information: Theory and evidence from two randomized controlled trials. *National Bureau of Economic Research Working Paper*, 2016.

A. Banerjee, E. Breza, E. Duflo, and C. Kinnan. Can microfinance unlock a poverty trap for some entrepreneurs? Technical report, National Bureau of Economic Research, 2019.

A. Banerjee, A. G. Chandrasekhar, S. Dalpath, E. Duflo, J. Floretta, M. O. Jackson, H. Kannan, F. N. Loza, A. Sankar, A. Schrimpf, et al. Selecting the most effective nudge: Evidence from a large-scale experiment on immunization. Technical report, National Bureau of Economic Research, 2021.

A. V. Banerjee and E. Duflo. Giving credit where it is due. *Journal of Economic Perspectives*, 24(3):61–80, 2010.

H. Bastani, K. Drakopoulos, V. Gupta, I. Vlachogiannis, C. Hadjicristodoulou, P. Lagiou, G. Magiorkinis, D. Paraskevis, and S. Tsiodras. Efficient and targeted COVID-19 border testing via reinforcement learning. *Nature*, 599(7883):108–113, 2021.

N. Bau and J. Das. Teacher value added in a low-income country. *American Economic Journal: Economic Policy*, 12(1):62–96, 2020.

M. Bayati, N. Hamidi, R. Johari, and K. Khosravi. Unreasonable effectiveness of greedy algorithms in multi-armed bandit with many arms. *Advances in Neural Information Processing Systems*, 33:1713–1723, 2020.

C. Beeri, R. Fagin, D. Maier, and M. Yannakakis. On the desirability of acyclic database schemes. *Journal of the ACM (JACM)*, 30(3):479–513, 1983.

C. Bénard and J. Josse. Variable importance for causal forests: breaking down the heterogeneity of treatment effects. *arXiv preprint arXiv:2308.03369*, 2023.

O. Besbes, Y. Gur, and A. Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in neural information processing systems*, 27, 2014.

P. J. Bickel and B. Li. Regularization in statistics. *Test*, 15:271–344, 2006.

P. G. Bissiri, C. C. Holmes, and S. G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78 (5):1103–1130, 2016.

E. Black, M. Raghavan, and S. Barocas. Model multiplicity: Opportunities, concerns, and solutions. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 850–863, 2022.

D. F. Blackburn and T. W. Wilson. Antihypertensive medications and blood sugar: theories and implications. *Canadian Journal of Cardiology*, 22(3):229–233, 2006.

L. Bolzoni, L. Real, and G. De Leo. Transmission heterogeneity and control strategies for infectious disease emergence. *PLoS One*, 2(8):e747, 2007.

L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001a.

L. Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, 2001b.

L. Breiman. *Classification and regression trees*. Routledge, 2017.

L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, 1984.

E. Breza, A. G. Chandrasekhar, T. H. McCormick, and M. Pan. Using aggregated relational data to feasibly identify network structure without network data. *American Economic Review*, 110(8):2454–84, 2020.

L. Brown, B. Needham, and J. Ailshire. Telomere length among older us adults: differences by race/ethnicity, gender, and age. *Journal of aging and health*, 29(8):1350–1366, 2017.

C. Browne, H. Gulbudak, and G. Webb. Modeling contact tracing in outbreaks with application to Ebola. *Journal of theoretical biology*, 384:33–49, 2015.

S. Bubeck, N. Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

R. Cai, Z. Zhang, and Z. Hao. SADA: A general framework to support robust causation discovery. In *International conference on machine learning*, pages 208–216. PMLR, 2013.

A. Carpentier and A. K. Kim. Adaptive and minimax optimal estimation of the tail coefficient. *Statistica Sinica*, pages 1133–1144, 2015.

A. Carpentier and M. Valko. Simple regret for infinitely many armed bandits. In *International Conference on Machine Learning*, pages 1133–1141. PMLR, 2015.

I. Cascorbi. Drug interactions – principles, examples and clinical consequences. *Deutsches Ärzteblatt International*, 109(33-34):546, 2012.

Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data, 1999-2002. https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes, 1999-2002. Accessed on 2024-01-12.

D. H. Chae, A. M. Nuru-Jeter, N. E. Adler, G. H. Brody, J. Lin, E. H. Blackburn, and E. S. Epel. Discrimination, racial bias, and telomere length in african-american men. *American journal of preventive medicine*, 46(2):103–111, 2014.

D. Chakrabarti, R. Kumar, F. Radlinski, and E. Upfal. Mortal multi-armed bandits. *Advances in Neural Information Processing Systems*, 21:273–280, 2008.

A. G. Chandrasekhar and M. O. Jackson. A network formation model based on subgraphs. *arXiv preprint arXiv:1611.07658*, 2016.

A. G. Chandrasekhar, H. Larreguy, and J. P. Xandri. Testing models of social learning on networks: Evidence from two experiments. *Econometrica*, 88(1):1–32, 2020.

O. Chapelle and L. Li. An empirical evaluation of Thompson sampling. *Advances in Neural Information Processing Systems*, 24, 2011.

C. Chatfield. Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 158(3):419–444, 1995.

Y. Chen and J. Tian. Finding the k-best equivalence classes of bayesian network structures for model averaging. In *Proceedings of the AAAI conference on artificial intelligence*, volume 28(1), 2014.

V. Chernozhukov, M. Demirer, E. Duflo, and I. Fernandez-Val. Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in India. Technical report, National Bureau of Economic Research, 2018.

R. Chetty, J. N. Friedman, and J. E. Rockoff. Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American economic review*, 104(9):2593–2632, 2014.

H. A. Chipman, E. I. George, and R. E. McCulloch. Bayesian CART model search. *Journal of the American Statistical Association*, 93(443):935–948, 1998.

H. A. Chipman, E. I. George, and R. E. McCulloch. Bayesian treed models. *Machine Learning*, 48:299–320, 2002.

H. A. Chipman, E. I. George, and R. E. McCulloch. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1), 2010.

H. Cho, D. Ippolito, and Y. W. Yu. Contact tracing mobile apps for covid-19: Privacy considerations and related trade-offs. *arXiv preprint arXiv:2003.11511*, 2020.

B. Chugg and D. E. Ho. Reconciling risk allocation and prevalence estimation in public health using batched bandits. *arXiv preprint arXiv:2110.13306*, 2021.

T. Claassen, J. Mooij, and T. Heskes. Learning sparse causal models is not np-hard. In *Conference on Uncertainty in Artificial Intelligence*, 2013.

M. Clyde. Model averaging. *Subjective and objective Bayesian statistics*, pages 636–642, 2003.

J. D. Cohen, S. M. McClure, and A. J. Yu. Should i stay or should i go? how the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1481):933–942, 2007.

B. Coker, C. Rudin, and G. King. A theory of statistical inference for ensuring the robustness of scientific results. *Management Science*, 67(10):6174–6197, 2021.

G. F. Cooper. A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. *Data Mining and Knowledge Discovery*, 1(2):203–224, 1997.

B. Crépon, F. Devoto, E. Duflo, and W. Parienté. Estimating the impact of microcredit on those who take it up: Evidence from a randomized experiment in morocco. *American Economic Journal: Applied Economics*, 7(1):123–150, 2015.

J. M. Currie and W. B. MacLeod. Understanding doctor decision making: The case of depression treatment. *Econometrica*, 88(3):847–878, 2020.

A. D'Amour, K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. D. Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *The Journal of Machine Learning Research*, 23(1):10237–10297, 2022.

L. O. Danquah, N. Hasham, M. MacFarlane, F. E. Conteh, F. Momoh, A. A. Tedesco, A. Jambai, D. A. Ross, and H. A. Weiss. Use of a mobile application for Ebola contact tracing and monitoring in northern Sierra Leone: A proof-of-concept study. *BMC Infectious Diseases*, 19(1):1–12, 2019.

M. E. Darden, D. Dowdy, L. Gardner, B. Hamilton, K. Kopecky, M. Marx, N. W. Papageorge, D. Polsky, K. Powers, E. Stuart, et al. Modeling to inform economy-wide pandemic policy: Bringing epidemiologists and economists together. Technical report, National Bureau of Economic Research, 2021.

I. Degtiar and S. Rose. A review of generalizability and transportability. *Annual Review of Statistics and Its Application*, 10:501–524, 2023.

D. G. Denison, B. K. Mallick, and A. F. Smith. A bayesian CART algorithm. *Biometrika*, 85(2):363–377, 1998.

T. A. DiPrete, A. Gelman, T. McCormick, J. Teitler, and T. Zheng. Segregation in social networks based on acquaintanceship and trust. *American journal of sociology*, 116(4): 1234–1283, 2011.

J. Dong and C. Rudin. Exploring the cloud of variable importance for the set of all good models. *Nature Machine Intelligence*, 2(12):810–824, 2020.

D. Dor and M. Tarsi. A simple algorithm to construct a consistent extension of a partially oriented graph. *Technicial Report R-185, Cognitive Systems Laboratory, UCLA*, page 45, 1992.

E. Duflo. The economist as plumber. *American Economic Review*, 107(5):1–26, 2017.

J. J. Filliben and A. N. Heckbert. Dataplot —- a statistical data analysis software system. *A Public Domain Software Released by NIST, Gaithersburg, MD 20899*, 2002.

A. Fisher, C. Rudin, and F. Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.*, 20(177):1–81, 2019.

P. Flajolet and R. Sedgewick. *Analytic Combinatorics*. Cambridge University Press, 2009.

A. G. Forster, H. G. van de Werfhorst, and T. Leopold. Who benefits most from college? dimensions of selection and heterogeneous returns to higher education in the united states and the netherlands. *Research in Social Stratification and Mobility*, 73:100607, 2021.

R. G. Frank and R. J. Zeckhauser. Custom-made versus ready-to-wear treatments: Behavioral propensities in physicians' choices. *Journal of health economics*, 26(6):1101–1127, 2007.

L. Galluzzi, J. Humeau, A. Buqué, L. Zitvogel, and G. Kroemer. Immunostimulation with chemotherapy in the era of immune checkpoint inhibitors. *Nature reviews Clinical oncology*, 17(12):725–741, 2020.

R.-X. Gao, T.-F. Wu, S.-C. Zhu, and N. Sang. Bayesian inference for layer representation with mixed markov random field. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 213–224. Springer, 2007.

M. Gardner, D. Bann, L. Wiley, R. Cooper, R. Hardy, D. Nitsch, C. Martin-Ruiz, P. Shiels, A. A. Sayer, M. Barbieri, et al. Gender and telomere length: systematic review and meta-analysis. *Experimental gerontology*, 51:15–27, 2014.

A. Gelman. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3):515 – 534, 2006. doi: 10.1214/06-BA117A. URL https://doi.org/10.1214/06-BA117A.

A. T. Geronimus, J. A. Pearson, E. Linnenbringer, A. J. Schulz, A. G. Reyes, E. S. Epel, J. Lin, and E. H. Blackburn. Race-ethnicity, poverty, urban stressors, and telomere length in a detroit community-based sample. *Journal of health and social behavior*, 56(2):199–224, 2015.

J. C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 41(2):148–164, 1979.

T. Graepel, J. Q. Candela, T. Borchert, and R. Herbrich. Web-Scale Bayesian Click-Through Rate Prediction for Sponsored Search Advertising in Microsoft's Bing Search Engine. In *Proceedings of the 27th International Conference on Machine Learning ICML*, pages 13–20. Omnipress, 2010.

H. Grushka-Cohen, R. Cohen, B. Shapira, J. Moran-Gilad, and L. Rokach. A framework for optimizing COVID-19 testing policy using a Multi Armed Bandit approach. *arXiv preprint arXiv:2007.14805*, 2020.

L. H. Gunn and D. B. Dunson. A transformation approach for incorporating monotone or unimodal constraints. *Biostatistics*, 6(3):434–449, 2005.

L. Haan and A. Ferreira. *Extreme value theory: an introduction*, volume 3. Springer, 2006.

V. Hadad, D. A. Hirshberg, R. Zhan, S. Wager, and S. Athey. Confidence intervals for policy evaluation in adaptive experiments. *Proceedings of the national academy of sciences*, 118 (15):e2014602118, 2021.

J. Hadamard. Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton University Bulletin*, pages 49–52, 1902.

T. Hagenaars, C. Donnelly, and N. Ferguson. Spatial heterogeneity and the persistence of infectious diseases. *Journal of theoretical biology*, 229(3):349–359, 2004.

P. R. Hahn, J. S. Murray, and C. M. Carvalho. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056, 2020.

R. Hamad, S. Tuljapurkar, and D. H. Rehkopf. Racial and socioeconomic variation in genetic markers of telomere length: a cross-sectional study of us older adults. *EBioMedicine*, 11: 296–301, 2016.

S. M. Hammer, K. E. Squires, M. D. Hughes, J. M. Grimes, L. M. Demeter, J. S. Currier, J. J. Eron Jr, J. E. Feinberg, H. H. Balfour Jr, L. R. Deyton, et al. A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and cd4 cell counts of 200 per cubic millimeter or less. *New England Journal of Medicine*, 337(11):725–733, 1997.

C. Hans, A. Dobra, and M. West. Shotgun stochastic search for "large p" regression. *Journal of the American Statistical Association*, 102(478):507–516, 2007.

C. Hansen, D. Kozbur, and S. Misra. Targeted undersmoothing. *arXiv preprint arXiv:1706.07328*, 2017.

B. M. Hill. A simple general approach to inference about the tail of a distribution. *The annals of statistics*, pages 1163–1174, 1975.

J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian model averaging: a tutorial (with comments by m. clyde, david draper and ei george, and a rejoinder by the authors. *Statistical science*, 14(4):382–417, 1999.

H. Hsu and F. Calmon. Rashomon capacity: A metric for predictive multiplicity in classification. *Advances in Neural Information Processing Systems*, 35:28988–29000, 2022.

X. Hu, C. Rudin, and M. Seltzer. Optimal sparse decision trees. *Advances in Neural Information Processing Systems*, 32, 2019.

Z. Hu and R. Evans. Faster algorithms for markov equivalence. In *Conference on Uncertainty in Artificial Intelligence*, pages 739–748, 2020.

R. Huerta and L. S. Tsimring. Contact tracing and epidemics control in social networks. *Physical Review E*, 66(5):056115, 2002.

J. M. Hyman, J. Li, and E. A. Stanley. Modeling the impact of random screening and contact tracing in reducing the spread of HIV. *Mathematical Biosciences*, 181(1):17–54, 2003.

A. Jaber, J. Zhang, and E. Bareinboim. Causal identification under markov equivalence: Completeness results. In *International Conference on Machine Learning*, pages 2981–2989, 2019.

F. V. Jensen and F. Jensen. Optimal junction trees. In *Uncertainty Proceedings 1994*, pages 360–366. Elsevier, 1994.

J. Jia and K. Rohe. Preconditioning the lasso for sign consistency. *Electronic Journal of Statistics*, 9:1150–1172, 2015.

D. Karlan and J. A. List. Does price matter in charitable giving? evidence from a large-scale natural field experiment. *American Economic Review*, 97(5):1774–1793, 2007.

D. Karlan and J. Zinman. Expanding credit access: Using randomized supply decisions to estimate the impacts. *The Review of Financial Studies*, 23(1):433–464, 2010.

M. Kasy and A. Sautmann. Adaptive treatment assignment in experiments for policy choice. *Econometrica*, 89(1):113–132, 2021.

J. Kawale, H. H. Bui, B. Kveton, L. Tran-Thanh, and S. Chawla. Efficient thompson sampling for online matrix-factorization recommendation. *Advances in neural information processing systems*, 28, 2015.

H. Kellerer, U. Pferschy, and D. Pisinger. *Knapsack Problems*. Springer Berlin, Heidelberg, 2003.

D. A. Kendrick, H. M. Amman, and M. P. Tucci. Learning about learning in dynamic economic models. In *Handbook of Computational Economics*, volume 3, pages 1–35. Elsevier, 2014.

K. Khare and B. Rajaratnam. Wishart distributions for decomposable covariance graph models. *The Annals of Statistics*, 39(1):514–555, 2011.

J.-h. Kim, M. Vojnovic, and S.-Y. Yun. Rotting infinitely many-armed bandits. In *International Conference on Machine Learning*, pages 11229–11254. PMLR, 2022.

K. Kobylińska, M. Krzyziński, R. Machowicz, M. Adamek, and P. Biecek. Exploration of rashomon set assists explanations for medical data. *arXiv preprint arXiv:2308.11446*, 2023.

M. E. Kretzschmar, G. Rozhnova, M. C. Bootsma, M. van Boven, J. H. van de Wijgert, and M. J. Bonten. Impact of delays on effectiveness of contact tracing strategies for COVID-19: a modelling study. *The Lancet Public Health*, 5(8):e452–e459, 2020.

T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

J. Langford and T. Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. *Advances in neural information processing systems*, 20(1):96–1, 2007.

T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

S. L. Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer Science & Business Media, 2006.

B. Letham, C. Rudin, T. H. McCormick, and D. Madigan. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, 9(3):1350–1371, 2015.

N. Levine, K. Crammer, and S. Mannor. Rotting bandits. *Advances in neural information processing systems*, 30, 2017.

Z. R. Li, T. H. McCormick, and S. J. Clark. Bayesian joint spike-and-slab graphical lasso. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3877–3885, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

Z. R. Li, T. H. McCormick, and S. J. Clark. Using Bayesian latent Gaussian graphical models to infer symptom associations in verbal autopsies. *Bayesian Analysis*, to appear.

J. Lin, C. Zhong, D. Hu, C. Rudin, and M. Seltzer. Generalized and scalable optimal sparse decision trees. In *International Conference on Machine Learning*, pages 6150–6160. PMLR, 2020.

A. Ling and R. S. Huang. Computationally predicting clinical drug combination efficacy with cancer cell line screens and independent drug action. *Nature communications*, 11(1): 5848, 2020.

F. Liu, Z. Zheng, and N. Shroff. Analysis of thompson sampling for graphical bandits without the graphs. *arXiv preprint arXiv:1805.08930*, 2018.

J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, and W. M. Getz. Superspreading and the effect of individual variation on disease emergence. *Nature*, 438(7066):355–359, 2005.

M. Lugo. Sum of the first k binomial coefficients for fixed N. https://mathoverflow.net/a/17236, 10 2017. URL https://mathoverflow.net/q/17236. Version 2017-10-01 accessed on 2023-10-26.

S. M. Lynch, M. Peek, N. Mitra, K. Ravichandran, C. Branas, E. Spangler, W. Zhou, E. D. Paskett, S. Gehlert, C. DeGraffinreid, et al. Race, ethnicity, psychosocial factors, and telomere length in a multicenter setting. *PloS one*, 11(1):e0146723, 2016.

Y. Ma, T.-K. Huang, and J. G. Schneider. Active search and bandits on graphs using sigma-optimality. In *UAI*, volume 542, page 551, 2015.

M. H. Maathuis and D. Colombo. A generalized back-door criterion. *The Annals of Statistics*, 43(3):1060–1088, 2015.

O. J. Maclaren and R. Nicholson. What can be estimated? identifiability, estimability, causal inference and ill-posed inverse problems. *arXiv preprint arXiv:1904.02826*, 2019.

K. Madhawa and T. Murata. Exploring partially observed networks with nonparametric bandits. In *International Conference on Complex Networks and their Applications*, pages 158–168. Springer, 2018.

K. Madhawa and T. Murata. A multi-armed bandit approach for exploring partially observed networks. *Applied Network Science*, 4(1):1–18, 2019.

D. Madigan and A. E. Raftery. Model selection and accounting for model uncertainty in graphical models using occam's window. *Journal of the American Statistical Association*, 89(428):1535–1546, 1994.

D. Madigan, A. E. Raftery, C. T. Volinsky, and J. A. Hoeting. Bayesian model averaging. *Integrating Multiple Learned Models (IMLM-96), (P. Chan, S. Stolfo, and D. Wolpert, eds)*, 1996.

D. Malinsky and P. Spirtes. Estimating bounds on causal effects in high-dimensional and possibly confounded systems. *International Journal of Approximate Reasoning*, 88:371–384, 2017.

J. Marecek. Screening for an infectious disease as a problem in stochastic control. *arXiv preprint arXiv:2011.00635*, 2020.

C. Marx, F. Calmon, and B. Ustun. Predictive multiplicity in classification. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6765–6774. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/marx20a.html.

M. B. Mathur, E. Epel, S. Kind, M. Desai, C. G. Parks, D. P. Sandler, and N. Khazeni. Perceived stress and telomere length: A systematic review, meta-analysis, and methodologic considerations for advancing the field. *Brain, behavior, and immunity*, 54:158–169, 2016.

J. W. McAllister. Model selection and the multiplicity of patterns in empirical data. *Philosophy of Science*, 74(5):884–894, 2007.

T. H. McCormick, M. J. Salganik, and T. Zheng. How many people do you know?: Efficiently estimating personal network size. *Journal of the American Statistical Association*, 105 (489):59–70, 2010.

R. Meager. Understanding the average impact of microcredit expansions: A bayesian hierarchical analysis of seven randomized experiments. *American Economic Journal: Applied Economics*, 11(1):57–91, 2019.

C. Meek. Causal inference and causal explanation with background knowledge. In *Conference on Uncertainty in Artificial Intelligence*, 1995.

M. Meister and J. Kleinberg. Optimizing the order of actions in contact tracing. *arXiv preprint arXiv:2107.09803*, 2021.

J. C. Miller. Epidemic size and probability in populations with heterogeneous infectivity and susceptibility. *Physical Review E*, 76(1):010101, 2007.

J. Mincer. Investment in human capital and personal income distribution. *Journal of Political Economy*, 66(4):281–302, 1958.

J. M. Mooij, S. Magliacane, and T. Claassen. Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 2020.

D. R. Morrison, S. H. Jacobson, J. J. Sauppe, and E. C. Sewell. Branch-and-bound algorithms: A survey of recent advances in searching, branching, and pruning. *Discrete Optimization*, 19:79–102, 2016.

B. R. Moulton. A Bayesian approach to regression selection and estimation, with application to a price index for radio services. *Journal of Econometrics*, 49(1-2):169–193, 1991.

P. Mozgunov and T. Jaki. An information theoretic approach for selecting arms in clinical trials. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(5): 1223–1247, 2020.

F. Murai, D. Rennó, B. Ribeiro, G. L. Pappa, D. Towsley, and K. Gile. Selective harvesting over networks. *Data Mining and Knowledge Discovery*, 32(1):187–217, 2018.

K. Muralidharan, M. Romero, and K. Wüthrich. Factorial designs, model selection, and (incorrect) inference in randomized experiments. *Review of Economics and Statistics*, pages 1–44, 2023.

N. U. Nair, P. Greninger, X. Zhang, A. A. Friedman, A. Amzallag, E. Cortez, A. D. Sahu, J. S. Lee, A. Dastur, R. K. Egan, et al. A landscape of response to drug combinations in non-small cell lung cancer. *Nature Communications*, 14(1):3830, 2023.

M. E. Newman and J. Park. Why social networks are different from other types of networks. *Physical review E*, 68(3):036122, 2003.

L. Nie and V. Ročková. Bayesian bootstrap spike-and-slab lasso. *Journal of the American Statistical Association*, 118(543):2013–2028, 2023.

L. Onorante and A. E. Raftery. Dynamic model averaging in large model spaces using dynamic occam's window. *European Economic Review*, 81:2–14, 2016.

H. Parikh, R. Ross, E. Stuart, and K. Rudolph. Who are we missing? a principled approach to characterizing the underrepresented population. *arXiv preprint arXiv:2401.14512*, 2024.

T. Park and G. Casella. The bayesian lasso. *Journal of the american statistical association*, 103(482):681–686, 2008.

M. Pawelczyk, K. Broelemann, and G. Kasneci. On counterfactual explanations under predictive multiplicity. In J. Peters and D. Sontag, editors, *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pages 809–818. PMLR, 03–06 Aug 2020. URL https://proceedings.mlr.press/v124/pawelczyk20a.html.

V. Perchet, P. Rigollet, S. Chassang, and E. Snowberg. Batched bandit problems. *The Annals of Statistics*, 44(2):660 – 681, 2016. doi: 10.1214/15-AOS1381. URL https://doi.org/10.1214/15-AOS1381.

E. Perković. Identifying causal effects in maximally oriented partially directed acyclic graphs. In J. Peters and D. Sontag, editors, *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pages 530–539. PMLR, 03–06 Aug 2020. URL https://proceedings.mlr.press/v124/perkovic20a.html.

E. Perković, J. Textor, M. Kalisch, and M. H. Maathuis. A complete generalized adjustment criterion. In *Conference on Uncertainty in Artificial Intelligence*, 2015.

E. Perković, J. Textor, M. Kalisch, and M. H. Maathuis. Complete graphical characterization and construction of adjustment sets in markov equivalence classes of ancestral graphs. *Journal of Machine Learning Research*, 18, 2018.

J. J. Pfeiffer, J. Neville, and P. Bennett. Active sampling of networks. In *Proceedings of the ICML 2012 Workshop on Mining and Learning with Graphs*, 2012.

J. J. Pfeiffer III, J. Neville, and P. N. Bennett. Active exploration in networks: using probabilistic relationships for learning and inference. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 639–648, 2014.

J. Pickands III. Statistical inference using extreme order statistics. *the Annals of Statistics*, pages 119–131, 1975.

E. Protsenko, D. Rehkopf, A. A. Prather, E. Epel, and J. Lin. Are long telomeres better than short? relative contributions of genetically predicted telomere length to neoplastic and non-neoplastic disease risk and population health burden. *PloS one*, 15(10):e0240185, 2020.

R. Rabbany, D. Bayani, and A. Dubrawski. Active search of connections for case building and combating human trafficking. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2120–2129, 2018.

A. E. Raftery, D. Madigan, and J. A. Hoeting. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437):179–191, 1997.

K. Rantanen, A. Hyttinen, and M. Järvisalo. Maximal ancestral graph structure learning via exact search. In *Conference on Uncertainty in Artificial Intelligence*, pages 1237–1247, 2021.

P. Ray, S. S. Reddy, and T. Banerjee. Various dimension reduction techniques for high dimensional data analysis: a review. *Artificial Intelligence Review*, 54(5):3473–3515, 2021.

P. B. Reverdy, V. Srivastava, and N. E. Leonard. Modeling human decision making in generalized gaussian multiarmed bandits. *Proceedings of the IEEE*, 102(4):544–571, 2014.

T. Richardson and P. Spirtes. Ancestral graph markov models. *The Annals of Statistics*, 30 (4):962–1030, 2002.

H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

F. Rossiello, D. Jurk, J. F. Passos, and F. d'Adda di Fagagna. Telomere dysfunction in ageing and age-related diseases. *Nature cell biology*, 24(2):135–147, 2022.

D. B. Rubin. Estimation in parallel randomized experiments. *Journal of Educational Statistics*, 6(4):377–401, 1981.

C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

D. Russo and B. Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

D. J. Russo, B. Van Roy, A. Kazerouni, I. Osband, Z. Wen, et al. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

S. Saurabh and S. Prateek. Role of contact tracing in containing the 2014 Ebola outbreak: A Review. *African Health Sciences*, 17(1):225–236, 2017.

E. M. Schwartz, E. T. Bradlow, and P. S. Fader. Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(4):500–522, 2017.

S. L. Scott. Multi-armed bandit experiments in the online service economy. *Applied Stochastic Models in Business and Industry*, 31(1):37–45, 2015.

L. Semenova, C. Rudin, and R. Parr. On the existence of simpler machine learning models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1827–1858, 2022.

J. Seznec, A. Locatelli, A. Carpentier, A. Lazaric, and M. Valko. Rotting bandits are no harder than stochastic ones. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2564–2572. PMLR, 2019.

I. Shalev, S. Entringer, P. D. Wadhwa, O. M. Wolkowitz, E. Puterman, J. Lin, and E. S. Epel. Stress and telomere biology: a lifespan perspective. *Psychoneuroendocrinology*, 38 (9):1835–1842, 2013.

M. Song and H. Zhong. Efficient weighted univariate clustering maps outstanding dysregulated genomic zones in human cancers. *Bioinformatics*, 36(20):5027–5036, 2020.

P. Spirtes and T. Richardson. A polynomial time algorithm for determining dag equivalence in the presence of latent variables and selection bias. In *Proceedings of the 6th International Workshop on Artificial Intelligence and Statistics*, pages 489–500, 1996.

P. Spirtes, C. Meek, and T. S. Richardson. *Computation, Causation and Discovery*, chapter An algorithm for causal inference in the presence of latent variables and selection bias, pages 211–252. MIT Press, 1999.

P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman. *Causation, prediction, and search*. MIT press, second edition, 2000.

N. Srinivas, S. Rachakonda, and R. Kumar. Telomeres and telomere length: a general overview. *Cancers*, 12(3):558, 2020.

M. Strens. A bayesian framework for reinforcement learning. In *ICML*, volume 2000, pages 943–950, 2000.

E. V. Strobl, S. Visweswaran, and P. L. Spirtes. Fast causal inference with non-random missingness by test-wise deletion. *International journal of data science and analytics*, 6 (1):47–62, 2018.

R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

A. Tarozzi, J. Desai, and K. Johnson. The impacts of microcredit: Evidence from ethiopia. *American Economic Journal: Applied Economics*, 7(1):54–89, 2015.

W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

J. Tian. Generating markov equivalent maximal ancestral graphs by single edge replacement. In *Conference on Uncertainty in Artificial Intelligence*, pages 591–598, 2005.

J. Tian and R. He. Computing posterior probabilities of structural features in bayesian networks. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 538–547, Arlington, Virginia, USA, 2009. AUAI Press. ISBN 9780974903958.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

A. N. Tikhonov et al. On the stability of inverse problems. In *Dokl. akad. nauk sssr*, volume 39(5), pages 195–198, 1943.

K. Topley. Computationally efficient bounds for the sum of catalan numbers. *arXiv preprint arXiv:1601.04223*, 2016.

S. Tracà, C. Rudin, and W. Yan. Reducing exploration of dying arms in mortal bandits. In *Uncertainty in Artificial Intelligence*, pages 156–163. PMLR, 2020.

S. Triantafillou and I. Tsamardinos. Constraint-based causal discovery from multiple interventions over overlapping variable sets. *Journal of Machine Learning Research*, 16(1): 2147–2205, 2015.

S. Triantafillou and I. Tsamardinos. Score-based vs constraint-based causal learning in the presence of confounders. In *Cfa@ uai*, pages 59–67, 2016.

C. Triplitt. Drug interactions of medications commonly used in diabetes. *Diabetes Spectrum*, 19(4):202, 2006.

W. T. Trotter. *Combinatorics and partially ordered sets*. Johns Hopkins University Press, 1992.

K. Tsirlis, V. Lagani, S. Triantafillou, and I. Tsamardinos. On scoring maximal ancestral graphs with the max–min hill climbing algorithm. *International Journal of Approximate Reasoning*, 102:74–85, 2018.

R. Tu, C. Zhang, P. Ackermann, K. Mohan, H. Kjellström, and K. Zhang. Causal discovery in the presence of missing data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1762–1770. PMLR, 2019.

T. Tulabandhula and C. Rudin. Robust optimization using machine learning for uncertainty sets. *arXiv preprint arXiv:1407.1097*, 2014.

L. Van Der Maaten, E. O. Postma, H. J. van den Herik, et al. Dimensionality reduction: A comparative review. *Journal of Machine Learning Research*, 10(66-71):13, 2009.

A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

A. Venkateswaran and E. Perković. Towards complete causal explanation with expert knowledge. *arXiv preprint arXiv:2407.07338*, 2024.

A. Venkateswaran, J. Das, and T. H. McCormick. Feasible contact tracing. *arXiv preprint arXiv:2312.05718*, 2023.

A. Venkateswaran, A. Sankar, A. G. Chandrasekhar, and T. H. McCormick. Robustly estimating heterogeneity in factorial data using rashomon partitions. *arXiv preprint arXiv:2404.02141*, 2024.

C. M. Vyas, S. Ogata, C. F. Reynolds, D. Mischoulon, G. Chang, N. R. Cook, J. E. Manson, M. Crous-Bou, I. De Vivo, and O. I. Okereke. Telomere length and its relationships with lifestyle and behavioural factors: variations by sex and race/ethnicity. *Age and ageing*, 50 (3):838–846, 2021.

S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

M. A. Wambaugh, S. T. Denham, M. Ayala, B. Brammer, M. A. Stonhill, and J. C. Brown. Synergistic and antagonistic drug interactions in the treatment of systemic fungal infections. *Elife*, 9:e54160, 2020.

M. Wang, R. S. Herbst, and C. Boshoff. Toward personalized treatment approaches for non-small-cell lung cancer. *Nature medicine*, 27(8):1345–1356, 2021.

T.-Z. Wang, T. Qin, and Z.-H. Zhou. Sound and complete causal identification with latent variables given local background knowledge. *Advances in Neural Information Processing Systems*, 35:10325–10338, 2022.

T.-Z. Wang, T. Qin, and Z.-H. Zhou. Estimating possible causal effects with latent variables via adjustment. In *International Conference on Machine Learning*, pages 36308–36335, 2023.

Y. Wang, J.-Y. Audibert, and R. Munos. Algorithms for infinitely many-armed bandits. *Advances in Neural Information Processing Systems*, 21, 2008.

Y. Wang, I. Yahav, and B. Padmanabhan. Whom to Test? Active Sampling Strategies for Managing COVID-19. *arXiv preprint arXiv:2012.13483*, 2020.

J. Watson-Daniels, D. C. Parkes, and B. Ustun. Predictive multiplicity in probabilistic classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37(9), pages 10306–10314, 2023.

M. Wienöbst, M. Bannach, and M. Liśkiewicz. A new constructive criterion for Markov equivalence of MAGs. In *Uncertainty in Artificial Intelligence*, pages 2107–2116, 2022.

C. M. Wu, E. Schulz, M. Speekenbrink, J. D. Nelson, and B. Meder. Generalization guides human exploration in vast decision spaces. *Nature human behaviour*, 2(12):915–924, 2018.

Y. Wu, H. Tjelmeland, and M. West. Bayesian CART: Prior specification and posterior simulation. *Journal of Computational and Graphical Statistics*, 16(1):44–66, 2007.

J. Wyatt. *Exploration and inference in learning from reinforcement*. PhD thesis, University of Edinburgh. College of Science and Engineering., 1998.

X. Xie and Z. Geng. A recursive method for structural learning of directed acyclic graphs. *The Journal of Machine Learning Research*, 9:459–483, 2008.

R. Xin, C. Zhong, Z. Chen, T. Takagi, M. Seltzer, and C. Rudin. Exploring the whole rashomon set of sparse decision trees. *Advances in Neural Information Processing Systems*, 35:14071–14084, 2022.

R. Zhan, Z. Ren, S. Athey, and Z. Zhou. Policy learning with adaptively collected data. *Management Science*, 2023.

C.-H. Zhang and T. Zhang. A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, 27(4):576–593, 2012.

J. Zhang. *Causal inference and reasoning in causally insufficient systems.* PhD thesis, Citeseer, 2006.

J. Zhang. Causal reasoning with ancestral graphs. *Journal of Machine Learning Research*, 9:1437–1474, 2008a.

J. Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896, 2008b.

J. Zhang and P. Spirtes. A transformational characterization of Markov equivalence for directed maximal ancestral graphs. In *Conference on Uncertainty in Artificial Intelligence*, 2005.

R. Zhang, R. Xin, M. Seltzer, and C. Rudin. Optimal sparse regression trees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37(9), pages 11270–11279, 2023.

W. Zhang. *Branch-and-bound search algorithms and their computational complexity.* University of Southern California, Information Sciences Institute, 1996.

H. Zhao, Z. Zheng, and B. Liu. On the markov equivalence of maximal ancestral graphs. *Science in China Series A: Mathematics*, 48(4):548–562, 2005.

P. Zhao and B. Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

Q. Zhao, H. Wen, Z. Lin, D. Xuan, and N. Shroff. On the accuracy of measured proximity of bluetooth-based contact tracing apps. In *International Conference on Security and Privacy in Communication Systems*, pages 49–60. Springer, 2020.

C. Zhong, Z. Chen, M. Seltzer, and C. Rudin. Exploring and interacting with the set of good sparse generalized additive models. *arXiv e-prints*, pages arXiv–2303, 2023.

Appendix A

SUPPLEMENT TO CHAPTER 2

"The story wasn't about him trying to be a hero. It was about him trying to be a dragon. In which, pointedly, he *failed*."

— Wit, *Rhythm of War*

This appendix contains technical details, proofs, and additional simulations to supplement Chapter 2.

A.1 Technical Details of Adaptive Greedy Sampling

A.1.1 Proof of Theorem 2.3.4

Before we prove Theorem 2.3.4, we first present a useful result.

Theorem A.1.1 (Bayesian Regret for Adaptive Greedy)**.** *Under the assumptions (A1)-(A4), for any $\epsilon \in (0, 1/3)$, the Bayesian regret of the adaptive greedy algorithm is given by*

$$BR_{T,N}(AG) \leq T\left(\mathbb{E}_\Gamma\left[\mathbb{I}(1-\mu \leq \epsilon)\mathbb{P}(\exists t : \widehat{\mu}^t < 1-2\epsilon)\right]\right)^N + 3\epsilon T + N\mathbb{E}_\Gamma\left[\mathbb{I}(1-\mu > 3\epsilon)\min\left\{1 + \min\{T/N_m, L\}(1-\mu)(1-\epsilon) + \frac{3}{C_1(1-\mu-2\epsilon)}, \min\{L,T\}(1-\mu)\right\}\right] \tag{A.1}$$

Proof of Theorem A.1.1. Let $\mu^\star = \max_{i\in N}\mu_i$ and $\Delta_i = \mu^\star - \mu_i$. Fix $\epsilon \in (0, \mu^\star/3)$. In the adaptive greedy sampling strategy, we exploit the arm with the largest sample mean with probability $\max_i \widehat{\mu}_i^t$. And with probability $1 - \max_i \widehat{\mu}_i^t$, we randomly choose an arm to explore. Here, $\widehat{\mu}_i^t$ is the sample mean reward of arm i after time t. Following the idea from

Bayati et al. (2020), we will assume that $\mu^\star = 1$. This will loosen the bound but also make it easier to handle integration over priors.

Let us call all arms i with $\Delta_i \leq \epsilon$ as ϵ-optimal. And all arms i such that $\Delta_i > 3\epsilon$ are called sub-optimal. Following standard bandit literature, let us define a bad event to distinguish the randomness of the distributions from the sampling. The bad event is

$$\mathcal{G}^c = \bigcap_{k:\Delta_k<\epsilon} \left\{\exists t : \widehat{\mu}_k^t < 1 - 2\epsilon\right\}$$

Essentially, the bad event happens when for every ϵ-optimal arm, there is at least one time when its sample mean drops below $1 - 2\epsilon$. So the good event happens when there is at least one ϵ-optimal arm whose sample mean remains above $1 - 2\epsilon$. Using the fact that rewards from different arms are independent,

$$\begin{aligned}\mathbb{P}(\mathcal{G}^c) &= \prod_{i:1-\mu_i\leq\epsilon} \mathbb{P}(\exists t : \widehat{\mu}_i^t < 1 - 2\epsilon) \\ &= \prod_{i=1}^{N} \mathbb{I}(1 - \mu_i \leq \epsilon)\mathbb{P}(\exists t : \widehat{\mu}_i^t < 1 - 2\epsilon)\end{aligned}$$

Next, let us estimate the number of times each arm is pulled under a good event,

$$\mathbb{E}[N_i(T) \mid \mathcal{G}] \leq 1 + \sum_{t=1}^{T} \mathbb{P}(\text{arm } i \text{ is chosen} \mid \mathcal{G})$$

$$\begin{aligned}\mathbb{P}(\text{arm } i \text{ is chosen} \mid \mathcal{G}) = \mathbb{P}(\text{arm } i \text{ is chosen} \mid \text{explored}, \mathcal{G})\mathbb{P}(\text{explored} \mid \mathcal{G}) + \\ \mathbb{P}(\text{arm } i \text{ is chosen through exploitation} \mid \mathcal{G})\end{aligned}$$

Since we are in the good event, there must be at least one ϵ-optimal arm k such that $\widehat{\mu}_k^t \geq 1 - 2\epsilon$ for all t. So if we chose arm i through exploitation, it must be that $\widehat{\mu}_i^t \geq 1 - 2\epsilon$

$$\mathbb{P}(\text{arm } i \text{ is chosen} \mid \mathcal{G}) \leq \frac{1}{N_t}(1 - \max_j \widehat{\mu}_j^t) + \mathbb{P}(\widehat{\mu}_i^t \geq 1 - 2\epsilon)$$

Notice that in a good event, $\max_j \widehat{\mu}_j^t \geq 1 - 2\epsilon$. Further, assume that there are always at least N_m arms left to play. This allows us to write

$$\begin{aligned}
\mathbb{P}(\text{arm } i \text{ is chosen} \mid \mathcal{G}) &\leq \frac{1-\epsilon}{N_m} + \mathbb{P}(\widehat{\mu}_i^t \geq 1 - 2\epsilon) \\
&= \frac{1-\epsilon}{N_m} + \mathbb{P}(\widehat{\mu}_i^t - \mu_i \geq \Delta_i - 2\epsilon) \\
&\leq \frac{1-\epsilon}{N_m} + \exp\left\{-t\frac{(\Delta_i - 2\epsilon)^2}{2}\right\}
\end{aligned}$$

where in the last inequality, we used the fact that $\widehat{\mu}_i^t$ is $1/t$-subgaussian. Going back to the counts,

$$\begin{aligned}
\mathbb{E}[N_i(T) \mid \mathcal{G}] &\leq 1 + \sum_{t=1}^{T} \frac{1-\epsilon}{N_m} + \exp\left\{-t\frac{(\Delta_i - 2\epsilon)^2}{2}\right\} \\
&\leq 1 + T\frac{1-\epsilon}{N_m} + \sum_{t=1}^{\infty} \exp\left\{-t\frac{(\Delta_i - 2\epsilon)^2}{2}\right\} \\
&\leq 1 + \min\{T/N_m, L_i\}(1-\epsilon) + \frac{1}{1 - \exp\left\{\frac{(\Delta_i - 2\epsilon)^2}{2}\right\}} \\
&\leq 1 + \min\{T/N_m, L_i\}(1-\epsilon) + \frac{1}{C_1(\Delta_i - 2\epsilon)^2}
\end{aligned}$$

where in the third inequality we bounded the number of times an arm i can be explored by L_i and in the last line we used the fact that $\exp(-x) \leq 1 - 2C_1 x, C_1 = (1 - \exp(-1))/2$ for $x \in [0, 1]$.

This implies,

$$\begin{aligned}
(1-\mu_i)\mathbb{E}[N_i(T) \mid \mathcal{G}] &\leq 1 - \mu_i + \min\{T/N_m, L_i\}(1-\mu_i)(1-\epsilon) + \frac{1-\mu_i}{C_1(\Delta_i - 2\epsilon)^2} \\
&\leq 1 + \min\{T/N_m, L_i\}(1-\mu_i)(1-\epsilon) + \frac{3}{C_1(\Delta_i - 2\epsilon)}
\end{aligned}$$

as $1 - \mu_i > 3\epsilon \implies 1 - \mu_i \leq 3(1 - 2\epsilon - \mu_i)$. Finally, we will bound the number of pulls on

the arm by $\min\{L_i, T\}$,

$$(1-\mu_i)\mathbb{E}[N_i(T) \mid \mathcal{G}] \leq \min\left\{1 + \min\{T/N_m, L_i\}(1-\mu_i)(1-\epsilon) + \frac{3}{C_1(\Delta_i - 2\epsilon)}, \min\{L_i, T\}(1-\mu_i)\right\}$$

Now,

$$\begin{aligned}
\mathbb{E}[R_T \mid \mathcal{G}] &= \sum_{i:\Delta_i \leq 3\epsilon} \Delta_i \mathbb{E}[N_i(T) \mid \mathcal{G}] + \sum_{i:\Delta_i > 3\epsilon} \Delta_i \mathbb{E}[N_i(T) \mid \mathcal{G}] \\
&\leq 3\epsilon T + \sum_{i:\Delta_i > 3\epsilon} \min\left\{1 + \min\{T/N_m, L_i\}(1-\mu_i)(1-\epsilon) + \frac{3}{C_1(\Delta_i - 2\epsilon)}, \min\{L_i, T\}(1-\mu_i)\right\} \\
&= 3\epsilon T + \sum_{i=1}^{N} \mathbb{I}(1-\mu_i > 3\epsilon) \min\left\{1 + \min\{T/N_m, L_i\}(1-\mu_i)(1-\epsilon) + \frac{3}{C_1(\Delta_i - 2\epsilon)}, \right. \\
&\qquad \left. \min\{L_i, T\}(1-\mu_i)\right\}
\end{aligned}$$

Therefore, the regret is

$$\begin{aligned}
\mathbb{E}[R_T] &= \mathbb{E}[R_T \mid \mathcal{G}^c]\mathbb{P}(\mathcal{G}^c) + \mathbb{E}[R_T \mid \mathcal{G}]\mathbb{P}(\mathcal{G}) \\
&\leq T\mathbb{P}(\mathcal{G}^c) + \mathbb{E}[R_T \mid \mathcal{G}] \\
\implies BR_T &= \mathbb{E}_\Gamma \mathbb{E}[R_T] \\
&\leq T\left(\mathbb{E}_\Gamma\left[\mathbb{I}(1-\mu \leq \epsilon)\mathbb{P}(\exists t : \widehat{\mu}^t < 1 - 2\epsilon)\right]\right)^N + 3\epsilon T + \\
N\mathbb{E}_\Gamma\Big[\mathbb{I}(1-\mu > 3\epsilon) &\min\left\{1 + \min\{T/N_m, L\}(1-\mu)(1-\epsilon) + \frac{3}{C_1(1-\mu-2\epsilon)}, \min\{L, T\}(1-\mu)\right\}\Big]
\end{aligned}$$

□

Proof of Theorem 2.3.4. The general strategy is similar to the one used by Bayati et al. (2020). The regret is,

$$\begin{aligned}
&BR_{T,N}(AG) \leq T\left(\mathbb{E}_\Gamma\left[\mathbb{I}(1-\mu \leq \epsilon)\mathbb{P}(\exists t : \widehat{\mu}^t < 1 - 2\epsilon)\right]\right)^N + 3\epsilon T + \\
&N\mathbb{E}_\Gamma\left[\mathbb{I}(1-\mu > 3\epsilon) \min\left\{1 + \min\{T/N_m, L\}(1-\mu)(1-\epsilon) + \frac{3}{C_1(1-\mu-2\epsilon)}, \min\{L, T\}(1-\mu)\right\}\right]
\end{aligned}$$

Consider the first term. Since μ comes from a γ-regular prior, $\mathbb{P}(\mu > 1-\epsilon) \geq c_{\min}\epsilon^{\gamma}$ for some absolute constant $c_{\min}$. From Lemma A.7.1,

$$\begin{aligned}
\mathbb{E}_{\Gamma}\left[\mathbb{I}(1-\mu \leq \epsilon)\mathbb{P}(\exists t : \widehat{\mu}^t < 1-2\epsilon)\right] &\leq (1-\exp(-0.5)/3)\,\mathbb{E}_{\Gamma}[\mathbb{I}(1-\mu \leq \epsilon)] \\
&\leq c_{\min}\epsilon^{\gamma}\,(1-\exp(-0.5)/3) \\
&\leq 1 - c_{\min}\epsilon^{\gamma}\frac{\exp(-0.5)}{3} \\
&= 1 - c_0\epsilon^{\gamma} \\
&\leq \exp\left\{-c_0\epsilon^{\gamma}\right\} \\
\implies \text{Term 1} &= T\exp\left\{-Nc_0\epsilon^{\gamma}\right\}
\end{aligned}$$

where $c_0 = c_{\min}\exp(-0.5)/3$ and we used $1-x \leq \exp(-x)$. Our strategy now is to control the first term as $\mathcal{O}(1)$ and treat the third term using Lemma A.7.3.

To analyze the third term, we will use the fact that $\min\{f+g, h\} \leq \min\{f,h\} + g$ to pull out the $\min\{T/L_m, L\}(1-\mu)(1-\epsilon)$ outside the outer min operator. Bounding $\min\{T/N_m, L\} \leq L$, we have

$$\mathbb{E}_{\Gamma}\left[\mathbb{I}(1-\mu > 3\epsilon)L(1-\mu)(1-\epsilon)\right] \leq L(1-\epsilon)(1-c_{\min}3^{\gamma}\epsilon^{\gamma})$$

We will also upper bound $\min\{T, L\} \leq T$. So, an application of Lemma A.7.3 takes care of the remaining pieces of the third term.

In particular, if $\gamma = 1$, then the Bayesian regret is upper bounded by

$$T\exp\left\{-Nc_0\epsilon^{\gamma}\right\} + 3\epsilon T + NL(1-\epsilon)(1-c_{\min}3^{\gamma}\epsilon^{\gamma}) + \frac{3C_0}{C_1}N(5+\log(1/\epsilon))$$

where C_0 is the constant in Lemma A.7.3. Now, if we choose $\epsilon^{\gamma} = C_2\log T/(Nc_0)$ where $C_2 \geq 1$, then

$$BR_{N,T} = \mathcal{O}\left(T^{1-C_2} + N^{-1}T\log T + N + \log T + N^{-1}(\log T)^2 + N\log\log T + N\log N\right)$$

$$= \mathcal{O}\left(N^{-1}T\log T + N(\log\log T + \log N) + \log T\right)$$
$$= \tilde{\mathcal{O}}\left(TN^{-1} + N\right)$$

And when $\gamma > 1$, if we choose the same ϵ as above,

$$BR_{N,T} = \mathcal{O}\left(T^{1-C_2} + N^{-1/\gamma}T(\log T)^{1/\gamma} + N + N^{1-1/\gamma}(\log T)^{1/\gamma} + \log T + N^{-1/\gamma}(\log T)^{1+1/\gamma} + N\right)$$
$$= \mathcal{O}(N^{-1/\gamma}(\log T)^{1/\gamma}(T+N) + N + \log T)$$
$$= \tilde{\mathcal{O}}(TN^{-1/\gamma} + N)$$

Finally, when $\gamma < 1$ and choosing the same ϵ,

$$BR_{N,T} = \mathcal{O}\left(N^{-1/\gamma}(\log T)^{1/\gamma}(T+N) + N + \log T + N\min(\sqrt{T}, N^{1/\gamma}(\log T)^{-1/\gamma})^{1-\gamma}\right)$$
$$= \tilde{\mathcal{O}}\left(TN^{-1/\gamma} + N\min(\sqrt{T}, N^{1/\gamma})^{1-\gamma}\right)$$

Thus,

$$BR_{T,N} = \begin{cases} \mathcal{O}\left(N^{-1/\gamma}(\log T)^{1/\gamma}(T+N) + N + \log T + N\min(\sqrt{T}, N^{1/\gamma}(\log T)^{-1/\gamma})^{1-\gamma}\right), & \gamma < 1 \\ \mathcal{O}\left(N^{-1}T\log T + N(\log\log T + \log N) + \log T\right), & \gamma = 1 \\ \mathcal{O}(N^{-1/\gamma}(\log T)^{1/\gamma}(T+N) + N + \log T), & \gamma > 1 \end{cases}$$
$$= \begin{cases} \tilde{\mathcal{O}}\left(TN^{-1/\gamma} + N\min(\sqrt{T}, N^{1/\gamma})^{1-\gamma}\right), & \gamma < 1 \\ \tilde{\mathcal{O}}\left(TN^{-1/\gamma} + N\right), & \gamma \geq 1 \end{cases}$$

□

A.1.2 Mean Rewards from a Beta Prior

Under assumption (A6), we have $\gamma = \alpha + \beta - 1$ for $\beta > 1$ and $\gamma = \alpha$ otherwise. Corollary A.1.2 gives an explicit bound under (A5) and (A6).

Corollary A.1.2 (Adaptive Greedy - Bayesian Regret with Beta priors)**.** *Under the setup of Theorem 2.3.4 with an additional assumption (A6), the Bayesian regret of the adaptive greedy algorithm is given by*

$$BR_T \leq T\exp\{-Nc_0\epsilon^\gamma\} + 3\epsilon T + N\min\left\{L\frac{\beta+1}{\alpha+\beta+1}(1-F_{\alpha,\beta+1}(3\epsilon)),\right.$$
$$\left.\left(1+\frac{3}{C_1\epsilon}\right)(1-F_{\alpha,\beta}(3\epsilon)) + \min\{T/N_m, L\}(1-\epsilon)\frac{\beta+1}{\alpha+\beta+1}(1-F_{\alpha,\beta+1}(3\epsilon))\right\}$$

where $\gamma = \alpha+\beta-1$ for $\beta > 1$ and $\gamma = \alpha$ otherwise, and $F_{a,b}$ is the distribution of Beta(a,b) and $L := \mathbb{E}[L_i]$.

Proof of Corollary A.1.2. Following the same idea as the proof of Theorem A.1.1, the first term is,

$$\mathbb{E}_\Gamma\left[\mathbb{I}(1-\mu \leq \epsilon)\mathbb{P}(\exists t : \widehat{\mu}^t < 1-2\epsilon)\right] \leq \exp\{-c_0\epsilon^\gamma\}$$

where $\gamma = \alpha+\beta-1$. When analyzing the second expectation, we will bring the min operator outside the expectation and treat each inner expectation separately. The first expectation is,

$$\begin{aligned}
&\mathbb{E}_\Gamma\left[\mathbb{I}(1-\mu > 3\epsilon)\left(1+\min\{T/N_m, L\}(1-\mu)\frac{1-\epsilon}{N_m} + \frac{3}{C_1(1-\mu-2\epsilon)}\right)\right] \\
&\leq \mathbb{E}_\Gamma\left[\mathbb{I}(1-\mu > 3\epsilon)\left(1+\min\{T/N_m, L\}(1-\mu)\frac{1-\epsilon}{N_m} + \frac{3}{C_1\epsilon}\right)\right] \\
&= \int_{3\epsilon}^1 \left(1+\min\{T/N_m, L\}(1-\mu)\frac{1-\epsilon}{N_m} + \frac{3}{C_1\epsilon}\right)\frac{\mu^{\alpha-1}(1-\mu)^{\beta-1}}{B(\alpha,\beta)}d\mu \\
&= \left(1+\frac{3}{C_1\epsilon}\right)(1-F_{\alpha,\beta}(3\epsilon)) + \min\{T/N_m, L\}(1-\epsilon)\frac{\beta+1}{\alpha+\beta+1}(1-F_{\alpha,\beta+1}(3\epsilon))
\end{aligned}$$

And the second is,

$$\mathbb{E}_\Gamma\left[\mathbb{I}(1-\mu > 3\epsilon)\min\{L,T\}(1-\mu)\right] = \min\{L,T\}\frac{\beta+1}{\alpha+\beta+1}(1-F_{\alpha,\beta+1}(3\epsilon))$$

Assuming that $\min\{T, L\} = L$ gives us the desired result. □

A.1.3 Sampling by Lifetime

Corollary A.1.3 describes the upper bounds of the Bayesian regret of Adaptive Greedy when sampling by lifetimes. Essentially this allows us to conclude that previously established asymptotics of the upper bounds are not affected thereby proving Corollary 2.3.8.

Corollary A.1.3 (Adaptive Greedy sampling by lifetime)**.** *Under the assumptions (A1)-(A4), for any $\epsilon \in (0, 1/3)$, the Bayesian regret of the adaptive greedy algorithm where we sample by lifetime is given by*

$$BR_{T,N}(AG) \leq T\left(\mathbb{E}_\Gamma\left[\mathbb{I}(1-\mu \leq \epsilon)\mathbb{P}(\exists t : \widehat{\mu}^t < 1-2\epsilon)\right]\right)^N + 3\epsilon T +$$
$$N\mathbb{E}_\Gamma\mathbb{I}(1-\mu > 3\epsilon)\min\left\{1 + \min\{\frac{T}{\sum_{i=1}^{N_m} L_{(i)}}, 1\}L(1-\mu)(1-\epsilon) + \frac{3}{C_1(1-\mu-2\epsilon)}, \min\{L,T\}(1-\mu)\right\}$$

where $\{L_{(i)}\}_{i=1}^N$ are the order statistics of $\{L_i\}_{i=1}^N$.

If assumption (A5) also holds, then the asymptotics in Theorem 2.3.4 still hold. Further, if we assume (A5) and (A6) hold, then the result of Corollary A.1.2 naturally extends to

$$BR_T \leq T\exp\{-Nc_0\epsilon^\gamma\} + 3\epsilon T + N\min\left\{L\frac{\beta+1}{\alpha+\beta+1}(1-F_{\alpha,\beta+1}(3\epsilon)),\right.$$
$$\left.\left(1+\frac{3}{C_1\epsilon}\right)(1-F_{\alpha,\beta}(3\epsilon)) + \mathbb{E}\left[\min\{\frac{T}{\sum_{i=1}^{N_m} L_{(i)}}, 1\}L\right](1-\epsilon)\frac{\beta+1}{\alpha+\beta+1}(1-F_{\alpha,\beta+1}(3\epsilon))\right\}$$

where $\gamma = \alpha+\beta-1$ and $F_{a,b}$ is the distribution of $Beta(a,b)$ and the expectation is over $\{L_i\}_{i=1}^N$.

Proof of Corollary A.1.3. The key difference from the previous proofs lies in the exploration phase. Instead of sampling uniformly, we are sampling proportional to lifetimes.

Therefore,

$$\mathbb{P}(\text{arm } i \text{ is chosen} \mid \mathcal{G}) \leq \frac{L_i}{\sum_{j=1}^{N_t} L_j}(1 - \max_j \widehat{\mu}_j^t) + \mathbb{P}(\widehat{\mu}_i^t \geq 1 - 2\epsilon)$$

Since there are at least N_m arms to play and $\sum_{j=1}^{N_m} L_j \geq \sum_{j=1}^{N_m} L_{(j)}$ where $L_{(1)}, \ldots, L_{(N)}$ are the order statistics of $L_1, \ldots, L_N$, we have

$$\mathbb{P}(\text{arm } i \text{ is chosen} \mid \mathcal{G}) \leq \frac{L_i}{\sum_{j=1}^{N_m} L_{(j)}}(1 - \max_j \widehat{\mu}_j^t) + \mathbb{P}(\widehat{\mu}_i^t \geq 1 - 2\epsilon)$$

Pushing this through the remainder of the steps gives the desired result and the extensions of Theorems 2.3.4 and A.1.2. □

A.2 Technical Details of Pilot Sampling

A.2.1 Proof of Theorem 2.3.6

Proof of Theorem 2.3.6. Let $\mu^\star = \max_{i \in N} \mu_i$ and $\Delta_i = \mu^\star - \mu_i$. In the pilot strategy, we get to pull an arm $\min\{K, L_i\}$ times. For simplicity of analysis, we will assume that $\min\{K, L_i\} = K$. If there is at least one positive in the K pulls, then we pull until the arm dies (or we run out of budget). Again, we will assume that $\mu^\star = 1$. This will loosen the bound but also make it easier to handle integration over priors.

Let $\mathbb{E}[R_{N,T}]$ be the mean regret when we have N arms and a budget of T. Immediately, we have $\mathbb{E}R_{N,T} \leq T$. Therefore, for $T_1 \geq T_2$, $\mathbb{E}R_{N,T_1} - \mathbb{E}R_{N,T_2} \leq T_1 - T_2$. And if we have $\{L_i\}_{i=1}^N$ arms initially and we remove arm L_j to get $N_j = N - 1$ arms, then for the same budget T, then $\mathbb{E}R_{N,T} - \mathbb{E}R_{N_j,T} \leq L_j$.

Suppose we pick arm i first. Then the conditional mean regret is,

$$\begin{aligned}
\mathbb{E}\left[R_{N,T} \mid \text{choose } i\right] &\leq \left[L_i(1-\mu_i) + \mathbb{E}R_{N_i,T-L_i}\right](1 - (1-\mu_i)^K) + \\
&\quad \left[K(1-\mu_i) + \mathbb{E}R_{N_i,T-K}\right](1-\mu_i)^K \\
&\leq L_i(1-\mu_i) - (L_i - K)(1-\mu_i)^{K+1} + \mathbb{E}R_{N_i,T-L_i} +
\end{aligned}$$

$$
\begin{aligned}
&(\mathbb{E}R_{N_i,T-K} - \mathbb{E}R_{N_i,T-L_i})(1-\mu_i)^K \\
&\leq L_i(1-\mu_i) + (L_i - K)\left[(1-\mu_i)^K - (1-\mu_i)^{K+1}\right] + \mathbb{E}R_{N_i,T-L_i}
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\mathbb{E}R_{N,T} &= \sum_{i=1}^{N} \mathbb{P}(\text{choose } i)\mathbb{E}\left[R_{N,T} \mid \text{choose } i\right] \\
&\leq \sum_{i=1}^{N} \mathbb{P}(i)\left(L_i(1-\mu_i) + (L_i - K)\left[(1-\mu_i)^K - (1-\mu_i)^{K+1}\right]\right) + \sum_{i=1}^{N} \mathbb{P}(i)\mathbb{E}R_{N_i,T-L_i}
\end{aligned}
$$

If we sample arms uniformly, then $\mathbb{P}(i) = 1/N$. Therefore,

$$
\begin{aligned}
&\mathbb{E}R_{N,T} \leq \frac{1}{N}\sum_{i=1}^{N}\left(L_i(1-\mu_i) + (L_i - K)\left[(1-\mu_i)^K - (1-\mu_i)^{K+1}\right]\right) + \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}R_{N_i,T-L_i} \\
\implies &\mathbb{E}_\Gamma\mathbb{E}R_{N,T} \leq \mathbb{E}_\Gamma\left(L_i(1-\mu_i) + (L_i - K)\left[(1-\mu_i)^K - (1-\mu_i)^{K+1}\right]\right) + \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_\Gamma\mathbb{E}R_{N_i,T-L_i}
\end{aligned}
$$

Notice that the first term is independent of T and N (technically, we should apply a min operator to account for $T < L_i$, but we will come back to that later). And the second term can be recursively expanded at most $\mathbb{E}\lfloor T/L_{(1)}\rfloor$ times on average where $L_{(1)} = \min_{i=1,\ldots,N} L_i$ is the smallest lifetime of N samples. Thus, we have

$$
\begin{aligned}
BR_T &= \mathbb{E}_\Gamma\mathbb{E}R_{N,T} \\
&\leq \mathbb{E}\left\lfloor\frac{T}{L_{(1)}}\right\rfloor \mathbb{E}_\Gamma\left(L(1-\mu) + (L-K)\left[(1-\mu)^K - (1-\mu)^{K+1}\right]\right) + \mathbb{E}_\Gamma L(1-\mu)
\end{aligned}
$$

where the last term captures any remaining available pulls after exhausting $\lfloor T/L\rfloor$ arms. Notice that this naturally captures the case where $T < L$, which we ignored above. And to finally account for $N < T/L$,

$$
BR_T \leq \min\left\{N-1, \mathbb{E}\lfloor T/L_{(1)}\rfloor\right\}\mathbb{E}_\Gamma\left(L(1-\mu) + (L-K)\left[(1-\mu)^K - (1-\mu)^{K+1}\right]\right) + \mathbb{E}_\Gamma L(1-\mu)
$$

This gives us that $BR_{T,N}(\text{Pilot}) = \mathcal{O}(\min\{N, T\})$. □

A.2.2 Mean Rewards from a Beta Prior

Theorem 2.3.6 immediately allows us to apply Beta priors described in Assumption (A6). This is presented in Corollary A.2.1.

Corollary A.2.1 (Pilot Sampling - Bayesian Regret under Beta Priors)**.** *Consider the setup of Theorem 2.3.6. Additionally assume that (A4) and (A6) hold. Then, the Bayesian regret of the pilot sampling algorithm is given by*

$$BR_{T,N}(Pilot) \leq \min\left\{N-1, \mathbb{E}\lfloor T/L_{(1)}\rfloor\right\}\left[L\frac{\beta}{\alpha+\beta} + (L-K)\frac{\alpha}{\beta+\alpha+K+1}\prod_{r=0}^{K}\frac{\beta+r}{\beta+\alpha+r}\right] + L\frac{\beta}{\alpha+\beta}$$

where $L := \mathbb{E}[L_i]$.

Proof of Corollary A.2.1. Suppose that $\mu_i \sim \text{Beta}(\alpha, \beta)$. Therefore, $1-\mu_i \sim \text{Beta}(\beta, \alpha)$. Then the m-th moment of $1-\mu_i$ is given by

$$\mathbb{E}[(1-\mu_i)^m] = \prod_{r=0}^{m-1}\frac{\beta+r}{\beta+\alpha+r}$$

Therefore,

$$BR_T \leq \min\left\{N-1, \mathbb{E}\lfloor T/L_{(1)}\rfloor\right\}\left[L\frac{\beta}{\alpha+\beta} + (L-K)\frac{\alpha}{\beta+\alpha+K+1}\prod_{r=0}^{K}\frac{\beta+r}{\beta+\alpha+r}\right] + L\frac{\beta}{\alpha+\beta}$$

where $L := \mathbb{E}[L_i]$. □

A.2.3 Sampling by Lifetime

Corollary A.2.2 provides an upper bound on the Bayesian regret of pilot sampling when we sample by lifetimes. At first glance, the asymptotic order does not match with Theorem

2.3.6. Here, we have $BR = \mathcal{O}(\min\{N, T\}N/N_m)$ with an inflation factor of N/N_m. To gain tractability over the problem, we sacrificed lower-bounded $\sum_{i=1}^{N} L_i$ by $\sum_{i=1}^{N_m} L_{(i)}$. This is the reason we see the sum of the order statistics in the regret resulting in an apparently larger bound (as $N \geq N_m$). Since N_m is the number of arms available to play at any time $t < T$, it makes sense that $N_m = o(N - T/L)$ (we assume $T < NL$ as otherwise any policy is good). Therefore, $N/N_m = \mathcal{O}(N/(N - T/L)) = \mathcal{O}(1)$, which implies that $BR = \mathcal{O}(\min\{N, T\})$. Thus, the asymptotic behavior matches that of the variant that uniformly samples arms. This is exactly what we state in Corollary 2.3.9.

Corollary A.2.2 (Pilot sampling by lifetime)**.** *Suppose that assumptions (A1), (A2), and (A7) hold. Then the Bayesian regret of the pilot sampling algorithm when choosing arms by lifetime is given by*

$$BR_{T,N}(Pilot) < \min\{N-1, \mathbb{E}\lfloor T/L_{(1)}\rfloor\}\mathbb{E}_\Gamma\left[\left(\frac{NL^2}{\sum_{i=1}^{N_m} L_{(i)}} - K\right)\left((1-\mu)^K - (1-\mu)^{K+1}\right) + \frac{NL^2}{\sum_{i=1}^{N_m} L_{(i)}}(1-\mu)\right] + \mathbb{E}_\Gamma L_{(N)}(1-\mu),$$

where $\{L_{(i)}\}_{i=1}^{N}$ *are the order statistics of* $\{L_i\}_{i=1}^{N}$.

Additionally, if we assume (A4) and (A6), then the Bayesian regret of the pilot sampling algorithm is given by

$$BR_{T,N}(Pilot) \leq \min\{N-1, \mathbb{E}\lfloor T/L_{(1)}\rfloor\}\mathbb{E}_\Gamma\left[\left(\frac{NL^2}{\sum_{i=1}^{N_m} L_{(i)}} - K\right)\frac{\alpha}{\beta+\alpha+K+1}\prod_{r=0}^{K}\frac{\beta+r}{\beta+\alpha+r} + \frac{NL^2}{\sum_{i=1}^{N_m} L_{(i)}}\frac{\beta}{\alpha+\beta}\right] + \mathbb{E}_\Gamma L_{(N)}\frac{\beta}{\alpha+\beta}$$

Proof of Corollary A.2.2. From the proof of Theorem 2.3.6, we have

$$\mathbb{E}R_{N,T} \leq \sum_{i=1}^{N}\mathbb{P}(i)\left(L_i(1-\mu_i) + (L_i - K)\left[(1-\mu_i)^K - (1-\mu_i)^{K+1}\right]\right) + \sum_{i=1}^{N}\mathbb{P}(i)\mathbb{E}R_{N_i,T-L_i}$$

$$\leq \sum_{i=1}^{N} \frac{L_i}{\sum_{j=1}^{N} L_j} \left(L_i(1-\mu_i) + (L_i - K)\left[(1-\mu_i)^K - (1-\mu_i)^{K+1}\right]\right) + \sum_{i=1}^{N} \frac{L_i}{\sum_{j=1}^{N} L_j} \mathbb{E}R_{N_i, T-L_i}$$

$$= -K\left((1-\mu_i)^K - (1-\mu_i)^{K+1}\right) + \sum_{i=1}^{N} \frac{L_i^2}{\sum_{j=1}^{N} L_j} \left((1-\mu_i) + (1-\mu_i)^K - (1-\mu_i)^{K+1}\right) + \sum_{i=1}^{N} \frac{L_i}{\sum_{j=1}^{N} L_j} \mathbb{E}R_{N_i, T-L_i}$$

Here, we will use the assumption (A2) which says that there are always at least N_m arms left to play. Therefore, $\sum_{i=1}^{N} L_i \geq \sum_{i=1}^{N_m} L_{(i)}$ where $\{L_{(i)}\}_{i=1}^{N}$ are the order statistics of $\{L_i\}_{i=1}^{N}$. Let us define $L^{N_m} = \sum_{i=1}^{N_m} L_{(i)}$ for brevity. Therefore,

$$\mathbb{E}R_{N,T} \leq -K\left((1-\mu_i)^K - (1-\mu_i)^{K+1}\right) + \sum_{i=1}^{N} \frac{L_i^2}{L^{N_m}} \left((1-\mu_i) + (1-\mu_i)^K - (1-\mu_i)^{K+1}\right) + \sum_{i=1}^{N} \frac{L_i}{\sum_{i=1}^{N} L_i} \mathbb{E}R_{N_i, T-L_i}$$

Recursively expanding $\mathbb{E}R_{N_i, T-L_i}$ once,

$$\mathbb{E}R_{N,T} \leq -2K\left((1-\mu_i)^K - (1-\mu_i)^{K+1}\right) + \sum_{i=1}^{N} \frac{L_i^2}{L^{N_m}} \left((1-\mu_i) + (1-\mu_i)^K - (1-\mu_i)^{K+1}\right) + \sum_{i=1}^{N} \frac{L_i}{\sum_{i=1}^{N} L_i} \sum_{j \neq i}^{N} \frac{L_j^2}{L^{N_m}} \left((1-\mu_j) + (1-\mu_j)^K - (1-\mu_j)^{K+1}\right) + \frac{L_j}{\sum_{k \neq i}^{N} L_k} \mathbb{E}R_{N_{i,j}, T-L_i-L_j}$$

Now, look at the second and third terms. Let us call $\tilde{\mu}_j = (1-\mu_j) + (1-\mu_j)^K - (1-\mu_j)^{K+1}$ for brevity. There are N copies each of $L_i^2 \tilde{\mu}_i / L^{N^m}$ for all $i = 1, \ldots, N$ where the sum of coefficients on copies is $2 - L_i / \sum_{j=1}^{N} L_j$. Therefore,

$$\sum_{i=1}^{N} \frac{L_i^2}{L^{N_m}} \tilde{\mu}_i + \sum_{i=1}^{N} \frac{L_i}{\sum_{i=1}^{N} L_i} \sum_{j \neq i}^{N} \frac{L_j^2}{L^{N_m}} \tilde{\mu}_j = \sum_{i=1}^{N} \sum_{j \neq i}^{N} \frac{L_i^2}{L^{N_m}} \tilde{\mu}_i + \frac{L_i}{\sum_{i=1}^{N} L_i} \frac{L_j^2}{L^{N_m}} \tilde{\mu}_j$$

$$= \sum_{i=1}^{N} \left(2 - \frac{L_i}{\sum_{j=1}^{N} L_j}\right) \frac{L_i^2}{L^{N_m}} \tilde{\mu}_i$$

Thus,

$$\mathbb{E}R_{N,T} \leq -2K\left((1-\mu_i)^K - (1-\mu_i)^{K+1}\right) + \sum_{i=1}^{N}\left(2 - \frac{L_i}{\sum_{i=1}^{N} L_i}\right)\frac{L_i^2}{L^{N_m}}\tilde{\mu}_i + \sum_{i=1}^{N}\sum_{j\neq i}^{N}\frac{L_i}{\sum_{i=1}^{N} L_i}\frac{L_j}{\sum_{k\neq i}^{N} L_k}\mathbb{E}R_{N_{i,j},T-L_i-L_j}$$

At this point, it is easy to see where this recursion is going. If we expand $\mathbb{E}R_{N,T}$ m times, then the coefficient on the first term will m; the coefficient of the terms within the second sum will be strictly smaller than m, and; the final term will have m summations. As in our strategy for Theorem 2.3.6, we can recursively expand out $\mathbb{E}R_{N_i,T-L_i}$ for at most $\lfloor T/L_{(1)}\rfloor$ times. The behavior of the last term is hard to analyze. There are $\lfloor T/L_{(1)}\rfloor$ summations. The kth sum is a weighted average over $N-k+1$ items. To make things concrete, note that the last term corresponds to any leftover budget that is not enough to fully play an arm. Therefore, we can bound the last term over the arm with the longest lifetime. So, applying expectation over $(\mu, L) \sim \Gamma$ before expanding out the recursion, we have

$$\begin{aligned}\mathbb{E}_\Gamma\mathbb{E}R_{N,T} &< \mathbb{E}_\Gamma\left\lfloor\frac{T}{L_{(1)}}\right\rfloor\mathbb{E}_\Gamma\left[\sum_{i=1}^{N}\frac{L_i^2}{L^{N_m}}\tilde{\mu}_i - K\left((1-\mu_i)^K - (1-\mu_i)^{K+1}\right)\right] + \mathbb{E}_\Gamma L_i(1-\mu_i)\\ &= \mathbb{E}_\Gamma\left\lfloor\frac{T}{L_{(1)}}\right\rfloor E_\Gamma\left[N\frac{L^2}{L^{N_m}}\tilde{\mu} - K\left((1-\mu)^K - (1-\mu)^{K+1}\right)\right] + \\ &\quad \mathbb{E}_\Gamma L_{(N)}(1-\mu)\end{aligned}$$

If we assume that $\mu \perp L$ as in (A4) and that $\mu \sim \text{Beta}(\alpha, \beta)$ as in (A6), then

$$BR_{N,T} \leq \mathbb{E}_\Gamma\left\lfloor\frac{T}{L_{(1)}}\right\rfloor E_\Gamma\left[\frac{NL^2}{L^{N_m}}\frac{\beta}{\alpha+\beta} + \left(\frac{NL^2}{L^{N_m}} - K\right)\frac{\alpha}{\beta+\alpha+K+1}\prod_{r=0}^{K}\frac{\beta+r}{\beta+\alpha+r}\right] + \mathbb{E}L_{(N)}\frac{\beta}{\alpha+\beta}$$

Finally, we need to account for the case $N < T/L_{(1)}$. We can do this by replacing $\mathbb{E}(T/L)$ with $\min\{N-1, \mathbb{E}(T/L)\}$. This completes the result. □

A.2.4 Choosing a Pilot Group Size

Let $\mu \sim \text{Beta}(\alpha, \beta)$ to denote the mean reward. Chakrabarti et al. (2008) impose an assumption that the lifetime of arms is exponentially distributed with mean L. They define the following reward function,

$$\mathcal{R}(x) = \frac{\mathbb{E}[\mu] + (1 - F(x))(L-1)\mathbb{E}[\mu \mid \mu \geq x]}{1 + (1 - F(x))(L-1)} \tag{A.2}$$

where F is the distribution of μ.

Equation A.2 captures the trade-off between exploration and exploitation. The first term corresponds to the action where we pull an arm once. Suppose we are given a threshold x. The second term corresponds to the action that we pull arm i until its death given that its mean reward is larger than x. Together, Equation A.2 describes the average reward per pull when following the strategy of pulling an arm $n \leq L_i$ times and pulling it until its death if the reward from the first n pulls is at least nx.

Therefore, choosing $x^* = \text{argmax}_x \mathcal{R}(x)$ gives us the optimal threshold. When choosing the pilot size K, Chakrabarti et al. (2008) show that if $K = \mathcal{O}(\log L/\epsilon^2)$ the average reward per step is $\mathcal{R}(x^* - \epsilon)$. We choose K (rounded to the nearest integer greater than K) such that $Kx^* = 1$. In other words, K is the optimal number of pulls before we decide to abandon or exploit the arm.

Under Assumption (A6), Lemma A.2.3 gives the exact form of the reward. Equation A.3 is concave for $x \in [0, 1]$ and thus has a unique maximizer, x^*, that can be calculated using standard optimization algorithms like gradient descent.

Lemma A.2.3. *For $\mu \sim$ Beta(α, β), Equation A.2 simplifies as*

$$\mathcal{R}(x) = \frac{\alpha}{\alpha + \beta} \frac{1 + (1 - F_{\alpha+1,\beta}(x))(L-1)}{1 + (1 - F_{\alpha,\beta}(x)(L-1)} \tag{A.3}$$

where $F_{a,b}$ is the distribution of Beta(a, b).

Proof of Lemma A.2.3. We have $\mu \sim \text{Beta}(\alpha, \beta)$. Let f denote the density of μ. Then,

Equation A.2 simplifies as

$$
\begin{aligned}
f(\mu \mid \mu \geq x) &= \frac{f(\mu, \mu \geq x)}{1 - F(x)} \\
\implies \mathbb{E}[\mu \mid \mu \geq x] &= \frac{1}{1 - F(x)} \int_x^1 \frac{1}{B(\alpha, \beta)} t^{\alpha} (1-t)^{\beta - 1} dt \\
&= \frac{1}{1 - F(x)} \frac{\alpha}{\alpha + \beta} \int_x^1 \frac{1}{B(\alpha + 1, \beta)} t^{\alpha} (1-t)^{\beta - 1} dt \\
&= \frac{\mathbb{E}[\mu]}{1 - F(x)} (1 - F_{\alpha+1,\beta}(x))
\end{aligned}
$$

where B is the beta function and $F_{a,b}$ is the distribution of a Beta(a, b) random variable. Thus,

$$
\begin{aligned}
\mathcal{R}(x) &= \mathbb{E}[\mu] \frac{1 + (1 - F_{\alpha+1,\beta}(x))(L-1)}{1 + (1 - F(x)(L-1)} \\
&= \frac{\alpha}{\alpha + \beta} \frac{1 + (1 - F_{\alpha+1,\beta}(x))(L-1)}{1 + (1 - F_{\alpha,\beta}(x)(L-1)}
\end{aligned}
$$

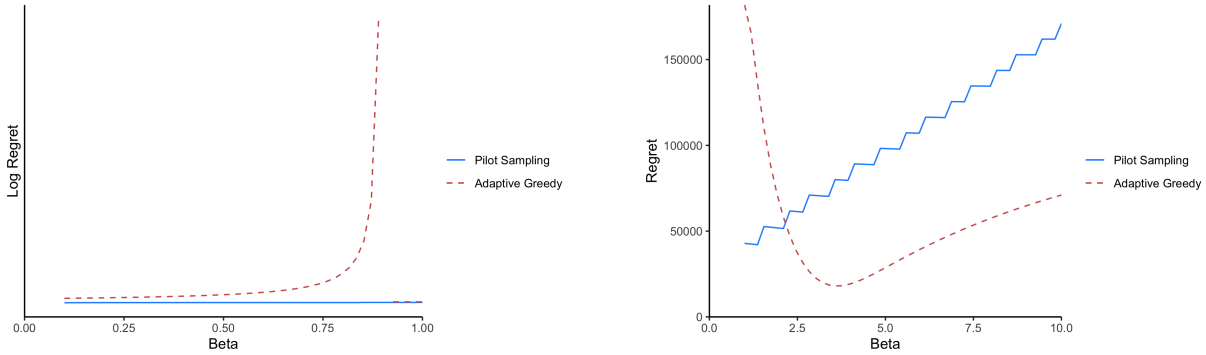
□

A.3 Asymptotic Behavior of Regret Bounds

In Figure A.1, we compare the regret bounds between Adaptive Greedy and Pilot Sampling algorithms for Beta priors.

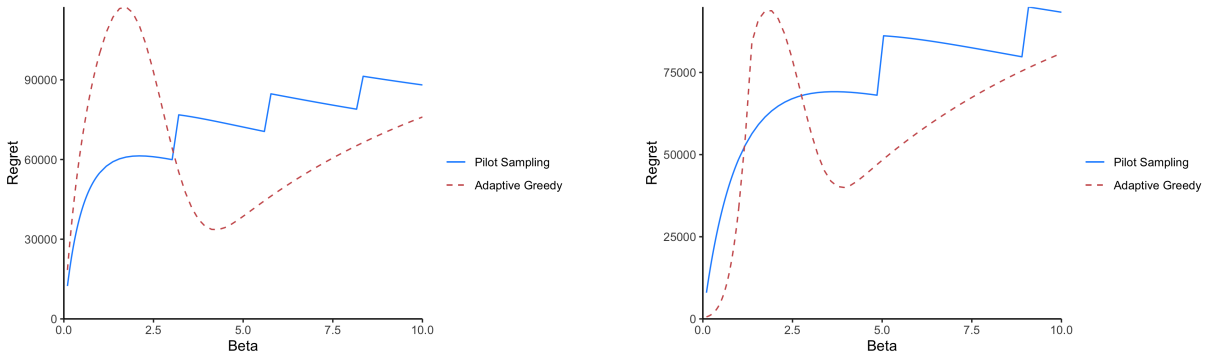
Here, we study the asymptotic behavior of the Bayesian regret bounds for a variety of Beta priors. In particular, we look at four different scenarios: (i) $\alpha < 1$, (ii) $\alpha = 1$, (iii) $\alpha > 1$, and (iv) $\beta < \alpha < 1$. These are shown in Figures A.2-A.5 respectively. In these plots, we used the bounds presented in Corollaries A.1.2 and A.2.1. In all of these plots, we fixed the ratio $T/N = 0.5$ and varied N to identify when the asymptotic behavior is achieved.

From these simulations, we can immediately learn three things. First, for a fixed α, as we increase β, the upper bounds of Pilot sampling and Adaptive Greedy get closer to each other. Second, for a fixed β, as we increase α, the upper bounds of Pilot sampling Adaptive



(a) $\alpha = 0.1, \beta < 1$. Regret is presented on a log scale.

(b) $\alpha = 0.1, \beta > 1$.



(c) $\alpha = 1$

(d) $\alpha = 2$

Figure A.1: Bayesian regret bounds for Adaptive Greedy and Pilot Sampling algorithms presented in Theorems A.1.2 and A.2.1 respectively. Here, we fixed $N = 10,000, N_m = 10, L = 20, T = 50000$. In each panel, we fixed $\alpha \in \{0.1, 1, 2\}$ and varied $\beta \in [0.1, 10]$. When $\alpha = 0.1$, we split the regret bounds into two plots to show the difference in the two algorithms at a meaningful scale. It is clear that when we have a heavily right-skewed distribution, Pilot Sampling has a much smaller regret bound compared to Adaptive Greedy (see panel (a)). This is exactly what we demonstrate in our simulations in Section 2.4. When there is a heavy left-skew, Adaptive Greedy has a smaller regret bound which is also in line with our simulations in Section 2.4. For other cases, when the regret bounds are comparable, we see little to no difference in performance in our simulations.

Greedy get closer to each other. Third, the asymptotic behavior occurs in the range of N between 10^3 (generally for smaller α) and 10^6 (generally for larger α) depending on the prior

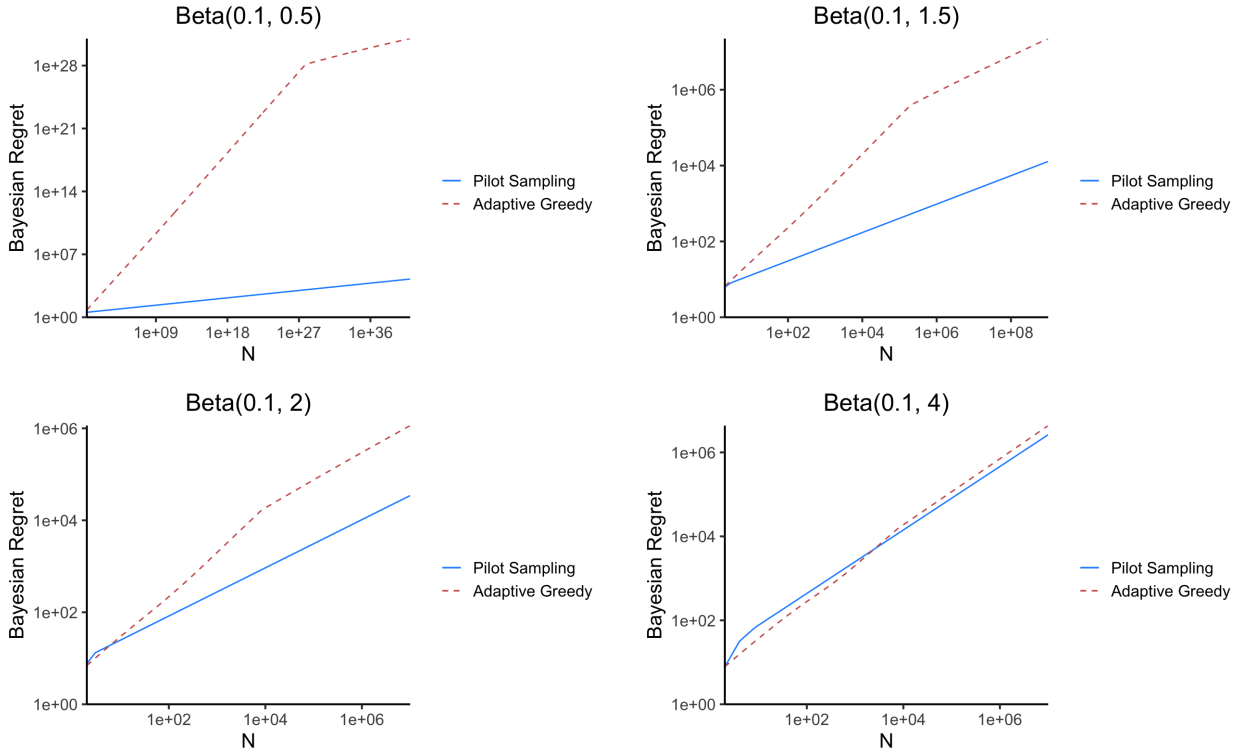


Figure A.2: In these figures, we fix $\alpha = 0.1$ and vary $\beta \in \{0.5, 1.5, 2, 4\}$. Except for the case in the top left with $\alpha < \beta < 1$, the asymptotic behavior is reached at a reasonable $N = 10^4$. Based on this, we can conclude that our datasets lie in this asymptotic regime. Here, we fixed $T/N = 0.5$.

distribution. The only exception is when $\alpha < \beta < 1$ where sometimes we need $N > 10^{27}$ for the asymptotic behavior to occur.

Now, we attempt to translate some of the patterns revealed by the asymptotic bounds to our empirical simulations. First, the size of the "asymptotic gap" between Pilot sampling and Adaptive Greedy tells us how they compare. Note that this gap is primarily composed of the logarithmic factors present in Theorem 2.3.4. Observe that, usually, the size of the gap corresponds to how much better Pilot sampling is. Particularly, consider $\alpha < \beta < 1$, and $\alpha = 1, \beta = 4$. Compared with simulations, we see that, in the former case, both Pilot sampling is better and the gap is larger. Of course, as the gap becomes smaller, it is more

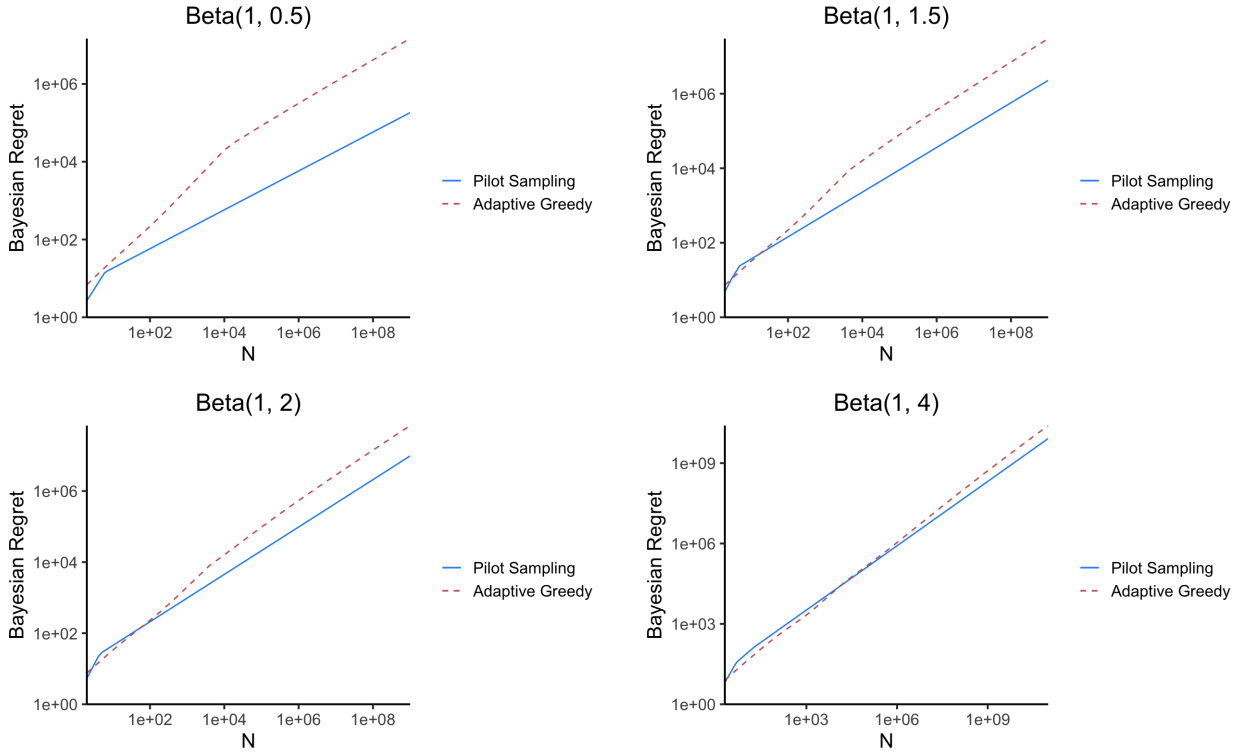


Figure A.3: In these figures, we fix $\alpha = 1$ and vary $\beta \in \{0.5, 1.5, 2, 4\}$. Again, we can conclude that the asymptotic behavior is achieved quickly. Here, we fixed $T/N = 0.5$.

likely that Adaptive Greedy performs better than Pilot sampling in practice.

The size of this gap also corresponds with the variance of the distribution of the mean rewards. The larger the variance, the larger the gap. When the variability (heterogeneity) is larger, Pilot sampling performs better. This is seen in the case with $\alpha < \beta < 1$. With a smaller variance ($\alpha = 1, \beta = 4$), the heterogeneity is not large enough for Pilot sampling to beat Adaptive Greedy.

A.4 Additional Comparison Algorithms

A naive approach We consider a naive approach that one might take when playing the mortal bandit. In this method, an agent may choose an arm uniformly at random and play it until it dies. Then, the agent repeats the procedure on all remaining arms. While this

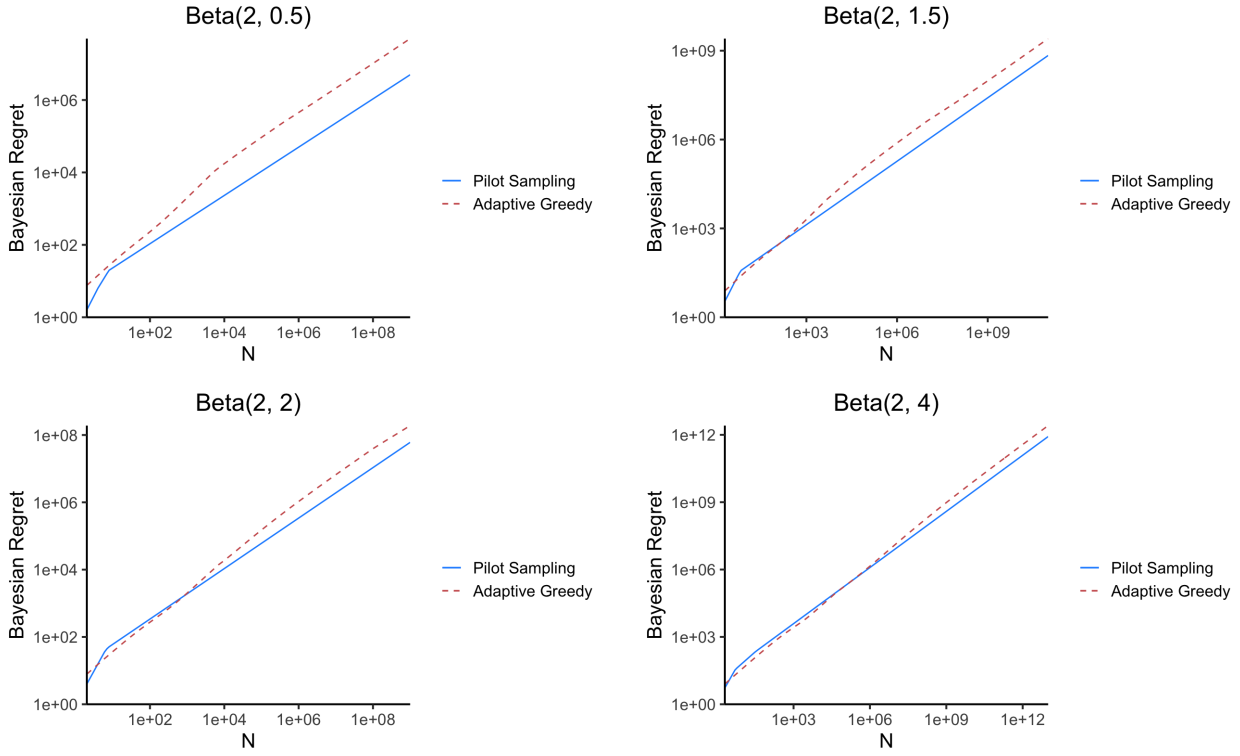


Figure A.4: In these figures, we fix $\alpha = 2$ and vary $\beta \in \{0.5, 1.5, 2, 4\}$. Again, the conclusions remain the same. Here, we fixed $T/N = 0.5$.

sampling method is very simple to understand and implement in practice, this method is sub-optimal as the agent does not do any kind of learning of the mean rewards and is agnostic to the lifetimes of arms. Hence, we dub this the "naive sampler." This method is described in Algorithm 7.

Algorithm 7 Naive sampling

Input: T budget, N arms

Set $t = 0$
while $t < T$ **do**
    randomly choose an available arm i
    Pull arm i until it dies (L_i times) or we exhaust budget
    $t = t + L_i$

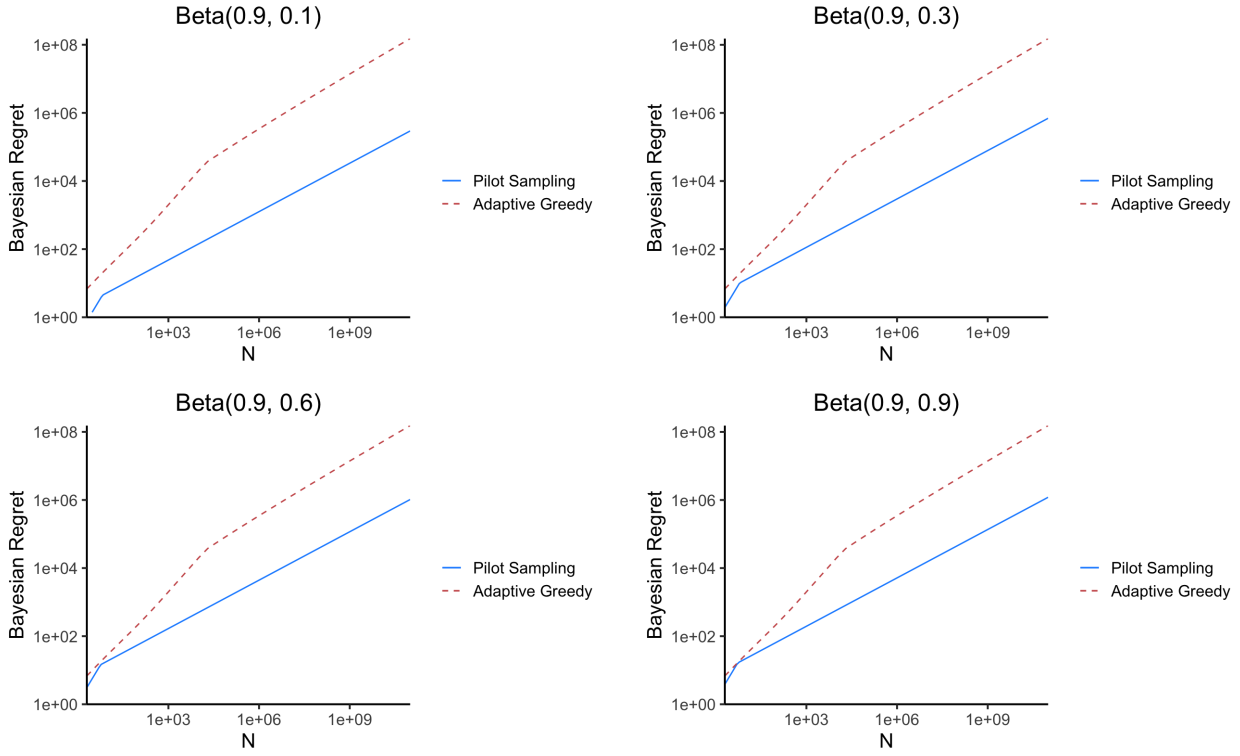


Figure A.5: In these figures, we fix $\alpha = 0.9$ and vary $\beta \in \{0.1, 0.3, 0.6, 0.9\}$. Again, the conclusions remain the same. Here, we fixed $T/N = 0.5$.

Thompson sampling Thompson sampling is a sequential decision-making algorithm that uses a Bayesian model (Thompson, 1933). While this method is very simple, its empirical performance makes it highly competitive and has been shown to theoretically achieve optimal performance by Chapelle and Li (2011) and Agrawal and Goyal (2012) in the standard bandit setting.

Consider the Beta-Bernoulli model, $\mu_i \sim \text{Beta}(\alpha_i, \beta_i)$ and $X_i \mid \mu_i \sim \text{Bernoulli}(\mu_i)$. In our problem, this models the infectiousness of a person i with μ representing their per-contact infectivity and X representing whether a random contact of person i gets infected.

The algorithm works the following way. First, we assign the mean reward of all arms the same Beta prior. This can, and often will, be different from the prior from which μ_i originally came from. For simplicity, we may choose the uniform prior i.e., $\text{Beta}(1, 1)$. Then,

we draw $\widetilde{\mu}_i$ for every arm. We pick the arm with the largest $\widetilde{\mu}_i$ and pull it once. Using the observed X_i, we update arm i's prior. Algorithm 8 formally describes the method.

Algorithm 8 Thompson sampling

Input: T budget, N arms, (α, β) initial priors

For every person $i \in \mathcal{I}$, assign them prior parameters (α, β)
$t = 0$
while $t < T$ **do**
    $t = t + 1$
    Sample $\widetilde{\mu}_i \sim \text{Beta}(\alpha_i, \beta_i)$ for arm i that is alive
    Find $k = \text{argmax}_i \, \widetilde{\mu}_i$
    Pull arm k to get reward X_k
    if $X_k = 1$ **then**
        Add infection as a new arm and assign priors (α, β)
        Update $\alpha_k \leftarrow \alpha_k + 1$
    else
        Update $\beta_k \leftarrow \beta_k + 1$
    Remove arm k if is dead

Thompson sampling leverages heterogeneity in mean rewards by assigning it a prior distribution. This allows us to model each arm's mean reward independently. By randomly sampling from and updating the Beta model during each iteration, Thompson sampling enables us to efficiently explore the pool of arms.

Despite being mathematically simple, Thompson sampling can be cumbersome to implement in practice. Theoretically, Thompson sampling invests a lot of resources in pulling randomly chosen arms, one at a time, before identifying the arm with the largest reward. In a setting such as contact tracing. it can be logistically difficult to test randomly during an outbreak of a disease.

A.5 Additional Simulations

A.5.1 Visualizing Beta Distributions

In Figure A.6, we visualize the probability densities of different Beta distributions we use in our simulations.

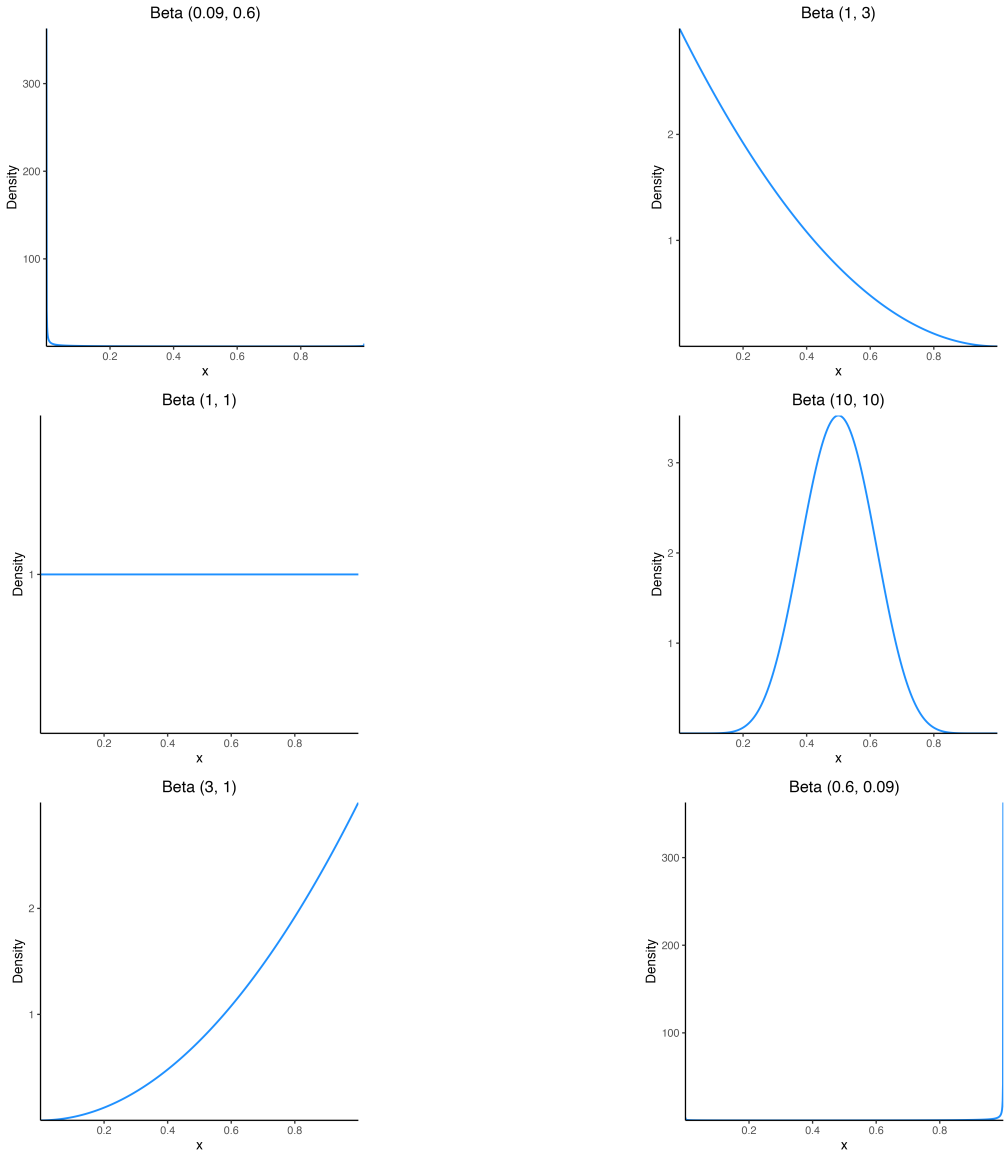


Figure A.6: Densities of the various Beta distributions used in the simulations.

A.5.2 Simulations with Branching

Recall that in Section 2.3, we worked under the assumption that new arms do not appear for the tractability of the problem. Here, we confront that assumption as well and show that the conclusions we draw are robust. In particular, we perform a second set of simulations where we introduce branching i.e., when we get a reward $X_i = 1$, this creates a new arm. This kind of branching is common in applications such as contact tracing where uncovering an infection presents a new set of contacts to test.

As in the simulations without branching, we will consider two regimes for the lifetime of arms – Poisson and heavy-tailed Pareto. In the first set of simulations, we set the initial number of arms $N = 10$, the average lifetime to $\lambda = 500$, and the budget to $T = 50,000$. The results are shown in Figure A.7. The trends largely agree with the simulations without branching in 2.1. A key distinction we note is that in the simulations without branching, there appeared to be a critical point when Pilot Sampling began outperforming Adaptive Greedy. And as the distribution became more and more right-skewed, that critical point kept moving further toward the end of the time horizon. In the simulations with branching, there does not seem to be such a critical point. This can be attributed to how Adaptive Greedy sampling works. At the beginning of the game, Adaptive Greedy sampling estimates the mean reward for *all* arms that are playable for the entire time horizon (see line 5 of Algorithm 1). This gives it a distinct advantage over the other methods, which do not perform this preparation, allowing it to shine in the non-branching case. This advantage is lost when new arms appear for which Adaptive Greedy does not have a mean reward estimate.

We perform similar simulations with $\lambda = 10$ and $T = 1000$ in Appendix A.5.4.

In the final set of simulations, the lifetime was drawn from a Pareto distribution with shape parameter 0.6 and location parameter 1. The results are shown in Figure A.8 and also agree with the simulations without branching.

A.5.3 Choosing a Pilot Group Size with a Misspecified Prior

Now, we turn to a key assumption regarding Pilot sampling. In order to choose an appropriate pilot group size, K, we need to know the distribution of the mean rewards. Of course, one could choose an arbitrary K and still successfully run Pilot sampling, but we are not guaranteed to achieve optimal performance. To demonstrate this, we will perform some simulations. Consider the case where Pilot sampling is the best in Figure 2.1. The mean rewards are drawn from Beta$(0.09, 0.61)$. We repeat the data generation with $S = 10$, $\lambda = 10$, and $N = 1000$. We compare the performance of Pilot sampling with $K = \{1, \ldots, 8\}$ in Figure A.9. Although the plot is cluttered, it is easy to see that the difference in performance is large. While they all seem to outperform Naive sampling (increasing $K \to \infty$ will make Pilot sampling identical to Naive sampling), the choice of optimal K is unclear. This is a big drawback of Pilot sampling when we do not know the distribution of mean rewards. In such cases, Thompson sampling or Adaptive Greedy is favorable as they do not need any parameter tuning. (Thompson sampling needs initial conditions for a prior, but in our simulations, we found that the choice of initial conditions does not impact performance, even when the correct priors are provided.)

A.5.4 Simulations with Smaller Average Degree

In these simulations, we fix $N = 10$, $\lambda = 10$, and $T = 1000$. We varied the parameters of the Beta distribution. For Thompson sampling, we initialized the prior distribution at Beta$(1, 1)$, a uniform prior. The results without branching are shown in Figure A.10. And the results with branching are shown in Figure A.11.

A.6 Real Data

Properties of these datasets are summarized in Table A.1. Estimated PCI parameters are described in Table A.2. We estimated the PCI using the Bayes shrinkage estimator and used the method of moments to estimate the parameters of the Beta model. The optimal pilot

group size was chosen as described in Appendix A.2.4.

Region	People traced	Tests administered	Positive infections
Punjab, Pakistan	165,072	1,911,669	36,868
Punjab, India	2,077	18,284	1,620
Southern India	88,616	649,990	27,196

Table A.1: Properties of data collected as a part of contact tracing efforts of COVID-19 in 2020

Figure A.12 is identical to Figure 2.4 but the axes reflect absolute numbers i.e., they are not normalized.

To compare with the Poisson and Pareto lifetime distributions we used in the simulations in Section 2.4, we found the maximum likelihood estimates of the parameters in Poisson and Pareto distributions. These are shown in Table A.3. To visualize how well they compare to the observed lifetimes, we use a P-P plot to compare the theoretical and empirical CDFs in Figure A.13. Just from these plots, it is unclear if either distribution is a good fit for the observed data. For instance, the Pareto distribution captures the third quantile in all cases and Poisson captures the median in the Punjab, Pakistan data but only the third quantile for the other two. The Poisson distribution tries to capture the median of the observation but is unable to account for the overdispersion. Pareto distribution does a better job of accounting for the overdispersion but loses out on the behavior around the median. It does appear to approximate the dataset from southern India well. Overall, these plots indicate that the distribution of the observed lifetimes does not have a heavy tail, which is indicated by the blue curve (corresponding to the Pareto distribution) lying below the grey curve as we approach 1 in all panels of Figure A.13.

A.7 Some Useful Results

Lemma A.7.1 (Lemma 4.2 of Bayati et al. (2020))**.** *Let* $\{X_i\}_{i=1}^{\infty}$ *be an i.i.d. sequence of Bern*(μ) *random variables. Let* $\sum_{i=1}^{n} X_i/n$ *as the sample mean of the first* n *random*

variables. For $\epsilon < 1/6$ *and* $\mu \geq 1 - \epsilon$,

$$\mathbb{P}(\exists n : M_n < 1 - 2\epsilon) \leq 1 - \frac{\exp(-0.3)}{2}$$

Lemma A.7.2 (Lemma 4.3 of Bayati et al. (2020))**.** *Suppose that* P_μ *is a distribution with mean* μ *and support* $[0, 1]$. *Let* $\{X_i\}_{i=1}^{\infty}$ *be a sequence of i.i.d. random variables with distribution* P_μ. *Define* $M_n = \sum_{i=1}^{n} X_i/n$ *as the sample mean of the first* n *random variables. If* P_μ *is 1-subgaussian, then for any* $\epsilon > 0$,

$$\mathbb{P}(\exists n : M_n < \mu - \epsilon) \leq \exp(-\epsilon^2/2)$$

Lemma A.7.3 (Lemma D.3 of Bayati et al. (2020))**.** *Suppose that the distribution of* $\mu \sim Q$ *is* γ*-regular i.e.,* $\mathbb{P}(\mu > 1 - \epsilon) = \Theta(\epsilon^\gamma)$. *Then,*

$$\mathbb{E}_Q\left[\mathbb{I}(1 - \mu > 3\epsilon) \min\left\{\left(1 + \frac{1}{1 - \mu - 2\epsilon}\right), T(1 - \mu)\right\}\right] \leq C_0 \begin{cases} 5 + \log(1/\epsilon), & \gamma = 1 \\ C(\gamma), & \gamma > 1 \\ C(\gamma) \min\left(\sqrt{T}, 1/\epsilon\right)^{1-\gamma}, & \gamma < 1 \end{cases}$$

where $C(\gamma)$ *is a constant that depends only on* γ.

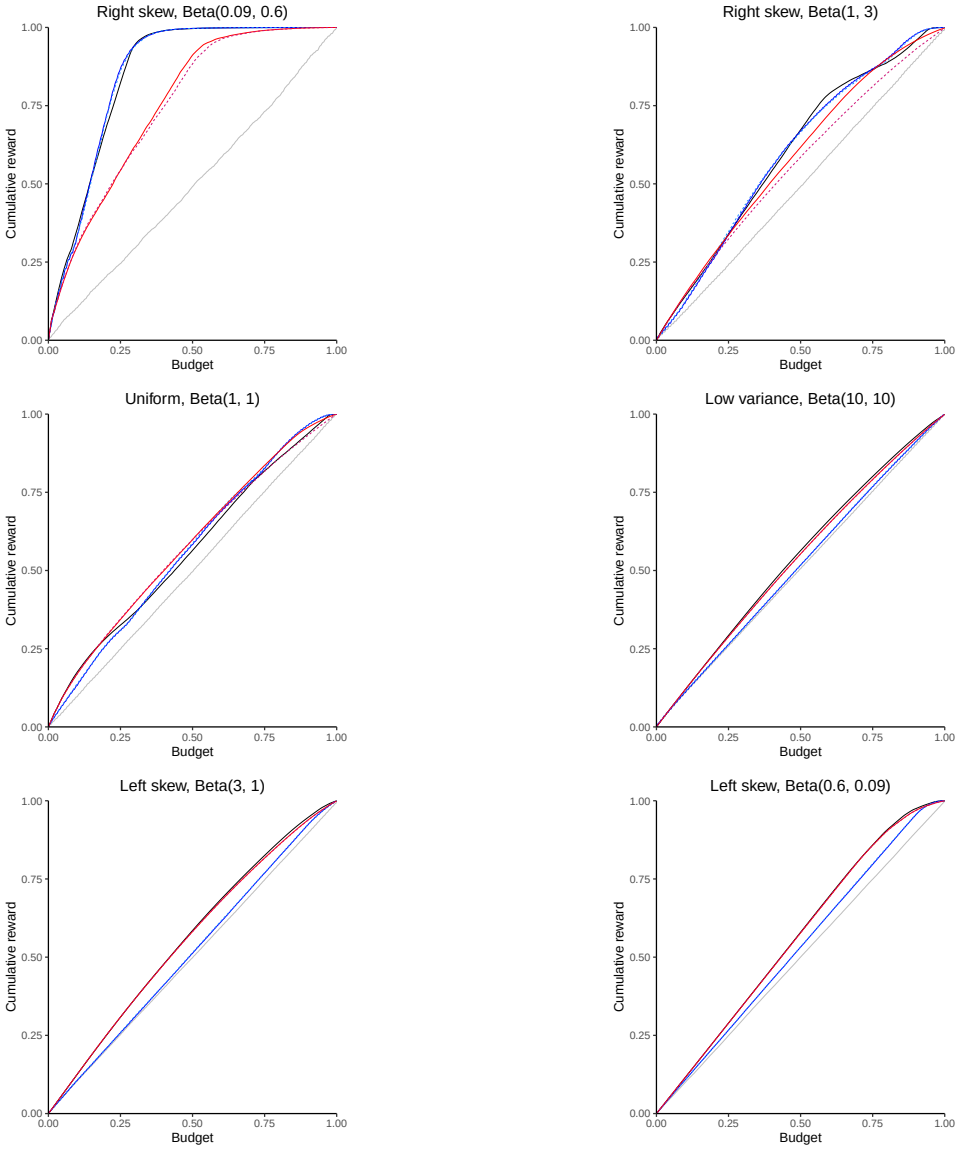


Figure A.7: Cumulative reward over the total time horizon for different policies and various reward distributions using data simulated with branching and Poisson(500) lifetime. The axes are normalized. Thompson sampling is in black; pilot sampling with uniform sampling and lifetime sampling are in dark blue and light blue respectively; adaptive greedy with uniform sampling and sampling by lifetime are in red and pink respectively, and; naive sampling is in grey. Pilot sampling performs better when the rewards are right-skewed while adaptive greedy performs better when rewards are left-skewed or rewards have low variance. Thompson sampling appears to perform consistently well in all scenarios. Sampling by degree is identical to sampling uniformly for pilot sampling while sampling uniformly is better for adaptive greedy. These conclusions are similar to Figure 2.1.

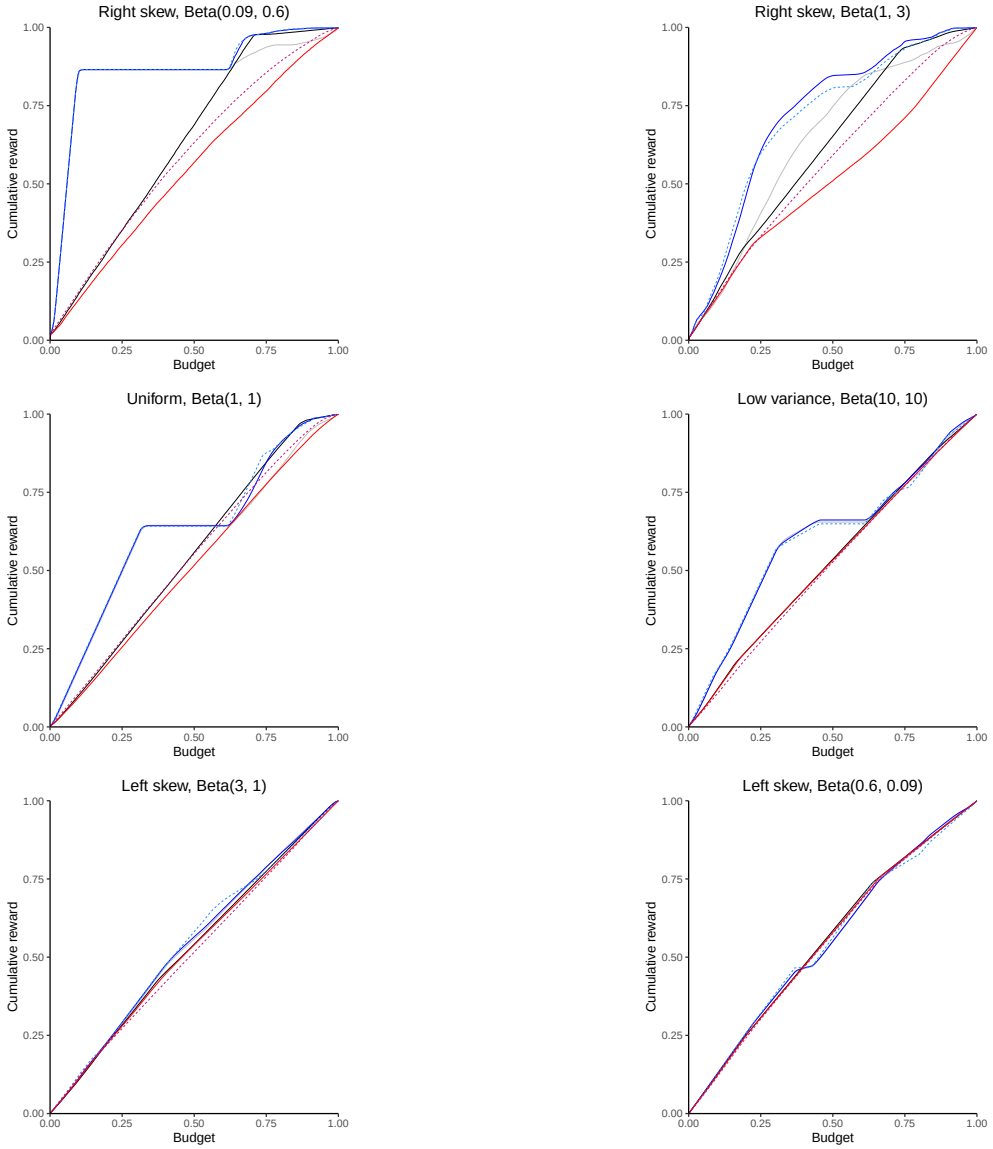


Figure A.8: Cumulative reward over the total time horizon for different policies and various reward distributions based on data simulated with branching and Pareto(1, 0.6) lifetime. The axes are normalized to facilitate visual comparison. Thompson sampling is in black; pilot sampling with uniform sampling and lifetime sampling are in dark blue and light blue respectively; adaptive greedy with uniform sampling and sampling by lifetime are in red and pink respectively, and; naive sampling is in grey. Pilot sampling performs in all scenarios except for heavily left-skewed rewards. In this setup, there is no clear winner between sampling by degree and sampling arms uniformly for pilot sampling while sampling by lifetime is better for adaptive greedy. These results are identical to simulations without branching in Figure 2.2.

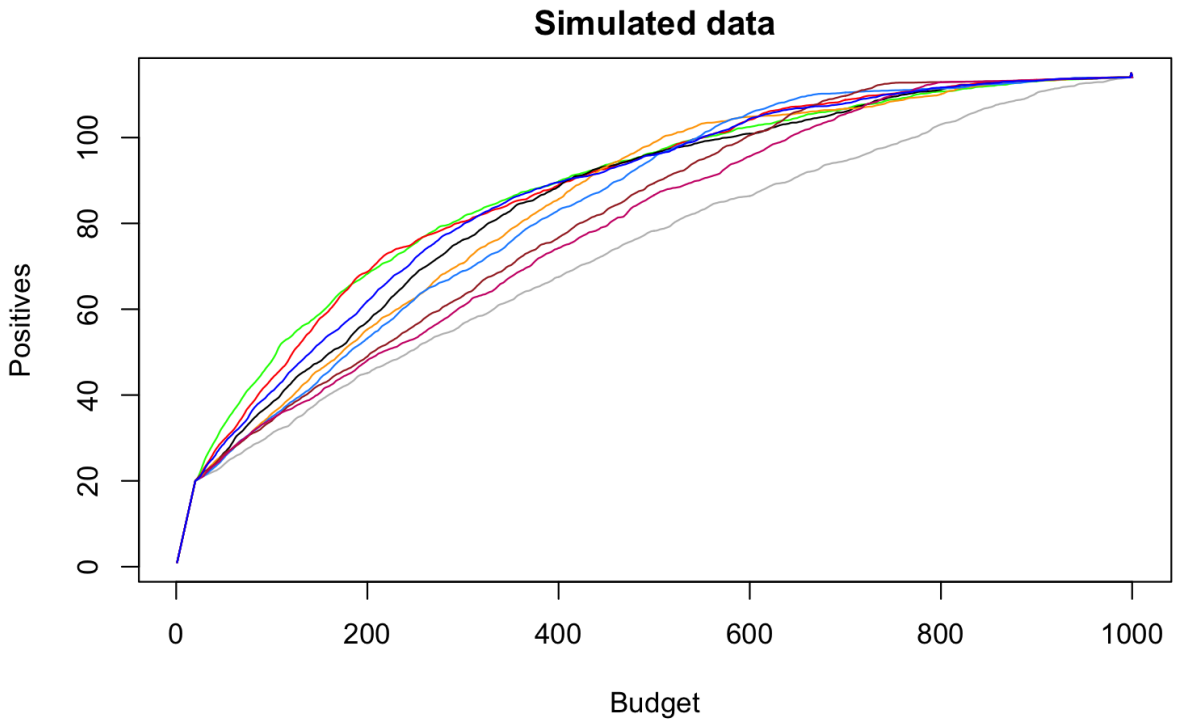


Figure A.9: Performance of Pilot Sampling with various group sizes, K. Here, $\mu \sim \text{Beta}(0.09, 0.61)$ and average degree was 10. $K = 1$ is green, $K = 2$ is red, $K = 3$ is blue, $K = 4$ is black, $K = 5$ is orange, $K = 6$ is light blue, $K = 7$ is brown, and $K = 8$ is pink. The grey line corresponds to Naive sampling. There is a large variability in our performance as we change the size of the pilot group. So when the priors are incorrectly specified, we might end up choosing a sub-optimal pilot group size.

Region	Beta parameters (α, β)	Optimal x^* ($\Gamma(x^*)$)	Pilot group size
Punjab, Pakistan	(0.0877, 2.0681)	0.1902 (0.1902)	6
Punjab, India	(0.1382, 0.8830)	0.5902 (0.5902)	2
Southern India	(0.1228 0.9709)	0.3415 (0.3415)	3

Table A.2: Estimated parameters for each dataset. All PCI distributions are right-skewed.

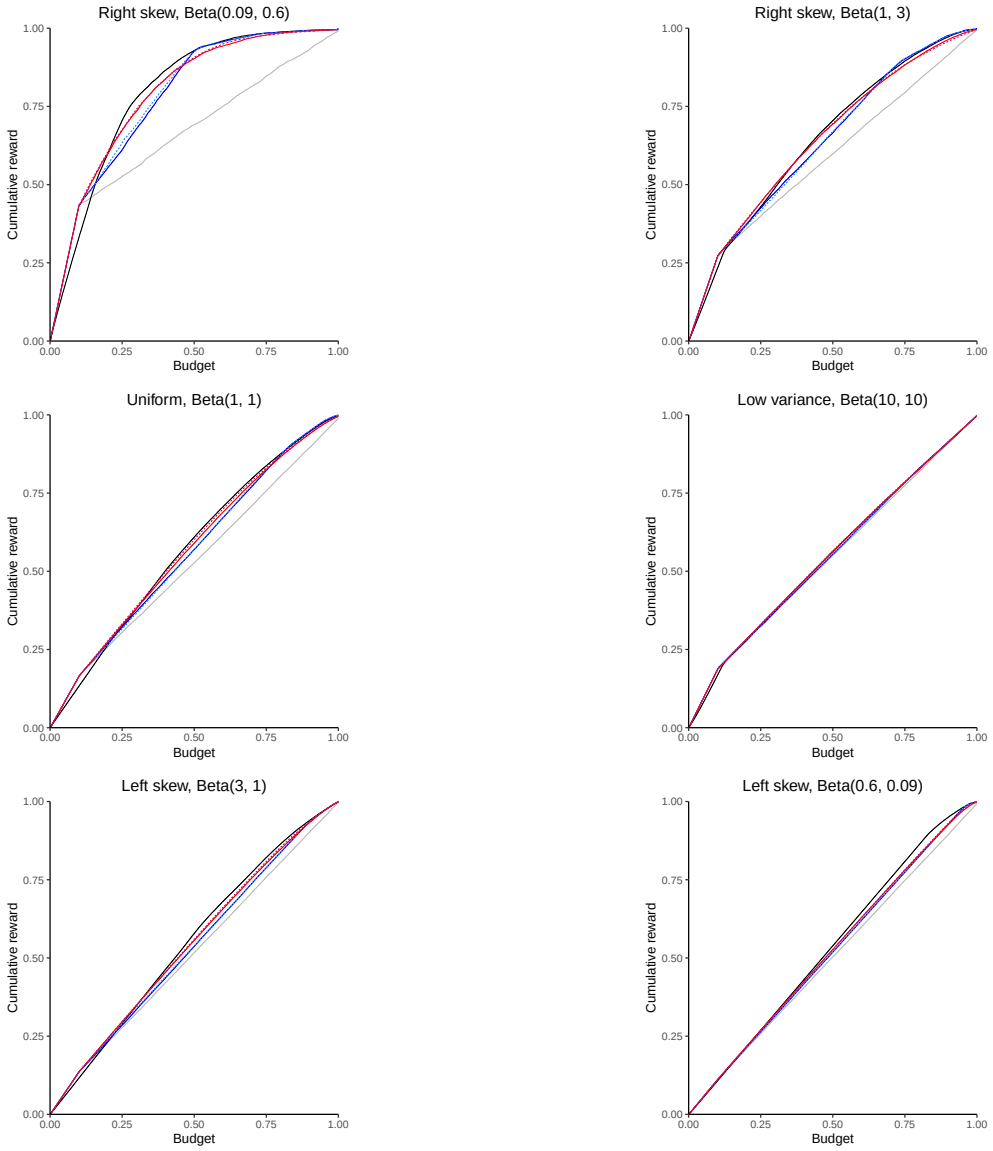


Figure A.10: Cumulative reward over the total time horizon for different policies and various reward distributions based on data simulated without branching and Poisson(10) lifetime. The axes are normalized to facilitate visual comparison. Thompson sampling is in black; pilot sampling with uniform sampling and lifetime sampling are in dark blue and light blue respectively; adaptive greedy with uniform sampling and sampling by lifetime are in red and pink respectively, and; naive sampling is in grey. These results are identical to simulations with a larger mean degree in Figure 2.1.

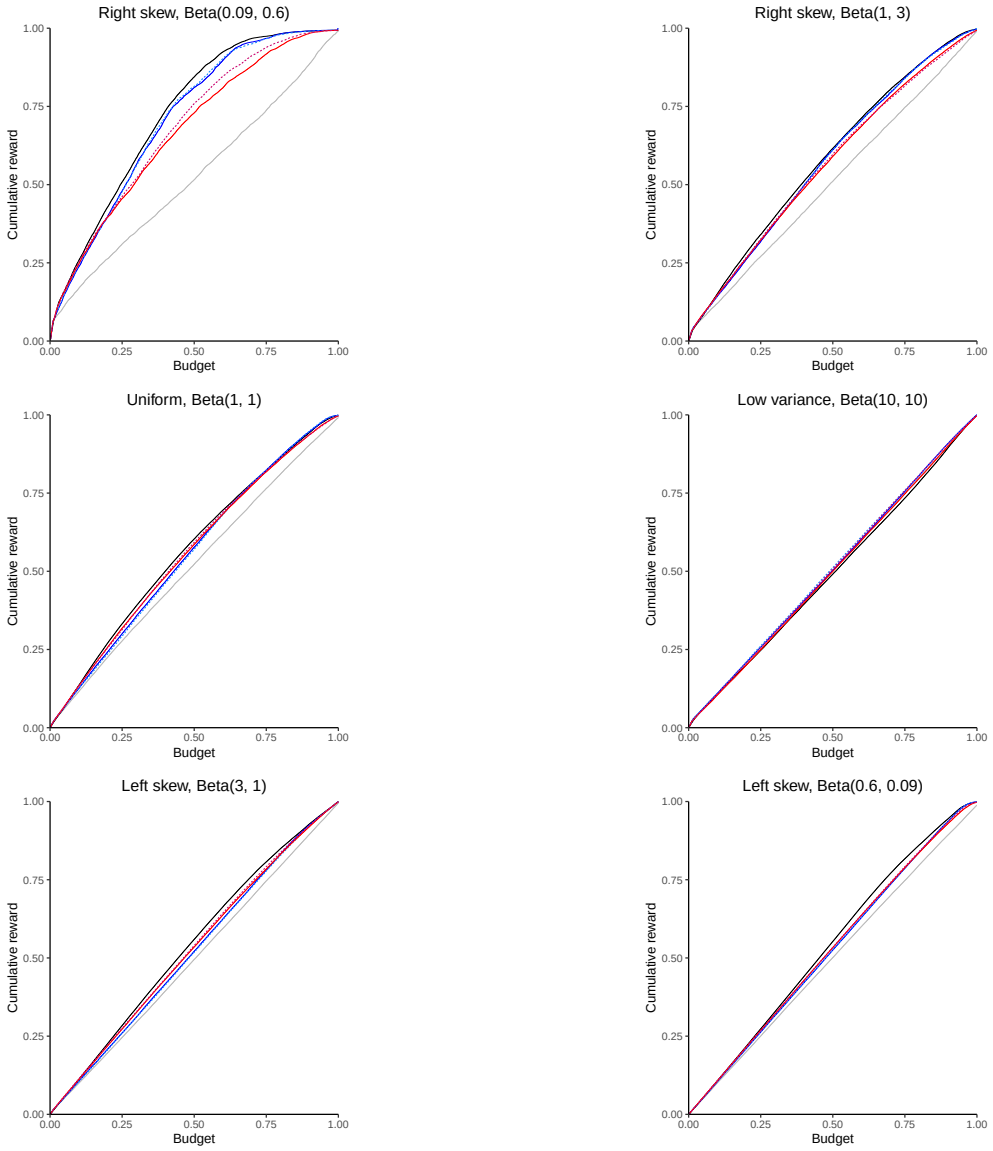


Figure A.11: Cumulative reward over the total time horizon for different policies and various reward distributions based on data simulated with branching and Poisson(10) lifetime. The axes are normalized to facilitate visual comparison. Thompson sampling is in black; pilot sampling with uniform sampling and lifetime sampling are in dark blue and light blue respectively; adaptive greedy with uniform sampling and sampling by lifetime are in red and pink respectively, and; naive sampling is in grey. These results are identical to simulations with a larger mean degree in Figure A.7.

Region	Estimated Poisson mean	Estimated Pareto location and shape
Punjab, Pakistan	10.58	(1, 0.51)
Punjab, India	26.62	(1, 0.39)
Southern India	6.36	(1, 0.71)

Table A.3: Estimated lifetime, or degree distribution, parameters for each dataset.

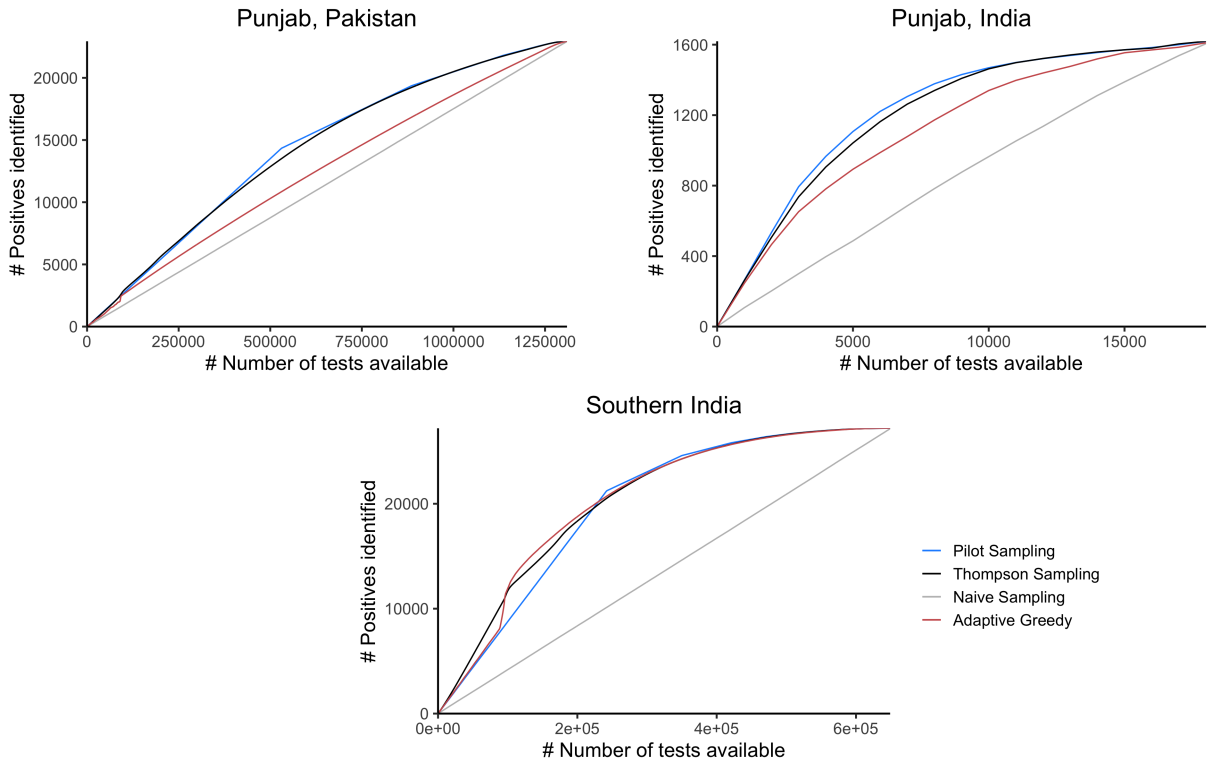


Figure A.12: The top left figure shows results on Punjab, Pakistan dataset. The top right figure shows results on the Punjab, India dataset. The bottom figure shows results on the southern India dataset. Pilot sampling and Thompson sampling are clearly doing better when we have branching data (both Punjab datasets). However, Thompson sampling may be logistically difficult making pilot sampling favorable when implementing contact tracing. This is identical to Figure 2.4 except for the axes' scales.

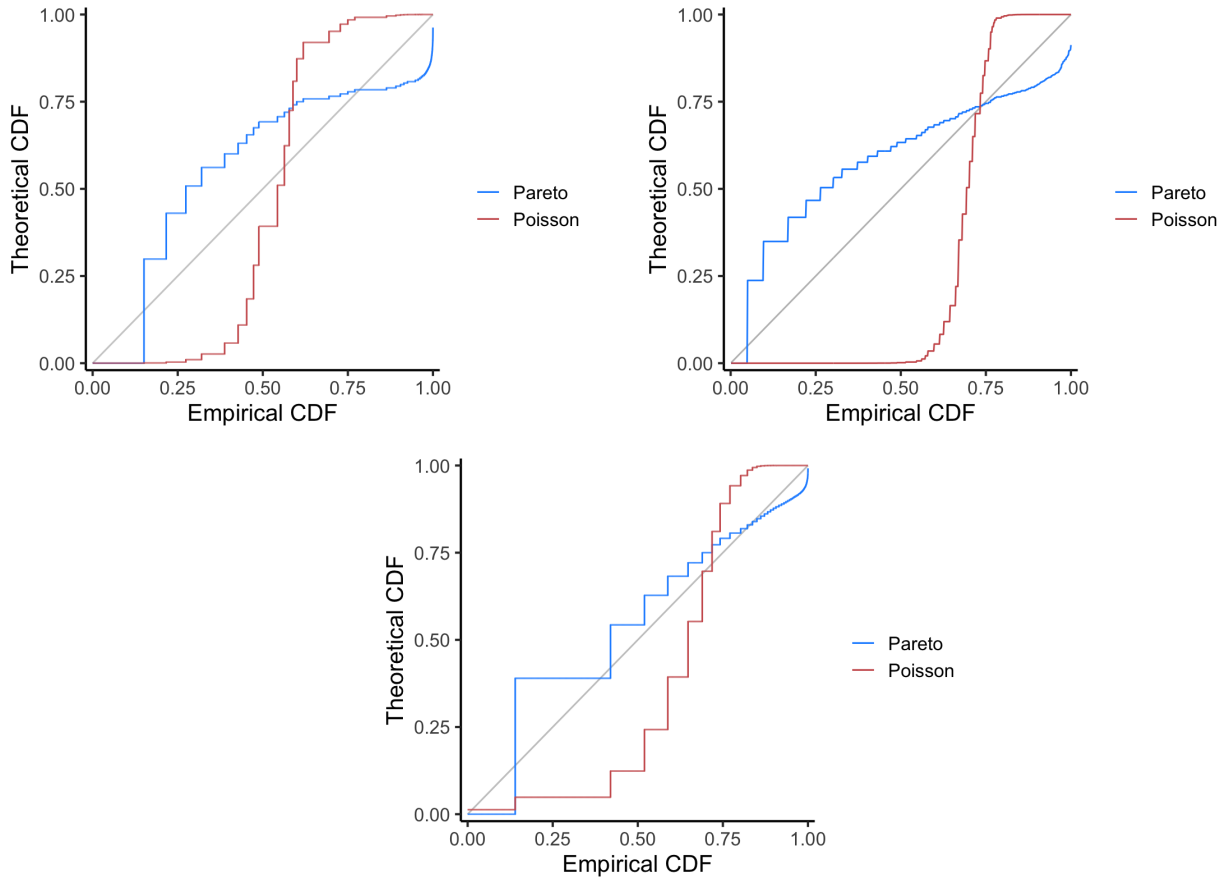


Figure A.13: The top left figure shows the P-P plot for the Punjab, Pakistan dataset. The top right figure shows the P-P plot for the Punjab, India dataset. The bottom figure shows the P-P plot for the southern India dataset. Pareto distribution seems to be a better fit than the Poisson distribution in all datasets reinforcing our belief that real-world networks tend to have heavy tails.

Appendix B

SUPPLEMENT TO CHAPTER 3

"This is your choice. You cannot pick the destination, only the path."

— Jasnah, *Oathbringer*

This appendix contains technical details, proofs, and additional discussion on relevant methods for the work presented in Chapter 3.

B.1 Permissibility and Hasse diagrams

One way to understand permissibility is by arranging the data of feature combination assignments into a *feature variant aggregation design matrix*, $\mathbf{F} \in \{0,1\}^{n,K}$. The entries of the matrix are as follows. If $k(i)$ is the feature combination that i is assigned to, we set

$$F_{i\ell} := \mathbb{I}\left\{k(i) \geq \ell \ \cap \ \rho(k(i)) = \rho(\ell)\right\}.$$

So the variant design matrix switches on a dummy variable for all variants that are subordinate to $k(i)$. The utility is that it allows for us to understand the marginal value of climbing the ordering up from $k(i)$, as in the *treatment variant aggregation* (TVA) procedure of Banerjee et al. (2021).

To see this, it is useful to rewrite Equation (3.1) in its variant form,

$$\mathbf{y} = \mathbf{F}\boldsymbol{\alpha} + \epsilon, \tag{B.1}$$

which is just a linear transformation of Equation (3.1), with $\boldsymbol{\beta}$ described by Equation (3.8),

$$\beta_k = \sum_{k' \leq k; \rho(k) = \rho(k')} \alpha_{k'}.$$

It is useful to represent our framework in a Hasse diagram. We imagine that moving *up* from one node to its adjacent node in Hasse inherits a value that corresponds to the marginal change in the outcome moving from an immediate subordinate variant to the present variant.

Of course, in this particular parameterization of $\boldsymbol{\beta}$, we chose to climb *up* the Hasse. We could have alternatively chosen to climb *down* the Hasse as

$$\mathbf{y} = \mathbf{G}\boldsymbol{\gamma} + \epsilon, \tag{B.2}$$

$$\beta_k = \sum_{k' \geq k; \rho(k) = \rho(k')} \gamma_{k'}, \tag{B.3}$$

where $G_{i\ell} := \mathbb{I}\left\{k\left(i\right) \leq \ell \ \cap \ \rho\left(k\left(i\right)\right) = \rho\left(\ell\right)\right\}$.

Consider the following definition of permissibility that places topological restrictions on any partition of the Hasse.

Definition B.1.1 (Permissible partition of a profile)**.** *A partition* Π_0 *of a profile* ρ_0 *is* permissible *if and only if*

(1) every $\pi \in \Pi_0$ *is a* pool *(cf. Definition 3.4.2),*

(2) every $\pi \in \Pi_0$ *is* strongly convex *i.e.,* $k, k' \in \pi$ *and* $k \geq k'' \geq k'$ *implies* $k'' \in \pi$*, and* $\min \pi$ *and* $\max \pi$ *both exist and are unique (and possibly equal to each other), and*

(3) Π_0 *respects* parallel splits*, i.e., for every pair of distinct pools* $\pi_i, \pi_j \in \Pi_0$

(a) if $\min \pi_i \lneqq \min \pi_j$*, then there exists a* $\pi' \in \Pi_0$ *such that* $\min \pi' = p'$*, where for each feature* m*,* $p'_m := \max\{p_m^{(i)}, p_m^{(j)}\}$ *where* $p^{(i)} = \min \pi_i$ *and* $p^{(j)} = \min \pi_j$*, and*

(b) if $\max \pi_i \not\leqslant \max \pi_j$*, then there exists a* $\pi'' \in \Pi_0$ *such* $\max \pi'' = p''$*, where for each feature* m*,* $p''_m := \min\{\tilde{p}^{(i)}_m, \tilde{p}^{(j)}_m\}$ *where* $\tilde{p}^{(i)} = \max \pi_i$ *and* $\tilde{p}^{(j)} = \max \pi_j$*.*

Condition (1) is obvious. Condition (2) says pools must contain contiguous features. Condition (3) is more technical but says that pools must be parallel on the Hasse.

The geometric definition of permissibility presented in Definition B.1.1 is equivalent to the statistically measurable definition we presented in Definition 3.4.8. We formally state this and prove it below.

Lemma B.1.2. *Definitions B.1.1 and 3.4.8 are equivalent to each other.*

Proof of Lemma B.1.2. We first look at permissibility as defined in Definition B.1.1. Condition (1) simply comes from the definition of a pool (cf. Definition 3.4.2). To understand the necessity of strong convexity in condition (2) and the parallel splitting criteria in condition (3), we need to understand how the marginal increments α_k affect the overall outcome $\beta_{k'}$. Observe that α_k affect $\beta_{k'}$ for all feature combinations $k' \geq k$ by Equation 3.8. We can define this "sphere of influence" of k as $A_k = \{k' \in \mathcal{K} \mid \rho(k) = \rho(k'), k' \geq k\}$. Further, when we are interested in the outcomes $\beta_{k'}$, it is sufficient to consider only spheres A_k where $\alpha_k > 0$ i.e., the "active spheres." Therefore, the intersection of all active spheres (taken either directly or through its complement) will give rise to a set of feature combinations with the same outcome i.e., a pool. It is easy to see that any such sphere is strongly convex in our sense. Therefore, any pool will also be strongly convex.

The parallel splitting criteria "from above" in condition ((3)a) of Definition B.1.1, also follows from these spheres of influence interpretation of the active α_k. Specifically, tracking the active α_k ensures that if a segment through a Hasse is pooled, then any segment both parallel to it and below it must be pooled as well. For the sake of contradiction, assume to the contrary that the top segment is pooled while a parallel bottom segment is cleaved. There must be some node along the bottom segment that was responsible for this cleaving through its marginal effect. However, the sphere of influence of this marginal effect cuts through the top segment too, cleaving the top segment into distinct pooled sets, a contradiction.

Of course, as mentioned in Section 3.4, it is possible to have exact marginal increments exactly offset each other so that despite two feature combinations k_1, k_2 influenced by two different spheres of action, $\beta_{k_1} = \beta_{k_2}$. However, we do not want to pool these features together because adding very little noise to one of the corresponding active $\alpha_{k'}$ will immediately render $\beta_{k_1} \neq \beta_{k_2}$ i.e., this is a measure zero event.

The above shows that permissibility (with condition (3) limited to condition ((3)a)) is necessary from just using the spheres of influence of marginal effects, and nothing more. However, this version of permissibility is also sufficient: any permissible partition Π_0 (as per Definition B.1.1) can be shown to derive from a common set of nodes $k_1, ...k_n$ so that each $\pi \in \Pi_0 = A_{k_1}^{a_1} \cap .... \cap A_{k_n}^{a_n}$, where $a_i \in \{1, c\}$, i.e. denoting either the sphere of influence or its complement. A quick proof sketch is as follows. First, we define the spheres through Π_0. For each $\pi_i \in \Pi_0$, take $k_i = \min \pi_i$ its unique minimum. These k_i will be the nodes that generate the spheres of influence. We will show that for any $p_1, p_2 \in \pi_i \in \Pi_0$, $p_1 \in A_{k_j} \iff p_2 \in A_{k_j}$. Observe that for any $p \in \pi_i$, then trivially $p \in A_{k_i}$. Now consider the condition when $p \in A_{k_j}$ for $k_j \neq k_i$. Then, $p > k_j$. It is not possible that $k_j \not\leqslant k_i$ as this would violate the parallel splits criteria (one can show there would be another $\pi' \in \Pi_0$ containing p). It is also not possible for $k_j > k_i$ as this would violate convexity ($p > k_j > k_i$ would imply $k_j \in \pi_i$). Thus, if $p \in \pi_i$ and $p \in A_{k_j}$, then necessarily $k_j \leq k_i$. Then, it follows that if $p_1, p_2 \in \pi_i$, then $p_1 \in A_{k_j} \iff p_2 \in A_{k_j}$. Therefore, all members of each $\pi \in \Pi_0$ are in the same unique intersection of spheres of influence, so each $pi \in \Pi_0$ is uniquely represented as $\pi = A_{k_1}^{a_1} \cap .... \cap A_{k_n}^{a_n}$.

It is easy to see with an analogous sphere of influence argument with $\boldsymbol{\gamma}$ that permissibility (With condition (3) limited to ((3)b) this time) is necessary and sufficient characterization of pools from active marginals in $\boldsymbol{\gamma}$.

□

When we wish to learn heterogeneity in $\boldsymbol{\beta}$, there is no reason to prefer one parameterization of climbing the Hasse over the other. Seeing that the parallel splitting criteria is linked

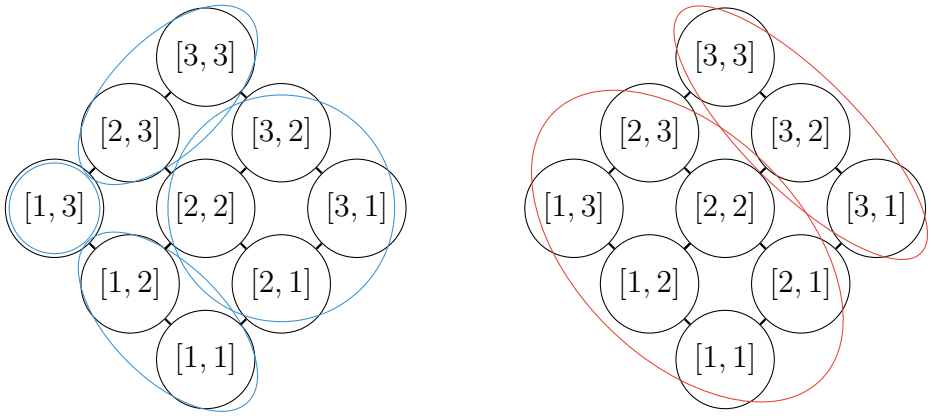


Figure B.1: Hasse diagram for Examples B.1.3 and B.1.4. The partition described in Example B.1.3 is shown in blue ellipses on the left panel. The right panel describes a different admissible partition in red ellipses seen in Example B.1.4

to robustly estimating the pools of heterogeneity, we want to obey both of them together at the same time. This does run the risk of generating more granular partitions as a result of stronger restrictions, but this is a small price to pay for robustly estimating heterogeneity when one wishes to be agnostic about the system. Hence the full criterion for permissibility Definition B.1.1, respecting parallel splits from both above (condition ((3)a)) and below (condition ((3)b)).

One might imagine that asking for more restrictions can complicate the search process. However, a by-product of being agnostic to the direction of Hasse traversal is that there is a bijective mapping between the $\mathbf{\Sigma}$ partition matrices and permissible partitions. We show in Proposition 3.5.1 that this significantly reduces the size of the model class, and later show in Theorem 3.5.3 that the size of the RPS, which is our primary estimation goal, is only polynomial. This has very important practical implications for computational feasibility.

We discuss specific examples of using the $\mathbf{\Sigma}$ matrix to represent partitions in Examples B.1.3, B.1.4, B.1.5 below.

Example B.1.3. Consider an example with $M = 2$ features, each with $R = 3$ discrete values, $\{1, 2, 3\}$. Then there are $K = R^M = 9$ different feature combinations. The Hasse

diagram is shown in Figure B.1. So, we end up pooling $(2,2)$ with $(3,2)$ and $(2,3)$ with $(3,3)$. The corresponding $\mathbf{\Sigma} \in \{0,1\}^{2\times 2}$ matrix for this profile is

$$\mathbf{\Sigma} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

This indicates that we split variants with value 1 from value 2 in the first feature (by $\Sigma_{11} = 0$) and pool variants of value 2 with value 3 in the first feature (by $\Sigma_{12} = 1$). Further, we pool variants with value 1 and value 2 in the second feature (by $\Sigma_{21} = 1$) and split variants with value 2 from value 3 in the second features (by $\Sigma_{22} = 0$).

□

Example B.1.4. Consider the same setup in Example B.1.3 with $M = 2$ features, each with $R = 3$ discrete values, $\{1,2,3\}$. Another permissible partition can be defined by the matrix

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}.$$

The pools are $\pi_1 = \{(a,b) \mid a = \{1,2\}, b = \{1,2,3\}\}$ and $\pi_2 = \{(a,b) \mid a = \{3\}, b = \{1,2,3\}\}$. This is illustrated in the right panel of Figure B.1. □

Example B.1.5. Consider a different setup with $M = 2$ features, The first feature takes on $R_1 = 5$ discrete values $\{1,2,3,4,5\}$ and the second feature takes on $R_2 = 3$ discrete values, $\{1,2,3\}$. An permissible partition can be defined by the matrix

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & - & - \end{bmatrix},$$

where we use "$-$" to denote that the second feature does not have dosages corresponding to those entries in the $\mathbf{\Sigma}$ matrix. The pools are $\pi_1 = \{(a,b) \mid a,b \leq 2\}$, $\pi_2 = \{(a,b) \mid a \leq 2, b =$

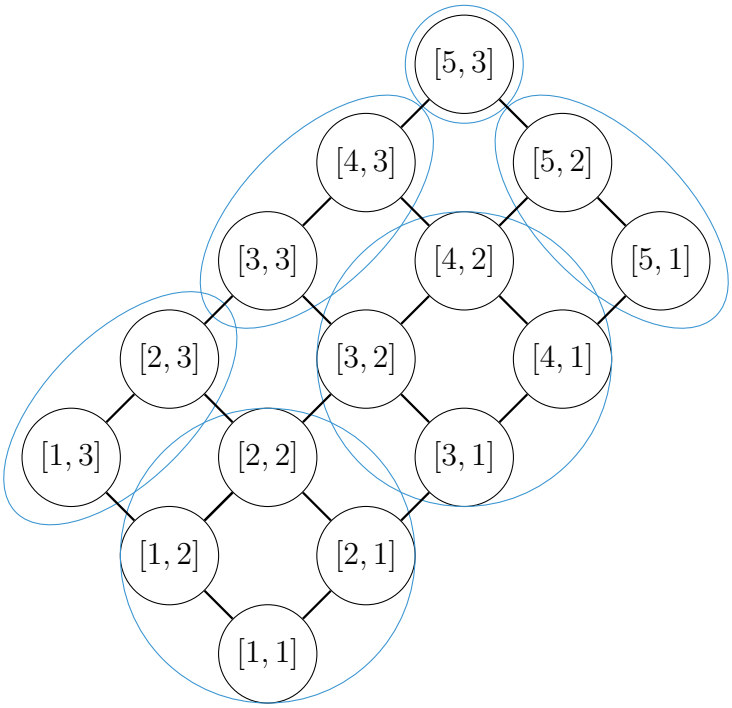


Figure B.2: Hasse diagram for Example B.1.5. The admissible partition is shown in blue ellipses.

$3\}$, $\pi_3 = \{(a,b) \mid a = 3,4, b \leq 2\}$, $\pi_4 = \{(a,b) \mid a = 3,4, b = 3\}$, $\pi_5 = \{(a,b) \mid a = 5, b \leq 2\}$, and $\pi_6 = \{(5,3)\}$. This is illustrated in Figure B.2. □

One can quickly verify that Examples B.1.3 - B.1.5 satisfy permissibility as defined in Definition B.1.1 by visual inspection and identifying the corresponding $\mathbf{\Sigma}$ matrices. In Example B.1.6 below, we show an example of a partition that is not permissible. Interestingly, there is a valid decision tree that arrives at this partition.

Example B.1.6. Consider the same setup in Example B.1.3 with $M = 2$ features, each with $R = 3$ discrete values, $\{1,2,3\}$. In Figure B.3, we illustrate a partition that is not permissible. This is not permissible because we have pools $\pi_1 = \{(1,1),(1,2),(1,3)\}$, $\pi_2 = \{(2,1),(2,2)\}$, $\pi_3 = \{(3,1),(3,2)\}$, and $\pi_4 = \{(2,3),(3,3)\}$. Permissibility (see condition (3) of Definition B.1.1) says that if π_1 is in the partition, then feature combinations $(\cdot,2)$ and $(\cdot,3)$ should

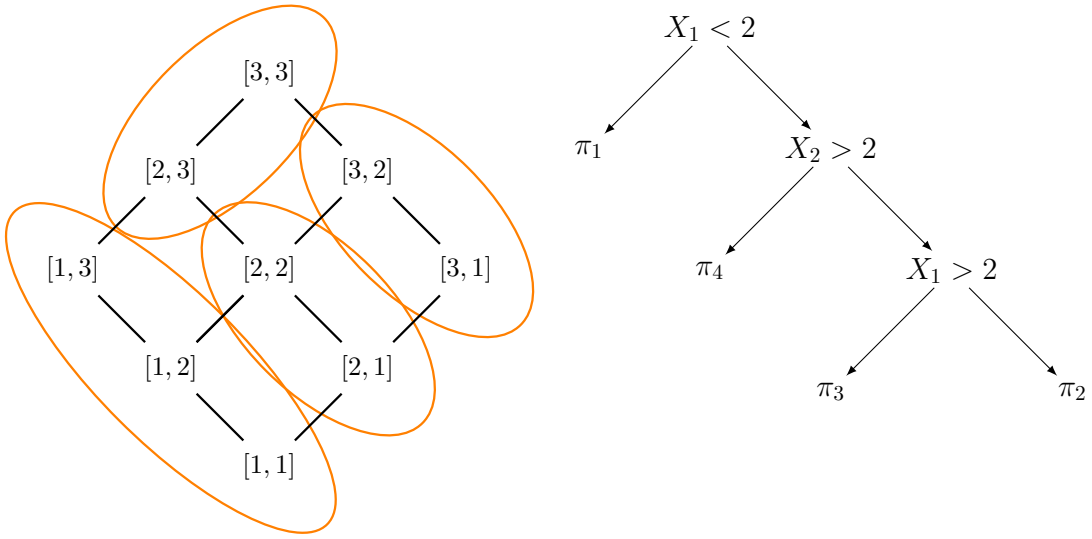


Figure B.3: Hasse diagram with the partition that is not permissible described in Example B.1.6. The pools are $\pi_1 = \{(1,1),(1,2),(1,3)\}$, $\pi_2 = \{(2,1),(2,2)\}$, $\pi_3 = \{(3,1),(3,2)\}$, and $\pi_4 = \{(2,3),(3,3)\}$. The decision tree illustrates how to generate this partition.

always be pooled together. This contradicts what we observe in π_2, π_3, and π_4. Similarly, if π_4 is in the partition, permissibility would require that feature combinations $(2,\cdot)$ and $(3,\cdot)$ need to be pooled together which contradicts π_2 and π_4. Since this partition is not permissible, we cannot represent it using the $\mathbf{\Sigma}$ matrix.

To see why this is not permissible from the marginal perspective, let us look at π_3 and π_4. From these pools, it is evident that $\alpha_{2,3} \neq 0$, $\alpha_{3,1} \neq 0$, and $\alpha_{3,3} \neq 0$. And we know that,

$$\beta_{3,3} = \alpha_{3,3} + \alpha_{3,2} + \alpha_{3,1} + \alpha_{2,3} + C$$
$$\beta_{2,3} = \alpha_{2,3} + C,$$

where the term C is common to both equations. In the pooling currently, it so happens that $\beta_{3,3} = \beta_{2,3}$ – the terms $\alpha_{3,3}, \alpha_{3,2}, \alpha_{3,1}$ jointly make this true by $\alpha_{3,3} + \alpha_{3,2} + \alpha_{3,1} = 0$. However, we know that $\alpha_{3,1} \neq 0$. So if we add some noise $\varepsilon > 0$ to $\alpha_{3,1}$ to get $\alpha'_{3,1} := \alpha_{3,1} + \varepsilon$. Then, $\beta_{3,3} \neq \beta_{2,3}$ anymore. In other words, the pool π_4 is not robust to noise in the non-zero

marginals as any noise will almost surely break π_4 into $\{(2,3)\}$ and $\{(3,3)\}$ as separate pools. Hence, this partition is not permissible.

Decision trees are not robust in this sense as they may generate partitions that are not permissible. The right panel of Figure B.3 illustrates a decision tree that generates the partition that is not permissible discussed in this example. □

B.1.1 Pooling across profiles

So far, we have been describing how to pool different feature combinations if they belong to the same profile. Now, we turn our attention to pooling *across* profiles. Definition 3.4.8 captures permissibility within a single profile, but we also want to consider pooling across profiles. For example, Definition B.1.1 does not speak to the question of pooling decisions for adding ibuprofen, as a temporary pain reliever, to a prescription of amoxicillin against a bacterial infection. Does introducing ibuprofen make an appreciable difference (offering the patient relief while waiting for the bacterial infection to work) or not (because the antibiotic itself offers pain relief by attacking the root cause)?

In order to reason about this, we consider partially ordering of the profiles themselves using their binary representation in Definition 3.4.5. This also allows us to embed the profiles in an M-d unit hypercube with profiles as the vertices. By the same intuition behind convexity, we can pool two profiles if they are reachable on this hypercube. We can generalize the marginal re-parameterization to allow for marginal gains when moving between profiles through a $\boldsymbol{\delta}$ vector,

$$\mathbf{y} = \mathbf{F}\boldsymbol{\alpha} + \mathbf{A}\boldsymbol{\delta} + \epsilon, \tag{B.4}$$

$$A_{i,(\ell,\rho)} = \mathbb{I}\left\{k(i) > \ell \ \cap \ \rho(k(i)) > \rho(\ell) = \rho\right\},$$

$$\beta_k = \sum_{k' \leq k; \rho(k) = \rho(k')} \alpha_{k'} + \sum_{k' < k} \sum_{\rho; \rho(k') < \rho(k)} \delta_{k',\rho}. \tag{B.5}$$

Here, observe that $\boldsymbol{\delta}$ is indexed by the profile ρ as well as the feature k'. Being feature-specific

gives the freedom to pool across profiles without imposing strong cross-profile restrictions that prevent measure zero events.

By setting $\boldsymbol{\delta} = \mathbf{0}$, we can immediately see that Equation (B.4) is a generalization of Equation (B.1). In fact, if depending on the context, we do not want to pool profile ρ_1 with ρ_2, then this corresponds to setting the appropriate entries in $\boldsymbol{\delta}$ to 0. This is exactly what Banerjee et al. (2021) do in their analysis of cross-randomized behavioral nudges for improving immunization.

We formalize this in Definition B.1.7.

Definition B.1.7 (Permissible partition)**.** *A* partition Π of the entire feature space $\mathcal{K}$ is permissible *if:*

(1) for every profile ρ_0*, the partition induced by* Π *on* ρ_0*,* $\Pi_0 = \{\pi \backslash \{k \mid \rho(k) \neq \rho_0\} \mid \pi \in \Pi\}$ *is permissible by Definition B.1.1, and*

(2) for every $\boldsymbol{\beta}$ *that generates* Π*, with respect to the Lebesgue measure, the support of* $\boldsymbol{\delta}(\boldsymbol{\beta})$*,* $S_{\boldsymbol{\delta}(\boldsymbol{\beta})} = \{\delta_{k,\rho} \neq 0 \mid \delta_{k,\rho} \in \boldsymbol{\delta}(\boldsymbol{\beta})\}$*, is measurable.*

Specifically, by allowing to pool across different profiles, Definition B.1.7 naturally allows us to explore heterogeneity in treatment effects where treatment and control are two distinct profiles. We illustrate this in the empirical data analysis of microcredit access in Section 3.8.

Definition B.1.8 gives an equivalent geometric interpretation of Definition B.1.7 through the Hasse.

Definition B.1.8 (Permissible partition)**.** *A* partition Π of the entire feature space $\mathcal{K}$ is permissible *if and only if the following hold true:*

(1) for every profile ρ_0*, the partition induced by* Π *on* ρ_0*,* $\Pi_0 = \{\pi \backslash \{k \mid \rho(k) \neq \rho_0\} \mid \pi \in \Pi\}$ *is permissible by Definition B.1.1,*

(2) *every* $\pi \in \Pi$ *is connected in feature levels across profiles i.e., if* $k_1, k_2 \in \pi$ *such that* $\rho_1 = \rho(k_1)$ *and* $\rho_2 = \rho(k_2)$ *are adjacent on the hypercube, then there are feature combinations* $k_1', k_2' \in \pi$ *such that* $\rho(k_1') = \rho_1$, $\rho(k_2') = \rho_2$ *and* $\|k_1' - k_2'\|_1 = 1$,[1] *and*

(3) *every* $\pi \in \Pi$ *is connected in profiles i.e., if* π *contains feature combinations from profiles* ρ_0 *and* ρ_k *where* $\rho_0 < \rho_k$, *then* π *also contains features in profiles* $\rho_1, \ldots, \rho_{k-1}$ *such that* $\|\rho_i - \rho_{i+1}\|_0 = 1$ *for* $i = 0, \ldots, k-1$.[2]

This representation agrees with our permissibility in Definition B.1.8. Case (1) follows from the fact that this is a generalization of Equation B.1. Cases (2) and (3) follow from the definition of the $\mathbf{A}$ matrix. For example, consider two features k_1, k_2 that belong to two different profiles. We can only pool variants i.e., $||k_1 - k_2||_1 = 1$. If they are variants, then the two profiles must be adjacent on the M-d hypercube.

At this point, it is important to note that there are no restrictions such as the parallel splitting criteria across different profiles. This is because the marginal δ_k only contributes to the outcome across profiles i.e., from the perspective within a profile, the sphere of influence of δ_k is indistinguishable from the sphere of influence of $\alpha_{k'}$ where k' is at the lower boundary of the Hasse adjacent to the Hasse of k. Since each pair of feature variants from different profiles have different across-profile marginals δ_k, they are not coupled together like the $\boldsymbol{\alpha}$ marginals are.

Just like the $\boldsymbol{\Sigma}$ matrix within each profile, we can also construct the intersection matrix $\boldsymbol{\Sigma}^{\cap}$ to denote how features are pooled across two adjacent profiles. Consider partitions induced by Π on two profiles ρ_1 and ρ_2. Let us call these Π_1, Π_2 respectively. $\boldsymbol{\Sigma}^{\cap} = \{0, 1, \infty\}^{|\Pi_1| \times |\Pi_2|}$ where $\boldsymbol{\Sigma}^{\cap}_{i,j} = 0$ means that pools $\pi_i \in \Pi_1$ and $\pi_j \in \Pi_2$ are poolable according to (2) of Definition B.1.8 but are not pooled together in Π. $\boldsymbol{\Sigma}^{\cap}_{i,j} = 1$ means that

[1] Along with (1), this means that we can reach k_2 from k_1 by traversing the Hasse for ρ_1 to k_1', then jumping to k_2' along an edge on the M-d hypercube, and then moving from k_2' to k' while respecting the Hasse for ρ_2.

[2] Along with (1) and (2), this means that we can reach ρ_k from ρ_0 by traversing the M-d hypercube while staying within π and respecting the Hasse at each vertex of the hypercube.

these pools are poolable and are indeed pooled in Π. Finally, $\mathbf{\Sigma}^{\cap}_{i,j} = \infty$ means that these pools are not poolable by Definition B.1.8. Observe that if $\mathbf{\Sigma}^{\cap}_{i,j} = 1$, then $\mathbf{\Sigma}^{\cap}_{i,-j} = \infty$ and $\mathbf{\Sigma}^{\cap}_{-i,j} = \infty$ in order to respect (1) of Definition B.1.8. This object is useful in our enumeration step in Algorithm 3.

B.1.2 Examples of other permissibility restrictions

Estimation strategies, in general, implicitly take some stand on partition structure. They impose permissibility restrictions, though they are not often presented formally as such. These restrictions are generated by the choice of technique, rather than through any specific scientific consideration. In the following examples, we show how these techniques can be framed as permissibility restrictions by defining them as partitions on the Hasse. This involves identifying sets of equivalent edges for each technique that, when removed together, generate corresponding partitions learned by that technique.

Example B.1.9 (Long, Short, and Lasso regression)**.** First, take a saturated or "long" regression. Here, every possible feature combination is its own pool i.e., the partition is the most granular partition possible. For example, consider the treatment outcome as a function of dosages of two drugs, A, B, and weight of the treated individual, W. Suppose that each variable takes on three discrete levels. Then, $\Pi^{\text{long}} = \{(a, b, w) : \quad \text{for every } a, b, w \in \{0, 1, 2\}^2 \times \{\text{low, med, high}\}\}$ with 27 elements.

Second, consider a "short" regression where the researcher does not include all relevant variables in the regression. Then, the partition generated is identical to long regression for variables included in the model i.e., it pools across all excluded variables. For example, if the researcher ignores weight, then $\Pi^{\text{short}} = \{(a, b, :) : \quad \text{for every } a, b \in \{0, 1, 2\}^2\}$ which pools across weight, with 9 elements.

Third, say the researcher uses Lasso to regularize the data to set marginal dosage or weight increase effects to 0, generating pools. Then the Π^{Lasso} is bijectively determined by

the support of the Lasso: the zero'ed elements generate the pooling structure.[3] □

Example B.1.10 (Decision Trees)**.** This is perhaps the most common approach used beyond imposed short and long regressions. In a decision tree, at every node, the statistician chooses whether or not to split based on the value of a given arm (e.g., amoxicillin greater than 250mg). Conditional on this split, following a given decision, another variable is selected, and the process repeats recursively until termination (maybe defined by some maximum number of splits or until no more splits are possible). It is useful to note that this procedure generates parallel splits in the Hasse, conditional on previously made splits. Therefore, decision trees generate convex partitions.

It is useful to note that the decision trees are captured by the equivalent edge framing in a conditional setting. Initially, the set of equivalent edges for a decision tree is identical to those of the robust partitions we consider in this paper. However, the edges to split upon are chosen sequentially (rather than jointly, as in the case of robust partitions). Thus, after each split, the set of equivalent edges changes. Specifically, each set of equivalent edges could get decomposed into two smaller sets of equivalent edges upon splitting a different equivalence class (this means that there are more than 2^n possible partitions where there were n equivalence classes originally). In other words, edges are equivalent conditional on previously made splits. The binary σ vector we used to represent splitting of jointly equivalent edges can be generalized to a *list* of indices that alternatingly indicate splitting and pooling along the edges. For example, we can order the edges in equivalence class E_i. Then, the pooling decisions for E_i is represented by a list σ_i such that all edges until $e_{\sigma_{i,1}}$ are pooled, all edges from $e_{\sigma_{i,1}}$ to $e_{\sigma_{i,2}}$ are split, all edges from $e_{\sigma_{i,2}}$ to $e_{\sigma_{i,3}}$ are pooled, and so on. This is essentially a tree data structure, i.e., the σ data structure is a tree in the limit where we only have conditionally equivalent edges (but not jointly). The key point here is the perspective of encoding the partition by keeping track of splits (made in equivalence

[3] In a frequentist perspective, under "beta-min" classical assumptions and irrepresentability, this will correctly identify the true generative partition with probability tending to one (as seen in Banerjee et al. (2021)).

classes) is general, takes us to scientifically relevant settings beyond robust partitions and decision trees, and generates vast improvements in refinements as well. However, the pooling restrictions imposed by trees suffer from a coherency issue, which we explore in our discussion of causal trees. We walk through a detailed example in Example B.1.6. □

Example B.1.11 (Causal Trees and Causal Random Forests: conditionally convex splits)**.** Decision trees cannot natively estimate heterogeneities in treatment effects. Causal trees (Athey and Imbens, 2016) and causal random forests (Wager and Athey, 2018) can do this natively by modifying the fit criteria used to make splits. The fit criteria for causal trees is the MSE of the treatment effect. So each leaf of the tree needs to contain both treatment and control observations to estimate the treatment effect. The partitions generated by causal trees are identical to decision trees if we ignore the treatment indicator as $(1, x)$ and $(0, x)$ are always in the same leaf or pool. Here, x is the feature combination, and $1/0$ is the binary treatment indicator variable.

This differs from our robust partitions because we allow for pools with $\{(1, x_1), (1, x_2), (0, x_1)\}$ and $\{(0, x_2)\}$ by the cross-Hasse pooling rules. On the other hand, causal trees will split them as $\{(1, x_1), (0, x_1)\}$ and $\{(1, x_2), (0, x_2)\}$ (or pool them together) even though $(0, x_1)$ and $(1, x_2)$ has the same outcome. This raises a conceptual issue. If $(1, x_2)$ and $(1, x_1)$ are equivalent to each other, and $(1, x_1)$ and $(0, x_1)$ are equivalent to each other, then by transitivity, $(1, x_2)$ and $(0, x_1)$ should be equivalent to each other as well. However, causal trees force them apart because $(0, x_1)$ and $(0, x_2)$ are not equivalent to each other. Such a pooling restriction appears incoherent. Besides this, since we allow for flexibility through the transitivity of equivalences, we enjoy statistical properties such as lower bias and variance.

Causal random forests are an aggregation over sampled partitions, $\{\Pi_b : \ b = 1, \ldots, B\}$, each of which is bijective with a partition created by causal trees. □

Example B.1.12 (Treatment Variant Aggregation: Non-Convex Robust Splits)**.** Here the authors study a Hasse directly and impose permissibility settings slightly different from that

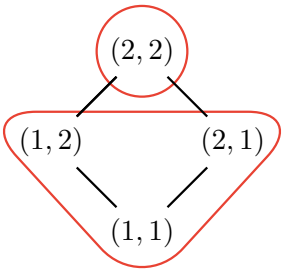


Figure B.4: Hasse diagram admissible that is not strongly convex but robust. This cannot be represented by a decision tree.

of the present paper. In this setting, the natural structure is to allow only one-sided convexity. Note that this is disallowed by a decision tree and scientifically desirable in certain settings.

Here, the equivalent edges framework does not hold. Consider the 2×2 Hasse with $\{11, 12, 21, 22\}$ as the four feature combinations. TVA allows for the following partitions of the Hasse:

- $\Pi_1 = \{\{11, 12, 21\}, \{22\}\}$
- $\Pi_2 = \{\{11\}, \{21\}, \{12, 22\}\}$
- $\Pi_3 = \{\{11\}, \{12\}, \{21, 22\}\}$
- $\Pi_4 = \{\{11, 12, 21, 22\}\}$
- $\Pi_5 = \{\{11\}, \{12\}, \{21\}, \{22\}\}$
- $\Pi_6 = \{\{11, 12\}, \{21, 22\}\}$
- $\Pi_7 = \{\{11, 21\}, \{12, 22\}\}$

In this 2×2 Hasse, there are four edges $\{\langle 11, 12\rangle, \langle 11, 21\rangle, \langle 12, 22\rangle, \langle 21, 22\rangle\}$. The partitions listed above illustrate that no edge is truly equivalent to another edge in the

Hasse despite TVA imposing convexity restrictions. In other words, the set of equivalent edges appears to be degenerate i.e., each each equivalence class is a singleton. However, there is a well-defined structure. We are not free to arbitrarily split on edges. For example, $\Pi_8 = \{\{11, 12, 22\}, \{21\}\}$ is not permissible as it violates convexity. This is why we say that the equivalent edges framing does not hold here.

However, we can still generalize the σ data structure to efficiently store this partition. This is because these partitions are "parallel from below," i.e., if a feature combination $k = [r_1, \ldots, r_i, \ldots, r_m]$ and $k' = [r_1, \ldots, r_i + 1, \ldots, r_m]$ are split, then all pairs of feature combinations $k_1 = [s_1, \ldots, r_i, \ldots, s_m]$ and $k_2 = [s_1, \ldots, r_i + 1, \ldots, s_m]$ where $k_1 > k$ and $k_2 > k'$ are also split. Therefore, using the same set of equivalent edge decomposition as our robust partitions (described below), we allow σ_i to denote a vector of largest levels in all features j, ℓ_j, besides the i-th one such that $k_1 = [\ldots, \ell_j, \ldots, r_i, \ldots]$ and $k_2 = [\ldots, \ell_j, \ldots, r_i + 1, \ldots]$ are pooled. □

B.2 Laplace approximation and generalized Bayesian inference

B.2.1 Laplace approximation

Here, we briefly outline how to approximate the full posterior using the Rashomon set and Laplace's method. Our goal is to estimate

$$p(\boldsymbol{\beta} \mid \mathbf{Z}) = \sum_{\Pi \in \mathcal{P}_\theta} p(\boldsymbol{\beta} \mid \mathbf{Z}, \Pi) \mathbb{P}(\Pi \mid \mathbf{Z})$$

We will do this by constructing a specific data-generating process. Consider the following data-generating process after fixing a partition Π. For each pool $\pi_j \in \Pi$, draw $\gamma_j \overset{i.i.d.}{\sim} \mathcal{N}(\mu_0, \tau^2)$ i.e., draw $\boldsymbol{\gamma} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0)$.[4] Here, $\boldsymbol{\gamma} \in \mathbb{R}^{|\Pi|}$ and $\boldsymbol{\Lambda}_0 = \tau^2 \mathcal{I}_{|\Pi|}$ where and $\mathcal{I}_m$ is an identity matrix of size m.

Then, we can define a transformation matrix $\boldsymbol{P} \in \{0, 1\}^{K \times |\Pi|}$, where K is the number of

[4] In this case, these draws need not be independent of identical. The computations just become a little more tedious.

possible feature combinations, that assigns each γ of each pool to the feature combinations in that pool,

$$P_{ij} = \begin{cases} 1, & \text{feature combination } i \in \pi_j \\ 0, & \text{else} \end{cases}.$$

The mean vector for the feature combinations is given by $\boldsymbol{\beta} = \boldsymbol{P}\boldsymbol{\gamma}$. By properties of the multivariate normal, we have $\boldsymbol{\beta} \mid \Pi \sim \mathcal{N}(\boldsymbol{\mu}_\Pi, \boldsymbol{\Lambda}_\Pi)$, where $\boldsymbol{\mu}_\Pi = \boldsymbol{P}\boldsymbol{\mu}_0$ and $\boldsymbol{\Lambda}_\Pi = \boldsymbol{P}\boldsymbol{\Lambda}_0\boldsymbol{P}^\intercal$. Specifically, note that the means of all feature combinations in a given pool don't just share the same mean, but are effectively equivalent to each other.

Then, given some feature combinations $\mathbf{D}$, we draw the outcomes as

$$\mathbf{y} \mid \mathbf{D}, \boldsymbol{\beta}, \Pi \sim \mathcal{N}(\mathbf{D}\boldsymbol{\beta}, \boldsymbol{\Sigma}) \implies \mathbf{y} \mid \mathbf{D}, \boldsymbol{\gamma}, \Pi \sim \mathcal{N}(\mathbf{D}\boldsymbol{P}\boldsymbol{\gamma}, \boldsymbol{\Sigma}).$$

Therefore, $\boldsymbol{\gamma} \mid \mathbf{Z}, \Pi \sim \mathcal{N}\left(\boldsymbol{\mu}_n, \boldsymbol{\Lambda}_n^{-1}\right)$ where

$$\begin{aligned} \boldsymbol{\mu}_n &= \boldsymbol{\Lambda}_n \left((\mathbf{D}\boldsymbol{P})^\top (\mathbf{D}\boldsymbol{P}) \widehat{\boldsymbol{\gamma}} + \boldsymbol{\Lambda}_0 \boldsymbol{\mu}_0 \right) \\ \boldsymbol{\Lambda}_n^{-1} &= (\mathbf{D}\boldsymbol{P})^\top (\mathbf{D}\boldsymbol{P}) + \boldsymbol{\Lambda}_0 \\ \widehat{\boldsymbol{\gamma}} &= ((\mathbf{D}\boldsymbol{P})^\top (\mathbf{D}\boldsymbol{P}))^{-1} (\mathbf{D}\boldsymbol{P})^\top \mathbf{y} \end{aligned}$$

Next, $\mathbb{P}(\Pi \mid \mathbf{Z}) = \mathbb{P}(\Pi) \int_{\boldsymbol{\gamma}'} p(\mathbf{Z} \mid \boldsymbol{\gamma}', \Pi) p(\boldsymbol{\gamma}' \mid \Pi) d\boldsymbol{\gamma}'$. We know that $\Pi(\Pi) = C \exp\{-\lambda |\Pi|\}$ where C is the normalization constant (or the partition function). Therefore, we have

$$\begin{aligned} \mathbb{P}(\Pi \mid \mathbf{Z}) &= C A_{1,\Pi} A_{2,\Pi} e^{-\lambda |\Pi|} \int_{\boldsymbol{\gamma}'} \exp\left\{-g(\boldsymbol{\gamma}')\right\} d\boldsymbol{\gamma}', \\ g(\boldsymbol{\gamma}') &= \frac{1}{2} (\boldsymbol{\gamma}' - \boldsymbol{\mu}_0)^\top \boldsymbol{\Lambda}_0 (\boldsymbol{\gamma}' - \boldsymbol{\mu}_0) + \frac{1}{2} (\mathbf{y} - \mathbf{D}\boldsymbol{P}\boldsymbol{\gamma}')^\top \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mathbf{D}\boldsymbol{P}\boldsymbol{\gamma}') \end{aligned}$$

where $A_{1,\Pi}, A_{2,\Pi}$ are known constants from the normal distributions. It is easy to verify that

$$\nabla g(\boldsymbol{\gamma}') = \boldsymbol{\Lambda}_0\boldsymbol{\gamma}' + (\mathbf{D}\boldsymbol{P})^\top\boldsymbol{\Sigma}^{-1}\mathbf{D}\boldsymbol{P}\boldsymbol{\gamma}' - \boldsymbol{\Lambda}_0\boldsymbol{\mu}_0 - (\mathbf{D}\boldsymbol{P})^\top\boldsymbol{\Sigma}^{-1}\mathbf{y}$$
$$\nabla^2 g(\boldsymbol{\gamma}') = \boldsymbol{\Lambda}_0 + (\mathbf{D}\boldsymbol{P})^\top\boldsymbol{\Sigma}^{-1}\mathbf{D}\boldsymbol{P}$$

Since $\nabla^2 g(\boldsymbol{\gamma}')$ is positive semi-definite for all $\boldsymbol{\gamma}'$, g is convex. Therefore, solving for $\nabla g(\boldsymbol{\gamma}') = 0$ allows us to find the minimum, $\boldsymbol{\gamma}^\star = (\boldsymbol{\Lambda}_0 + (\mathbf{D}\boldsymbol{P})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{D}\boldsymbol{P}))^{-1}(\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0 + (\mathbf{D}\boldsymbol{P})^\top\boldsymbol{\Sigma}^{-1}\mathbf{y})$.

Using Laplace approximation, we get

$$\mathbb{P}(\Pi \mid \mathbf{Z}) \approx C A_{1,\Pi} A_{2,\Pi} e^{-\lambda|\Pi| - g(\gamma^\star)} (2\pi)^{|\Pi|/2} \det(\boldsymbol{\Lambda}_0 + (\mathbf{D}\boldsymbol{P})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{D}\boldsymbol{P}))^{1/2}.$$

This allows us to approximate the original quantity of interest, $p(\boldsymbol{\beta} \mid \mathbf{Z})$ through a variable transformation of $\boldsymbol{\gamma}$ where all the constants are known except for C. Exactly computing C is NP-Hard and is reminiscent of partition functions used in graphical models (not to be confused with the "partition" that we are using in this work). See Agrawal et al. (2021) for a survey of methods used to estimate or approximate the constant C.

B.2.2 Generalized Bayesian inference

We have our mean squared error for a given partition Π,

$$\begin{aligned} \mathcal{L}(\Pi; \mathbf{Z}) &= \frac{1}{n}\left(\mathbf{y} - \widehat{\mathbf{y}}\right)^\intercal \left(\mathbf{y} - \mathbf{y}\right), \\ \widehat{y}_i &= \frac{\sum_{\pi\in\Pi} \mathbb{I}\{k(i) \in \pi\} \sum_j \mathbb{I}\{k(j) \in \pi\} y_j}{\sum_{\pi\in\Pi} \mathbb{I}\{k(i) \in \pi\} \sum_j \mathbb{I}\{k(j) \in \pi\}} \end{aligned} \tag{B.6}$$

where $\widehat{y}_i$ is the mean outcome in the pool $\pi \in \Pi$ containing the feature combination of unit i, $k(i)$.

Our goal is to show that minimizing $\mathcal{L}(\Pi; \mathbf{Z})$ corresponds to maximizing the likelihood $\mathbb{P}(\mathbf{y} \mid \mathbf{D}, \Pi)$. Consider the same data-generating process in Appendix B.2.1. Specifically, we require independence over γ_i and we will assume that the prior over $\boldsymbol{\gamma}$ is diffuse i.e., $\tau^2 \gg 1$.

As before, given some feature combinations $\mathbf{D}$, we draw the outcomes as

$$\mathbf{y} \mid \mathbf{D}, \boldsymbol{\beta}, \Pi \sim \mathcal{N}(\mathbf{D}\boldsymbol{\beta}, \boldsymbol{\Sigma}) \implies \mathbf{y} \mid \mathbf{D}, \boldsymbol{\gamma}, \Pi \sim \mathcal{N}(\mathbf{D}\boldsymbol{P}\boldsymbol{\gamma}, \boldsymbol{\Sigma}).$$

This allows us to find the likelihood,

$$\begin{aligned}
\mathbb{P}(\mathbf{y} \mid \mathbf{D}, \Pi) &= \int_{\boldsymbol{\beta}} \mathbb{P}(\mathbf{y} \mid \mathbf{D}, \Pi, \boldsymbol{\beta}) \mathbb{P}(\boldsymbol{\beta} \mid \Pi) d\boldsymbol{\beta} = \int_{\boldsymbol{\gamma}} \mathbb{P}(\mathbf{y} \mid \mathbf{D}, \Pi, \boldsymbol{\gamma}) \mathbb{P}(\boldsymbol{\gamma} \mid \Pi) d\boldsymbol{\gamma} \\
&= \int_{\boldsymbol{\gamma}} \mathbb{P}(\mathbf{y} \mid \mathbf{D}, \boldsymbol{P}, \boldsymbol{\gamma}) \mathbb{P}(\boldsymbol{\gamma} \mid \Pi) d\boldsymbol{\gamma} \\
&\propto \int_{\boldsymbol{\gamma}} \exp\left\{ -\frac{1}{2} (\mathbf{y} - \mathbf{D}\boldsymbol{P}\boldsymbol{\gamma})^{\intercal} \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mathbf{D}\boldsymbol{P}\boldsymbol{\gamma}) \right\} \exp\left\{ -\frac{1}{2} (\boldsymbol{\gamma} - \boldsymbol{\mu}_0)^{\intercal} \boldsymbol{\Lambda}_0^{-1} (\boldsymbol{\gamma} - \boldsymbol{\mu}_0) \right\} d\boldsymbol{\gamma}
\end{aligned}$$

After re-arranging the terms in the exponent, we have

$$-\frac{1}{2} \left(\left(\boldsymbol{\gamma} - \mathbf{M}^{-1} \boldsymbol{u} \right)^{\intercal} \mathbf{M} \left(\boldsymbol{\gamma} - \mathbf{M}^{-1} \boldsymbol{u} \right) + \mathbf{y}^{\intercal} \boldsymbol{\Sigma}^{-1} \mathbf{y} + \boldsymbol{\mu}_0^{\intercal} \boldsymbol{\Lambda}_0^{-1} \boldsymbol{\mu}_0 - \boldsymbol{u}^{\intercal} \mathbf{M}^{-1} \boldsymbol{u} \right),$$

where $\mathbf{M} = \boldsymbol{P}^{\intercal} \mathbf{D}^{\intercal} \boldsymbol{\Sigma}^{-1} \boldsymbol{P} \mathbf{D} + \boldsymbol{\Lambda}_0^{-1}$ and $\boldsymbol{u} = \boldsymbol{P}^{\intercal} \mathbf{D}^{\intercal} \boldsymbol{\Sigma}^{-1} \mathbf{y} + \boldsymbol{\Lambda}_0^{-1} \boldsymbol{\mu}_0$. Notice that when integrating with respect to $\boldsymbol{\gamma}$, the first quadratic term becomes a constant in $\mathbf{y}$. Therefore,

$$\mathbb{P}(\mathbf{y} \mid \mathbf{D}, \Pi) \propto \exp\left\{ -\frac{1}{2} \left(\mathbf{y}^{\intercal} \boldsymbol{\Sigma}^{-1} \mathbf{y} + \boldsymbol{\mu}_0^{\intercal} \boldsymbol{\Lambda}_0^{-1} \boldsymbol{\mu}_0 - \boldsymbol{u}^{\intercal} \mathbf{M}^{-1} \boldsymbol{u} \right) \right\}.$$

Now, as the prior over $\boldsymbol{\gamma}$ becomes more diffuse i.e., as $\tau^2 \to \infty$, we have that $\boldsymbol{\Lambda}_0^{-1} \to \mathbf{0}$. Therefore, $\boldsymbol{\mu}_0^{\intercal} \boldsymbol{\Lambda}_0^{-1} \boldsymbol{\mu}_0 \to \mathbf{0}$, $\mathbf{M} \to \boldsymbol{P}^{\intercal} \mathbf{D}^{\intercal} \boldsymbol{\Sigma}^{-1} \boldsymbol{P} \mathbf{D}$, and $\boldsymbol{u} \to \boldsymbol{P}^{\intercal} \mathbf{D}^{\intercal} \boldsymbol{\Sigma}^{-1} \mathbf{y}$. This allows us to simplify,

$$\begin{aligned}
\mathbb{P}(\mathbf{y} \mid \mathbf{D}, \Pi) &\propto \exp\left\{ -\frac{1}{2} \left(\mathbf{y}^{\intercal} \boldsymbol{\Sigma}^{-1} \mathbf{y} - (\boldsymbol{P}^{\intercal} \mathbf{D}^{\intercal} \boldsymbol{\Sigma}^{-1} \mathbf{y})^{\intercal} (\boldsymbol{P}^{\intercal} \mathbf{D}^{\intercal} \boldsymbol{\Sigma}^{-1} \mathbf{D} \boldsymbol{P})^{-1} (\boldsymbol{P}^{\intercal} \mathbf{D}^{\intercal} \boldsymbol{\Sigma}^{-1} \mathbf{y}) \right) \right\} \\
&\propto \exp\left\{ -\frac{1}{2} \mathbf{y}^{\intercal} \left(\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1} \mathbf{D} \boldsymbol{P} (\boldsymbol{P}^{\intercal} \mathbf{D}^{\intercal} \boldsymbol{\Sigma}^{-1} \mathbf{D} \boldsymbol{P})^{-1} \boldsymbol{P}^{\intercal} \mathbf{D}^{\intercal} \boldsymbol{\Sigma}^{-1} \right) \mathbf{y} \right\}
\end{aligned}$$

The likelihood is maximized when the log-likelihood is maximized,

$$\begin{aligned}
\log \mathbb{P}(\mathbf{y} \mid \mathbf{D}, \Pi) &= -\frac{1}{2}\mathbf{y}^\intercal \left(\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{D}\boldsymbol{P}(\boldsymbol{P}^\intercal\mathbf{D}^\intercal\boldsymbol{\Sigma}^{-1}\mathbf{D}\boldsymbol{P})^{-1}\boldsymbol{P}^\intercal\mathbf{D}^\intercal\boldsymbol{\Sigma}^{-1}\right)\mathbf{y} + \mathrm{c} \\
\frac{\partial \log \mathbb{P}(\mathbf{y} \mid \mathbf{D}, \Pi)}{\partial \mathbf{y}} &\overset{\text{set}}{=} 0 \\
\implies \mathbf{y} &= \mathbf{D}\boldsymbol{P}(\boldsymbol{P}^\intercal\mathbf{D}^\intercal\boldsymbol{\Sigma}^{-1}\mathbf{D}\boldsymbol{P})^{-1}\boldsymbol{P}^\intercal\mathbf{D}^\intercal\boldsymbol{\Sigma}^{-1}\mathbf{y} \\
&= \mathbf{D}\boldsymbol{P}\widehat{\boldsymbol{\gamma}} = \mathbf{D}\widehat{\boldsymbol{\beta}},
\end{aligned}$$

where $\widehat{\boldsymbol{\beta}} = \boldsymbol{P}\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\gamma}} = (\boldsymbol{P}^\intercal\mathbf{D}^\intercal\boldsymbol{\Sigma}^{-1}\mathbf{D}\boldsymbol{P})^{-1}\boldsymbol{P}^\intercal\mathbf{D}^\intercal\boldsymbol{\Sigma}^{-1}\mathbf{y}$. This is exactly the solution to is the solution to the following ordinary least-squares problem,

$$\min_{\boldsymbol{\gamma}} \|\mathbf{y} - \mathbf{D}\boldsymbol{P}\boldsymbol{\gamma}\|_2^2 .$$

Next, we will show that this ordinary least squares problem is identical to $\mathcal{L}(\Pi; \mathbf{Z})$. In order to make this argument cleaner, we will assume that $\boldsymbol{\Sigma} = \sigma^2\mathcal{I}_n$. Now, observe the structure of $\mathbf{D}$. In any row i, $D_{ik} = 1$ if observation i is assigned to feature combination k, and $D_{ik} = 0$ otherwise. So, $\mathbf{D}^\intercal\mathbf{D}$ is a diagonal matrix of size $K \times K$ where $(\mathbf{D}^\intercal\mathbf{D})_{kk}$ is the number of observations assigned to feature combination k, n_k. And $\mathbf{D}^\intercal\mathbf{y}$ sums all outcomes y_i corresponding to each feature combination k.

Similar to how $\mathbf{D}$ collects all observations into their respective feature combinations, $\boldsymbol{P}$ collects all feature combinations into their respective pools. Therefore $(\boldsymbol{P}^\intercal\mathbf{D}^\intercal\mathbf{D}\boldsymbol{P})^{-1}\boldsymbol{P}^\intercal\mathbf{D}^\intercal\mathbf{y}$ is the average outcome in each pool. This is exactly our estimated $\widehat{\mathbf{y}}$ in Equation B.6. In other words, $\mathcal{L}(\Pi; \mathbf{Z})$ is exactly the minimized squared error (up to some scaling constant).

Therefore, maximizing the posterior $\mathbb{P}(\mathbf{y} \mid \mathbf{D}, \Pi)\mathbb{P}(\Pi)$ corresponds to minimizing the mean-squared error with the ℓ_0 penalty. This has connections to loss-based generalized Bayesian inference (Bissiri et al., 2016). Here, we have described one possible prior on $\boldsymbol{\beta}$ to recover the mean-squared error. However, other such priors exist that describe such analytic loss functions. We refer the reader to Section 4 of Chipman et al. (2010) for examples of such priors used for BARTs.

B.3 Approximating the posterior

Proof of Theorem 3.3.1. By the triangle inequality, we can write

$$\begin{aligned}
\sup_{\boldsymbol{t}} \left| F_{\boldsymbol{\beta}|\mathbf{Z},\mathcal{P}_\theta}(\boldsymbol{t}) - F_{\boldsymbol{\beta}|\mathbf{Z}}(\boldsymbol{t}) \right| &= \sup_{\boldsymbol{t}} \left| \sum_{\Pi \in \mathcal{P}_\theta} F_{\boldsymbol{\beta}|\mathbf{Z}}(\boldsymbol{t} \mid \Pi) \frac{\mathbb{P}(\Pi \mid \mathbf{Z})}{\sum_{\Pi' \in \mathcal{P}_\theta} \mathbb{P}(\Pi' \mid \mathbf{Z})} - \sum_{\Pi \in \mathcal{P}^\star} F_{\boldsymbol{\beta}|\mathbf{Z}}(\boldsymbol{t} \mid \Pi) \mathbb{P}(\Pi \mid \mathbf{Z}) \right|, \\
&\leq (\mathrm{I}) + (\mathrm{II}) \\
\mathrm{I} &= \sup_{\boldsymbol{t}} \left| \sum_{\Pi \in \mathcal{P}_\theta} F_{\boldsymbol{\beta}|\mathbf{Z}}(\boldsymbol{t} \mid \Pi) \frac{\mathbb{P}(\Pi \mid \mathbf{Z})}{\sum_{\Pi' \in \mathcal{P}_\theta} \mathbb{P}(\Pi' \mid \mathbf{Z})} - \sum_{\Pi \in \mathcal{P}_\theta} F_{\boldsymbol{\beta}|\mathbf{Z}}(\boldsymbol{t} \mid \Pi) \mathbb{P}(\Pi \mid \mathbf{Z}) \right|, \\
\mathrm{II} &= \sup_{\boldsymbol{t}} \left| \sum_{\Pi \in \mathcal{P}_\theta} F_{\boldsymbol{\beta}|\mathbf{Z}}(\boldsymbol{t} \mid \Pi) \mathbb{P}(\Pi \mid \mathbf{Z}) - \sum_{\Pi \in \mathcal{P}^\star} F_{\boldsymbol{\beta}|\mathbf{Z}}(\boldsymbol{t} \mid \Pi) \mathbb{P}(\Pi \mid \mathbf{Z}) \right|
\end{aligned}$$

Let us denote $K = \sum_{\Pi' \in \mathcal{P}_\theta} \mathbb{P}(\Pi' \mid \mathbf{Z})$. Then the first term is,

$$\begin{aligned}
(\mathrm{I}) &= \sup_{\boldsymbol{t}} \left| \sum_{\Pi \in \mathcal{P}_\theta} F_{\boldsymbol{\beta}|\mathbf{Z}}(\boldsymbol{t} \mid \Pi) \frac{\mathbb{P}(\Pi \mid \mathbf{Z})}{K} - \sum_{\Pi \in \mathcal{P}_\theta} F_{\boldsymbol{\beta}|\mathbf{Z}}(\boldsymbol{t} \mid \Pi) \mathbb{P}(\Pi \mid \mathbf{Z}) \right| \\
&\leq \left| \frac{1}{K} - 1 \right| \sup_{\boldsymbol{t}} \left| \sum_{\Pi \in \mathcal{P}_\theta} F_{\boldsymbol{\beta}|\mathbf{Z}}(\boldsymbol{t} \mid \Pi) \mathbb{P}(\Pi \mid \mathbf{Z}) \right| \\
&\leq \left| \frac{1}{K} - 1 \right| \sup_{\boldsymbol{t}} \left| \sum_{\Pi \in \mathcal{P}_\theta} \mathbb{P}(\Pi \mid \mathbf{Z}) \right| \\
&\leq \left| \frac{1}{K} - 1 \right| = \frac{1}{K} - 1,
\end{aligned}$$

where in the third line, we trivially bound $F_{\boldsymbol{\beta}|\mathbf{Z}}(\boldsymbol{t} \mid \Pi) \leq 1$ as it is a distribution function and in the last time we bound $\sum_{\Pi \in \mathcal{P}_\theta} \mathbb{P}(\Pi \mid \mathbf{Z}) \leq 1$ since it is a probability mass function. Note that we were able to remove the absolute values because $K \leq 1$ giving us $1/K - 1 > 0$. Note that, by definition, $K \geq |\mathcal{P}_\theta|\, \theta$. Therefore,

$$(\mathrm{I}) \leq \frac{1}{|\mathcal{P}_\theta|\, \theta} - 1.$$

Moving on to the second term,

$$\begin{aligned}
(\mathrm{II}) &= \sup_{\boldsymbol{t}} \left| \sum_{\Pi \in \mathcal{P}_\theta} F_{\boldsymbol{\beta}|\mathbf{Z}}(\boldsymbol{t} \mid \Pi) \cdot \mathbb{P}(\Pi \mid \mathbf{Z}) - \sum_{\Pi \in \mathcal{P}^\star} F_{\boldsymbol{\beta}|\mathbf{Z}}(\boldsymbol{t} \mid \Pi) \cdot \mathbb{P}(\Pi \mid \mathbf{Z}) \right| \\
&= \sup_{\boldsymbol{t}} \left| \sum_{\Pi \in \mathcal{P}^\star \setminus \mathcal{P}_\theta} F_{\boldsymbol{\beta}|\mathbf{Z}}(\boldsymbol{t} \mid \Pi) \cdot \mathbb{P}(\Pi \mid \mathbf{Z}) \right| \\
&\leq \sup_{\boldsymbol{t}} \left| \sum_{\Pi \in \mathcal{P}^\star \setminus \mathcal{P}_\theta} 1 \cdot \mathbb{P}(\Pi \mid \mathbf{Z}) \right| \\
&= \sum_{\Pi \in \mathcal{P}^\star \setminus \mathcal{P}_\theta} \mathbb{P}(\Pi \mid \mathbf{Z}) \\
&\leq 1 - |\mathcal{P}_\theta|\, \theta,
\end{aligned}$$

where in the third line, we again bound $F_{\boldsymbol{\beta}|\mathbf{Z}}(\boldsymbol{t} \mid \Pi) \leq 1$, and in the final step, we use the definition of $\mathcal{P}_\theta$.

Therefore,

$$\sup_{\boldsymbol{t}} \left| F_{\boldsymbol{\beta}|\mathbf{Z},\mathcal{P}_\theta}(\boldsymbol{t}) - F_{\boldsymbol{\beta}|\mathbf{Z}}(\boldsymbol{t}) \right| \leq (\mathrm{I}) + (\mathrm{II}) \leq \frac{1}{|\mathcal{P}_\theta|\, \theta} - |\mathcal{P}_\theta|\, \theta.$$

□

Proof of Corollary 3.3.2. This argument is identical to Theorem 3.3.1 except for how we bound the expectations. We have

$$\begin{aligned}
\left\| \mathbb{E}_{\Pi|\mathcal{P}_\theta} \boldsymbol{\beta} - \mathbb{E}_{\Pi,\mathcal{P}_\theta} \boldsymbol{\beta} \right\| &= \left\| \sum_{\Pi \in \mathcal{P}_\theta} \boldsymbol{\beta}_\Pi \frac{\mathbb{P}(\Pi \mid \mathbf{Z})}{\sum_{\Pi' \in \mathcal{P}_\theta} \mathbb{P}(\Pi' \mid \mathbf{Z})} - \sum_{\Pi \in \mathcal{P}_\theta} \boldsymbol{\beta}_\Pi \mathbb{P}(\Pi \mid \mathbf{Z}) \right\| \\
&= \left| \frac{1}{\sum_{\Pi' \in \mathcal{P}_\theta} \mathbb{P}(\Pi' \mid \mathbf{Z})} - 1 \right| \left\| \sum_{\Pi \in \mathcal{P}_\theta} \boldsymbol{\beta}_\Pi \mathbb{P}(\Pi \mid \mathbf{Z}) \right\| \\
&= \left| \frac{1}{K} - 1 \right| \left\| \mathbb{E}_{\Pi,\mathcal{P}_\theta} \boldsymbol{\beta} \right\|,
\end{aligned}$$

where $K = \sum_{\Pi' \in \mathcal{P}_\theta} \mathbb{P}(\Pi' \mid \mathbf{Z})$. Note that by definition, $K \geq |\mathcal{P}_\theta|\, \theta$. Further, $K \leq 1$ gives us

$1/K - 1 > 0$. Therefore,

$$\begin{aligned} \left\|\mathbb{E}_{\Pi|\mathcal{P}_\theta}\boldsymbol{\beta} - \mathbb{E}_{\Pi,\mathcal{P}_\theta}\boldsymbol{\beta}\right\| &\leq \left(\frac{1}{|\mathcal{P}_\theta|\,\theta} - 1\right) \|\mathbb{E}_{\Pi,\mathcal{P}_\theta}\boldsymbol{\beta}\| \\ \implies \frac{\left\|\mathbb{E}_{\Pi|\mathcal{P}_\theta}\boldsymbol{\beta} - \mathbb{E}_{\Pi,\mathcal{P}_\theta}\boldsymbol{\beta}\right\|}{\|\mathbb{E}_{\Pi,\mathcal{P}_\theta}\boldsymbol{\beta}\|} &= \mathcal{O}\left(\frac{1}{|\mathcal{P}_\theta|\,\theta} - 1\right). \end{aligned}$$

If we assume that $\|\boldsymbol{\beta}_\Pi\| < \infty$, then define $C = \max_{\Pi\in\mathcal{P}^\star} \|\boldsymbol{\beta}_\Pi\| < \infty$. Then, we have

$$\left\|\mathbb{E}_{\Pi|\mathcal{P}_\theta}\boldsymbol{\beta} - \mathbb{E}_{\Pi,\mathcal{P}_\theta}\boldsymbol{\beta}\right\| = \mathcal{O}\left(\frac{1}{|\mathcal{P}_\theta|\,\theta} - 1\right).$$

Further,

$$\begin{aligned} \|\mathbb{E}_{\Pi,\mathcal{P}_\theta}\boldsymbol{\beta} - \mathbb{E}_\Pi\boldsymbol{\beta}\| &= \left\| \sum_{\Pi\in\mathcal{P}_\theta} \boldsymbol{\beta}_\Pi \mathbb{P}(\Pi \mid \mathbf{Z}) - \sum_{\Pi\in\mathcal{P}^\star} \boldsymbol{\beta}_\Pi \mathbb{P}(\Pi \mid \mathbf{Z}) \right\| \\ &= \left\| \sum_{\Pi\in\mathcal{P}^\star\setminus\mathcal{P}_\theta} \boldsymbol{\beta}_\Pi \mathbb{P}(\Pi \mid \mathbf{Z}) \right\| \\ &\leq C \sum_{\Pi\in\mathcal{P}^\star\setminus\mathcal{P}_\theta} \mathbb{P}(\Pi \mid \mathbf{Z}) \\ &= \mathcal{O}\left(1 - |\mathcal{P}_\theta|\,\theta\right), \end{aligned}$$

where in the last line we used the definition of Rashomon sets.. Therefore,

$$\begin{aligned} \left\|\mathbb{E}_{\Pi|\mathcal{P}_\theta}\boldsymbol{\beta} - \mathbb{E}_\Pi\boldsymbol{\beta}\right\| &\leq \left\|\mathbb{E}_{\Pi|\mathcal{P}_\theta}\boldsymbol{\beta} - \mathbb{E}_{\Pi,\mathcal{P}_\theta}\boldsymbol{\beta}\right\| + \|\mathbb{E}_{\Pi,\mathcal{P}_\theta}\boldsymbol{\beta} - \mathbb{E}_\Pi\boldsymbol{\beta}\| \\ &= \mathcal{O}\left(\frac{1}{|\mathcal{P}_\theta|\,\theta} - |\mathcal{P}_\theta|\,\theta\right). \end{aligned}$$

□

Proof of Theorem 3.3.3. For any prior $P \in \mathcal{Q}_{\mathcal{P}|h}$, we have,

$$\sup_{Q\in\mathcal{Q}_{\mathcal{P}|h}} \delta(\mathrm{P}_{P,\mathbf{Z}}, \mathrm{P}_{Q,\mathbf{Z}}) = \sup_{Q\in\mathcal{Q}_{\mathcal{P}|h}} \sup_{\Pi\in\mathcal{P}_{|h}} |\mathrm{P}_{P,\mathbf{Z}}(\Pi) - \mathrm{P}_{Q,\mathbf{Z}}(\Pi)|$$

Table B.1: Notation used in Theorem 3.3.3.

Notation	Definition
$\mathcal{P}_{\|h}$	Set of permissible partitions with h pools
$Q \in \mathcal{Q}$	Prior over all $\boldsymbol{\beta}$
$Q \in \mathcal{Q}_{\|h}$	Prior over $\boldsymbol{\beta}$ such that there is some partition $\Pi_{\boldsymbol{\beta}} \in \mathcal{P}_{\|h}$
$Q \in \mathcal{Q}_{\mathcal{P}\|h}$	Prior for partitions $\Pi \in \mathcal{P}_{\|h}$
P_{ℓ_0}	The uniform prior over $\mathcal{P}_{\|h}$ (induced by ℓ_0 over $\mathcal{P}^\star$)
$\mathrm{P}_{Q,\mathbf{z}}$	Posterior density (over partitions or $\boldsymbol{\beta}$) with prior Q
$\delta(P, Q)$	Total variation distance between P and Q

$$
\begin{aligned}
&= \sup_{\Pi \in \mathcal{P}_{|h}} \sup_{Q \in \mathcal{Q}_{\mathcal{P}|h}} \left| \mathrm{P}_{P,\mathbf{z}}(\Pi) - \mathrm{P}_{Q,\mathbf{z}}(\Pi) \right| \\
&= \frac{1}{\mathbb{P}(\mathbf{y} \mid \mathbf{X})} \sup_{\Pi \in \mathcal{P}_{|h}} \mathbb{P}(\mathbf{y} \mid \mathbf{X}, \Pi) \sup_{Q \in \mathcal{Q}_{\mathcal{P}|h}} \left| P(\Pi) - Q(\Pi) \right|.
\end{aligned}
$$

First, consider the ℓ_0 prior.

$$
\sup_{Q \in \mathcal{Q}_{\mathcal{P}|h}} \delta(\mathrm{P}_{P_{\ell_0},\mathbf{z}}, \mathrm{P}_{Q,\mathbf{z}}) = \frac{1}{\mathbb{P}(\mathbf{y} \mid \mathbf{X})} \sup_{\Pi \in \mathcal{P}_{|h}} \mathbb{P}(\mathbf{y} \mid \mathbf{X}, \Pi) \sup_{Q \in \mathcal{Q}_{\mathcal{P}|h}} \left| \frac{1}{N(h)} - Q(\Pi) \right|.
$$

Choose an adversarial prior $Q^\star$ such that $Q^\star(\Pi^\star) = 1$ for some arbitrary $\Pi^\star \in \mathcal{P}_{|h}$. Then,

$$
\begin{aligned}
&\sup_{Q \in \mathcal{Q}_{\mathcal{P}|h}} \left| \frac{1}{N(h)} - Q(\Pi) \right| = \left| \frac{1}{N(h)} - Q^\star(\Pi^\star) \right| = 1 - \frac{1}{N(h)} \\
\implies &\sup_{Q \in \mathcal{Q}_{\mathcal{P}|h}} \delta(\mathrm{P}_{P_{\ell_0},\mathbf{z}}, \mathrm{P}_{Q,\mathbf{z}}) = \left(1 - \frac{1}{N(h)}\right) \frac{\sup_{\Pi \in \mathcal{P}_{|h}} \mathbb{P}(\mathbf{y} \mid \mathbf{X}, \Pi)}{\mathbb{P}(\mathbf{y} \mid \mathbf{X})}
\end{aligned}
$$

Next, consider any other prior $P \in \mathcal{Q}_{\mathcal{P}|h}$, $P \neq P_{\ell_0}$. Let $\Pi_m = \operatorname{argmin}_\Pi P(\Pi)$. Denote $P(\Pi_m) = p$. Observe that $p < 1/N(h)$ because $P \neq P_{\ell_0}$. Construct an adversarial prior $Q^\star$ such that $Q^\star(\Pi_m) = 1$. Therefore,

$$
\begin{aligned}
&\sup_{Q \in \mathcal{Q}_{\mathcal{P}|h}} |P(\Pi) - Q(\Pi)| = |P(\Pi_m) - Q^\star(\Pi_m)| = 1 - p \\
\implies &\sup_{Q \in \mathcal{Q}_{\mathcal{P}|h}} \delta(\mathrm{P}_{P,\mathbf{z}}, \mathrm{P}_{Q,\mathbf{z}}) = \frac{1}{\mathbb{P}(\mathbf{y} \mid \mathbf{X})} \sup_{\Pi \in \mathcal{P}_{|h}} \mathbb{P}(\mathbf{y} \mid \mathbf{X}, \Pi)(1 - p)
\end{aligned}
$$

$$= (1-p)\frac{\sup_{\Pi\in\mathcal{P}_{|h}} \mathbb{P}(\mathbf{y} \mid \mathbf{X}, \Pi)}{\mathbb{P}(\mathbf{y} \mid \mathbf{X})}$$
$$> \sup_{Q\in\mathcal{Q}_{\mathcal{P}|h}} \delta(\mathrm{P}_{P_{\ell_0},\mathbf{Z}}, \mathrm{P}_{Q,\mathbf{Z}}).$$

Thus, the ℓ_0 prior is minimax optimal,

$$\sup_{Q\in\mathcal{Q}_{\mathcal{P}|h}} \delta(\mathrm{P}_{P_{\ell_0},\mathbf{Z}}, \mathrm{P}_{Q,\mathbf{Z}}) = \inf_{P\in\mathcal{Q}_{\mathcal{P}|h}} \sup_{Q\in\mathcal{Q}_{\mathcal{P}|h}} \delta(\mathrm{P}_{P,\mathbf{Z}}, \mathrm{P}_{Q,\mathbf{Z}}).$$

□

Proof of Proposition 3.3.4. We can easily see that

$$\begin{aligned} & \xi(\Pi, \Pi_0) \leq \epsilon \\ \iff & \exp(-Q(\Pi)) \geq \exp\{-Q(\Pi_0)(1+\epsilon)\} = \exp\{-q_0(1+\epsilon)\} \\ \iff & \mathbb{P}(\Pi \mid \mathbf{Z}) \geq \frac{e^{-q_0(1+\epsilon)}}{c}, \end{aligned}$$

where $c := c(\mathbf{Z})$ is the normalization constant. □

B.4 Appendix to Size of the Rashomon Set

Proof of Proposition 3.5.1. To count the number of all possible partitions, we cast this as a decision tree problem. There are $(R-1)^m$ possible treatment policies in the profile with all arms turned on. These constitute possible nodes in a binary decision tree. The leaves in the decision tree are the pools. The number of binary trees with n nodes is given by

$$C_n = \frac{1}{n+1}\binom{2n}{n},$$

where C_n is the Catalan number (see *An Invitation to Analytic Combinatorics* from Flajolet and Sedgewick, 2009). Therefore, the number of trees we can construct (that may or may

not be admissible) is

$$T = \sum_{n=1}^{(R-1)^m} C_n = \mathcal{O}\left(2^{2(R-1)^m}\right),$$

where the big-O bound is given by Topley (2016).

To count the number of permissible partitions, conceptualize the binary matrix, $\mathbf{\Sigma} \in \{0,1\}^{m\times(R-2)}$ again. Each element of $\mathbf{\Sigma}$ tells us whether a particular pair of adjacent levels in a feature is pooled. In particular, we define $\Sigma_{ij} = 1$ if and only if feature combinations with dosage j are pooled with feature combinations with factor $j+1$ in feature i. Therefore, $\mathbf{\Sigma}$ enumerates all admissible partitions. This gives us the desired result.

□

Proof of Lemma 3.5.2. From the definition of the Rashomon set, if $\Pi \in \mathcal{P}_\theta$, then

$$\begin{aligned} & \mathbb{P}(\Pi \mid \mathbf{Z}) \geq \theta \\ \implies & \frac{\exp\{-\mathcal{L}(\Pi) - \lambda H(\Pi)\}}{c} \geq \theta \\ \implies & \exp\{-\lambda H(\Pi)\} \geq c\theta \\ \implies & H(\Pi) \leq -\frac{\ln(c\theta)}{\lambda}, \end{aligned}$$

which gives our desired result. □

Proof of Theorem 3.5.3. We know that the total number of pools in any partition is bounded by $H := H_\theta(\lambda)$.

Suppose there are k profiles. There is at least one pool in each profile. Therefore, the number of profiles is bounded by $1 \leq k \leq H$. The profiles can be generated by making a tree-like partition of the feature space. In each feature, there are $R-1$ places to split. Therefore, there are $M(R-1)$ places to split overall. To generate k profiles, we need to choose $k-1$ positions to split. This gives us $\binom{M(R-1)}{k-1}$ possibilities.

For a given set of k profiles, Lemma B.4.5 bounds the number of partitions as

$$\left|\mathcal{P}^{(k)}\right| = \begin{cases} \mathcal{O}\left(M^k R^{H-k}\right), & R > M^{c_{\text{crit}}} \\ \mathcal{O}\left((MR)^{k\log_2 H/k}(\log_2(MR))^{-1}\right), & \text{else} \end{cases},$$

where $c_{\text{crit}} = (\log_2 3 - 1)/(2 - \log_2 3)$.

Now, all that remains is to count across the number of profiles. This gives us,

$$\begin{aligned} |\mathcal{P}_\theta| &\leq \sum_{k=1}^{H} \binom{M(R-1)}{k-1} \left|\mathcal{P}^{(k)}\right| \\ &\leq C\sum_{k=1}^{H}(MR)^{k-1} \times \begin{cases} \mathcal{O}\left(M^k R^{H-k}\right), & R > M^{c_{\text{crit}}} \\ \mathcal{O}\left((MR)^{k\log_2 H/k}(\log_2(MR))^{-1}\right), & \text{else} \end{cases} \\ &\leq \begin{cases} \mathcal{O}\left(\sum_{k=1}^{H} M^{2k-1}R^{H-1}\right), & R > M^{c_{\text{crit}}} \\ \mathcal{O}\left(\sum_{k=1}^{H}(MR)^{k(1+\log_2 H/k)-1}(\log_2(MR))^{-1}\right), & \text{else} \end{cases} \end{aligned}$$

The first case simplifies as

$$R^{H-1}\sum_{k=1}^{H} M^{2k-1} = \mathcal{O}\left(M^{2H-1}R^{H-1}\right).$$

Next, look at the exponent on (MR) in the second case, $k(1+\log_2 H/k) - 1$. Using the second derivative test, we can see that this is maximized when $k = H$:

$$\begin{aligned} &\frac{\partial}{\partial k} : \log_2 H - \log_2 k \overset{\text{set}}{=} 0 \implies k = H \\ &\frac{\partial^2}{\partial k^2} : -\frac{1}{k} < 0. \end{aligned}$$

Therefore, the second case can be bounded as,

$$\sum_{k=1}^{H}(MR)^{k(1+\log_2 H/k)-1} \leq \sum_{k=1}^{H}(MR)^{H-1} = \mathcal{O}\left((MR)^{H-1}\right).$$

Therefore, we have

$$|\mathcal{P}_\theta| = \begin{cases} \mathcal{O}\left(M^{2H-1}R^{H-1}\right), & R > M^{c_{\text{crit}}} \\ \mathcal{O}\left((MR)^{H-1}(\log_2(MR))^{-1}\right), & \text{else} \end{cases},$$

where $c_{\text{crit}} = (\log_2 3 - 1)/(2 - \log_2 3)$.

□

B.4.1 Helpful results

We state a useful result that helps us count the number of pools generated by a partition matrix $\mathbf{\Sigma}$.

Lemma B.4.1. *Let $\mathbf{\Sigma}$ be the partition matrix for a profile with m active features. Suppose there are z_i 1's in the i-th row of $\mathbf{\Sigma}$. Then the number of pools created by $\mathbf{\Sigma}$ is,*

$$H(\mathbf{\Sigma}) = (R-1)^m - (R-1)^{m-1}\sum_i z_i + (R-1)^{m-2}\sum_{i_1<i_2} z_{i_1}z_{i_2} - (R-1)^{m-3}\sum_{i_1<i_2<i_3} z_{i_1}z_{i_2}z_{i_3} + \cdots + (-1)^m z_1 \ldots z_m.$$

Proof of Lemma B.4.1. Observe that there are $(R-1)^m$ feature combinations in total ($R-1$ because we are assuming the R discrete values include the control). Suppose, we set $\mathbf{\Sigma}_{ij} = 1$, then we are pooling policies of type $[r_1, \ldots, r_{i-1}, j, r_{i+1}, \ldots, r_m]$ with $[r_1, \ldots, r_{i-1}, j-1, r_{i+1}, \ldots, r_m]$, where $r_{i'}$ can take on $R-1$ values. Therefore, $(R-1)^{m-1}$ policies are pooled. So, if there are in $\text{nnz}(\mathbf{\Sigma}) = \sum_i z_i$, then $(R-1)^{m-1}\sum_i z_i$ policies are pooled.

However, if some of those 1's are in a different treatment arm, then we end up double

counting those. For example, if $\mathbf{\Sigma}_{i_1,j} = 1$ and $\mathbf{\Sigma}_{i_2,j'} = 1$, then we remove policies of type $[r_1, \ldots, j, \ldots, j', \ldots, r_m]$ *twice* where j and j' are at indices i_1 and i_2. So, we need to add them back. Similarly, the remaining non-linear terms account for this "double counting." $\square$

Lemma B.4.1 tells us how to count the number of pools given a partition matrix. We now state another result that bounds the sparsity of the partition matrix given some number of pools in Lemma B.4.2.

Lemma B.4.2. *Let $\mathbf{\Sigma}$ be the matrix defined in Proposition 3.5.1 for a profile with m active features. Suppose there are H pools. Then,*

$$\sum_i z_i \leq \frac{(2R-3)^m + 1 - 2H}{2(R-1)^{m-1}}$$

Proof of Lemma B.4.2. Rearranging and dropping negative terms from Lemma B.4.1,

$$\begin{aligned}
(R-1)^{m-1} \sum_i z_i &\leq -H + (R-1)^m + (R-1)^{m-2} \sum_{i_1<i_2} z_{i_1} z_{i_2} \\
&\quad + (R-1)^{m-4} \sum_{i_1<\cdots<i_4} z_{i_1} z_{i_2} z_{i_3} z_{i_4} + \ldots \\
&\leq -H + (R-1)^m + (R-1)^{m-2}(R-2)^2 \sum_{i_1<i_2} 1 \\
&\quad + (R-1)^{m-4}(R-2)^4 \sum_{i_1<\cdots<i_4} 1 + \ldots \\
&= -H + \sum_{n \text{ even}} \binom{m}{n} (R-1)^{m-n}(R-2)^n \\
\implies \sum_i z_i &\leq \frac{(2R-3)^m + 1 - 2H}{2(R-1)^{m-1}},
\end{aligned}$$

where the second inequality uses $z_j \leq R-2$ and the last step uses the well-known identity

$$\sum_{k \text{ even}} \binom{n}{k} a^{n-k} b^k = \frac{1}{2}\left((a+b)^n + (a-b)^n\right).$$

□

We state a stronger result in Lemma B.4.3 that tells us exactly how many partition matrices could have generated a given number of pools. This result is crucial in bounding the size of the RPS in a practically meaningful way as in Theorem 3.5.3.

Lemma B.4.3. *Let $\mathbf{\Sigma}$ be the matrix defined in Proposition 3.5.1 for a profile with m active features. Then the number of $\mathbf{\Sigma}$ matrices that generate h pools is given by*

$$N(h) = \sum_{k=0}^{m} \binom{m}{k} \sum_{\prod_{i=1}^{k}(z_i+1)=h} \prod_{i=1}^{k} \binom{R-2}{z_i},$$

where we define $N(1) = 1$ and $\binom{n}{k} = 0$ for $k > n$.

As $m, R \to \infty$, we have

$$N(h) = \begin{cases} \mathcal{O}(mR^{h-1}), & R > m^{c_{\text{crit}}} \\ \mathcal{O}((mR)^{\log_2 h}), & \textit{else} \end{cases},$$

where $c_{\text{crit}} = (\log_2 3 - 1)/(2 - \log_2 3)$.

Proof of Lemma B.4.3. This is an exercise in counting. When we make z splits in one feature, we generate $z+1$ pools. When we make z_i splits in feature i and z_j splits in feature j, we generate $(z_i+1)(z_j+2)$ pools.

When we want to generate h pools, we first choose the features where we want the splits to occur. This is what the first summation is doing. Suppose that we've chosen k features where we want to perform splits. Next, we need to identify how many splits can be made in each feature. This is what the inner summation is doing with the condition $\prod_{i=1}^{k}(z_i+1) = h$. Finally, we need to identify where those splits are made, which is where the binomial coefficient comes in.

To get the asymptotic bound, we first consider the term where the exponent on R is the largest. This is when we choose all splits in the same feature. Next, we consider the term where the exponent on m is the largest. For this to happen, we need to choose as many arms as possible i.e., make the smallest number of non-zero splits in each feature. This corresponds to making 1 split in each feature i.e., selecting $\log_2 h$ feature. Hence, we get the asymptotic bound $\mathcal{O}(\max\{mR^{h-1}, (mR)^{\log_2 h}\})$.

Observe that

$$mR^{h-1} > (mR)^{\log_2 h} \iff R > m^{\frac{\log_2 h - 1}{h - \log_2 h - 1}}.$$

The exponent on m is a decreasing function in h. When $h = 2$, $mR^{h-1} = (mR)^{\log_2 h}$. And when $h = 3$, the exponent is $c_{\text{crit}} = (\log_2 3 - 1)/(2 - \log_2 3)$. Therefore $mR^{h-1} > (mR)^{\log_2 h}$ whenever $R > m^{c_{\text{crit}}}$, which gives our desired result. □

Lemma B.4.3 has a nice implication. When h is a prime number, we expect $N(h)$ to be small because all of the splits need to be made in the same feature. On the other hand, when $h = 2^k$ is a power-of-two, we expect $N(h)$ to be very large since we can make splits in multiple features at the same time.

Lemma B.4.4. *For $a > 1$,*

$$\int_x a^{\log_2 x} dx = \frac{x a^{\log_2 x}}{1 + \log_2 a} + C.$$

Proof of Lemma B.4.4. We use integration by parts to solve this,

$$\begin{aligned}
\int_x a^{\log_2 x} dx &= a^{\log_2 x} \int_x dx - \int_x x \cdot \frac{a^{\log_2 x} \log_2 a}{x} dx \\
&= x a^{\log_2 x} - \log_2 a \int_x a^{\log_2 x} dx \\
\implies \int_x a^{\log_2 x} dx &= \frac{x a^{\log_2 x}}{1 + \log_2 a} + C.
\end{aligned}$$

□

Lemma B.4.5. *Suppose there are $k \geq 1$ fixed profiles across M features each taking on R discrete ordered values. Suppose that the maximum number of pools in any partition is H. Then, the number of permissible partitions is bounded by,*

$$\left|\mathcal{P}^{(k)}\right| = \begin{cases} \mathcal{O}\left(M^k R^{H-k}\right), & R > M^{c_{\text{crit}}} \\ \mathcal{O}\left((MR)^{k \log_2 H/k} (\log_2(MR))^{-1}\right), & \text{else} \end{cases},$$

where $c_{\text{crit}} = (\log_2 3 - 1)/(2 - \log_2 3)$.

Proof of Lemma B.4.5. Let h_i denote the number of pools within each profile. Then, we know that $k \leq \sum_{i=1}^{k} h_i \leq h$ where $1 \leq h_i \geq h - k + 1$ for every profile i and $h \leq H$. Observe that partitions within each profile are strongly convex. By Lemma B.4.3, we have a bound on number of partitions of size h_i, $N_i(h_i)$,

$$N_i(h_i) = \max\left\{\mathcal{O}(MR^{h_i - 1}), \mathcal{O}((MR)^{\log_2 h_i})\right\}.$$

We also know from Lemma B.4.3 that MR^{h_i-1} beats $(MR)^{\log_2 h_i}$ whenever $R > m^{c_{\text{crit}}}$ where $c_{\text{crit}} = (\log_2 3 - 1)/(2 - \log_2 3)$. For now, we will suppress this condition for readability. We will re-introduce this at the end.

For a given set of k profiles, the number of partitions is given by,

$$\left|\mathcal{P}^{(k)}\right| = \sum_{h=k}^{H} \prod_{\sum_{i=1}^{k} h_i = h} N_i(h_i).$$

There are $\binom{h-1}{k-1}$ positive integral solutions to the equation $\sum_{i=1}^{k} h_i = h$. Thus,

$$\left|\mathcal{P}^{(k)}\right| = \sum_{h=k}^{H} \prod_{\sum_{i=1}^{k} h_i = h} N_i(h_i)$$

$$= \sum_{h=k}^{H} \binom{h-1}{k-1} \max\left\{\mathcal{O}\left(M^k R^{\sum_{i=1}^k h_i - 1}\right), \mathcal{O}\left((MR)^{\sum_{i=1}^k \log_2 h_i}\right)\right\}$$
$$= \sum_{h=k}^{H} \binom{h-1}{k-1} \max\left\{\mathcal{O}\left(M^k R^{h-k}\right), \mathcal{O}\left((MR)^{\sum_{i=1}^k \log_2 h_i}\right)\right\}$$

To bound $\sum_{i=1}^k \log_2 h_i$ when $\sum_{i=1}^k h_i = h$, we use the arithmetic-geometric mean inequality,

$$\left(\prod_{i=1}^k h_i\right)^{1/k} \leq \frac{\sum_{i=1}^k h_i}{k} \implies \prod_{i=1}^k h_i \leq \left(\frac{h}{k}\right)^k$$
$$\implies \sum_{i=1}^k \log_2 h_i \leq k(\log_2 h - \log_2 k).$$

Therefore,

$$\left|\mathcal{P}^{(k)}\right| = \sum_{h=k}^{H} \binom{h-1}{k-1} \max\left\{\mathcal{O}\left(M^k R^{h-k}\right), \mathcal{O}\left((MR)^{k(\log_2 h/k)}\right)\right\}$$
$$= \max\left\{\mathcal{O}\left(M^k \sum_{h=k}^{H} R^{h-k}\right), \mathcal{O}\left(\sum_{h=k}^{H} (MR)^{k(\log_2 h/k)}\right)\right\}$$

The first term simplifies as

$$M^k \sum_{h=k}^{H} R^{h-k} = \mathcal{O}(M^k R^{H-k}).$$

The second term can be bounded by the integral,

$$\sum_{h=k}^{H} (MR)^{k(\log_2 h/k)} \leq \int_{h=k}^{H} (MR)^{k(\log_2 h/k)} dh = \frac{H(MR)^{k\log_2 H/k} - k}{1 + k\log_2(MR)}$$
$$= \mathcal{O}\left((MR)^{k\log_2 H/k}(\log_2(MR))^{-1}\right)$$

where the integral is evaluated in Lemma B.4.4.

Thus, for a given set of k profiles, the number of partitions is bounded by,

$$\left|\mathcal{P}^{(k)}\right| = \max\left\{\mathcal{O}\left(M^k R^{H-k}\right), \mathcal{O}\left((MR)^{k\log_2 H/k}(\log_2(MR))^{-1}\right)\right\}.$$

Re-introducing the condition for when the first term beats the second gives us the desired result. □

B.5 Appendix to Rashomon set enumeration and Generalizations

We organize this appendix into proofs for results in Section 3.6, additional algorithms used in Section 3.6, and proofs for results in Section B.7.

B.5.1 Proofs in Section 3.6

Proof of Theorem 3.6.1. By definition,

$$b(\mathbf{\Sigma}, \mathcal{M}; \mathbf{Z}) \leq \frac{1}{n}\sum_{\pi\in\Pi_\mathfrak{f}}\sum_{k(i)\in\pi}(y_i - \widehat{\mu}_\pi)^2 + \lambda H(\Pi, \mathcal{M})$$

Notice that $|\Pi| \geq H(\Pi, \mathcal{M})$. Further, by making more splits, we can only reduce the total mean-squared error incurred. Therefore,

$$\begin{aligned}
Q(\Pi; \mathbf{Z}) &= \mathcal{L}(\Pi; \mathbf{Z}) + \lambda|\Pi| \\
&= \frac{1}{n}\sum_{\pi\in\Pi}\sum_{k(i)\in\pi}(y_i - \widehat{\mu}_\pi)^2 + \lambda|\Pi| \\
&\geq \frac{1}{n}\sum_{\pi\in\Pi}\sum_{k(i)\in\pi}\mathbb{I}\left\{k(i)\cap\pi_\mathfrak{f} \neq \varnothing\right\}(y_i - \widehat{\mu}_\pi)^2 + \lambda|\Pi| \\
&\geq \frac{1}{n}\sum_{\pi\in\Pi_\mathfrak{f}}\sum_{k(i)\in\pi}\mathbb{I}\left\{k(i)\cap\pi_\mathfrak{f} \neq \varnothing\right\}(y_i - \widehat{\mu}_\pi)^2 + \lambda|\Pi| \\
&\geq \frac{1}{n}\sum_{\pi\in\Pi_\mathfrak{f}}\sum_{k(i)\in\pi}\mathbb{I}\left\{k(i)\cap\pi_\mathfrak{f} \neq \varnothing\right\}(y_i - \widehat{\mu}_\pi)^2 + \lambda H(\Pi, \mathcal{M}) \\
&= b(\mathbf{\Sigma}_\mathfrak{f}; \mathbf{Z}).
\end{aligned}$$

So if $b(\boldsymbol{\Sigma}, \mathcal{M}; \mathbf{Z}) > \theta_\epsilon$, then $\boldsymbol{\Sigma}$ is not in the Rashomon set. Now consider $\boldsymbol{\Sigma}' \in \text{child}(\boldsymbol{\Sigma}, \mathcal{M})$. Notice that the size of the fixed set of indices $\mathcal{M}'$ in any child of $\boldsymbol{\Sigma}$ increases (because there are fewer places to make further splits). With any further split we make in $\mathcal{M}$, the number of pools increases. Finally, the loss is non-negative. These together imply,

$$\begin{aligned} & b(\boldsymbol{\Sigma}', \mathcal{M}'; \mathbf{Z}) \geq b(\boldsymbol{\Sigma}, \mathcal{M}; \mathbf{Z}) \\ \implies & Q(\Pi(\boldsymbol{\Sigma}'); \mathbf{Z}) \geq b(\boldsymbol{\Sigma}', \mathcal{M}'; \mathbf{Z}) \geq b(\boldsymbol{\Sigma}, \mathcal{M}; \mathbf{Z}). \end{aligned}$$

Therefore, if $b(\boldsymbol{\Sigma}, \mathcal{M}; \mathbf{Z}) > \theta_\epsilon$, then $\boldsymbol{\Sigma}$ and all $\boldsymbol{\Sigma}' \in \text{child}(\boldsymbol{\Sigma}, \mathcal{M})$ are not in the Rashomon set. □

Proof of Theorem 3.6.2. By definition of b_{eq},

$$b_{eq}(\boldsymbol{\Sigma}, \mathcal{M}; \mathbf{Z}) \leq \frac{1}{n} \sum_{\pi \in \Pi} \sum_{k(i) \in \pi} \mathbb{I}\left\{k(i) \cap \pi_{\mathfrak{f}}^c \neq \varnothing\right\} (y_i - \widehat{\mu}_\pi)^2.$$

The idea in the inequality above is that any further split we make must obey the splits made at $\mathcal{M}$.

$$\begin{aligned} Q(\Pi; \mathbf{Z}) &= \mathcal{L}(\Pi; \mathbf{Z}) + \lambda |\Pi| \\ &= \frac{1}{n} \sum_{\pi \in \Pi} \sum_{k(i) \in \pi} \mathbb{I}\left\{k(i) \cap \pi_{\mathfrak{f}} \neq \varnothing\right\} (y_i - \widehat{\mu}_\pi)^2 + \lambda |\Pi| + \\ &\qquad \frac{1}{n} \sum_{\pi \in \Pi} \sum_{k(i) \in \pi} \mathbb{I}\left\{k(i) \cap \pi_{\mathfrak{f}}^c \neq \varnothing\right\} (y_i - \widehat{\mu}_\pi)^2 \\ &\geq b(\boldsymbol{\Sigma}, \mathcal{M}; \mathbf{Z}) + b_{eq}(\boldsymbol{\Sigma}, \mathcal{M}; \mathbf{Z}) \\ &= B(\boldsymbol{\Sigma}, \mathcal{M}; \mathbf{Z}). \end{aligned}$$

Therefore, if $B(\boldsymbol{\Sigma}, \mathcal{M}; \mathbf{Z}) > \theta_\epsilon$, then $Q(\Pi; \mathbf{Z}) > \theta_\epsilon$ and $\boldsymbol{\Sigma}' \in \text{child}(\boldsymbol{\Sigma}, \mathcal{M})$. Let $\Pi' := \Pi(\boldsymbol{\Sigma}')$.

Then,

$$
\begin{aligned}
Q(\Pi'; \mathbf{Z}) &= \mathcal{L}(\Pi'; Z) + \lambda \left|\Pi'\right| \\
&= \frac{1}{n} \sum_{\pi \in \Pi'} \sum_{k(i) \in \pi} \mathbb{I}\left\{k(i) \cap \pi_{\mathfrak{f}} \neq \varnothing\right\} (y_i - \widehat{\mu}_{\pi})^2 + \lambda \left|\Pi'\right| + \\
&\qquad \frac{1}{n} \sum_{\pi \in \Pi'} \sum_{k(i) \in \pi} \mathbb{I}\left\{k(i) \cap \pi_{\mathfrak{f}}^c \neq \varnothing\right\} (y_i - \widehat{\mu}_{\pi})^2 \\
&\geq b(\boldsymbol{\Sigma}, \mathcal{M}; \mathbf{Z}) + \frac{1}{n} \sum_{\pi \in \Pi'} \sum_{k(i) \in \pi} \mathbb{I}\left\{k(i) \cap \pi_{\mathfrak{f}}^c \neq \varnothing\right\} (y_i - \widehat{\mu}_{\pi})^2 \\
&\geq b(\boldsymbol{\Sigma}, \mathcal{M}; \mathbf{Z}) + b_{eq}(\boldsymbol{\Sigma}, \mathcal{M}; \mathbf{Z}) \\
&= B(\boldsymbol{\Sigma}, \mathcal{M}; \mathbf{Z}).
\end{aligned}
$$

In the steps above, we used the fact that making any split will increase the number of pools to say that $|\Pi'| \geq |\Pi|$. We also used the definition of b_{eq} and the idea of a minimum loss incurred by equivalent units in the final step.

Therefore, if $B(\boldsymbol{\Sigma}, \mathcal{M}; \mathbf{Z}) > \theta_\epsilon$, then $Q(\Pi'; \mathbf{Z}) > \theta_\epsilon$ for any $\boldsymbol{\Sigma}' \in \text{child}(\boldsymbol{\Sigma}, \mathcal{M})$. So $\boldsymbol{\Sigma}$ and all such $\boldsymbol{\Sigma}'$ are not in the Rashomon set. □

Proof of Theorem 3.6.3. First note that Algorithm 9 correctly enumerates the Rashomon set for any given profile. This follows directly from Lemma 3.5.2, and Theorems 3.6.1 and 3.6.2.

Next, Algorithm 12 performs a breadth-first search starting at the control profile. Since the M-d hypercube has a unique source (the control profile) and sink (the profile with all features active), the breadth-first search will terminate after a finite time and traverse every possible path in the hypercube. When traversing an edge in the hypercube, Algorithm 11 attempts to pool adjacent profiles using the intersection matrix $\boldsymbol{\Sigma}^{\cap}$ while obeying (1) and (2) of Definition B.1.8. This pooling attempt is done recursively guaranteeing that all permissible partitions are considered for the Rashomon set.

The choice of Rashomon thresholds for each profile, described in Line 5, is justified by

the usage of Theorem 3.6.2.

Correctness of Algorithm 3 immediately follows. □

B.5.2 Additional algorithms

Algorithm 3 calls upon two important algorithms and uses a specific caching object that we describe here. First, Algorithm 9 describes how to enumerate the Rashomon partitions for a single profile. The choice of starting position in Algorithm 9 is arbitrary. No matter with which feature we start our search, Algorithm 9 will eventually explore the feature space sufficiently to identify partitions outside the Rashomon set, at which point we abandon that search. Theorems 3.6.1 and 3.6.2 guarantee this. Algorithm 9 will correctly enumerate the RPS independently of our starting search position.[5]

Second, we describe how to pool across profiles as defined by Definition B.1.8. The key insight here is the construction of the intersection matrix $\mathbf{\Sigma}^{\cap}$ we discussed earlier. Algorithm 12 describes a breadth-first search to enumerate partitions across different profiles by traversing the M-d hypercube. This algorithm in turn relies on Algorithm 10 to obtain the intersection matrix between partitions of adjacent profiles and Algorithm 11 to pool adjacent profiles recursively. Since Algorithm 12 is a breadth-first search, it can also be parallelized.

Finally, Algorithm 13 describes the implementation of the caching object used in Algorithm 3.

B.6 Appendix to simulations

B.6.1 Simulation 1: An example in medicine

Imagine a setting where a pharmacist is interested in finding the best treatment as a combination of two drugs, A and B. In particular, we consider a scenario where each drug needs

[5]Obviously, some starting positions may be computationally favorable i.e., we do not need to search for too long before we encounter low posterior partitions. We believe domain experts will have a better understanding of the context and may be able to choose a starting location that can reduce computation costs.

Algorithm 9 `EnumerateRPS_profile`$(M, R, H, \mathbf{Z}, \theta_\epsilon)$

Input: M features, R factors per feature, max pools H, data $\mathbf{Z}$, Rashomon threshold θ_ϵ
Output: Rashomon set $\mathcal{P}_{q,\epsilon}$

$\mathcal{P}_{q,\epsilon} = \varnothing$
$\mathcal{S} =$ `cache()` ▷ See Algorithm 13
$\mathcal{Q} =$ `queue()`
$\mathbf{\Sigma} = \{1\}^{M \times (R-2)}$
$\mathcal{Q}$.`push`$(\mathbf{\Sigma}, 1, 1)$ ▷ Can start at any arbitrary arm
while $\mathcal{Q}$ is not empty **do**
    $(\mathbf{\Sigma}, i, j) = \mathcal{Q}$.`dequeue()`
    if $\mathcal{S}$.`seen`$(\mathbf{\Sigma}, i, j)$ **then** continue
    $\mathcal{S}$.`insert`$((\mathbf{\Sigma}, i, j))$
    if $H(\mathbf{\Sigma}) > H$ **then** continue
    $\mathbf{\Sigma}_1 = \mathbf{\Sigma}$, $\mathbf{\Sigma}_{1,i,j} = 1$
    $\mathbf{\Sigma}_0 = \mathbf{\Sigma}$, $\mathbf{\Sigma}_{0,i,j} = 1$
    for $m \leq M$ **do** ▷ Branch and search
        $j_1 = \min\{j \leq R-2 \mid$ not $\mathcal{S}$.`seen`$(\mathbf{\Sigma}_1, m, j)\}$
        if $j_1 \neq \varnothing$ **then** $\mathcal{Q}$.`enqueue`$(\mathbf{\Sigma}_1, m, j_1)$
        $j_0 = \min\{j \leq R-2 \mid$ not $\mathcal{S}$.`seen`$(\mathbf{\Sigma}_0, m, j)\}$
        if $j_0 \neq \varnothing$ **then** $\mathcal{Q}$.`enqueue`$(\mathbf{\Sigma}_0, m, j_1)$
    if $B(\mathbf{\Sigma}, i, j; \mathbf{Z}) > \theta_\epsilon$ **then** continue
    if $Q(\mathbf{\Sigma}_1) \leq \theta_\epsilon$ **then** $\mathcal{P}_{q,\epsilon}.add(\mathbf{\Sigma}_1)$
    if $Q(\mathbf{\Sigma}_0) \leq \theta_\epsilon$ and $H(\mathbf{\Sigma}_0) \leq H$ **then** $\mathcal{P}_{q,\epsilon}.add(\mathbf{\Sigma}_0)$
    if $j < R-2$ **then** ▷ Search deeper
        if not $\mathcal{S}$.`seen`$(\mathbf{\Sigma}_1, i, j+1)$ **then** $\mathcal{Q}$.`enqueue`$(\mathbf{\Sigma}_1, i, j+1)$
        if not $\mathcal{S}$.`seen`$(\mathbf{\Sigma}_0, i, j+1)$ **then** $\mathcal{Q}$.`enqueue`$(\mathbf{\Sigma}_0, i, j+1)$
return $\mathcal{P}_{q,\epsilon}$

a minimum dosage to be effective. However, when the dosages become too strong, the interaction between the drugs results in a reduced treatment effect.[6]

In this setup, suppose that each drug has 4 possible non-zero dosages $d \in \{1, 2, 3, 4\}$. We will denote each of $4^2 = 16$ unique drug cocktails by their dosage (d_A, d_B). The treatment

[6]Such *antagonistic interactions* where one drug impedes the effect of another drug are not uncommon (see Cascorbi, 2012; Triplitt, 2006; Wambaugh et al., 2020). For example, ACE inhibitors (Angiotensin-converting enzyme inhibitors), used in treating high blood pressure, and Metformin, used in treating diabetes, are known to have negative interactions (Blackburn and Wilson, 2006); ACE inhibitors and NSAIDs (Non-steroidal anti-inflammatory drugs), and Aspirin and Ibuprofen have reduced effects when taken together (Cascorbi, 2012), and; effects of drug interactions in non-small cell lung cancer are highly context-specific (Nair et al., 2023).

Algorithm 10 `IntersectionMatrix`(Π, ρ_i, ρ_j)

Input: Partition Π, Adjacent profiles ρ_i, ρ_j such that $\rho_i < \rho_j$
Output: Intersection matrix $\mathbf{\Sigma}^{\cap}$

$\mathbf{m} = \rho_i \wedge \rho_j$ ▷ Indices of features active in both profiles
$m' = \rho_i \oplus \rho_j$ ▷ Index where ρ_i, ρ_j differ
$\Pi_{\rho_i} = \{\pi \setminus \{k \mid \rho(k) \neq \rho_i\} \mid \pi \in \Pi\}$
$\Pi_{\rho_j} = \{\pi \setminus \{k \mid \rho(k) \neq \rho_j\} \mid \pi \in \Pi\}$
$\mathbf{\Sigma}^{\cap} = [\infty]^{|\Pi_{\rho_i}| \times |\Pi_{\rho_j}|}$
for $\pi_k \in \Pi_{\rho_i}$ **do**
  for $\pi_{k'} \in \Pi_{\rho_j}$ **do** ▷ Features with lowest non-zero level in m'
    $\mathcal{A} = \sum_{a_1 \in \pi_k} \sum_{a_2 \in \pi_{k'}} \mathbb{I}\{\|\mathbf{x}(a_1) - \mathbf{x}(a_2)\|_1 = 1\}$
    if $\mathcal{A} \neq \varnothing$ **then**
      $\mathbf{\Sigma}^{\cap}_{k,k'} = 0$
return $\mathbf{\Sigma}^{\cap}$

Algorithm 11 `PoolAdjacentProfiles`$(\mathcal{P}_{q,\epsilon}, \Pi, \mathbf{z}, \mathbf{\Sigma}^{\cap}, \mathbf{Z}, \theta)$

Input: Rashomon set $\mathcal{P}_{q,\epsilon}$, partition Π, list of pools that can be pooled across profiles $\mathbf{z}$, data $\mathbf{Z}$, Rashomon threshold θ, intersection matrices already seen $\mathcal{S}$
Output: Rashomon set $\mathcal{P}_{q,\epsilon}$

while $\mathbf{z} \neq \varnothing$ **do**
  $(k, k') = \mathbf{z}.\texttt{pop}()$
  $\mathcal{P}_{q,\epsilon} = \texttt{PoolAdjacentProfiles}(\mathcal{P}_{q,\epsilon}, \Pi, \mathbf{z}, \mathbf{\Sigma}^{\cap}, \theta)$
  $\mathbf{\Sigma}^{\cap,\prime} = \mathbf{\Sigma}^{\cap}$
  $\mathbf{\Sigma}^{\cap,\prime}_{k,k'} = 1$
  $\mathbf{\Sigma}^{\cap,\prime}_{k,-k'} = \infty$, $\mathbf{\Sigma}^{\cap,\prime}_{-k,k'} = \infty$ ▷ Cannot pool π_k or $\pi_{k'}$ with any other pool
  $\Pi' = (\Pi \setminus \{\pi_k, \pi_{k'}\}) \cup (\pi_k \cup \pi_{k'})$ ▷ Update Π
  if $Q(\Pi'; \mathbf{Z}) \leq \theta$ **then**
    $\mathcal{P}_{q,\epsilon} = \Pi' \cup \texttt{PoolAdjacentProfiles}(\mathcal{P}_{q,\epsilon}, \Pi', \mathbf{z}, \mathbf{\Sigma}^{\cap,\prime}, \theta)$
return $\mathcal{P}_{q,\epsilon}$

is not effective when $d_A < 4$ and $d_B = 1$. As we increase the dosages of the drugs, they become more effective. The treatment is most effective when drug A is maxed and drug B has a medium dosage i.e., $d_A = 4$, $d_B = 2, 3$. When both drugs are maxed, a drug interaction produces a sub-optimal effect. Our parameters and partition are summarized in Figure B.5. Observe that, by design, all non-zero marginal increases in outcomes as we increase dosage

Algorithm 12 `PoolProfiles`$(\mathcal{P}, \rho_0, \mathbf{Z}, \theta)$

Input: Candidates for Rashomon set $\mathcal{P}$, control profile ρ_0, data $\mathbf{Z}$, Rashomon threshold θ
Output: Rashomon set $\mathcal{P}_{q,\epsilon}$

$\mathcal{P}_{q,\epsilon} = \varnothing$
$\mathcal{Q} =$ `queue()`
while $\mathcal{Q} \neq \varnothing$ **do**
    $\rho_i = \mathcal{Q}$.`dequeue()`
    $\mathcal{N}(\rho_i) = \{\rho_j \mid \|\rho_i - \rho_j\|_0 = 1, \rho_j > \rho_i\}$ ▷ Neighbors of ρ_i with additional active feature
    for $\rho_j \in \mathcal{N}(\rho_i)$ **do**
        $\mathcal{Q}$.`enqueue`(ρ_j)
        for $\Pi \in \mathcal{P}'$ **do**
            $\mathbf{\Sigma}^{\cap} =$ `IntersectionMatrix`(Π, ρ_i, ρ_j) ▷ See Algorithm 10
            $\mathbf{z} = \{(k, k') \mid \mathbf{\Sigma}^{\cap}_{k,k'} = 0\}$
            $\mathcal{P}_{q,\epsilon} =$ `PoolAdjacentProfiles`$(\mathcal{P}_{q,\epsilon}, \Pi, \mathbf{z}, \mathbf{\Sigma}^{\cap}, \mathbf{Z}, \theta)$ ▷ See Algorithm 11
return $\mathcal{P}_{q,\epsilon}$

Algorithm 13 Implementation of caching object used in Algorithm 9

$\mathcal{S} =$ `cache()` ▷ Initialize caching object
    $C = \{\}$
$\mathcal{S}$.`insert`$(\mathbf{\Sigma}, i, j)$ ▷ Extract and insert $\mathbf{\Sigma}_{\mathfrak{f}}$
    $\mathbf{\Sigma}[i, j : (R-2)] =$ `NA`
    $C = C \cup \{\mathbf{\Sigma}\}$
$\mathcal{S}$.`seen`$(\mathbf{\Sigma}, i, j)$ ▷ Extract and check presence of $\mathbf{\Sigma}_{\mathfrak{f}}$
    $\mathbf{\Sigma}[i, j : (R-2)] =$ `NA`
    return $\mathbf{\Sigma} \in C$

levels are correlated. Therefore, we expect Lasso to perform poorly as it penalizes selecting correlated features. However, the ℓ_0 penalty does not presume any such false independence assumptions.

In each dataset, we fixed the number of samples per feature combination to n_k and outcomes for each feature were drawn from a $\mathcal{N}(\beta_i, 1)$ distribution. We varied $n_k \in \{10, 100, 1000, 5000\}$ and for each n_k simulated $r = 100$ datasets. For each dataset, we fit the Lasso model and found the RPS. The Rashomon threshold was chosen to be $1.5 \times \text{MSE}_{\text{Lasso}}$

Algorithm 14 `select_feasible_combinations`(K, θ)

Input: K list of n sorted lists containing a numerical score, θ threshold
Output: F, list of lists of length n with indices of elements from each of K_i such that their sum is less than θ

$F = \{\}$
$n = \texttt{len}(K)$
if $n = 0$ **then return** $\{\}$
$K_{1,\text{feasible indices}} = \{i \mid K_1[i] \leq \theta\}$
if $n = 1$ **then**
    $F = \{K_{1,\text{feasible indices}}\}$
    return F
$x = \sum_{j=2}^{n} K_j[1]$
for $i \in K_{1,\text{feasible indices}}$ **do**
    $\theta_i = \theta - K_1[i]$
    if $\theta_i < x$ **then break**
    $F_i = \texttt{select\_feasible\_combinations}(K[2:], \theta_i)$
    for $f \in F_i$ **do**
        $F.\texttt{insert}([i].\texttt{append}(f))$
    return F

where $\text{MSE}_{\text{Lasso}}$ is the MSE of the Lasso model with $n_k = 10$. We use the same regularization parameter $\lambda = 0.1$ for the Rashomon and Lasso models.

The results from the simulations are presented in Figure B.6. The metrics reported here are described in Appendix B.6. It is clear that the "average" model in the RPS not only performs much better than Lasso in terms of overall MSE, but it is also able to recover the true best policy set. By looking only for the optimal model, Lasso consistently misses out on coverage for the best policies by incorrectly selecting features. However, when we look at the RPS of near-optimal models, we can recover the full best policy coverage almost always. Our results also reveal that the poor performance of Lasso is not a sample size issue. Lasso is simply not the right tool in situations with correlated parameters. Appendix B.6 also visualizes the the Rashomon set through a 2D heatmap.

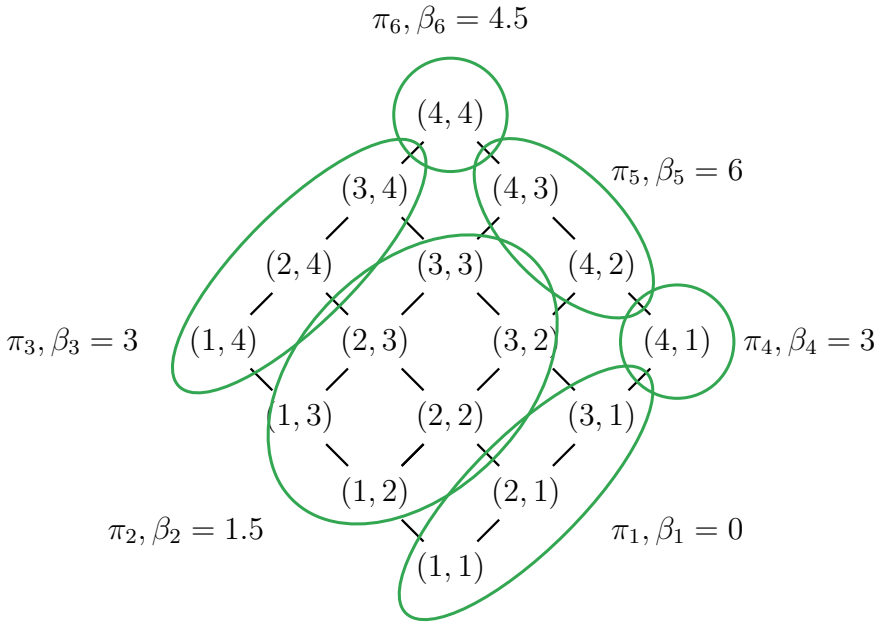


Figure B.5: Hasse diagram illustrating partition used in Simulation 1.

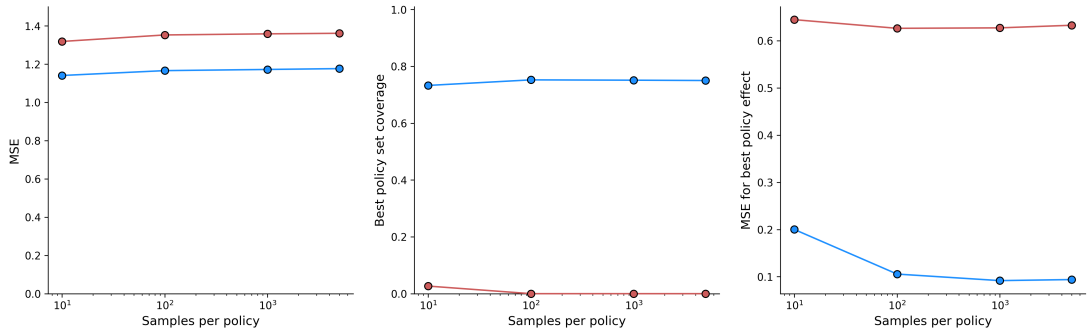


Figure B.6: Results for Simulation 1. The blue points correspond to models in the Rashomon set and the red points correspond to Lasso estimates. From left to right: mean squared error, best policy set coverage and best policy mean squared error.

B.6.2 Performance metrics for Simulation 1

We used the following performance metrics to evaluate Lasso and models in the RPS in Figure B.6 in Section 3.7:

1. Overall mean-squared error (MSE): Suppose $\widehat{y}_i$ and y_i are the estimated and true outcomes for unit i, then the overall MSE is defined as

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\widehat{y}_i - y_i)^2.$$

2. Best policy MSE: Let $y_{\max}$ be the true best policy effect and $\widehat{y}_{max}$ be the estimated best policy effect. Then the best policy MSE is

$$\text{MSE}_{\text{best}} = (\widehat{y}_{\max} - y_{\max})^2.$$

3. Best policy coverage: Let $\pi^\star$ and $\widehat{\pi}^\star$ be the true and estimated set of policies with the highest effect. Then, we define the best policy coverage as the intersection-over-union of these two sets

$$\text{IOU} = \frac{|\pi^\star \cap \widehat{\pi}^\star|}{|\pi^\star \cup \widehat{\pi}^\star|}.$$

These metrics are easily understood for a single model. For the RPS, we reported the performance metric averaged across all partitions in the RPS.

We visualize the RPS through a heat map. An example heatmap with instructions on how to read it is shown in Figure B.7. We also use these heatmaps in our empirical data examples in Appendix B.9.

For Simulation 1, we visualize the RPS in a heatmap in Figure B.8. As the size of the dataset increases, the Rashomon set becomes smaller as we become more confident in our estimates.

B.7 Generalizations

Here, we consider two generalizations of the methods discussed so far. First, we consider a family of heterogeneous effects functions beyond just heterogeneity splits. For example,

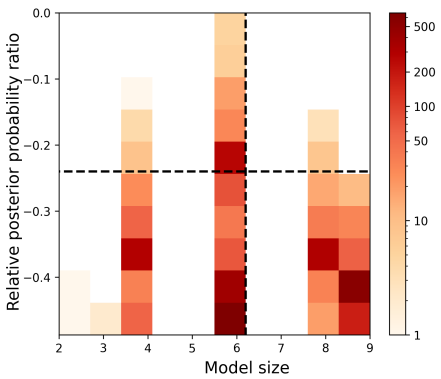


Figure B.7: Visualizing the Rashomon set through a heat map. This heatmap actually reflects a 2D histogram binned by the model size (number of pools in a partition) and the relative posterior probability ratio i.e., $(\mathbb{P}(\Pi \mid \mathbf{Z}) - \max \mathbb{P}(\Pi \mid \mathbf{Z}))/\max \mathbb{P}(\Pi \mid \mathbf{Z})$. The color of the bin reflects the number of times, averaged per simulation, a model at that sparsity and probability (distinct models may be in the same bin) appear in some Rashomon set. One might refine the set of partitions further by the probability and the sparsity. For example, if we want models with a relative probability of at least -0.25, then we look only at models that are above the dashed black horizontal line. If we want models with fewer than 6 pools, then we look only at models to the left of the dashed black vertical line. If we want both criteria to be satisfied, we look at the top left box.

there might be some heterogeneity in *slopes* (and slopes of slopes, and so on). Second, we extend our method to pool on the space of the covariance between coefficients, rather than on the coefficients themselves. This means that coefficients no longer need to be exactly equal but, instead, related through a sparse covariance structure.

B.7.1 Pooling higher order derivatives

We ask whether, given some feature combination $k = (k_1, \ldots, k_M)$, the marginal effect of increasing, say, k_1 has a linear effect. That is, we can just as simply allow for outcomes as we increase the intensity up a given feature that is not just a step function, but one that checks if there is a linear relationship.[7] In this case, there is no "large" versus "small" effect and no

[7]Extensions of this kind can be made to accommodate higher order derivatives and other bases as well, e.g., sinusoidal effects.

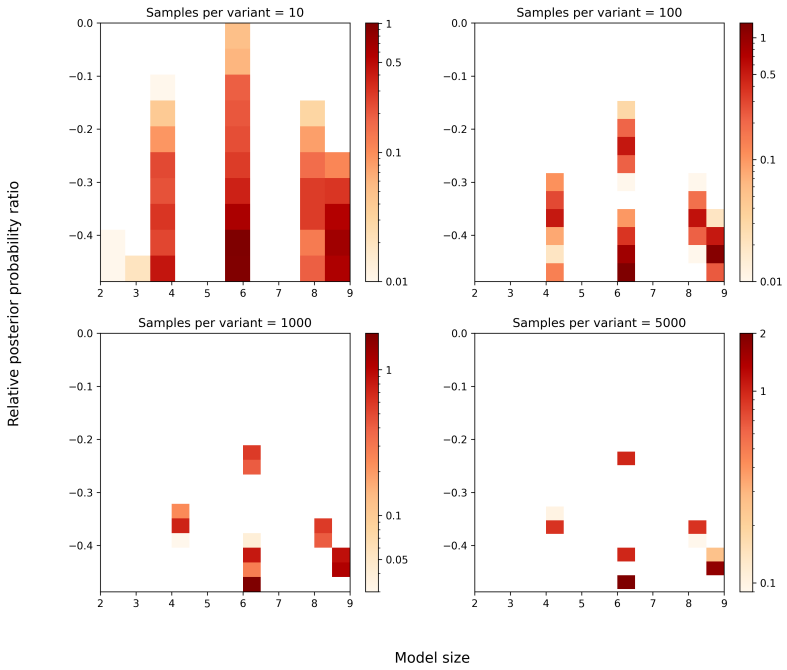


Figure B.8: Visualizing the Rashomon set in Simulation 1. Notice how as the size of the data set grows, the Rashomon set concentrates around a few very good models, one of which corresponds to the data generating process.

natural pool in the space of levels. However, there is a natural low dimensional effect and even pools when considering the space of slopes. The result is a framework that captures extensions of Bayesian treed models (e.g., Chipman et al. (2002)).

Before we proceed, we first generalize the notion of pools described in Definition 3.4.2.

Definition B.7.1 (Generalization of pools)**.** *Given M features taking on R partially ordered values each and some function $g(k, \boldsymbol{\beta})$, a pool π is a set of feature combinations k whose outcomes are given by $g(k, \boldsymbol{\beta}_\pi)$ where $\boldsymbol{\beta}_\pi$ depends on π.*

It is easy to see that the original pool defined in Definition 3.4.2 is recovered by setting $g(k, \beta_\pi) = \beta_\pi$.

For instance, suppose we are interested in linear effects. Then the regression equation for

pool π

$$y = g(k, \boldsymbol{\beta}_\pi) = \beta_{\pi,0} + \sum_{m=1}^{M} \beta_{\pi,m} k_m = \beta_{\pi,0} + \boldsymbol{\beta}_\pi^\intercal k$$

where $\boldsymbol{\beta}_\pi$ is the linear slope within that pool. The estimated outcome for feature combination $k \in \pi$ is $\widehat{y} = f(k, \widehat{\boldsymbol{\beta}}_\pi; \Pi)$, where $\widehat{\boldsymbol{\beta}}_\pi$ is estimated within each pool using some procedure like least squares.

For some partition Π, define the block vector $\boldsymbol{\beta} = [\boldsymbol{\beta}_{\pi_1}, \ldots, \boldsymbol{\beta}_{\pi_{|\Pi|}}]$ where $\pi_i \in \Pi$. Then, the general outcome function for any feature combination k can be written as

$$y = g(k, \boldsymbol{\beta}; \Pi) = \sum_{\pi \in \Pi} \mathbb{I}\{k \in \pi\} g(k, \boldsymbol{\beta}_\pi).$$

The practitioner is free to choose any domain-specific parametric function. For example, g could be a higher-order Taylor series-like expansion. Or, g could even be sinusoidal because the practitioner believes the outcomes are (piece-wise) sinusoidal. Of course, the more complex the estimation procedure for $\boldsymbol{\beta}$, the harder it is to enumerate the RPS.

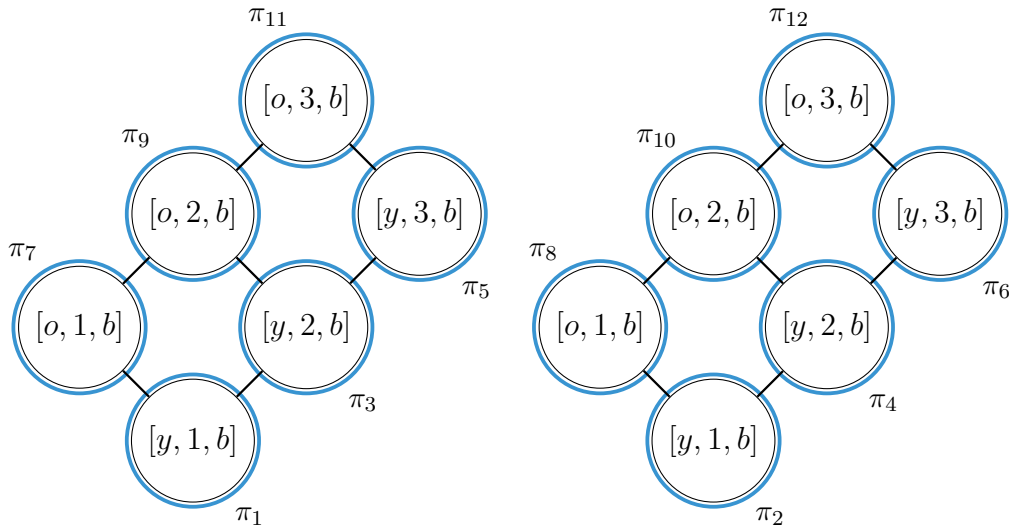
Observe that the form of the posterior remains the same,

$$\mathbb{P}(\Pi \mid \mathbf{Z}) \propto \exp\left\{-\mathcal{L}(\mathbf{Z}) + \lambda H(\Pi)\right\} = \exp\left\{-\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 + \lambda H(\Pi)\right\}.$$

Therefore, the results in Section 3.3 still hold. We can freely choose any other non-negative loss function, $\mathcal{L}(\mathbf{Z})$, and still use the same framework and algorithm to enumerate the RPS.

Further, the results in Section 3.6 are also valid when using an arbitrary parametric outcome function as discussed here. We summarize this in Theorem B.7.2.

Theorem B.7.2. *Suppose the outcome function is $g(k, \boldsymbol{\beta}; \Pi)$ for feature combination k, admissible partition Π, and some unknown parameter $\boldsymbol{\beta}$. Let us denote the estimated outcome for unit i with feature combination k by $\widehat{y}_i = g(k, \widehat{\boldsymbol{\beta}}; \Pi)$ where $\widehat{\boldsymbol{\beta}}$ is estimated from the data. If we use $\widehat{y}_i$ instead of $\widehat{\mu}_\pi$ in Equations 3.9 and 3.11, then*



(a) Drug B dosages, $b \in \{1,2\}$ (b) Drug B dosages, $b \in \{3,4,5\}$

Figure B.9: Hasse diagram for simulation with linear outcomes. y is young and o is old.

(i) Theorem 3.6.1 is still true,

(ii) Theorem 3.6.2 is still true, and

(iii) Algorithm 3 correctly enumerates the Rashomon partitions for outcome function f.

Proof of Theorem B.7.2. The results follow directly from Theorems 3.6.1, 3.6.2, and 3.6.3. □

To see the usefulness of the generalization, we motivate a simple example where we are interested in how a person's age affects their response to a treatment consisting of a combination of two drugs, A and B. Suppose that there are four possible dosages for drug A, $\{0,1,2,3\}$, six possible dosages for drug B, $\{0,1,2,3,4,5\}$, and people are classified as young aged or old aged where 0 indicates control. We assume that there is no treatment

effect unless drug A and drug B are taken together. The partition matrix is

$$\boldsymbol{\Sigma} = \begin{bmatrix} 0 & - & - & - \\ 0 & 0 & - & - \\ 1 & 0 & 1 & 1 \end{bmatrix}.$$

We visualize the twelve pools in Figure B.9 indicating heterogeneity in age and the dosages of drugs A and B.

Suppose that the treatment effects are piecewise linear (which generalizes the stepwise effects that we've assumed in previous simulations),

$$\begin{aligned}
\boldsymbol{\beta}_1 &= [0, -1, 0, 1] \\
\boldsymbol{\beta}_2 &= [1.5, -4, 0, 1.5] \\
\boldsymbol{\beta}_3 &= [0, -1, -0, 1] \\
\boldsymbol{\beta}_4 &= [4.5, -4, 0, 0.5] \\
\boldsymbol{\beta}_5 &= [4, -2, -1, 1] \\
\boldsymbol{\beta}_6 &= [1, 1, 1, -1] \\
\boldsymbol{\beta}_7 &= [-3, 2, -3, 1] \\
\boldsymbol{\beta}_8 &= [0, 0, 0, 0] \\
\boldsymbol{\beta}_9 &= [4, 2, -3, -1] \\
\boldsymbol{\beta}_{10} &= [0, 0, 0, 0] \\
\boldsymbol{\beta}_{11} &= [5, 2, -3, 0] \\
\boldsymbol{\beta}_{12} &= [5, -1, 0, -1],
\end{aligned}$$

where the first coefficient is the intercept and the remaining elements are slopes on each feature. For feature profiles with zero treatment effect, we set the effect to be 0, a constant. For the feature profile where drugs A and B are administered together, a random error

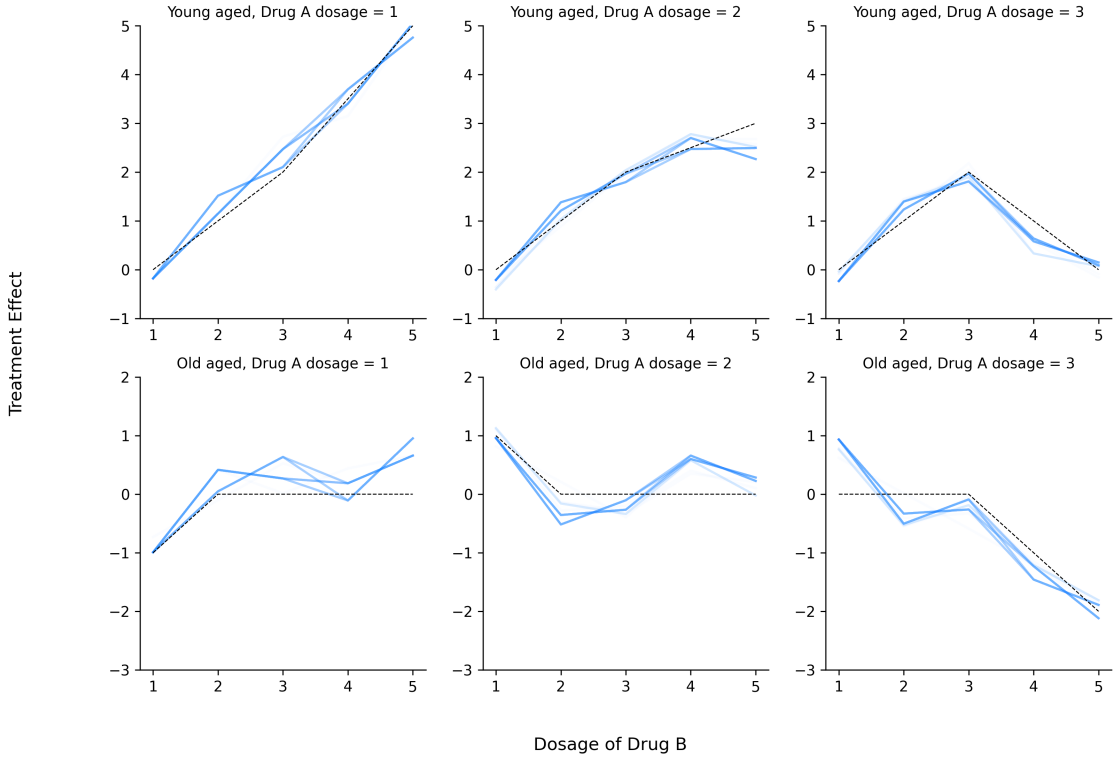


Figure B.10: The black line corresponds to the true data-generating process and the blue lines correspond to effects estimated in each model in the Rashomon set. We estimate the outcome of each pool as a linear function of the features. The denser the blue line, the more often it appears in the Rashomon set.

is drawn independently and identically from $\mathcal{N}(0, 1)$. We draw 10 measurements for each feature combination. We set $\lambda = 4 \times 10^{-3}$.

We illustrate the treatment effects for different combinations in black dashed lines in Figure B.10. By choosing a linear function as the outcome for each pool, we can find the Rashomon set. In Figure B.10, we show the estimated linear curves in 100 models present in the Rashomon set ($\epsilon \approx 5 \times 10^{-4}$) in blue. The denser the blue line, the more often it appears in the Rashomon set.

B.7.2 Sparse correlation structure between coefficients

Next, we explore the space of potential (sparse) covariance matrices between the coefficients. We now apply the Hasse structure to the elements of the variance-covariance matrix and pool on the space of covariances rather than the coefficients themselves. This generalization requires an additional distributional assumption on the coefficients. Specifically, assume that

$$\boldsymbol{\beta} \mid \mu, \boldsymbol{\Lambda} \sim \mathcal{N}(\mu, \boldsymbol{\Lambda})$$

where μ is some mean matrix and $\boldsymbol{\Lambda}$ is some covariance matrix. Then the posterior has the form

$$\mathbb{P}(\boldsymbol{\beta}, \boldsymbol{\Lambda}, \Pi \mid \mathbf{Z}) \propto \mathbb{P}(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\Lambda}, \mathbf{D}, \Pi) \cdot \mathbb{P}(\boldsymbol{\beta}, \boldsymbol{\Lambda}, \Pi)$$

The likelihood component of the loss is

$$\mathbb{P}(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\Lambda}, \mathbf{D}, \Pi) \propto \exp\left\{-\frac{1}{N}(\mathbf{y} - \mathbf{D}\boldsymbol{\beta})^{\intercal}\boldsymbol{\Lambda}(\mathbf{y} - \mathbf{D}\boldsymbol{\beta})\right\},$$

where N is the number of observed data points.

We do not have additional information about the covariance structure (though this could of course also be included in a prior) beyond the following three assumptions. First, we think that $\boldsymbol{\Lambda}$ is dense i.e., $\boldsymbol{\Lambda}$ is sparse in the number of uncorrelated outcomes. Second, we neither know nor want to know the correlation: it is an ℓ_0 problem. Third, we assume independence across the mean and correlation conditional on the covariance pooling. That is, Π is sufficient for the existence of dependence. Then

$$\mathbb{P}(\boldsymbol{\beta}, \boldsymbol{\Lambda}, \Pi) = \mathbb{P}(\boldsymbol{\beta} \mid \boldsymbol{\Lambda}, \Pi) \cdot \mathbb{P}(\boldsymbol{\Lambda} \mid \Pi) \cdot \mathbb{P}(\Pi)$$

Suppose that we have a partition $\Pi = \{\pi_1, \ldots, \pi_H\}$ where $H = |\Pi|$ and Π now is defined in the space of covariance matrices, so pooling means that the covariance values within a

pool are non-zero. Then, consider the following procedure for drawing the covariance matrix, $\mathbf{\Lambda} \in \mathbb{R}^{K \times K}$. For each pool $\pi_i \in \Pi$, draw $\mathbf{\Lambda}_i \sim f_i$ independently where f_i is some prior (for example, inverse Wishart). Then, $\mathbf{\Lambda} = \text{diag}(\mathbf{\Lambda}_i, \ldots, \mathbf{\Lambda}_H)$. The number of non-zero elements of $\mathbf{\Lambda}$ is given by $\|\mathbf{\Lambda}\|_0 = \sum_{i=1}^{H} h_i^2$. Therefore, we penalize the number of zero elements, $K^2 - \sum_{i=1}^{H} h_i^2$. Thus, the prior is

$$\mathbb{P}(\Pi) \propto \exp\left\{-\lambda\left(K^2 - \sum_{i=1}^{H} h_i^2\right)\right\}.$$

So our penalized loss function is just weighted mean-squared error penalized differently,

$$Q(\Pi; \mathbf{Z}) = \mathcal{L}(\Pi; \mathbf{Z}) + \lambda H(\Pi) = \frac{1}{n}(\mathbf{y} - \mathbf{D}\boldsymbol{\beta})^\intercal \mathbf{\Lambda}(\mathbf{y} - \mathbf{D}\boldsymbol{\beta}) + \lambda\left(K^2 - \sum_{i=1}^{H} h_i^2\right). \tag{B.7}$$

Theorem B.7.3. *Consider the same setup in Section 3.6 except the loss function is weighted mean squared error penalized by the number of zeros in the covariance matrix as in Equation B.7. Specifically, Equations (3.9) and (3.10) are modified, respectively, as*

$$b(\mathbf{\Sigma}, \mathcal{M}; \mathbf{Z}) = \frac{1}{n}\sum_{\pi \in \Pi_\mathfrak{f}}\sum_{k(i) \in \pi} \mathbb{I}\left\{k(i) \in \pi_\mathfrak{f}\right\} \widehat{\Lambda}^2_{k(i),k(i)}(y_i - \widehat{\mu}_\pi)^2 + \lambda H(\Pi, \mathcal{M}), \tag{B.8}$$

$$b_{eq}(\mathbf{\Sigma}, \mathcal{M}; \mathbf{Z}) = \frac{1}{n}\sum_{\pi \in \Pi_\mathfrak{f}}\sum_{k(i) \in \pi} \mathbb{I}\left\{k(i) \in \pi_\mathfrak{f}^c\right\} \widehat{\Lambda}^2_{k(i),k(i)}(y_i - \widehat{\mu}_\pi)^2. \tag{B.9}$$

where $\widehat{\Lambda}^2_{k,k}$ is the estimated variance of feature combination k.

Then

(i) Theorem 3.6.1 is still true,

(ii) Theorem 3.6.2 is still true, and

(iii) Algorithm 3 correctly enumerates the Rashomon partitions.

Proof of Theorem B.7.3. The results follow directly from Theorems 3.6.1, 3.6.2, and

3.6.3. □

B.8 Further Details on Related Work

It is useful to contrast our method with several other (some recent) approaches to study heterogeneity. Specifically, we are interested in their application to settings with partial orderings (e.g., factorial structure and admissibility) which is easily interpretable.

We will focus on four main related approaches: (1) canonical Bayesian Hierarchical Models (BHM) (Rubin, 1981; Gelman, 2006; Meager, 2019); (2) ℓ_1 regularization of marginal effects to identify heterogeneity (Banerjee et al., 2021); (3) causal forests (Wager and Athey, 2018); and (4) machine learned proxies (Chernozhukov et al., 2018). We intend this discussion to be a guide for practitioners considering implementing our proposed method or one of these state-of-the-art alternatives. We discuss conceptually related work (e.g. Bayesian decision trees) in previous sections. an Let us for the moment set aside the following immediate differences. Our focus on robustness, profiles, and enumerating the entire Rashomon Partition are all novel. Instead, it is useful to identify the philosophical differences across the various approaches and how they relate to us. Every approach, as we will note, effectively uses partitions Π at some point to determine which data to pool or not. The specific techniques create distributions, possibly degenerate, over these partitions, and these distributions are sampled from and marginalized to estimate treatment effects β_k. The interesting thing therefore is how one builds a distribution over Π.

B.8.1 Bayesian Hierarchical Models

We now discuss how our work relates to a canonical representation of a Bayesian Hierarchical Model. As discussed previously, our work is more similar to Bayesian Tree(d) models than to other methods for accounting for learning heterogeneity, such as Bayesian Model Averaging. For context, however, we present how our approach compares to the canonical Bayesian approach. The Bayesian perspective provides a compromise between complete and partial pooling. Partial pooling occurs by encouraging similarity in the values for parameters without

requiring strict equality. Using the notation from our model, for example, we could construct a model where

$$\mathbf{y} \sim N(\mathbf{D}\boldsymbol{\beta}, \sigma_y^2)$$

and, for the sake of exposition, all β are draw independently from

$$\boldsymbol{\beta} \sim N(\mu_\beta, \sigma_\beta^2).$$

Requiring that all values of $\boldsymbol{\beta}$ come from the same distribution encourages sharing information across potential feature combinations and encourages the effects on heterogeneity to be similar (but not identical). Meager (2019) uses this approach when comparing treatment effects across multiple domains. In that paper, the goal is not to pool across potentially similar treatment conditions but instead to (partially) pool across geographic areas.

As one example of the classical model, Meager (2019) has outcomes for household i in study k modeled as

$$y_{ik} \sim N(\mu_k + \tau_k \mathbf{T}_{ik}, \sigma_{yk}^2) \ \forall \ i, k$$

$$\begin{pmatrix} \mu_k \\ \tau_k \end{pmatrix} \sim N\left[\begin{pmatrix} \mu \\ \tau \end{pmatrix}, \begin{pmatrix} \sigma_\mu^2 & \sigma_{\mu\tau} \\ \sigma_{\mu\tau} & \sigma_\tau^2 \end{pmatrix} \right] \ \forall \ i, k$$

where τ_k, μ_k are the overall mean and treatment effect at area k, respectively. The vector $\mathbf{T_{ik}}$ is the treatment indicator for household i in study k.

One way to measure the degree of pooling is the (partial) "pooling factor" metric defined in Gelman (2006), $\omega(\boldsymbol{\beta}) = \sigma_y^2/\sigma_y^2 + \sigma_\beta^2$. The partial pooling metric quantifies how much the effect of treatment combinations varies compared to the overall heterogeneity in the outcomes. The partial pooling metric, the Meager (2019) context refers to the relative

variation related to differences between studies compared to sampling variability.

In contrast, we could think of our approach as using a prior on $\boldsymbol{\beta}$ conditional on the partitions that potentially force some values of β_k, β'_k to be equal. In Appendix B.2.2, we show that the objective function we use in Equation 3.5 corresponds to a hierarchical model where we draw the $\boldsymbol{\beta}$ vector as

$$\boldsymbol{\beta} \mid \Pi \sim \mathcal{N}(\boldsymbol{\mu}_\Pi, \boldsymbol{\Lambda}),$$

where $\boldsymbol{\mu}_\Pi$ is structured such that $\mu_k = \mu_{k'}$ for any $k, k' \in \pi \in \Pi$. Then, given some feature combinations $\mathbf{D}$, we draw the outcomes as

$$\mathbf{y} \mid \mathbf{D}, \boldsymbol{\beta} \sim \mathcal{N}(\mathbf{D}\boldsymbol{\beta}, \boldsymbol{\Sigma}).$$

To understand the variation within the $\boldsymbol{\beta}$ vector, we need to average across potential partitions, since some partitions will set $\beta_k = \beta'_k$ and others will not. This amounts to replacing the σ^2_β in the pooling factor with the variance of the distribution of $\mathbb{P}(\boldsymbol{\beta}|Z)$, which is defined in Equation 3.3.

We could also conceptualize the above derivation in terms of equality on $\boldsymbol{\beta}$ rather than the means μ_Π. If, for example, we replace $\boldsymbol{\Lambda}$ with $\boldsymbol{\Lambda}_\Pi$ where $Var(\mu_k, \mu'_k) = 0$ for any $k, k' \in \pi \in \Pi$ (or equivalently, when $\mu_k = \mu_{k'}$) then we enforce that $\beta_k = \beta'_k$. Of course, if we go the opposite direction and let the diagonal of $\boldsymbol{\Lambda}_\Pi$ be unconstrained then there is essentially no sharing of information across feature combinations.

Finally, hierarchical models of this type are, of course, quite flexible and we could construct more complex models that capture features of our pooling approach. Among those options would be to use the Bayesian version of penalized regression, such as the Bayesian Lasso, which would be philosophically related to the approach we describe in the next section.

B.8.2 Lasso regularization

This is the approach taken in prior work by several of the authors of the present paper, in Banerjee et al. (2021). There the setting was one in which the researcher faced a factorial experimental design: a crossed randomized controlled trial (RCT). The paper developed the Hasse structure described above and an approach that required transforming $\mathbf{D}$ into an equivalent form presented in Equation (B.1) in Appendix B.1. Here every parameter α_k represents the marginal difference between β_k and $\beta_{k'}$ where $\rho(k) = \rho(k')$ (they are the same profile) and k exactly differs from k' on one arm by one dose. The parameter vector $\boldsymbol{\alpha}$ records the marginal effects. Notice the support of $\boldsymbol{\alpha}$ therefore identifies Π (since non-zero entries determine splits).

The first difficulty in applying this to general settings of heterogeneity is that ℓ_1 regularization requires irrepresentability: that there is limited correlation between the regressors so that the support may be consistently recovered (Zhao and Yu, 2006). Unfortunately, the regression implied by the Hasse does not satisfy this so some pre-processing is required. Banerjee et al. (2021) apply the Puffer transformation of Jia and Rohe (2015) to retain irrepresentability and estimate the Lasso model. However, this is not free: the approach requires conditions on the minimum singular value of the design matrix. The authors leverage the structure of a crossed randomized controlled trial (which places considerable restriction on the design matrix) to argue that indeed these conditions are met. There is no guarantee and it is unlikely to be the case that these conditions are met for general factorial data of arbitrary covariates. So, tackling the much more general structure required moving away from regression (we use decision trees) and changing the regularization (we use ℓ_0).

The second key observation is that the Bayesian lasso means that the ℓ_1 penalty corresponds to priors $\mathbb{P}(\boldsymbol{\alpha})$ that are i.i.d. Laplace on every dimension k. That is

$$-\log \mathbb{P}(\boldsymbol{\alpha}) = \log \prod_k \mathbb{P}(\alpha_k) = \log \prod_k \exp(-\lambda \left|\alpha_k\right|) = \lambda \sum_k \left|\alpha_k\right| .$$

Note that this is true whether one uses regular lasso, Puffer transformed lasso, spike-and-slab

lasso, group lasso (up to the group level), and so on. No matter at whatever level the ℓ_1 sum is being taken, it corresponds to independence at that level in the prior.

In practice what this means is that given two partitions Π and Π', which have the same number of pools and which have the same loss value, if one is more consistent with independent values of α_k than the other, it will receive a higher posterior. There are at least two problems.

The main philosophical problem is that there is no reason to place the meta-structure that the marginal differences between adjacent variants should have an i.i.d. distribution. In fact, one might think that the basic science or social science dictates *exactly the opposite*. Independence means that a marginal increase in dosage of drug A, holding fixed B and C at some level, is thought to be *independent* of increasing A holding fixed B and C at (potentially very similar) different levels. Similarly, the marginal value of receiving a slightly larger loan given that the recipient has 10 years of schooling and started 5 previous businesses is *independent* of receiving a slightly larger loan if the recipient had 10 years of schooling and started 6 previous businesses. Independence is unreasonable in both examples.

There is a second issue in that if an object of interest is Π, this approach provides no way forward. Regularization delivers posteriors over $\boldsymbol{\alpha}$: $\mathbb{P}(\boldsymbol{\alpha} \mid \mathbf{y}, \mathbf{X})$. This implies a posterior over $S_{\boldsymbol{\alpha}}$. The map from $S_{\boldsymbol{\alpha}}$ to Π is deterministic, and is given by some $\phi(S_{\boldsymbol{\alpha}}) = \Pi$, which means that

$$\mathbb{P}(\Pi \mid \mathbf{y}, \mathbf{X}) = \int_{\boldsymbol{\alpha}} \mathbf{1}\left\{\phi(S_{\boldsymbol{\alpha}}) = \Pi\right\} \cdot \mathbb{P}(\boldsymbol{\alpha} \mid \mathbf{y}, \mathbf{X})$$

is the actual calculation of interest.

So the regularization approach requires the statistician to take all the marginal parameters to be i.i.d., and given this, integrate over possible coefficient vectors that are consistent with this specific aggregation. This makes calculating an RPS very difficult.

B.8.3 Causal Random Forests

We now compare our approach to Causal Random Forests (CRFs) introduced by Wager and Athey (2018). CRFs construct regression trees over the space of potential combinations of covariates. Trees partition the space of covariates into "leaves." Unlike our setting, trees are heirarchical; the procedure to construct trees involves splitting the observed data in two based on X_i being above or below a threshold. They then partition recursively, dividing each subsequent group until the leaves contain very few observations. This approach can also be thought of as finding nearest neighbors, where the number of neighbors is the number of observations in the leaf and using distance on the tree as the closeness metric. CRFs construct a conditional average treatment effect at a pre-determined point $X = x$, $\tau(x) = \mathbb{E}[Y(1) - Y(0)|X = x]$ where $Y(1)$ is potential outcome for the treated and $Y(0)$ is the potential outcome for the control.

Relating this back to our work, take T to be a tree and $\pi \in \Pi(T)$ to be a leaf in the tree, which corresponds to a pool in our language. Then, the estimated expected outcomes for each leaf is

$$\widehat{\beta}_\pi = \frac{1}{|\{i : X_i \in \pi\}|} \sum_{\{i:X_i \in \pi\}} Y_i.$$

Further, taking τ_π to be the treatment effect of observations in pool (leaf) π and W_i as the treatment indicator, which we assume orthogonal to X and Y, the estimated treatment effect for π is

$$\widehat{\tau}_\pi = \frac{1}{|\{i : W_i = 1, X_i \in \pi\}|} \sum_{\{i:W_i=1,X_i \in \pi\}} Y_i \quad - \quad \frac{1}{|\{i : W_i = 0, X_i \in \pi\}|} \sum_{\{i:W_i=0,X_i \in \pi\}} Y_i.$$

To summarize, the approach for forming trees splits the observed covariate space into partitions, known as leaves. Each leaf consists of a mix of people in treatment and control groups and, in fact, the specification of the tree depends on this balance across treatment

and control groups since the algorithm requires that splitting be done in a way that preserves a minimum number of treatment and control in each leaf. To compute a treatment effect conditional on a particular value of X, look at the difference in outcome between treated and control people in a given leaf. Outcomes are not considered with constructing the tree (in contrast to our proposed approach) and treatment status is not used to split explicitly but does influence the construction of the tree through the sample size restriction.

Despite being similar in that we both use geometric objects that partition the space of covariates, there are three fundamental differences between our approach and CRFs. The first difference is geometric. CRFs use regression trees, whereas we use Hasse diagrams. Regression trees are appealing in many settings because of their flexibility in representing complex, nonlinear, relationships between variables. Regression trees, however, require imposing a hierarchy between variables that is not supported by the data. This hierarchy is "baked in" to the structure of the trees and is evident from how we describe constructing trees in the previous paragraph. The data, however, are not fully hierarchical and are instead partially ordered.

This mismatch creates an identification issue. Within education and within income, there are clear orderings. There is, however, no heirarchy between education and income. One tree may, therefore, split first based on income and then split on education conditional on income while another tree does the opposite. In both cases, we can trace the trees to end up with the same estimated treatment effects for any group of covariates (as shown in Wager and Athey (2018)). The trees themselves, however, arise from this arbitrary ordering and are, thus, not interpretable. Work such as Bénard and Josse (2023) describe measures of variable importance in CRFs, but the problem of an arbitrarily imposed hierarchy is still present. Hasse diagrams, in contrast, are the natural geometry for partially ordered sets, alleviating this issue and allowing the researcher to interpret the pooling structure on the domain of the covariates directly.

The second difference is computational but has conceptual implications. In both our approach and CRFs, we do not take the stucture of the partition as known. Both approaches

must, therefore, account for additional uncertainty in treatment effect estimates that arises from not knowing the partition. In CRFs, bootstrap samples over the data propagate this uncertainty. CRFs then aggregate over trees using Monte Carlo averaging over $b = 1, .., B$ boostrap samples of the covariates and outcomes, $\{Z_1, ..., Z_n\}$,

$$RF(\pi; Z_1, ..., Z_n) \approx \frac{1}{B} \sum_{b=1}^{B} T(\pi; \xi_b; Z^*_{b1}, ..., Z^*_{bs}),$$

where π represents a pool or leaf specifying a combination of features and levels. The ξ_b term is an additional stochastic component. The trees sampled as part of this process create a "forest" are, by definition, random draws given the data. That is, given a different set of data, the distribution of likely trees would change. They are also not guaranteed to be optimal or nearly optimal. If the goal is to estimate average treatment effects, this approach represents a principled way to explore the space of trees. If the goal, however, is to identify potential models of heterogeneity, then sampling randomly is very unlikely to produce high quality trees. With Rashomon partitions, by definition, we guarantee that all models in our set are of high posterior.

To this point, we have not discussed inference in CRFs. A key contribution of Wager and Athey (2018) is forming so-called "honest" trees that account for issues that arise when using the same data to learn trees and then to make inference conditional on the group of trees. In our work, we use a Bayesian framework to address this issue, which also has the advantage of being able to estimate functions of treatment effects (see 3.3.1). Future work, however, could consider Rashomon sets for honest regression trees. This work would build upon our own work as well as Xin et al. (2022) that introduces Rashomon sets for classification trees. The algorithm for inference would begin with splitting as proposed by Wager and Athey (2018) to preserve honest inference, then construct Rashomon sets using the algorithm from Xin et al. (2022). Since the space of trees is enormous, finding the "best" tree is impossible, which creates issues for finding the Rashomon set since it is used to define the reference partition. Fortunately, a recent paper by Hu et al. (2019) provides an algorithm. While this approach

would allow the CRF framework to find optimal trees, it does not address the identifiability issue that arises when using trees for data that are only partially ordered. Similar work was explored in Hahn et al. (2020), who estimate heterogeneous treatment effects using a sum of Bayesian regression trees, which they refer to as the Bayesian causal forest. They decompose the outcome into a mean outcome and a treatment effect. Since they are only interested in the treatment effect, the mean outcome becomes a nuisance parameter. They impose a vague prior on the mean and a strong prior on the treatment effect. Otherwise, the tree estimation procedure is identical to Bayesian Additive Regression Trees (Chipman et al., 2010).

Third, both our approach and CRFs impose regularization but do so in philosophically very different ways. We take the perspective that we do not know and cannot fully enumerate correlation structure in a high dimensional space. So we use the ℓ_0 prior, which we show is the least informative prior in Theorem 3.3.3. In other words, we regularize, and impose a prior, on the size of the partition. In doing so, we are trading off information on full distribution to robustly identify partitions. On the other hand, causal forests regularize on the number of observations in each leaf of the tree. Specifically, they require at least k samples in each leaf. This choice is sensible because with insufficient data there is no information. At the same time, this is odd as the regularization depends directly on the data. Elaborating on this, we can write the posterior for some tree T given data $\mathbf{y}, \mathbf{X}$ as

$$\mathbb{P}(T \mid \mathbf{y}, \mathbf{X}) \propto \mathbb{P}(\mathbf{y} \mid T, \mathbf{X})\mathbb{P}(T \mid \mathbf{X}) \text{ where } \mathbb{P}(T \mid \mathbf{X}) \propto \exp\left\{-\frac{\lambda}{\min_{\pi \in \Pi(T)} n_\pi(\mathbf{X})}\right\},$$

where $\Pi(T)$ is the set of pools (leaves) in T and $n_\pi(\mathbf{X})$ is the number of observations in $\mathbf{X}$ that belong to pool π. This prior down-weights and discards partitions where for *some* π the observations are low. In that sense, the prior effectively assumes that in the background there is a kind of stratification – that observations are sampled from some process such that all pools have enough observations, though of course *the true partition is unknown.* This feels awkward as there is a relationship between the data collection process and the actual

true partitioning wherein the user of the causal forest is assuming that they have effectively stratified data collection against the unknown partitions.

Together, these differences mean that the scope of our method is wider than CRFs. While both methods can estimate heterogeneity in treatment effects and control for multiple testing, we also produce interpretable explanations of heterogeneity. For the reasons outlined above, namely identification and sampling, it is not possible to extract information on the relationship between covariates from elements of the random forest. We can, of course, test for any hypothesis about potential heterogneeity between arbitrary combinations of features, but CRFs require that we specify the hypothesis *a priori*. In our setting, however, finding the set of high posterior probability partitions gives a policymaker or researcher as set of potential models of heterogeneity and interaction between the covariates that can be used to design future policies or generate new research hypotheses. On the other hand, our method assumes that the posterior has separated modes. If the posterior distribution is very flat or has many (many) very similar modes, then the Rashomon set will be very large and our benefits in terms of interpretability will diminish.

B.8.4 Treatment heterogeneity via Machine Learning Proxies

Chernozhukov et al. (2018) propose a general framework for using machine learning proxies to explore treatment effect heterogeneity. They allow for estimation of multiple outcomes, including conditional average treatment effects and treatment effect heterogeneity between the most and least impacted groups. Rather than search the space of covariates directly, Chernozhukov et al. (2018) uses a machine learning method to create a "proxy" for the heterogeneous effects. This approach has the advantage that it can be applied in high dimensional settings. A downside, however, is that the machine learning proxies are often uninterpretable in terms of the original covariates, making it necessary to post-process the treatment effect distributions to gain insights about particular covariates.

We now give a brief overview to unify notation but do not exhasutively cover all the estimators presented in Chernozhukov et al. (2018). Say that $s_0(Z)$ is the true conditional

average treatment effect, $\mathbb{E}[Y(1)|Z] - \mathbb{E}[Y(0)|Z]$ Ascertaining the functional form of the relationship between the non-intervention covariates X and the outcome Y, though, is complicated when X is high dimensional. In response, Chernozhukov et al. (2018) use a machine learning method (e.g. neural networks, random forests, etc) to construct a proxy for $s_0(Z)$ using an auxillary dataset. In a heuristic sense, this proxy serves the role of a partition π, in that it aggregates across covariates to separate the data based on the treatment effect. This analogy is most direct when the machine learning model is a decision tree (which it need not be) since in that case leaves of the tree would correspond to partitions of the covariate space based on treatment effect. After computing the machine learning proxy, Chernozhukov et al. (2018) then project it back to the space of the observed outcomes. It is then also possible to construct clusterings based on the proxies and related those clusterings back to the outcomes. Chernozhukov et al. (2018) differs from our approach in many of the same ways as the comparison with Wager and Athey (2018), namely that we focus on identifying multiple explanations for heterogeneity and that we utlize the Hasse diagram as a geometric representation of partial ordering. We also find that this structure is sufficient to explore models for heterogeneity on the space of the covariates without the need to use proxies.

B.9 Appendix to Empirical Data Examples

B.9.1 Charitable giving and telomere lengths

Figures B.11 and B.12 visualize the Rashomon sets for the charitable giving datasets of Karlan and List (2007) and the NHANES telomere lengths using the 2D histogram that is described in Figure B.7 in Appendix B.6.

B.9.2 Heterogeneity in the impact of microcredit access

For the microcredit data from Banerjee et al. (2015), we present the results for all profiles in Figure B.13. This includes the robust profiles we discussed in Figure 3.6 as well as the non-robust ones.

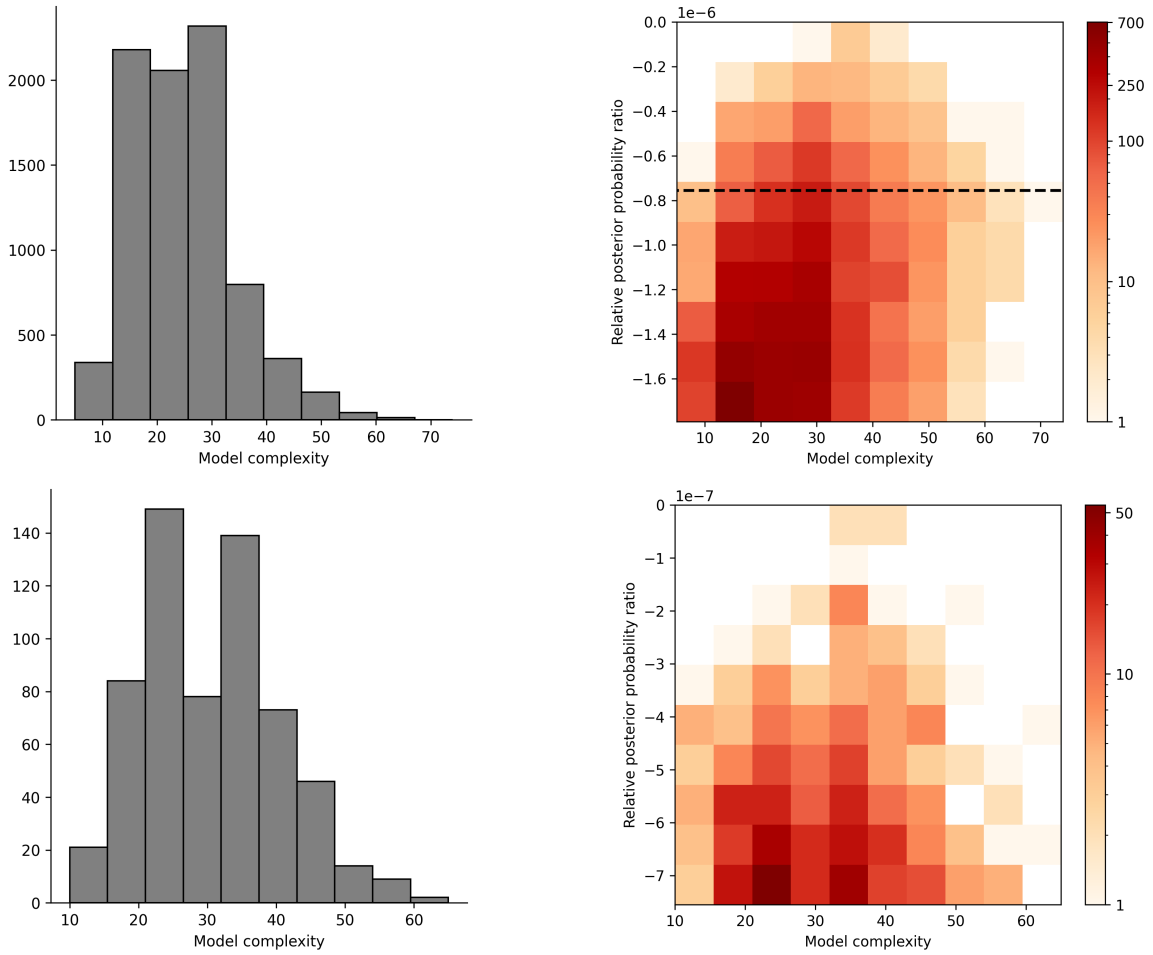


Figure B.11: Visualizing the Rashomon set for Karlan and List (2007) charitable donations dataset. The top two panels show the distribution of partition sizes and a 2D histogram of how partition sizes and relative posterior probabilities vary. The black dotted line in the 2D histogram shows our chosen Rashomon threshold. The bottom two panels show the same after pruning low-posterior models.

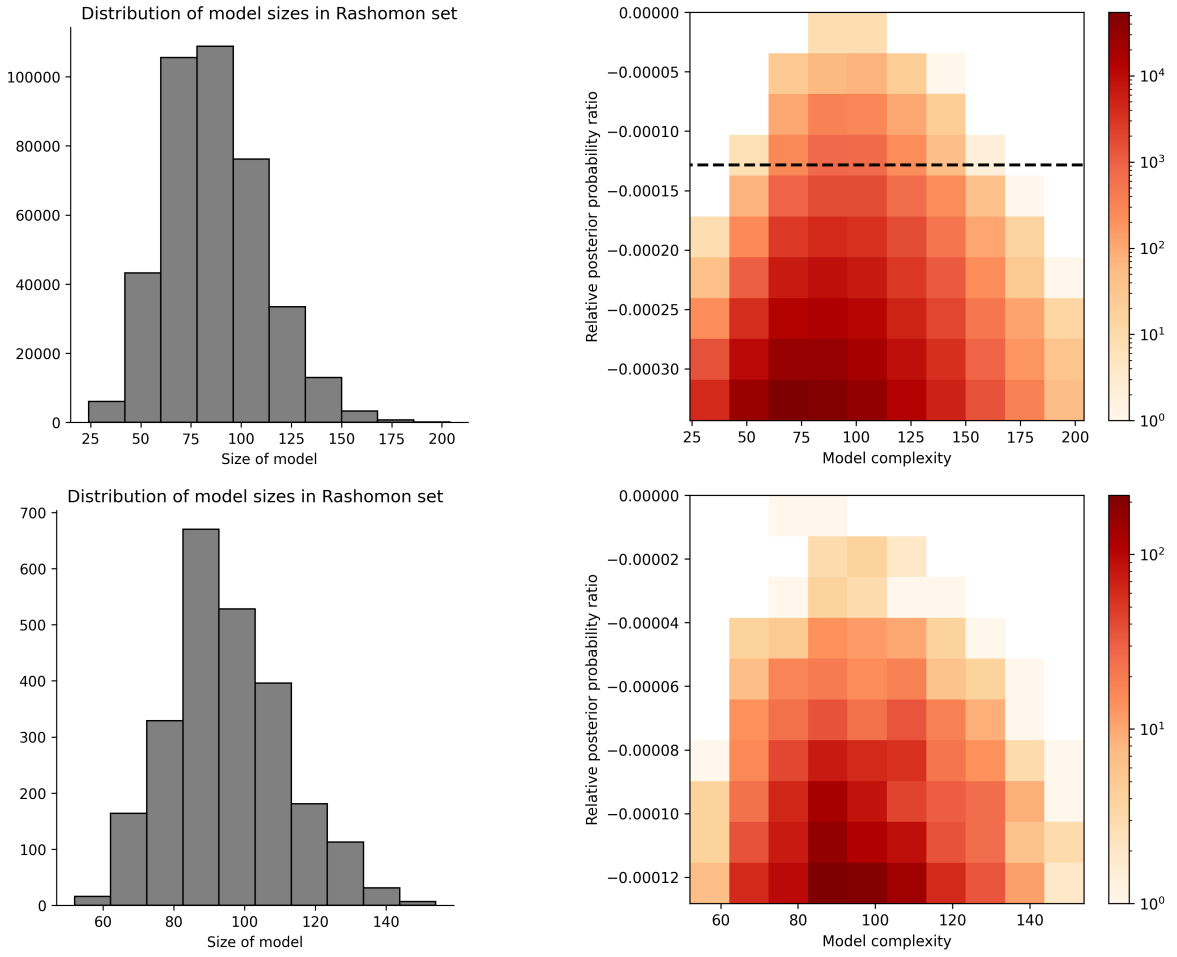


Figure B.12: Visualizing the Rashomon set for NHANES telomeres dataset. The top two panels show the distribution of size of models and their relative posterior probability relative. The black dashed vertical and horizontal lines show the sparsity cutoff and Rashomon cutoff respectively. The bottom two panels show the same after pruning low-posterior models.

Additionally, we look at the treatment effect heterogeneity across genders,

$$\text{HTE}(\mathbf{x}) = \mathbb{E}\left[\{Y_i(1, F, \mathbf{x}) - Y_i(0, F, \mathbf{x})\} - \{Y_i(1, M, \mathbf{x}) - Y_i(0, M, \mathbf{x})\}\right],$$

where $Y_i(\cdot, F, \cdot)$ is interpreted as the potential outcome of household i were it headed by a woman, and $Y_i(\cdot, M, \cdot)$ is the potential outcome of household i were it headed by a man. As before, we use the sample means $\widehat{y}(\cdot)$ to find $\widehat{\text{HTE}}(\mathbf{x})$ and $\text{sign}\{\widehat{\text{HTE}}(\mathbf{x})\}$. Again, we sort $\mathbf{x}$ into profiles and repeat the same counting and averaging exercise. We visualize the results as before in Figure B.14. For most profiles, we see essentially no robust conclusions about gender heterogeneity in treatment effects. We highlight a few robust items below.

We see an increase in loans procured by households headed by women with past business experience when compared to households headed by men. When these households are already in debt with no previous experience, they tend to borrow less. We see no heterogeneity by gender in the amount of informal loans procured.

Households headed by women tend to consistently spend more. However, they spend more money on durable goods than households headed by men. We also see that, in the absence of past experience, there is a decline in expenditure on tempting goods compared to households headed by men. We also see a higher tendency for women to invest in business assets more than men.

We find that households headed by women with no past experience have a lower revenue than men. But this effect is reversed when the households do have previous business experience. However, there is no heterogeneity by gender in the profit or the number of employees. We also find that households headed by women tend to spend fewer hours working when they are in debt or when there is regional competition. But this makes a negligible difference in the profits.

We find that in households headed by women, there is less participation in the business by women if the household is in debt and there is competition from neighbors. We also find that fewer girls attend school in households headed by women with no previous experience

than in households headed by men.

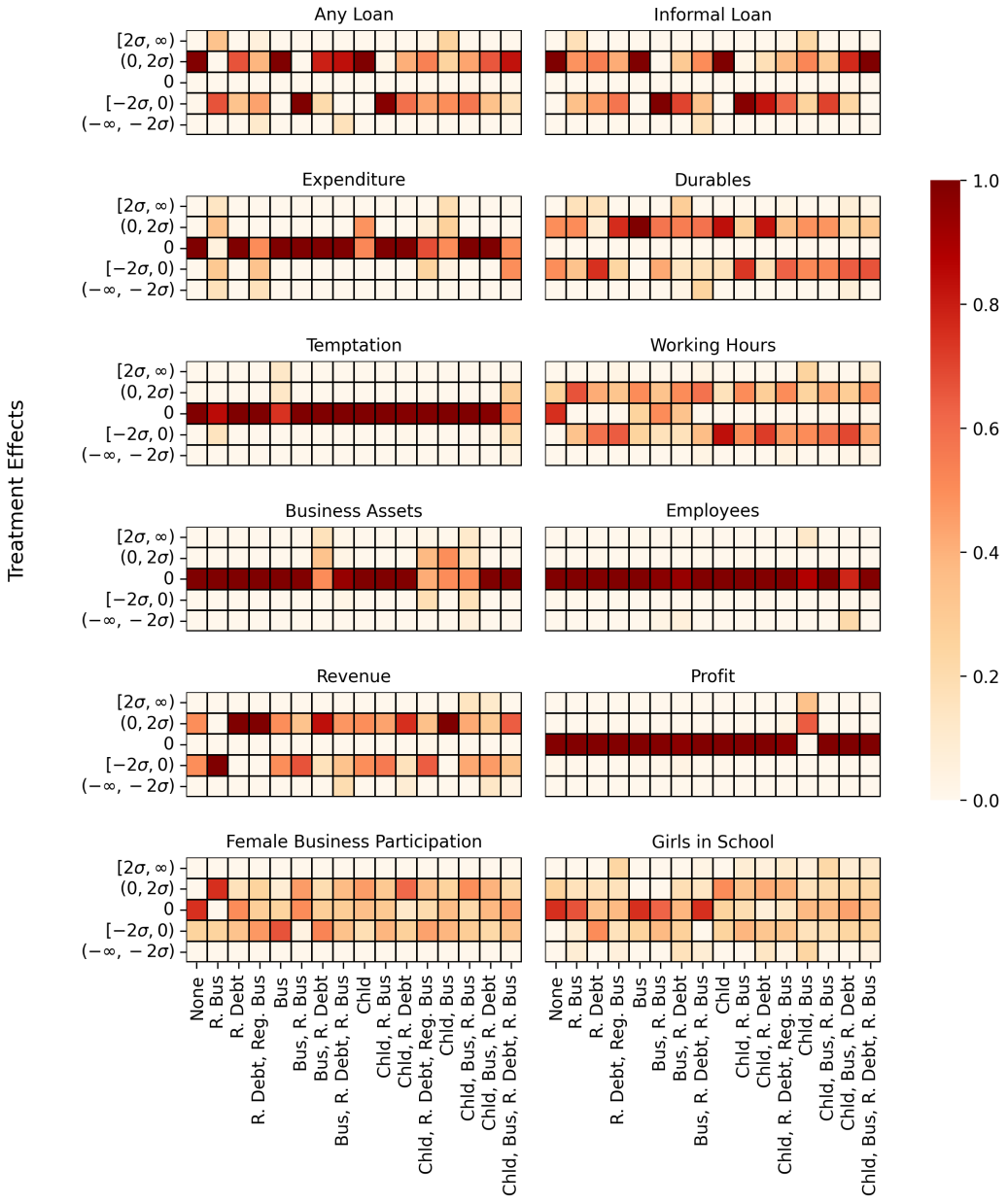


Figure B.13: Here, we visualize the average number of models in the Rashomon set indicating a positive, zero, or negative effect. Each column corresponds to a different feature profile where the label denotes which features are active (i.e., do not take the lowest level). "None" means that all features are taking these lowest values. We also allow the gender of the household head and education status of the household head to take on any value in all of the sixteen feature profiles.

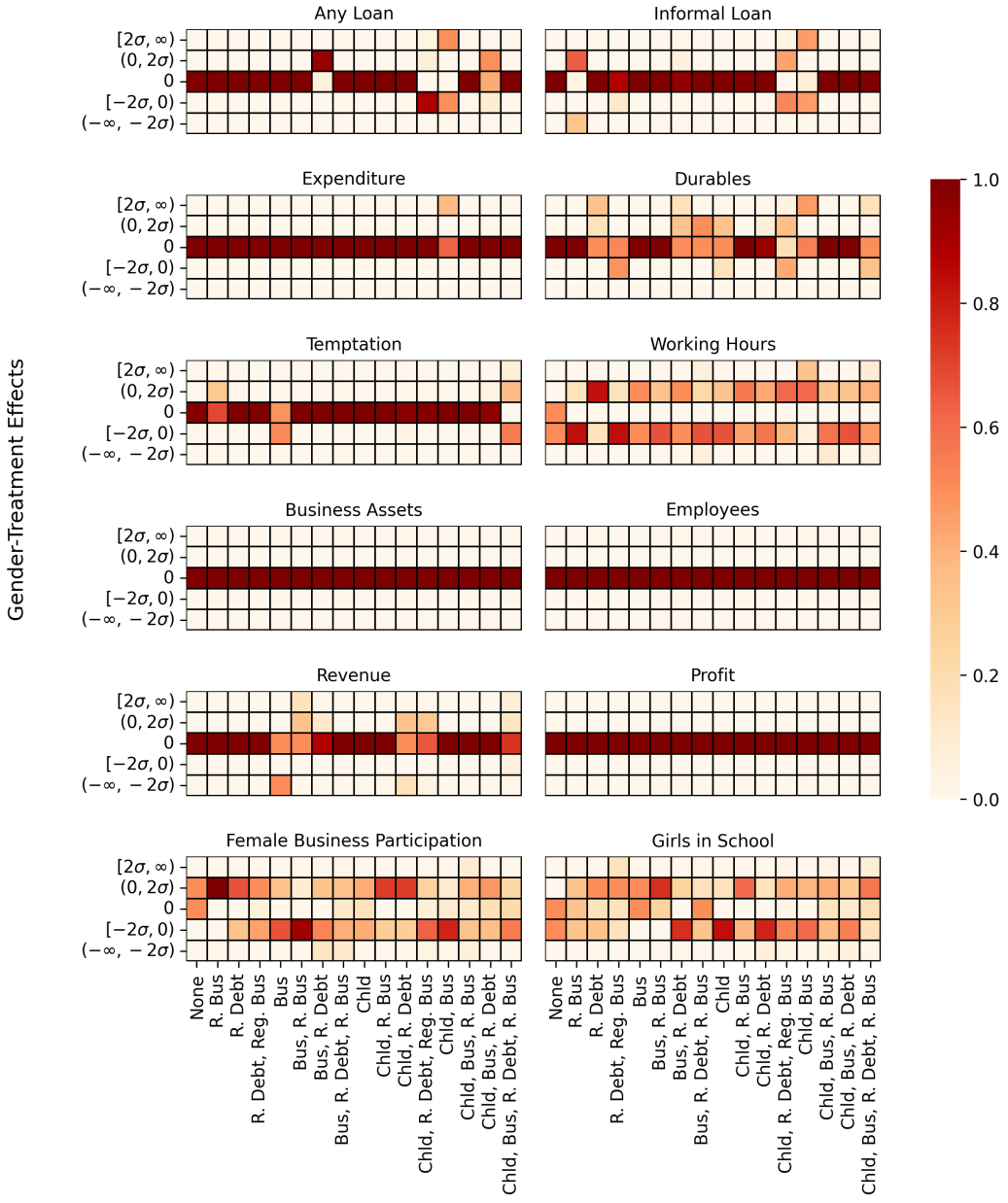


Figure B.14: Here, we visualize the average number of models in the Rashomon set indicating a positive, zero, or negative effect. The axis labels should be read as in Figure B.13.

Appendix C

SUPPLEMENT TO CHAPTER 4

> "Contradiction! Wonderful and blessed contradiction of nonsense and human complication to be alive!"
>
> — Mosaic, *Rhythm of War*

This appendix contains all technical details and proofs for results presented in Chapter 4.

C.1 Additional Preliminaries and Existing Results

C.1.1 Additional Definitions

Paths. If $p = \langle X_1, X_2, \ldots, X_k, \rangle, k \geq 2$, then with $-p$ we denote the path $\langle X_k, \ldots, X_2, X_1 \rangle$. For two disjoint subsets $\mathbf{X}$ and $\mathbf{Y}$ of $\mathbf{V}$, a path from $\mathbf{X}$ to $\mathbf{Y}$ is a path from some $X \in \mathbf{X}$ to some $Y \in \mathbf{Y}$. If $\mathcal{G}$ and $\mathcal{G}^*$ are two graphs with identical adjacencies and p is a path in $\mathcal{G}$, then the *corresponding path* p^* is the path in $\mathcal{G}^*$ constituted by the same sequence of nodes as p. The *length* of a path $p = \langle X_1, X_2, \ldots, X_k, \rangle, k \geq 2$, is equal to the number of edges on the path p, and denoted by $|p|$, that is, $|p| = k - 1$.

Concatenation of paths. We denote the concatenation of paths by $\oplus$, so that for example $p = p(X_1, X_k) \oplus p(X_k, X_m)$. In this paper, we only concatenate paths if the result of the concatenation is again a path.

Definition C.1.1 (Chordal Graph)**.** *A graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ is chordal if for every path $\langle P_1, P_2, \ldots, P_k \rangle$, $k > 3$ in $\mathcal{G}$ such that edge $\langle P_1, P_k \rangle$ is in $\mathcal{G}$, there is an edge $\langle P_i, P_j \rangle$, $1 \leq i < j \leq k$ in $\mathcal{G}$, such that $j - i > 1$.*

C.1.2 Existing Results

Theorem C.1.2 (Theorem 2.1 of Zhao et al., 2005)**.** *Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be two MAGs on the same set of vertices $\mathbf{V}$. Then $\mathcal{M}_1$ and $\mathcal{M}_2$ are Markov equivalent if and only if $\mathcal{M}_1$ and $\mathcal{M}_2$ have the same skeleton and the same minimal collider paths.*

Corollary C.1.3 (Lemma 4.1 of Zhang (2008b))**.** *Let $\mathcal{G}$ be an essential ancestral graph. Then, the circle component of $\mathcal{G}$ i.e., a subgraph of $\mathcal{G}$ containing only edges of type $\circ\!\!-\!\!\circ$ is a union of disconnected chordal graphs.*

Lemma C.1.4 (Lemmas B.4, B.5 and Corollary B.6 of Zhang (2008b))**.** *Let $\mathcal{G}$ be an essential ancestral graph. If path $p = \langle V_1, \ldots, V_k \rangle, k > 1$, does not contain any edge of the form $V_i \leftarrow\!\!\bullet V_{i+1}, 1 \leq i \leq k-1$ and if there is an edge $\langle V_1, V_k \rangle$ in $\mathcal{G}$, then $V_1 \to V_k$, or $V_1 \circ\!\!-\!\!\bullet V_k$ is in $\mathcal{G}$.*

Furthermore, if $V_{k-1} \bullet\!\!\rightarrow V_k$ is in $\mathcal{G}$, then $V_1 \to V_k$, or $V_1 \circ\!\!\rightarrow V_k$ is in $\mathcal{G}$.

Lemma C.1.5 (Lemmas B.7 of Zhang (2008b))**.** *Let $\mathcal{G}$ be an essential ancestral graph. If path $p = \langle V_1, \ldots, V_k \rangle, k > 1$, is of the form $V_1 \circ\!\!-\!\!\circ V_2 \circ\!\!-\!\!\circ \ldots \circ\!\!-\!\!\circ V_k$ and there is an edge $\langle V_1, V_k \rangle$ in $\mathcal{G}$, then $V_1 \circ\!\!-\!\!\circ V_k$ is in $\mathcal{G}$.*

Lemma C.1.6 (Lemmas B.8 of Zhang (2008b))**.** *Let $\mathcal{G}$ be an essential ancestral graph. If path $p = \langle V_1, \ldots, V_k \rangle, k > 3$, is an unshielded path of the form $V_1 \circ\!\!-\!\!\circ V_2 \circ\!\!-\!\!\circ \ldots \circ\!\!-\!\!\circ V_k$ in $\mathcal{G}$, then there is no edge edge $\langle V_i, V_j \rangle$ in $\mathcal{G}$, where $1 \leq i < j \leq k$.*

Lemma C.1.7 (Lemma A.1 of Zhang, 2008b)**.** *Let $\mathcal{P}$ be an essential ancestral graph, and let A, B, and C be three distinct nodes in $\mathcal{P}$. If $A \bullet\!\!\rightarrow B \circ\!\!-\!\!\bullet C$ is in $\mathcal{P}$, then $A \bullet\!\!\rightarrow C$ is also in $\mathcal{P}$. Furthermore, if $A \to B$ is in $\mathcal{P}$, then $A \to C$, or $A \circ\!\!\rightarrow C$ is in $\mathcal{P}$.*

Lemma C.1.8 (Lemma 7.5 of Maathuis and Colombo (2015))**.** *Let X and Y be two distinct nodes in an essential graph $\mathcal{P}$. If edge $X \leftarrow\!\!\bullet Y$ is in $\mathcal{P}$ then any path $p = \langle X = P_1, P_2, \ldots, P_k = Y \rangle, k > 1$ from X to Y, must contain at least one edge of the form $P_i \leftarrow\!\!\bullet P_{i+1}$, $i \in \{1, \ldots, k-1\}$.*

Conversely, if a path $q = \langle Q_1, \ldots, Q_r \rangle, r > 1$ *does not contain any edge of the form* $Q_j \leftarrow\!\bullet Q_{j+1}$, $j \in \{1, \ldots, r-1\}$, *then* q *is a possibly directed path from* Q_1 *to* Q_r.

C.2 Auxiliary Results

To reason with partial mixed graphs that are not essential ancestral graphs, we need to make some tweaks to definitions that are generally used in essential ancestral graphs, particularly the definition of a possibly directed path (Definition C.2.1). Definition C.2.1 is also listed in the main text preliminaries, but we also include it here for emphasis.

Definition C.2.1 (Possibly directed/causal path)**.** *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a partial mixed graph, and let $p = \langle V_1, \ldots, V_k \rangle, k \geq 2$ be a path in $\mathcal{G}$. Then p is a possibly directed path in $\mathcal{G}$ if for every $i, j \in \{1, \ldots, k\}, i \neq j$, $V_i \leftarrow\!\bullet V_j$ is not in $\mathcal{G}$.*

We also reformulate and prove a few important and well known essential ancestral graph results for our general partial mixed graph setting. Note that Lemma C.2.2 is our equivalent Lemma B.1 of Zhang (2008b), Corollary C.2.3 is our version of Lemma B.2 of Zhang (2008b), and Lemma C.2.4 is our version of Lemma 1 of Meek (1995) and Lemma A.1 of Zhang (2008b) (given in Lemma C.1.7 above).

Lemma C.2.2 (Unshielded subsequence forms a possibly directed path)**.** *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a partial mixed graph. Let $p = \langle V_1, \ldots, V_k \rangle, k > 1$ be a possibly directed path in $\mathcal{G}$. Then there is a subsequence of p called p', $p' = \langle V_1 = V'_1, V'_2, \ldots, V'_\ell = V_k \rangle$, $\ell > 1$, such that p' is an unshielded possibly directed path.*

Proof of Lemma C.2.2. Observe that any subsequence of p is also a possibly directed path. Consider the following subsequence construction algorithm that iteratively removes shielded triples:

(i) Set $p' = p$.

(ii) While there are shielded triples in p':

  (a) Find a shielded triple $\langle V_{i-1}, V_i, V_{i+1} \rangle$.

  (b) Construct $p' = p'(V_1, V_{i-1}) \oplus p'(V_{i+1}, V_k)$.

Thus, we have a subsequence of p, $p' = \langle V_1 = V'_1, \ldots, V'_\ell = V_k \rangle$ that is unshielded and possibly directed. □

Corollary C.2.3. *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a partial mixed graph such that the orientations in $\mathcal{G}$ are closed under rule R1. Let $p = \langle A, \ldots, B \rangle$ be an unshielded possibly directed path in $\mathcal{G}$. Then:*

(i) If there is a $\circ\!\!\rightarrow$ or $\rightarrow$ edge on p, then all edges after that edge on p are of type $\rightarrow$

(ii) If there is a $\circ\!\!-\!\!\circ$ edge on p, this edge occurs before a $\circ\!\!\rightarrow$ or $\rightarrow$ edge on p.

(iii) There is at most one $\circ\!\!\rightarrow$ edge on p

Proof of Corollary C.2.3. Follows from the fact that orientations in $\mathcal{G}$ are completed under R1 and the fact that p is an unshielded possibly directed path. □

Lemma C.2.4 (Property P1)**.** *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a partial mixed graph such that the orientations in $\mathcal{G}$ are closed under rules R1 and R2. For any three vertices $A, B, C \in \mathcal{G}$ such that $A \bullet\!\!\rightarrow B \circ\!\!-\!\!\bullet C$. Then there is an edge between A and C in $\mathcal{G}$ that is not of the form $A \leftarrow C$. Moreover, if $A \rightarrow B \circ\!\!-\!\!\bullet C$ is in $\mathcal{G}$, then the edge between A and C is also not of the form $A \leftrightarrow C$.*

Proof of Lemma C.2.4. First, there must be an edge between A and C. Otherwise, the circle at $B \circ\!\!-\!\!\bullet C$ will be invariant by R1. The edge between A and C cannot be of the form $A \leftarrow C$, since that would imply that orientations in $\mathcal{G}$ are not complete with respect to R2. Similarly, if the edge between A and B in $\mathcal{G}$ is $A \rightarrow B$, then due to R2, $A \leftrightarrow C$ is also not in $\mathcal{G}$. □

C.3 Supplement to Section 4.3

C.3.1 Omitted Proofs from Section 4.3

Proof of Theorem 4.3.1. To begin, we note that p forms a collider path. If p is a minimal collider path, then $\langle Q_k, B, C\rangle$ will be a collider in every $\mathcal{M}^*$ that is Markov equivalent to $\mathcal{M}$ (Theorem 2.1 of Zhao et al., 2005). Thus, suppose for the rest of the proof that p is not a minimal collider path. Then there is a subsequence $p' = \langle A = Q_{n_0}, Q_{n_1}, \ldots, Q_{n_m}, B, C\rangle$ of p, such that p' is a minimal collider path and if $m > 0$, $\{Q_{n_j}\}_{j=1}^{m} \subset \{Q_i\}_{i=1}^{k}$. Note that B must be in p' as $Q_i \to C$ for all i by definition of discriminating path.

There are two possibilities for p':

(i) $n_m = k$: Then, by Theorem 2.1 of Zhao et al. (2005) (see Theorem C.1.2), $\langle Q_k, B, C\rangle$ forms a collider in every MAG $\mathcal{M}^*$ that is Markov equivalent to $\mathcal{M}$.

(ii) $n_m \neq k$: Then, we have $Q_{n_m} \bullet\!\!\to B \leftarrow\!\!\bullet C$. By Theorem C.1.2, we have $Q_{n_m} \bullet\!\!\to B \leftarrow\!\!\bullet C$ in every $\mathcal{M}^*$ that is Markov equivalent to $\mathcal{M}$. So, we now only need to show that $Q_k \bullet\!\!\to B$ in every $\mathcal{M}^*$ that is Markov equivalent to $\mathcal{M}$.

For the sake of contradiction, assume that there is at least one MAG Markov equivalent to $\mathcal{M}$ that does not contain $Q_k \bullet\!\!\to B$. Therefore, in the essential ancestral graph $\mathcal{P}$ of $\mathcal{M}$, we have $Q_k \bullet\!\!\circ B \leftarrow\!\!\bullet C$. Then by Lemma A.1 of Zhang (2008b) (see Lemma C.1.7), the edge between Q_k and C of the form $C \bullet\!\!\to Q_k$. This is a contradiction to the assumption that p is a discriminating path in $\mathcal{M}$ as that implies that Q_k is a parent of C. Therefore, we must have $Q_k \bullet\!\!\to B \leftarrow\!\!\bullet C$ in the essential ancestral graph $\mathcal{P}$ and therefore, $\langle Q_k, B, C\rangle$ will be a collider in every $\mathcal{M}^*$ that is Markov equivalent to $\mathcal{M}$.

□

Proof of Theorem 4.3.2. Observe that if we had used the original Zhang-R4 with the arrow completing part, the result of the theorem follows immediately from Theorem C.1.2 and completeness result of Zhang (2008b). So, it is sufficient to show that we do not need

the arrow completing part of Zhang-R4, therefore giving us only the tail completing part described in Zhao-R4.

It is sufficient to show that all colliders on discriminating paths in $\mathcal{M}$ are also colliders in $\mathcal{G}$ (as this directly implies that arrowheads oriented by Zhang-R4 are already complete). We will use the following notation: If $p_{\mathcal{M}}$ is a path in $\mathcal{M}$, then $p_{\mathcal{G}}$ denotes the corresponding path in $\mathcal{G}$. Consider a discriminating path $p_{\mathcal{M}} = \langle X = Q_0, Q_1, \dots, Q_k, Y \rangle$, $k \geq 2$ in $\mathcal{M}$ such that Q_k is a collider on $p_{\mathcal{M}}$.

If $p_{\mathcal{M}}$ is a minimal collider path, then Q_k is a collider on $p_{\mathcal{G}}$ and we are done. Hence, for the rest of the proof suppose that $p_{\mathcal{M}}$ is not a minimal collider path, and let $p'_{\mathcal{M}}$ be a subsequence of $p_{\mathcal{M}}$ that forms a minimal collider path in $\mathcal{M}$.

Since $X \notin \mathrm{Adj}(Y, \mathcal{M})$ and since $Q_i \to Y$ is in $\mathcal{M}$ for all $i \in \{1, \dots, k-1\}$ it follows that $p'_{\mathcal{M}}$ is of the form $p'_{\mathcal{M}} = \langle X = Q_{n_0}, Q_{n_1}, \dots, Q_{n_\ell}, Q_k, Y \rangle$, $\ell \geq 0$. Let $p'_{\mathcal{G}}$ be the corresponding minimal collider path in $\mathcal{G}$. Hence, $Q_k \leftarrow\!\bullet Y$ is in $\mathcal{G}$, by Algorithm 4.

Therefore, none of the conclusions made by completing the Zhao-R4 in Algorithm 4 are incorrect. Furthermore, this implies that all arrowhead and tail marks in $\mathcal{G}$ are also in $\mathcal{M}$, meaning that $\mathcal{G}$ is an ancestral partial mixed graph that does not contain any inducing paths.

We now only need to show that $Q_{k-1} \bullet\!\!\to Q_k$ is in $\mathcal{G}$. Note that since $Q_{k-1} \leftrightarrow Q_k$ is in $\mathcal{M}$, the only options are that $Q_{k-1} \bullet\!\!\to Q_k$ or $Q_{k-1} \bullet\!\!-\!\!\circ Q_k$ are in $\mathcal{G}$. Furthermore, note that if $Q_{k-1} \bullet\!\!-\!\!\circ Q_k$ is in $\mathcal{G}$, then since $p'_{\mathcal{G}} = \langle X = Q_{n_0}, Q_{n_1}, \dots, Q_{n_\ell}, Q_k, Y \rangle$ is a minimal collider path in $\mathcal{G}$ Lemma C.3.1 would imply that $Y \bullet\!\!\to Q_{k-1}$ is in $\mathcal{G}$. However, this would contradict that $Y \leftarrow Q_{k-1}$ is in $\mathcal{M}$. Hence, $Q_{k-1} \bullet\!\!\to Q_k$ is in $\mathcal{G}$. □

Now, we state and prove an important result in Lemma C.3.1. This is key to many results in the main text. It was already used to prove Theorem 4.3.2 and will later be used in arguing Lemma 4.4.9.

Lemma C.3.1. *Let $\mathcal{G}$ be an ancestral partial mixed graph that does not contain inducing paths and such that orientations in $\mathcal{G}$ are closed under R1-R3, Zhao-R4. Furthermore, let X and Y be distinct nodes in $\mathcal{G}$ such that $X \notin \mathrm{Adj}(Y, \mathcal{G})$. Suppose that there is a minimal*

collider path $p = \langle X = Q_{l_k}, \ldots, Q_{l_1} Q, Q_{r_1}, \ldots, Q_{r_m} = Y \rangle$, $k, m \geq 1$, $k + m > 2$ *in* $\mathcal{G}$ *and a node* W *not on* p *such that* $W \bullet\!\!-\!\!\circ Q$ *is in* $\mathcal{G}$. *Then the following hold:*

(i) Either $X \bullet\!\!\rightarrow W$ *is in* $\mathcal{G}$, *or there is an* $i \in \{1, \ldots, k-1\}$ *such that* $Q_{l_i} \leftrightarrow W$ *is in* $\mathcal{G}$.

(ii) Either $Y \bullet\!\!\rightarrow W$ *is in* $\mathcal{G}$, *or there is an* $j \in \{1, \ldots, m-1\}$ *such that* $Q_{r_j} \leftrightarrow W$ *is in* $\mathcal{G}$.

Proof of Lemma 4.3.3. Note that p is of the form $P_1 \bullet\!\!\rightarrow P_2 \leftrightarrow \cdots \leftrightarrow P_{k-1} \leftarrow\!\!\bullet P_k$. If $P_{i-1} \notin \mathrm{Adj}(P_{i+1}, \mathcal{G})$, then we are in case (i) and we are done. Otherwise, $P_{i-1} \in \mathrm{Adj}(P_{i+1}, \mathcal{G})$, so by Lemma C.3.3, we have that either $P_{i-1} \rightarrow P_{i+1}$ or $P_{i-1} \leftarrow P_{i+1}$ is in $\mathcal{G}$.

Assume without loss of generality that $P_{i-1} \rightarrow P_{i+1}$ is in $\mathcal{G}$. We will show that in this case, we end up having a discriminating collider path for P_i of the form in (ii). If $P_{i-1} \leftarrow P_{i+1}$ was in $\mathcal{G}$, an analogous argument can be used to show the existence of a discriminating collider path for P_i of the form in (iii).

Since $P_{i-1} \rightarrow P_{i+1}$ is in $\mathcal{G}$, by Lemma C.3.3, we have that $i - 1 \neq 1$, that is $i > 2$.

If $i = 3$, we have that $P_{i-2} = P_1$. Since we also know that $P_{i-2} \bullet\!\!\rightarrow P_{i-1} \rightarrow P_{i+1}$ is in $\mathcal{G}$ and we know that $\mathcal{G}$ is an ancestral graph it follows that if there is any edge between P_{i-2} and P_{i+1} it would need to be of the form $P_{i-2} \bullet\!\!\rightarrow P_{i+1}$ which contradicts Lemma C.3.3. Therefore, $P_{i-2} \notin \mathrm{Adj}(P_{i+1}, \mathcal{G})$ and $p(P_{i-2}, P_{i+1})$ is a discriminating collider path of the form (ii).

If $i > 3$, and $P_{i-2} \notin \mathrm{Adj}(P_{i+1}, \mathcal{G})$ we also have that $p(P_{i-2}, P_{i+1})$ is a discriminating collider path of the form (ii). Otherwise, $P_{i-2} \in \mathrm{Adj}(P_{i+1}, \mathcal{G})$, and Lemma C.3.3 and ancestrality of $\mathcal{G}$ imply that $P_{i-2} \rightarrow P_{i+1}$ is in $\mathcal{G}$.

If $P_{i-3} \notin \mathrm{Adj}(P_{i+1}, \mathcal{G})$, we can now obtain analogously to above that $p(P_{i-3}, P_{i+1})$ is a discriminating collider path of the form (ii). Otherwise, $P_{i-3} \rightarrow P_{i+1}$ is in $\mathcal{G}$ and we can continue this argument moving backward through nodes on p until we reach a first node $P_l, l \geq 1$, not adjacent to P_{i+1}. All the nodes $P_{l+1}, \rightarrow P_{i-1}$ have an directed edge pointing into P_{i+1} and $p(P_l, P_{i+1})$ will be a a discriminating collider path of the form (ii). Such a node will definitely exist, since if $P_2 \rightarrow P_{i+1}, \ldots, P_{i-1} \rightarrow P_{i+1}$ are all in $\mathcal{G}$, then $P_1 \notin \mathrm{Adj}(P_{i+1}, \mathcal{G})$, otherwise either $\mathcal{G}$ is not ancestral, or we have a contradiction with Lemma C.3.3. □

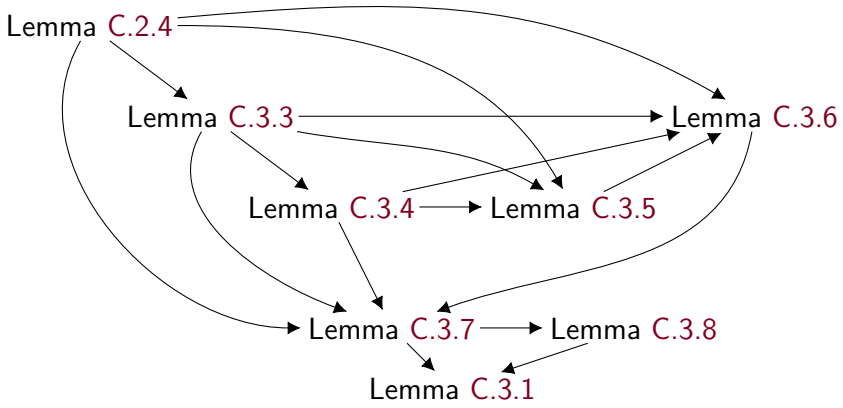


Figure C.1: Proof structure of Lemma C.3.1

C.3.2 Proof of Lemma C.3.1

Figure C.1 includes the proof structure for Lemma C.3.1.

Proof of Lemma C.3.1. First note, that by Lemma C.3.7, we have that $Q_{l_1} \bullet\!\!\rightarrow W \leftarrow\!\!\bullet Q_{r_1}$. The claim then follows by iterative application of Lemma C.3.8. □

Definition C.3.2 (Distance to $\mathbf{Z}$; cf. Zhang, 2006, Perković et al., 2018)**.** *Let p be a path in a partial mixed graph $\mathcal{G}$ and $\mathbf{Z}$ a node set in $\mathcal{G}$. Suppose that every node on $p = \langle V_1, \dots V_k \rangle$ is in* $\text{PossAn}(\mathbf{Z}, \mathcal{G})$*. Then the distance to $\mathbf{Z}$ for each node V_i, $i \in \{1, \dots, k\}$ on p is the length of a shortest possibly causal path from V_i to $\mathbf{Z}$. The distance to $\mathbf{Z}$ for the entire path p is equal to the sum of the distances to $\mathbf{Z}$ for each node on p.*

The following Lemma is similar to Lemma 2.1 of Zhao et al. (2005).

Lemma C.3.3. *Let $p = \langle X = Q_0, Q_1, \dots, Q_k, Q_{k+1} = Y \rangle, k > 1$ be a minimal collider path in an ancestral partial mixed graph that $\mathcal{G}$ such that the edge mark orientations in $\mathcal{G}$ are closed under R1, R2, Zhao-R4. Then the following hold*

(i) If edge $\langle Q_i, X \rangle$ is in $\mathcal{G}$ for some $i \in \{2, \dots, k\}$, then this edge is of the form $Q_i \rightarrow X$.

(ii) *If edge* $\langle Q_i, Y\rangle$ *is in* $\mathcal{G}$ *for some* $i \in \{1, \dots, k-1\}$*, then it is of the form* $Q_i \to Y$*.*

(iii) *If edge* $\langle Q_i, Q_j\rangle$ *is in* $\mathcal{G}$ *for some* $i, j \in \{1, \dots, k-1\}, i < j-1$*, then this edge is either* $Q_i \to Q_j$ *or* $Q_j \to Q_i$*.*

(iv) *If* $\langle Q_i, Q_j\rangle$ *and* $\langle Q_i, Q_{j+1}\rangle$ *are edges in* $\mathcal{G}$ *for some* $i, j \in \{1, \dots, k-1\}, i \neq j$*, then these edges are either* $Q_j \to Q_i \leftarrow Q_{j+1}$*, or* $Q_j \leftarrow Q_i \to Q_{j+1}$ *in* $\mathcal{G}$*.*

Proof of Lemma C.3.3. (i)-(ii): We only prove the claim for $\langle Q_i, X\rangle$, since the proof for $\langle Q_i, Y\rangle$ is analogous. Note that the edge $\langle Q_i, X\rangle$ cannot be of the form $Q_i \leftarrow\!\bullet X$, since in this case, p is not a minimal collider path in $\mathcal{G}$. Hence, we only need to show that this edge is also not of the form $Q_i \circ\!\!-\!\!\bullet X$.

Since $Q_{k+1} = Y$ is not adjacent to X in $\mathcal{G}$, there is at least one node on $p(Q_{i+1}, Q_{k+1})$ that is not adjacent to X. Let $Q_r, i < r \leq k+1$ be the closest node to Q_i on $p(Q_i, Q_{k+1})$ such that $Q_r \notin \text{Adj}(X, \mathcal{G})$. Then $Q_j \in \text{Adj}(X, \mathcal{G})$ for all $j \in \{i, \dots, r-1\}$. Additionally, $Q_j \leftarrow\!\bullet X$ is not in $\mathcal{G}$ for any $j \in \{i, \dots, r-1\}$ as that would contradict that p is a minimal collider path. If $Q_{r-1} \circ\!\!-\!\!\bullet X$ was in $\mathcal{G}$, $Q_r \bullet\!\!\to Q_{r-1} \circ\!\!-\!\!\bullet X$ and $Q_r \notin \text{Adj}(X, \mathcal{G})$ would contradict that orientations in $\mathcal{G}$ are completed according to R1. Hence, $X \leftarrow Q_{r-1}$ is in $\mathcal{G}$.

If $i = r-1$ we are done. Otherwise, consider the path $Q_r \bullet\!\!\to Q_{r-1} \leftrightarrow Q_{r-2}$ and edge $Q_{r-1} \to X$ in $\mathcal{G}$. Since orientations in $\mathcal{G}$ are completed by Zhao-R4 and since $Q_{r-2} \in \text{Adj}(X, \mathcal{G})$, $X \leftarrow Q_{r-2}$ is in $\mathcal{G}$. We can apply this same reasoning iteratively for all (if any) remaining $j \in \{i, \dots, r-2\}$ to show that $Q_j \to X$ is in $\mathcal{G}$.

(iii): Since p is a minimal collider path in $\mathcal{G}$, it is clear that $Q_i \leftrightarrow Q_j$ is not in $\mathcal{G}$. Hence, we only need to show that $Q_i \circ\!\!-\!\!\bullet Q_j$ and $Q_i \bullet\!\!-\!\!\circ Q_j$ are not in $\mathcal{G}$. We will do this by contradiction.

Suppose first that $Q_i \circ\!\!-\!\!\bullet Q_j$ is in $\mathcal{G}$. Since $i \geq 1$, $Q_{i-1} \bullet\!\!\to Q_i$ is in $\mathcal{G}$. Hence, by Lemma C.2.4, $Q_{i-1} \to Q_j$, $Q_{i-1} \circ\!\!-\!\!\bullet Q_j$ or $Q_{i-1} \leftarrow\!\!\circ Q_j$ is in $\mathcal{G}$. Then if $i = 1$, by (i) above, we immediately reach a contradiction.

If $Q_{i-1} \to Q_j$, or $Q_{i-1} \circ\!\!-\!\!\bullet Q_j$ is in $\mathcal{G}$, then consider that $Q_{i-2} \bullet\!\!\to Q_{i-1}$ is also in $\mathcal{G}$, and

since orientations in $\mathcal{G}$ are closed under R1 and Zhao-R4 it follows that $\langle Q_{i-2}, Q_j\rangle$ must be in $\mathcal{G}$. Similarly, by the ancestral property of $\mathcal{G}$ and by Lemma C.2.4, $Q_{i-2} \leftarrow Q_j$ is not in $\mathcal{G}$. Hence, by (i) $X \neq Q_{i-2}$, that is $i > 2$ and $Q_{i-2} \leftarrow\!\circ Q_j$, $Q_{i-2} \rightarrow Q_j$, or $Q_{i-2} \circ\!\!-\!\!\bullet Q_j$ is in $\mathcal{G}$.

If $Q_{i-1} \leftarrow\!\circ Q_j$ is in $\mathcal{G}$, then by (ii), $Q_j \neq Y$ and hence, $j < k+1$. Therefore, in this case, we can consider that $Q_{i-1} \leftarrow\!\circ Q_j \leftarrow\!\bullet Q_{j+1}$ implies by Lemma C.2.4 that edge $\langle Q_{i-1}, Q_{j+1}\rangle$ is in $\mathcal{G}$ and it is not of the form $Q_{i-1} \rightarrow Q_{j+1}$. Hence, by (ii), $Y \neq Q_{j+1}$, that is $j < k$, and $Q_{i-1} \leftarrow\!\circ Q_{j+1}$, $Q_{i-1} \leftarrow Q_{j+1}$, or $Q_{i-1} \circ\!\!-\!\!\bullet Q_{j+1}$ is in $\mathcal{G}$.

Next, we can apply the same reasoning as above to conclude that $i > 3$, and or $j < k-1$, and so forth. Since $i < j$, we will eventually run into a contradiction.

Analogously we can derive a contradiction when assuming that $Q_i \bullet\!\!-\!\!\circ Q_j$ is in $\mathcal{G}$. Hence, $Q_i \rightarrow Q_j$, or $Q_i \leftarrow Q_j$ are in $\mathcal{G}$.

(iv): This case follows for the fact that $\mathcal{G}$ is ancestral and (i)-(iii). □

Lemma C.3.4. *Let $\mathcal{G}$ be an ancestral partial mixed graph that does not contain inducing paths. Furthermore, suppose that the edge orientations in $\mathcal{G}$ are closed under rules R1, R2, and Zhao-R4, and let $p = \langle X = Q_0, Q_1, \ldots, Q_k, Q_{k+1} = Y\rangle, k > 2$ be a minimal collider path in $\mathcal{G}$. Then the following hold*

(i) For any subpath $p(Q_i, Q_j), 0 \leq i < j-1 \leq k$, there is at least one non-endpoint node Q_l, $l \in \{i+1, \ldots, j-1\}$ such that $Q_l \notin \mathrm{An}(\{Q_i, Q_j\}, \mathcal{G})$.

(ii) There is at least one unshielded triple on p.

(iii) Suppose that there is an edge $Q_i \rightarrow Q_j$, $i, j \in \{1, \ldots, k+1\}$, $i < j$ in $\mathcal{G}$. Then there is a node Q_l, $0 \leq l < i$, such that $Q_l \notin \mathrm{Adj}(Q_j, \mathcal{G})$ and $Q_{l_1} \rightarrow Q_j$ is in $\mathcal{G}$ for all $l_1 \in \{l+1, \ldots, i\}$.

Proof of Lemma C.3.4. (i): Suppose for a contradiction that there is a subpath $p(Q_i, Q_j)$, of p such that for all $l \in \{i+1, \ldots, j-1\}$, $Q_l \in \mathrm{An}(\{Q_i, Q_j\}, \mathcal{G})$. Since there are no inducing paths in $\mathcal{G}$, $Q_i \in \mathrm{Adj}(Q_j, \mathcal{G})$. Then by Lemma C.3.3, either $Q_i \rightarrow Q_j$, or $Q_i \leftarrow Q_j$ is in

$\mathcal{G}$. However both options, $Q_i \to Q_j \bullet\!\!\to Q_{j-1} \to \cdots \to Q_i$ or $Q_j \to Q_i \bullet\!\!\to Q_{i+1} \to \cdots \to Q_j$ contradict that $\mathcal{G}$ is an ancestral graph.

(ii): Suppose for a contradiction that every consecutive triple on p is shielded. Then by Lemma C.3.3 it follows that $Q_0 \leftarrow Q_2$ and $Q_{k-1} \to Q_{k+1}$ is in $\mathcal{G}$. Then $k > 2$ otherwise, we immediately reach a contradiction with (i).

Since $Q_1 \leftrightarrow Q_2 \leftrightarrow Q_3$ is a shielded triple, it follows that $Q_1 \leftarrow Q_3$ or $Q_1 \to Q_3$ is in $\mathcal{G}$ (Lemma C.3.3). However, since $Q_0 \leftarrow Q_2$ is also in $\mathcal{G}$ (and since $\mathcal{G}$ does not contain inducing paths) by (i), we can conclude that $Q_1 \leftarrow Q_3$ is in $\mathcal{G}$. In fact, by applying this argument iteratively to consecutive shielded triples on p, we can show that $Q_{k-2} \leftarrow Q_k$ is in $\mathcal{G}$. But now we have that subpath $p(Q_{k-2}, Q_{k+1})$ contradicts (i).

(iii): Suppose for contradiction, that $\langle Q_{l_1}, Q_j \rangle$ is in $\mathcal{G}$ for all $l_1 \in \{0, \ldots, i\}$. Note that (iv) and (i) in Lemma C.3.3 then implies that $Q_{l_1} \to Q_j$ for all $l_1 \in \{1, \ldots, i\}$ and $Q_0 \leftarrow Q_j$ respectively. However, $Q_0 \leftarrow Q_j \leftarrow Q_1 \leftrightarrow Q_0$ contradicts that $\mathcal{G}$ is ancestral.

Hence, there is a node Q_l, $0 \leq l < i$, such that $Q_l \notin \text{Adj}(Q_j, \mathcal{G})$. Then (iv) in Lemma C.3.3 implies that $Q_{l_1} \to Q_j$ for all $l_1 \in \{l+1, \ldots, i\}$. □

Lemma C.3.5. *Let $\mathcal{G}$ be an ancestral partial mixed graph that does not contain inducing paths. Furthermore, suppose that the edge orientations in $\mathcal{G}$ are closed under rules R1 - R3, and Zhao-R4. Let X and Y be distinct nodes in $\mathcal{G}$ such that $X \notin \text{Adj}(Y, \mathcal{G})$. Suppose that there is a minimal collider path $p = \langle X = Q_{l_k}, \ldots, Q_{l_1}, Q = Q_{r_0}, Q_{r_1}, \ldots, Q_{r_m} = Y \rangle$, $m \geq 1$, $k > 1$, and a node W not on p such that*

- *$W \circ\!\!-\!\!\circ Q$, and*
- *$W \circ\!\!-\!\!\bullet Q_{r_1}$, and*
- *$W \leftarrow\!\!\circ Q_{l_i}$, or $W \leftarrow Q_{l_i}$, for $i \in \{1, \ldots, k_1\}$, $1 \leq k_1 \leq k-2$ and*
- *$W \bullet\!\!-\!\!\circ Q_{l_{k_1+1}}$ is in $\mathcal{G}$.*

Then

- $Q_{l_{k_1+2}} \in \mathrm{Adj}(W, \mathcal{G})$, *and* $W \circ\!\!-\!\!\bullet Q_{l_{k_1+2}}$ *is not in* $\mathcal{G}$.

Proof of Lemma C.3.5. Note that $Q_{l_{k_1+2}} \in \mathrm{Adj}(W, \mathcal{G})$ by Lemma C.2.4, since $Q_{l_{k_1+2}} \bullet\!\!\rightarrow Q_{l_{k_1+1}} \circ\!\!-\!\!\bullet W$ is in $\mathcal{G}$. Suppose for a contradiction that $W \circ\!\!-\!\!\bullet Q_{l_{k_1+2}}$ is in $\mathcal{G}$.

- Suppose first that $k_1 > 1$. If $W \circ\!\!-\!\!\circ Q_{l_{k_1+1}}$ is in $\mathcal{G}$, then $Q_{l_1} \bullet\!\!\rightarrow W \circ\!\!-\!\!\circ Q_{l_{k_1+1}}$ together with Lemmas C.2.4 and C.3.3 implies that $Q_{l_1} \rightarrow Q_{l_{k_1+1}}$ is in $\mathcal{G}$. By the same reasoning $Q_{l_i} \bullet\!\!\rightarrow W \circ\!\!-\!\!\circ Q$ implies that $Q_{l_i} \rightarrow Q$ is in $\mathcal{G}$, for $i \in \{2, \ldots, k_1\}$. However now, $p(Q_{l_{k_1+1}}, Q)$ contradicts (i) of Lemma C.3.4.

  Otherwise, $W \leftarrow\!\!\circ Q_{l_{k_1+1}}$ is in $\mathcal{G}$. But in this case, $Q_{l_1} \bullet\!\!\rightarrow W \circ\!\!-\!\!\bullet Q_{l_{k_1+2}}$ together with Lemmas C.2.4 and C.3.3 implies that $Q_{l_1} \rightarrow Q_{l_{k_1+2}}$ is in $\mathcal{G}$. By the same reasoning $Q_{l_i} \bullet\!\!\rightarrow W \circ\!\!-\!\!\circ Q$ implies that $Q_{l_i} \rightarrow Q$ is in $\mathcal{G}$, for $i \in \{2, \ldots, k_1 + 1\}$. However now, $p(Q_{l_{k_1+2}}, Q)$ contradicts (i) of Lemma C.3.4.

- Next, consider the case when $k_1 = 1$. Since having $Q \leftrightarrow Q_{l_1} \rightarrow W$ and $W \circ\!\!-\!\!\circ Q$ in $\mathcal{G}$ would contradict that orientations in $\mathcal{G}$ are completed under R2, we must have that $Q_{l_1} \circ\!\!\rightarrow W$ is in $\mathcal{G}$.

  Since $Q_{l_1} \circ\!\!\rightarrow W \circ\!\!-\!\!\bullet Q_{r_1}$ and $Q_{l_1} \circ\!\!\rightarrow W \circ\!\!-\!\!\bullet Q_{l_3}$ are in $\mathcal{G}$, Lemmas C.2.4 and C.3.3 imply that $Q_{l_1} \rightarrow Q_{r_1}$ and $Q_{l_1} \rightarrow Q_{l_3}$ are in $\mathcal{G}$.

  If $W \leftarrow\!\!\circ Q_{l_2}$ is in $\mathcal{G}$, then since $Q_{l_2} \circ\!\!\rightarrow W \circ\!\!-\!\!\circ Q$ is in $\mathcal{G}$, Lemmas C.2.4 and C.3.3 would lead us to conclude that $Q_{l_2} \rightarrow Q$ is in $\mathcal{G}$, making $p(Q_{l_3}, Q)$ contradict (i) of Lemma C.3.4. Alternatively, if $Q_{l_2} \circ\!\!-\!\!\circ W$ is in $\mathcal{G}$, then $Q_{l_2} \circ\!\!-\!\!\circ W \circ\!\!-\!\!\circ Q$, $Q_{l_2} \bullet\!\!\rightarrow Q_{l_1} \leftrightarrow Q$, and $Q_{l_1} \circ\!\!-\!\!\bullet W$, together with R3 and Lemma C.3.3, would imply that either $Q_{l_2} \rightarrow Q$, or $Q_{l_2} \leftarrow Q$ are in $\mathcal{G}$. Having both $Q_{l_2} \rightarrow Q$ and $Q_{l_1} \rightarrow Q_{l_3}$ in $\mathcal{G}$, would make $p(Q_{l_3}, Q)$ contradict (i) of Lemma C.3.4. Alternatively, having both $Q_{l_2} \leftarrow Q$ and $Q_{l_1} \rightarrow Q_{r_1}$ in $\mathcal{G}$, would make $p(Q_{l_2}, Q_{r_1})$ contradict (i) of Lemma C.3.4.

□

Lemma C.3.6. *Let $\mathcal{G}$ be an ancestral partial mixed graph that does not contain inducing paths and such that orientations in $\mathcal{G}$ are closed under R1-R3, Zhao-R4. Furthermore, let X and Y be distinct nodes in $\mathcal{G}$ such that $X \notin \text{Adj}(Y, \mathcal{G})$. Suppose that there is a minimal collider path $p = \langle X = Q_{l_k}, \ldots, Q_{l_1}, Q = Q_{r_0}, Q_{r_1}, \ldots, Q_{r_m} = Y \rangle$, $m \geq 1$, $k > 1$, and a node W not on p such that*

- $W \bullet\!\!-\!\!\circ Q$, *and*
- $W \circ\!\!-\!\!\bullet Q_{r_1}$, *and*
- $W \bullet\!\!-\!\!\circ Q_{l_i}$, *or* $W \leftarrow Q_{l_i}$, *for* $i \in \{1, \ldots, k_1\}$, $k_1 < k$ *and*
- $Q_{l_i} \rightarrow Q_{r_1}$ *are in* $\mathcal{G}$, *for* $i \in \{1, \ldots, k_1\}$, $k_1 < k$.

Then

- $Q_{l_{k_1+1}} \in \text{Adj}(W, \mathcal{G})$, *and* $W \rightarrow Q_{l_{k_1+1}}$ *is not in* $\mathcal{G}$, *and*
- $Q_{l_{k_1+1}} \rightarrow Q_{r_1}$ *is in* $\mathcal{G}$.

Proof of Lemma C.3.6. This proof is split into three cases (a)-(c) below.

(a) First, suppose that $Q_{l_{k_1}} \rightarrow W$ is in $\mathcal{G}$. Let $i_1 \in \{1, \ldots, k_1\}$ be the largest index such that $Q_{l_{i_1}} \circ\!\!-\!\!\bullet W$ is in $\mathcal{G}$. If such an index does not exist then let $i_1 = 0$ and $Q_{l_0} = Q$ since $Q_{l_0} \circ\!\!-\!\!\bullet W$ is in $\mathcal{G}$. Since $k > k_1$, we now have that $Q_{l_{k_1+1}} \bullet\!\!\rightarrow Q_{l_{k_1}} \leftrightarrow \cdots \leftrightarrow Q_{l_{i_1}} \circ\!\!-\!\!\bullet W$ is in $\mathcal{G}$. Furthermore, $Q_{l_{i'}} \rightarrow W$ is in $\mathcal{G}$ for all $i' \in \{i_1 + 1, \ldots, k_1\}$. Hence, since orientations in $\mathcal{G}$ are completed under Zhao-R4, $Q_{l_{k_1+1}} \in \text{Adj}(W, \mathcal{G})$. Furthermore, $Q_{l_{k_1+1}} \leftarrow W$ is not in $\mathcal{G}$ since $\mathcal{G}$ is ancestral. In fact, since orientations in $\mathcal{G}$ are completed under R2, $Q_{l_{k_1+1}} \bullet\!\!\rightarrow W$ is in $\mathcal{G}$. Now, $Q_{l_{k_1+1}} \bullet\!\!\rightarrow W \circ\!\!-\!\!\bullet Q_{r_1}$ implies that $Q_{l_{k_1+1}} \rightarrow Q_{r_1}$ is in $\mathcal{G}$, by Lemmas C.2.4 and C.3.3.

(b) Next, suppose that $Q_{l_{k_1}} \circ\!\!-\!\!\bullet W$ and $W \leftarrow\!\!\circ\, Q$ are in $\mathcal{G}$. Since $k > k_1$, and $Q_{l_{k_1+1}} \bullet\!\!\rightarrow Q_{l_{k_1}} \circ\!\!-\!\!\bullet W$ is in $\mathcal{G}$, Lemma C.2.4 implies that $Q_{l_{k_1+1}} \in \text{Adj}(W, \mathcal{G})$ and that $Q_{l_{k_1+1}} \leftarrow W$ is not in $\mathcal{G}$.

Note also that $Q_{l_{k_1+1}} \bullet\!\!-\!\!\circ W$ is not possible, since $Q_{l_{k_1+1}} \bullet\!\!-\!\!\circ W \leftarrow\!\!\circ\, Q$ would by Lemmas C.2.4 and C.3.3 imply that $Q_{l_{k_1+1}} \leftarrow Q$ thus, together with $Q_{l_i} \rightarrow Q_{r_1}$ for all $i \in \{1, \ldots, k_1\}$, making $p(Q_{l_{k_1+1}}, Q_{r_1})$ contradict (i) of Lemma C.3.4. Hence, $Q_{l_{k_1+1}} \bullet\!\!\rightarrow W \circ\!\!-\!\!\bullet Q_{r_1}$ is in $\mathcal{G}$ implying that $Q_{l_{k_1+1}} \rightarrow Q_{r_1}$ is also in $\mathcal{G}$ by Lemmas C.2.4 and C.3.3.

(c) Lastly, suppose that $Q_{l_{k_1}} \circ\!\!-\!\!\bullet W$ and $W \circ\!\!-\!\!\circ\, Q$ are in $\mathcal{G}$. Let $Q_{l_0} = Q$. As in the above cases, note that since $k > k_1$, $Q_{l_{k_1+1}} \bullet\!\!\rightarrow Q_{l_{k_1}} \circ\!\!-\!\!\bullet W$ is in $\mathcal{G}$. Therefore, Lemma C.2.4 implies that $Q_{l_{k_1+1}} \in \text{Adj}(W, \mathcal{G})$ and $Q_{l_{k_1+1}} \leftarrow W$ is not in $\mathcal{G}$. If $Q_{l_{k_1+1}} \bullet\!\!\rightarrow W$ is in $\mathcal{G}$, we can use exactly the same argument as in (b) to show that $Q_{l_{k_1+1}} \rightarrow Q_{r_1}$ is in $\mathcal{G}$.

Otherwise, $Q_{l_{k_1+1}} \bullet\!\!-\!\!\circ W$ is in $\mathcal{G}$. Suppose first that $k_1 = 1$. Then $Q_{l_2} \bullet\!\!-\!\!\circ W \circ\!\!-\!\!\circ Q$, $Q_{l_2} \bullet\!\!\rightarrow Q_{l_1} \leftrightarrow Q$, and $Q_{l_1} \circ\!\!-\!\!\bullet W$, together with R3 and Lemma C.3.3, imply that $Q_{l_2} \rightarrow Q$, or $Q_{l_2} \leftarrow Q$ is in $\mathcal{G}$. Since $Q_{l_2} \leftarrow Q$ together with $Q_{l_1} \rightarrow Q_{r_1}$ would imply that $p(Q_{l_2}, Q_{r_1})$ contradicts (i) of Lemma C.3.4, it must be that $Q_{l_2} \rightarrow Q$ is in $\mathcal{G}$. We can now apply R3 and Lemma C.3.3 to $Q_{l_2} \bullet\!\!-\!\!\circ W \circ\!\!-\!\!\bullet Q_{r_1}$, $Q_{l_2} \rightarrow Q \leftarrow\!\!\bullet Q_{r_1}$, and $Q \circ\!\!-\!\!\circ W$ to conclude that $Q_{l_2} \rightarrow Q_{r_1}$ must be in $\mathcal{G}$.

Next, suppose that $k_1 > 1$. Note, that if there is any edge $Q_{l_{i_1}} \circ\!\!\rightarrow W$, or $Q_{l_{i_1}} \rightarrow W$ in $\mathcal{G}$, for $i_1 \in \{1, \ldots, k_1 - 1\}$, we can construct a contradiction with Lemma C.3.5. Hence, all edges $\langle Q_{l_{i_1}}, W \rangle, i_1 \in \{0, \ldots, k_1 - 1\}$ must be of the form $Q_{l_{i_1}} \circ\!\!-\!\!\circ W$ in $\mathcal{G}$.

Note that $Q_{l_{k_1+1}} \bullet\!\!-\!\!\circ W \circ\!\!-\!\!\circ Q_{l_{k_1-1}}$ and $Q_{l_{k_1}} \circ\!\!-\!\!\bullet W$ with $Q_{l_{k_1+1}} \bullet\!\!\rightarrow Q_{l_{k_1}} \leftrightarrow Q_{l_{k_1-1}}$ and R3 imply that $Q_{l_{k_1+1}} \in \text{Adj}(Q_{l_{k_1-1}}, \mathcal{G})$. Due to Lemmas C.3.3, and C.3.4, this edge must be of the form $Q_{l_{k_1+1}} \rightarrow Q_{l_{k_1-1}}$.

Then $Q_{l_{k_1+1}} \bullet\!\!-\!\!\circ W \circ\!\!-\!\!\circ Q_{l_{k_1-2}}$, $Q_{l_{k_1}} \circ\!\!-\!\!\circ W$, and $Q_{l_{k_1+1}} \rightarrow Q_{l_{k_1-1}} \leftrightarrow Q_{l_{k_1-2}}$ are in $\mathcal{G}$. Hence, by R3, Lemma C.3.3, and Lemma C.3.4, $Q_{l_{k_1+1}} \rightarrow Q_{l_{k_1-2}}$ is in $\mathcal{G}$. Since $Q_{l_{i_1}} \circ\!\!-\!\!\circ W$ for all $i_1 \in \{0, \ldots, k_1\}$, we can keep iterating the above procedure until we get that

$Q_{l_{k_1+1}} \to Q$ is in $\mathcal{G}$. The conclusion that $Q_{l_{k_1+1}} \to Q_{r_1}$ is in $\mathcal{G}$ is in $\mathcal{G}$, then follows from the above paragraph.

□

Lemma C.3.7. *Let $\mathcal{G}$ be an ancestral partial mixed graph that does not contain inducing paths and such that orientations in $\mathcal{G}$ are closed under R1-R3, Zhao-R4. Furthermore, let X and Y be distinct nodes in $\mathcal{G}$ such that $X \notin \text{Adj}(Y, \mathcal{G})$. Suppose that there is a minimal collider path $p = \langle X = Q_{l_k}, \ldots, Q_{l_1}, Q = Q_{r_0}, Q_{r_1}, \ldots, Q_{r_m} = Y \rangle$, $k, m, \geq 1$, $m + k > 2$, and a node W not on p such that $W \bullet\!\!-\!\!\circ Q$ is in $\mathcal{G}$. Then edges $\langle Q_{l_1}, W \rangle$, and $\langle W, Q_{r_1} \rangle$ are in $\mathcal{G}$. Furthermore, both of these edges are into W.*

Proof of Lemma C.3.7. First note that edges $\langle Q_{l_1}, W \rangle$, and $\langle W, Q_{r_1} \rangle$ are in $\mathcal{G}$ by Lemma C.2.4. Furthermore, by the same lemma, neither $W \to Q_{l_1}$ nor $W \to Q_{r_1}$ is in $\mathcal{G}$. Hence, either at least one of the two edges $\langle Q_{l_1}, W \rangle$ and $\langle W, Q_{r_1} \rangle$ is into W in $\mathcal{G}$, or $Q_{l_1} \bullet\!\!-\!\!\circ W \circ\!\!-\!\!\bullet Q_{r_1}$ is in $\mathcal{G}$.

Suppose for a contradiction that at least one of the edges $Q_{l_1} \bullet\!\!-\!\!\circ W$ and $W \circ\!\!-\!\!\bullet Q_{r_1}$ is in $\mathcal{G}$ If $Q_{l_1} \bullet\!\!-\!\!\circ W \leftarrow\!\!\bullet Q_{r_1}$ is in $\mathcal{G}$ then $Q_{l_1} \leftarrow Q_{r_1}$ is in $\mathcal{G}$, by Lemmas C.2.4 and C.3.3. Hence, in this case, we either have that $m \neq 1$, or we have reached a contradiction with Lemma C.3.3. If $Q_{l_1} \bullet\!\!\to W \circ\!\!-\!\!\bullet Q_{r_1}$ is in $\mathcal{G}$, then $Q_{l_1} \to Q_{r_1}$ is in $\mathcal{G}$, by Lemmas C.2.4 and C.3.3. Hence, in this case, we either have that $k \neq 1$, or we have reached a contradiction with Lemma C.3.3.

Otherwise, if $Q_{l_1} \bullet\!\!-\!\!\circ W \circ\!\!-\!\!\bullet Q_{r_1}$ is in $\mathcal{G}$, then since $W \bullet\!\!-\!\!\circ Q$ and $Q_{l_1} \bullet\!\!\to Q \leftarrow\!\!\bullet Q_{r_1}$ are also in $\mathcal{G}$ and since orientations in $\mathcal{G}$ are closed under R3, it follows that $Q_{l_1} \in \text{Adj}(Q_{r_1}, \mathcal{G})$ is in $\mathcal{G}$. By Lemma C.3.3, $Q_{l_1} \to Q_{r_1}$, or $Q_{l_1} \leftarrow Q_{r_1}$ is in $\mathcal{G}$. If $k = 1$, then by the same lemma $Q_{l_1} \leftarrow Q_{r_1}$, and if $m = 1$, $Q_{l_1} \to Q_{r_1}$ otherwise, $m \neq 1 \neq k$ and both options are possible. Since $m + k > 2$, either $m > 1$, or $k > 1$.

For the rest of the proof we will assume that $W \circ\!\!-\!\!\bullet Q_{r_1}$ and $Q_{l_1} \to Q_{r_1}$ is in $\mathcal{G}$. The proof of the case when $Q_{l_1} \bullet\!\!-\!\!\circ W$ and $Q_{l_1} \leftarrow Q_{r_1}$ is exactly symmetric. We now split the proof into two parts and show a contradiction in each.

(a) There is no $i \in \{1, \ldots, k\}$ such that $Q_{l_i} \leftarrow\!\bullet W$ is in $\mathcal{G}$. In this case, $Q_{l_1} \circ\!\!-\!\!\bullet W$, or $Q_{l_1} \rightarrow W$ is in $\mathcal{G}$ and by assumption $Q_{l_1} \rightarrow Q_{r_1}$ and $W \circ\!\!-\!\!\bullet Q_{r_1}$ are also in $\mathcal{G}$. We can now use Lemma C.3.6 iteratively to show that $Q_{l_i} \circ\!\!-\!\!\bullet W$, or $Q_{l_i} \rightarrow W$ is in $\mathcal{G}$, for all $i \in \{1, \ldots, k\}$. Additionally, by the same lemma, we will also have that $Q_{l_i} \rightarrow Q_{r_1}$, for all $i \in \{1, \ldots, k\}$. Since $Q_{l_k} = X$, we now reach a contradiction with Lemma C.3.3.

(b) There is an $i \in \{1, \ldots, k\}$ such that $Q_{l_i} \leftarrow\!\bullet W$ is in $\mathcal{G}$, and $Q_{l_{i_1}}$ is the closest such node to Q on $p(X, Q)$. In this case, $Q_{l_{i_1}} \leftarrow\!\bullet W$ is in $\mathcal{G}$ and $Q_{l_i} \circ\!\!-\!\!\bullet W$ or $Q_{l_i} \rightarrow W$ is in $\mathcal{G}$, for all $i \in \{1, \ldots, i_1 - 1\}$. Furthermore, by Lemma C.3.6, $Q_{l_i} \rightarrow Q_{r_1}$ is in $\mathcal{G}$, for all $i \in \{1, \ldots, i_1\}$. Since $Q_{l_{i_1}} \leftrightarrow Q_{l_{i_1-1}} \rightarrow W$, or $Q_{l_{i_1}} \leftrightarrow Q_{l_{i_1-1}} \circ\!\!-\!\!\bullet W$, by the ancestral property of $\mathcal{G}$ and Lemma C.2.4, $Q_{l_{i_1}} \leftarrow W$ is not in $\mathcal{G}$. Hence, $Q_{l_{i_1}} \leftarrow\!\bullet W$ is either $Q_{l_{i_1}} \leftrightarrow W$ or $Q_{l_{i_1}} \leftarrow\!\!\circ W$.

Now, since $Q_{l_{i_1}} \rightarrow Q_{r_1}$ is in $\mathcal{G}$, either $i_1 = k$ and we have reached a contradiction with Lemma C.3.3, or by Lemma C.3.4, there is a node $Q_{l_{i_2}}$ on $p(X, Q_{l_{i_1}})$ such that $Q_{l_{i_2}} \notin \text{Adj}(Q_{r_1}, \mathcal{G})$, and $Q_{l_i} \rightarrow Q_{r_1}$, for all $i \in \{i_1, \ldots, i_2 - 1\}$. But in this case, we also have the path $p(Q_{l_{i_2}}, Q_{l_{i_1}}) \oplus \langle Q_{l_{i_1}}, W \rangle \oplus \langle W, Q_{r_1} \rangle$ which contradicts that orientations in $\mathcal{G}$ are completed under Zhao-R4.

□

Lemma C.3.8. *Let $\mathcal{G}$ be an ancestral partial mixed graph that does not contain inducing paths and such that orientations in $\mathcal{G}$ are closed under R1-R3, Zhao-R4. Furthermore, let X and Y be distinct nodes in $\mathcal{G}$ such that $X \notin \text{Adj}(Y, \mathcal{G})$. Suppose that there is a minimal collider path $p = \langle X = Q_{l_k}, \ldots, Q_{l_1} Q, Q_{r_1}, \ldots, Q_{r_m} = Y \rangle$, $k, m \geq 1$, $k + m > 2$ in $\mathcal{G}$ and a node W not on p such that*

- *$W \bullet\!\!\circ Q$ is in $\mathcal{G}$, and*
- *$Q_{l_i} \circ\!\!-\!\!\bullet W$ or $Q_{l_i} \rightarrow W$ is in $\mathcal{G}$ for $i \in \{1, \ldots, k_1\}$, $k_1 < k$.*

Then $Q_{l_{k_1+1}} \bullet\!\!\rightarrow W$ *is in* $\mathcal{G}$.

Proof. Let $Q_{l_0} \equiv Q$. Suppose first that $Q_{l_{k_1}} \rightarrow W$ is in $\mathcal{G}$ and let $i_1 \in \{0, \ldots, k_1 - 1\}$ be such that $Q_{l_{i_1}}$ is the closest node to $Q_{l_{k_1}}$ on $p(Q_{l_{k_1}}, Q)$ such that $Q_{l_{i_1}} \circ\!\!-\!\!\bullet W$ is in $\mathcal{G}$. Now, $Q_{l_i} \rightarrow W$ for all $i \in \{i_1 + 1, \ldots, k_1\}$ and $Q_{l_{k_1+1}} \bullet\!\!\rightarrow Q_{l_{k_1}} \leftrightarrow \cdots \leftrightarrow Q$ is in $\mathcal{G}$, so since orientations in $\mathcal{G}$ are closed under Zhao-R4, it follows that $Q_{l_{k_1+1}} \in \text{Adj}(W, \mathcal{G})$. Furthermore, since $\mathcal{G}$ is ancestral and $Q_{l_{k_1+1}} \bullet\!\!\rightarrow Q_{l_{k_1}} \rightarrow W$ is in $\mathcal{G}$, $Q_{l_{k_1+1}} \leftarrow W$ is not in $\mathcal{G}$. Additionally, since orientations in $\mathcal{G}$ are closed under R2, $Q_{l_{k_1+1}} \bullet\!\!-\!\!\circ W$ is also not in $\mathcal{G}$. Hence, $Q_{l_{k_1+1}} \bullet\!\!\rightarrow W$ is in $\mathcal{G}$.

Otherwise, $Q_{l_{k_1}} \circ\!\!\rightarrow W$ is in $\mathcal{G}$. Then directly by Lemma C.3.7, $Q_{l_{k_1+1}} \bullet\!\!\rightarrow W$ is in $\mathcal{G}$. □

C.4 Supplement to Section 4.4

Proof of Lemma 4.4.9. Suppose for contradiction that there is a possible inducing path in $\mathcal{P}$. Then there is also a minimal collider path in $\mathcal{P}$ that is a possible inducing path. Among all possible inducing paths that are minimal collider paths, choose a shortest path in $\mathcal{P}$ that has the shortest distance to its endpoints (Definition C.3.2). Let this path be $p = \langle X, Q_1, \dots Q_k, Y \rangle, k > 1$. Then $Q_i \in \text{PossAn}(\{X, Y\}, \mathcal{P})$ for all $i \in \{1, \dots, k\}$ and there is at least one $i \in \{1, \dots, k\}$ such that $Q_i \notin \text{An}(\{X, Y\}, \mathcal{P})$ (otherwise p is an inducing path).

Let Q_j, $j \in \{1, \dots, k\}$, be the closest node to X on p, such that $Q_j \notin \text{An}(\{X, Y\}, \mathcal{P})$ and suppose without loss of generality that $Q_j \in \text{PossAn}(Y, \mathcal{P})$. Hence, let $q = \langle Q_j = Q_{j,1}, Q_{j,2}, \dots, Q_{j,k_j} = Y \rangle, k_1 \geq 2$ be a shortest possibly directed path from Q_j to Y in $\mathcal{G}$. By Lemma C.2.2, q is then an unshielded possibly directed path. Hence, by Lemma C.3.3 (ii) on path p, $k_j > 2$ (otherwise, $Q_j \in \text{An}(Y, \mathcal{P})$). Furthermore, by Corollary C.2.3, q must start with edge $Q_j \circ\!\!-\!\!\bullet Q_{j,2}$ in $\mathcal{P}$. The contradiction now follows by applying Lemma C.3.1.

Now, by Lemma C.3.1, either $Y \bullet\!\!\rightarrow Q_{j,2}$ or there is some Q_{j_+}, $j_+ \in \{j+1, \dots, k\}$, such that $Q_{j_+} \leftrightarrow Q_{j,2}$. We cannot have $Y \bullet\!\!\rightarrow Q_{j,2}$ as that contradicts q being an unshielded possible directed path from Q_j to Y. Similarly, either $X \bullet Q_{j,2}$ or there is some Q_{j_-}, $j_- \in \{1, \dots, j-1\}$, such that $Q_{j_-} \leftrightarrow Q_{j,2}$. In the former case, consider the path $p_1 = \langle X, Q_{j,2}, Q_{j_+} \rangle \oplus p(Q_{j_+}, Y)$ and in the latter case, consider the path $p_2 = p(X, Q_{j_-}) \oplus \langle Q_{j_-}, Q_{j,2}, Q_{j_+} \rangle \oplus p(Q_{j_+}, Y)$. Either way, we now have a contradiction that p is a minimal collider path between X and Y with the shortest distance to its endpoints.

□

Proof of Corollary 4.4.10. Follows from Definition 4.4.5, the definition of a MAG and Lemma 4.4.9. □

Proof of Lemma 4.4.11. The fact that $[\mathcal{G}'] \subseteq [\mathcal{G}]$ follows directly from the definition of $[\mathcal{G}']$ and $[\mathcal{G}]$, and Theorem C.1.2. Suppose for a contradiction that $\mathcal{G}'$ is not maximal, that is, there is a possible inducing path in $\mathcal{G}'$. Then there is also a minimal collider path that is

a possible inducing path in $\mathcal{G}'$. The corresponding path in $\mathcal{G}$ must then also be a minimal collider path and a possible inducing path. But this contradicts that $\mathcal{G}$ is maximal (Corollary 4.4.10). □

C.5 Supplement to Section 4.5

Proof of Lemma 4.5.1. If no MAG is represented by $\mathcal{G}$, the theorem immediately holds. Hence, suppose that there is a MAG represented by $\mathcal{G}$.

We prove the theorem by contradiction while considering different possibilities for the orientation of the $A \bullet\!\!\bullet B$ edge. Hence, suppose for a contradiction that there is a MAG $\mathcal{M}$ represented by $\mathcal{G}$ that contains $A \leftarrow\!\!\bullet D$ and (i) $A \leftarrow\!\!\bullet B$, or (ii) $A \rightarrow B$.

(i) We immediately have the contradiction in this case, as $D \bullet\!\!\rightarrow A \leftarrow\!\!\bullet B$ is an unshielded collider in $\mathcal{M}$ that is not in $\mathcal{G}$. Hence, $\mathcal{M}$ cannot be represented by $\mathcal{G}$.

(ii) We assume that $A \rightarrow B$ and $A \leftarrow\!\!\bullet D$ are in $\mathcal{M}$. Then $D \rightarrow A$ cannot be in $\mathcal{M}$, as $C \rightarrow D \rightarrow A \rightarrow B \bullet\!\!\rightarrow C$ is either a directed or an almost directed cycle. Hence, $D \leftrightarrow A$ is in $\mathcal{M}$. Furthermore, using similar reasoning, $C \rightarrow D \leftrightarrow A \rightarrow B$ implies that $B \leftrightarrow C$ is in $\mathcal{M}$, and $C \rightarrow D \leftrightarrow A$ implies that $A \leftrightarrow C$. But this gives us an inducing path $D \leftrightarrow A \leftrightarrow C \leftrightarrow B$ in $\mathcal{M}$, which is a contradiction.

□

Proof of Theorem 4.5.2. If no MAG is represented by $\mathcal{G}$, the theorem immediately holds. Hence, suppose that there is a MAG represented by $\mathcal{G}$ and let $p = \langle V_1, \ldots, V_i \rangle$, and $q = \langle V_i, V_{i+1}, V_1 \rangle$. Suppose for a contradiction that there exists a MAG $\mathcal{M}$ represented by $\mathcal{G}$ such that $V_1 \rightarrow V_2$ is in $\mathcal{M}$.

Since $\mathcal{M}$ contains only those unshielded colliders already present in $\mathcal{G}$ and since p is an unshielded possibly directed path in $\mathcal{G}$, we will have that the path corresponding to p in $\mathcal{M}$ is of the form $V_1 \rightarrow V_2 \rightarrow \cdots \rightarrow V_i$. Hence, the paths corresponding to p and q in $\mathcal{M}$ an almost directed cycle, which is a contradiction with $\mathcal{M}$ being an ancestral graph. □

Proof of Theorem 4.5.6. If there is no MAG represented by $\mathcal{G}$ the theorem immediately holds. Hence, suppose that there is a MAG represented by $\mathcal{G}$.

Let $p_{\mathcal{M}}$ be the path in $\mathcal{M}$ corresponding to p in $\mathcal{G}$. Note that p is not a collider path. Moreover, there cannot be a subsequence of p that forms a collider path in $\mathcal{G}$ since that would require an edge of the form $Q_j \leftarrow\!\bullet Y$, $j \in \{0, \ldots, k\}$, and by choice of p there is no such edge in $\mathcal{G}$.

For the sake of contradiction, assume that there is a MAG $\mathcal{M}$ represented by $\mathcal{G}$ that contains $Q_k \leftarrow\!\bullet Y$. We will derive the contradiction by proving that there is a subsequence of $p_{\mathcal{M}}$ that forms a collider path from X to Y in $\mathcal{M}$. Hence, there is also a subsequence of $p_{\mathcal{M}}$ that forms a minimal collider path from X to Y, which ultimately gives us the contradiction with $\mathcal{M}$ being represented by $\mathcal{G}$ by Definition 4.4.1.

Note that since $Q_k \leftarrow\!\bullet Y \leftarrow Q_{k-1}$ is in $\mathcal{M}$, and since $\mathcal{M}$ is ancestral it follows that $Q_k \leftarrow\!\bullet Q_{k-1}$ is in $\mathcal{M}$, that is Q_k is a collider on $p_{\mathcal{M}}$. If the remaining non-endpoint nodes on $p_{\mathcal{M}}$ are colliders, then the contradiction is immediate. Otherwise, there is at least one non-endpoint node on $p_{\mathcal{M}}$ that is a non-collider. Let $\{Q_{k_1}, \ldots, Q_{k_m}\}$, $m \geq 1$ and $1 \leq k_i < k_j \leq k-1$, $1 \leq i < j \leq m$, be the non-colliders on $p_{\mathcal{M}}$. We will show how to "skip over" one or two of these non-colliders and construct a subsequence of $p_{\mathcal{M}}$ called $p^1_{\mathcal{M}}$ that has one fewer non-collider, or a subsequence of $p_{\mathcal{M}}$ called $p^2_{\mathcal{M}}$ that has two fewer non-colliders. This process can then be applied again on the obtained subsequence until we reach a subsequence of $p_{\mathcal{M}}$ called $p^m_{\mathcal{M}}$ that is a collider path, thereby deriving the contradiction.

Hence, let $i = k_j$. Since Q_i is a non-collider on $p_{\mathcal{M}}$, Q_i satisfies (i)(b), (i)(c), (ii)(b), (ii)(b), (iii)(b), or (iii)(c) of Definition 4.5.4 on p. We now discuss each of these cases and show how to construct $p^1_{\mathcal{M}}$.

(i)(b), that is $Q_0 \bullet\!\rightarrow Q_1 \circ\!\rightarrow Q_2$ and $Q_0 \bullet\!\!-\!\!\circ Q_2$ is in $\mathcal{G}$. Since Q_1 is a non-collider on $p_{\mathcal{M}}$, $Q_0 \bullet\!\rightarrow Q_1 \rightarrow Q_2$ is in $\mathcal{M}$. Additionally, since $\mathcal{M}$ is an ancestral graph, the edge between Q_0 and Q_2 is $Q_0 \bullet\!\rightarrow Q_2$. Hence, let $p^1_{\mathcal{M}} = \langle Q_0, Q_2 \rangle \oplus p_{\mathcal{M}}(Q_2, Y)$.

If Q_2 is a collider on both $p_{\mathcal{M}}$ and $p^1_{\mathcal{M}}$, then $p^1_{\mathcal{M}}$ has one fewer non-collider. If however, Q_2 is a non-collider on $p_{\mathcal{M}}$, then $Q_2 \rightarrow Q_3$ is on $p_{\mathcal{M}}$ as well. Therefore, $Q_1 \circ\!\rightarrow Q_2 \circ\!\rightarrow Q_3$ is on p. And by choice of p, $Q_1 \leftarrow\!\circ Q_3$ would need to be in $\mathcal{G}$. Then $Q_1 \rightarrow Q_2 \rightarrow Q_3$ and $Q_1 \leftarrow\!\bullet Q_3$ would imply that $\mathcal{M}$ is not ancestral, which is a contradiction.

(i)(c), that is $Q_0 \bullet\!\!-\!\!\circ Q_1 \leftarrow\!\!\bullet Q_2$ and $Q_0 \circ\!\!\rightarrow Q_2$ is in $\mathcal{G}$. Since Q_1 is a non-collider on $p_{\mathcal{M}}$, $Q_0 \leftarrow Q_1 \leftarrow\!\!\bullet Q_2$ is in $\mathcal{M}$. Additionally, since $\mathcal{M}$ is an ancestral graph, the edges between Q_0 and Q_2, and Q_1 and Q_2 must be $Q_0 \leftrightarrow Q_2$, $Q_1 \leftrightarrow Q_2$. Now, if Q_2 is a collider on $p_{\mathcal{M}}$, let as above $p^1_{\mathcal{M}} = \langle Q_0, Q_2 \rangle \oplus p_{\mathcal{M}}(Q_2, Y)$ and we are done.

Otherwise, Q_2 a non-collider on $p_{\mathcal{M}}$, meaning that $Q_0 \leftarrow Q_1 \leftrightarrow Q_2 \rightarrow Q_3$ is in $\mathcal{M}$. Consider what this implies in $\mathcal{G}$, we know that $Q_0 \bullet\!\!-\!\!\circ Q_1 \leftarrow\!\!\bullet Q_2$ is in $\mathcal{G}$ and we know that $Q_2 \rightarrow Q_3$ is in $\mathcal{M}$. By properties of p as an almost discriminating path, $Q_2 \circ\!\!\rightarrow Q_3$ must be in $\mathcal{G}$. This furthermore implies that $Q_1 \leftrightarrow Q_2 \circ\!\!\rightarrow Q_3$, and $Q_1 \leftarrow\!\!\circ Q_3$ is in $\mathcal{G}$. Hence, since $Q_0 \leftarrow Q_1 \leftrightarrow Q_2 \rightarrow Q_3$ is in $\mathcal{M}$, for $\mathcal{M}$ to be ancestral, $Q_1 \leftrightarrow Q_3$ is also in $\mathcal{M}$.

Therefore, we have that $Q_0 \leftrightarrow Q_2 \leftrightarrow Q_1 \leftrightarrow Q_3$, and $Q_2 \rightarrow Q_3$, $Q_1 \rightarrow Q_0$ are in $\mathcal{M}$. Now, since $\mathcal{M}$ is a maximal graph, edge $\langle Q_0, Q_3 \rangle$ is in $\mathcal{M}$. Furthermore, for $\mathcal{M}$ to be ancestral, it must be of the form $Q_0 \leftrightarrow Q_3$.

Now, there are two possibilities – either $Q_3 \leftarrow\!\!\bullet Q_4$ is on p, or $Q_3 \circ\!\!-\!\!\bullet Q_4$ and $Q_2 \leftarrow\!\!\circ Q_4$ are on p. In the first case, Q_3 is already a collider on p. In the second case, since we also have that $Q_2 \rightarrow Q_3$, for $\mathcal{M}$ to be ancestral it must be that $Q_3 \leftarrow\!\!\bullet Q_4$ is in $\mathcal{M}$. Therefore, Q_3 is collider on $p_{\mathcal{M}}$ regardless of its status on p. Hence, let $p^2_{\mathcal{M}} = \langle Q_0, Q_3 \rangle \oplus p_{\mathcal{M}}(Q_3, Y)$. Then $p^2_{\mathcal{M}}$ has two fewer non-colliders than $p_{\mathcal{M}}$.

(ii)(b), that is $Q_{i-1} \bullet\!\!\rightarrow Q_i \circ\!\!\rightarrow Q_{i+1}$, and $Q_{i-1} \leftarrow\!\!\circ Q_{i+1}$ are in $\mathcal{G}$ and $i \in \{2, \ldots, k-2\}$. Since Q_i is a non-collider on $p_{\mathcal{M}}$, $Q_{i-1} \bullet\!\!\rightarrow Q_i \rightarrow Q_{i+1}$ is in $\mathcal{M}$. Additionally, since $Q_{i-1} \bullet\!\!\rightarrow Q_i \rightarrow Q_{i+1}$, $\mathcal{M}$ is an ancestral graph, and $Q_{i-1} \leftarrow\!\!\circ Q_{i+1}$ is in $\mathcal{G}$, the edges between Q_{i-1} and Q_{i+1} and Q_{i-1} and Q_i are $Q_{i-1} \leftrightarrow Q_{i+1}$, $Q_{i-1} \leftrightarrow Q_i$.

Now, we know that $Q_{i-1} \leftrightarrow Q_i \rightarrow Q_{i+1}$ and $Q_{i-1} \leftrightarrow Q_{i+1}$ are in $\mathcal{M}$. First we show that Q_{i+1} is a collider on $p_{\mathcal{M}}$. Note that Q_{i+1} is either already a collider on p, or $Q_i \circ\!\!\rightarrow Q_{i+1} \circ\!\!-\!\!\bullet Q_{i+2}$ and $Q_i \leftarrow\!\!\circ Q_{i+2}$ are in $\mathcal{G}$. In the latter case, since $Q_i \rightarrow Q_{i+1}$ is in $\mathcal{M}$ and since $\mathcal{M}$ is ancestral, $Q_{i+1} \leftarrow\!\!\bullet Q_{i+2}$ is in $\mathcal{M}$. Hence, Q_{i+1} is a collider on $p_{\mathcal{M}}$.

Note that $Q_{i-1} \leftrightarrow Q_i$ is on $p_{\mathcal{M}}$, so if Q_{i-1} is also a collider on $p_{\mathcal{M}}$, let $p^1_{\mathcal{M}} = p^1_{\mathcal{M}}(X, Q_{i-1}) \oplus \langle Q_{i-1}, Q_{i+1} \rangle \oplus p_{\mathcal{M}}(Q_{i+1}, Y)$ and we are done.

Otherwise, Q_{i-1} is a non-collider on $p_{\mathcal{M}}$, so since $Q_{i-1} \leftrightarrow Q_i$ is in $\mathcal{M}$, it follows

that $Q_{i-2} \bullet\!\!\rightarrow Q_{i-1}$ cannot on p. Since p is an almost discriminating path it must be that $Q_{i-2} \bullet\!\!-\!\!\circ Q_{i-1} \leftrightarrow Q_i$ and $Q_{i-2} \circ\!\!\rightarrow Q_i$ are in $\mathcal{G}$. Then for Q_{i-1} to be a non-collider on $p_{\mathcal{M}}$, we have that $Q_{i-1} \leftarrow Q_{i-1} \leftrightarrow Q_i$ in $\mathcal{M}$, and since $\mathcal{M}$ is ancestral, and $Q_{i-2} \circ\!\!\rightarrow Q_i$ is in $\mathcal{G}$, $Q_{i-2} \leftrightarrow Q_i$ is in $\mathcal{M}$.

Consider that now we know that $Q_{i-2} \leftrightarrow Q_i \leftrightarrow Q_{i-1} \leftrightarrow Q_{i+1}$, $Q_i \rightarrow Q_{i+1}$ and $Q_{i-1} \rightarrow Q_{i-2}$ are in $\mathcal{M}$. Hence, since $\mathcal{M}$ is maximal $\langle Q_{i-2}, Q_{i+1} \rangle$ must also be in $\mathcal{M}$. Furthermore, since $\mathcal{M}$ is ancestral this edge between Q_{i-2} and Q_{i+1} is of the form $Q_{i-2} \leftrightarrow Q_{i+1}$.

If $i = 2$, let $p^2_{\mathcal{M}} = \langle Q_{i-2}, Q_{i+1} \rangle \oplus p_{\mathcal{M}}(Q_{i+1}, Y)$ and we are done. Otherwise, $i > 2$, so edge $Q_{i-2} \bullet\!\!-\!\!\circ Q_{i-1}$ is of the form $Q_{i-2} \leftarrow\!\!\circ Q_{i-1}$ on p. Furthermore, then either Q_{i-2} is a collider on p, or $Q_{i-3} \bullet\!\!-\!\!\circ Q_{i-2} \leftarrow\!\!\circ Q_{i-1}$ and $Q_{i-3} \circ\!\!\rightarrow Q_{i-1}$ is in $\mathcal{G}$. In the latter case, since $\mathcal{M}$ is an ancestral graph and since $Q_{i-2} \leftarrow Q_{i-1}$ is in $\mathcal{M}$, $Q_{i-3} \leftrightarrow Q_{i-2}$ and $Q_{i-3} \leftrightarrow Q_{i-1}$ are also in $\mathcal{M}$. Hence, under both options, we have that Q_{i-2} is a collider on $p_{\mathcal{M}}$. Hence, $p^2_{\mathcal{M}} = p_{\mathcal{M}}(X, Q_{i-2}) \oplus \langle Q_{i-2}, Q_{i+1} \rangle \oplus p_{\mathcal{M}}(Q_{i+1}, Y)$ is a subsequence of $p_{\mathcal{M}}$ with two fewer non-colliders.

(ii)(c), that is $Q_{i-1} \leftarrow\!\!\circ Q_i \leftarrow\!\!\bullet Q_{i+1}$, and $Q_{i-1} \circ\!\!\rightarrow Q_{i+1}$ are in $\mathcal{G}$ and $i \in \{2, \ldots, k-2\}$. This case is exactly symmetric to the case (ii)(b). Using a symmetric argument we can conclude that Q_{i-1} is always a collider on $p_{\mathcal{M}}$. Additionally, if Q_{i+1} is not a collider on $p_{\mathcal{M}}$, then Q_{i+2} will be a collider on $p_{\mathcal{M}}$. So we either show that $Q_{i-1} \leftrightarrow Q_{i+1}$ is in $\mathcal{M}$ and construct the path $p^1_{\mathcal{M}} = p_{\mathcal{M}}(X, Q_{i-1}) \oplus \langle Q_{i-1}, Q_{i+1} \rangle \oplus p_{\mathcal{M}}(Q_{i+1}, Y)$ with one fewer non-collider compared to $p_{\mathcal{M}}$, or show that $Q_{i-1} \leftrightarrow Q_{i+2}$ is in $\mathcal{M}$ and construct the path $p^2_{\mathcal{M}} = p_{\mathcal{M}}(X, Q_{i-1}) \oplus \langle Q_{i-1}, Q_{i+2} \rangle \oplus p_{\mathcal{M}}(Q_{i+2}, Y)$ with two fewer non-colliders.

(iii)(b), that is $Q_{k-2} \bullet\!\!\rightarrow Q_{k-1} \circ\!\!\rightarrow Q_k$, and $Q_{k-2} \leftarrow\!\!\circ Q_k$ is in $\mathcal{G}$. This case is symmetric to (i)(c) and holds by an analogous argument.

(iii)(c), that is $Q_{k-2} \leftarrow\!\!\circ Q_{k-1} \leftarrow\!\!\bullet Q_k$, and $Q_{k-2} \circ\!\!-\!\!\bullet Q_k$ is in $\mathcal{G}$. This case is symmetric to (i)(b) and holds by an analogous argument.

□

C.6 Supplement to: Section 4.6

In this appendix, we illustrate the technical details for Lemma 4.6.3 and Theorem 4.6.7. In Appendix C.6.1, we show the proof of Lemma 4.6.3. In Appendix C.6.4, we give an overview of the proof of Theorem 4.6.7 and discuss the main argument. We devote devote Appendix C and Appendix C for showing the auxiliary results that Theorem 4.6.7 depends upon.

C.6.1 Lemma 4.6.3

Proof of Lemma 4.6.3. Completeness of orientations with respect to R9 follows from Lemma C.6.1. Completeness of orientations with respect to R3 follows from the fact that we are adding consistent background knowledge to $\mathcal{G}$, which means we never elicit a new unshielded collider in $\mathcal{G}'$. For R3 to be invoked a new unshielded collider would be needed. □

Lemma C.6.1. *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be an ancestral partial mixed graph and $\mathcal{P}$ be an essential ancestral graph such that they have the same skeleton, share the same set of minimal collider paths, and the set of all invariant edge marks in $\mathcal{P}$ is a subset of the invariant edge marks in $\mathcal{G}$. Suppose furthermore that the edge mark at A on edge $A \circ\!\!\rightarrow C$ is in $\mathcal{G}$ and that there is an unshielded possbily directed path p, from A to C, $p = \langle A = P_1, P_2, \ldots, P_k = C \rangle$, $k > 3$ in $\mathcal{G}$. Then $P_2 \in \text{Adj}(C, \mathcal{G})$.*

Proof of Lemma C.6.1. Let p^* be the path in $\mathcal{P}$ that corresponds to p in $\mathcal{G}$. Since $\mathcal{P}$ and $\mathcal{G}$ have same skeleton, p^* is an unshielded path in $\mathcal{P}$. Furthermore, since $\mathcal{G}$ has additional edge orientations compared to $\mathcal{P}$, any possibly directed path in $\mathcal{G}$ corresponds to a possibly directed path in $\mathcal{P}$. Therefore, p^* is a possibly directed unshielded path in $\mathcal{P}$.

Suppose first that $A \circ\!\!\rightarrow C$ is in $\mathcal{P}$. Then $P_2 \in \text{Adj}(C, \mathcal{P}) \equiv \text{Adj}(C, \mathcal{G})$ because otherwise, orientations in $\mathcal{P}$ are not complete under R9. Next, suppose $A \circ\!\!-\!\!\circ C$ is in $\mathcal{P}$. Then Lemma C.1.4 and Corollary C.2.3 together imply that p^* is an unshielded path of the form $A \circ\!\!-\!\!\circ P_2 \circ\!\!-\!\!\circ \ldots \circ\!\!-\!\!\circ C$ in $\mathcal{P}$. Furthermore, since by assumption $|p^*| \geq 3$ and since $A \circ\!\!-\!\!\circ C$ we obtain a contradiction with Lemma Lemma C.1.6. □

C.6.2 Lemma 4.6.8

Proof of Lemma 4.6.8. Let $p_{\mathcal{G}} = \langle P_1, P_2, \ldots, P_k \rangle$ be a path in $\mathcal{G}$ that makes up a shortest directed or an almost directed cycle with edge $\langle P_1, P_k \rangle$. If $k = 3$, we are done. Hence, suppose for a contradiction that $k > 3$ and let $p_{\mathcal{P}}$ be the path in $\mathcal{P}$ that corresponds to path $p_{\mathcal{G}}$ in $\mathcal{P}$.

Note that $p_{\mathcal{G}}$ must be an unshielded path since due to the completion of orientations in $\mathcal{G}$ under R2 and R8, any shield $\langle P_i, P_{i+2} \rangle$ would imply the existence of a shorter directed or almost directed cycle in $\mathcal{G}$. Therefore, $p_{\mathcal{P}}$ is an unshielded path of length $k > 3$. Hence, it cannot be a circle path (Lemma C.1.6).

By Corollary C.2.3, it follows that $P_{k-1} \bullet\!\!\rightarrow P_k$ is in $\mathcal{P}$. Using the same reasoning as in the previous paragraph, we can also conclude that $P_2 \notin \text{Adj}(P_k, \mathcal{P})$ and that $P_1 \notin \text{Adj}(P_{k-1}, \mathcal{P})$. Since orientations in $\mathcal{P}$ are closed under R9, it therefore follows that we cannot have $P_1 \circ\!\!\rightarrow P_k$ in $\mathcal{P}$.

Hence $P_1 \leftarrow P_k$, or $P_1 \leftarrow\!\!\circ P_k$, or $P_1 \circ\!\!-\!\!\circ P_k$ is in $\mathcal{P}$. Since $P_1 \notin \text{Adj}(P_{k-1}, \mathcal{P})$, and since $\mathcal{P}$ is ancestral, Lemma C.1.8 implies that $P_1 \rightarrow P_2 \cdots \rightarrow P_k$ cannot be in $\mathcal{P}$. Hence $P_1 \circ\!\!-\!\!\bullet P_2$ is in $\mathcal{P}$.

But now $P_1 \circ\!\!-\!\!\bullet P_2$, $P_2 \notin \text{Adj}(P_k, \mathcal{P})$, and Lemma C.1.7, imply that $P_1 \leftarrow\!\!\bullet P_k$ is not in $\mathcal{P}$. Hence, $P_1 \circ\!\!-\!\!\circ P_k$ is in $\mathcal{P}$.

But now $P_1 \circ\!\!-\!\!\circ P_k$ and the path $p_{\mathcal{P}}$ from P_1 to P_k that does not contain $P_i \leftarrow\!\!\bullet P_{i+1}$, $i \in \{1, \ldots, k-1\}$ and ends with $P_{k-1} \bullet\!\!\rightarrow P_k$ contradict Lemma C.1.4.

□

C.6.3 Theorem 4.6.5 and Corollary 4.6.6

Proof of Corollary 4.6.6. This follows from Theorem 4.6.5 and from the fact that $\mathcal{K}$ is a set of background knowledge edge marks consistent with $\mathcal{G}$. □

Proof of Theorem 4.6.5. Consider the following procedure. First, we identify the circle component of $\mathcal{G} = (\mathbf{V}, \mathbf{E})$. This is the subgraph of $\mathcal{G}$ containing only $\circ\!\!-\!\!\circ$ edges, $\mathbf{E}_C$. Call this

$\mathcal{G}_C = (\mathbf{V}, \mathbf{E}_C)$. Consider the same edges present in $\mathcal{G}' = (\mathbf{V}, \mathbf{E}')$, which might potentially have different edge mark orientations, $\mathbf{E}'_C$. Note that by Corollary C.1.3, $\mathcal{G}_C = (\mathbf{V}, \mathbf{E}_C)$ is a collection of undirected connected chordal components $\mathcal{G}_{C_1}, \ldots \mathcal{G}_{C_k}$, $k \geq 1$, each of which is an induced subgraph of $\mathcal{G}$. We will refer to the corresponding induced subgraphs of $\mathcal{G}'$ as $\mathcal{G}'_{C_1}, \ldots, \mathcal{G}'_{C_k}$. Theorem C.7.12 tells us that each individual induced subgraph $\mathcal{G}'_{C_i}$, $i \geq 1$ of $\mathcal{G}'_C$ is a restricted essential ancestral graph. That is, each $\mathcal{G}'_{C_i}$ can be oriented into a MAG $\mathcal{M}_i$ with no minimal collider paths, and with the desired edge orientation of a particular edge $\langle A, B \rangle$.

Now, suppose we construct a new directed mixed graph $\mathcal{M} = (\mathbf{V}, \mathbf{E}_\mathcal{M})$ obtained by taking the union of all invariant edge marks in $\mathcal{G}'$ and $\mathcal{M}_i$ for all $i \in \{1, \ldots, k\}$. We will now show that $\mathcal{M}$ is a MAG represented by $\mathcal{G}'$. That is $\mathcal{M}$ is an ancestral graph with the same minimal collider paths as $\mathcal{G}'$ (Lemma 4.4.11). In particular, it suffices to show that there are no directed cycles or almost directed cycles in $\mathcal{M}$ that contain some edges from $\mathcal{M}_i$, $i \in \{1, \ldots, k\}$ and some edges from $\mathcal{G}'$ that are not in any $\mathcal{M}_i$, $i \in \{1, \ldots, k\}$, and also that there are no minimal collider paths in $\mathcal{M}$ that are made up of edges from $\mathcal{M}_i$, $i \in \{1, \ldots, k\}$ and edges outside of $\mathcal{M}_i$, $i \in \{1, \ldots, k\}$ that are in $\mathcal{G}'$.

First, we show that $\mathcal{M}$ is ancestral. By Lemma 4.6.8, it is enough to show that there are no directed cycles or almost directed cycles of length 3 in $\mathcal{M}$. For sake of contradiction, we will suppose that there is a triple $A \to B \to C$ and edge $A \leftarrow\!\bullet C$ in $\mathcal{M}$. Furthermore, since $\mathcal{G}'$ and $\mathcal{M}_i$, for all i are ancestral, and since $\mathcal{G}'_{C_i}$ are induced subgraphs of $\mathcal{G}'_C$ (Lemmas B.4 and B.8 of Zhang, 2008b), $\forall i$, exactly two of the nodes A, B, C are in $\mathcal{G}'_{C_j}$ for some $j \geq 1$. We consider the options below:

(a) Suppose that A, C are in $\mathcal{G}'_{C_j}$, and $B \notin \mathcal{G}_C$. Note again that $A \to B \to C$ and $A \leftarrow\!\bullet C$ is in $\mathcal{M}$. Furthermore, since $A, C \in \mathcal{G}'_{C_j}$ and $B \notin \mathcal{G}_C$, we have that $A \circ\!\!-\!\!\circ C$ is in $\mathcal{G}$, and also that $B \to C$ or $B \circ\!\!\to C$ is in $\mathcal{G}$. Now Lemma C.1.7, implies that $B \bullet\!\!\to A$ must have been in $\mathcal{G}$, which leads us to a contradiction.

(b) Suppose that A, B are in $\mathcal{G}'_{C_j}$, and $C \notin \mathcal{G}_C$. Again, consider that $A \to B \to C$ and

$A \leftarrow\!\bullet C$ are in $\mathcal{M}$. Therefore, similarly to above, we have that $A \circ\!\!-\!\!\circ B$ is in $\mathcal{G}$ and since $C \notin \mathcal{G}_C$, $A \leftarrow\!\bullet C$ is in $\mathcal{G}$. Hence, we obtain a contradiction with Lemma C.1.7 as in the previous case.

(c) Suppose that B, C are in $\mathcal{G}'_{C_j}$, and $A \notin \mathcal{G}_C$. Now again $A \to B \to C$ and $A \leftarrow\!\bullet C$ are in $\mathcal{M}$. Now, $C \to A \to B$ or $C \leftrightarrow A \to B$ are in $\mathcal{G}'$. So since edge mark orientations in $\mathcal{G}'$ are closed under R2, the edge $\langle C, B\rangle$ must have an arrowhead at B in $\mathcal{G}'$. But, this contradicts that $B \to C$ is in $\mathcal{M}$.

Therefore, $\mathcal{M}$ is ancestral. It remains to prove that $\mathcal{M}$ has the same minimal collider paths as $\mathcal{G}'$. Suppose for a contradiction, there is a minimal collider path $p_{\mathcal{M}} = \langle V_1, \ldots, V_r\rangle$, $r \geq 3$ in $\mathcal{M}$ such that the corresponding path $p_{\mathcal{G}'}$ in $\mathcal{G}'$ is not a collider path. Furthermore, we will choose the shortest such path $p_{\mathcal{M}}$ and denote the corresponding paths (same sequences of nodes) in $\mathcal{G}'$ as $p_{\mathcal{G}'}$ and in $\mathcal{G}$ as $p_{\mathcal{G}}$.

Since there are no minimal collider paths in $\mathcal{M}_i$, $i \in \{1 \ldots, k\}$, and since a node in $\mathcal{M}_i$ is not in $\mathcal{M}_j$, for $i \neq j$, $i, j \in \{1, \ldots, k\}$, we know that at least one edge on $p_{\mathcal{M}}$ is in $\mathcal{G}'$, but not in $\mathcal{G}'_C$. Since $\mathcal{G}'$ contains exactly the same minimal collider paths as $\mathcal{G}$, there is also at least one edge mark on $p_{\mathcal{M}}$ that is in $\mathcal{M}_i$, $i \in \{1 \ldots, k\}$, but not in $\mathcal{G}'$.

Note first that $p_{\mathcal{M}}$ cannot be an unshielded collider itself, and that $p_{\mathcal{M}}$ cannot contain an unshielded collider that is not on $p_{\mathcal{G}'}$. This is because none of the $\mathcal{M}_i$, $i \in \{1 \ldots, k\}$, graphs contain unshielded colliders, and $\mathcal{G}'$ itself does not contain unshielded collider that are not already in $\mathcal{G}$. Furthermore, we cannot have a path $\langle A, B, C\rangle$ in $\mathcal{M}$, where $\langle A, B\rangle$ is in $\mathcal{M}_i$, and $\langle B, C\rangle$ is in $\mathcal{M}_j$, where $i, j \in \{1, \ldots, k\}$, and $i \neq j$ (due to Corollary C.1.3). Furthermore, we know that $\mathcal{G}'$ does not contain any unshielded collider $A \bullet\!\!\to B \leftarrow\!\bullet C$, where $\langle A, B\rangle$ is in $\mathcal{G}_C^{prime}$, and $\langle B, C\rangle$ is in $\mathcal{G}'$ but not in $\mathcal{G}'_C$, or vice versa (based on Lemma C.1.7 the fact that $\mathcal{G}'$ does not contain new unshielded colliders compared to $\mathcal{G}$) and also that $\mathcal{G}'$ also cannot contain $A \bullet\!\!\to B \circ\!\!-\!\!\bullet C$, where $A \notin \text{Adj}(C, \mathcal{G}')$, due to orientations in $\mathcal{G}'$ being completed under R1.

Hence, any consecutive triple of nodes on $p_{\mathcal{M}}$ is either shielded, or the corresponding triple is already an unshielded collider on $p_{\mathcal{G}}$. In particular, any triple $\langle V_l, V_{l+1}, V_{l+2}\rangle$, $l \in \{1, \ldots, r-2\}$ on $p_{\mathcal{M}}$ such that $\langle V_l, V_{l+1}\rangle$ is in $\mathcal{G}'$, but not in $\mathcal{G}'_C$ and $\langle V_{l+1}, V_{l+2}\rangle$ is in $\mathcal{G}'_{C_i}$ for some $i \in \{1, \ldots, k\}$ is shielded.

Since $p_{\mathcal{G}'}$ is not a collider path we now consider the following options for choosing a triple on $p_{\mathcal{G}'}$ which will be used to derive our desired contradiction.

(a) Choose a triple $\langle V_l, V_{l+1}, V_{l+2}\rangle$, with the smallest index $l \in \{1, \ldots, r-2\}$ on $p_{\mathcal{G}'}$ that is of one of the following forms in $\mathcal{G}'$:

(a1) $V_l \bullet\!\!\rightarrow V_{l+1} \circ\!\!-\!\!\bullet V_{l+2}$ such that $V_l \in \text{Adj}(V_{l+2}, \mathcal{G}')$ and such that $\langle V_l, V_{l+1}\rangle$ is in $\mathcal{G}'$, but not in $\mathcal{G}'_C$ and $\langle V_{l+1}, V_{l+2}\rangle$ is in $\mathcal{G}'_{C_i}$ for some $i \in \{1, \ldots, k\}$, or

(a2) $V_l \bullet\!\!\rightarrow V_{l+1} \leftarrow\!\!\bullet V_{l+2}$ such that $V_l \bullet\!\!\rightarrow V_{l+2}$ is also in $\mathcal{G}'$, and such that $\langle V_l, V_{l+1}\rangle$ is in $\mathcal{G}'$, but not in $\mathcal{G}'_C$ and $\langle V_{l+1}, V_{l+2}\rangle$ is in $\mathcal{G}'_{C_i}$ for some $i \in \{1, \ldots, k\}$, or

(a3) $V_l \bullet\!\!\rightarrow V_{l+1} \leftarrow\!\!\bullet V_{l+2}$ such that $V_l \bullet\!\!\rightarrow V_{l+2}$ is also in $\mathcal{G}'$, and such that $\langle V_l, V_{l+1}\rangle$ is in $\mathcal{G}'_{C_i}$ for some $i \in \{1, \ldots, k\}$ and $\langle V_{l+1}, V_{l+2}\rangle$ is in $\mathcal{G}'$, but not in $\mathcal{G}'_C$.

(b) Choose a triple $\langle V_l, V_{l+1}, V_{l+2}\rangle$, with the largest index $l \in \{1, \ldots, r-2\}$ on $p_{\mathcal{G}'}$, that is of one of the following forms in $\mathcal{G}'$:

(b1) $V_l \bullet\!\!-\!\!\circ V_{l+1} \leftarrow\!\!\bullet V_{l+2}$ in $\mathcal{G}'$ such that $V_l \in \text{Adj}(V_{l+2}, \mathcal{G}')$ and such that $\langle V_l, V_{l+1}\rangle$ is in $\mathcal{G}'_{C_i}$ for some $i \in \{1, \ldots, k\}$, and $\langle V_{l+1}, V_{l+2}\rangle$ is in $\mathcal{G}'$ but not in $\mathcal{G}'_C$, or

(b2) $V_l \bullet\!\!\rightarrow V_{l+1} \leftarrow\!\!\bullet V_{l+2}$ such that $V_l \leftarrow\!\!\bullet V_{l+2}$ is also in $\mathcal{G}'$, and such that $\langle V_l, V_{l+1}\rangle$ is in $\mathcal{G}'_{C_i}$ for some $i \in \{1, \ldots, k\}$ and $\langle V_{l+1}, V_{l+2}\rangle$ is in $\mathcal{G}'$, but not in $\mathcal{G}'_C$, or

(b3) $V_l \bullet\!\!\rightarrow V_{l+1} \leftarrow\!\!\bullet V_{l+2}$ such that $V_l \leftarrow\!\!\bullet V_{l+2}$ is also in $\mathcal{G}'$, and such that $\langle V_l, V_{l+1}\rangle$ is in $\mathcal{G}'$, but not in $\mathcal{G}'_C$ and $\langle V_{l+1}, V_{l+2}\rangle$ is in $\mathcal{G}'_{C_i}$ for some $i \in \{1, \ldots, k\}$.

Note that cases (a) and (b) cover all options for the form of the triple $\langle V_l, V_{l+1}, V_{l+2}\rangle$ on $p_{\mathcal{G}'}$, so we are assured that one of the above options will exist on $p_{\mathcal{G}'}$. Also, note that case

(b) is symmetric to case (a), and the proof will be using exactly the same arguments Hence, without loss of generality, we only derive a contradiction for cases (a).

(a) We discuss all three possible forms of the triple $\langle V_l, V_{l+1}, V_{l+2}\rangle$ below and derive a contradiction in each case.

(a1) or (a2) In this case we assume that either:

- $V_l \bullet\!\!\rightarrow V_{l+1} \circ\!\!-\!\!\bullet V_{l+2}$ is in $\mathcal{G}'$ and $V_l \in \text{Adj}(V_{l+2}, \mathcal{G}')$ and moreover, $\langle V_l, V_{l+1}\rangle$ is in $\mathcal{G}'$, but not in $\mathcal{G}'_C$ and $\langle V_{l+1}, V_{l+2}\rangle$ is in $\mathcal{G}'_{C_i}$ for some $i \in \{1, \ldots, k\}$.
- Or that $V_l \bullet\!\!\rightarrow V_{l+1} \leftarrow\!\!\bullet V_{l+2}$ and $V_l \bullet\!\!\rightarrow V_{l+2}$ are in $\mathcal{G}'$, and moreover, $\langle V_l, V_{l+1}\rangle$ is in $\mathcal{G}'$, but not in $\mathcal{G}'_C$ and $\langle V_{l+1}, V_{l+2}\rangle$ is in $\mathcal{G}'_{C_i}$ for some $i \in \{1, \ldots, k\}$.

Hence, consider the form of edge $\langle V_l, V_{l+1}\rangle$ in $\mathcal{G}$. If this edge is of the form $V_l \rightarrow V_{l+1}$, $V_l \leftarrow\!\!\circ V_{l+1}$, or $V_l \leftrightarrow V_{l+1}$ in $\mathcal{G}$, then Lemma C.6.2 tells us that the form of the edge $\langle V_l, V_{l+2}\rangle$ in $\mathcal{G}'$ and $\mathcal{M}$ would allow us construct a shorter minimal collider path than $p_{\mathcal{M}}$ by skipping over V_{l+1}, which leads us to a contradiction.

Next, we consider the case where $\langle V_l, V_{l+1}\rangle$ is of the form $V_l \circ\!\!\rightarrow V_{l+1}$ in $\mathcal{G}$. Then Lemma C.6.2 implies that $V_l \rightarrow V_{l+2}$ or $V_l \leftrightarrow V_{l+2}$ is in $\mathcal{G}'$ and $\mathcal{M}$. In the latter case, we again get a contradiction with $p_{\mathcal{M}}$ being a minimal collider path, as we could replace $\langle V_l, V_{l+1}, V_{l+2}\rangle$ with $\langle V_l, V_{l+2}\rangle$. Similarly, we get the same contradiction if $\langle V_l, V_{l+1}\rangle$ is the first edge on $p_{\mathcal{G}'}$ and $p_{\mathcal{M}}$, regardless of the form of the $\langle V_l, V_{l+2}\rangle$ edge in $\mathcal{G}'$ and $\mathcal{M}$.

Hence, suppose that $V_l \rightarrow V_{l+2}$ is in $\mathcal{G}'$ and $\mathcal{M}$ and that $l > 1$, meaning that $V_l \leftrightarrow V_{l+1}$ is in $\mathcal{G}'$ and $\mathcal{M}$ (corresponding to $V_l \circ\!\!\rightarrow V_{l+1}$ in $\mathcal{G}$). Next, note that if $p_{\mathcal{G}'}(V_1, V_{l+1})$ is of the form $V_1 \bullet\!\!\rightarrow V_2 \leftrightarrow \cdots \leftrightarrow V_{l+1}$, case (iv) of Lemma C.6.2 would imply that we can choose a subsequence of $p_{\mathcal{M}}$ as a shorter minimal collider path, which is a contradiction. Otherwise, there is at least one edge $\langle V_j, V_{j+1}\rangle$, $1 \leq j < l$ on $p_{\mathcal{G}'}(V_1, V_{l+1})$ that corresponds to $V_j \circ\!\!-\!\!\circ V_{j+1}$ in $\mathcal{G}$, and also by case (iv) of Lemma C.6.2, there are edges $V_i \rightarrow V_{l+2}$ in $\mathcal{G}'$ for every $j + 1 < i \leq l$, and also that

$V_{j+1} \circ\!\!\rightarrow V_{l+2}$ or $V_{j+1} \rightarrow V_{l+2}$ is in $\mathcal{G}'$. Let $\langle V_j, V_{j+1} \rangle$, $1 \leq j < l$ be indeed such an edge on $p_{\mathcal{G}'}(V_1, V_{l+1})$ chosen so that the index j is the largest possible.

Now, consider the triple $\langle V_j, V_{j+1}, V_{j+2} \rangle$ in $\mathcal{G}'$. By choice of our original triple $\langle V_l, V_{l+1}, V_{l+2} \rangle$, we can conclude that the triple $\langle V_j, V_{j+1}, V_{j+2} \rangle$ must be of one of the forms in (b), and more precisely, either of the form described in case (b1) or case (b2).

In either case, we have that either $V_j \leftrightarrow V_{j+2}$ or $V_j \leftarrow V_{j+2}$ is in $\mathcal{G}'$ by Lemma C.6.2. If $V_j \leftrightarrow V_{j+2}$ is in $\mathcal{G}'$, we obtain our desired contradiction by constructing a shorter collider path $p_{\mathcal{M}}(V_1, V_j) \oplus \langle V_j, V_{j+2} \rangle \oplus p_{\mathcal{M}}(V_{j+2}, V_r)$. If $V_j \leftarrow V_{j+2}$ is in $\mathcal{G}'$, then we must be in case (iv) of Lemma C.6.2, so that $V_j \leftarrow V_s$, or $V_j \leftrightarrow V_s$, $j + 2 \leq s \leq l$ and either $V_j \leftarrow V_{l+1}$, $V_j \leftrightarrow V_{l+1}$, or $V_j \leftarrow\!\circ V_{l+1}$ is in $\mathcal{G}'$. If any of the mentioned edges is of the form $\leftrightarrow$ in $\mathcal{G}'$, we obtain a contradiction. Otherwise, we consider the edges between the following nodes in $\mathcal{G}$: $V_j, V_{j+1}, V_{l+1}, V_{l+2}$.

We know that $V_j \circ\!\!-\!\!\circ V_{j+1}$ and $V_{l+1} \circ\!\!-\!\!\circ V_{l+2}$ is in $\mathcal{G}$. We also know that $V_{l+1} \rightarrow V_j$ or $V_{l+1} \circ\!\!\rightarrow V_j$ are in $\mathcal{G}'$ and that similarly $V_{j+1} \rightarrow V_{l+2}$ or $V_{j+1} \circ\!\!\rightarrow V_{l+2}$ is in $\mathcal{G}'$.

If $V_{l+1} \rightarrow V_j \circ\!\!-\!\!\circ V_{j+1}$ or $V_{l+1} \rightarrow V_j \circ\!\!-\!\!\circ V_{j+1}$ is in $\mathcal{G}$, then Lemma C.1.7 and completeness of R2 in $\mathcal{G}$ imply that $V_{l+1} \rightarrow V_{j+1}$ or $V_{l+1} \circ\!\!\rightarrow V_{j+1}$ is in $\mathcal{G}$. Similarly, if $V_{j+1} \rightarrow V_{l+2} \circ\!\!-\!\!\circ V_{l+1}$ or $V_{j+1} \rightarrow V_{l+2} \circ\!\!-\!\!\circ V_{l+1}$ are in $\mathcal{G}$, then Lemmas C.1.7 and completeness of R2 in $\mathcal{G}$ imply that $V_{j+1} \rightarrow V_{l+1}$ or $V_{j+1} \circ\!\!\rightarrow V_{l+1}$ is in $\mathcal{G}$. Both of these cannot be true at the same time, so at least one of the edges $\langle V_{l+1}, V_j \rangle$ or $\langle V_{j+1}, V_{l+2} \rangle$ are of the form $\circ\!\!-\!\!\circ$ in $\mathcal{G}$.

Furthermore, if $V_{l+2} \circ\!\!-\!\!\circ V_{l+1} \circ\!\!-\!\!\circ V_j \circ\!\!-\!\!\circ V_{j+1}$ is in $\mathcal{G}$, then the edge $\langle V_{l+2}, V_{j+1} \rangle$ must also be of the form $\circ\!\!-\!\!\circ$ in $\mathcal{G}$ (Lemma C.1.5). Analogously, if $V_j \circ\!\!-\!\!\circ V_{j+1} \circ\!\!-\!\!\circ V_{l+2} \circ\!\!-\!\!\circ V_{l+1}$, is in $\mathcal{G}$, we conclude that $V_j \circ\!\!-\!\!\circ V_{l+1}$ is in $\mathcal{G}$ as well.

Hence, now we have an undirected cycle of length 4 in $\mathcal{G}$. Then by the chordality of the circle component of $\mathcal{G}$ (Corollary C.1.3), either $V_j \circ\!\!-\!\!\circ V_{l+2}$ or $V_{j+1} \circ\!\!-\!\!\circ V_{l+1}$ is in $\mathcal{G}$. Let us assume without loss of generality that $V_j \circ\!\!-\!\!\circ V_{l+2}$ is in $\mathcal{G}$, and consider the form

of this edge in $\mathcal{M}$. If $V_j \to V_{l+2}$ is in $\mathcal{M}$, then this edge together with $V_{l+2} \bullet\!\!\to V_{l+1} \to V_j$ contradicts that $\mathcal{M}$ is ancestral. If $V_j \leftarrow V_{l+2}$ is in $\mathcal{M}$, then this edge together with $V_j \bullet\!\!\to V_{j+1} \to V_{l+2}$ contradicts that $\mathcal{M}$ is ancestral. Hence, the only option is for $V_j \leftrightarrow V_{l+2}$ to be in $\mathcal{M}$, in which case $p_{\mathcal{M}}(V_1, V_j) \oplus \langle V_j, V_{l+2} \rangle \oplus p_{\mathcal{M}}(V_{l+2}, V_r)$ is a subsequence of $p_{\mathcal{M}}$ in $\mathcal{M}$ that forms a shorter collider path, which is a contradiction.

(a3) $V_l \bullet\!\!\to V_{l+1} \leftarrow\!\!\bullet V_{l+2}$ such that $V_l \bullet\!\!\to V_{l+2}$ is also in $\mathcal{G}'$, and $\langle V_l, V_{l+1} \rangle$ is in $\mathcal{G}'_{C_i}$ for some $i \in \{1, \ldots, k\}$ and $\langle V_{l+1}, V_{l+2} \rangle$ is in $\mathcal{G}'$, but not in $\mathcal{G}'_C$. By Lemma C.6.4, we have that either $V_l \leftrightarrow V_{l+2}$ or $V_l \to V_{l+2}$ is in $\mathcal{G}'$. In the former case, we again get a contradiction with $p_{\mathcal{M}}$ being a minimal collider path, as we could replace $\langle V_l, V_{l+1}, V_{l+2} \rangle$ with $\langle V_l, V_{l+2} \rangle$. Similarly, we get the same contradiction if $\langle V_l, V_{l+1} \rangle$ is the first edge on $p_{\mathcal{G}'}$ and $p_{\mathcal{M}}$, regardless of the form of the $\langle V_l, V_{l+2} \rangle$ edge in $\mathcal{G}'$ and $\mathcal{M}$.

Hence, suppose that $V_l \to V_{l+2}$ is in $\mathcal{G}'$ and $\mathcal{M}$ and that $l > 1$, meaning that $V_l \leftrightarrow V_{l+1}$ is in $\mathcal{G}'$ and $\mathcal{M}$ (corresponding to $V_l \circ\!\!-\!\!\circ V_{l+1}$ in $\mathcal{G}$). Suppose first that V_{l-1} is also in $\mathcal{G}'_{C_i}$.

Since $V_{l+2} \bullet\!\!\to V_{l+1}$ is in $\mathcal{G}'$, if $V_{l+2} \notin \text{Adj}(V_{l-1}, \mathcal{G})$, we have that $V_{l+1} \to V_{l-1}$ is in $\mathcal{G}'$ by R1, and therefore, $\langle V_{l+2}, V_{l+1}, V_l, V_{l-1} \rangle$ would be a minimal discriminating collider path for V_l that is in $\mathcal{G}'$ but not in $\mathcal{G}$, therefore giving us our contradiction.

Otherwise, $V_{l+2} \in \text{Adj}(V_{l-1}, \mathcal{G})$. In this case consider again the edge $\langle V_{l-1}, V_{l+1} \rangle$ in $\mathcal{M}$. If $V_{l-1} \leftrightarrow V_{l+1}$ is in $\mathcal{M}$ we obtain a contradiction with our choice of path. If $V_{l-1} \leftarrow V_{l+1}$, then due to ancestrality of $\mathcal{M}$, we have that $V_{l-1} \leftrightarrow V_{l+2}$ is in $\mathcal{M}$, then again there is a subsequence of $p_{\mathcal{M}}$ that forms a collider path in $\mathcal{M}$ which also gives us a contradiction.

Otherwise, $V_{l+2} \in \text{Adj}(V_{l-1}, \mathcal{G})$ and $V_{l-1} \to V_{l+1}$ is in $\mathcal{M}$. Let j be chosen as the smallest index on $p_{\mathcal{M}}(V_1, V_{l-1})$ such that $V_j, \ldots, V_l, V_{l+1}$ are all in $\mathcal{G}'_{C_i}$. Then all of the nodes in $V_j, \ldots, V_{l+1}$ must be in the same clique since we do not create any minimal collider paths in $\mathcal{M}_i$. Furthermore, if any edge $V_d \leftrightarrow V_s$, $j \leq d < d+1 <$

$s \leq l+1$ is in $\mathcal{M}$, we can choose a subsequence of $p_{\mathcal{M}}$ that is a shorter collider path. Moreover, since $\mathcal{M}_i$ and $\mathcal{M}$ are ancestral, it follows that either $V_d \to V_s$ for all pairs $j \leq d < d+1 < s \leq l+1$ or $V_d \to V_s$. If $j = 1$, we now have that $\langle V_1, V_{l+1} \rangle \oplus p_{\mathcal{M}}(V_{l+1}, V_r)$ is a collider path, which is a contradiction.

Otherwise, $j \neq 1$, and consider the triple $\langle V_{j-1}, V_j, V_{j+1} \rangle$ in $\mathcal{G}'$. Note that $\langle V_{j-1}, V_j \rangle$ cannot be in $\mathcal{G}'_C$, otherwise it would be in $\mathcal{G}'_{C_i}$ (Corollary C.1.3). Hence, $\langle V_{j-1}, V_j \rangle$ is in $\mathcal{G}'$ but not in $\mathcal{G}'_C$, and $\langle V_j, V_{j+1} \rangle$ is in $\mathcal{G}'_{C_i}$. By choice of our original triple $\langle V_l, V_{l+1}, V_{l+2} \rangle$, we can conclude that the triple $\langle V_{j-1}, V_j, V_{j+1} \rangle$ must be of the form in case (b3), that is $V_{j-1} \leftarrow\!\bullet V_{j+1}$ is in $\mathcal{M}$.

If $V_{j-1} \leftrightarrow V_{j+1}$ is in $\mathcal{M}$, then of course, $p_{\mathcal{M}}(V_1, V_{j-1}) \oplus \langle V_{j-1}, V_{j+1} \rangle \oplus p_{\mathcal{M}}(V_{j+1}, V_r)$ is a subsequence of $p_{\mathcal{M}}$ that forms a minimal collider path and give us our contradiction.

Otherwise, $V_{j-1} \leftarrow V_{j+1}$ is in $\mathcal{M}$ and we focus on the subpath $\langle V_{j-1}, V_j, V_{j+1}, V_{j+2} \rangle$. Since $V_j \to V_{j+2}$ is in $\mathcal{M}$ by assumption, we have that if edge $\langle V_{j-1}, V_{j+2} \rangle$ is in $\mathcal{G}$, then due to the ancestral property of $\mathcal{M}$, $V_{j-1} \leftrightarrow V_{j+2}$ is in $\mathcal{M}$ and then similarly to above, $p_{\mathcal{M}}(V_1, V_{j-1}) \oplus \langle V_{j-1}, V_{j+2} \rangle \oplus p_{\mathcal{M}}(V_{j+2}, V_r)$ is a subsequence of $p_{\mathcal{M}}$ that forms a minimal collider path and give us our contradiction. If however $V_{j-1} \notin \text{Adj}(V_{j+2}, \mathcal{G})$, then consider that $\langle V_{j-1}, V_j, V_{j+1}, V_{j+2} \rangle$ is an inducing path and a minimal collider path in $\mathcal{M}$. Since $\langle V_{j-1}, V_j, V_{j+1}, V_{j+2} \rangle$ is an inducing path in $\mathcal{M}$, this path cannot be collider path in $\mathcal{G}'$ (otherwise, it would be a possibly inducing path and contradict Lemma 4.4.11). Furthermore, since $1 \leq j \leq l-1 < r$, $\langle V_{j-1}, V_j, V_{j+1}, V_{j+2} \rangle$ is shorter than $p_{\mathcal{M}}$, so our choice of a minimal collider path is incorrect and we obtain our contradiction.

□

Supporting Results for the Proof of Theorem 4.6.5

Lemma C.6.2. *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be an ancestral partial mixed graph and $\mathcal{P}$ be an essential ancestral graph such that they have the same skeleton, share the same set of minimal collider*

paths, and the set of all invariant edge marks in $\mathcal{P}$ *is a subset of the invariant edge marks in* $\mathcal{G}$*. Suppose furthermore, that all* $A \circ\!\!\rightarrow B$ *edges in* $\mathcal{P}$ *correspond to* $A \rightarrow B$ *or* $A \leftrightarrow B$ *edges in* $\mathcal{G}$ *and that orientations in* $\mathcal{G}$ *are closed under R1-R4, R8-R12. Suppose that* $E \bullet\!\!\rightarrow C$ *is in* $\mathcal{G}$*, where this edge is of one of the following forms in* $\mathcal{P}$*:* $E \leftarrow\!\!\circ C$*,* $E \rightarrow C$*,* $E \circ\!\!\rightarrow C$*, or* $E \leftrightarrow C$*. Furthermore, suppose that there is an edge* $\langle C, D \rangle$ *in* $\mathcal{G}$ *that corresponds to* $C \circ\!\!-\!\!\circ D$ *in* $\mathcal{P}$*, and also suppose that edge* $\langle E, D \rangle$ *is in* $\mathcal{G}$*.*

(1) If the form of the edge $\langle C, D \rangle$ *is* $C \circ\!\!-\!\!\bullet D$ *in* $\mathcal{G}$*, or*

(2) if the form of the edge $\langle C, D \rangle$ *is* $C \leftarrow\!\!\bullet D$ *in* $\mathcal{G}$*, while the form of the edge* $\langle E, D \rangle$ *is* $E \bullet\!\!\rightarrow D$ *in* $\mathcal{G}$*,*

then the following hold:

(i) If $E \rightarrow C$ *is in* $\mathcal{P}$*, then* $E \rightarrow C$ *is in* $\mathcal{G}$ *and also* $E \rightarrow D$ *or* $E \leftrightarrow D$ *is in* $\mathcal{G}$*.*

(ii) If $E \leftarrow\!\!\circ C$ *is in* $\mathcal{P}$*, then* $E \leftrightarrow C$ *is in* $\mathcal{G}$ *and also* $E \leftrightarrow D$ *is in* $\mathcal{G}$*.*

(iii) If $E \leftrightarrow C$ *is in* $\mathcal{P}$*, then* $E \leftrightarrow C$ *is in* $\mathcal{G}$ *and also* $E \leftrightarrow D$ *is in* $\mathcal{G}$*.*

(iv) If $E \circ\!\!\rightarrow C$ *is in* $\mathcal{P}$*, then either*

- $E \rightarrow C$ *and* $E \rightarrow D$ *are in* $\mathcal{G}$*, or*
- $E \rightarrow C$ *and* $E \leftrightarrow D$ *are in* $\mathcal{G}$*, or*
- $E \leftrightarrow C$ *and* $E \leftrightarrow D$ *are in* $\mathcal{G}$*, or*
- $E \rightarrow C$ *and* $E \leftrightarrow D$ *are in* $\mathcal{G}$*. Furthermore, in this setting, we have that*

  (a) for every P_1 *in* $\mathcal{G}$ *such that* $P_1 \bullet\!\!\rightarrow E$ *is in* $\mathcal{G}$*,* $P_1 \bullet\!\!\rightarrow D$ *is in* $\mathcal{G}$*, and*

  (b) for every $P_1 \bullet\!\!\rightarrow P_2 \leftrightarrow \cdots \leftrightarrow P_k$*,* $P_k \equiv E$*,* $k > 1$ *either there is an* $i \in \{1, \ldots, k\}$ *such that* $P_i \leftrightarrow D$ *and* $P_j \rightarrow D$*, for all* $j \in \{i{+}1, \ldots k\}$ *or* $P_1 \bullet\!\!\rightarrow D$ *and* $P_i \rightarrow D$ *for all* $i \in \{2, \ldots, k\}$ *is in* $\mathcal{G}$*.*

Proof of Lemma C.6.2. (i) Since $E \to C \circ\!\!-\!\!\circ D$ is in $\mathcal{P}$, Lemma C.1.7 implies that $E \to D$ or $E \circ\!\!\to D$ are in $\mathcal{P}$. Since all $\circ\!\!\to$ edges in $\mathcal{P}$ correspond to $\to$ or $\leftrightarrow$ edges in $\mathcal{G}$, we know that $E \to D$, or $E \leftrightarrow D$ is in $\mathcal{G}$.

(ii) Since $E \leftarrow\!\!\circ C \circ\!\!-\!\!\circ D$ is in $\mathcal{P}$, and $E \in \text{Adj}(D, \mathcal{P})$, we have by Lemmas C.1.7 and the fact that R2 is completed in $\mathcal{P}$, that $E \leftarrow\!\!\circ D$ or $E \leftarrow D$ is in $\mathcal{P}$. Then $E \leftrightarrow D$ or $E \leftarrow D$ is in $\mathcal{G}$.

In case (2), we then immediately have that $E \leftrightarrow D$ is in $\mathcal{G}$. Now, in case (1), $E \leftrightarrow C \circ\!\!-\!\!\bullet D$ in $\mathcal{G}$, and the fact that orientations in $\mathcal{G}$ are completed with respect to R2 would imply that $E \leftarrow D$ cannot be in $\mathcal{P}$. Hence, $E \leftarrow\!\!\circ D$ is in $\mathcal{P}$ and therefore, $E \leftrightarrow D$ is in $\mathcal{G}$.

(iii) If $E \leftrightarrow C$ is in $\mathcal{P}$, then since $C \circ\!\!-\!\!\circ D$ is in $\mathcal{P}$, Lemma C.1.7 and completeness of R2 in $\mathcal{P}$ imply that $E \leftrightarrow D$ is also in $\mathcal{P}$. Hence, $E \leftrightarrow C$ and $E \leftrightarrow D$ are also in $\mathcal{G}$.

(iv) If $E \circ\!\!\to C$ is in $\mathcal{P}$, then since $C \circ\!\!-\!\!\circ D$ is in $\mathcal{P}$, Lemma C.1.7 implies $E \circ\!\!\to D$ or $E \to D$ is in $\mathcal{P}$. Then we have the combination of cases as listed above. In particular, if $E \leftrightarrow C$ and $E \to D$ are in $\mathcal{G}$, we also have that cases (iv)a and (iv)b hold because $\mathcal{G}$ is ancestral and that $\mathcal{G}$ has the same minimal collider paths as $\mathcal{P}$. Note that Corollary C.6.3, Theorem 4.3.1 and the fact that all minimal collider paths in $\mathcal{G}$ are also minimal collider paths in $\mathcal{P}$ imply that every last collider on a discriminating collider path in $\mathcal{G}$ must be a collider in $\mathcal{P}$ that every collider on a discriminating collider path in $\mathcal{G}$, must already be a collider in $\mathcal{P}$. Since we know that $C \circ\!\!-\!\!\bullet D$ is in $\mathcal{P}$, we know that the paths of the form $P_i \bullet\!\!\to P_{i+1} \leftrightarrow \cdots \leftrightarrow P_k \leftrightarrow E \leftrightarrow C \leftarrow\!\!\bullet D$, $i \in \{1, \ldots, k\}$ in $\mathcal{G}$ cannot be discriminating paths, hence $P_i \in \text{Adj}(D, \mathcal{G})$. The rest of the argument follows by using completeness of orientation rules R1, R2, and R4 in $\mathcal{G}$. □

Corollary C.6.3. *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be an ancestral partial mixed graph. Furthermore, suppose edge orientations in $\mathcal{G}$ are closed under R1, R2, R4. Let $p = \langle P_1, P_2, \ldots, P_k \rangle, k \geq 3$ be a minimal collider path in $\mathcal{G}$. Then for every $i \in \{2, \ldots, k-1\}$, one of the following holds:*

(i) $P_{i-1} \bullet\!\!\to P_i \leftarrow\!\!\bullet P_{i+1}$ and $P_{i-1} \notin \text{Adj}(P_{i+1}, \mathcal{G})$, or

(ii) $\exists\, l \in \{1, \ldots, i-2\}$, such that $P_l \bullet\!\!\to P_{l+1} \leftrightarrow \cdots \leftrightarrow P_i \leftarrow\!\!\bullet P_{i+1}$ is a discriminating collider path from P_l to P_{i+1} for P_i, or

(iii) $\exists\ r \in \{i+2, \dots k\}$ *such that* $P_r \bullet\!\!\rightarrow P_{r-1} \leftrightarrow \cdots \leftrightarrow P_{i+1} \leftrightarrow P_i \leftarrow\!\!\bullet P_{i-1}$ *is a discriminating collider path from* P_r *to* P_{i-1} *for* P_i.

Proof of Corollary C.6.3. Follows from Lemma 4.6.9 and the fact that R4 subsumes Zhao-R4. □

Lemma C.6.4. *Let* $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ *be an ancestral partial mixed graph and* $\mathcal{P}$ *be an essential ancestral graph such that they have the same skeleton, share the same set of minimal collider paths, and the set of all invariant edge marks in* $\mathcal{P}$ *is a subset of the invariant edge marks in* $\mathcal{G}$. *Suppose furthermore, that all* $A \circ\!\!\rightarrow B$ *edges in* $\mathcal{P}$ *correspond to* $A \rightarrow B$ *or* $A \leftrightarrow B$ *edges in* $\mathcal{G}$ *and that orientations in* $\mathcal{G}$ *are closed under R1-R4, R8-R12. Suppose that* $C \leftarrow\!\!\bullet D$ *is an edge in* $\mathcal{G}$ *that corresponds to* $C \circ\!\!-\!\!\circ D$ *in* $\mathcal{P}$, *and also that* $E \bullet\!\!\rightarrow C$ *is in* $\mathcal{G}$, *where this edge is of one of the following forms in* $\mathcal{P}$: $E \leftarrow\!\!\circ C$, $E \rightarrow C$, $E \circ\!\!\rightarrow C$, *or* $E \leftrightarrow C$, *then there is an edge* $\langle E, D \rangle$ *in* $\mathcal{G}$ *and suppose that this edge is of the form* $E \leftarrow\!\!\bullet D$ *is in* $\mathcal{G}$. *Then*

(i) If $E \rightarrow C$ *is in* $\mathcal{P}$, *then* $E \rightarrow C$ *is in* $\mathcal{G}$ *and* $E \leftrightarrow D$ *is in* $\mathcal{G}$.

(ii) If $E \leftrightarrow C$ *is in* $\mathcal{P}$, *then* $E \leftrightarrow C$ *is in* $\mathcal{G}$ *and also* $E \leftrightarrow D$ *is in* $\mathcal{G}$.

(iii) If $E \circ\!\!\rightarrow C$ *is in* $\mathcal{P}$, *then either* $E \rightarrow C$ *or* $E \leftrightarrow C$ *is in* $\mathcal{G}$ *and* $E \leftrightarrow D$ *is in* $\mathcal{G}$.

(iv) If $E \leftarrow\!\!\circ C$ *is in* $\mathcal{P}$, *then* $E \leftrightarrow C$ *is in* $\mathcal{G}$ *and* $E \leftrightarrow D$ *or* $E \leftarrow D$ *is in* $\mathcal{G}$.

Proof of Lemma C.6.4. (i) Since $E \rightarrow C \circ\!\!-\!\!\circ D$ is in $\mathcal{P}$, Lemma C.1.7 implies that $E \rightarrow D$ or $E \circ\!\!\rightarrow D$ are in $\mathcal{P}$. Since all $\circ\!\!\rightarrow$ edges in $\mathcal{P}$ correspond to $\rightarrow$ or $\leftrightarrow$ edges in $\mathcal{G}$, we know that $E \rightarrow D$, or $E \leftrightarrow D$ is in $\mathcal{G}$. By assumption, we know $E \leftarrow\!\!\bullet D$ is in $\mathcal{G}$, and hence, $E \leftrightarrow D$ must be in $\mathcal{G}$.

(ii) If $E \leftrightarrow C$ is in $\mathcal{P}$, then since $C \circ\!\!-\!\!\circ D$ is in $\mathcal{P}$, Lemma C.1.7 and completeness of R2 in $\mathcal{P}$, imply that $E \leftrightarrow D$ is also in $\mathcal{P}$. Hence, $E \leftrightarrow C$ and $E \leftrightarrow D$ are also in $\mathcal{G}$.

(iii) If $E \circ\!\!\rightarrow C$ is in $\mathcal{P}$, then since $C \circ\!\!-\!\!\circ D$ is in $\mathcal{P}$, Lemma C.1.7 implies $E \circ\!\!\rightarrow D$ or $E \rightarrow D$ is in $\mathcal{P}$. Since we know, that $E \leftarrow\!\!\bullet D$ is in $\mathcal{G}'$, it must be that $E \circ\!\!\rightarrow D$ is in $\mathcal{P}$ and $E \leftrightarrow D$ is in $\mathcal{G}$.

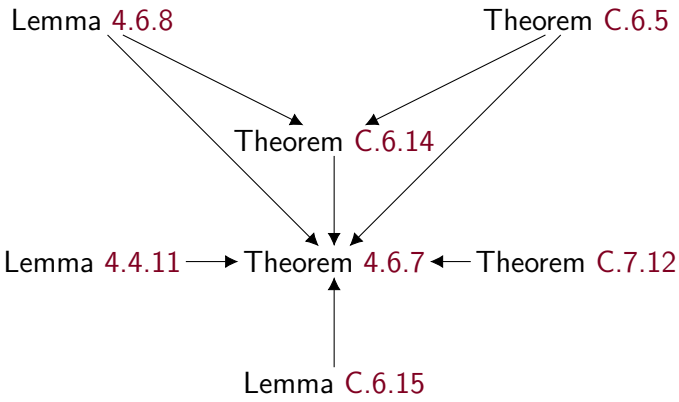


Figure C.2: Proof structure of Theorem 4.6.7

(iv) Since $E \leftarrow\!\circ C \circ\!\!-\!\!\circ D$ is in $\mathcal{P}$, and $E \in \text{Adj}(D, \mathcal{P})$, we have by Lemmas C.1.7 and the fact that R2 is completed in $\mathcal{P}$ that $E \leftarrow\!\circ D$ or $E \leftarrow D$ is in $\mathcal{P}$. Since we assume that $E \leftarrow\!\bullet D$ is in $\mathcal{G}$, this implies that $E \leftrightarrow D$ or $E \leftarrow D$ is in $\mathcal{G}$. □

C.6.4 Theorem 4.6.7

The proof of Theorem 4.6.7 can become cumbersome. To help the reader understand, we give an overview of the primary supporting arguments:

1. Theorem C.6.5 says that a very specific tail-orientation process that we describe in the proof of Theorem 4.6.7 results in a graph closed under the orientation rules.

2. Theorem C.6.14 says that the same process preserves ancestrality as well as the set of minimal collider paths.

3. Lemma C.6.15 says that it is feasible to look at the chordal and non-chordal components of the graph

4. Theorem C.7.12 shows how to orient the chordal component like Meek (1995).

We illustrate this dependence in Figure C.2. In this appendix, we only show Theorems C.6.5, C.6.14, and Lemma C.6.15 (and their supporting arguments). Theorem C.7.12 is sufficiently

complicated and relies on substantially different arguments that we study it separately in Appendix C.7.

Proof of Theorem 4.6.7. Consider the following procedure. First, we identify the circle component of $\mathcal{G} = (\mathbf{V}, \mathbf{E})$. This is the subgraph of $\mathcal{G}$ containing only $\circ\!\!-\!\!\circ$ edges, $\mathbf{E}_C$. Call this $\mathcal{G}_C = (\mathbf{V}, \mathbf{E}_C)$. Consider the same edges present in $\mathcal{G}' = (\mathbf{V}, \mathbf{E}')$, which might potentially have different edge mark orientations, $\mathbf{E}'_C$. Note that by Corollary C.1.3, $\mathcal{G}_C = (\mathbf{V}, \mathbf{E}_C)$ is a collection of undirected connected chordal components $\mathcal{G}_{C_1}, \dots \mathcal{G}_{C_k}$, $k \geq 1$, each of which is an induced subgraph of $\mathcal{G}$. We will refer to the corresponding induced subgraphs of $\mathcal{G}'$ as $\mathcal{G}'_{C_1}, \dots, \mathcal{G}'_{C_k}$.

Now, construct the graph $\mathcal{G}''$ by replacing all edges $\langle X, Y \rangle$ in $\mathcal{G}'$ that are of the form $X \circ\!\!\rightarrow Y$ in both $\mathcal{G}'$ and $\mathcal{G}$ with $X \rightarrow Y$.

By Theorems C.6.5 and C.6.14, $\mathcal{G}''$ is ancestral, has the same minimal collider paths as $\mathcal{G}'$, and edge mark orientations in $\mathcal{G}''$ are complete under R1-R4 and R8-R12.

Furthermore, Theorem C.7.12 allows us to orient $\mathcal{G}'_{C_i}$, $i \geq 1$ of $\mathcal{G}'_C$ into a MAG $\mathcal{M}_i$ with no minimal collider paths, and with the desired edge orientation of $\langle A, B \rangle$ indicated in cases (i) or (ii) of Theorem 4.6.7.

Now, suppose we construct a new directed mixed graph $\mathcal{M} = (\mathbf{V}, \mathbf{E}_{\mathcal{M}})$ obtained by taking the union of all invariant edge marks in $\mathcal{G}''$ and $\mathcal{M}_i$ for all $i \in \{1, \dots, k\}$. We will now show that $\mathcal{M}$ is a MAG represented by $\mathcal{G}'$. That is $\mathcal{M}$ is an ancestral graph with the same minimal collider paths as $\mathcal{G}'$ (Lemma 4.4.11). In particular, it suffices to show that there are no directed cycles or almost directed cycles in $\mathcal{M}$ that contain some edges from $\mathcal{M}_i$, $i \in \{1, \dots, k\}$ and some edges from $\mathcal{G}''$, and also that there are no minimal collider paths in $\mathcal{M}$ that are made up of edges from $\mathcal{M}_i$, $i \in \{1, \dots, k\}$ and edges from $\mathcal{G}''$.

First, we show that $\mathcal{M}$ is ancestral. By Lemma 4.6.8, it is enough to show that there are no directed cycles or almost directed cycles of length 3 in $\mathcal{M}$. For sake of contradiction, we will suppose that there is a triple $A \rightarrow B \rightarrow C$ and edge $A \leftarrow\!\!\bullet C$ in $\mathcal{M}$. Furthermore, since $\mathcal{G}''$ and $\mathcal{M}_i$, for all i are ancestral, and since $\mathcal{G}'_{C_i}$ are induced subgraphs of $\mathcal{G}'_C$ (Lemmas B.4

and B.8 of Zhang, 2008b), $\forall i$, exactly two of the nodes A, B, C are in $\mathcal{G}'_{C_j}$ for some $j \geq 1$. We consider the options below:

(a) Suppose that A, C are in $\mathcal{G}'_{C_j}$, and $B \notin \mathcal{G}_C$. Note again that $A \to B \to C$ and $A \leftarrow\!\bullet C$ is in $\mathcal{M}$. Furthermore, since $A, C \in \mathcal{G}'_{C_j}$ and $B \notin \mathcal{G}_C$, we have that $A \circ\!\!-\!\!\circ C$ is in $\mathcal{G}$, and also that $B \to C$ or $B \circ\!\!\to C$ is in $\mathcal{G}$. Now Lemma C.1.7, implies that $B \bullet\!\!\to A$ must have been in $\mathcal{G}$, which leads us to a contradiction.

(b) Suppose that A, B are in $\mathcal{G}'_{C_j}$, and $C \notin \mathcal{G}_C$. Again, consider that $A \to B \to C$ and $A \leftarrow\!\bullet C$ are in $\mathcal{M}$. Therefore, similarly to above, we have that $A \circ\!\!-\!\!\circ B$ is in $\mathcal{G}$ and since $C \notin \mathcal{G}_C$, $A \leftarrow\!\bullet C$ is in $\mathcal{G}$. Hence, we obtain a contradiction with Lemma C.1.7 as in the previous case.

(c) Suppose that B, C are in $\mathcal{G}'_{C_j}$, and $A \notin \mathcal{G}_C$. Now again $A \to B \to C$ and $A \leftarrow\!\bullet C$ are in $\mathcal{M}$. Now, since $B \circ\!\!-\!\!\circ C$ is in $\mathcal{G}$ and $A \circ\!\!\to B$ or $A \to B$ must be is in $\mathcal{G}$. Therefore, Lemma C.1.7 implies that $A \bullet\!\!\to C$ is in $\mathcal{G}$. If $A \leftarrow C$ is in $\mathcal{M}$, we immediately arrive at a contradiction, so suppose that $A \leftrightarrow C$ is in $\mathcal{M}$. Since the same edge $\langle A, C \rangle$ from $\mathcal{M}$ is in $\mathcal{G}''$ and we do not introduce bidirected edges in $\mathcal{G}''$ that are not already in $\mathcal{G}$, we have that $A \leftrightarrow C$ is in $\mathcal{G}$.

Now, we have that $A \leftrightarrow C$ is in $\mathcal{G}$ and either $A \to B \circ\!\!-\!\!\circ C$ or $A \circ\!\!\to B \circ\!\!-\!\!\circ C$ is in $\mathcal{G}$. However, both cases lead to a contradiction. The former contradicts that orientations in $\mathcal{G}$ are complete under R2 while the latter case contradicts Lemma C.1.7.

Therefore, $\mathcal{M}$ is ancestral. It remains to prove that $\mathcal{M}$ has the same minimal collider paths as $\mathcal{G}'$. Suppose for a contradiction, there is a minimal collider path $p_{\mathcal{M}} = \langle V_1, \ldots, V_r \rangle$, $r \geq 3$ in $\mathcal{M}$ such that the corresponding path $p_{\mathcal{G}'}$ in $\mathcal{G}'$ is not a collider path. Since there are no minimal collider paths in $\mathcal{M}_i$, $i \in \{1 \ldots, k\}$, and since a node in $\mathcal{M}_i$ is not in $\mathcal{M}_j$, for $i \neq j$, $i, j \in \{1, \ldots, k\}$, we know that at least one edge on $p_{\mathcal{M}}$ is in $\mathcal{G}''$ (but not in any of the $\mathcal{M}_i$'s). Since no arrowheads are oriented in the process of orienting $\mathcal{G}''$, there is also at least one

edge on $p_{\mathcal{M}}$ that is not in $\mathcal{G}''$. Hence, let $\langle V_l, V_{l+1}, V_{l+2}\rangle$, $l \in \{1, \ldots, r-2\}$ be a triple on $p_{\mathcal{M}}$ such that

(a) $\langle V_l, V_{l+1}\rangle$ is in $\mathcal{G}''$, and $\langle V_{l+1}, V_{l+2}\rangle$ is in $\mathcal{M}_i$ for some $i \in \{1, \ldots, k\}$, or

(b) $\langle V_l, V_{l+1}\rangle$ is in $\mathcal{M}_i$ for some $i \in \{1, \ldots, k\}$, and $\langle V_{l+1}, V_{l+2}\rangle$ is in $\mathcal{G}''$.

Note that case (b) is symmetric to case (a) and the proof will be using exactly the same arguments. Hence, without loss of generality, suppose we are in case (a), that is $\langle V_l, V_{l+1}\rangle$ is in $\mathcal{G}''$, and $\langle V_{l+1}, V_{l+2}\rangle$ is in $\mathcal{M}_i$ for some $i \in \{1, \ldots, k\}$. Note that since the arrowhead on $V_l \bullet\!\!\rightarrow V_{l+1}$ is in $\mathcal{M}$ and in $\mathcal{G}''$ it is also in $\mathcal{G}'$ (and furthermore, in $\mathcal{G}$ as well). Since $\langle V_{l+1}, V_{l+2}\rangle$ is in $\mathcal{M}_i$, this edge corresponds to $V_{l+1} \circ\!\!-\!\!\circ V_{l+2}$ in $\mathcal{G}$. Now, Lemma C.6.15 implies that either

- $V_l \to V_{l+1}$ and $V_l \to V_{l+2}$ are in $\mathcal{M}$, or
- $V_l \leftrightarrow V_{l+1}$ and $V_l \leftrightarrow V_{l+2}$ are in $\mathcal{M}$.

If $l = 1$, we have a contradiction with $p_{\mathcal{M}}$ being a minimal collider path since $\langle V_1, V_3\rangle \oplus p_{\mathcal{M}}(V_3, V_r)$ is a subsequence of $p_{\mathcal{M}}$ that is also a collider path. Additionally, if $l \neq 1$, then we are in the latter of the two cases above, meaning that $V_l \leftrightarrow V_{l+1}$ and $V_l \leftrightarrow V_{l+2}$ are in $\mathcal{M}$. Hence, we again have a contradiction with $p_{\mathcal{M}}$ being a minimal collider path since $p_{\mathcal{M}}(V_1, V_l) \oplus \langle V_l, V_{l+2}\rangle \oplus p_{\mathcal{M}}(V_{l+2}, V_r)$ is a subsequence of $p_{\mathcal{M}}$ that is also a collider path. □

Supporting results for Theorem 4.6.7, Part 1

We first prove the main results for the graph $\mathcal{G}''$ described in the proof of Theorem 4.6.7. As a reminder, this graph is obtained by orienting all $\circ\!\!\rightarrow$ edges in the essential ancestral graph as tails i.e., $\to$. In Theorem C.6.5, we show that this graph is complete under the orientation rules R1-R4, R8-R12. This is the most difficult of the auxiliary results and we dedicate this entire section to studying it.

In Figure C.3, we illustrate the intricate dependence of the auxiliary results that the two main theorems build upon.

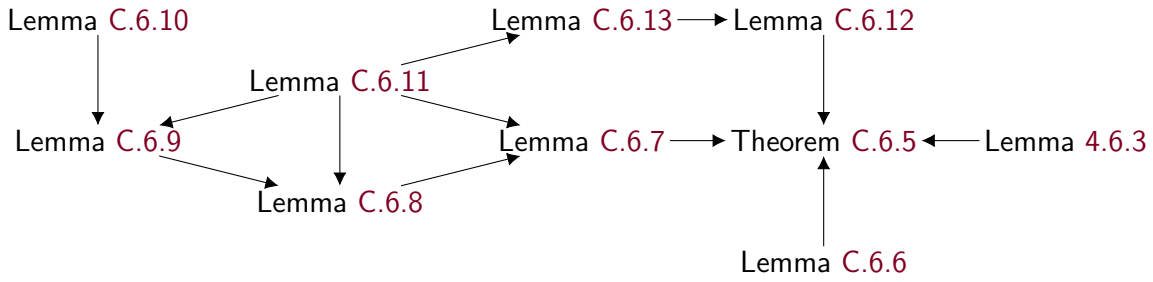


Figure C.3: Proof structure of Theorem C.6.5

Theorem C.6.5. *Let $\mathcal{G}' = (\mathbf{V}, \mathbf{E})$ be an ancestral partial mixed graph and $\mathcal{G}$ be an essential ancestral graph such that $\mathcal{G}$ and $\mathcal{G}'$ have the same skeleton, the same set of minimal collider paths, and all invariant edge marks in $\mathcal{G}$ exist and are identical in $\mathcal{G}'$. Suppose furthermore, that every edge $A \circ\!\!\rightarrow B$ in $\mathcal{G}$ corresponds either to $A \rightarrow B$ or to $A \circ\!\!\rightarrow B$ in $\mathcal{G}'$ and that edge mark orientations in $\mathcal{G}'$ are closed under R1-R4, R8-R12. Let $\mathcal{G}''$ be identical to $\mathcal{G}'$ except all $A \circ\!\!\rightarrow B$ edges in $\mathcal{G}$ correspond to $A \rightarrow B$ edges in $\mathcal{G}''$. Then edge mark orientations in $\mathcal{G}''$ are closed under R1-R4, R8-R12.*

Proof of Theorem C.6.5. It is enough to consider each orientation rule and show that the antecedent for any rule will not occur in $\mathcal{G}''$ directly. A lot of the arguments below will use the fact that $\mathcal{G}''$ does not contain any new arrowhead edge marks compared to $\mathcal{G}'$ and that edge mark orientations in $\mathcal{G}'$ are already complete under R1-R4, R8-R12.

First, note that completeness of edge marks under R3 and R9 follows immediately by Lemma 4.6.3.

R1 The antecedent of R1 requires a triple $X \bullet\!\!\rightarrow Y \circ\!\!\bullet Z$ to exist in $\mathcal{G}''$, and $X \notin \text{Adj}(Z, \mathcal{G}'')$. We know this type of triple cannot exist in $\mathcal{G}''$ because we do not introduce any arrowhead edge marks in $\mathcal{G}''$ compared to $\mathcal{G}'$, and edge mark orientations in $\mathcal{G}'$ are closed under R1-R4, R8-R12.

R2 Having the antecedent of R2 in $\mathcal{G}''$ but not in $\mathcal{G}'$ would require that there is a triple X, Y, Z in $\mathcal{G}$ such that

- $X \bullet\!\!-\!\!\circ Z$ is in $\mathcal{G}''$, $\mathcal{G}'$, and in $\mathcal{G}$, and
- $X \bullet\!\!\rightarrow Y \rightarrow Z$ or $X \rightarrow Y \bullet\!\!\rightarrow Z$ is in $\mathcal{G}''$, but
- $X \bullet\!\!\rightarrow Y \circ\!\!\rightarrow Z$ or $X \circ\!\!\rightarrow Y \bullet\!\!\rightarrow Z$ is in $\mathcal{G}'$, and by assumption
- $X \bullet\!\!\rightarrow Y \circ\!\!\rightarrow Z$, or $X \circ\!\!-\!\!\circ Y \circ\!\!\rightarrow Z$ or $X \circ\!\!\rightarrow Y \bullet\!\!\rightarrow Z$, or $X \circ\!\!\rightarrow Y \circ\!\!-\!\!\circ Z$ is in $\mathcal{G}$.

Note that if either $X \bullet\!\!\rightarrow Y \circ\!\!\rightarrow Z$, or $X \circ\!\!\rightarrow Y \circ\!\!-\!\!\circ Z$ are in $\mathcal{G}$, then $X \bullet\!\!-\!\!\circ Z$ cannot be in $\mathcal{G}$ by Lemma C.1.7. Similarly, having either $X \circ\!\!-\!\!\circ Y \circ\!\!\rightarrow Z$ or $X \circ\!\!\rightarrow Y \bullet\!\!\rightarrow Z$ in $\mathcal{G}$, together with edge $X \bullet\!\!-\!\!\circ Z$ would imply a contradiction with Lemma C.1.7, as Lemma C.1.7 would insist on an arrowhead at X on edge $\langle X, Y \rangle$.

R4 The antecedent of R4 would require the presence of:

- an almost discriminating path $p = \langle X, Q_1, \ldots, Q_k, Q_{k+1} \rangle$ for Q_k in $\mathcal{G}''$, $X \notin \text{Adj}(Q_{k+1}, \mathcal{G})$, with
- $Q_k \circ\!\!-\!\!\bullet Y$ also being in $\mathcal{G}''$.
- Then $p(X, Q_k)$ is then an almost collider path in $\mathcal{G}''$, and by inspecting the definition of an almost collider path (Definition 4.5.4), it is clear that
- $\langle X, Q_1, \ldots, Q_k \rangle$ must also be an almost collider path in $\mathcal{G}'$.

However, since $\langle X, Q_1, \ldots, Q_k, Q_{k+1} \rangle$ is not a discriminating path in $\mathcal{G}'$ (otherwise, $Q_k \rightarrow Y$ would be in $\mathcal{G}'$), at least one of the edges $\langle Q_i, Y \rangle$ is of the form $Q_i \circ\!\!-\!\!\bullet Y$, $i \in \{1, \ldots, k-1\}$ in $\mathcal{G}'$. Note that since all edges $\langle Q_i, Y \rangle$, $i \in \{1, \ldots, k-1\}$ are of the form $Q_i \rightarrow Y$ in $\mathcal{G}''$, the form of all of these edges in $\mathcal{G}'$ is either $\rightarrow$ or $\circ\!\!\rightarrow$. Let $Q_j \circ\!\!\rightarrow Y$, $j \in \{1, \ldots, k-1\}$ be an edge in $\mathcal{G}'$, chosen such that there is no edge of that form with a smaller index than j.

If $j = 1$, then we know that $X \bullet\!\!\rightarrow Q_1$ cannot be in $\mathcal{G}'$, otherwise, edge mark orientations in $\mathcal{G}'$ would not be complete by R1. Examining the definition of an almost collider path, we now know that $X \bullet\!\!-\!\!\circ Q_1 \leftarrow\!\!\bullet Q_2$ and $X \circ\!\!\rightarrow Q_2$ are in $\mathcal{G}'$, Furthermore, $X \circ\!\!\rightarrow Q_2$ implies

$Q_2 \to Y$ is in $\mathcal{G}'$ by R1. Now consider the relationships between nodes X, Q_1, Q_2 and Y in $\mathcal{G}'$:

- $X \bullet\!\!\to Q_1 \leftarrow\!\!\bullet Q_2 \to Y$ is in $\mathcal{G}'$ and so are
- $X \bullet\!\!\to Q_2$, and
- $Q_1 \circ\!\!\to Y$, and in addition,
- $X \notin \text{Adj}(Y, \mathcal{G}')$.

Now, the above implies that edge mark orientations in $\mathcal{G}'$ are not complete under R11, which is a contradiction.

Next, suppose that $j > 1$ and $Q_j \circ\!\!\to Y$ is in $\mathcal{G}'$. Now, Lemma C.6.6 implies that $\langle X, Q_1, \ldots, Q_j \rangle \oplus \langle Q_j, Y \rangle$ is an almost discriminating path for Q_j in $\mathcal{G}'$. However, this now implies that edge mark orientations in $\mathcal{G}'$ are not complete under R4, which is a contradiction.

R8 Having the antecedent of R8 in $\mathcal{G}''$ but not in $\mathcal{G}'$ would require that there is a triple X, Y, Z in $\mathcal{G}$ such that

- $X \circ\!\!\to Z$ is in $\mathcal{G}''$, and
- $X \to Y \to Z$ is in $\mathcal{G}''$, but
- $X \circ\!\!\to Y \to Z$, or $X \to Y \circ\!\!\to Z$, or $X \circ\!\!\to Y \circ\!\!\to Z$ is in $\mathcal{G}'$.

Also, note that

- since $X \circ\!\!\to Z$ is in $\mathcal{G}''$, it must be that $X \circ\!\!-\!\!\circ Z$ is in $\mathcal{G}$, and
- either $X \circ\!\!\to Y \circ\!\!-\!\!\circ Z$, or $X \circ\!\!-\!\!\circ Y \circ\!\!\to Z$, or $X \circ\!\!\to Y \circ\!\!\to Z$ is in $\mathcal{G}$.

But for all the above combinations of edges in $\mathcal{G}$, we would have a contradiction with Lemma C.1.7.

R10 For the antecedent of R10 to exist in $\mathcal{G}''$, by Lemma C.6.7, we must have the following:

- $B \to C \leftarrow D$, $A \circ\!\!\to C$, $M_{11} \bullet\!\!\to C \leftarrow\!\!\bullet M_{21}$, are in $\mathcal{G}''$ and $M_{11} \notin \text{Adj}(M_{21}, \mathcal{G})$, and
- $A \circ\!\!-\!\!\bullet M_{11}$, or $A \to M_{11}$, and $A \circ\!\!-\!\!\bullet M_{21}$, or $A \to M_{21}$, and are in $\mathcal{G}''$.
- Then also, $M_{11} \bullet\!\!\to C \leftarrow\!\!\bullet M_{21}$, is also in $\mathcal{G}$ and in $\mathcal{G}'$ since we do not introduce new unshielded colliders into $\mathcal{G}'$ or into $\mathcal{G}''$, and
- similarly $A \circ\!\!-\!\!\bullet M_{11}$, or $A \to M_{11}$, and $A \circ\!\!-\!\!\bullet M_{21}$, or $A \to M_{21}$, and are in $\mathcal{G}'$ and $\mathcal{G}$.
- also, by construction of $\mathcal{G}''$, it must be that $A \circ\!\!-\!\!\circ C$ is in $\mathcal{G}$.

Now, focus on the triple A, C, and M_{11} in $\mathcal{G}$. We know that $M_{11} \bullet\!\!\to C \circ\!\!-\!\!\circ A$ is in $\mathcal{G}$ and since $M_{11} \in \text{Adj}(A, \mathcal{G})$, Lemma C.1.7 implies that $M_{11} \bullet\!\!\to A$ is in $\mathcal{G}$ as well. But that contradicts that $A \circ\!\!-\!\!\bullet M_{11}$, or $A \to M_{11}$ is in $\mathcal{G}$.

R11 For the antecedent of R11 consider the left panel of Figure 4.1. To have this graph as an induced subgraph of $\mathcal{G}''$, but not of $\mathcal{G}'$, edge $C \to D$ must have been $C \circ\!\!\to D$ in $\mathcal{G}'$. However, this would contradict that edge mark orientations in $\mathcal{G}'$ are under R1.

R12 For the antecedent of R12 to exist in $\mathcal{G}''$ we must have the following:

- $V_1 \leftrightarrow V_{k+1} \leftarrow V_k$, $k > 2$ is in $\mathcal{G}''$ and by Lemma C.6.12, $V_1 \notin \text{Adj}(V_k, \mathcal{G})$,
- $V_1 \leftrightarrow V_{k+1} \leftarrow\!\!\circ V_k$ is in $\mathcal{G}'$
- $V_1 \leftrightarrow V_{k+1} \leftarrow\!\!\circ V_k$ is in $\mathcal{G}$, since $\mathcal{G}'$ does not contain new unshielded collider compared to $\mathcal{G}$ and since we do not orient any $\circ\!\!\to$ edge in $\mathcal{G}$ as $\leftrightarrow$ in $\mathcal{G}'$.

Additionally, by the antecedent of R12, $\mathcal{G}''$, must also contain an unshielded possibly directed path from V_1 to V_k of the form $V_1 \circ\!\!-\!\!\circ V_2 \circ\!\!-\!\!\circ \ldots \circ\!\!-\!\!\circ V_{k-1} \circ\!\!-\!\!\bullet V_k$. This path is of that same form in $\mathcal{G}'$ and in $\mathcal{G}$. However, we not have a contradiction with Lemma C.1.8 as in $\mathcal{G}$ we have both a possibly directed path $V_1 \circ\!\!-\!\!\circ V_2 \circ\!\!-\!\!\circ \ldots \circ\!\!-\!\!\circ V_{k-1} \circ\!\!-\!\!\bullet V_k \circ\!\!\to V_{k+1}$ as well as the edge $V_1 \leftrightarrow V_{k+1}$.

□

R4 completeness in Theorem C.6.5

Lemma C.6.6. *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a partial mixed graph. Let $p = \langle X = Q_0, Q_1, \dots Q_k, Q_{k+1} = Y \rangle, k > 3$ be an almost discriminating path for Q_k in $\mathcal{G}$. Then $p(X, Q_k) \oplus \langle Q_k, Y \rangle$ is an almost discriminating path for Q_{k-1}.*

Proof of Lemma C.6.6. Follows from Definition 4.5.5. □

R10 completeness in Theorem C.6.5

Lemma C.6.7 (R10 Requires an Unshielded Collider)**.** *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be an ancestral partial mixed graph and $\mathcal{P}$ be an essential ancestral graph such that they have the same skeleton, share the same set of minimal collider paths, and the set of all invariant edge marks in $\mathcal{P}$ is a subset of the invariant edge marks in $\mathcal{G}$. Suppose that the edge marks in $\mathcal{G}$ are complete under R1, R2, R3, R8, R9, R11, R12. Suppose furthermore that the edge mark at A on edge $A \circ\!\!-\!\!\bullet C$ is not complete according to R10 in $\mathcal{G}$. That is, there are edges $A \circ\!\!\rightarrow C$ and $B \rightarrow C \leftarrow D$ in $\mathcal{G}$, and unshielded possibly directed paths $p_1 = \langle A, M_{11}, \dots, M_{1l} = B \rangle, l \geq 1$ and $p_2 = \langle A, M_{21}, \dots, M_{2r} = D \rangle, r \geq 1$ such that $M_{11} \neq M_{21}$ and $M_{11} \notin \text{Adj}(M_{21}, \mathcal{G})$. Then $M_{11} \bullet\!\!\rightarrow C \leftarrow\!\!\bullet M_{21}$ is an unshielded collider in $\mathcal{G}$.*

Proof of Lemma C.6.7. Without loss of generality, we will only show that $M_{11} \bullet\!\!\rightarrow C$ is in $\mathcal{G}$. If $M_{11} \equiv B$ we are done since $B \rightarrow C$ is already in $\mathcal{G}$ by assumption. Hence, suppose that $M_{11} \neq B$, that is $l > 1$ on p_1.

By Lemma C.6.8, $q_1 = p_1 \oplus \langle B, C \rangle$ is a possibly directed path from A to B in $\mathcal{G}$. Let $M_{1i}, i \in \{1, \dots, l\}$ be chosen as the node on p with a smallest index i, such that $M_{1i} \in \text{Adj}(C, \mathcal{G})$. Then $q = q_1(A, M_{1i}) \oplus \langle M_{1i}, C \rangle$ is also a possibly directed path from A to C and if $M_{1i} \neq M_{11}$, q is an unshielded path from A to C that together with $A \circ\!\!\rightarrow C$ contradicts that orientations in $\mathcal{G}$ are completed by R9. Therefore, $M_{11} \in \text{Adj}(C, \mathcal{G})$ and moreover, $M_{11} \circ\!\!-\!\!\bullet C$, or $M_{11} \rightarrow C$ is in $\mathcal{G}$ (because q is a possibly directed path).

Consider next the edge $\langle M_{11}, M_{12} \rangle$ in $\mathcal{G}$. If this edge is of the form $M_{11} \rightarrow M_{12}$ in $\mathcal{G}$, then $p_1(M_{11}, B)$ must be a directed path from M_{11} to B, due to this path being unshielded

and orientations in $\mathcal{G}$ being completed by R1. Hence, $M_{11} \to \cdots \to B \to C$ is in $\mathcal{G}$, which by Lemma C.6.11 implies that $M_{11} \to C$ must be in $\mathcal{G}$ and we are done.

Otherwise, the edge $M_{11} \circ\!\!-\!\!\bullet M_{12}$ is in $\mathcal{G}$. Then, by Corollary C.2.3, the edge $\langle A, M_{11} \rangle$ is of the form $A \circ\!\!-\!\!\circ M_{11}$.

Now consider that in the case where $l = 2$, that is $A \circ\!\!-\!\!\circ M_{11} \circ\!\!-\!\!\bullet B$ is in $\mathcal{G}$ it holds that $A \notin \text{Adj}(B, \mathcal{G})$ (since p_1 is unshielded), and that in turn implies that $A \circ\!\!\to C \leftarrow B$ is an unshielded collider. Now the fact that orientations in $\mathcal{G}$ are complete under R3, leads us to conclude that $M_{11} \bullet\!\!\to C$ is in $\mathcal{G}$, and we are done.

Otherwise, $l > 2$. Suppose next that $l = 3$. Then because p is unshielded, there is no edge between M_{11} and B. Hence, since $B \to C$ and orientations in $\mathcal{G}$ being completed under R1 implies that $M_{11} \circ\!\!\to C$, or $M_{11} \to C$ is in $\mathcal{G}$ and we are done.

Lastly, consider $l > 4$. If $M_{11} \notin \text{Adj}(B, \mathcal{G})$, we conclude that $M_{11} \circ\!\!\to C$, or $M_{11} \to C$ is in $\mathcal{G}$ by the same argument as in the previous paragraph. So suppose that $M_{11} \in \text{Adj}(B, \mathcal{G})$. Since $p_1(M_{11}, B)$ is a possibly directed path from M_{11} to B in $\mathcal{G}$, there edge between M_{11} and B is $M_{11} \circ\!\!-\!\!\circ B$ or $M_{11} \circ\!\!\to B$ or $M_{11} \to B$.

Now, let p_1^* be the path in $\mathcal{P}$ that consists of the same sequence of nodes as p_1 in $\mathcal{G}$. If $M_{11} \circ\!\!-\!\!\circ B$ is in $\mathcal{G}$, then $M_{11} \circ\!\!-\!\!\circ B$ is also in $\mathcal{P}$ and by (contrapositive of) Lemma C.1.4, the edge $M_{1l-1} \circ\!\!-\!\!\circ B$ must be on p_1^*. Then, Corollary C.2.3 implies that $p_1^*(M_{11}, B)$ is of the form $M_{11} \circ\!\!-\!\!\circ M_{12} \circ\!\!-\!\!\circ \ldots \circ\!\!-\!\!\circ B$ and since $|p_1^*(M_{11}, B)| > 2$ and $p_1^*(M_{11}, B)$ is unshielded, this leads us to a contradiction with Lemma C.1.6.

Hence, the edge $M_{11} \circ\!\!\to B$ or $M_{11} \to B$ must be in $\mathcal{G}$. Now having $M_{11} \circ\!\!\to B \to C$ or $M_{11} \to B \to C$ and orientations in $\mathcal{G}$ being closed under R2, implies that $M_{11} \circ\!\!\to B$ or $M_{11} \to B$ is in $\mathcal{G}$. □

Lemma C.6.8 (Towards Possibly Directed Path Concatenation)**.** *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be an ancestral partial mixed graph and $\mathcal{P}$ be an essential ancestral graph such that they have the same skeleton, share the same set of minimal collider paths, and the set of all invariant edge marks in $\mathcal{P}$ is a subset of the invariant edge marks in $\mathcal{G}$. Suppose that the edge marks in*

$\mathcal{G}$ are complete under R1, R2, R3, R8, R9, R11, R12. If $p = \langle P_1, \ldots, P_k \rangle$, $k \geq 1$ is an unshielded possibly directed path in $\mathcal{G}$ and if $P_k \to P_{k+1}$ is in $\mathcal{G}$, then $p \oplus \langle P_k, P_{k+1} \rangle$ is a possibly directed path in $\mathcal{G}$.

Proof of Lemma C.6.8. It is enough to show that $P_i \leftarrow\!\bullet P_{k+1}$, $i \in \{1, \ldots, k-1\rangle$ is not in $\mathcal{G}$.

This claim holds for $i = k-1$ since otherwise, $P_{k-1} \leftarrow\!\bullet P_{k+1} \leftarrow P_k$ and the fact that $\mathcal{G}$ is ancestral and that orientations are closed under R2 would imply that $P_{k-1} \leftarrow\!\bullet P_k$ is on p. And this fact would in turn contradict that p is possibly directed from P_1 to P_k.

Hence, suppose for a contradiction that $P_i \leftarrow\!\bullet P_{k+1}$ is in $\mathcal{G}$ for some $i \in \{1, \ldots, k-2\}$, and let P_j be the closest such node to P_k on p. Furthermore, note that $P_j \leftarrow\!\circ P_{k+1}$ is not in $\mathcal{G}$, as $P_j \leftarrow\!\circ P_{k+1} \leftarrow P_k$ and the fact that orientations in $\mathcal{G}$ are closed under R1 implies that P_k and P_j are adjacent. Further, R2 implies that $P_j \leftarrow\!\bullet P_k$ is in $\mathcal{G}$ thus contradicting the assumption that p is a possibly directed path in $\mathcal{G}$.

Then $P_j \leftarrow P_{k+1}$ or $P_j \leftrightarrow P_{k+1}$ is in $\mathcal{G}$. Since $\mathcal{G}$ is an ancestral graph, we can also conclude that $p(P_j, P_k)$ is not a directed path from P_j to P_k. Furthermore, since orientations in $\mathcal{G}$ are completed by R1 and since p is an unshielded path, this also implies that $P_1 \circ\!\!-\!\!\circ P_2 \circ\!\!-\!\!\circ \ldots \circ\!\!-\!\!\circ P_j \circ\!\!-\!\!\bullet P_{j+1}$ is in $\mathcal{G}$ by Corollary C.2.3.

Now, let P_l be the closest node to P_1 on p such that $P_l \to \cdots \to P_k$ is in $\mathcal{G}$ and if no such node is in p, then let $P_l \equiv P_k$. Consider paths $P_j \circ\!\!-\!\!\circ \ldots \circ\!\!-\!\!\circ P_l$ and $q = p(P_l, P_k) \oplus \langle P_k, P_{k+1}, P_j \rangle$, where by construction, q is of one of the following forms $P_l \to \cdots \to P_{k+1} \to P_j$, or $P_l \to \cdots \to P_{k+1} \leftrightarrow P_j$. If $l > j+1$, these two paths contradict Lemma C.6.9. If $l = j+1$, then since orientation in $\mathcal{G}$ are completed by R1, $P_{k+1} \in \text{Adj}(P_l, \mathcal{G})$ and furthermore, Lemma C.6.11 implies that $P_l \to P_{k+1}$ is in $\mathcal{G}$. But closure under R2 implies that $P_j \leftarrow\!\bullet P_l$ contradicting our assumption that p is a possibly directed path. □

Lemma C.6.9. *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be an ancestral partial mixed graph and $\mathcal{P}$ be an essential ancestral graph such that they have the same skeleton, share the same set of minimal collider paths, and the set of all invariant edge marks in $\mathcal{P}$ is a subset of the invariant edge marks*

in $\mathcal{G}$. Suppose that the edge marks in $\mathcal{G}$ are complete under R1, R2, R3, R8, R9, R11, R12. Then there are no two paths $p = \langle V_1, \ldots, V_i \rangle, i > 1$ and $q = \langle V_i, \ldots, V_k, V_1 \rangle$, $k > i$ in $\mathcal{G}$ such that p and q have the same endpoint nodes and are of the following forms:

(1) p is an unshielded path of the form $V_1 \circ\!\!-\!\!\circ V_2 \ldots \circ\!\!-\!\!\circ V_{i-1} \circ\!\!-\!\!\bullet V_i$, and

(2) q is one of the following forms

(i) $V_i \to \cdots \to V_k \to V_1$, or

(ii) $V_i \to \cdots \to V_k \bullet\!\!\to V_1$, or

(iii) $V_i \bullet\!\!\to V_{i+1} \to \cdots \to V_k \to V_1$, or

(iv) $V_i \to \cdots \to V_j \bullet\!\!\to V_{j+1} \to \cdots \to V_k \to V_1$, $k > j > i$.

Proof of Lemma C.6.9. Suppose for a contradiction that there are two paths with the same endpoints that are of the forms as discussed in (1) and (2) in $\mathcal{G}$. Choose among all such pairs in $\mathcal{G}$ the paths p and q with endpoints V_1 and V_i such that for any other pair of paths p' and q' with endpoints $V_1^{'}$ and $V_i^{'}$ and such that p' is of the form (1), and q' is of the form (2), the following hold: either $|p| < |p'|$ and $|q| \leq |q'|$, or $|p| = |p'|$ and $|q| \leq |q'|$.

By choice of p and q, there cannot be any subsequence of q that forms a path in $\mathcal{G}$, that is of one of the forms: (2)(i) - (2)(iv). In conjunction with Lemma C.6.11, we then have that there cannot be any edge between any two non-consecutive nodes on q. Hence, q is an unshielded path. This further implies that $V_i \notin \text{Adj}(V_1, \mathcal{G})$, and hence, $i > 2$ on p.

Next, consider path p. By assumption p is an unshielded path and above we concluded that $|p| > 1$. Additionally, by Lemma C.6.10, there is no edge of the form $V_l \leftarrow\!\!\bullet V_r$, $1 \leq l < r \leq i$ in $\mathcal{G}$. By the same reasoning, there is also no edge of the form $V_l \bullet\!\!\to V_r$, for $1 \leq l < r \leq i - 1$. Furthermore, by choice of p and q there also cannot be an edge $V_l \to V_i$, or $V_l \circ\!\!\to V_i$, $1 \leq l < i$ in $\mathcal{G}$. Lastly, by choice of p there also cannot be an edge of the form $V_l \circ\!\!-\!\!\circ V_r$, $1 \leq l < r \leq i$ in $\mathcal{G}$. Hence, not only is p unshielded, but similarly to q, there is no edge between any two non-consecutive nodes on p.

Revisiting the fact that q is unshielded, together with the assumption that orientations in $\mathcal{G}$ are complete by R1, the $\bullet\!\!\rightarrow$ edge on q must be either $\rightarrow$, or (if q starts with $\circ\!\!\rightarrow$ as in case (2)(iii), then p must end with $\circ\!\!-\!\!\circ$ and we just redefine p to include the $\circ\!\!\rightarrow$ edge). We now break the rest of the proof up into cases depending on the form of q.

(i) Since $V_k \rightarrow V_1 \circ\!\!-\!\!\circ V_2$ is in $\mathcal{G}$, and since orientations in $\mathcal{G}$ are closed under R1, $V_k \in \text{Adj}(V_2, \mathcal{G})$.The edge $\langle V_k, V_2\rangle$ cannot be of the form $V_k \bullet\!\!\rightarrow V_2$ as that contradicts the choice of q (this would fall under case (ii)). It also cannot be of the form $V_k \leftarrow\!\!\bullet V_2$ as that contradicts that orientations are completed under R2.

Hence, $V_k \circ\!\!-\!\!\circ V_2$ must be in $\mathcal{G}$. Since we now have that $V_{k-1} \rightarrow V_k \circ\!\!-\!\!\circ V_2$ is in $\mathcal{G}$, and since orientations in $\mathcal{G}$ are closed under R1, $\langle V_{k-1}, V_2\rangle$ is in $\mathcal{G}$. By the same reasoning as above, we now have that $V_{k-1} \circ\!\!-\!\!\circ V_2$ must be in $\mathcal{G}$. However, now $\mathcal{G}$ contains the unshielded triples $V_{k-1} \rightarrow V_k \rightarrow V_1$ and $V_{k-1} \circ\!\!-\!\!\circ V_2 \circ\!\!-\!\!\circ V_1$ and edge $V_k \circ\!\!-\!\!\circ V_2$ which contradicts that orientations in $\mathcal{G}$ are completed according to R11.

(ii) Since we already discussed the case when q is a directed path in (i), we will assume that $V_k \leftrightarrow V_1$ is in $\mathcal{G}$. Furthermore, since orientations in $\mathcal{G}$ are complete according to R12, we know that $|q| > 2$, that is, $k > i + 1$.

Since $V_k \leftrightarrow V_1 \circ\!\!-\!\!\circ V_2$ is in $\mathcal{G}$, and since orientations in $\mathcal{G}$ are closed under R1, $V_k \in \text{Adj}(V_2, \mathcal{G})$. Note, furthermore, that the edge $\langle V_k, V_2\rangle$ cannot be of the form $V_k \bullet\!\!\rightarrow V_2$ as that contradicts the choice of q. Additionally, $V_k \leftarrow V_2$ contradicts that orientations are completed under R2, since in this case $V_2 \rightarrow V_k \bullet\!\!\rightarrow V_1$ and $V_1 \circ\!\!-\!\!\circ V_2$ would be in $\mathcal{G}$.

Thus, $V_k \leftarrow\!\!\circ V_2$ or $V_k \circ\!\!-\!\!\circ V_2$ are in $\mathcal{G}$. Let us first consider the case where $V_k \circ\!\!-\!\!\circ V_2$ in $\mathcal{G}$. Now have that $V_{k-1} \rightarrow V_k \circ\!\!-\!\!\circ V_2$ is in $\mathcal{G}$, so that since orientations in $\mathcal{G}$ are closed under R1, $\langle V_{k-1}, V_2\rangle$ is in $\mathcal{G}$. Note that $V_{k-1} \leftarrow\!\!\bullet V_2$ contradicts that orientations are completed under R2, and $V_{k-1} \bullet\!\!\rightarrow V_2$ contradicts the choice of q. Hence $V_{k-1} \circ\!\!-\!\!\circ V_2$ is in $\mathcal{G}$. But now, the unshielded collider $V_{k-1} \rightarrow V_k \leftrightarrow V_1$, and $V_{k-1} \circ\!\!-\!\!\circ V_2 \circ\!\!-\!\!\circ V_1$ and $V_k \circ\!\!-\!\!\circ V_2$ contradict that orientations in $\mathcal{G}$ are closed under R3.

Hence, it is left to consider the case when $V_k \hookleftarrow V_2$ is in $\mathcal{G}$. Consider that $p(V_2, V_i)$ is a possibly directed path in $\mathcal{G}$ and that $q(V_i, V_k)$ is a directed path in $\mathcal{G}$ and moreover, that there cannot be any edge $V_l \hookleftarrow V_r$, $2 \leq l < r \leq k$ in $\mathcal{G}$ as that contradicts either that $\mathcal{G}$ is ancestral, or the choice of p and q. Hence $t = p(V_2, V_i) \oplus q(V_i, V_k)$ is a possibly directed path in $\mathcal{G}$.

Note that if there is any edge $\langle V_l, V_r \rangle$, $2 \leq l < i < r \leq k$ in $\mathcal{G}$, by choice of p and q, this edge cannot be of the form $V_l \rightarrow V_r$, or $V_l \hookleftarrow V_r$. Hence, any such edge must be of the form $V_l \circ\!\!-\!\!\bullet V_r$.

Furthermore, consider any edge $\langle V_l, V_k \rangle$ $2 \leq l < i$. Then since $V_1 \leftrightarrow V_k$ is in $\mathcal{G}$ and $V_1 \notin \text{Adj}(V_l, \mathcal{G})$ and since orientations in $\mathcal{G}$ are completed under R1, we can conclude that the $\bullet$ on edge $V_l \circ\!\!-\!\!\bullet V_k$ must be an arrowhead, that is $V_l \circ\!\!\rightarrow V_k$. Now, let $V_s, s \in \{2, \ldots, i-1\}$ be the closest node to V_i on t such that $V_s \circ\!\!\rightarrow V_k$ is in $\mathcal{G}$.

Consider again that any edge $\langle V_l, V_r \rangle$, $2 \leq l < i < r \leq k$ in $\mathcal{G}$ must be of the form $V_l \circ\!\!-\!\!\bullet V_r$. Since orientations in $\mathcal{G}$ are closed under R12 and $V_{k-1} \rightarrow V_k \leftrightarrow V_1$ is in $\mathcal{G}$ and $V_1 \notin \text{Adj}(V_{k-1}, \mathcal{G})$ this implies that we cannot have an edge $\langle V_2, V_{k-1} \rangle$ in $\mathcal{G}$. Moreover, if $i > 3$, then since $p(V_1, V_3)$ is an unshielded path of the form $V_1 \circ\!\!-\!\!\circ V_2 \circ\!\!-\!\!\circ V_3$ and since $V_1, V_2 \notin \text{Adj}(V_{k-1}, \mathcal{G})$, we also cannot have an edge $\langle V_3, V_{k-1} \rangle$ in $\mathcal{G}$. We can apply the same reasoning to conclude that $V_2, \ldots, V_{i-1} \notin \text{Adj}(V_{k-1}, \mathcal{G})$. Hence, also $V_s, V_{s+1}, \ldots, V_{i-1} \notin \text{Adj}(V_{k-1}, \mathcal{G})$.

Now we have the following $V_s \circ\!\!\rightarrow V_k$ is in $\mathcal{G}$, $V_s, \ldots V_{i-1} \notin \text{Adj}(V_{k-1}, \mathcal{G})$, $V_{s+1}, \ldots, V_i \notin \text{Adj}(V_k, \mathcal{G})$ and $t(V_s, V_k)$ is a possibly directed path of the form, $V_s \circ\!\!-\!\!\circ \ldots \circ\!\!-\!\!\circ V_{i-1} \circ\!\!-\!\!\bullet V_i \rightarrow \cdots \rightarrow V_{k-1} \rightarrow V_k$. Now we can choose nodes V_a and V_b such that V_a is on $t(V_s, V_i)$, $a \neq s$, and V_b is on $t(V_i, V_k)$, $b \notin \{k-1, k\}$, and $V_a \circ\!\!-\!\!\bullet V_b$ is in $\mathcal{G}$ ($a = i - 1$ and $b = i$ is a valid choice, so such pairs a, b exist). Moreover, we can choose V_a, V_b such that $w = t(V_s, V_a) \oplus \langle V_a, V_b \rangle \oplus t(V_b, V_k)$ is an unshielded possibly directed path in $\mathcal{G}$. Then note that $V_s \circ\!\!\rightarrow V_k$ is also in $\mathcal{G}$ and that $V_{s+1} \notin \text{Adj}(V_k, \mathcal{G})$ by choice of s, which contradicts with orientations in $\mathcal{G}$ being closed under R9.

(iii) and (iv) Since $V_k \rightarrow V_1 \circ\!\!-\!\!\circ V_2$ is in $\mathcal{G}$ and since orientations in $\mathcal{G}$ are closed under R1, $V_k \in \text{Adj}(V_2, \mathcal{G})$. As in the proof of case (i), the edge $\langle V_k, V_2 \rangle$ must be of the form $V_k \circ\!\!-\!\!\bullet V_2$ is in $\mathcal{G}$. Now, $V_{k-1} \bullet\!\!\rightarrow V_k \circ\!\!-\!\!\bullet V_2$ and orientations in $\mathcal{G}$ being completed under R1 imply that edge $\langle V_{k-1}, V_2 \rangle$ is in $\mathcal{G}$. Furthermore, as q is unshielded, we know that $V_{k-1} \notin \text{Adj}(V_1, \mathcal{G})$. Putting it all together, we now have that $V_2 \bullet\!\!\bullet V_{k-1} \bullet\!\!\rightarrow V_k \rightarrow V_1$, $V_1 \circ\!\!-\!\!\circ V_2 \bullet\!\!-\!\!\circ V_k$, and $V_{k-1} \notin \text{Adj}(V_1, \mathcal{G})$, which contradicts that orientations in $\mathcal{G}$ are completed under R11.

□

Lemma C.6.10 (Possibly Directed Status of an Unshielded Path)**.** *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be an ancestral partial mixed graph and $\mathcal{P}$ be an essential ancestral graph such that they have the same skeleton, share the same set of minimal collider paths, and the set of all invariant edge marks in $\mathcal{P}$ is a subset of the invariant edge marks in $\mathcal{G}$. Suppose that the edge marks in $\mathcal{G}$ are complete under R2, R9, R12. Suppose furthermore that there is an unshielded path $q = \langle V_1, V_2, \ldots, V_k \rangle, k \geq 3$ in $\mathcal{G}$ of the form $V_1 \circ\!\!-\!\!\circ V_2 \circ\!\!-\!\!\circ \ldots V_{k-1} \circ\!\!-\!\!\bullet V_k$. Then there is no edge $V_1 \leftarrow\!\!\bullet V_k$ in $\mathcal{G}$.*

Proof of Lemma C.6.10. If $k = 3$, then since q is unshielded, $V_1 \notin \text{Adj}(V_3, \mathcal{G})$. For the rest of the proof, suppose that $k > 3$ and let q^* be the path in $\mathcal{P}$ that corresponds to q in $\mathcal{G}$. Additionally, suppose for a contradiction that $V_1 \leftarrow\!\!\bullet V_k$ is in $\mathcal{G}$.

Consider the case where $V_{k-1} \circ\!\!-\!\!\circ V_k$ is in $\mathcal{P}$. Since q^* is of the form $V_1 \circ\!\!-\!\!\circ \ldots V_{k-1} \circ\!\!-\!\!\circ V_k$, $k > 3$ in $\mathcal{P}$, the edge $\langle V_1, V_k \rangle$ is of the form $V_1 \circ\!\!-\!\!\circ V_k$ in $\mathcal{P}$ by Lemma C.1.5. But now due to chordality of the circle component of $\mathcal{P}$ (Corollary C.1.3, Lemma C.1.6), q^* cannot be an unshielded path in $\mathcal{P}$. Since $\mathcal{P}$ and $\mathcal{G}$ have the same skeleton we reach a contradiction.

For the rest of the proof, we consider the case where $V_{k-1} \circ\!\!\rightarrow V_k$ is in $\mathcal{P}$, and therefore also in $\mathcal{G}$. By Lemma C.1.8, path q^* is an unshielded possibly directed path from V_1 to V_k in $\mathcal{P}$. Further, it also ends with an arrowhead pointing to V_k. Hence, Lemma C.1.4 implies that edge $\langle V_1, V_k \rangle$ in $\mathcal{P}$ is of the form $V_1 \circ\!\!\rightarrow V_k$, or $V_1 \rightarrow V_k$. Since $V_1 \leftarrow\!\!\bullet V_k$ is supposed to be in $\mathcal{G}$, we now conclude that $V_1 \circ\!\!\rightarrow V_k$ must be in $\mathcal{P}$, which further implies that $V_1 \leftrightarrow V_k$ is in $\mathcal{G}$.

Furthermore, since q^* is an unshielded possibly directed path from V_1 to V_k in $\mathcal{P}$ (Lemma C.1.8), and $k > 3$, and since $V_1 \circ\!\!\rightarrow V_k$ is in $\mathcal{P}$ and orientations in $\mathcal{P}$ are completed by R9, it follows that $\langle V_2, V_k \rangle$ is in $\mathcal{P}$. If $k = 4$, we now reach a contradiction with q being an unshielded path. Otherwise, $k > 4$, and by Lemma C.1.4 edge $\langle V_2, V_k \rangle$ is of the form $V_2 \circ\!\!\rightarrow V_k$, or $V_2 \rightarrow V_k$ in $\mathcal{P}$. Note that, $V_2 \rightarrow V_k$ cannot be in $\mathcal{P}$, otherwise $V_2 \rightarrow V_k \leftrightarrow V_1 \circ\!\!-\!\!\circ V_2$ is in $\mathcal{G}$ which contradicts that orientations in $\mathcal{G}$ are closed under R2. Hence, $V_2 \circ\!\!\rightarrow V_k$ is in $\mathcal{P}$.

Now, similarly to above, consider that $q^*(V_2, V_k)$ is an unshielded possibly directed path from V_2 to V_k in $\mathcal{G}$, and that $k > 4$, and that $V_2 \circ\!\!\rightarrow V_k$ is in $\mathcal{P}$ and orientations in $\mathcal{P}$ are completed by R9. Hence, it follows that $\langle V_3, V_k \rangle$ is in $\mathcal{P}$. If $k = 5$, we now reach a contradiction with q being an unshielded path. Otherwise, $k > 5$, and by Lemma C.1.4 edge $\langle V_3, V_k \rangle$ is of the form $V_3 \circ\!\!\rightarrow V_k$, or $V_3 \rightarrow V_k$ in $\mathcal{P}$. Note that, $V_3 \rightarrow V_k$ cannot be in $\mathcal{P}$, otherwise $V_3 \rightarrow V_k \leftrightarrow V_1$ is in $\mathcal{G}$ and $V_1 \circ\!\!-\!\!\circ V_2 \circ\!\!-\!\!\circ V_3$ is an unshielded path in $\mathcal{G}$ which contradicts that orientations in $\mathcal{G}$ are closed under R12. Hence, $V_3 \circ\!\!\rightarrow V_k$ is in $\mathcal{P}$.

Note that the above argument can be repeated to conclude that $V_4 \circ\!\!\rightarrow V_k, \ldots, V_{k-2} \circ\!\!\rightarrow V_k$ are all in $\mathcal{P}$, which ultimately leads to a contradiction with the assumption that q is an unshielded path. □

Lemma C.6.11 (Agreeable Orientations Maintain the Ancestral Property)**.** *Suppose that $\mathcal{G}$ is an ancestral partial mixed graph with orientations completed according to R1, R2, R8, R9. Suppose that there is a path $p = \langle P_1, \ldots, P_k \rangle$, $k \geq 3$ and edge $\langle P_1, P_k \rangle$ in $\mathcal{G}$. Then the following hold*

(i) If p is a directed path from P_1 to P_k, then $P_1 \rightarrow P_k$ is in $\mathcal{G}$.

(ii) If $P_i \rightarrow P_{i+1}$ for all $i \in \{1, \ldots, k-1\} \setminus \{j\}, 1 \leq j \leq k-1$ and $P_j \bullet\!\!\rightarrow P_{j+1}$, then $P_1 \bullet\!\!\rightarrow P_k$ is in $\mathcal{G}$.

Proof of Lemma C.6.11. We prove the two statements by induction on the length of p. For the base case of the induction $k = 3$, and we have that both cases (i) and (ii) hold

because $\mathcal{G}$ is an ancestral partial mixed graph and because orientations in $\mathcal{G}$ are completed under R2 and R8.

Next, we show the induction step in each of the two cases.

(i) Suppose that claim (i) holds for all paths p' of length $n \leq k$, where $k \geq 3$. Let p be a directed path with $k+1$ nodes, $p = \langle P_1, \ldots, P_{k+1} \rangle$ such that the edge $\langle P_1, P_{k+1} \rangle$ is also in $\mathcal{G}$. Let $p' = \langle P_1 = Q_1, \ldots, Q_m = P_{k+1} \rangle, m > 1$ be a shortest subsequence of p that forms a directed path from P_1 to P_{k+1} in $\mathcal{G}$. If $m \leq k$, then $P_1 \to P_{k+1}$ is in $\mathcal{G}$ by the induction assumption. Otherwise $m > k$, that is $m = k+1$ and $p' \equiv p$, meaning that p is an unshielded path in $\mathcal{G}$. Since $\mathcal{G}$ is ancestral, this edge cannot be $P_1 \leftarrow P_{k+1}$ or $P_1 \leftrightarrow P_{k+1}$. Below we argue by contradiction that edge $\langle P_1, P_{k+1} \rangle$ cannot be $P_1 \circ\!\!-\!\!\bullet P_{k+1}$ or $P_1 \circ\!\!\to P_{k+1}$ in $\mathcal{G}$.

Suppose for a contradiction that $P_1 \circ\!\!-\!\!\bullet P_{k+1}$ is in $\mathcal{G}$. Since $P_1 \circ\!\!-\!\!\bullet P_{k+1} \leftarrow P_k$ is in $\mathcal{G}$, and since orientations in $\mathcal{G}$ are completed under R1 it follows that $P_k \in \text{Adj}(P_1, \mathcal{G})$. Hence, by the induction assumption, $P_1 \to P_k$ is in $\mathcal{G}$. But this further implies that $P_1 \to P_k \to P_{k+1}$ which is a subsequence of p that is a directed path is in $\mathcal{G}$, and that contradicts that $p' \equiv p$.

Next, suppose for a contradiction that $P_1 \circ\!\!\to P_{k+1}$ is in $\mathcal{G}$. Note that since $P_1 \to \cdots \to P_k \to P_{k+1}$ is an unshielded directed path in the ancestral graph $\mathcal{G}$ and since edge mark orientations in $\mathcal{G}$ are complete according to R9, it follows that $P_2 \in \text{Adj}(P_{k+1}, \mathcal{G})$. Since $P_2 \in \text{Adj}(P_{k+1}, \mathcal{G})$ and $P_2 \to \cdots \to P_{k+1}$ is in $\mathcal{G}$, by the induction assumption, $P_2 \to P_{k+1}$ is in $\mathcal{G}$. But now $P_1 \to P_2 \to P_{k+1}$ is a subsequence of p that is a directed path is in $\mathcal{G}$. This contradicts that $p' \equiv p$.

(ii) Suppose that claim (ii) holds for all paths p' of length $n \leq k$, where $k \geq 3$. Let p be a path with $k+1$ nodes, $p = \langle P_1, \ldots, P_{k+1} \rangle$ such that $P_j \bullet\!\!\to P_{j+1}$, for some $j \in \{1, \ldots, k\}$, but $P_i \to P_{i+1}$ for all $i \in \{1, \ldots, k\} \setminus \{j\}$ and also such that the edge $\langle P_1, P_{k+1} \rangle$ is in $\mathcal{G}$.

Since $\mathcal{G}$ is ancestral, the edge $\langle P_1, P_{k+1} \rangle$ cannot be of the form $P_1 \leftarrow P_{k+1}$. Hence, for the claim (ii), it is enough to show that this edge is also not of the form $P_1 \circ\!\!-\!\!\bullet P_{k+1}$ in $\mathcal{G}$.

Suppose for a contradiction that $P_1 \circ\!\!-\!\!\bullet P_{k+1}$ is in $\mathcal{G}$. Since $P_1 \circ\!\!-\!\!\bullet P_{k+1} \leftarrow\!\!\bullet P_k$ is in $\mathcal{G}$ and since orientations in $\mathcal{G}$ are closed under R1 it follows that $P_k \in \text{Adj}(P_1, \mathcal{G})$. If $p(P_1, P_k)$ is a directed path, then by (i) above, we have that $P_1 \to P_k$ is in $\mathcal{G}$. But then $P_1 \to$

$P_k \bullet\!\!\rightarrow P_{k+1}$ together with $P_1 \bullet\!\!-\!\!\circ P_{k+1}$ contradicts that orientations in $\mathcal{G}$ are completed under R2. Otherwise, $p(P_1, P_k)$ contains either $\circ\!\!\rightarrow$ or a $\leftrightarrow$ edge, so by the induction step $P_1 \bullet\!\!\rightarrow P_k$ is in $\mathcal{G}$. However, in this case $P_1 \bullet\!\!\rightarrow P_k \rightarrow P_{k+1}$ and $P_1 \bullet\!\!-\!\!\circ P_{k+1}$ are in $\mathcal{G}$, which contradicts that orientations in $\mathcal{G}$ are closed under R2. □

R12 completeness in Theorem C.6.5

Lemma C.6.12 (R12 Requires an Unshielded Collider)**.** *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be an ancestral partial mixed graph and $\mathcal{P}$ be an essential ancestral graph such that they have the same skeleton, share the same set of minimal collider paths, and the set of all invariant edge marks in $\mathcal{P}$ is a subset of the invariant edge marks in $\mathcal{G}$. Suppose that the edge marks in $\mathcal{G}$ are complete under R1, R2, R8, R9, R11. Suppose furthermore that the edge mark at V_1 on some edge $V_1 \circ\!\!-\!\!\circ V_2$ is not complete according to R12 in $\mathcal{G}$. That is, there is an unshielded path of the form $V_1 \circ\!\!-\!\!\circ V_2 \circ\!\!-\!\!\circ \dots V_{i-1} \circ\!\!-\!\!\bullet V_i$, $i > 2$ in $\mathcal{G}$, as well as a path $V_i \rightarrow V_{i+1} \leftrightarrow V_1$ in $\mathcal{G}$. Then $V_1 \notin \mathrm{Adj}(V_i, \mathcal{G})$, that is $V_i \rightarrow V_{i+1} \leftrightarrow V_1$ is an unshielded collider.*

Proof of Lemma C.6.12. Follows directly from Lemma C.6.13. □

Lemma C.6.13. *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be an ancestral partial mixed graph and $\mathcal{P}$ be an essential ancestral graph such that they have the same skeleton, share the same set of minimal collider paths, and the set of all invariant edge marks in $\mathcal{P}$ is a subset of the invariant edge marks in $\mathcal{G}$.*

Suppose that the edge marks in $\mathcal{G}$ are complete under R1, R2, R8, R9, R11. Suppose furthermore that there is an unshielded path of the form $V_1 \circ\!\!-\!\!\circ V_2 \circ\!\!-\!\!\circ \dots V_{i-1} \circ\!\!-\!\!\bullet V_i$, $i > 2$ in $\mathcal{G}$, as well as a path $p = \langle V_i, V_{i+1}, V_1 \rangle$ that is of one of the following forms in $\mathcal{G}$: $V_i \rightarrow V_{i+1} \bullet\!\!\rightarrow V_1$, or $V_i \bullet\!\!\rightarrow V_{i+1} \rightarrow V_1$. Then $V_1 \notin \mathrm{Adj}(V_i, \mathcal{G})$.

Proof of Lemma C.6.13. If $i = 3$, then $V_1 \notin \mathrm{Adj}(V_3, \mathcal{G})$ by assumption that $V_1 \circ\!\!-\!\!\circ V_2 \circ\!\!-\!\!\bullet V_3$ is an unshielded path.

Hence, suppose that $i > 3$ and suppose for a contradiction that $V_1 \in \mathrm{Adj}(V_i, \mathcal{G})$. Let the path $V_1 \circ\!\!-\!\!\circ V_2 \circ\!\!-\!\!\circ \dots V_{i-1} \circ\!\!-\!\!\bullet V_i$ be called q in $\mathcal{G}$ and q^* in $\mathcal{P}$, $q = q^* = \langle V_1, \dots, V_i \rangle$.

We will first assume that $V_{i-1} \circ\!\!-\!\!\circ V_i$ is in $\mathcal{P}$. Since q^* is of the form $V_1 \circ\!\!-\!\!\circ \dots V_{i-1} \circ\!\!-\!\!\circ V_i$, $i > 3$ in $\mathcal{P}$, the edge $\langle V_1, V_i \rangle$ is of the form $V_1 \circ\!\!-\!\!\circ V_i$ in $\mathcal{P}$ by Lemma C.1.5. But now due to chordality of the circle component of $\mathcal{P}$ (Corollary C.1.3, Lemma C.1.6), q^* cannot be an unshielded path in $\mathcal{P}$. Since $\mathcal{P}$ and $\mathcal{G}$ have the same skeleton we reach a contradiction.

For the rest of the proof, we consider the case where $V_{i-1} \circ\!\!\rightarrow V_i$ is in $\mathcal{P}$, and therefore also in $\mathcal{G}$. By Lemma C.1.8, path q^* is an unshielded possibly directed path from V_1 to V_i in $\mathcal{P}$. Further, it also ends with an arrowhead pointing to V_i. Hence, Lemma C.1.4 implies that edge $\langle V_1, V_i \rangle$ in $\mathcal{P}$ is of the form $V_1 \circ\!\!\rightarrow V_i$, or $V_1 \rightarrow V_i$. In the latter case, we obtain a contradiction, because $V_1 \rightarrow V_i$ would also be in $\mathcal{G}$ and together with p and completed orientations under R2, R8 in $\mathcal{G}$, it would imply that $\mathcal{G}$ is not ancestral. Hence, $V_1 \circ\!\!\rightarrow V_i$ is in $\mathcal{P}$.

Now, consider the edge $\langle V_1, V_i \rangle$ and path $\langle V_i, V_{i+1}, V_1 \rangle$ in $\mathcal{G}$. Since $V_i \leftarrow\!\!\circ V_1$ is in $\mathcal{P}$, and since $\mathcal{G}$ is ancestral $V_i \leftarrow\!\!\circ V_1$, or $V_i \leftrightarrow V_1$ is in $\mathcal{G}$. Furthermore, since $V_i \rightarrow V_{i+1} \bullet\!\!\rightarrow V_1$, or $V_i \bullet\!\!\rightarrow V_{i+1} \rightarrow V_1$ is in $\mathcal{G}$ and since orientations in $\mathcal{G}$ are completed by R2, it must be that $V_i \leftrightarrow V_1$ is in $\mathcal{G}$. By analogous reasoning we furthermore have that path $\langle V_i, V_{i+1}, V_1 \rangle$ in $\mathcal{G}$, must be of one of the following forms $V_i \rightarrow V_{i+1} \leftrightarrow V_1$, or $V_i \leftrightarrow V_{i+1} \rightarrow V_1$ otherwise, we have a contradiction with $\mathcal{G}$ being ancestral, or with the orientations in $\mathcal{G}$ being completed under R2. Hence, for the rest of the proof, note that $V_i \rightarrow V_{i+1} \leftrightarrow V_1$, or $V_i \leftrightarrow V_{i+1} \rightarrow V_1$ is in $\mathcal{G}$.

Since $V_1 \circ\!\!\rightarrow V_i$ is in $\mathcal{P}$, q^* is an unshielded and possibly directed path from V_1 to V_i in $\mathcal{P}$ and since orientations in $\mathcal{P}$ are complete under R9 it follows that $V_2 \in \text{Adj}(V_i, \mathcal{G})$. If $i = 4$ this leads us to our final contradiction since this would imply that q^* (and therefore q) is not unshielded. Otherwise, $i > 4$ and by Lemma C.1.4, $V_2 \circ\!\!\rightarrow V_i$, or $V_2 \rightarrow V_i$ is in $\mathcal{P}$. Note that in the later case, $V_2 \rightarrow V_i$ would also be in $\mathcal{G}$ and we would have that $V_2 \rightarrow V_i \rightarrow V_{i+1} \leftrightarrow V_1 \circ\!\!-\!\!\circ V_2$, or $V_2 \rightarrow V_i \leftrightarrow V_{i+1} \rightarrow V_1 \circ\!\!-\!\!\circ V_2$ is in $\mathcal{G}$, which contradicts Lemma C.6.11. Hence, $V_2 \circ\!\!\rightarrow V_i$ is in $\mathcal{P}$.

Now, we can use the same argument as above iteratively to conclude that $V_3 \circ\!\!\rightarrow V_i, \dots, V_{i-2} \circ\!\!\rightarrow V_i$ are in $\mathcal{P}$. Hence, we obtain a contradiction with q^* and therefore q

being an unshielded path. Hence, our original supposition that $V_1 \in \text{Adj}(V_i, \mathcal{G})$ is incorrect. □

Supporting results for Theorem 4.6.7, Part 2: Ancestrality and maximality

As a reminder, in Theorem 4.6.7, we obtain a graph $\mathcal{G}''$ by orienting all $\circ\!\!\rightarrow$ edges in the essential ancestral graph as tails i.e., $\rightarrow$. Earlier, in Theorem C.6.5, we showed that the edge mark orientations in this graph are complete under the orientation rules. In Theorem C.6.14 below, we use the results of Theorem C.6.5 to show that this graph is ancestral and preserves the minimal collider paths present in the essential ancestral graph.

Theorem C.6.14. *Let $\mathcal{G}' = (\mathbf{V}, \mathbf{E})$ be an ancestral partial mixed graph and $\mathcal{G}$ be an essential ancestral graph such that $\mathcal{G}$ and $\mathcal{G}'$ have the same skeleton, the same set of minimal collider paths, and the set of all invariant edge marks in $\mathcal{G}$ is a subset of the invariant edge marks in $\mathcal{G}'$. Suppose furthermore, that every edge $A \circ\!\!\rightarrow B$ in $\mathcal{G}$ corresponds either to $A \rightarrow B$ or to $A \circ\!\!\rightarrow B$ in $\mathcal{G}'$ and that edge mark orientations in $\mathcal{G}'$ are closed under R1-R4, R8-R12. Let $\mathcal{G}''$ be identical to $\mathcal{G}'$ except all $A \circ\!\!\rightarrow B$ edges in $\mathcal{G}$ correspond to $A \rightarrow B$ edges in $\mathcal{G}''$. Then $\mathcal{G}''$ is ancestral, shares the same set of minimal collider paths with $\mathcal{G}'$ and $\mathcal{G}$ and edge mark orientations in $\mathcal{G}''$ are closed under R1-R4, R8-R12.*

Proof of Theorem C.6.14. By Theorem C.6.5, the orientations in $\mathcal{G}''$ are complete under R1-R4, R8-R12, so we only need to show that $\mathcal{G}''$ is ancestral and also that it does not contain any new minimal collider paths compared to $\mathcal{G}'$. The latter follows immediately since we do not introduce any arrowheads in $\mathcal{G}''$, or remove edges compared to $\mathcal{G}'$.

Suppose for a contradiction that there is a directed or almost directed cycle in $\mathcal{G}''$. By Lemma 4.6.8, there is also one such cycle of length 3 in $\mathcal{G}''$. Let $X \rightarrow Y \rightarrow Z$, $X \leftarrow\!\!\circ Z$ be one such cycle in $\mathcal{G}''$. Since $\mathcal{G}'$ is ancestral, we know that the corresponding edges in $\mathcal{G}'$ are in one of the following categories:

(a) $X \circ\!\!\rightarrow Y \rightarrow Z$ and $X \leftarrow\!\!\circ Z$.

(b) $X \to Y \circ\!\!\to Z$ and $X \leftarrow\!\!\bullet Z$.

(c) $X \circ\!\!\to Y \circ\!\!\to Z$ and $X \leftarrow Z$.

(d) $X \circ\!\!\to Y \circ\!\!\to Z$ and $X \leftarrow\!\!\circ Z$.

(e) $X \circ\!\!\to Y \circ\!\!\to Z$ and $X \leftrightarrow Z$.

Note that cases (a)-(c) contradict that edge mark orientations in $\mathcal{G}'$ are complete under R2 and R8. The edges in case (d)-(e) must be of that same form in $\mathcal{G}$. However, (d) contradicts Lemma C.1.8 and (e) contradicts Lemma C.1.7.

□

Lemma C.6.15 is another final auxiliary result used in proving Theorem 4.6.7. This result is crucial in showing why we can work with the circle and non-circle components independently.

Lemma C.6.15. *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be an ancestral partial mixed graph and $\mathcal{P}$ be an essential ancestral graph such that they have the same skeleton, share the same set of minimal collider paths, and the set of all invariant edge marks in $\mathcal{P}$ is a subset of the invariant edge marks in $\mathcal{G}$. Suppose furthermore, that all $A \circ\!\!\to B$ edges in $\mathcal{P}$ correspond to $A \to B$ in $\mathcal{G}$ and that orientations in $\mathcal{G}$ are closed under R1-R4, R8-R12. Suppose that $C \bullet\!\!\bullet D$ is an edge in $\mathcal{G}$ that corresponds to $C \circ\!\!\circ D$ in $\mathcal{P}$ and that $E \bullet\!\!\to C$ is an edge in $\mathcal{G}$ such that the corresponding edge in $\mathcal{P}$ is one of the following: $E \circ\!\!\to C$, $E \to C$, or $E \leftrightarrow C$. Then either*

(i) $E \to C$ and $E \to D$ are in $\mathcal{G}$, or

(ii) $E \leftrightarrow C$ and $E \leftrightarrow D$ are in $\mathcal{G}$.

Proof of Lemma C.6.15. Since $E \bullet\!\!\to C$ is in $\mathcal{G}$ and and since the arrowhead at C on $\langle E, C \rangle$ is also in $\mathcal{P}$ it follows by Lemma C.1.7 that $E \bullet\!\!\to D$ is in $\mathcal{P}$ as well. The arrowhead at D

on edge $\langle E, D\rangle$ is also in $\mathcal{G}$. Moreover, edge $\langle E, D\rangle$ is of one of the following forms in $\mathcal{P}$, $E \to D$, $E \circ\!\!\to D$, $E \leftrightarrow D$.

Furthermore, if $E \leftrightarrow C$ is in $\mathcal{P}$ then $E \circ\!\!\to D$ cannot be in $\mathcal{P}$, because the $E \leftrightarrow D \circ\!\!-\!\!\circ C$ and $C \leftarrow\!\!\circ E$ would contradict Lemma C.1.7. For the same reason, we cannot have $E \leftrightarrow D$ and $E \circ\!\!\to C$ together in $\mathcal{P}$.

Hence, one of the following appears in $\mathcal{P}$, $D \leftarrow\!\!\circ E \circ\!\!\to C$, $D \leftarrow E \circ\!\!\to C$, $D \leftarrow\!\!\circ E \to C$, $D \leftarrow E \to C$, or $D \leftrightarrow E \leftrightarrow C$. By construction of $\mathcal{G}$, we then must have either $D \leftarrow E \to C$, or $D \leftrightarrow E \leftrightarrow C$ in $\mathcal{G}$. □

C.6.5 Lemma 4.6.9

Proof of Lemma 4.6.9. We consider both directions below.

⇐: First, assume that conditions (i) and (ii) are satisfied. Consider a discriminating collider path $q_{\mathcal{G}'} = \langle X, Q_1, \ldots, Q_k, B, Y\rangle$ in $\mathcal{G}'$.

Since we know that $B \leftarrow\!\!\bullet Y$ is in $\mathcal{G}$ and that edges $\langle Q_k, B\rangle$ and $\langle Q_k, Y\rangle$ are also in $\mathcal{G}$, we can reason about the edge mark at B on edge $\langle Q_k, B\rangle$ in $\mathcal{G}$. Note that $Q_k \bullet\!\!-\!\!\circ B$ cannot be in $\mathcal{G}$, since otherwise Lemma C.1.7 would imply that $Q_k \leftarrow\!\!\bullet Y$ is also in $\mathcal{G}$. This in turn would require that $Q_k \leftarrow\!\!\bullet Y$ is also in $\mathcal{G}'$ by construction, which we know is not the case due to $q_{\mathcal{G}'}$ being a discriminating collider path in $\mathcal{G}'$. Similarly, $Q_k \leftarrow B$ cannot be in $\mathcal{G}$, since $Q_k \leftrightarrow B$ is in $\mathcal{G}'$. Therefore, we can conclude that $Q_k \bullet\!\!\to B \leftarrow\!\!\bullet Y$ is in $\mathcal{G}'$.

Then, since $\mathcal{G}'$ is ancestral and edge mark orientations in $\mathcal{G}'$ are completed by R1, R2, and R4 it follows by Lemma C.6.3 that every triple $\langle V_{i-1}, V_i, V_{i+1}\rangle$, $1 < i < k$ on a minimal collider path $p_{\mathcal{G}'} = \langle V_1, \ldots, V_k\rangle, k \geq 3$ in $\mathcal{G}'$ is either an unshielded collider in $\mathcal{G}'$ or the last collider on a discriminating collider path in $\mathcal{G}'$. By the assumptions (i) and (ii), and the reasoning above, it then follows that the edge marks at V_i on $\langle V_{i-1}, V_i, V_{i+1}\rangle$ are identical in $\mathcal{G}$.

Since this holds for every triple on $p_{\mathcal{G}'}$, that implies that the same sequence of nodes

$\langle V_1, \ldots, V_k \rangle$ in $\mathcal{G}$ is a collider path. Furthermore, since this holds for every minimal collider path $p_{\mathcal{G}'} = \langle V_1, \ldots, V_k \rangle$ in $\mathcal{G}'$, and since $\mathcal{G}'$ contains more invariant edge marks than $\mathcal{G}$, we have that the all of the corresponding sequences of nodes in $\mathcal{G}$, $p_{\mathcal{G}} = \langle V_1, \ldots, V_k \rangle$, must also be minimal collider paths in $\mathcal{G}$.

$\Rightarrow$: Suppose that every minimal collider path $p_{\mathcal{G}'} = \langle V_1, \ldots, V_k \rangle, k > 1$ in $\mathcal{G}'$ corresponds to a minimal collider path $p_{\mathcal{G}} = \langle V_1, \ldots, V_k \rangle, k > 1$ in $\mathcal{G}$. Note that by the construction of $\mathcal{G}'$, not only is every minimal collider path in $\mathcal{G}$ a minimal collider path in $\mathcal{G}'$, but also every collider path in $\mathcal{G}$ is a collider path in $\mathcal{G}'$.

Since every unshielded collider is a minimal collider path, the unshielded colliders in $\mathcal{G}$ and $\mathcal{G}'$ must be identical. Hence, (i) holds.

Furthermore, every discriminating collider path $q_{\mathcal{G}'} = \langle X, Q_1, \ldots, Q_m, B, Y \rangle, m \geq 1$ in $\mathcal{G}'$ that is also a minimal collider path in $\mathcal{G}'$, will definitely satisfy (ii) in $\mathcal{G}$. Now, suppose that $q_{\mathcal{G}'} = \langle X, Q_1, \ldots, Q_m, B, Y \rangle, m \geq 1$ is a discriminating collider path in $\mathcal{G}'$, but not a minimal collider path in $\mathcal{G}'$. Then there must be a subsequence $q'_{\mathcal{G}'}$ of $q_{\mathcal{G}'}$ that is a minimal collider path in $\mathcal{G}'$. Then, the edge $\langle B, Y \rangle$ must be on $q'_{\mathcal{G}'}$. This is because, $Q_i \to Y$ is in $\mathcal{G}'$, for all $i \in \{1, \ldots, m\}$ and $X \notin \text{Adj}(Y, \mathcal{G}')$ by definition of a discriminating collider path. The corresponding path to $q'_{\mathcal{G}'}$ in $\mathcal{G}$ is then a collider path, and since $\langle B, Y \rangle$ is an edge on this path and B is not an endpoint on it $B \leftarrow\!\bullet Y$ is in $\mathcal{G}$.

□

C.7 Completeness of Edge Mark Orientations in Ancestral Partial Mixed Graphs with no Minimal Collider Paths

Consider a maximal ancestral graph $\mathcal{G}'$ that is obtained as the output of Algorithm 5. We examine edge orientations of $\mathcal{G}'_C$, which is the induced subgraph of $\mathcal{G}'$ that corresponds to the circle component of the essential ancestral graph $\mathcal{G}$. We show that edge orientations within these graphs are complete using an argument similar to Meek (1995).

Since the skeleton of $\mathcal{G}'_C$ is chordal (Zhang, 2008b), we can construct join trees on their maximal cliques (see Chapter 3.2 of Lauritzen (1996), and Theorem C.7.8 below). Similar to Meek (1995), we define a total ordering of maximal cliques in a join tree and show that this ordering induces a partial ordering of nodes in $\mathcal{G}'_C$ that is consistent with prior edge mark orientations, maintains the ancestral property and does not introduce any minimal collider paths. Then, we show how to select two MAGs represented by $\mathcal{G}'_C$ as extensions of these orderings with the required orientations of an edge in question.

Our main result is presented in Theorem C.7.12 in Section C.7.2. A map of how all results in this section are used to prove Theorem C.7.12 is given in Figure C.4 of Section C.7.2. Throughout this section we also include examples for all intermediate results and algorithms, concluding with Example C.7.34, which demonstrates the constructive process for obtaining the MAGs described in Theorem C.7.12.

In Table C.1 below, we make explicit the connections between our results and that of Meek (1995). The second column provides locations or specific references to results in this manuscript that are somewhat analogous to those of Meek (1995). As a consequence of Theorem C.7.12 holding in a more general case, it subsumes the result of Theorem 4 of Meek (1995) and the proof of Theorem C.7.12 provides an alternative proof to Theorem 4 of Meek (1995).

Meek (1995)	Our Results	Examples	Location
Lemma 6	Lemma C.7.13	Example C.7.14	Section C.7.3
Lemma 7	Lemma C.7.15	Examples C.7.16-C.7.18	Section C.7.4
Lemmas 4 and 8	Algorithm 18 and Lemma C.7.31	–	Section C.7.5

Table C.1: Locating Analogous Results to Meek (1995).

C.7.1 Section Specific Preliminaries

Definition C.7.1 (Partial Order)**.** *Consider a set of elements* $\mathbf{V}$*. A relation* $\leq$*, between the elements of* $\mathbf{V}$ *is called a partial order if and only if for every* $X, Y, Z \in \mathbf{V}$

(i) reflexive: $X \leq X$*,*

(ii) antisymmetric: if $X \leq Y$ *and* $Y \leq X$*, then* $X = Y$*, and*

(iii) transitive: if $X \leq Y$ *and* $Y \leq Z$*, then* $X \leq Z$*.*

Remark. If a pairwise relation π on a set of elements of $\mathbf{V}$ is a partial ordering, then for elements $A, B \in \mathbf{V}$ such that $\pi(A, B)$ holds, we will also write $A \leq_\pi B$. Note also, that not every two elements of $\mathbf{V}$ need to be comparable to have a partial ordering on $\mathbf{V}$. For distinct elements X and Y in $\mathbf{V}$, if $X \not\leq Y$ and $Y \not\leq X$, then we say that X and Y are incomparable and we denote this by $X \not\leqslant Y$ or, equivalently, $Y \not\leqslant X$ (Trotter, 1992).

Definition C.7.2 (Extending Orders)**.** *A partial order* π_1 *is an extension of a partial order* π_2 *if and only if* $A \leq_{\pi_2} B$ *implies* $A \leq_{\pi_1} B$*.*

Definition C.7.3 (Compatible Order)**.** *Let* $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ *be a partial mixed graph. A partial order* π *over* $\mathbf{V}$ *is compatible with* $\mathcal{G}$ *if and and only if for any pair of nodes* A *and* B *in* $\mathcal{G}$

- *if* $A \to B$ *is in* $\mathcal{G}$*, then* $A \leq_\pi B$*,*
- *if* $A \leftarrow\!\bullet B$ *is in* $\mathcal{G}$*, then* $A \not\leq_\pi B$*.*

Definition C.7.4 (Induced Orientation)**.** *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a partially directed mixed graph and let $\leq_\alpha$ be a partial order on $\mathbf{V}$ that is compatible with $\mathcal{G}$. Then $\leq_\alpha$ induces a partial orientation as follows:*

- *if $A \circ\!\!-\!\!\bullet B$ is in $\mathcal{G}$ and $A \leq_\alpha B$, or $\alpha(A, B)$, then orient $A \to B$.*

The graph resulting from applying the above procedure is called $\mathcal{G}_\alpha$.

Lemma C.7.5 (Ancestral Property Induces a Compatible Partial Order)**.** *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a partially directed ancestral mixed graph. Let π be a relation on the nodes of $\mathcal{G}$ induced by the ancestral relationships. That is $\pi(A, B)$ if and only if $A \in \mathrm{An}(B, \mathcal{G})$. Then π is a partial ordering of $\mathbf{V}$ that is compatible with $\mathcal{G}$.*

Proof of Lemma C.7.5. By definition, every node in $\mathcal{G}$ is an ancestor of itself, hence π is a reflexive relationship. To show that π is antisymmetric note that $\mathcal{G}$ is ancestral, so if $A \in \mathrm{An}(B, \mathcal{G})$, that is $\pi(A, B)$ and $B \in \mathrm{An}(A, \mathcal{G})$, that is $\pi(B, A)$ holds, we must have $A \equiv B$. The transitive property also holds by definition. Therefore, π is a partial ordering that is naturally compatible with $\mathcal{G}$. □

Definition C.7.6 (Tree Graph)**.** *A graph $\mathcal{T} = (\mathbf{V}, \mathbf{E})$ is a tree if for any pair of nodes $A, B \in \mathbf{V}$, there is exactly one path $p = \langle A = V_1, \ldots, B = V_k \rangle$ in $\mathcal{T}$.*

Definition C.7.7 (Join Tree Graph)**.** *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a graph. A join tree graph $\mathcal{T} = (\mathbf{C}, \mathbf{E}')$ for $\mathcal{G}$ is an undirected tree graph whose nodes $\mathbf{C}$ are a partition of $\mathbf{V}$ with the following properties:*

(i) for set of nodes $\mathbf{A} \subseteq \mathbf{V}$ that forms a maximal clique in $\mathcal{G}$, $\mathbf{A} \equiv \mathcal{C}_i$, for some $\mathcal{C}_i \in \mathbf{C}$, and

(ii) (running intersection) for each pair $\mathcal{C}_i, \mathcal{C}_j \in \mathbf{C}$ such that $A \in (\mathcal{C}_i \cap \mathcal{C}_j) \subseteq \mathbf{V}$, each node $\mathcal{C}_k$ on the unique path between $\mathcal{C}_i$ and $\mathcal{C}_j$ in $\mathcal{T}$ also contains A.

Remark. Join trees are sometimes also called junction trees or chordal trees, due to the fact that only chordal graphs have a join tree. We state the original result of Beeri et al. (1983) in Lemma C.7.8. We refer the reader to Jensen and Jensen (1994) and Lauritzen (1996) for a modern treatment of join trees and how to construct them.

Lemma C.7.8 (Theorem 3.4 of Beeri et al. (1983))**.** *A graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ has a join tree if and only if $\mathcal{G}$ is chordal.*

Λ_{ij} notation. Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a graph with a chordal skeleton. For maximal cliques $\mathcal{C}_i, \mathcal{C}_j \subseteq \mathbf{V}$, we will use Λ_{ij} to denote their intersection, that is, $\Lambda_{ij} = \mathcal{C}_i \cap \mathcal{C}_j$.

Definition C.7.9 (γ-relation)**.** *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be an ancestral partial mixed graph such that the skeleton of $\mathcal{G}$ is chordal and $\mathcal{G}$ contains no minimal collider paths. Let $\mathcal{T} = (\mathbf{C}, \mathbf{E}')$ be an undirected join tree graph for $\mathcal{G}$. Let $\mathcal{C}_i, \mathcal{C}_j \in \mathbf{C}$, and $\Lambda_{ij} = \mathcal{C}_i \cap \mathcal{C}_j$. We define a relation γ on the nodes of $\mathcal{T}$ as follows: $\gamma(\mathcal{C}_i, \mathcal{C}_j)$ if and only if*

(i) $\Lambda_{ij} \neq \varnothing$,

(ii) for all $B \in \Lambda_{ij}$ and $C \in \mathcal{C}_j \setminus \Lambda_{ij}$, $B \to C$ is in $\mathcal{G}$, and

(iii) there exist nodes $A \in \mathcal{C}_i \setminus \Lambda_{ij}$ and $B \in \Lambda_{ij}$ such that $A \circ\!\!\to B$ is in $\mathcal{G}$.

Definition C.7.10 (Partially Directed Join Tree)**.** *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be an ancestral partial mixed graph such that the skeleton of $\mathcal{G}$ is chordal and $\mathcal{G}$ contains no minimal collider paths. Let $\mathcal{T} = (\mathbf{C}, \mathbf{E}')$ be an undirected join tree graph for $\mathcal{G}$ and let γ be a relation on the nodes of $\mathcal{T}$ defined in Definition C.7.9. We define a partially directed join tree graph $\mathcal{T}_\gamma = (\mathbf{C}, \mathbf{E}'')$ as follows:*

(i) The skeleton of $\mathcal{T}_\gamma$ is identical to the skeleton of $\mathcal{T}$.

(ii) Edge $\langle \mathcal{C}_i, \mathcal{C}_j \rangle$ in $\mathcal{T}$ corresponds to:

- *$\mathcal{C}_i \to \mathcal{C}_j$ in $\mathcal{T}_\gamma$ if $\gamma(\mathcal{C}_i, \mathcal{C}_j)$,*

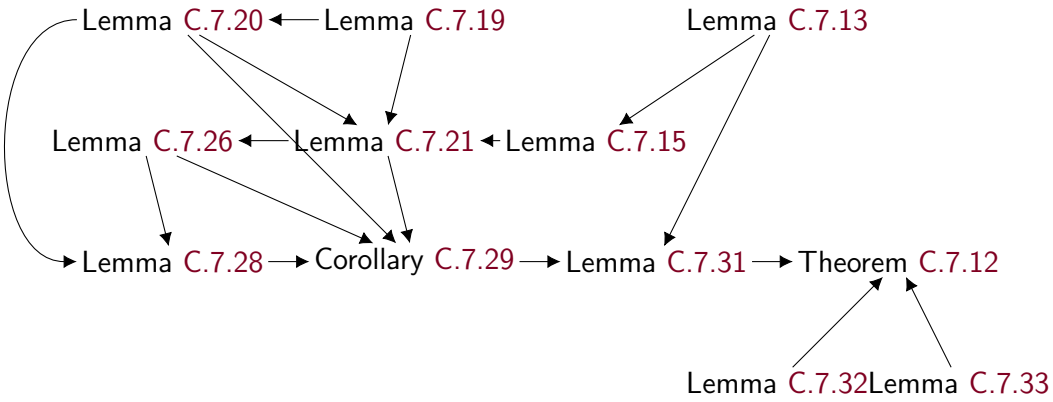


Figure C.4: Proof structure of Theorem C.7.12

- $\mathcal{C}_i \leftarrow \mathcal{C}_j$ *in* $\mathcal{T}_\gamma$ *if* $\gamma(\mathcal{C}_j, \mathcal{C}_i)$, *and*
- $\mathcal{C}_i - \mathcal{C}_j$ *in* $\mathcal{T}_\gamma$ *if neither* $\gamma(\mathcal{C}_i, \mathcal{C}_j)$ *nor* $\gamma(\mathcal{C}_j, \mathcal{C}_i)$.

Remark. Note that the partially or fully directed trees we consider are not always arborescences in the graph theory sense. Meaning that our definition of a partially directed tree allows for more than one root node.

Definition C.7.11 (Join Tree Induced Edge Orientations)**.** *Let* $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ *be an ancestral partial mixed graph such that the skeleton of* $\mathcal{G}$ *is chordal and* $\mathcal{G}$ *contains no minimal collider paths. Let* $\mathcal{T} = (\mathbf{C}, \mathbf{E}')$ *be a partially directed join tree graph for* $\mathcal{G}$ *(Definition C.7.10) and suppose that* $\pi_\mathcal{T}$ *is a partial ordering compatible with* $\mathcal{T}$*, such that* $\mathcal{T}_{\pi_\mathcal{T}}$ *is a directed graph with no colliders. Then,* $\pi_\mathcal{T}$ *induces orientations on the nodes of* $\mathcal{G}$ *using the following rule:*

(i) if $\pi_\mathcal{T}(\mathcal{C}_i, \mathcal{C}_j)$, *then for all* $B \in \mathcal{C}_i \cap \mathcal{C}_j$ *and* $C \in \mathcal{C}_j \setminus \mathcal{C}_i$, *orient* $B \to C$.

The graph obtained as a result of this operation is called $\mathcal{G}_\pi$.

C.7.2 Main result

Theorem C.7.12. *Suppose* $\mathcal{G}$ *is a maximal and ancestral partial mixed graph with no minimal collider paths, such that the skeleton of* $\mathcal{G}$ *is chordal and such that the edge orientations*

in $\mathcal{G}$ are complete under R1-R4, R8-R12.

(i) If $A \circ\!\!-\!\!\circ B$ is in $\mathcal{G}$, then there are MAGs $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ represented by $\mathcal{G}$ such that $A \to B$ is in $\mathcal{M}_1$, $A \leftarrow B$ is in $\mathcal{M}_2$, and $A \leftrightarrow B$ is in $\mathcal{M}_3$.

(ii) If $A \circ\!\!\to B$ is in $\mathcal{G}$, then there are two MAGs $\mathcal{M}_1$ and $\mathcal{M}_2$ represented by $\mathcal{G}$ such that $A \to B$ is in $\mathcal{M}_1$, and $A \leftrightarrow B$ is in $\mathcal{M}_2$.

Proof of Theorem C.7.12. Let $\mathcal{T}_0$ be a partially directed join tree of $\mathcal{G}$ (Definition C.7.10). There is at least one clique $\mathcal{C}_0$ such that $A, B \in \mathcal{C}_0$. Next, let $\mathcal{T}_1 = \texttt{transformTree}(\mathcal{T}_0, \mathcal{C}_0)$ (Algorithm 17), $\mathcal{T} = \texttt{orientTree}(\mathcal{T}_1, \mathcal{C}_0)$ (Algorithm 18), and $\pi_{\mathcal{T}}$ be a partial order compatible with $\mathcal{T}$.

By case (ii) of Lemma C.7.31, $\pi_{\mathcal{T}}$ induces edge orientations that are compatible with $\mathcal{G}$ through the process described in Definition C.7.11. Call the graph so obtained as $\mathcal{G}_\pi$. Then $\langle A, B\rangle$ is of the same form in $\mathcal{G}$ and $\mathcal{G}_\pi$ ((iii) of Lemma C.7.31).

Furthermore, by case (vii) of Lemma C.7.31, $\mathcal{G}_\pi$ is an ancestral partial mixed graph with no minimal collider paths and edge orientations completed under R2, and R8. Additionally, any ancestral directed mixed graph $\mathcal{M}$ that is represented by $\mathcal{G}_\pi$ will be a MAG represented by $\mathcal{G}$.

Observe that all edges in $\mathcal{G}_\pi$ that are between two cliques are invariant. All variant edges ($\circ\!\!-\!\!\bullet$) are only present inside cliques. Therefore, to ensure that a directed mixed graph $\mathcal{M}$ that is represented by $\mathcal{G}_\pi$ is ancestral, it is enough to ensure that no directed or almost directed cycle is created within its maximal cliques. Lemmas C.7.32 and C.7.33 give us two alternate procedures for orienting partially oriented cliques in $\mathcal{G}_\pi$ to obtain ancestral directed mixed graphs $\mathcal{M}_1$ and $\mathcal{M}_2$ with desired edge marks on $\langle A, B\rangle$. Although the procedure in Lemma C.7.33 is stronger than the one in Lemma C.7.32, we still state Lemma C.7.32 as it provides an alternate proof for Meek (1995)'s result. □

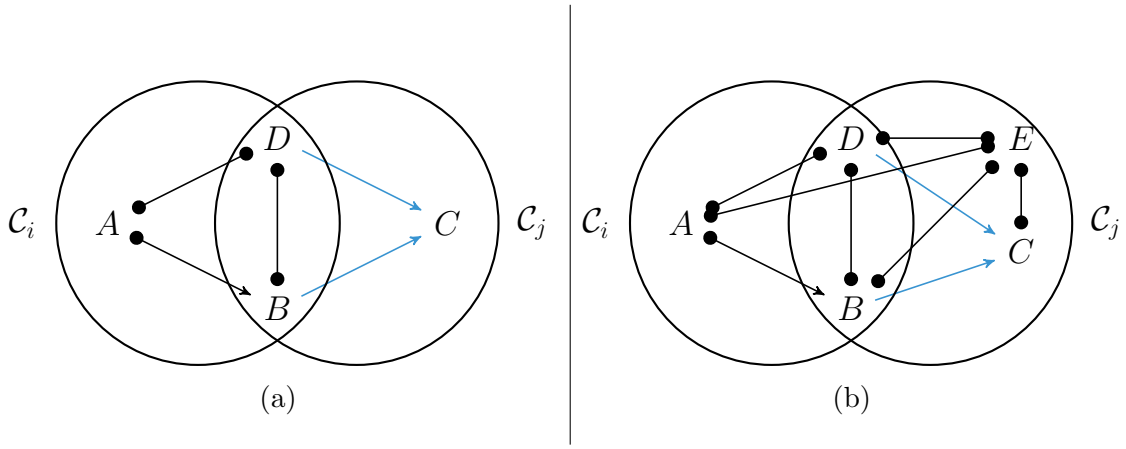


Figure C.5: Used in proof of Lemma C.7.13.

C.7.3 General Partially Directed Join Tree Properties

Lemma C.7.13. *Let $\mathcal{G}$ be an ancestral partial mixed graph with a chordal skeleton such that $\mathcal{G}$ has no minimal collider paths such that the orientations in $\mathcal{G}$ are closed under R1 and R11. Let $\mathcal{T}$ be a join tree for $\mathcal{G}$ and γ a relation as defined in Definition C.7.9. Let $\mathcal{C}_i$ and $\mathcal{C}_j$ be adjacent in $\mathcal{T}$, and suppose that there is an unshielded triple $\langle A, B, C\rangle$ such that $A \bullet\!\!\rightarrow B$ in $\mathcal{G}$, and $A, B \in \mathcal{C}_i$, $B, C \in \mathcal{C}_j$, $A \notin \mathcal{C}_j$, $C \notin \mathcal{C}_i$. Then $\gamma(\mathcal{C}_i, \mathcal{C}_j)$.*

Proof of Lemma C.7.13. Since $A \bullet\!\!\rightarrow B$ is in $\mathcal{G}$ and since $\Lambda_{ij} \neq \varnothing$, for $\gamma(\mathcal{C}_i, \mathcal{C}_j)$ it is enough to show that for all $D \in \Lambda_{ij}$, $E \in \mathcal{C}_j \setminus \mathcal{C}_i$, $D \rightarrow E$ is in $\mathcal{G}$. There are three cases:

(i) If $D \equiv B$, then for $E \equiv C$, or $E \notin \text{Adj}(A, \mathcal{G})$, $\langle A, D, E\rangle$, forms an unshielded triple in $\mathcal{G}$. Since $\mathcal{G}$ does not contain unshielded colliders, by R1, we conclude that $D \rightarrow E$ is in $\mathcal{G}$.

(ii) For $D \not\equiv B$ but $E \equiv C$, we have that $\langle B, D\rangle$ edge is in $\mathcal{G}$ since $B, D \in \Lambda_{ij}$. Since $\mathcal{G}$ does not contain unshielded colliders or longer minimal collider paths, we conclude by R11 (see Figure C.5a) that $D \rightarrow C$, that is $D \rightarrow E$ is in $\mathcal{G}$.

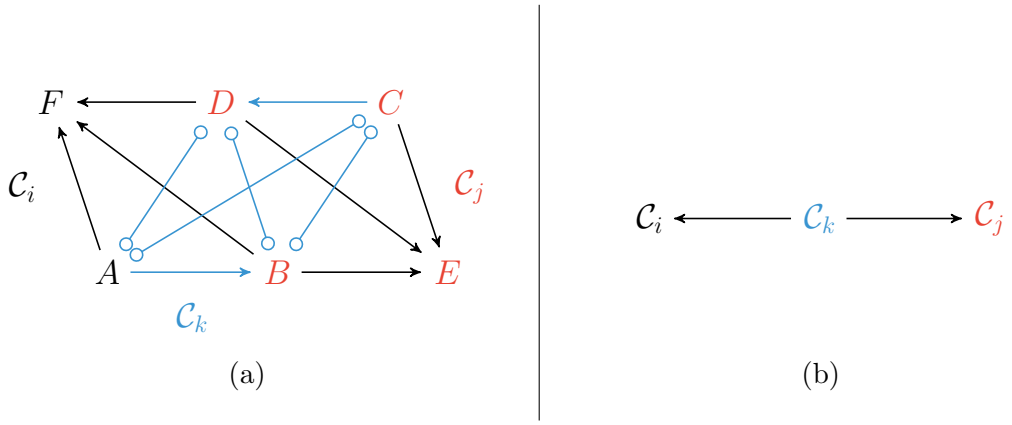


Figure C.6: C.6a partially mixed graph $\mathcal{G}$, C.6b partially directed join tree $\mathcal{T}$ of $\mathcal{G}$. These graphs are used in Example C.7.14.

(iii) For $D \not\equiv B$ and $E \not\equiv C$, we know that, $\langle B, D\rangle$ edge is in $\mathcal{G}$ since $B, D \in \Lambda_{ij}$ and also that $\langle D, C\rangle$, $\langle E, C\rangle$ are in $\mathcal{G}$, since $B, C, D, E \in \mathcal{C}_j$. If $A \notin \text{Adj}(E, \mathcal{G})$, then as in the cases above, by R1 $B \to E$ is in $\mathcal{G}$ and by R11, $D \to E$ is also in $\mathcal{G}$ and we are done.

Otherwise, $A \in \text{Adj}(E, \mathcal{G})$ as in Figure C.5b. However, this case is not possible. For sake of contradiction assume that this is possible. Note that A, B, D, E form a clique in $\mathcal{G}$, but since $E \notin \mathcal{C}_i$, there must be another maximal clique in $\mathcal{G}$, $\mathcal{C}_k$ that is a node in $\mathcal{T}$, such that $A, B, D, E \in \mathcal{C}_k$. Furthermore, $C \notin \mathcal{C}_k$, because $A \notin \text{Adj}(C, \mathcal{G})$.

There cannot be a path from $\mathcal{C}_k$ to $\mathcal{C}_j$ in $\mathcal{T}$ that contains $\mathcal{C}_i$ as that violates the running intersection property ($\mathcal{C}_k \cap \mathcal{C}_j \subseteq \{B, D, E\} \not\subseteq \mathcal{C}_i$ as $E \notin \mathcal{C}_i$).

Similarly, there is no path from $\mathcal{C}_i$ to $\mathcal{C}_k$ in $\mathcal{T}$ that contains $\mathcal{C}_j$ as that also violates the running intersection property ($\mathcal{C}_i \cap \mathcal{C}_k \subseteq \{A, B, D\} \not\subseteq \mathcal{C}_j$ since $A \notin \mathcal{C}_j$).

And since we assume that $\mathcal{C}_i$ and $\mathcal{C}_j$ are adjacent in $\mathcal{T}$ there cannot be a path from $\mathcal{C}_i$ to $\mathcal{C}_j$ that contains $\mathcal{C}_k$. Thus, we have a contradiction to $A \in \text{Adj}(E, \mathcal{G})$.

Therefore, we have shown that $\gamma(\mathcal{C}_i, \mathcal{C}_j)$. □

Example C.7.14. The condition of $\mathcal{C}_i$ and $\mathcal{C}_j$ being adjacent in the join tree is necessary for Lemma C.7.13 to hold. As an example illustrating this, consider the graphs in Figure C.6. A partially directed ancestral mixed graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ in Figure C.6a has orientations that are complete with respect to R1-R4, R8-R12 and a chordal skeleton. In fact, the essential graph of $\mathcal{G}$ is fully undirected.

Three maximal cliques make up $\mathbf{V}$. These are $\mathcal{C}_i = \{A, B, D, F\}, \mathcal{C}_k = \{A, B, C, D\}$, and $\mathcal{C}_j = \{B, C, D, E\}$. A partially directed join tree of $\mathcal{G}$, called $\mathcal{T}$ is given in Figure C.6b. In fact, $\mathcal{T}$ is the only valid join tree of $\mathcal{G}$, since $\mathcal{C}_k$ is a separator for $\mathcal{C}_i$ and $\mathcal{C}_j$.

Now, note that $\mathcal{C}_i$ and $\mathcal{C}_j$ are not adjacent in $\mathcal{T}$, but otherwise, satisfy conditions of Lemma C.7.13. Notably also, $\Lambda_{ij} = \{B, D\}, \mathcal{C}_i \setminus \Lambda_{ij} = \{A, F\}$, and $\mathcal{C}_j \setminus \Lambda_{ij} = \{C, E\}$. However, looking at $\mathcal{G}$, we can conclude that $\neg\gamma(\mathcal{C}_i, \mathcal{C}_j)$ because $D \leftarrow C$ is in $\mathcal{G}$, and also $\neg\gamma(\mathcal{C}_j, \mathcal{C}_i)$ because $A \rightarrow B$ is in $\mathcal{G}$. Hence, this adjacency condition is necessary for Lemma C.7.13 to hold. □

Based on the result of Lemma C.7.13, one may assume that paths $\mathcal{C}_1 \rightarrow \mathcal{C}_2 - \mathcal{C}_3$, or $\mathcal{C}_1 \rightarrow \mathcal{C}_2 \leftarrow \mathcal{C}_3$ cannot occur in some partially directed join tree $\mathcal{T}$. We consider this in Lemma C.7.15, and show that contrary to the above intuition, the general join tree properties do not preclude such paths from existing. Subsequently, in Examples C.7.16-C.7.18 and, later, in Example C.7.25, we showcase a few partially directed join trees where such paths do occur.

We follow up Examples C.7.16-C.7.18 with a result (Lemma C.7.19) that shows how to move within the partially directed join tree space to a different partially directed join tree of $\mathcal{G}$ where some of these paths do not occur. Algorithm 15 operationalizes this result, and we show in Lemma C.7.21 that the result of applying Algorithm 15 is a partially directed join tree with our desired properties. Moreover, case (iv) of Lemma C.7.21 shows that the partially directed join tree resulting from the application of Algorithm 15 does not contain colliders. We demonstrate the Algorithm 15 in Examples C.7.22-C.7.24.

Lemma C.7.15. *Let $\mathcal{G}$ be an ancestral partial mixed graph with a chordal skeleton such that $\mathcal{G}$ has no minimal collider paths such that the orientations in $\mathcal{G}$ are closed under R1 and*

R11. Let $\mathcal{T}$ be a partially directed join tree for $\mathcal{G}$ as defined in Definition C.7.10.

Consider any two nodes $\mathcal{C}_i$ and $\mathcal{C}_j$ adjacent in $\mathcal{T}$, such that $\gamma(\mathcal{C}_i, \mathcal{C}_j)$. If there is node $\mathcal{C}_k$ in $\mathcal{T}$ that is distinct from $\mathcal{C}_i$ and such that $\mathcal{C}_j$ and $\mathcal{C}_k$ are adjacent in $\mathcal{T}$, then one of the following holds:

(i) $\gamma(\mathcal{C}_j, \mathcal{C}_k)$, or

(ii) $\neg\gamma(\mathcal{C}_j, \mathcal{C}_k)$ and $\Lambda_{ik} = \Lambda_{jk} \subseteq \Lambda_{ij}$, or

(iii) $\neg\gamma(\mathcal{C}_j, \mathcal{C}_k)$ and $\Lambda_{ik} = \Lambda_{ij} \subset \Lambda_{jk}$. In this case, $\gamma(\mathcal{C}_i, \mathcal{C}_k)$ holds.

Proof of Lemma C.7.15. Let $\Lambda_{ij} = \mathcal{C}_i \cap \mathcal{C}_j$, and $\Lambda_{jk} = \mathcal{C}_j \cap \mathcal{C}_k$. By assumption, $\Lambda_{ij} \neq \varnothing \neq \Lambda_{jk}$. Furthermore, by definition of γ, for all $B \in \Lambda_{ij}$ and $C \in \mathcal{C}_j \setminus \Lambda_{ij}$, $B \to C$ and there is at least one $A \in \mathcal{C}_i \setminus \Lambda_{ij}$ and $B \in \Lambda_{ij}$, such that $A \bullet\!\!\to B$ is in $\mathcal{G}$. Note that $\mathcal{C}_i$ and $\mathcal{C}_k$ are not adjacent in $\mathcal{T}$, because $\mathcal{T}$ is a tree. We also know that $\mathcal{C}_i \cap \mathcal{C}_k = \Lambda_{ik} \subseteq \mathcal{C}_j$ by the running intersection property. Consider the following possibilities:

(1) $(\mathcal{C}_j \setminus \Lambda_{ij}) \cap \Lambda_{jk} = \varnothing$.

(2) $(\mathcal{C}_j \setminus \Lambda_{ij}) \cap \Lambda_{jk} \neq \varnothing$ and $(\mathcal{C}_j \setminus \Lambda_{jk}) \cap \Lambda_{ij} \neq \varnothing$.

(3) $(\mathcal{C}_j \setminus \Lambda_{ij}) \cap \Lambda_{jk} \neq \varnothing$ and $(\mathcal{C}_j \setminus \Lambda_{jk}) \cap \Lambda_{ij} = \varnothing$.

Cases (1)-(3) are mutually disjoint by construction and exhaust all possibilities for the relationship between $\mathcal{C}_i, \mathcal{C}_j$, and $\mathcal{C}_k$. We will show that they correspond to certain cases of Lemma C.7.15.

We will make use of the following three set identities:

$$\text{For any two sets } \mathcal{X}, \mathcal{Y} \text{ such that } \mathcal{Y} \subseteq \mathcal{X}, \text{ then } \mathcal{Y} = \mathcal{Y} \cap \mathcal{X}. \tag{C.1}$$

$$\text{For any three sets } \mathcal{X}, \mathcal{Y}, \mathcal{Z} \text{ such that } \mathcal{Y} \subset \mathcal{Z}, \text{ then } (\mathcal{X} \setminus \mathcal{Z}) \subset \mathcal{X} \setminus \mathcal{Y}. \tag{C.2}$$

For any three sets $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ such that $\mathcal{Y}, \mathcal{Z} \subseteq \mathcal{X}$, then $(\mathcal{X} \setminus \mathcal{Y}) \cap \mathcal{Z} = \varnothing \iff \mathcal{Z} \subseteq \mathcal{Y}$. (C.3)

(1) By Equation (C.3) on $(\mathcal{C}_j, \Lambda_{ij}, \Lambda_{jk})$, we have $\Lambda_{jk} \subseteq \Lambda_{ij}$. This, along with Equation (C.1), allows us to write

$$\Lambda_{jk} = (\Lambda_{jk} \cap \mathcal{C}_k) \subseteq (\Lambda_{ij} \cap \mathcal{C}_k) = \Lambda_{ik}.$$

Running intersection ($\Lambda_{ik} \subseteq \mathcal{C}_j$) tells us $\Lambda_{ik} = (\Lambda_{ik} \cap \mathcal{C}_k) \subseteq (\mathcal{C}_j \cap \mathcal{C}_k) = \Lambda_{jk}$.

Hence, we have that $\Lambda_{ik} = \Lambda_{jk} \subseteq \Lambda_{ij}$. Therefore, if $\gamma(\mathcal{C}_j, \mathcal{C}_k)$ we are in case (i) and otherwise, we are in case (ii).

(2) There is a node $A \in (\mathcal{C}_j \setminus \Lambda_{jk}) \cap \Lambda_{ij}$ and also a node $B \in (\mathcal{C}_j \setminus \Lambda_{ij}) \cap \Lambda_{jk}$ and for any such pair of nodes (A, B), $A \to B$ is in $\mathcal{G}$ by assumption that $\gamma(\mathcal{C}_i, \mathcal{C}_j)$ holds. Now, Lemma C.7.13 tells us that $\gamma(\mathcal{C}_j, \mathcal{C}_k)$. Hence, we are in case (i).

(3) By Equation (C.3) on $(\mathcal{C}_j, \Lambda_{jk}, \Lambda_{ij})$, we have $\Lambda_{ij} \subseteq \Lambda_{jk}$.

This, along with Equation (C.1), allows us to write

$$\Lambda_{ij} = (\Lambda_{ij} \cap \mathcal{C}_i) \subseteq (\Lambda_{jk} \cap \mathcal{C}_i) = (\mathcal{C}_j \cap \mathcal{C}_k \cap \mathcal{C}_i) = (\mathcal{C}_j \cap \Lambda_{ik}) = \Lambda_{ik},$$

where we used the running intersection property ($\Lambda_{ik} \subseteq \mathcal{C}_j$) and Equation (C.1) the last step. Running intersection also tells us $\Lambda_{ik} = (\Lambda_{ik} \cap \mathcal{C}_i) \subseteq \Lambda_{ij}$.

Hence, $\Lambda_{ij} = \Lambda_{ik} \subseteq \Lambda_{jk}$.

Additionally, by Equation (C.3) on $(\mathcal{C}_j, \Lambda_{ij}, \Lambda_{jk})$, we have $\Lambda_{jk} \not\subseteq \Lambda_{ij}$. Therefore, $\Lambda_{ij} = \Lambda_{ik} \subset \Lambda_{jk}$.

To show that we are now either in case (i) or (iii), we will prove that $\gamma(\mathcal{C}_i, \mathcal{C}_k)$ holds. Let A, B be nodes such that $A \in \mathcal{C}_i \setminus \mathcal{C}_j$, $B \in \Lambda_{ij}$, and $A \bullet\!\!\to B$ is in $\mathcal{G}$. Since $\Lambda_{ik} \subset \mathcal{C}_j$, Equation (C.2) says $\mathcal{C}_i \setminus \mathcal{C}_j \setminus \mathcal{C}_i \setminus \Lambda_{ik}$. Therefore, $A \in \mathcal{C}_i \setminus \mathcal{C}_k$. Further, since $\Lambda_{ij} = \Lambda_{ik}$, $B \in \Lambda_{ik}$.

Furthermore, note that for any $C \in \mathcal{C}_k \setminus \mathcal{C}_i$, $A \notin \mathrm{Adj}(C, \mathcal{G})$. For sake of contradiction,

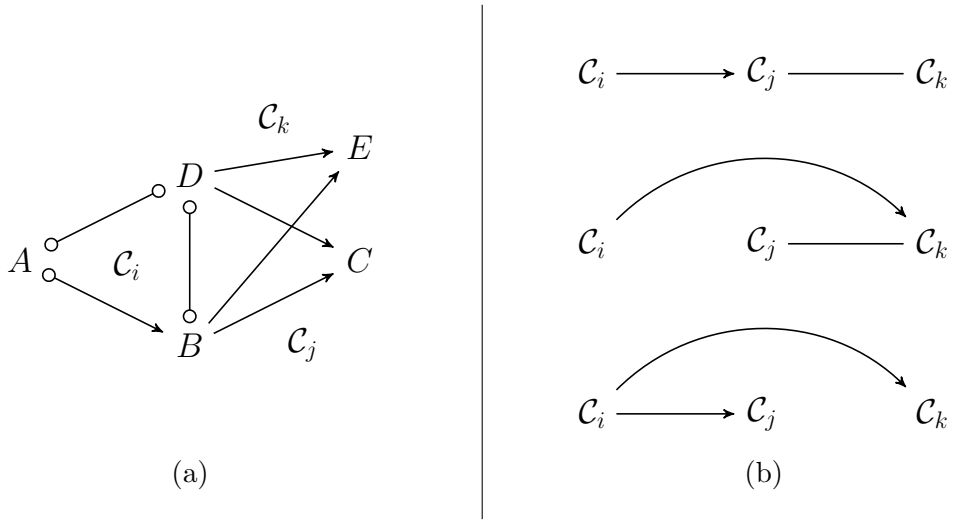


Figure C.7: C.7a Partially mixed graph $\mathcal{G}$, C.7b Three partially directed join trees for $\mathcal{G}$. These graphs are explored in Examples C.7.16 and C.7.22.

assume that there is some $C \in \mathcal{C}_k \setminus \mathcal{C}_i$, such that $A \in \text{Adj}(C, \mathcal{G})$, then there also must be a maximal clique $\mathcal{C}_r$ in $\mathcal{G}$, such that $A, B, C \in \mathcal{C}_r$. However, we know that $\langle \mathcal{C}_i, \mathcal{C}_j, \mathcal{C}_k \rangle$ is in $\mathcal{T}$ meaning that either (a) every path from $\mathcal{C}_r$ to $\mathcal{C}_i$ contains $\mathcal{C}_j$, or (b) every path from $\mathcal{C}_r$ to $\mathcal{C}_k$ contains $\langle \mathcal{C}_i, \mathcal{C}_j, \mathcal{C}_k \rangle$. Now the contradiction follows from the running intersection property since we have that $\mathcal{C}_r \cap \mathcal{C}_i \supseteq \{A\} \not\subseteq \mathcal{C}_j$ and $\mathcal{C}_r \cap \mathcal{C}_k \supseteq \{C\} \not\subseteq \mathcal{C}_i$.

Since every node in $C \in \mathcal{C}_k \setminus \mathcal{C}_i$ is not adjacent to A, $B \to C$ is in $\mathcal{G}$, and for every other node $D \in \Lambda_{ik}$, $\langle B, D \rangle$ is in $\mathcal{G}$ and $D \to C$ is in $\mathcal{G}$ using the fact that orientations in $\mathcal{G}$ are complete under R1 and R11. Therefore, $\gamma(\mathcal{C}_i, \mathcal{C}_k)$ holds. □

Example C.7.16. A chordal and ancestral partial mixed graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ in Figure C.7a has orientations that are complete with respect to R1-R4, R8-R12. In fact, the essential graph of $\mathcal{G}$ is fully undirected.

Three maximal cliques make up $\mathbf{V}$. These are $\mathcal{C}_i = \{A, B, D\}, \mathcal{C}_j = \{B, C, D\}$, and $\mathcal{C}_k = \{B, D, E\}$. Three different partially directed join trees for $\mathcal{G}$ are given in Figure C.7b. From top to bottom, these join trees are $\mathcal{T}_1$, $\mathcal{T}_2$ and $\mathcal{T}_3$. As can be seen from the figure,

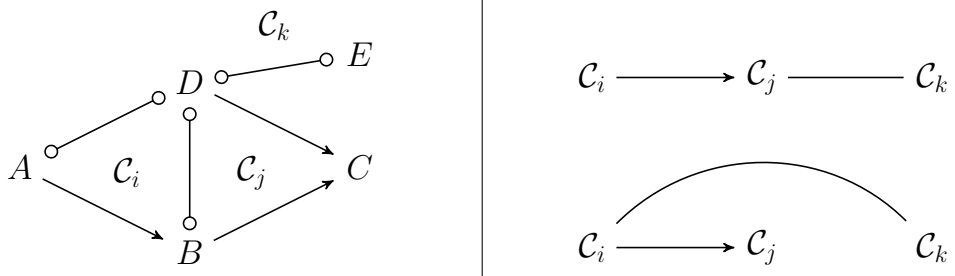


Figure C.8: C.8a Partially mixed graph $\mathcal{G}$, C.8b Two partially directed join trees for $\mathcal{G}$. These graphs are explored in Examples C.7.17 and C.7.23.

orientations in these join trees are not necessarily complete with respect to R1. Based on $\mathcal{G}$, we have that $\gamma(\mathcal{C}_i, \mathcal{C}_j)$ and $\gamma(\mathcal{C}_i, \mathcal{C}_k)$. However, neither $\gamma(\mathcal{C}_j, \mathcal{C}_k)$, nor $\gamma(\mathcal{C}_k, \mathcal{C}_j)$ hold.

□

Example C.7.17. A chordal and ancestral partial mixed graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ in Figure C.8a has orientations that are complete with respect to R1-R4, R8-R12. In fact, the essential graph of $\mathcal{G}$ is fully undirected.

Three maximal cliques make up $\mathbf{V}$. These are $\mathcal{C}_i = \{A, B, D\}, \mathcal{C}_j = \{B, C, D\}$, and $\mathcal{C}_k = \{D, E\}$. Two partially directed join trees for $\mathcal{G}$ are given in Figure C.8b. From top to bottom, these join trees are $\mathcal{T}_1$, $\mathcal{T}_2$. As can be seen from the figure, orientations in $\mathcal{T}_1$ are not complete with respect to R1. Based on $\mathcal{G}$, we have that $\gamma(\mathcal{C}_i, \mathcal{C}_j)$ but that is the only valid γ-relation on the maximal cliques of $\mathcal{G}$. □

Example C.7.18. A chordal and ancestral partial mixed graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ in Figure C.9a has orientations that are complete with respect to R1-R4, R8-R12. In fact, the essential graph of $\mathcal{G}$ is fully undirected.

Four maximal cliques make up $\mathbf{V}$. These are $\mathcal{C}_i = \{A, B, D\}, \mathcal{C}_j = \{B, C, D\}$, $\mathcal{C}_k = \{D, E\}$, and $\mathcal{C}_l = \{E, F\}$. Two partially directed join trees for $\mathcal{G}$ are given in Figure C.9b.

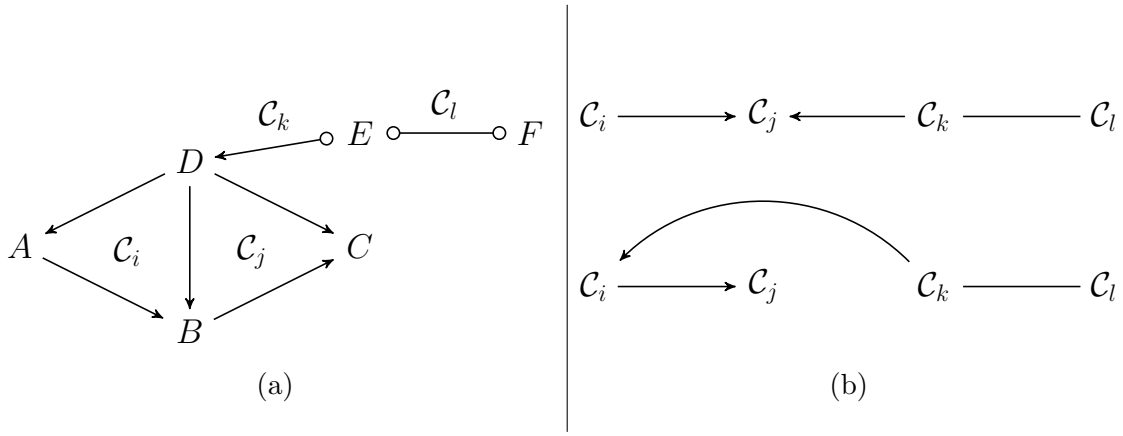


Figure C.9: C.9a Partially mixed graph $\mathcal{G}$, C.9b Two partially directed join trees for $\mathcal{G}$. These graphs are explored in Examples C.7.18 and C.7.24.

From top to bottom, these join trees are $\mathcal{T}_1$, $\mathcal{T}_2$. As can be seen from the figure, $\mathcal{T}_1$ contains an unshielded collider. Based on $\mathcal{G}$, the only valid γ-relations on the maximal cliques of $\mathcal{G}$ are $\gamma(\mathcal{C}_i, \mathcal{C}_j)$, $\gamma(\mathcal{C}_k, \mathcal{C}_i)$, and $\gamma(\mathcal{C}_k, \mathcal{C}_j)$.

□

C.7.4 Finding the Appropriate Partially Directed Join Tree

Lemma C.7.19. *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be an ancestral partial mixed graph with a chordal skeleton and such that $\mathcal{G}$ does not contain minimal collider paths. Let $\mathcal{T}_0 = (\mathbf{C}, \mathbf{E}_0)$ be a partially directed join tree for $\mathcal{G}$ (Definition C.7.10). Consider a triple $\langle \mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3 \rangle$ in $\mathcal{T}_0$ such that $\Lambda_{13} = \Lambda_{23} \subseteq \Lambda_{12}$. Suppose that $\gamma(\mathcal{C}_1, \mathcal{C}_2)$ holds, but not $\gamma(\mathcal{C}_2, \mathcal{C}_3)$. Then, the graph $\mathcal{T}$ obtained from $\mathcal{T}_0$ by removing edge $\langle \mathcal{C}_2, \mathcal{C}_3 \rangle$ and adding edge*

- $\mathcal{C}_1 \leftarrow \mathcal{C}_3$, *if* $\gamma(\mathcal{C}_3, \mathcal{C}_1)$, *or*
- $\mathcal{C}_1 \rightarrow \mathcal{C}_3$, *if* $\gamma(\mathcal{C}_1, \mathcal{C}_3)$, *or*
- $\mathcal{C}_1 - \mathcal{C}_3$, *if neither* $\gamma(\mathcal{C}_1, \mathcal{C}_3)$, *nor* $\gamma(\mathcal{C}_3, \mathcal{C}_1)$,

is still a partially directed join tree for $\mathcal{G}$.

Proof of Lemma C.7.19. It is easy to see that $\mathcal{T}$ is a tree: we replace edge $\langle \mathcal{C}_2, \mathcal{C}_3 \rangle$ with edge $\langle \mathcal{C}_1, \mathcal{C}_3 \rangle$ in $\mathcal{T}_0$, and since $\mathcal{T}_0$ is a tree, in doing so we do not create any cycles in the skeleton of $\mathcal{T}$.

The nodes of $\mathcal{T}$ are still maximal cliques of $\mathcal{G}$, and the orientations of edges in $\mathcal{T}$ still follow the γ relation by construction. So to show that $\mathcal{T}$ is a join tree for $\mathcal{G}$, we need to show that the running intersection property still holds.

Specifically, consider two maximal cliques $\mathcal{C}_i, \mathcal{C}_j$ in $\mathcal{G}$ such that $\Lambda_{ij} \neq \varnothing$. Suppose the unique path between $\mathcal{C}_i$ and $\mathcal{C}_j$ in $\mathcal{T}_0$ is p. If p does not contain edge $\langle \mathcal{C}_2, \mathcal{C}_3 \rangle$ then, p also exists in $\mathcal{T}$ and the running intersection holds for this path because $\mathcal{T}_0$ is a join tree.

Suppose that p contains the subpath $\langle \mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3 \rangle$ (with $\mathcal{C}_1$ or $\mathcal{C}_3$ possibly being the endpoints). Then, in $\mathcal{T}$, the unique path between $\mathcal{C}_i$ and $\mathcal{C}_j$ is $q = p(\mathcal{C}_i, \mathcal{C}_1) \oplus \langle \mathcal{C}_1, \mathcal{C}_3 \rangle \oplus p(\mathcal{C}_3, \mathcal{C}_j)$. Since $\Lambda_{ij} \subseteq \mathcal{C}_1$ and $\Lambda_{ij} \subseteq \mathcal{C}_3$ holds already in $\mathcal{T}_0$, the running intersection property is also satisfied in $\mathcal{T}$. A symmetric argument can be made when p contains the subpath $\langle \mathcal{C}_3, \mathcal{C}_2, \mathcal{C}_1 \rangle$.

Next, suppose that p contains the edge $\langle \mathcal{C}_2, \mathcal{C}_3 \rangle$ but does not contain node $\mathcal{C}_1$. Then, in $\mathcal{T}$, the unique path between $\mathcal{C}_i$ and $\mathcal{C}_j$ is $q = p(\mathcal{C}_i, \mathcal{C}_2) \oplus \langle \mathcal{C}_2, \mathcal{C}_1, \mathcal{C}_3 \rangle \oplus p(\mathcal{C}_3, \mathcal{C}_j)$. Then, in the new tree $\mathcal{T}$, the path must contain node $\mathcal{C}_1$. It is sufficient to show that $\Lambda_{ij} \subseteq \mathcal{C}_1$. Since $\Lambda_{ij} \subseteq \mathcal{C}_2$ and $\Lambda_{ij} \subseteq \mathcal{C}_3$, we have $\Lambda_{ij} \subseteq \Lambda_{23}$. This implies $\Lambda_{ij} \subseteq \Lambda_{12}$ by assumption. As $\Lambda_{12} \subseteq \mathcal{C}_1$, we have $\Lambda_{ij} \subseteq \mathcal{C}_1$. Therefore, the running intersection property still holds. A symmetric argument can be made when p contains the edge $\langle \mathcal{C}_3, \mathcal{C}_2 \rangle$ but not the node $\mathcal{C}_1$. □

Algorithm 15 presents a procedure leverages Lemma C.7.19 to remove triples $\langle \mathcal{C}_i, \mathcal{C}_j, \mathcal{C}_k \rangle$ such that $\Lambda_{ik} = \Lambda_{jk} \subseteq \Lambda_{ij}$ from the join tree $\mathcal{T}$. The key idea for this algorithm is that we make an exhaustive list of triples in the join tree, $\mathcal{Q}$ (line 2). Then, we go through every triple and check whether it meets the antecedent of Lemma C.7.19 (line 6). If it does, then we operate on the tree as Lemma C.7.19 suggests (line 12). This results in a new tree where the set of triples have changed. Therefore, we update the set of triples, $\mathcal{Q}$ (line 14). When we update $\mathcal{Q}$, we remove any triples present in the tree before the operation and add only the newly formed triples. This ensures that a triple present before the operation that we've

Algorithm 15 transformTreeHelper

Input: Partially directed join tree $\mathcal{T} = (\mathbf{C}, \mathbf{E})$ for an ancestral partial mixed graph $\mathcal{G}$ with a chordal skeleton, with edge orientations closed under R1-R4, R8-R12 and such that $\mathcal{G}$ is without minimal collider paths.
Output: Another join tree $\mathcal{T}' = (\mathbf{C}, \mathbf{E}')$ for $\mathcal{G}$.

1: $\mathcal{T}' \leftarrow \mathcal{T}$
2: $\mathcal{Q} \leftarrow \{\langle \mathcal{C}_i, \mathcal{C}_j, \mathcal{C}_k \rangle \mid \langle \mathcal{C}_i, \mathcal{C}_j \rangle, \langle \mathcal{C}_j, \mathcal{C}_k \rangle \in \mathbf{E}\}$ ▷ Set of triples yet to be verified
3: **while** $\mathcal{Q} \neq \varnothing$ **do**
4: $\quad \langle \mathcal{C}_i, \mathcal{C}_j, \mathcal{C}_k \rangle \leftarrow \mathcal{Q}_1$ ▷ Remove the first triple from $\mathcal{Q}$
5: $\quad \mathcal{Q} \leftarrow \mathcal{Q} \setminus \mathcal{Q}_1$
6: $\quad$ **if** $\gamma(\mathcal{C}_i, \mathcal{C}_j)$ and $\neg\gamma(\mathcal{C}_j, \mathcal{C}_k)$ **then**
7: $\quad\quad \Lambda_{ij} \leftarrow \mathcal{C}_i \cap \mathcal{C}_j$
8: $\quad\quad \Lambda_{jk} \leftarrow \mathcal{C}_j \cap \mathcal{C}_k$
9: $\quad\quad \Lambda_{ik} \leftarrow \mathcal{C}_i \cap \mathcal{C}_k$
10: $\quad\quad$ **if** $\Lambda_{ik} = \Lambda_{jk} \subseteq \Lambda_{ij}$ **then**
11: $\quad\quad\quad \mathcal{A} \leftarrow \{\langle \mathcal{C}_u, \mathcal{C}_v, \mathcal{C}_w \rangle \mid \langle \mathcal{C}_u, \mathcal{C}_v \rangle, \langle \mathcal{C}_v, \mathcal{C}_w \rangle \in \mathbf{E}'\}$
12: $\quad\quad\quad \mathbf{E}' \leftarrow (\mathbf{E}' \cup \langle \mathcal{C}_i, \mathcal{C}_k \rangle) \setminus \langle \mathcal{C}_j, \mathcal{C}_k \rangle$ ▷ Transform as in Lemma C.7.19
13: $\quad\quad\quad \mathcal{B} \leftarrow \{\langle \mathcal{C}_u, \mathcal{C}_v, \mathcal{C}_w \rangle \mid \langle \mathcal{C}_u, \mathcal{C}_v \rangle, \langle \mathcal{C}_v, \mathcal{C}_w \rangle \in \mathbf{E}'\}$
14: $\quad\quad\quad \mathcal{Q} \leftarrow (\mathcal{Q} \setminus (\mathcal{A} \setminus \mathcal{B})) \cup (\mathcal{B} \setminus \mathcal{A})$ ▷ Update $\mathcal{Q}$ with triples present *only* in $\mathcal{B}$
15: **return** $\mathcal{T}'$

already verified in line 6 does not get added back. We show that Algorithm 15 terminates in Lemma C.7.20 and prove some important properties of its output in Lemma C.7.21.

Lemma C.7.20. *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be an ancestral partial mixed graph with a chordal skeleton such that $\mathcal{G}$ has no minimal collider paths. Let $\mathcal{T}_0$ be any partially directed join tree for $\mathcal{G}$ (Definition C.7.10) and γ a relation as defined in Definition C.7.9. Then Algorithm 15 terminates with input $\mathcal{T}_0$.*

Proof of Lemma C.7.20. For sake of contradiction, suppose that Algorithm 15 does not terminate. Observe that there are only a finite number of possible triples in $\mathcal{T}$, $|\mathbf{C}| \times (|\mathbf{C}| - 1) \times (|\mathbf{C}| - 2)$. As Algorithm 15 does not terminate, it must be that Line 6 encounters some triple $\langle \mathcal{C}_i, \mathcal{C}_j, \mathcal{C}_k \rangle$ again after previously operating on it according to Lemma C.7.19.

The first time we encounter this triple, we operate as in Lemma C.7.19 to construct a new triple $\langle \mathcal{C}_k, \mathcal{C}_i, \mathcal{C}_j \rangle$. In order to have encountered the triple $\langle \mathcal{C}_i, \mathcal{C}_j, \mathcal{C}_k \rangle$ again, there must

be another triple $\langle \mathcal{C}_j, \mathcal{C}_{j_2}, \mathcal{C}_k \rangle$ (or $\langle \mathcal{C}_k, \mathcal{C}_{j_2}, \mathcal{C}_j \rangle$), in a tree $\mathcal{T}_1$, that gets operated on it as in Lemma C.7.19 to construct the triple $\langle \mathcal{C}_k, \mathcal{C}_j, \mathcal{C}_{j2} \rangle$ (or $\langle \mathcal{C}_j, \mathcal{C}_k, \mathcal{C}_{j2} \rangle$).

However, this must mean that there is an undirected cycle in the skeleton of $\mathcal{T}_1$ made up by $p = \langle \mathcal{C}_j, \mathcal{C}_{j2}, \mathcal{C}_k \rangle$ and $q = \langle \mathcal{C}_k, \ldots, \mathcal{C}_i, \mathcal{C}_j \rangle$. Here q must contain the edge $\langle \mathcal{C}_i, \mathcal{C}_j \rangle$ in $\mathcal{T}_1$. Further, $q(\mathcal{C}_k, \mathcal{C}_i)$ is either the edge $\langle \mathcal{C}_k, \mathcal{C}_i \rangle$ that was obtained from operating on $\langle \mathcal{C}_i, \mathcal{C}_j, \mathcal{C}_k \rangle$ the first time and is still present in $\mathcal{T}_1$, or $\langle \mathcal{C}_k, \mathcal{C}_i \rangle$ was removed by some prior application of Lemma C.7.19 in which case a longer path $q(\mathcal{C}_k, \mathcal{C}_i) = \langle \mathcal{C}_k, \ldots, \mathcal{C}_i \rangle$ is present in $\mathcal{T}_1$. Such a cycle with p and q, of course, is a contradiction with $\mathcal{T}_0$ being a tree, or the result of Lemma C.7.19. □

Lemma C.7.21. *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be an ancestral partial mixed graph with a chordal skeleton such that $\mathcal{G}$ has no minimal collider paths and such that orientations in $\mathcal{G}$ are closed under R1 and R11. Let $\mathcal{T}_0 = (\mathbf{C}, \mathbf{E_0})$ be any partially directed join tree for $\mathcal{G}$ (Definition C.7.10) and γ a relation as defined in Definition C.7.9. Let $\mathcal{T} = (\mathbf{C}, \mathbf{E})$ be the output of Algorithm 15 i.e., $\mathcal{T} =$ `transformTreeHelper`$(\mathcal{T}_0)$. Then*

(i) $\mathcal{T}$ is also a join tree for $\mathcal{G}$, and

(ii) for any pair of cliques, if $\gamma(\mathcal{C}_i, \mathcal{C}_j)$ in $\mathcal{T}_0$, then $\gamma(\mathcal{C}_i, \mathcal{C}_j)$ in $\mathcal{T}$ as well, and

(iii) for any path $\langle \mathcal{C}_i, \mathcal{C}_j, \mathcal{C}_k \rangle$ in $\mathcal{T}$ such that $\gamma(\mathcal{C}_i, \mathcal{C}_j)$ but not $\gamma(\mathcal{C}_j, \mathcal{C}_k)$, then $\Lambda_{ik} = \Lambda_{ij} \subset \Lambda_{jk}$ and $\gamma(\mathcal{C}_i, \mathcal{C}_k)$ holds.

(iv) $\mathcal{T}$ does not contain any path of the form $\mathcal{C}_i \to \mathcal{C}_j \leftarrow \mathcal{C}_k$ for any $\mathcal{C}_i, \mathcal{C}_j, \mathcal{C}_k \in \mathbf{C}$.

Proof of Lemma C.7.21. Algorithm 15 terminates by Lemma C.7.20. This allows us to talk about the properties of its output, $\mathcal{T}$.

(i) $\mathcal{T}$ is a join tree by Lemma C.7.19.

(ii) Since we do not change any orientations of edges in $\mathcal{G}$ during the course of Algorithm 15, γ ordering is preserved.

(iii) By Lemma C.7.15, if $\langle \mathcal{C}_i, \mathcal{C}_j, \mathcal{C}_k \rangle$ in $\mathcal{T}$ such that $\gamma(\mathcal{C}_i, \mathcal{C}_j)$ but not $\gamma(\mathcal{C}_j, \mathcal{C}_k)$, then either $\Lambda_{ik} = \Lambda_{jk} \subseteq \Lambda_{ij}$ or $\Lambda_{ik} = \Lambda_{ij} \subset \Lambda_{jk}$. However, it is not the case that $\Lambda_{ik} = \Lambda_{jk} \subseteq \Lambda_{ij}$ (otherwise $\mathcal{Q} \neq \varnothing$ in Algorithm 15). Therefore, $\Lambda_{ik} = \Lambda_{ij} \subset \Lambda_{jk}$ and $\gamma(\mathcal{C}_i, \mathcal{C}_k)$ holds.

(iv) Suppose for a contradiction that $\mathcal{T}$ does contain a path of the form $\mathcal{C}_i \to \mathcal{C}_j \leftarrow \mathcal{C}_k$. By case (iii) above , we then have that $\Lambda_{ik} = \Lambda_{ij} \subset \Lambda_{jk}$, and also that $\gamma(\mathcal{C}_i, \mathcal{C}_k)$ holds. But, also, since $\mathcal{C}_k \to \mathcal{C}_j \leftarrow \mathcal{C}_i$, case (iii) above leads us to conclude that $\Lambda_{ik} = \Lambda_{jk} \subset \Lambda_{ij}$, and $\gamma(\mathcal{C}_k, \mathcal{C}_i)$ hold in $\mathcal{G}$, which a contradiction.

□

Example C.7.22. Consider again graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ in Figure C.7a used in Example C.7.18 above. As discussed in Example C.7.16, Figure C.7b contains three partially directed join trees for $\mathcal{G}$. From top to bottom, these join trees are $\mathcal{T}_1$, $\mathcal{T}_2$ and $\mathcal{T}_3$.

Applying Algorithm 15 to $\mathcal{T}_1$ or to $\mathcal{T}_2$ leads to $\mathcal{T}_3$ as output. Note that $\Lambda_{ij} = \{B, D\} = \Lambda_{jk} = \Lambda_{ik}$ and that therefore in $\mathcal{T}_1$, $\Lambda_{ik} = \Lambda_{jk} \subseteq \Lambda_{ij}$, and in $\mathcal{T}_2$, $\Lambda_{ij} = \Lambda_{jk} \subseteq \Lambda_{ik}$. So both $\mathcal{T}_1$ and $\mathcal{T}_2$ satisfy conditions of Lemma C.7.19.

For $\mathcal{T}_1$, this is since line 10 calls for Lemma C.7.19 to be applied applied to triple $\mathcal{C}_i \to \mathcal{C}_j - \mathcal{C}_k$. That is edge $\mathcal{C}_j - \mathcal{C}_k$ is removed from $\mathcal{T}_1$ and edge $\mathcal{C}_i \to \mathcal{C}_k$ is added to create $\mathcal{T}_3$.

For $\mathcal{T}_2$, Lemma C.7.19 is applied to $\mathcal{C}_i \to \mathcal{C}_k - \mathcal{C}_i$. It removes $\mathcal{C}_k - \mathcal{C}_i$ from $\mathcal{T}_2$ and adds $\mathcal{C}_i \to \mathcal{C}_k$ to create $\mathcal{T}_3$. □

Example C.7.23. Consider again graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ in Figure C.8a used in Example C.7.18 above. As discussed in Example C.7.18, Figure C.8b contains two partially directed join trees for $\mathcal{G}$. From top to bottom, these join trees are $\mathcal{T}_1$, $\mathcal{T}_2$.

Applying Algorithm 15 to $\mathcal{T}_1$ leads to $\mathcal{T}_2$ as output. This is because line 10 calls for Lemma C.7.19 to be applied to triple $\mathcal{C}_i \to \mathcal{C}_j - \mathcal{C}_k$. Note that $\Lambda_{ij} = \{B, D\}$, and $\Lambda_{jk} = \{D\} = \Lambda_{ik}$. Therefore in $\mathcal{T}_1$, $\Lambda_{ik} = \Lambda_{jk} \subset \Lambda_{ij}$, so $\mathcal{T}_1$ satisfies conditions of Lemma C.7.19. That is edge $\mathcal{C}_j - \mathcal{C}_k$ is removed from $\mathcal{T}_1$ and edge $\mathcal{C}_i - \mathcal{C}_k$ is added to create $\mathcal{T}_2$. □

Example C.7.24. Consider again graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ in Figure C.9a used in Example C.7.18 above. As discussed in Example C.7.18, Figure C.9b contains two partially directed join trees for $\mathcal{G}$. From top to bottom, these join trees are $\mathcal{T}_1$, $\mathcal{T}_2$.

Applying Algorithm 15 to $\mathcal{T}_1$ leads to $\mathcal{T}_2$ as output. This is because line 10 calls for Lemma C.7.19 to be applied to triple $\mathcal{C}_i \to \mathcal{C}_j \leftarrow \mathcal{C}_k$. Note that $\Lambda_{ij} = \{B, D\}$, $\Lambda_{jk} = \{D\} = \Lambda_{ik}$ and $\Lambda_{kl} = \{E\}$. Therefore in $\mathcal{T}_1$, $\Lambda_{ik} = \Lambda_{jk} \subset \Lambda_{ij}$, so $\mathcal{T}_1$ satisfies conditions of Lemma C.7.19. That is edge $\mathcal{C}_j \leftarrow \mathcal{C}_k$ is removed from $\mathcal{T}_1$ and edge $\mathcal{C}_i \leftarrow \mathcal{C}_k$ is added to create $\mathcal{T}_2$. □

In all the examples above there always exists a partially directed join tree $\mathcal{T}$ for a graph $\mathcal{G}$ such that paths $\mathcal{C}_i \to \mathcal{C}_j - \mathcal{C}_k$ and $\mathcal{C}_i \to \mathcal{C}_j \leftarrow \mathcal{C}_k$ do not occur in $\mathcal{T}$. While by case (iv) of Lemma C.7.21 it is true that a partially directed join tree without colliders will always exist for an ancestral and chordal partially directed mixed graph $\mathcal{G}$ with no minimal collider paths, the same is not true for paths of the form $\mathcal{C}_i \to \mathcal{C}_j - \mathcal{C}_k$. Example C.7.25 presents one case where all partially directed join trees for $\mathcal{G}$ contain such paths.

Lemma C.7.26 discusses how such paths can be transformed in $\mathcal{T}$, but they will not necessarily disappear entirely from the transformed join tree. Instead, we devise Algorithm 17 that in addition to colliders, removes all paths of the form $\mathcal{C}_{i_1} \to \mathcal{C}_{i_2} - \cdots - \mathcal{C}_{i_k} \leftarrow \mathcal{C}_{i_{k+1}}$ from a partially directed join tree for an ancestral and chordal partially directed mixed graph $\mathcal{G}$ with no minimal collider paths. Additionally, Algorithm 17 ensures that for a specified maximal clique $\mathcal{C}_0$, no path of the form $\mathcal{C}_{i_1} \to \mathcal{C}_{i_2} - \cdots - \mathcal{C}_{i_k} \to \cdots \to \mathcal{C}_r$ with $\mathcal{C}_r \equiv \mathcal{C}_0$ occurs in the resulting partially directed join tree. We prove these properties in Corollary C.7.29 and demonstrate Algorithm 17 in Example C.7.34.

Example C.7.25. A chordal and ancestral partial mixed graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ in Figure C.10a has orientations that are complete with respect to R1-R4, R8-R12. In fact, the essential graph of $\mathcal{G}$ as in all previous examples in this section is fully undirected.

Four maximal cliques make up $\mathbf{V}$. These are $\mathcal{C}_i = \{E, F\}, \mathcal{C}_j = \{C, D, F\}$, $\mathcal{C}_k = \{B, C, F\}$, and $\mathcal{C}_l = \{A, B, F\}$. Three partially directed join trees for $\mathcal{G}$ are given in Figure C.10b. From top to bottom, these join trees are $\mathcal{T}_1$, $\mathcal{T}_2$, and $\mathcal{T}_3$. As can be seen from

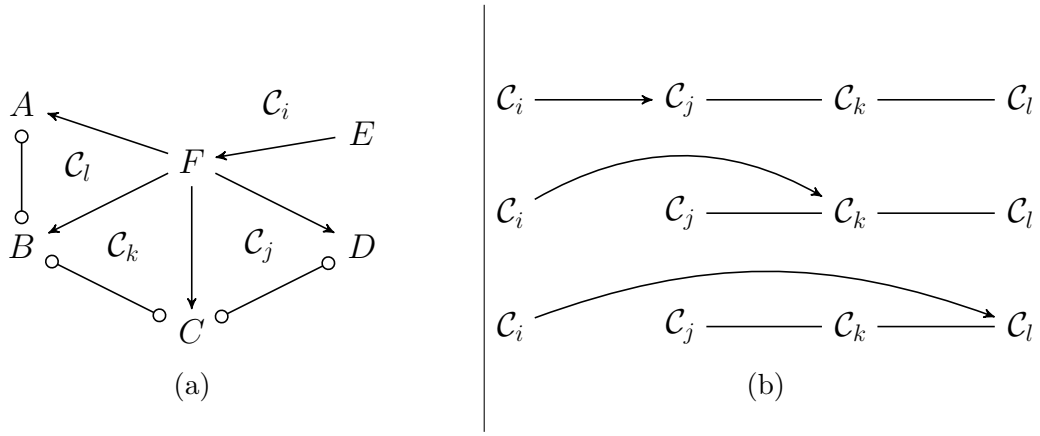


Figure C.10: C.10a Partially mixed graph $\mathcal{G}$, C.10b Three partially directed join trees for $\mathcal{G}$. These graphs are explored in Example C.7.25.

the figure, none of these partially directed join trees have orientations complete under R1. Based on $\mathcal{G}$, the only valid γ-relations on the maximal cliques of $\mathcal{G}$ are $\gamma(\mathcal{C}_i, \mathcal{C}_j)$, $\gamma(\mathcal{C}_i, \mathcal{C}_k)$, and $\gamma(\mathcal{C}_i, \mathcal{C}_l)$.

Note that $\Lambda_{ij} = \Lambda_{ik} = \Lambda_{il} = \{F\}$, $\Lambda_{jk} = \{C, F\}$, and $\Lambda_{kl} = \{B, F\}$. Therefore in $\mathcal{T}_1$, $\Lambda_{ik} = \Lambda_{ij} \subset \Lambda_{jk}$, so applying Algorithm 15 to $\mathcal{T}_1$ results in $\mathcal{T}_1$ as output. Similarly $\mathcal{T}_2 = \texttt{transformTree}(\mathcal{T}_2)$, and $\mathcal{T}_3 = \texttt{transformTree}(\mathcal{T}_3)$.

Note also that since $\Lambda_{jk} \not\subseteq \mathcal{C}_i$, $\mathcal{C}_j \leftarrow \mathcal{C}_i \rightarrow \mathcal{C}_k$ cannot be a path in a valid join tree for $\mathcal{G}$. A similar issue arises with path $\mathcal{C}_k \leftarrow \mathcal{C}_i \rightarrow \mathcal{C}_l$. Hence, the list of join trees in Figure C.10b is exhaustive for $\mathcal{G}$. This example demonstrates that while Algorithm 15 deals with some properties of a general partially directed join tree, it is not enough to ensure that the resulting partially directed join tree for $\mathcal{G}$ has a single root node. □

Lemma C.7.26. *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be an ancestral partial mixed graph with a chordal skeleton and such that $\mathcal{G}$ does not contain minimal collider paths and such that orientations in $\mathcal{G}$ are closed under R1 and R11. Let $\mathcal{T}_0 = (\mathbf{C}, \mathbf{E}_0)$ be a partially directed join tree for $\mathcal{G}$ (Definition C.7.10). Furthermore, suppose that applying Algorithm 15 to $\mathcal{T}_0$ results in the same tree, that is $\mathcal{T}_0 =$ `transformTreeHelper`$(\mathcal{T}_0)$. Consider a triple $\langle \mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3 \rangle$ in $\mathcal{T}_0$ that is of the*

form $\mathcal{C}_1 \to \mathcal{C}_2 - \mathcal{C}_3$. *Then the graph* $\mathcal{T}$ *obtained from* $\mathcal{T}_0$ *by removing edge* $\langle \mathcal{C}_1, \mathcal{C}_2 \rangle$ *and adding edge,* $\mathcal{C}_1 \to \mathcal{C}_3$ *is still a partially directed join tree for* $\mathcal{G}$.

Proof of Lemma C.7.26. It is easy to see that $\mathcal{T}$ is a tree: we replace the edge $\langle \mathcal{C}_1, \mathcal{C}_2 \rangle$ with $\langle \mathcal{C}_1, \mathcal{C}_3 \rangle$, which will not create any undirected cycles in the graph skeleton since the original graph $\mathcal{T}_0$ did not have any undirected cycles in the graph skeleton.

The nodes of $\mathcal{T}$ are still maximal cliques of $\mathcal{G}$, and by Lemma C.7.21, the γ property is maintained in $\mathcal{T}$. Hence, to show that $\mathcal{T}$ is a partially directed join tree for $\mathcal{G}$, we need to show that the running intersection property still holds. Specifically, consider two maximal cliques $\mathcal{C}_i, \mathcal{C}_j$ in $\mathcal{G}$ such that $\Lambda_{ij} \neq \varnothing$ and suppose the unique path between $\mathcal{C}_i$ and $\mathcal{C}_j$ in $\mathcal{T}_0$ is p. If p does not contain edge $\langle \mathcal{C}_1, \mathcal{C}_2 \rangle$ then, p also exists in $\mathcal{T}$ and the running intersection holds for this path because $\mathcal{T}_0$ is a join tree.

Suppose that p contains the subpath $\langle \mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3 \rangle$ (with $\mathcal{C}_1$ or $\mathcal{C}_3$ possibly being the endpoints). Then the unique path between $\mathcal{C}_i$ and $\mathcal{C}_j$ in $\mathcal{T}$ is $q = p(\mathcal{C}_i, \mathcal{C}_1) \oplus \langle \mathcal{C}_1, \mathcal{C}_3 \rangle \oplus p(\mathcal{C}_3, \mathcal{C}_j)$. Since $\Lambda_{ij} \subseteq \mathcal{C}_1$ and $\Lambda_{ij} \subseteq \mathcal{C}_3$ already holds in $\mathcal{T}_0$, q still satisfies the running intersection property in $\mathcal{T}$. A symmetric argument holds if p contains the subpath $\langle \mathcal{C}_3, \mathcal{C}_2, \mathcal{C}_1 \rangle$.

Next, suppose that p contains the edge $\langle \mathcal{C}_1, \mathcal{C}_2 \rangle$ but does not contain node $\mathcal{C}_3$. Then the unique path between $\mathcal{C}_i$ and $\mathcal{C}_j$ in $\mathcal{T}$ is $q = p(\mathcal{C}_i, \mathcal{C}_1) \oplus \langle \mathcal{C}_1, \mathcal{C}_3, \mathcal{C}_2 \rangle \oplus p(\mathcal{C}_2, \mathcal{C}_j)$. That is, the path must contain node $\mathcal{C}_3$. It is sufficient to show that $\Lambda_{ij} \subseteq \mathcal{C}_3$. Since $\Lambda_{ij} \subseteq \mathcal{C}_1$ and $\Lambda_{ij} \subseteq \mathcal{C}_2$, we have $\Lambda_{ij} \subseteq \Lambda_{12}$. This implies $\Lambda_{ij} \subseteq \Lambda_{23}$ by assumption, and $\Lambda_{23} \subseteq \mathcal{C}_3$. Therefore, $\Lambda_{ij} \subseteq \mathcal{C}_3$, and the running intersection property still holds. A symmetric argument holds if p contains the subpath $\langle \mathcal{C}_2, \mathcal{C}_1 \rangle$ but not the node $\mathcal{C}_3$. □

As we already discussed, the goal of Algorithm 17 is to remove all paths of the form $\mathcal{C}_{i_1} \to \mathcal{C}_{i_2} - \cdots - \mathcal{C}_{i_k} \leftarrow \mathcal{C}_{i_{k+1}}$ and $\mathcal{C}_{i_1} \to \mathcal{C}_{i_2} - \cdots - \mathcal{C}_{i_k} \to \cdots \to \mathcal{C}_r$ with $\mathcal{C}_r \equiv \mathcal{C}_0$, for a specified maximal clique $\mathcal{C}_0$ in the join tree. The intuition behind this algorithm is repeated application of the operation described in Lemma C.7.26. Specifically, we need to be careful about the order in which we apply this operation. Otherwise, we open ourselves to an infinite loop – for instance, in Example C.7.25, by applying this operation on randomly chosen triples

Algorithm 16 relevantPaths

Input: Partially directed join tree $\mathcal{T} = (\mathbf{C}, \mathbf{E})$ and node $\mathcal{C}_0 \in \mathbf{C}$
Output: List of paths $\mathbf{P}$ relevant to Corollary C.7.29

$\mathbf{A} \leftarrow \{\mathcal{C}_1 \to \mathcal{C}_2 - \cdots - \mathcal{C}_k \to \cdots \to \mathcal{C}_r \mid \langle \mathcal{C}_i, \mathcal{C}_{i+1} \rangle \in \mathbf{C}, r \geq k > 2, \mathcal{C}_r \equiv \mathcal{C}_0\}$
$\mathbf{B} \leftarrow \{\mathcal{C}_1 \to \mathcal{C}_2 - \cdots - \mathcal{C}_{k-1} \leftarrow \mathcal{C}_k \mid \langle \mathcal{C}_i, \mathcal{C}_{i+1} \rangle \in \mathbf{C}, k > 3\}$
$\mathbf{P} \leftarrow \mathbf{A} \cup \mathbf{B}$
return $\mathbf{P}$

Algorithm 17 transformTree

Input: Partially directed join tree $\mathcal{T} = (\mathbf{C}, \mathbf{E})$, and node $\mathcal{C}_0 \in \mathbf{C}$ for an ancestral partial mixed graph $\mathcal{G}$ with a chordal skeleton, with edge orientations closed under R1-R4, R8-R12 and such that $\mathcal{G}$ is without minimal collider paths.
Output: Another join tree $\mathcal{T}' = (\mathbf{C}, \mathbf{E}')$ for $\mathcal{G}$.

$\mathcal{T}' \leftarrow \texttt{transformTreeHelper}(\mathcal{T})$
$\mathbf{P} \leftarrow \texttt{relevantPaths}(\mathcal{T}, \mathcal{C}_0)$ ▷ Algorithm 16
while $\mathbf{P} \neq \varnothing$ **do**
    $p = \langle \mathcal{C}_1, \ldots, \mathcal{C}_k \rangle \in \mathbf{P}$ such that $p = \operatorname{argmax}_{p' \in \mathbf{P}} d(\mathcal{C}_1, \mathcal{C}_0)$ ▷ Definition C.7.27
    $\mathbf{E}' \leftarrow (\mathbf{E}' \cup (\mathcal{C}_1 \to \mathcal{C}_3)) \setminus (\mathcal{C}_1 \to \mathcal{C}_2)$ ▷ Transform as in Lemma C.7.26
    $\mathcal{T}' \leftarrow \texttt{transformTreeHelper}(\mathcal{T}')$
    $\mathbf{P} \leftarrow \texttt{relevantPaths}(\mathcal{T}', \mathcal{C}_0)$ ▷ Update paths in $\mathcal{T}'$
return $\mathcal{T}'$

we will traverse the space of the three join trees infinitely. To prevent such infinite loops, we will anchor the two kinds of paths we wish to remove to some node in the tree. When applying the operation in Lemma C.7.26, we will always prioritize a path that is *farthest* from this anchor. We will use Definition C.7.27 to characterize how far the endpoints from the paths are. For convenience, we will choose $\mathcal{C}_0$ as the anchor (any node in the tree will serve as a valid anchor as long as the tree is connected). We describe the technical details in Lemma C.7.28 and Algorithm 17.

Definition C.7.27 (Distance between nodes, d)**.** *For any two nodes, X, Y in a graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, the distance between them along a path $p = \langle X, \ldots, Y \rangle$ is the number of edges on p. We denote this by $d(X, Y; p)$. We say $d(X, X) = 0$ and if there is no path from X to Y, then $d(X, Y) = \infty$.*

Remark. Observe that in a tree graph, $\mathcal{T} = (\mathbf{V}, \mathbf{E})$, there is only one path between X and Y. Therefore, the distance between X and Y is unique and we will refer to this as $d(X, Y) := d(X, Y; p)$.

Lemma C.7.28. *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be an ancestral partial mixed graph with a chordal skeleton such that $\mathcal{G}$ has no minimal collider paths and such that orientations in $\mathcal{G}$ are closed under R1 and R11. Let $\mathcal{T}_0$ be any join tree for $\mathcal{G}$, $\mathcal{C}_0$ a node in $\mathcal{T}_0$, and γ a relation as defined in Definition C.7.9. Then Algorithm 17 terminates on input $(\mathcal{T}_0, \mathcal{C}_0)$.*

Proof of Lemma C.7.28. By Lemma C.7.20, we know that Algorithm 15 terminates. Furthermore, Algorithm 16 also terminates since we only consider graphs defined on a finite number of nodes in this manuscript. Therefore, to show the termination of Algorithm 17, we only need to show that the set $\mathbf{P}$ will be empty at some point. Note that every path in $\mathbf{P}$ starts with a triple of the form $\mathcal{C}_1 \to \mathcal{C}_2 - \mathcal{C}_3$. Hence, for the set $\mathbf{P}$ to become empty it is enough to show that once a path starting with a triple $\langle \mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3 \rangle$ is removed from $\mathbf{P}$ by applications of Lines 5-7, it will not be added again in a subsequent pass through the while loop.

For sake of contradiction, assume that Line 5 sees a triple $\langle \mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3 \rangle$ that was processed in a previous while loop iteration. During the previous encounter of this triple in the while loop, $\mathcal{C}_1 \to \mathcal{C}_2 - \mathcal{C}_3$ must have been transformed into $\mathcal{C}_1 \to \mathcal{C}_3 - \mathcal{C}_2$ by Line 5. Observe that since $\Lambda_{13} = \Lambda_{12} \subset \Lambda_{23}$, Algorithm 15 will not operate on this triple. Therefore, in order to re-encounter the triple $\mathcal{C}_1 \to \mathcal{C}_2 - \mathcal{C}_3$, one of the following must be true:

(i) there must have been some triple $\mathcal{C}_1 \to \mathcal{C}_\ell - \mathcal{C}_2$, $\ell \neq 3$, that got operated on by either Algorithm 15 or by Line 5 to create the edge $\langle \mathcal{C}_1, \mathcal{C}_2 \rangle$.

(ii) there must have been some path in $\mathbf{P}$ that started with the triple $\langle \mathcal{C}_1, \mathcal{C}_3, \mathcal{C}_2 \rangle$ and therefore got operated on as per Lemma C.7.26.

Case (i) indicates the presence of an undirected cycle in the skeleton of the tree, which leads to a contradiction. Therefore, in the rest of the proof we suppose case (ii) is true.

The fact that we encountered the triple $\langle \mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3 \rangle$ the first time, in some tree $\mathcal{T}_1$, indicates the presence of one of these two paths:

(A1) $\mathcal{C}_1 \to \mathcal{C}_2 - \mathcal{C}_3 - \cdots - \mathcal{C}_k \to \cdots \to \mathcal{C}_0$ ($\mathcal{C}_3 \equiv \mathcal{C}_0$ or $\mathcal{C}_k \equiv \mathcal{C}_0$ possibly), or

(A2) $\mathcal{C}_1 \to \mathcal{C}_2 - \mathcal{C}_3 - \cdots - \mathcal{C}_{k-1} \leftarrow \mathcal{C}_k$.

Now, when we encounter the triple $\langle \mathcal{C}_1, \mathcal{C}_3, \mathcal{C}_2 \rangle$, later on in some other tree $\mathcal{T}_2$, in Line 5, this indicates the presence of one of these two paths in $\mathcal{T}_2$:

(B1) $\mathcal{C}_1 \to \mathcal{C}_3 - \mathcal{C}_2 - \cdots - \mathcal{C}_{k'} \to \cdots \to \mathcal{C}_0$ ($\mathcal{C}_2 \equiv \mathcal{C}_0$ or $\mathcal{C}_{k'} \equiv \mathcal{C}_0$ possibly), or

(B2) $\mathcal{C}_1 \to \mathcal{C}_3 - \mathcal{C}_2 - \mathcal{C}_{i'} \cdots - \mathcal{C}_{k'-1} \leftarrow \mathcal{C}_{k'}$.

Clearly if (A1) was true, then (B1) cannot be true as this indicates the presence of a path from $\mathcal{C}_1$ to $\mathcal{C}_0$ that passes through $\langle \mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3 \rangle$ in $\mathcal{T}_1$ and another that passes through $\langle \mathcal{C}_1, \mathcal{C}_3, \mathcal{C}_2 \rangle$ in $\mathcal{T}_2$. Observe that after applying Lemma C.7.26 on path (A1), the path from $\mathcal{C}_1$ to $\mathcal{C}_0$ does not pass through $\mathcal{C}_2$. For (B1) to be present, there must have already been another path from $\mathcal{C}_2$ to $\mathcal{C}_0$ that does not pass through $\mathcal{C}_3$. This indicates the presence of cycles which contradicts that $\mathcal{T}_1$ is a tree.

Further, possibilities {(A1), (B2)} and {(A2), (B1)} are symmetric. So, without loss of generality, we only consider two cases below – {(A1), (B2)} and {(A2), (B2)}. In these cases, we rely on the fact that we have a fixed anchor ($\mathcal{C}_0$, here) and that we always choose a path from $\mathbf{P}$ that starts from a node that is farthest from the anchor (see Definition C.7.27 for definition of distance between nodes).

(A1) and (B2). Suppose that (A1) was present in $\mathcal{T}_1$ and (B2) is present in $\mathcal{T}_2$. Then, in $\mathcal{T}_2$, the path from $\mathcal{C}_{k'}$ to $\mathcal{C}_0$ must pass through $\mathcal{C}_3$. Therefore, this path is longer than the path from $\mathcal{C}_1$ to $\mathcal{C}_0$. Therefore, we would have had to operate on the triple $\langle \mathcal{C}_{k'}, \mathcal{C}_{k'-1}, \mathcal{C}_{k'-2} \rangle$ before $\langle \mathcal{C}_1, \mathcal{C}_3, \mathcal{C}_2 \rangle$ giving rise to a contradiction.

(A2) and (B2). Now, consider the case where (A2) was present in $\mathcal{T}_1$ and (B2) is present in $\mathcal{T}_2$. Since (A2) was in $\mathcal{T}_1$ and we operated on $\langle \mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3 \rangle$ in $\mathcal{T}_1$, it must be that $\mathcal{C}_1$ is farther away from $\mathcal{C}_0$ than $\mathcal{C}_k$. In other words, the path from $\mathcal{C}_1$ to $\mathcal{C}_0$ must pass through some subsequence of (A2). However, this must imply that, in $\mathcal{T}_2$, the path from $\mathcal{C}_{k'}$ to $\mathcal{C}_0$ must pass through $\mathcal{C}_3$. Therefore, $\mathcal{C}_{k'}$ is farther away from $\mathcal{C}_0$ than $\mathcal{C}_1$. So we would have had to operate on the triple $\langle \mathcal{C}_{k'}, \mathcal{C}_{k'-1}, \mathcal{C}_{k'-2} \rangle$ before $\langle \mathcal{C}_1, \mathcal{C}_3, \mathcal{C}_2 \rangle$ giving rise to a contradiction.

□

Corollary C.7.29. *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be an ancestral partial mixed graph with a chordal skeleton such that $\mathcal{G}$ has no minimal collider paths and such that orientations in $\mathcal{G}$ are closed under R1 and R11. Let $\mathcal{T}_0$ be any join tree for $\mathcal{G}$, $\mathcal{C}_0$ a node in $\mathcal{T}_0$ and let $\mathcal{T}$ be the output of Algorithm 17, that is $\mathcal{T} =$ `transformTree`$(\mathcal{T}_0, \mathcal{C}_0)$. Then*

(i) $\mathcal{T}$ is also a join tree for $\mathcal{G}$,

(ii) for any pair of cliques, if $\gamma(\mathcal{C}_i, \mathcal{C}_j)$ in $\mathcal{T}_0$, then $\gamma(\mathcal{C}_i, \mathcal{C}_j)$ in $\mathcal{T}$ as well,

(iii) $\mathcal{T}$ does not contain any colliders, or paths of the form $\mathcal{C}_{i_1} \to \mathcal{C}_{i_2} - \cdots - \mathcal{C}_{i_k} \leftarrow \mathcal{C}_{i_{k+1}}$, $k > 2$,

(iv) $\mathcal{T}$ does not contain paths of the form $\mathcal{C}_{i_1} \to \mathcal{C}_{i_2} - \cdots - \mathcal{C}_{i_k} \to \cdots \to \mathcal{C}_{i_r}$, $r \geq k > 2$, where $\mathcal{C}_{i_r} \equiv \mathcal{C}_0$.

Proof of Corollary C.7.29. From Lemmas C.7.20 and C.7.28 we know that Algorithm 17 terminates. Lemmas C.7.21 and C.7.26 tell us that cases (i) and (ii) are true. Cases (iii) and (iv) are true by construction of Algorithm 17. □

C.7.5 Orienting a Partially Directed Join Tree and its Cliques

Before we discuss Algorithm 18, we state and prove a useful set identity.

Algorithm 18 orientTree

Input: Partially directed join tree $\mathcal{T} = (\mathbf{C}, \mathbf{E})$, and node $\mathcal{C}_0 \in \mathbf{C}$ for an ancestral partial mixed graph $\mathcal{G}$ with a chordal skeleton, with edge orientations closed under R1-R4, R8-R12 and such that $\mathcal{G}$ is without minimal collider paths.

Output: Directed join tree $\mathcal{T}' = (\mathbf{C}, \mathbf{E}')$.

1: $\mathcal{T}' \leftarrow \texttt{transformTree}(\mathcal{T}, \mathcal{C}_0)$
2: **while** an undirected edge is in $\mathcal{T}'$ **do**
3: Let $p = \langle \mathcal{C}_{j_1}, \ldots, \mathcal{C}_{j_k} \rangle, k > 1$ be a longest undirected path in $\mathcal{T}'$
4: **if** $\mathcal{C}_{j_1} \in \mathrm{An}(\mathcal{C}_0, \mathcal{T}')$ or $\exists\, \mathcal{C}_j \in \mathbf{C}$, such that $\mathcal{C}_j \in \mathrm{Pa}(\mathcal{C}_{j_1}, \mathcal{T}')$ **then**
5: orient p as $\mathcal{C}_{j_1} \to \cdots \to \mathcal{C}_{j_k}$ in $\mathcal{T}'$
6: **else**
7: orient p as $\mathcal{C}_{j_1} \leftarrow \cdots \leftarrow \mathcal{C}_{j_k}$ in $\mathcal{T}'$
8: **return** $\mathcal{T}'$

Proposition C.7.30. *For any three subsets $A, B, C \subseteq \mathbf{V}$ of some finite set $\mathbf{V}$ i.e., $|\mathbf{V}| < \infty$, we have that*

$$B \setminus A \subseteq (B \setminus C) \cup (C \setminus A).$$

Proof of Proposition C.7.30. Since $\mathbf{V}$ is finite, set complements are well-defined. Specifically, $C^c \cup C = \mathbf{V}$. Further, we know that $B \setminus A = B \cap A^c$. Then,

$$\begin{aligned}
B \setminus A &= (B \setminus A) \cap \mathbf{V} \\
&= (B \setminus A) \cap (C^c \cup C) \\
&= ((B \setminus A) \cap C^c)) \cup ((B \setminus A) \cap C)) \\
&= (B \cap A^c \cap C^c)) \cup (B \cap A^c \cap C)) \\
&= (B \cap C^c \cap A^c)) \cup (C \cap A^c \cap B)) \\
&= ((B \setminus C) \cap A^c)) \cup ((C \setminus A) \cap B)) \\
&\subseteq (B \setminus C) \cup (C \setminus A)
\end{aligned}$$

□

Lemma C.7.31. *Let $\mathcal{G}$ be an ancestral partial mixed graph with a chordal skeleton such that $\mathcal{G}$ has no minimal collider paths such that the orientations in $\mathcal{G}$ are closed under R1-R4 and R8-R12. Let $\mathcal{T}_0$ be a partially directed join tree for $\mathcal{G}$ as defined in Definition C.7.10 and let $\mathcal{C}_0$ be a node in $\mathcal{T}_0$. Furthermore, let $\mathcal{T}_1 = \texttt{transformTree}(\mathcal{T}_0, \mathcal{C}_0)$ (Algorithm 17) and $\mathcal{T} = \texttt{orientTree}(\mathcal{T}_0, \mathcal{C}_0)$ (Algorithm 18). Also, let $\pi_{\mathcal{T}}$ be a partial order compatible with $\mathcal{T}$. Then the following hold:*

(i) *$\mathcal{T}$ is a directed join tree for $\mathcal{G}$ that does not contain colliders and $\mathrm{An}(\mathcal{C}_0, \mathcal{T}_1) = \mathrm{An}(\mathcal{C}_0, \mathcal{T})$.*

(ii) *$\pi_{\mathcal{T}}$ induces a edge orientations that are compatible with $\mathcal{G}$. Call this induced graph $\mathcal{G}_\pi$ (Definition C.7.11).*

(iii) *For any node $A \in \mathcal{C}_0$ there are no new edge marks into A in $\mathcal{G}_\pi$ compared to $\mathcal{G}$. Furthermore, for any pair of nodes $A, B \in \mathcal{C}_0$, $\langle A, B\rangle$ is of the same form in $\mathcal{G}$ and $\mathcal{G}_\pi$.*

(iv) *If path $\langle A, V_1, \ldots, V_k, D\rangle$, $k \geq 1$ is in $\mathcal{G}_\pi$ such that $\{A, V_1, \ldots, V_k\} \subseteq \mathcal{C}_i$, and $\{V_1, \ldots, V_k, D\} \subseteq \mathcal{C}_j$, for some maximal cliques $\mathcal{C}_i, \mathcal{C}_j$ in $\mathcal{G}_\pi$, and also $A \notin \mathrm{Adj}(D, \mathcal{G}_\pi)$, then at least one of the following holds:*

- *$V_t \to D$ is in $\mathcal{G}_\pi$, for all $t \in \{1, \ldots, k\}$.*
- *$V_t \to A$ is in $\mathcal{G}_\pi$, for all $t \in \{1, \ldots, k\}$.*

(v) *If $A \bullet\!\!\to B \to C$ is in $\mathcal{G}_\pi$ and $A \in \mathrm{Adj}(C, \mathcal{G}_\pi)$, where $B \to C$ is induced by $\pi_{\mathcal{T}}$, then $A \to C$ is in $\mathcal{G}_\pi$.*

(vi) *If $A \to B \bullet\!\!\to C$ is in $\mathcal{G}_\pi$ and $A \in \mathrm{Adj}(C, \mathcal{G}_\pi)$, where $A \to B$ is induced by $\pi_{\mathcal{T}}$, then $A \to C$ is in $\mathcal{G}_\pi$.*

(vii) $\mathcal{G}_\pi$ *is ancestral, and edge orientations in* $\mathcal{G}_\pi$ *are complete with respect to R2, and R8. Furthermore,* $\mathcal{G}_\pi$ *contains no minimal collider paths and neither does any directed mixed graph* $\mathcal{M}$ *that is represented by* $\mathcal{G}_\pi$.

Proof of Lemma C.7.31. (i) We have, $\mathcal{T}_1 = \texttt{transformTree}(\mathcal{T}_0, \mathcal{C}_0)$. By Corollary C.7.29 there are no paths in $\mathcal{T}_1$ that are of the forms:

- $\mathcal{C}_{i_1} \to \mathcal{C}_{i_2} \leftarrow \mathcal{C}_{i_3}$, or
- $\mathcal{C}_{i_1} \to \mathcal{C}_{i_2} - \cdots - \mathcal{C}_{i_k} \leftarrow \mathcal{C}_{i_{k+1}}$, $k > 2$, or
- $\mathcal{C}_{i_1} \to \mathcal{C}_{i_2} - \cdots - \mathcal{C}_{i_k} \to \cdots \to \mathcal{C}_{i_r}$, $r \geq k > 2$, where $\mathcal{C}_{i_r} \equiv \mathcal{C}_0$.

Orienting paths as in Algorithm 18 will not create colliders in $\mathcal{T}$. Further, we will not create new ancestors for $\mathcal{C}_0$ as we always orient paths away from existing ancestors of $\mathcal{C}_0$. By construction of Algorithm 18, all ancestors of $\mathcal{C}_0$ in $\mathcal{T}_1$ are also ancestors of $\mathcal{C}_0$ in $\mathcal{T}_0$. Therefore, $\text{An}(\mathcal{C}_0, \mathcal{T}_1) = \text{An}(\mathcal{C}_0, \mathcal{T})$.

(ii) Note that $\pi_\mathcal{T}$ only induces directed edges in $\mathcal{G}_\pi$ by Definition C.7.11. Hence, to show that edge orientations induced by $\pi_\mathcal{T}$ are compatible with $\mathcal{G}$, we need to show that it is possible to orient $A \to B$ for every $A \in \mathcal{C}_i \cap \mathcal{C}_j$, $B \in \mathcal{C}_j \setminus \mathcal{C}_i$ whenever $\pi_\mathcal{T}(\mathcal{C}_i, \mathcal{C}_j)$ holds in $\mathcal{T}$.

For any two maximal cliques $\mathcal{C}_i$ and $\mathcal{C}_j$ in $\mathcal{G}$ such that $\pi_\mathcal{T}(\mathcal{C}_i, \mathcal{C}_j)$ and $\mathcal{C}_i, \mathcal{C}_j$ are adjacent in $\mathcal{T}$, $\mathcal{C}_i \to \mathcal{C}_j$ is in $\mathcal{T}$ either because $\gamma(\mathcal{C}_i, \mathcal{C}_j)$ holds, or because this edge got oriented by Algorithm 18. In the former case, the induced orientations in $\mathcal{G}_\pi$ are surely compatible with orientations already in $\mathcal{G}$. In the latter case, it must be that $\neg\gamma(\mathcal{C}_i, \mathcal{C}_j)$ and $\neg\gamma(\mathcal{C}_j, \mathcal{C}_i)$. With $\neg\gamma(\mathcal{C}_j, \mathcal{C}_i)$, the contraposition of Lemma C.7.13 tells us that there is no edge $A \leftarrow\!\bullet B$, $A \in \mathcal{C}_i \cap \mathcal{C}_j, B \in \mathcal{C}_j \setminus \mathcal{C}_i$ in $\mathcal{G}$. Therefore, all such edges in $\mathcal{G}$ must be either $A \circ\!\!-\!\!\bullet B$ or $A \to B$. Thus, it is possible to orient all such edges as $A \to B$ in $\mathcal{G}_\pi$.

For any two maximal cliques $\mathcal{C}_i$ and $\mathcal{C}_k$ in $\mathcal{G}$ such that $\pi_\mathcal{T}(\mathcal{C}_i, \mathcal{C}_k)$ and $\mathcal{C}_i \cap \mathcal{C}_k \neq \varnothing$, but $\mathcal{C}_i$ and $\mathcal{C}_k$ are not adjacent in $\mathcal{T}$, there is a path $p = \langle \mathcal{C}_i = \mathcal{C}_{j_1}, \mathcal{C}_{j_2}, \ldots, \mathcal{C}_{j_r} = \mathcal{C}_k \rangle$, $r > 2$

of the form $\mathcal{C}_i \to \cdots \to \mathcal{C}_k$ in $\mathcal{T}$. We will now prove the rest of this claim by using an induction argument on the length of p. For clarity and conciseness, below we will use the following shorthand $\Lambda_{j_t j_s} \to \mathcal{C}_{j_s} \setminus \mathcal{C}_{j_t}$ for $t, s \in \{1, \ldots, r\}, t \neq r$, to say that it is *possible* to orient all edges $\langle A, B \rangle$, such that $A \in \Lambda_{j_t j_s}$, $B \in \mathcal{C}_{j_s} \setminus \mathcal{C}_{j_t}$ as $A \to B$ in $\mathcal{G}_\pi$.

For the base of the induction suppose that $r = 3$ i.e., p is of the form $\mathcal{C}_i \to \mathcal{C}_{j_2} \to \mathcal{C}_k$. If $\Lambda_{ik} = \varnothing$, we are done. Hence, suppose that $\Lambda_{ik} \neq \varnothing$.

From previous argument for adjacent nodes, we have that $\Lambda_{ij_2} \to \mathcal{C}_{j_2} \setminus \mathcal{C}_i$ and $\Lambda_{j_2 k} \to \mathcal{C}_k \setminus \mathcal{C}_{j_2}$. By the join tree running intersection property, we have that $\Lambda_{ik} \subseteq \mathcal{C}_{j_2}$. Therefore, $\Lambda_{ik} \subseteq \Lambda_{ij_2}$ and $\Lambda_{ik} \subseteq \Lambda_{j_2 k}$. Then, to show that $\Lambda_{ik} \to \mathcal{C}_k \setminus \mathcal{C}_i$ it is enough to show that $\mathcal{C}_k \setminus \mathcal{C}_i \subseteq (\mathcal{C}_k \setminus \mathcal{C}_{j_2}) \cup (\mathcal{C}_{j_2} \setminus \mathcal{C}_i)$. This follows from Proposition C.7.30 as the vertex set $\mathbf{V}$ is finite.

For the induction hypothesis suppose that the claim holds for every path of length t, $t \geq 3$. We will show that then it also holds for the path of length $t + 1$. Let $r = t + 1$ i.e., $p = \langle \mathcal{C}_i, \mathcal{C}_{j_2}, \ldots \mathcal{C}_{j_t}, \mathcal{C}_k \rangle$. If $\Lambda_{ik} = \varnothing$, we have nothing to prove, so suppose $\Lambda_{ik} \neq \varnothing$ The goal is then again to show that $\Lambda_{ik} \to \mathcal{C}_k \setminus \mathcal{C}_i$.

We know that $\Lambda_{j t_k} \to \mathcal{C}_k \setminus \mathcal{C}_{j_t}$ holds, and from the induction hypothesis, we also know that $\Lambda_{i j_t} \to \mathcal{C}_{j_t} \setminus \mathcal{C}_i$ holds. By the intersection property, we also have that $\Lambda_{ik} \subseteq C_{j_l}$, for every $l \in \{2, \ldots, t\}$. Therefore, $\Lambda_{ik} \subseteq \Lambda_{i j_t}$, and $\Lambda_{ik} \subseteq \Lambda_{j_t k}$. Similar to the base case, it is enough to show that $\mathcal{C}_k \setminus \mathcal{C}_i \subseteq (\mathcal{C}_k \setminus \mathcal{C}_{j_t}) \cup (\mathcal{C}_{j_t} \setminus \mathcal{C}_i)$. This, of course, follows from Proposition C.7.30 like before.

(iii) First, note that by construction, $\mathcal{T}_1$ is a partially directed join tree for $\mathcal{G}$ (Corollary C.7.29). Hence, case (ii) implies that the only way to obtain new edge marks into A in $\mathcal{G}_\pi$ is by adding new ancestors of $\mathcal{C}_0$ in $\mathcal{T}$, compared to $\mathcal{T}_1$. But we know by case (i), that no such edge marks are added.

For the statement about the form of $\langle A, B \rangle$, note that an edge is of different form in $\mathcal{G}_\pi$ compared to $\mathcal{G}$, only if its orientation is induced by $\pi_\mathcal{T}$. Also, since $\mathcal{T}_1$ is a partially

directed join tree for $\mathcal{G}$, only orientations added to $\mathcal{T}_1$ to create $\mathcal{T}$ would be able to orient $\langle A, B\rangle$ through $\pi_{\mathcal{T}}$.

Since $A, B \in \mathcal{C}_0$, the only way to orient $\langle A, B\rangle$ in some way in $\mathcal{G}_\pi$ is if there is a clique $\mathcal{C}_i$, such that $\mathcal{C}_i$ is an ancestor of $\mathcal{C}_0$ in $\mathcal{T}$, but not in $\mathcal{T}_1$. By case (i), $\text{An}(\mathcal{C}_0, \mathcal{T}) \backslash \text{An}(\mathcal{C}_0, \mathcal{T}_1) = \varnothing$. Hence, $\langle A, B\rangle$ must be of the same form in both $\mathcal{G}_\pi$ and $\mathcal{G}$.

(iv) Note that the mutually exclusive and collectively exhaustive options for $\mathcal{C}_i$ and $\mathcal{C}_j$ are

(a) $\pi_{\mathcal{T}}(\mathcal{C}_i, \mathcal{C}_j)$: Here, $V_t \to D$ for all $t \in \{1, \ldots, k\}$ by Definition C.7.11 and cases (i), and (ii).

(b) $\pi_{\mathcal{T}}(\mathcal{C}_j, \mathcal{C}_i)$: Here, $V_t \to A$ for all $t \in \{1, \ldots, k\}$ by Definition C.7.11 and cases (i), and (ii).

(c) $\neg\pi_{\mathcal{T}}(\mathcal{C}_i, \mathcal{C}_j) \wedge \neg\pi_{\mathcal{T}}(\mathcal{C}_j, \mathcal{C}_i)$: Here, by case (i) there exists a maximal clique $\mathcal{C}_l$ in $\mathcal{G}_\pi$ such that the path between $\mathcal{C}_i$ and $\mathcal{C}_j$ in $\mathcal{T}$ is of the form $\mathcal{C}_i \leftarrow \cdots \leftarrow \mathcal{C}_l \to \cdots \to \mathcal{C}_j$. By the running intersection property $\{V_1, \ldots, V_k\} \subseteq \mathcal{C}_l$. Case (ii) implies that we have that $\pi_{\mathcal{T}}(\mathcal{C}_l, \mathcal{C}_i)$ and $\pi_{\mathcal{T}}(\mathcal{C}_l, \mathcal{C}_j)$. Furthermore, at least one of the nodes A, D is not in $\mathcal{C}_l$ because $A \notin \text{Adj}(D, \mathcal{G}_\pi)$. Without loss of generality, assume $A \notin \mathcal{C}_l$. Then $\pi_{\mathcal{T}}(\mathcal{C}_l, \mathcal{C}_i)$ implies that $V_t \to A$ is in $\mathcal{G}_\pi$ for all $t \in \{1, \ldots, k\}$. A symmetric argument holds when $D \notin \mathcal{C}_l$.

(v) By assumption, $A \bullet\!\!\to B \to C$ is in $\mathcal{G}_\pi$, $A \in \text{Adj}(C, \mathcal{G})$ and $B \to C$ is induced by $\pi_{\mathcal{T}}$. Then there are maximal cliques $\mathcal{C}_i, \mathcal{C}_j$, and $\mathcal{C}_k$ in $\mathcal{G}_\pi$ such that the following holds:

- $\mathcal{C}_i \supseteq \{B\}$, and $C \notin \mathcal{C}_i$,
- $\mathcal{C}_j \supseteq \{B, C\}$, and $\pi_{\mathcal{T}}(\mathcal{C}_i, \mathcal{C}_j)$, and
- $\mathcal{C}_k \supseteq \{A, B, C\}$.

Next we consider whether A belongs to $\mathcal{C}_i, \mathcal{C}_j$. We have the following cases: (a) $A \in \mathcal{C}_j \setminus \mathcal{C}_i$, (b) $A \notin \mathcal{C}_i \cup \mathcal{C}_j$, (c)$A \in \mathcal{C}_j \cap \mathcal{C}_i$, or (d) $A \in \mathcal{C}_i \setminus \mathcal{C}_j$. For the rest of the

proof, we show that the cases (a) and (b) are in fact not possible, since they lead to a contradiction, while cases (c) and (d) lead us to conclude that $A \to C$ is in $\mathcal{G}_\pi$.

(a) Since $A \bullet\!\!\to B$ is in $\mathcal{G}_\pi$, we know that A cannot be in $\mathcal{C}_j \setminus \mathcal{C}_i$.

(b) $A \in \mathcal{C}_k \setminus (\mathcal{C}_i \cup \mathcal{C}_j)$: Since $B \in \mathcal{C}_k \cap \mathcal{C}_i$ and $A \bullet\!\!\to B$ is in $\mathcal{G}_\pi$, we know that $\neg\pi_{\mathcal{T}}(\mathcal{C}_i, \mathcal{C}_k)$ and $\neg\pi_{\mathcal{T}}(\mathcal{C}_j, \mathcal{C}_k)$. Since we also know that $\pi_{\mathcal{T}}(\mathcal{C}_i, \mathcal{C}_j)$, let us consider the options for paths between $\mathcal{C}_i, \mathcal{C}_j$ and $\mathcal{C}_k$. Let p_{ij} be the path from $\mathcal{C}_i$ to $\mathcal{C}_j$ in $\mathcal{T}$, p_{ik} the path from $\mathcal{C}_i$ to $\mathcal{C}_k$ and p_{jk} the path from $\mathcal{C}_j$ to $\mathcal{C}_k$ in $\mathcal{T}$. The only options are that: (1) $\mathcal{C}_i$ is on p_{jk}, or that (2) a node from p_{ij} other than $\mathcal{C}_i$ is on p_{ik}.

(1) Since $\mathcal{C}_k \cap \mathcal{C}_j \not\subseteq \mathcal{C}_i$, the running intersection property of $\mathcal{T}$ implies that $\mathcal{C}_i$ cannot be on p_{jk}.

(2) $\pi_{\mathcal{T}}(\mathcal{C}_i, \mathcal{C}_j)$ and $\neg\pi_{\mathcal{T}}(\mathcal{C}_i, \mathcal{C}_k)$ together imply that $\mathcal{C}_k$ is not on p_{ij}, and also that no other node from p_{ij} except $\mathcal{C}_i$ is on p_{ik}.

(c) If $A \in \mathcal{C}_j \cap \mathcal{C}_i$, then $A \to C$ is in $\mathcal{G}_\pi$ by $\pi_{\mathcal{T}}(\mathcal{C}_i, \mathcal{C}_j)$.

(d) $A \in (\mathcal{C}_i \cap \mathcal{C}_k) \setminus \mathcal{C}_j$, then as above, let p_{ij} be the path from $\mathcal{C}_i$ to $\mathcal{C}_j$ in $\mathcal{T}$, p_{ik} the path from $\mathcal{C}_i$ to $\mathcal{C}_k$ and p_{jk} the path from $\mathcal{C}_j$ to $\mathcal{C}_k$ in $\mathcal{T}$. The only options are that: (1) $\mathcal{C}_i$ is on p_{jk}, or that (2) a node from p_{ij} other than $\mathcal{C}_i$ is on p_{ik}.

(1) Since $\mathcal{C}_k \cap \mathcal{C}_j \supseteq \{B, C\} \not\subseteq \mathcal{C}_i$, the running intersection property of $\mathcal{T}$ implies that $\mathcal{C}_i$ cannot be on p_{jk}.

(2) Since $\pi_{\mathcal{T}}(\mathcal{C}_i, \mathcal{C}_j)$, having $\mathcal{C}_k$ on p_{ij}, implies $\pi_{\mathcal{T}}(\mathcal{C}_i, \mathcal{C}_k)$ and therefore, $A \to C$ is in $\mathcal{G}_\pi$. Similarly, having any node from p_{ij} except $\mathcal{C}_i$ on p_{ik} implies the same thing.

(vi) By assumption, $A \to B \bullet\!\!\to C$ is in $\mathcal{G}_\pi$, $A \in \mathrm{Adj}(C, \mathcal{G})$ and $A \to B$ is induced by $\pi_{\mathcal{T}}$. Then there are maximal cliques $\mathcal{C}_i, \mathcal{C}_j$, and $\mathcal{C}_k$ in $\mathcal{G}_\pi$ such that the following holds:

- $\mathcal{C}_i \supseteq \{A\}$, and $B \notin \mathcal{C}_i$,
- $\mathcal{C}_j \supseteq \{A, B\}$, and $\pi_{\mathcal{T}}(\mathcal{C}_i, \mathcal{C}_j)$, and

- $\mathcal{C}_k \supseteq \{A, B, C\}$.

Next we consider whether C belongs to $\mathcal{C}_i, \mathcal{C}_j$. We have the following cases: (b) $C \in \mathcal{C}_i \setminus \mathcal{C}_j$, (a)$C \in \mathcal{C}_j \cap \mathcal{C}_i$, (c) $C \in \mathcal{C}_j \setminus \mathcal{C}_i$, or (d) $C \notin \mathcal{C}_i \cup \mathcal{C}_j$. For the rest of the proof, we show that the cases (b) and (a) are in fact not possible, since they lead to a contradiction, while cases (c) and (d) lead us to conclude that $A \to C$ is in $\mathcal{G}_\pi$.

(a) If $C \in \mathcal{C}_j \cap \mathcal{C}_i$, then $B \in \mathcal{C}_j \setminus \mathcal{C}_i, \pi_{\mathcal{T}}(\mathcal{C}_i, \mathcal{C}_j)$ and $B \bullet\!\!\to C$ together imply a contradiction.

(b) $C \in (\mathcal{C}_i \cap \mathcal{C}_k) \setminus \mathcal{C}_j$. Since $B \in \mathcal{C}_k \setminus \mathcal{C}_i$, and $B \bullet\!\!\to C$ is in $\mathcal{G}_\pi$, we have that $\neg\pi_{\mathcal{T}}(\mathcal{C}_i, \mathcal{C}_k)$. Now, let p_{ij} be the path from $\mathcal{C}_i$ to $\mathcal{C}_j$ in $\mathcal{T}$, p_{ik} the path from $\mathcal{C}_i$ to $\mathcal{C}_k$ and p_{jk} the path from $\mathcal{C}_j$ to $\mathcal{C}_k$ in $\mathcal{T}$. The only options are that: (1) $\mathcal{C}_i$ is on p_{jk}, or that (2) a node from p_{ij} other than $\mathcal{C}_i$ is on p_{ik}.

(1) Since $\mathcal{C}_k \cap \mathcal{C}_j \supseteq \{A, B\} \not\subseteq \mathcal{C}_i$, the running intersection property of $\mathcal{T}$ implies that $\mathcal{C}_i$ cannot be on p_{jk}.

(2) Since $\pi_{\mathcal{T}}(\mathcal{C}_i, \mathcal{C}_j)$, having a node from p_{ij} other than $\mathcal{C}_i$ on p_{ik} would imply $\pi_{\mathcal{T}}(\mathcal{C}_i, \mathcal{C}_k)$ which is a contradiction.

(c) $C \in (\mathcal{C}_j \cap \mathcal{C}_k) \setminus \mathcal{C}_i$. Since $C \in \mathcal{C}_j \setminus \mathcal{C}_i$ and $A \in \mathcal{C}_i \cap \mathcal{C}_j$, then $\pi_{\mathcal{T}}(\mathcal{C}_i, \mathcal{C}_j)$ implies $A \to C$ is in $\mathcal{G}_\pi$.

(d) $C \in \mathcal{C}_k \setminus (\mathcal{C}_i \cup \mathcal{C}_j)$. Let us consider the options for paths between $\mathcal{C}_i, \mathcal{C}_j$ and $\mathcal{C}_k$. Let p_{ij} be the path from $\mathcal{C}_i$ to $\mathcal{C}_j$ in $\mathcal{T}$, p_{ik} the path from $\mathcal{C}_i$ to $\mathcal{C}_k$ and p_{jk} the path from $\mathcal{C}_j$ to $\mathcal{C}_k$ in $\mathcal{T}$. The only options are that: (1) $\mathcal{C}_i$ is on p_{jk}, or that (2) a node from p_{ij} other than $\mathcal{C}_i$ is on p_{ik}.

(1) Since $\mathcal{C}_k \cap \mathcal{C}_j \supseteq \{A, B\} \not\subseteq \mathcal{C}_i$, the running intersection property of $\mathcal{T}$ implies that $\mathcal{C}_i$ cannot be on p_{jk}.

(2) If a node on p_{ij} other than $\mathcal{C}_i$ is on p_{ik},t that implies that $\pi_{\mathcal{T}}(\mathcal{C}_i, \mathcal{C}_k)$. Since $A \in \mathcal{C}_i \cap \mathcal{C}_k$ and $C \in \mathcal{C}_k \setminus \mathcal{C}_i$, we have that $A \to C$ is in $\mathcal{G}_\pi$.

□

Lemma C.7.32. *Suppose an ancestral partial mixed graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ with edge orientations complete under R2 and R8 is a clique that contains no edges of the form $\circ\!\!\rightarrow$ or $\leftrightarrow$. Consider edge $A \circ\!\!-\!\!\circ B$ in $\mathcal{G}$ for some $A, B \in \mathbf{V}$. Then there is are total orderings π_1 and π_2 of $\mathbf{V}$ compatible with $\mathcal{G}$, such that $\mathcal{G}_{\pi_1}$ and $\mathcal{G}_{\pi_2}$ are DAGs and such that $A \rightarrow B$ is in $\mathcal{G}_{\pi_1}$ and $A \leftarrow B$ is in $\mathcal{G}_{\pi_2}$.*

Proof of Lemma C.7.32. We will show how to obtain π_1 using the sink elimination Algorithm of Dor and Tarsi (1992). The proof for π_2 is analogous.

Since $\mathcal{G}$ is ancestral and therefore, acyclic, there will always be at least one node V in $\mathcal{G}$ such that there are no edges out of V in $\mathcal{G}$. This type of node is called a potential sink node according to Dor and Tarsi (1992) algorithm since $\mathcal{G}$ is a clique.

To obtain π_1, we consider whether B is a potential sink node in $\mathcal{G}$.

(i) If B is a potential sink, let $\pi^{(1)}$ be a partial ordering that only states that $\pi^{(1)}(W, B)$ for every node $W \in \mathbf{V}$. Then consider, the induced subgraph $\mathcal{G}_{\mathbf{V}\setminus\{B\}} = (\mathbf{V}_{-B}, \mathbf{E}_{-B})$ where $\mathbf{V}_{-B} = \mathbf{V} \setminus \{B\}$ and $\mathbf{E}_{-B} = \{(X, Y) \in \mathbf{E} \mid X \neq B, Y \neq B\}$. $\mathcal{G}_{\mathbf{V}\setminus\{B\}}$ is also a clique that is ancestral and does not contain $\circ\!\!\rightarrow$ or $\leftrightarrow$ edges. We can then apply the Algorithm of Dor and Tarsi (1992) to $\mathcal{G}_{\mathbf{V}\setminus\{B\}}$ to obtain a total ordering $\pi^{(2)}$ of $\mathbf{V} \setminus \{B\}$. We can construct π_1 as follows:

$$\pi^{(1)}(V_1, V_2) \implies \pi_1(V_1, V_2),$$
$$\pi^{(2)}(V_1, V_2) \implies \pi_1(V_1, V_2).$$

It is easy to see that π_1 is compatible with $\mathcal{G}$ by construction and $\mathcal{G}_{\pi_1}$ is a DAG with $A \rightarrow B$.

(ii) If B is not a potential sink, then since a potential sink node must exist in $\mathcal{G}$, we only need to show that there is a potential sink node that is different from A in $\mathcal{G}$. Note that since B is not a potential sink there is a node $B \rightarrow V_2$, for some $V_2 \in \mathbf{V}$ in $\mathcal{G}$.

If A was the only potential sink node in $\mathcal{G}$, that would mean that there is a path $B \to V_2 \to \cdots \to V_k \to A$, $k \geq 2$ in $\mathcal{G}$. However, since $\mathcal{G}$ is an ancestral clique with edge orientations complete under R2 and R8, the successive edges $B \to V_3, \ldots B \to V_k$, $B \to A$ are in $\mathcal{G}$. This contradicts $A \circ\!\!-\!\!\circ B$ being in $\mathcal{G}$. Hence, there is at least one potential sink node that is different from A in $\mathcal{G}$.

Suppose this potential sink node in $\mathcal{G}$ that is different from A is called V. Let $\pi^{(1)}$ be a partial ordering that only states that $\pi^{(1)}(W, B)$ for every node $W \in \mathbf{V}$. Then consider, the induced subgraph $\mathcal{G}_{\mathbf{V}\setminus\{V\}}$ (defined like before) which is also an ancestral clique that does not contain $\circ\!\!\to$ or $\leftrightarrow$ edges.

If B is a potential sink in $\mathcal{G}_{\mathbf{V}\setminus\{V\}}$, we can apply step (i) to $\mathcal{G}_{\mathbf{V}\setminus\{V\}}$ to obtain a total ordering $\pi^{(2)}$ of $\mathbf{V} \setminus \{V\}$ that is compatible with $\mathcal{G}$. Then we extend $\pi^{(1)}$ to π_2 using $\pi^{(2)}$, as follows

$$\pi^{(1)}(V_1, V_2) \implies \pi_2(V_1, V_2),$$
$$\pi^{(2)}(V_1, V_2) \implies \pi_2(V_1, V_2).$$

This is the desired ordering: π_2 is compatible with $\mathcal{G}$ by construction and $\mathcal{G}_{\pi_2}$ is a DAG with $A \to B$.

If B is not a potential sink in $\mathcal{G}_{\mathbf{V}\setminus\{V\}}$, we can apply step (ii) on $\mathcal{G}_{\mathbf{V}\setminus\{V\}}$ to obtain $\pi^{(2)}$ and recursively continue obtaining $\pi^{(3)}, \ldots, \pi^{(l)}$ until $\mathcal{G}_{\mathbf{V}\setminus\mathbf{S}}$, for some $\mathbf{S} \supset \{V\}$ such that B is a potential sink in $\mathcal{G}_{\mathbf{V}\setminus\mathbf{S}}$. Then we apply step (i) to $\mathcal{G}_{\mathbf{V}\setminus\mathbf{S}}$ which gives us partial ordering $\pi^{(l+1)}$. Finally, we construct the desired total ordering π_2, where for any $V_1, V_2 \in \mathbf{V}$:

$$\pi^{(j)}(V_1, V_2) \implies \pi_2(V_1, V_2) \qquad \text{for all } j \in \{1, \ldots, l+1\}.$$

□

Lemma C.7.33. *Suppose an ancestral partial mixed graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ with edge orientations complete under R2 and R8 is a clique. Consider the graph $\mathcal{G}'$ obtained from $\mathcal{G}$ in one of the two following ways:*

(a) Orient all variant edge marks ($\circ$) as arrowheads. That is, orient edges of the form $V_i \circ\!\!-\!\!\circ V_j$ and of the form $V_i \circ\!\!\rightarrow V_j$ as $V_i \leftrightarrow V_j$.

(b) Choose an edge $A \circ\!\!-\!\!\bullet B$ in $\mathcal{G}$. Then

(1) orient $A \rightarrow B$ in $\mathcal{G}$, and

(2) for all C in $\mathcal{G}$ such that $B \rightarrow C$ is in $\mathcal{G}$, orient $A \rightarrow C$, and

(3) for all D in $\mathcal{G}$ such that $B \circ\!\!\rightarrow D$ or $B \leftrightarrow D$ is in $\mathcal{G}$, orient the edge mark at D on edge $\langle A, D \rangle$ as an arrowhead, that is, $A \bullet\!\!\rightarrow D$ and orient $B \leftrightarrow D$.

Then, orient all remaining $V_i \circ\!\!\rightarrow V_j$ or $V_i \circ\!\!-\!\!\circ V_j$ edges in $\mathcal{G}$ as $V_i \leftrightarrow V_j$.

Then $\mathcal{G}'$ is a MAG represented by $\mathcal{G}$.

Proof of Lemma C.7.33. Note that for $\mathcal{G}'$ to be a MAG represented by $\mathcal{G}$ it is enough to show that $\mathcal{G}'$ does not contain directed or almost directed cycles of length 3.

For case (a), we only need to worry about creating almost directed cycles in $\mathcal{G}'$. We know these cannot be created in $\mathcal{G}'$ since, $\mathcal{G}$ cannot contain $V_1 \rightarrow V_2 \rightarrow V_3$, and $V_1 \circ\!\!-\!\!\circ V_3$ for any three nodes V_1, V_2, V_3 due to orientations in $\mathcal{G}$ being complete under R2 and R8.

For case (b), note that steps in (1)-(3) ensure that orientations under R2 and R8 are closed after adding $A \rightarrow B$. Hence, as long as steps (1)-(3) can be performed and do not create a directed or almost directed cycle, the remainder of the proof follows by case (a) above.

By assumption, step (1) can be performed. Additionally, step (1) cannot in itself create a directed or almost directed cycle since $\mathcal{G}$ is a clique with edge mark orientations completed under R2 and R8.

As for step (2), note that for any C in $\mathcal{G}$, such that $A \circ\!\!-\!\!\bullet B \to C$ is in $\mathcal{G}$, $A \circ\!\!-\!\!\bullet C$ must be in $\mathcal{G}$ again, due to edge mark orientations being completed in $\mathcal{G}$ under R2 and R8. Hence, step (2) can be performed.

Furthermore, completing step (2) cannot create a directed cycle. To see why, observe that a directed cycle would imply that $C \to E \to A$ was already in $\mathcal{G}$ for some node E. This is because in steps (1) and (2) we do not create any new arrowheads into A and do not orient any edge marks on edges that do not contain A.

Since we know $C \to E \to A$ and $A \circ\!\!-\!\!\bullet C$ cannot both be in $\mathcal{G}$, we know that orienting $A \to C$ does not create a directed cycle. Using a similar reasoning we can conclude that neither $C \to F \bullet\!\!\to A$, nor $C \bullet\!\!\to F \to A$ can be in $\mathcal{G}$, for any node F, so orienting $A \to C$ also does not create an almost directed cycle.

Lastly, consider step (3). We first show that it can be performed, that is that $A \leftarrow D$ cannot occur for the mentioned configuration. Note that if we have $A \circ\!\!-\!\!\bullet B$ and $B \bullet\!\!\to D$ are in $\mathcal{G}$ we cannot also have $A \leftarrow D$ in $\mathcal{G}$ as that would imply that edge mark orientations in $\mathcal{G}$ are not complete with respect to R2. Hence, it is possible to orient edge $\langle A, D \rangle$ into D i.e., as $A \bullet\!\!\to D$, and by assumption, it is also possible to orient $B \leftrightarrow D$.

Now we only need to show that completing step (3) does not create an almost directed cycle. Orienting $B \circ\!\!\to D$ as $B \leftrightarrow D$ can only create an almost directed cycle if edge mark orientations in $\mathcal{G}$ are not complete under R8. Additionally, the only other way that completing step (3) could create an almost directed cycle, is if in completing step (3) we oriented $A \leftarrow\!\!\circ D$ as $A \leftrightarrow D$. But this type of an almost directed cycle would imply that $A \to F \to D$ was already in $\mathcal{G}$ for some F, which itself implies an almost directed cycle already exists in $\mathcal{G}$, which is a contradiction. □

Example C.7.34. Consider again the graphs in Figure C.10, where Figure C.10a represents the ancestral partial mixed graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ and Figure C.10b represents the partially oriented join trees for $\mathcal{G}$. From top to bottom, these join trees are $\mathcal{T}_1$, $\mathcal{T}_2$, and $\mathcal{T}_3$.

Suppose that we are interested in finding a MAG that contains a particular orien-

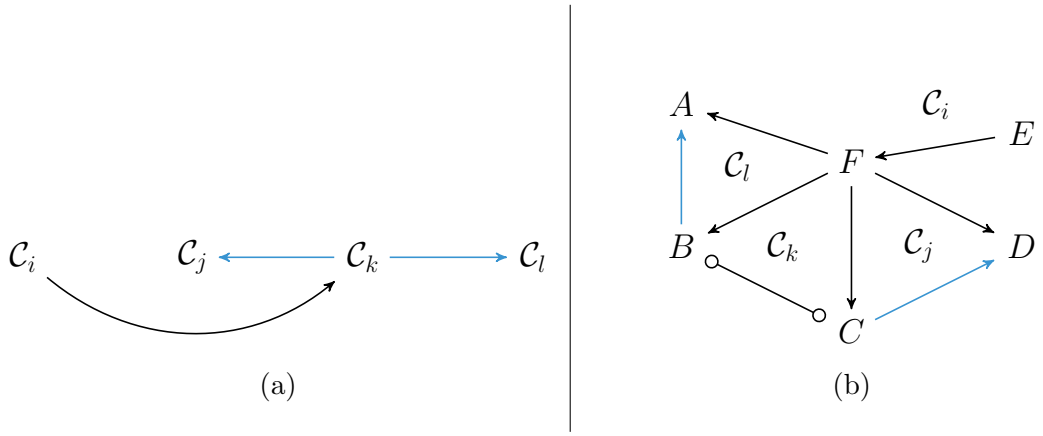


Figure C.11: Directed join tree used in Example C.7.34.

tation of edge $B \circ\!\!-\!\!\circ C$. Note that $\texttt{transformTree}(\mathcal{T}_1, \mathcal{C}_k)$ will return join tree $\mathcal{T}_3$, and so will $\texttt{transformTree}(\mathcal{T}_2, \mathcal{C}_k)$, and $\texttt{transformTree}(\mathcal{T}_3, \mathcal{C}_k)$. Then applying, for instance, $\texttt{orientTree}(\mathcal{T}_1, \mathcal{C}_k)$ (Algorithm 18) results in the directed join tree $\mathcal{T}$ in Figure C.11a. Let $\pi_{\mathcal{T}}$ be the partial ordering compatible with $\mathcal{T}$. Then $\pi_{\mathcal{T}}$ induces edge mark orientations in $\mathcal{G}$ as in Definition C.7.11 to create graph $\mathcal{G}_\pi$ in Figure C.11b. Now, we can use the result of Lemma C.7.33 to orient $B \circ\!\!-\!\!\circ C$ in $\mathcal{G}_\pi$ into any of the three options $B \to C$, $B \leftarrow C$, $B \leftrightarrow C$, thereby resulting in a valid MAG represented by $\mathcal{G}$ of Figure C.10a. □