Bidirectional insights between classical and quantum causal inference

by

Marina Maciel Ansanelli

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Physics

Waterloo, Ontario, Canada, 2026

© Marina Maciel Ansanelli 2026

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Roger Colbeck Professor, Dept. of Mathematics, King's College London
Supervisor(s):	Robert W. Spekkens Senior faculty, Perimeter Institute for Theoretical Physics Adjunct faculty, Dept. of Physics and Astronomy, University of Waterloo Kevin Resch Professor, Dept. of Physics and Astronomy, University of Waterloo
Internal Member:	Anton Burkov Professor, Dept. of Physics and Astronomy, University of Waterloo
Internal-External Member:	Doreen Fraser Professor, Dept. of Philosophy, University of Waterloo
Other Member(s):	Lucien Hardy Senior faculty, Perimeter Institute for Theoretical Physics Adjunct faculty, Dept. of Physics and Astronomy, University of Waterloo

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Chapters 2 and 3 are respectively taken from Refs. [3, 4], and were part of the same overarching project. The initial idea for this project came from Robert Spekkens and Elie Wolfe. Conceptual developments and technical proofs were developed by all authors. The code was developed by Elie Wolfe and me. Most of the writing was made by me, with comments and help from Robert Spekkens.

Chapter 4 is taken from Ref. [16]. The initial idea for this project came from Robin Evans, and the project was guided by Robin Evans and Elie Wolfe. The technical proofs were written mainly by Ryan Carey and me, with Ryan taking the lead in most of them and me contributing through detailed verification and many corrections. The writing was made equally by Ryan Carey and me.

Chapter 5 consists of original material, that will be turned into an individual publication in the near future. This project was proposed by Robert Spekkens for the PSI Winter School of 2022. Most of the conceptual progress for this project came from group discussions during the Winter School, where the supervisors were Robert Spekkens, Elie Wolfe, Thomas C. Fraser and myself, and the students were Sofia González-García, Matthew Fox and Manu Srivastava. The ideas for technical proofs were given by all authors, and their formalization as presented here was made by me. The code implementation was made by Elie Wolfe, Sofia González-García and me. The writing was made entirely by me.

Abstract

Causal inference is a sub-area of statistics that investigates the causal explanations underlying observed data. The last decade has seen growing recognition of its applications to quantum physics, which include the use of causal inference to certify quantumness through violations of Bell-like inequalities. This connection places quantum physicists in a unique position to contribute to classical causal inference by addressing open problems whose solutions benefit both fields. This thesis highlights this bidirectional interplay of fields, presenting both the study of foundational problems in classical causal inference from the perspective of quantum physics and the application of these results to quantum research.

An important problem in causal inference is attesting when two causal structures are in principle indistinguishable from data obtained under a given probing scheme on the visible variables. For example, passive observations on two variables cannot distinguish between a direct causal relation and a shared latent common cause. However, they *can* be distinguished by interventional probing schemes. The first question we address here is which causal structures remain indistinguishable even under the most informative interventional probing schemes. We derive a necessary and sufficient condition for such indistinguishability. We then consider the same question under access to passive observations only, compiling all known sufficient rules for indistinguishability and applying them to causal structures with three and four visible nodes, though a complete characterization remains open. Then, we extend these results to scenarios involving selection bias.

The works described above also help identify which causal structures impose nontrivial inequality constraints on their classically realizable distributions; an example is the Bell causal structure, that imposes Bell inequalities. This has particular interest for quantum physicists, because inequality constraints are a prerequisite for a causal structure to have a quantum-classical gap (QC gap), that is, for it to explain more distributions when its latent nodes are associated with quantum systems than when its latent nodes are associated with classical variables. This is important because, when we know that a phenomenon is governed by a causal structure that presents a QC gap, observing a probability distribution over the observed variables that cannot be explained using classical latent variables in that structure certifies that the phenomenon involves genuine nonclassicality. Thus, after identifying which causal structures could potentially exhibit a QC gap, we investigate which actually do. Using techniques developed here, we show that hundreds of new causal structures have a QC gap, while previously only around a dozen examples were known.

Acknowledgements

First of all, I would like to thank my supervisor, Rob Spekkens. Rob, it has been an immense privilege to work with you. You are not only a deep thinker in physics, but also thoughtful about how to cultivate an environment where everybody feels welcomed and encouraged to think critically and ask questions. Hearing you talk about physics has been one of the most interesting scientific experiences I've had, and I'm deeply grateful to you for teaching me to always strive to think and communicate as clearly as possible. I also want to thank you for believing in me even in moments when I didn't.

I am also deeply grateful to Elie Wolfe, who was in practice a co-supervisor during the first years of my PhD, and a cherished collaborator throughout it. Elie, your enthusiasm is contagious, and I think we work together very well. Thank you for always hyping up students' ideas and contributions, and being so available and generous with your time.

I am grateful to Márcia and Donato for showing me, since the beginning, how pursuing knowledge can be fun and fulfilling. Mom, thank you for telling me real stories from history whenever you couldn't dream up new fairy tales. Dad, thank you for taking me to more museums and exhibitions than I can count, and always being excited to show me something new. I am also grateful for the love and support from all of the rest of my family.

I am grateful to all of the teachers and mentors I had in my life, but I would like to highlight a few. Fuscella, Arderucio and Hugo, who led me to the lightbulb moment when I realized I needed to be a physicist. Ravi and Vicky, who led me to the lightbulb moment when I realized I needed to work in quantum foundations. I am also grateful to all my past and present collaborators, who have so often been my teachers, as well as to the many amazing researchers I have had the opportunity to interact with.

I am grateful to the quantum foundations group of Perimeter, which is the best group ever in objective terms.[1] Each and every one of you makes the environment of our group healthy, collaborative, and special. Among you, I would especially like to thank Yìlè Yīng. We are a great duo, and I hope we work together forever. And I will be very persistent in scheduling video calls to also talk about life (and continue loitering, but now online).

There are multiple other people who supported me throughout the PhD journey, to whom I am very grateful. To Mayumi and Ana, for bringing home closer, keeping our connection from childhood and even coming to visit me in Canada. Mayumi, thank you for the regular updates and gossip about your life, and Ana, thank you (?) for remaining mysterious and never sharing any gossip about yours. To Luiza, for our friendship growing

[1] In my opinion.

even stronger since I moved away (and for the day we made the two of us on The Sims). To Sruthi and Viola, for giving me community in Canada, and showing me that the creative wonders and struggles of a scientist and an artist are not so different. To Davide, for the ice skating and puppetry. To Carol and Bruna, for all the dance. To my therapist, Angélica, for the help in all of these years. To the staff of Perimeter Institute, for working hard to make PI an exceptional place to do science. To my cat Chinelo, for being a silent companion who grounded me during many moments of PhD desperation. To the many other friends and connections I have made along the way.

To arXiv and Sci-Hub, and to all who work for knowledge to be free.

To my favorite person, Léo. Over the past ten years you constantly showed me the extent of what love can be, and I am eternally grateful for it. During the PhD, your care was my home. I look forward to see what the future holds for us in this next chapter, and I am sure we will continue growing as a team. You are my oasis.

The chapters of this thesis come from works that were written with colleagues. Apart from my coauthors, these chapters have the following people to thank: Pedro Lauand, for sharing with us a technique that was crucial for the proof of the main result of Chapter 2. Robin Evans, for discussions about his work during the early stages of the project that led to Chapter 3. And an anonymous reviewer, for finding mistakes in two of the proofs of Chapter 3, as well as presenting a counter-example to a conjecture from a previous version of our manuscript.

Finally, I would like to thank whoever made the lo-fi playlist that I always listen to on Spotify. If not for you, I wouldn't have been able to get any work done.

Dedication

To my families,

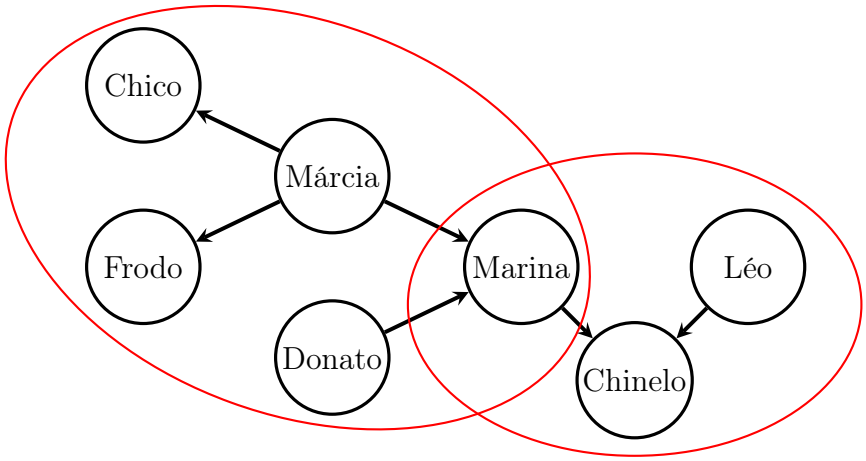


Table of Contents

Examining Committee Membership ii

Author's Declaration iii

Statement of Contributions iv

Abstract v

Acknowledgements vi

Dedication viii

List of Figures xiv

List of Tables xxiii

1 Introduction 1

1.1 Classical causal inference 2

1.2 Causal inference and quantum theory 5

2 Interventional distinguishability of causal structures in the presence of latent nodes 11

2.1 Preliminaries 15

2.1.1 Explaining Observed Data with a Causal Structure 16

2.1.2 Observational Equivalence and Dominance . . . 19
2.1.3 mDAGs . . . 20
2.1.4 Restriction of scope to sets of mDAGs consistent with a fixed nodal ordering . . . 26
2.1.5 3-pDAGs . . . 28
2.2 Full-SWIGs . . . 31
2.3 Probing Schemes and Shadows . . . 36
2.4 Equivalence and Dominance of causal structures under the Observe&Do Probing scheme . . . 39
2.5 Equivalence and Dominance under all patterns of do-interventions . . . 43
2.5.1 One-Value do-interventions . . . 51
2.6 Structural Partial Order for mDAGs with three or four visible nodes . . . 52
2.7 Conclusion . . . 55

3 Observational distinguishability of causal structures in the presence of latent nodes 62
3.1 Preliminaries . . . 67
3.2 Objectives and Methods . . . 69
3.2.1 Special Cases: Latent-confounder-free and Directed-edge-free mDAGs 69
3.2.2 Full Solution of Observational Partial Order for 3-node mDAGs . . . 71
3.2.3 Significance of observational partial order for causal discovery . . . 72
3.2.4 How to Express Partial Information about the Observational Equivalence Partition . . . 79
3.3 Rules for Establishing Observational Dominance . . . 81
3.3.1 HLP Edge-Adding Rule . . . 81
3.3.2 Facet-Adding Rules . . . 83
3.4 Rules for Establishing Observational Nondominance . . . 90
3.4.1 Skeletons . . . 91
3.4.2 d-separation relations . . . 92

3.4.3 e-separation relations . . . 93
3.4.4 Pairs of Densely Connected Nodes . . . 95
3.4.5 Directed-edge-free rule . . . 98
3.4.6 Comparison of unrealizable supports . . . 101
3.5 Deriving the Observational Partial Order for 3-node mDAGs . . . 105
3.5.1 Observational Dominances . . . 105
3.5.2 Observational Nondominances . . . 107
3.6 Partial Solution of Observational Partial Order for 4-node mDAGs . . . 112
3.6.1 Completely Identified Classes . . . 115
3.6.2 Importance of Inequality Constraints for Causal Discovery . . . 117
3.6.3 mDAGs that can Realize Every Support on Binary Variables . . . 120
3.7 Conclusion . . . 123

4 Interventional and observational distinguishability of causal structures in the presence of latent and selected nodes 127
4.1 Preliminaries . . . 129
4.1.1 Observe&Do in the presence of selection . . . 131
4.2 Transformations that maintain Observe&Do equivalence in the presence of selection . . . 134
4.3 Selected-marginalized directed graphs (smDGs) . . . 140
4.4 Interventional inequivalence of smDGs . . . 144
4.4.1 Distinct smDGs are Observe&Do distinguishable . . . 144
4.4.2 Distinct smDGs are not always All-patterns Observe-Or-Do distinguishable . . . 145
4.5 Rules to prove observational equivalence of smDGs . . . 146
4.5.1 Adding marginalized faces . . . 147
4.5.2 Removing special edges . . . 148
4.5.3 Removing selected faces . . . 149
4.5.4 Leveraging mDAG equivalence for smDG equivalence . . . 151
4.6 Conclusion . . . 153

5 Finding new causal structures that present quantum-classical gaps **155**

5.1 Examples of application of the piggybacks . . . 160

5.1.1 The Point Distribution Piggyback: Bell-with-mediary Scenario . . . 160

5.1.2 The Marginalization Piggyback . . . 161

5.1.3 The Fritz Piggyback: Triangle Scenario . . . 167

5.1.4 The Interruption Piggyback: Ghost Scenario . . . 171

5.1.5 The Postselection Piggyback: Flag Scenario . . . 173

5.2 **Formalizati**on of the piggybacks . . . 176

5.3 Observational dominance in the quantum case . . . 178

5.4 New QC gaps in causal structures with four visible nodes . . . 182

5.5 Conclusion . . . 185

References **188**

APPENDICES **196**

A Appendices for Chapter 2 **197**

A.1 Proof of Lemma 6 . . . 197

A.2 Proof of the Case 2 of Theorem 2 . . . 199

B Appendices for Chapter 3 **204**

B.1 Proof that latent-confounder-free pDAGs consistent with a fixed nodal ordering are observationally inequivalent . . . 204

B.2 Proof of Strong Facet-Adding . . . 206

B.3 Fraser's Algorithm for Feasible Supports . . . 210

B.4 Proofs that comparison of unrealizable supports subsumes graphical methods as rules to show observational nondominance . . . 213

C Appendices for Chapter 4 **217**

C.1 Proofs that the transformations of Section 4.2 maintain Observe&Do equivalence . . . 217

C.1.1 Proof of Proposition 15 . . . 217

C.1.2 Proof of Proposition 16 . . . 221

C.1.3 Proof of Proposition 17 . . . 226

C.1.4 Proof of Proposition 18 . . . 229

C.1.5 Proof of Proposition 19 . . . 231

C.2 Proof that different SLP implies interventional inequivalence (Theorem 5) . 233

C.3 A weaker probing scheme: Observe-or-do . . . 241

C.4 Proofs of rules to show observational equivalence . . . 243

C.4.1 Proofs for section 4.5.1 . . . 243

C.4.2 Proofs for Section 4.5.2 . . . 244

C.4.3 Proofs for section 4.5.3 . . . 246

D Appendices for Chapter 5 **253**

D.1 Proofs of Piggybacks . . . 253

D.2 Proofs of observational dominance rules in the quantum case . . . 257

List of Figures

1.2.1 The Bell causal structure. 6

1.2.2 (a) Causal structure for a Bell experiment that has superluminal causation. (b) Causal structure for a Bell experiment that has superdeterminism. (c) Causal structure for a Bell experiment that has retrocausality. 6

2.1.1 An example of pDAGs that are observationally equivalent. The first of this triple is generally known as the *instrumental scenario*. 20

2.1.2 (a) A pDAG. (b) The pDAG obtained from (a) by exogenizing the latent node α. (c) The pDAG obtained from that of (b) by removing the latent node α, which is redundant to β since the children of α are a subset of the children of β. By Lemmas 1 and 2, these three pDAGs are observationally equivalent. The pDAG in (c) is RE-reduced. 21

2.1.3 (a) A RE-reduced pDAG. (b) The mDAG associated with (a). This mDAG has simplicial complex $\mathcal{B} = \{\{a\}, \{b\}, \{c\}, \{d\}, \{a, b\}, \{a, d\}, \{b, d\}, \{a, b, d\}\}$. This is the mDAG associated with all of the pDAGs of Fig. 2.1.2. The inclusion-maximal element of $\mathcal{B}$ are indicated by the red loop on the mDAG. 24

2.1.4 (a) mDAG with trivial directed structure and with simplicial complex $\mathcal{B} = \{\{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}\}$. (b) mDAG with trivial directed structure and with simplicial complex $\mathcal{B}' = \{\{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}\}$.The facets (inclusion-maximal elements of $\mathcal{B}$ and $\mathcal{B}'$) are indicated by red loops. 25

2.1.5 (a) Canonical pDAG of Fig. 2.1.4(a). (b) Canonical pDAG of Fig. 2.1.4(b). (c) pDAG obtained by adding the latent nodes of (a) to (b). This is not a canonical pDAG. By Lemma 2, (b) and (c) are observationally equivalent. . 25

2.1.6 (a) Causal structure of the "fork" type that is consistent with the nodal ordering (a, b, c). (b) Causal structure of the "chain" type that is consistent with the nodal ordering (a, b, c). These two causal structures are *not* observationally equivalent. 28

2.2.1 (a) The pDAG `Cause-Effect&Confounder`. (b) The SWIG that represents the possibility of intervening on the node a of `Cause-Effect&Confounder`. 32

2.2.2 Example of the operation of splitting all visible nodes of a pDAG $\mathcal{G}$ to obtain the corresponding full-SWIG $\mathscr{G} = \texttt{split}(\mathcal{G})$. 33

2.2.3 Example of the operation of splitting all nodes of an mDAG $\mathfrak{G}$ to obtain the corresponding 3-mDAG $\mathbb{G} = \texttt{split}(\mathfrak{G})$. 33

2.3.1 Example used to understand the notion of shadow of the full conditional distribution $P(X_{\texttt{vis}(\mathcal{G})^\flat} | X_{\texttt{vis}(\mathcal{G})^\sharp})$ that is revealed by a specific probing scheme. 37

2.4.1 Structural partial order for the set of two-node mDAGs that are consistent with a fixed nodal ordering. This also expresses the partial order of equivalence classes under the O&D relation. The background colour of each mDAG indicates its observational equivalence class. 43

2.4.2 All two-node mDAGs consistent with a fixed nodal ordering, organized into the partial order of equivalence classes under the observational dominance relation. 44

2.5.1 mDAGs that appear in Fig. 16 of Ref [33]. 46

2.6.1 (a) Structural partial order for the set of two-node counfounder-free mDAGs that are consistent with a fixed nodal ordering. (b) Structural partial order that holds between all of the two-node directed-edge-free mDAGs. Note that a directed-edge-free mDAG is always consistent with any nodal ordering. 53

2.6.2 Compressed way of expressing the structural partial order that holds between two-node mDAGs. The decompressed structural partial order was given in Fig. 2.4.1. 54

2.6.3 (a) Structural partial order for the set of three-node confounder-free mDAGs that are consistent with a fixed nodal ordering. (b) Structural partial order for the set of three-node directed-edge-free mDAGs. Note that a directed-edge-free mDAG is always consistent with any nodal ordering. 55

2.6.4 Structural partial order for the set of three-node mDAGs consistent with a fixed nodal ordering. The line between a pair of large boxes (islands) is to be interpreted as every mDAG within the higher island dominating the *corresponding* mDAG within the lower island, i.e., the mDAG that has the same directed-edge structure. Two mDAGs of this picture have the same background colour and the same number if they are *observationally* equivalent. 56

2.6.5 Structural partial for the set of four-node confounder-free mDAGs consistent with a fixed nodal ordering. 57

2.6.6 Structural partial order that holds between all four-node directed-edge-free mDAGs. Next to each mDAG, we indicate the set of facets of its simplicial complex. Note that a directed-edge-free mDAG is always consistent with any nodal ordering. 58

2.7.1 Partial order of equivalence classes of the different probing schemes studied in this work, where the relation is inclusion of the image of the shadowing function of each probing scheme. The informationally complete probing schemes, that reveal the entirety of this conditional distribution $P(X^{\flat}_{\mathtt{vis}(\mathcal{G})}|X^{\sharp}_{\mathtt{vis}(\mathcal{G})})$, form an equivalence class. In the O&D Probing Scheme, for each visible variable one can first observe its natural value and then force it to take any desired value. In the all-patterns O-or-D probing scheme, for each visible variable one can either passively observe it or force it to take any desired value, without previously observing the natural value of the variable. In the all-patterns O-or-1D probing scheme, for each visible variable one can either passively observe it or force it to take *one* fixed value, without previously observing the natural value of the variable. 60

3.2.1 The 72 mDAGs with three visible nodes with nodal ordering (a, b, c) organized into observational equivalence classes and the observational dominance relations that hold between these classes. In each mDAG of this figure, the bottom node is a, the top-left node is b and the top-right node is c. 73

3.2.2 The proven-equivalence partition (pink loops) is found by applying the equivalence rules of Section 3.3. The proven-inequivalence partition (gray loops) is found by the inequivalence rules of Section 3.4. The final goal is to find the observational equivalence partition (green dashed loops). 80

3.3.1 Classification of the different dominance-proving rules that we use. Classification of the different nondominance-proving rules. An arrow from one rule to another indicates that the latter is strictly less powerful than the former (i.e., the latter is subsumed in the former). The rules indicated in thick green rectangles are original results of this chapter. 82

3.3.2 Two mDAGs that can be proven to be observationally equivalent by the HLP Edge-Adding Rule [47]. 82

3.3.3 Two mDAGs that can be proven observationally equivalent by the Weak Facet-Adding rule. Here, the sets B, C and D from the statement of Proposition 5 are respectively $B = \{a, b, d, e\}$, $C = \{a, b\}$ and $D = \{d, e\}$. 84

3.3.4 The first and third mDAG of this figure can be shown observationally equivalent by Evans' rule, as well as by the HLP Edge-Adding rule together with the Weak Facet-Adding rule. This exemplifies Lemma 9. 85

3.3.5 Two mDAGs that can be proven observationally equivalent by Moderate Facet-Adding, but not by Weak Facet-Adding. Here, the sets B, C and D from the statement of Proposition 5 are respectively $B = \{a, b, d, e\}$, $C = \{a, b\}$ and $D = \{d, e\}$. 86

3.3.6 mDAG obtained by starting from Fig. 3.3.5(a) and performing the steps i) and ii) described in the proof of Proposition 7. 87

3.3.7 Two mDAGs that can be shown observationally equivalent by the Strong Facet-Adding rule. 88

3.3.8 Example of the sequence of mDAGs that appears in Conjecture 2. One can see that $\mathfrak{G}_2$ and $\mathfrak{G}_3$ are observationally equivalent by applying the HLP Edge-Adding rule twice: first to remove the arrow $c \to d$ and second to add the arrow $d \to c$. Conjecture 2 says that every strict observational dominance can be obtained by a sequence like this one. 90

3.4.1 Classification of the different nondominance-proving rules. An arrow from one rule to another indicates that the latter is strictly less powerful than the former (i.e., the latter is subsumed in the former), and the Lemma that proves it is indicated on the arrow. We defined a shorthand name for each rule, shown inside the parentheses. 91

3.4.2 (a) An mDAG; and (b) Its skeleton. 92

3.4.3 Example of mDAGs that have the same skeleton, but different sets of d-separation relations. The mDAG (a) presents the d-separation relation $b \perp\!\!\!\perp_{\mathrm{d}} c$, while the mDAG (b) does not. In Fig. 3.2.1, the mDAG (a) is in the `Collider A` observational equivalence class, while (b) is in the `Evans` observational equivalence class. 93

3.4.4 Example of mDAGs that have the same set of d-separation relations (i.e., no d-separation relations), but different skeletons. In Fig. 3.2.1, the mDAG (a) is in the `Triangle` observational equivalence class, while (b) is in the `Evans` observational equivalence class. 93

3.4.5 The districts of this mDAG are $\{a, b, c, h\}$, $\{d\}$ and $\{e, f, g\}$. 96

3.4.6 Steps used to find the closure of the set $A = \{b, c, g\}$ in the mDAG of Fig. 3.4.5. 97

3.4.7 Example of a pair of mDAGs that have the same pairs of densely connected nodes, but different sets of e-separation relations (which in this case are simply d-separation relations). In Fig. 3.2.1, the mDAG (a) is in the `Collider A` observational equivalence class, while (b) is in the `Fork` observational equivalence class. 99

3.4.8 Example of a pair of mDAGs that have the same sets of e-separation relations, but different pairs of densely connected nodes. The e-separation relations of both of these mDAGs are $a \perp\!\!\!\perp_{\mathrm{d}} c|\neg b$, $a \perp\!\!\!\perp_{\mathrm{d}} c|\neg bd$, $a \perp\!\!\!\perp_{\mathrm{d}} c|b$, $a \perp\!\!\!\perp_{\mathrm{d}} c|b\neg d$, $a \perp\!\!\!\perp_{\mathrm{d}} c|d\neg b$, $a \perp\!\!\!\perp_{\mathrm{d}} d|\neg b$, $a \perp\!\!\!\perp_{\mathrm{d}} d|\neg bc$, $a \perp\!\!\!\perp_{\mathrm{d}} d|\neg c$ and $a \perp\!\!\!\perp_{\mathrm{d}} d|c\neg b$. Meanwhile, the pair of nodes (a, d) is densely connected in (b) but not in (a). 99

3.4.9 The directed-edge-free mDAGs are represented by blue dots, while the other mDAGs are red dots. Suppose that nondominance-proving rules other than the directed-edge-free rule have been used to obtain the proven-inequivalence partition depicted. The directed-edge-free rule says that the proven-inequivalence block on the left should split into the four proven-equivalence blocks that live inside it. On the other hand, the proven-inequivalence block on the right includes one distinguished proven-equivalence block that does not contain any directed-edge-free mDAG. Therefore, the directed-edge-free rule cannot be used to split the proven-inequivalence block on the right. 100

3.5.1 Three observationally equivalent mDAGs, that lie in the `Instrumental ABC` class. 105

3.5.2 Intermediary steps to show observational equivalence between the second and the third mDAGs of Fig. 3.5.1. 106

3.5.3 Representation of the information obtained by applying the comparison of skeletons to the proven-equivalence partition of 3-node mDAGs. In this figure, the loops represent blocks of the proven-inequivalence partition and the cut arrows ($\not\to$) represent proven nondominances. Double-headed cut arrows mean that there is proven nondominance in both directions, i.e., proven incomparability. Compare this to the set of observational *dominances* proven by Structural Dominance, the HLP Edge-Adding rule and Weak Facet-Adding, reproduced in the miniature on the top right. 108

3.5.4 Representation of the information obtained by applying the comparison of skeletons and the comparison of d-separation relations to the proven-equivalence partition of 3-node mDAGs. Here, we only show the blocks of the proven-equivalence partition that were not already shown to be observational equivalence classes by the comparison of skeletons alone. The loops represent blocks of the proven-inequivalence partition and the cut arrows ($\not\to$) represent proven nondominances. Double-headed cut arrows mean that there is proven nondominance on both directions, i.e., proven incomparability. Compare this to the set of observational *dominances* among these mDAGs proven by Structural Dominance, the HLP Edge-Adding rule and Weak Facet-Adding, reproduced in the miniature on the top right. 109

3.6.1 (a) Bell mDAG. (b) `Square` mDAG. Both of these belong to equivalence classes that are completely identified by d-separation, meaning that there is no mDAG that is observationally inequivalent to them while still imposing the same conditional independence constraints on the realizable distributions. 117

3.6.2 Seven mDAGs that do not appear in the same block as the saturated mDAGs in the proven-equivalence partition induced by the rules described in Section 3.3, but that nevertheless can realize any support on binary variables. mDAGs (a)-(d) are shown to *not* be saturated via analysis of supports on variables of higher cardinality. It is still unknown whether mDAGs (e)-(g) are observationally inequivalent to the saturated mDAGs. 121

3.6.3 Diagrammatic representation of the proven-equivalence and proven-inequivalence partitions of 4-node mDAGs consistent with a fixed nodal ordering that are induced by the rules indicated in the labels of the figure. The saturated proven-equivalence block includes 2134 mDAGs consistent with a fixed nodal ordering. The letters (a)-(g) in this diagram refer to the corresponding mDAGs of Fig. 3.6.2. 122

4.0.1 Two selected S-pDAGs that are interventionally distinguishable but correspond to the same MAG. 128

4.0.2 Can these S-pDAGs produce different data on a, b, c, d after marginalization of the m_i nodes and selection of the s_i nodes? Spoiler: as we will see, they can be distinguished by interventions but not by passive observations. . . . 129

4.2.1 Example of the application of Proposition 16. 136

4.2.2 First example of the application of Proposition 17. 138

4.2.3 Second example of the application of Proposition 17. 138

4.2.4 Example of the application of Proposition 19. 140

4.3.1 Under marginalization of m_1, m_2, m_3, m_4, m_5 and selection of s_1, s_2, s_3, s_4, we obtain the selected-latent projection shown in (b). 142

4.4.1 Two canonical S-pDAGs whose selected-latent projections differ in the self-loop $v \to v$. 146

4.5.1 Two smDGs that can be shown observationally equivalent by Proposition 21. 148

4.5.2 An smDG that can be shown observationally equivalent to Figure 4.5.1b by Proposition 22 (and Corollary 1). 149

4.5.3 An smDG that can be shown observationally equivalent to Figure 4.5.2 by Proposition 23. 151

4.5.4 Two smDGs that can be shown observationally equivalent by Proposition 23, for $V_s = \{c, d\}$. This is an example where Proposition 23 can be applied even though $\mathcal{G}_{\mathbf{an}_{\mathcal{G}}(V_s)}$ is not saturated. 151

4.5.5 An smDG that can be shown to be observationally equivalent to Figure 4.5.3 by Proposition 24 together with Proposition 5 of [33]. That is, we treat all of the selected nodes of the canonical S-pDAG as visible, apply Proposition 5 and then take the selected-latent projection. 152

5.0.1 (a) Bell pDAG with an unspecified semantics. (b) Bell pDAG with a classical semantics. (c) Bell pDAG with a quantum semantics. 157

5.0.2 Two pDAGs that correspond to the same mDAG, and thus are Observe&Do equivalent in the classical case, but that are observationally *inequivalent* in the quantum case. Example taken from Ref. [81]. 158

5.1.1 The Bell-with-mediary pDAG $\mathcal{G}_{\text{BellMed}}$. The nodes x, y, s, t and m are visible, and the node α is latent. 161

5.1.2 The Bell-with-broken-leg pDAG $\mathcal{G}_{\text{BrokenLeg}}$. 162
5.1.3 The pDAG $\mathcal{G}_{\text{EqToBell1}}$, obtained from $\mathcal{G}_{\text{BrokenLeg}}$ by turning a into a classical latent node. 163
5.1.4 The Charlie Brown Tree pDAG $\mathcal{G}_{\text{Charlie}}$. Here we look at semantics where μ is classical, because those are enough to show the QC gap. 164
5.1.5 The Instrumental pDAG $\mathcal{G}_{\text{Instr}}$. 164
5.1.6 The pDAG $\mathcal{G}_{\text{EqToInstr1}}$, obtained from $\mathcal{G}_{\text{Charlie}}$ by turning d into a classical latent node. 165
5.1.7 The Pike pDAG $\mathcal{G}_{\text{Pike}}$. 166
5.1.8 The pDAG $\mathcal{G}_{\text{EqToInstr2}}$, obtained from $\mathcal{G}_{\text{Pike}}$ by turning d into a classical latent node. 166
5.1.9 The Triangle pDAG $\mathcal{G}_{\text{Tri}}$. The nodes a, b and c are visible, and the nodes α, β and γ are latent. 167
5.1.10 DAG $\mathcal{G}_{\text{SplitTri}}$ obtained by splitting each visible variable of the Triangle pDAG into two variables of smaller cardinality. 168
5.1.11 DAG $\mathcal{G}_{\text{DropSplitTri}}$ obtained by dropping the arrows $\alpha \to a_2$, $\gamma \to c_1$, $\alpha \to b_1$ and $\beta \to c_2$ from $\mathcal{G}_{\text{SplitTri}}$. This pDAG can classically realize all of the distributions of $\mathcal{S}_C$. 169
5.1.12 pDAG $\mathcal{G}_{\text{EqToBell2}}$ obtained by deleting b_1 and a_2 from $\mathcal{G}_{\text{DropSplitTri}}$. 170
5.1.13 The Ghost pDAG $\mathcal{G}_{\text{Ghost}}$. 171
5.1.14 The Evans pDAG $\mathcal{G}_{\text{Evans}}$. 172
5.1.15 The Flag pDAG $\mathcal{G}_{\text{Flag}}$. 173
5.1.16 The pDAG $\mathcal{G}_{\text{EqToInstr3}}$, obtained from $\mathcal{G}_{\text{Flag}}$ by adding a latent parent of the parents of d and then deleting d. 174
5.4.1 Before this work, these were all of the 4-node mDAGs (up to relabelling of the nodes) that were known to have a QC gap. Now, the number of 4-node mDAGs (up to relabelling of the nodes) that are known to have a QC gap is 926. 185

A.1.1 Commutative diagram corresponding to the statement of Lemma 6. 197
A.1.2 Exogenizing a latent node commutes with the split-node map. 198

A.1.3Removing a redundant latent node commutes with the split-node map. . . 199

B.2.1By Lemma 22, we have $\mathcal{G}_2 \succeq \mathcal{G}_1$. Here, $C_1 = \{t, u\}$, $C_2 = \{u, v\}$, and $D = \{d\}$. Note that $\mathcal{G}_1$ does *not* observationally dominate $\mathcal{G}_2$: for example, $\mathcal{G}_1$ has the d-separation relation $x \perp\!\!\!\perp_{\mathrm{d}} d$, that is not presented by $\mathcal{G}_2$. Edges that point into a latent node are represented in red. 207

B.2.2The pDAGs $\mathcal{G}_1$ and $\mathcal{G}_3$ can be shown observationally equivalent by Proposition 30: Lemma 22 says that $\mathcal{G}_2 \succeq \mathcal{G}_1$, and Exogenization (Lemma 1) says that $\mathcal{G}_2 \cong \mathcal{G}_3$. We can see that $\mathcal{G}_1 \succeq \mathcal{G}_3$ by structural dominance (Lemma 3). Edges that point into a latent node are represented in red. 211

C.2.1(a) Canonical S-pDAG of the subgraph of $\mathcal{G}'$ over a and b. (b) Graph that represents the data realized by (a) after an intervention on a (when we omit the natural value $X_{a^\flat}$). (c) Graph that represents the data realized by (a) after an intervention on b (when we omit the natural value $X_{b^\flat}$). 236

D.2.1DAGs used in to exemplify the proof of Lemma 18. 259

List of Tables

3.1 Number of blocks of the proven-equivalence partition for 3-node mDAGs consistent with a fixed nodal ordering obtained by applying various combinations of the dominance-proving rules of Section 3.3. The row "None" gives the total number of 3-node mDAGs consistent with a fixed nodal ordering. These numbers give upper bounds on the number of observational equivalence classes. 112

3.2 Number of blocks of the proven-inequivalence partition for 3-node mDAGs consistent with a fixed nodal ordering obtained by applying various combinations of the nondominance-proving rules of Section 3.4. These numbers give lower bounds on the final number of equivalence classes. The fact that the last line reaches 15, which is the upper bound obtained from the proven-equivalence partition, shows that we have identified the observational equivalence partition for 3-node mDAGs consistent with a fixed nodal ordering. 113

3.3 Cardinalities of the proven-equivalence partitions induced by various combinations of the dominance-proving rules of Section 3.3 for the set of 4-node mDAGs consistent with a fixed nodal ordering. The row "None" gives the total number of 4-node mDAGs consistent with a fixed nodal ordering. These cardinalities give successively tighter upper bounds on the number of observational equivalence classes. 114

3.4 Cardinalities of the proven-inequivalence partitions induced by various combinations of the nondominance-proving rules of Section 3.4 for the set of 4-node mDAGs consistent with a fixed nodal ordering. These cardinalities give successively tighter lower bounds on the number of observational equivalence classes. All the support comparison applied assumed all the visible variables to be binary, i.e., $\vec{c}_{\text{vis}} = (2, 2, 2, 2)$. 114

3.5 Number of identified classes of 4-node mDAGs consistent with a fixed nodal ordering obtained by applying various combinations of the nondominance-proving rules of Section 3.4. These are sets of mDAGs that are both blocks of the proven-equivalence partition and of the proven-inequivalence partition. 115

5.1 Results obtained from applying the techniques developed in this chapter to 4-node mDAGs consistent with a fixed nodal ordering. The first column indicates the piggybacks used: PD is the point distribution piggyback (Proposition 25), Interr is the interruption piggyback (Proposition 28), Marg is the marginalization piggyback (Proposition 26), Post is the postselection piggyback (Proposition 29) and Fritz is the Fritz piggyback (Proposition 27). The second column indicates the number of 4-node mDAGs consistent with a fixed nodal ordering that are known to have a QC gap up to that point, and the third column indicates the number of remaining 4-node mDAGs consistent with a fixed nodal ordering that were not proven to be algebraic. The sum of the two numbers on a line is always 2807, which is the total number of 4-node mDAGs consistent with a fixed nodal ordering that are not proven to be algebraic. 184

Chapter 1

Introduction

Many times in science our interest does not lie in the *correlations* between events under study, but instead in their *causal* connections. For example, for a policy maker it is more important to know whether a certain educational program *causally affects* the career success of participants, than to simply know that there is a correlation between participating in the educational program and succeeding in one's career. Methods that are traditionally used in science, however, do not adequately treat causation: for example, when writing that the atmospheric pressure P is related to the reading of a barometer B by the equation $P = kB$ (where k is some constant), it is not clear if the pressure P causes the reading B, if the reading B causes the pressure P or if none holds and this correlation appears due to a hidden common cause[1]. In this case it is easy to conclude that the reading of the barometer is caused by the atmospheric pressure, but this information is not contained in the equation. Therefore, in more complicated scenarios, employing traditional methods based on correlations can induce one to error. The area of causal inference was born as a sub-area of statistics to fill this gap, answering questions about the underlying causal explanations for the observed statistical correlations.

In recent years, it has been shown that causal inference has many applications to quantum theory research. In particular, Bell's theorem can be understood as a challenge to our classical notions of causality [83]. To causally explain the correlations obtained from Bell experiments while still respecting intuitions that come from special relativity, generalizations of the framework of causal inference have been developed, culminating in the area of quantum causality.

[1] Example taken from Ref. [58], which is a science communication book on causal inference.

This thesis broadly has two themes: first, our intuition and motivations from quantum research lead us to ask fundamental questions about classical causal inference. Second, the insights and results we gain from solving these problems in the classical case are a step in developing new applications of causal inference to quantum foundations. Chapters 2 to 4 more naturally align with the first theme, while Chapter 5 aligns with the second theme. In the rest of this introduction, we will describe the topics of this thesis in the following way: in Section 1.1, we will summarize the main results and consequences of Chapters 2 to 4 under the lens of classical causal inference. In Section 1.2, we will talk about the relevance of causal inference to quantum foundations, connect it back to a quantum physicists' motivation to be interested in Chapters 2 to 4, and summarize the main results and consequences of Chapter 5.

1.1 Classical causal inference

In a causal discovery scenario, we want to adjudicate between candidate causal explanations for our data, appealing to different causal structures. A causal structure is represented by a directed acyclic graph (DAG), whose nodes represent variables of interest and whose arrows represent direct causal influences. These nodes could either be visible, i.e., associated with variables that appear in our data, or latent, i.e., associated with variables that are important for the causal explanation but that do not appear in the data. An example of a causal discovery scenario is a drug trial: is the correct causal structure one where the treatment causes the recovery, or one where the observed correlation comes entirely from a latent common cause (such as the health awareness of the subjects)?[2] These two causal structures are *observationally equivalent*: all of the probability distributions over visible variables that can be explained by one causal structure can also be explained by the other, which makes it impossible to distinguish them by passive observations alone. However, it is possible to settle the question by performing an *interventional* experiment: if we perform a randomized controlled trial where subjects are randomly assigned to the treatment or to a placebo, we can find out whether there is a direct causal connection between treatment and recovery.

As it turns out, however, there are causal structures that cannot be distinguished even by probing the visible variables with arbitrary interventions. In Ref. [3], that is presented in Chapter 2, we give a **necessary and sufficient graphical condition for two classical causal structures to be indistinguishable even when there is access to the**

[2] The presence of latent common causes is sometimes referred to as *latent confounding*.

strongest possible interventional probing scheme on the visible nodes. The answer is given by a graphical structure that was previously defined [33], the *marginalized DAG* (mDAG). That is, two causal structures that can involve latent confounding are distinguishable by a maximally informative probing scheme if and only if they are associated with different mDAGs. This implies that the most you can learn about the underlying causal structure governing a phenomenon via interaction with the visible variables is the mDAG to which the causal structure is associated. This work thus establishes the mDAG as a fundamental object in the causal discovery problem with latent variables.

Having understood the fundamental limits of causal discovery under access to arbitrary interventions, a next natural question is how much information about the causal structure can we still learn when there is only access to passive observations. This is because sometimes it is impossible or unethical to perform an interventional experiment; for instance, we cannot force a group of people to smoke to see if they will develop lung cancer. Our work in Ref. [4], that is presented in Chapter 3, **makes progress towards the goal of characterizing when two causal structures that correspond to different mDAGs cannot be distinguished by passive observations**, that is, when they are observationally equivalent. More specifically, for causal structures with three and four visible nodes, we explicitly construct a partial order over these where a causal structure $\mathcal{G}$ is dominated by another causal structure $\mathcal{G}'$ in the partial order if all of the probability distributions over visible variables that can be explained by $\mathcal{G}$ can also be explained by $\mathcal{G}'$. This work offers a foundational step toward a more complete understanding of the landscape of causal structures with latent variables, although further progress in this direction is still required for a complete solution: all of the existing results (including the new ones we introduced) are still not enough to provide a necessary and sufficient graphical condition for observational equivalence.

If a probability distribution over the visible variables obtained by passive observations can be explained by a causal structure $\mathcal{G}$, then it has to satisfy a set of constraints imposed by $\mathcal{G}$, called *causal-compatibility constraints*. For example, every probability distribution that can be explained by the DAG $a \to b \to c$ has to be such that the variable associated with a is independent of the variable associated with c when we know the value of the variable associated with b. The set of distributions that can be explained by a causal structure is completely determined by the set of causal-compatibility constraints, so two causal structures that are observationally equivalent will also share the same set of causal-compatibility constraints. These constraints can be of two types: they can take the form of equalities or inequalities. The equality constraints, such as the example given for $a \to b \to c$ above, are relatively easy to characterize, at least for causal structures with a small number of visible nodes. The inequality constraints, on the other hand, are much

harder to characterize. Therefore, it is useful to know *when* inequality constraints will be present. Our work in Ref. [48], that was not included in this thesis, looked at mDAGs associated with causal structures with 4 visible nodes and classified all but three of them as having inequality constraints or not (three mDAGs are left unresolved). The work of Chapter 3 on observational equivalence is also useful for characterizing which causal structures present inequality constraints: as it turns out, a causal structure has inequality constraints if and only if it is *not* observationally equivalent to any causal structure that lacks latent nodes [35]. The classification of causal structures into observational equivalence classes is particularly important for attesting how *frequent* the appearance of inequality constraints is. As will be argued in Chapter 3, the counting should be made by observational equivalence classes, i.e., one should ask what is the fraction of *observational equivalence classes* that present inequality constraints, not the fraction of mDAGs. With this, Chapter 3 also shows that **inequality constraints are ubiquitous**, that is, at least for causal structures with 4 visible nodes, inequality constraints are present in the majority of the observational equivalence classes.

The works mentioned above assume that the correlation between two variables comes from two possible types of causal explanations: direct causation or the presence of a hidden common cause. There is, however, a third way through which correlation between two variables can appear: selection effects.[3] This is the phenomena through which variables of interest become correlated when we condition on a specific value of a variable that is caused by our variables of interest. For a concrete example, imagine that the grass in a lawn could become wet for two reasons: either it is raining, or a sprinkler is on. Also imagine that this sprinkler is on at certain times regardless of whether it is raining or not; that is, the variables *rain* and *sprinkle* are initially independent. When we learn that the grass is wet, however, these variables become correlated: now, learning that it is not raining teaches me that the sprinkler must be on.[4]

Sometimes, all of the data we have access to is subject to selection effects. For example, if we ask about people's income and education level via an online questionnaire, all of our data is conditioned on the fact that the respondents have access to the internet and have interest in answering online forms, which might induce selection effects. Therefore, it is important to learn how to extract causal information from data that is subject to selection effects. Thus, in Ref. [16], which is presented in Chapter 4, we **generalized the results of Ref. [33] and of Chapters 2 and 3 to causal structures that can also involve selection effects**. That is, we first provide a necessary and sufficient graphical condition for two classical causal structures that can involve both latent confounding and selection

[3] Also referred to as *selection bias* or *postselection*.

[4] Example taken from Ref. [58].

effects to be indistinguishable under access to a maximally informative interventional probing scheme. This graphical condition is given by a graphical structure introduced in this work, the selected-marginalized directed graph (smDG). The smDG is the generalization of the mDAG to include selection effects: two causal structures that can involve both latent confounding and selection effects are indistinguishable under access to a maximally informative probing scheme if and only if they correspond to the same smDG. Thus, this work establishes the smDG as a fundamental object in the causal discovery problem with latent variables and selection effects. Furthermore, in this work we also study observational equivalence of causal structures that correspond to different smDGs, and we provide certain rules that show observational equivalence. This list of rules, however, is far from comprehensive: as seen in Chapter 3, the problem of deciding whether two causal structures are observationally equivalent is not completely solved even in the case without selection effects.

1.2 Causal inference and quantum theory

As mentioned before, Bell's theorem can be seen as a challenge to our classical notions of causality. In a Bell scenario, two parties Alice and Bob make measurements on a shared system. The measurement choice of Alice is determined by a setting variable X_s, and her measurement outcome is determined by a variable X_x. The measurement choice of Bob is determined by a setting variable X_t, and his measurement outcome is determined by a variable X_y. Alice and Bob make their measurements in space-like separation, so we do not expect there to be any causal influence going across the two wings of the experiment. We allow, however, for a common cause between the outcomes X_x and X_y. Furthermore, we assume that the choice of settings X_s and X_t is independent of any other variable that is relevant to the experiment. The most natural causal structure to describe a Bell scenario, thus, is the *Bell causal structure*, presented in Figure 1.2.1. There, visible nodes are black with white background and latent nodes are red with gray background, and the notation is such that a variable X_v is associated with a node v.

However, standard causal models with this causal structure *cannot* explain all the data obtained when the system shared by Alice and Bob is a quantum system! This is because some of the causal-compatibility inequalities imposed by the Bell causal structure are *Bell inequalities*, which are known to be violated by quantum experiments [8, 23].

If one still wants to find causal explanations for the data obtained from Bell experiments on quantum systems, there are a few options available. First, one could stick to the standard causal modeling framework, and alter the causal structure. Figure 1.2.2 shows a

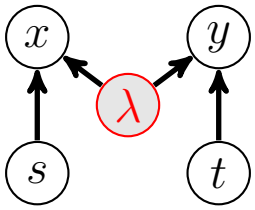
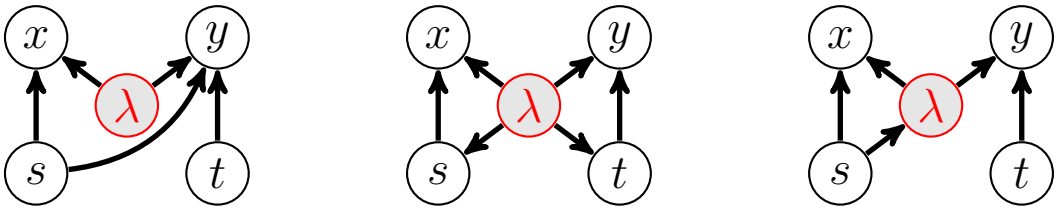


Figure 1.2.1: The Bell causal structure.

few examples, all of which *can* violate Bell inequalities and explain the data obtained from Bell experiments on quantum systems. All of these causal structures, however, violate some metaphysical assumption that we may regard as fundamental: Fig. 1.2.2(a) has causal influence outside of the light cone, Fig. 1.2.2(b) does not allow Alice and Bob to freely choose their settings, and Fig. 1.2.2(c) has causal influence back in time, because we assume the shared system is prepared before Alice makes her measurement choice. These violations are respectively referred to as superluminal causation, superdeterminism and retrocausality.



(a) (b) (c)

Figure 1.2.2: (a) Causal structure for a Bell experiment that has superluminal causation. (b) Causal structure for a Bell experiment that has superdeterminism. (c) Causal structure for a Bell experiment that has retrocausality.

Apart from violating the metaphysical principles mentioned above, the causal explanations of Figure 1.2.2 have another problem: all of them can also explain some probability distributions that have *signaling* outside of the light cone. For example, in Fig. 1.2.2(a) X_y can functionally depend on X_s, so this causal structure can explain some probability distributions where $P(X_y|X_sX_t)$ depends on the value of X_s. Of course, there are also probability distributions that can be explained by Fig. 1.2.2(a) while respecting $P(X_y|X_sX_t) = P(X_y|X_t)$, but these arise from a very small subset of the all possible ways in which X_y can functionally depend on X_s, X_t, X_λ. Therefore, we say that causal models that use Fig. 1.2.2(a) to explain the distributions obtained from Bell experiments

are *fine-tuned* [83][5], because their functional parameters must be chosen so precisely that they wash out the dependence of X_y on X_s. If one believes in special relativity (i.e., if one believes that we will never see superluminal signaling in nature) and wants to champion Fig. 1.2.2(a) as the correct causal explanation for the Bell experiment, one then needs too explain why nature only allows for certain choices for these functional parameters. As it turns out, Ref. [83] has shown that *all* of the classical causal models that can explain the data obtained from Bell experiments on quantum systems are fine-tuned.

One can be unsatisfied with these causal explanations that are both fine-tuned and violate important metaphysical principles. In this case, the way to causally explain the correlations obtained from Bell experiments is to reject the standard classical framework for causal modeling. In this alternative, one still defends that the the causal structure to explain a Bell experiment is Figure 1.2.1, but the way through which this causal structure generates probability distributions is not via simple functional dependencies anymore. That is, one is conservative with the causal structure, but radical about the meaning of causal influence. This alternative has led to the development of frameworks for quantum causality [47, 7, 25].

In the framework used in this thesis, following Ref. [47], visible nodes are always associated with classical variables, but latent nodes can be associated with quantum systems. Causal influence is defined by functions when the variables are classical, and by quantum channels when the cause is a quantum system. By associating the latent node λ with a quantum system, the Bell causal structure of Figure 1.2.1 *can* explain all of the distributions obtained from a Bell experiment on quantum systems. Furthermore, it does so *without* fine-tuning: the no-signaling relations $P(X_y|X_sX_t) = P(X_y|X_t)$ and $P(X_x|X_sX_t) = P(X_x|X_s)$ are causal-compatibility constraints imposed by the Bell causal structure (even when λ is quantum), so *all* choices of functional parameters on this causal structure yield distribution that *do not* have signaling outside of the light cone.

This reformulation of Bell's theorem in the language of causal inference opens the door for the exploration of quantum advantage in other causal structures. If a causal structure can explain more probability distributions over the visible variables when its latent nodes are associated with quantum systems than when its latent nodes are associated with classical variables, we say that this causal structure presents a quantum-classical gap (QC gap). As discussed above, the Bell causal structure is one example of causal structure that presents a QC gap. This property is important because, when we know that a phenomenon is governed by a causal structure that presents a QC gap, observing a probability distribution over the visible variables that cannot be explained using classical

[5] In causal inference literature, fine-tuning is referred to as *unfaithfulness* [74] or lack of *stability* [57].

latent variables in that structure certifies that the phenomenon involves genuine quantumness. In the last decade, other examples of causal structures that present QC gaps were found [15, 39, 18, 78, 50], and a lot of work in the area of network nonlocality came from that. However, the number of available examples was still relatively small. Finding more such examples is the main goal of Chapter 5.

A causal structure only has the potential to present a QC gap if, in the classical case, it has inequality constraints. This is because the only causal-compatibility constraints that can be violated when we go from classical models on a causal structure $\mathcal{G}$ to quantum models on $\mathcal{G}$ are inequality constraints, such as Bell inequalities. Therefore, our work on observational equivalence classes and its conclusion about the ubiquitousness of inequality constraints have important consequences for the search for new QC gaps: the results from Chapter 3 show that the potential for finding new causal structures that present a QC gap is huge. In fact, this is where our initial motivation for studying inequality constraints comes from. This is an example of how ideas from quantum physics can provide new insights into classical causal inference: compared to equality constraints, inequality constraints have generally received less attention in the classical causal inference community, but our results from Chapter 3 highlight their importance.

After filtering out the causal structures of 4 visible nodes that present inequality constraints (and thus have the *potential* for a QC gap), in Chapter 5 we ask which ones of these *actually* present a QC gap. The techniques we develop to do so will be called *piggybacks*, because they use the knowledge that a certain causal structure has a QC gap to prove that a new causal structure also has a QC gap. With these techniques, we show that **the majority of 4-node mDAGs that have inequality constraints also have a QC gap**: out of 996 4-node mDAGs (up to relabelling of the nodes) that might have inequality constraints,[6] we prove that 926 have a QC gap. The remaining 70 are left as open problems. This is a huge increase in the number of examples of QC gaps that we have available, that can have multiple foundational and practical applications in the future.

In general, the fact that two causal structures are classically observationally equivalent *does not* imply that they are also observationally equivalent in the quantum case. In fact, even if two causal structures are associated with the *same mDAG*, this does not imply that they are quantumly observationally equivalent: a counterexample was first presented in Ref. [81], and it is reproduced at the beginning of Chapter 5. This distinction highlights another interesting conclusion from Chapter 2: its results show that quantum and classical causal inference are *even more distinct* than previously thought, because there are examples of causal structures that can be distinguished by *passive observations* in the quantum case,

[6] 993 of these are proven to have inequality constraints, while 3 are open problems [48].

while in the classical case they are indistinguishable not only by passive observations, but also by any interventional probing scheme that interacts with the visible variables.

Another interesting question then is when are two causal structures observationally equivalent (or equivalent under access to other probing schemes) in the quantum case. An intermediary result of Chapter 5 is a list of rules that prove that two causal structures are quantumly observationally equivalent, or one quantumly observationally dominates the other. This question, however, is far from closed: there is still work left to do in the classical case (as shown in Chapter 3), and we know even less about the quantum case.

In this thesis, we only discuss how to witness quantum advantage in data obtained from *passive observations*, that is, the concept of QC gap is defined in a way that does not consider interventional data. There have been, however, recent works [41, 60] that explore how to leverage interventional data to witness quantum advantage in situations where passive observations would not be sufficient to do so. This research direction shows that the careful analysis of interventional probing schemes, that was the focus of Chapter 2, is also of interest to quantum physicists.

The questions asked in Chapter 3 are a natural follow-up to Chapter 2. Chapter 4 presents a generalization of the main result of Chapter 2 to incorporate selection effects, and takes initial steps toward an analogous generalization of Chapter 3. On the other hand, Chapter 5 is connected to Chapter 3, but not so tightly connected to Chapters 2 and 4. It presents one quantum observational equivalence rule that involves selection effects, but at the moment its main analysis (the search for new causal structures that present QC gaps) is focused on causal structures *without* selection effects. Furthermore, as described above, it only discusses witnessing quantum advantage on data obtained from passive observations. Therefore, a possible direction for future research is strengthening the ties between Chapter 5 and Chapters 2 and 4. This could be done for example by generalizing our piggybacks in a way that can both prove the existence of QC gaps in causal structures that include selection effects, and find more cases in which it is possible to use *both* observational and interventional data to detect nonclassicality, building on the work of Refs. [41, 60].

My PhD research has thus taken a bidirectional character: while I used tools and motivations from quantum research to get insights into classical causal inference, I also used causal inference tools to address problems in quantum foundations. Two of my papers that were not included in this thesis also exemplify this bidirectional research: in the first direction, Ref. [5] develops a **resource theory of causal influence** and a **resource theory of knowledge about causal influence** in the classical case, aiming to formalize the quantification of causal strength. This provides a clear example of application of a tool

from quantum research to classical causal inference: resource theories were developed in quantum information to quantify resources such as entanglement, athermality, coherence etc. In the second direction, in Ref. [84] we **used the framework of causal inference to better understand Extended Wigner's Friend Scenarios**, which are thought experiments where an observer is treated as a quantum system. There, we analyzed the Local Friendliness no-go theorem [12] using the same methods that were used in Ref. [83] to analyze the Bell no-go theorem.

Chapter 2

Interventional distinguishability of causal structures in the presence of latent nodes

The most common type of probing scheme considered in causal discovery is passive observation. It is well known, however, that allowing interventions in one's probing scheme (of which there are different varieties) generally allows one to draw stronger conclusions about the causal structure. In short, by availing oneself of interventional probing schemes, one can resolve ambiguities that hold for observational probing schemes.

In this chapter, we determine what ambiguity about the causal structure remains regardless of how the visible variables are probed, including when one makes use of an *informationally complete probing scheme*, which is one that provides us with all of the information about the causal structure and the causal parameters that is obtainable by interacting with the visible variables.[1] One example of an informationally complete probing scheme is a scheme where, for each visible variable, one observes its natural value and then implements a *do-intervention*, which forces the variable to take a fixed desired value. We refer to this as the *Observe&Do* probing scheme.

We find that variations of the causal structure that cannot be discriminated are all and only those that are consistent with the same *marginalized DAG* (mDAG) structure, a notion that was introduced in Ref. [33]. In other words, an informationally complete

[1] Edge interventions [71] are not considered among the probing schemes investigated in this work. Nonetheless, we believe that the conclusions of this chapter can be extended to the case where one has access to edge interventions.

probing scheme determines the causal structure up to its mDAG equivalence class: two causal structures that are associated with the same mDAG are indistinguishable and two causal structures that are associated with different mDAGs can always be distinguished. Our results therefore establish that the mDAG structure is a fundamental structure for causal analysis.

We should emphasize that, when we say that two causal structures $\mathcal{G}$ and $\mathcal{G}'$ are *distinguishable* under some probing scheme, this means that there is, *in principle*, the possibility of obtaining a set of data from said probing scheme that would tell us that the underlying causal structure is definitely one and not the other. In other words, there exists *one* set of data that is realizable by one of these causal structures but not by the other. This *does not* imply that the set of data we obtain from one specific experiment will always be able to tell us if the underlying causal structure is $\mathcal{G}$ or $\mathcal{G}'$. For example: a distribution that is completely factorizing, that is, where none of the visible variables depend on any other visible variable (be it under interventions or not) is realizable by *every* causal structure. Therefore, if one starts with such a distribution as their data, one certainly will not be able to tell if it was generated by $\mathcal{G}$ or $\mathcal{G}'$ (if one does not impose any model-selection principles such as faithfulness[74, 57]).

In this chapter we also investigate to what extent one can weaken the probing scheme and still maintain the discriminatory power of an informationally complete probing scheme. We focus on two probing schemes of this type.

Firstly, we consider the probing scheme where, for each visible variable, it is possible to perform a passive observation *or* a do-intervention, but not both. This is the most commonly studied type of interventional probing scheme: in general, in a blind drug trial the experimentalist does not bother to ask the subjects what is their preference regarding taking the drug or not before assigning them the drug or the placebo. If one partitions the ensemble of samples into subensembles and applies do-interventions on different subsets of the visible variables in each subensemble (with passive observations of the complementary set of visible variables) for *all* possible subsets of the set of visible variables, we say that they are applying the *all-patterns Observe-or-Do* probing scheme. For example, if there

are three visible variables X_a, X_b and X_c, the patterns are:

$$\begin{aligned}
&\text{Observe } X_a, X_b, X_c \\
&\text{Intervene on } X_a, \text{ Observe } X_b, X_c \\
&\text{Intervene on } X_a, X_b, \text{ Observe } X_c \\
&\text{Intervene on } X_a, X_c, \text{ Observe } X_b \\
&\text{Intervene on } X_b, \text{ Observe } X_a, X_c \\
&\text{Intervene on } X_b, X_c, \text{ Observe } X_a \\
&\text{Intervene on } X_c, \text{ Observe } X_a, X_c.
\end{aligned}$$

As we show here, distinguishability relative to the all-patterns Observe-or-Do probing scheme is *also* characterized by mDAGs: even this weaker probing scheme can distinguish between any two causal structures that correspond to different mDAGs. Therefore, even though the all-patterns Observe-or-Do probing scheme does not provide us with full information about the causal structure *and* the causal parameters, it does provide us with full information about the causal structure alone. (To see that it cannot distinguish between causal structures that correspond to the same mDAG, it suffices to note that it is no more informative than an informationally complete probing scheme.)[2]

Secondly, we consider a probing scheme that is even weaker than the one just described, now allowing only for do-interventions that set a visible variable to *one* of its allowed values, called *one-value do-interventions*. Here, one can imagine an example of a test of the effect of smoking on developing lung cancer. While the experimentalist can request people to stop smoking, thus setting the value of the "smoker" variable to 0, they cannot ethically request people to start smoking, so they are not allowed to make an intervention that sets the variable "smoker" to the value 1. When one partitions the ensemble of samples and applies one-value do-interventions on all possible subsets of visible nodes, we say that one is implementing the *all-patterns Observe-or-1Do* probing scheme. As it turns out, we prove that two causal structures are *also* indistinguishable under the *all-patterns Observe-or-1Do* probing scheme if and only if they correspond to the same mDAG.

Apart from solving this indistinguishability problem, in this chapter we also fully characterize when one causal structure can realize all the sets of data that are realizable by

[2] In Ref. [33], it was claimed that there are pairs of causal structures associated with different mDAGs which are indistinguishable when, for each visible variable, the experimenter performs a passive observation *or* a do-intervention (an example is given in Fig. 16 of Ref. [33]). Here, we argue against this conclusion. The discrepancy between our result and the claim of Ref. [33] comes from the fact that Ref. [33] did not consider the relationship between the data obtained from applying do-interventions on different subsets of the set of visible nodes.

another causal structure for each of the three different types of probing schemes studied here. In this case, we say that the first causal structure *dominates* the second relative to the probing scheme in question. As it turns out, the characterization of the dominance relations among causal structures, like the characterization of the equivalence relations, is the same for the three types of probing schemes studied here, and it is determined by the structure of the corresponding mDAGs.

In summary, in this chapter we show that two causal structures that are associated with the same mDAG are indistinguishable even when there is access to an informationally complete probing scheme, which corresponds to a very strong experimental power. Furthermore, two causal structures that are associated with different mDAGs can still be distinguished by the all-patterms Observe-or-1Do probing scheme, which corresponds to a much weaker experimental power.

The structure of this chapter is as follows: In Section 2.1, we will give the background about causal modelling that will be necessary for our examples and results. This background will also be relevant in all of the other chapters of this thesis. In Section 2.2, we will discuss the full-SWIG, a structure that will be central to frame most of our discussion. In Section 2.3, we will discuss in more detail the notion of different probing schemes and the information revealed by these. In Section 2.4, we will present our first main result: the mDAG structure completely captures equivalence and dominance with respect to any informationally complete probing scheme, of which the Observe&Do probing scheme is an example. In Section 2.5, we present our second and third main results. The second main result says that by restricting ourselves to the all-patterns Observe-or-Do probing scheme, we do not lose any power in discriminating causal structures. The third main result strengthens this by showing that even restricting ourselves to the all-patterns Observe-or-1Do probing scheme does not cost any discriminating power. In other words, the mDAG structure also completely captures equivalence and dominance with respect to the all-patterns Observe-or-Do and the all-patterns Observe-or-1Do probing schemes. In Section 2.6, we provide some concrete examples of the partial order of equivalence classes of causal structures that is implied by the dominance relation induced by any of the three probing schemes considered here (as noted above, the dominance relation is the same for all three). Specifically, we describe the partial orders for all causal structures over three visible variables and all causal structures over four visible variables. Finally, in Section 2.7, we describe some open problems and future directions for research.

2.1 Preliminaries

A causal structure is represented by a directed acyclic graph (DAG), which is given by $\mathcal{G} = (\texttt{nodes}(\mathcal{G}), \texttt{edges}(\mathcal{G}))$, where $\texttt{nodes}(\mathcal{G})$ is a collection of nodes and $\texttt{edges}(\mathcal{G})$ is a collection of directed edges between these nodes. In this work, we will use lower case letters to indicate nodes, and upper case letters to indicate sets of nodes.

Below, we define some terminology about DAGs.

Definition 1 (Children, Parents, Descendants, Ancestors)**.** *Let $\mathcal{G}$ be a DAG, and let $u, v, w \in \texttt{nodes}(\mathcal{G})$. If $\mathcal{G}$ includes a directed edge $v \to w$, then w is called a child of v. Conversely, a node u that has v as a child is called a parent of v. The set of all children of v is denoted $\texttt{ch}_{\mathcal{G}}(v)$, while the set of all parents of v is denoted $\texttt{pa}_{\mathcal{G}}(v)$. A node $w \in \texttt{nodes}(\mathcal{G})$ is a descendant of v if it can be reached from v by a sequence of nodes through directed edges pointing away from v. Conversely, a node $u \in \texttt{nodes}(\mathcal{G})$ that has v as a descendant is called an ancestor of v. The set of all descendants of v is denoted $\texttt{des}_{\mathcal{G}}(v)$, and the set of all ancestors of v is denoted $\texttt{an}_{\mathcal{G}}(v)$. Nodes with no parents in the DAG are termed* exogenous*, while those that do have parents in the DAG are termed* endogenous*.*

Definition 2 (Subgraph)**.** *Let $\mathcal{G}$ and $\mathcal{G}'$ be DAGs. If $\texttt{nodes}(\mathcal{G}') \subseteq \texttt{nodes}(\mathcal{G})$ and $\texttt{edges}(\mathcal{G}') \subseteq \texttt{edges}(\mathcal{G})$, we say that $\mathcal{G}'$ is a subgraph of $\mathcal{G}$.*

In particular, if we start from $\mathcal{G}$ and delete all of the nodes that are not *in a particular subset $S \in \texttt{nodes}(\mathcal{G})$ (hence also deleting their incoming and outgoing edges), the resulting induced subgraph will be denoted by $\mathcal{G}_S$.*

Some nodes of a causal structure are probed in experiments, while others are not. As noted earlier, we refer to this distinction with the terms *visible* and *latent*. In order to save the term "DAG" for the mathematical object composed of nodes and edges, we define a new object to represent a DAG with a partitioned set of nodes:[3]

Definition 3 (Partitioned DAG)**.** *A* partitioned DAG (pDAG) *$\mathcal{G}$ is a DAG together with a partition of its nodes into two subsets: the set of* visible nodes*, denoted $\texttt{vis}(\mathcal{G})$ and the set of* latent nodes*, denoted $\texttt{lat}(\mathcal{G})$, where $\texttt{vis}(\mathcal{G}) \cup \texttt{lat}(\mathcal{G}) = \texttt{nodes}(\mathcal{G})$ and $\texttt{vis}(\mathcal{G}) \cap \texttt{lat}(\mathcal{G}) = \varnothing$.*

[3]The notion of a pDAG is the same as the notion of a "latent structure" defined by Pearl (Definition 2.3.2 of Ref. [57]). Here we opted for the term "pDAG" since it is even-handed, putting latent and visible nodes on an equal footing. Also, this term allows for expressions such as "latent-confounder-free pDAG".

In a pDAG $\mathcal{G}$, the set of all *visible* parents of a node $a \in \texttt{nodes}(\mathcal{G})$ will be denoted by $\texttt{Vpa}_{\mathcal{G}}(a)$, while the set of all *latent* parents of a will be denoted by $\texttt{Lpa}_{\mathcal{G}}(a)$. Similarly, the sets of visible and latent children of a are denoted by $\texttt{Vch}_{\mathcal{G}}(a)$ and $\texttt{Lch}_{\mathcal{G}}(a)$ respectively. When depicting pDAGs in this article we follow the convention that visible nodes are circular with white backgrounds and latent nodes are circular with grey backgrounds. A pDAG that is composed only of visible nodes will be called a *latent-confounder-free* pDAG. Whenever two visible nodes of a pDAG are children of the same latent node, we will refer to the latent node as a *latent confounder* of the two children.

2.1.1 Explaining Observed Data with a Causal Structure

A natural experiment is one that passively observes a set of variables over many runs. Such an experiment yields a statistical estimate of the probability distribution over the possible values of these variables. Determining which causal structures could have realized this distribution is the causal discovery problem.

To do so, one has to specify the *semantics* of this causal model[4], a description of what type of entities are associated to the nodes, and what type of mechanism determines how each node depends causally on its parents in the DAG, and of how to parametrize statistical uncertainty.

In Chapters 2 to 4, we will only consider causal model semantics of the *classical* type, wherein the entities that are associated with nodes are classical random variables. An alternative possibility, however, is the semantics used in quantum causal modelling. There, the entities that can be associated with nodes are quantum systems, each of which is described by a Hilbert space[5]. See Refs. [19, 81, 72] for works on the causal discovery problem in causal models with quantum semantics. In Chapter 5 we will also consider causal models whose semantics allow latent nodes to be associated with quantum systems.

[4] A causal structure is analogous to syntax in linguistics, which is the structural form of a sentence, while the semantics of a causal model is the analogue of semantics in linguistics. Semantics refers to a specific choice of objects that follow the structure defined by the syntax. In linguistics, semantics provides the meanings of the terms in a sentence. In causal modelling, it provides the parameters of the model. We thank Nick Ormrod for the suggestion of this terminology.

[5] The analogue of the cardinality of a random variable is the dimension of this Hilbert space. The analogue of an error variable of a given cardinality is an ancillary quantum system of a given dimension, and the analogue of a distribution over the error variable is a quantum state of the ancillary system. Finally, the analogue of a functional dependence between classical random variables is a unitary channel between quantum systems.

The variables associated with visible nodes will be called *visible variables*, and the variables associated with latent nodes will be called *latent variables*. For each $a \in \texttt{nodes}(\mathcal{G})$, we denote the associated variable by X_a, and for a set S of nodes, we denote the associated joint variable by X_S. The set of all possible values of X_a, termed its *state space*, is denoted $\mathcal{X}_a$. In this chapter, we will assume that each variable has a discrete and finite number of possible outcomes.[6] Under this assumption, the cardinality of each $\mathcal{X}_a$ is a natural number, so that the set of such cardinalities can be specified by a vector $\vec{c} \in \mathbb{N}^{|\texttt{nodes}(\mathcal{G})|}$, where $|S|$ denotes the cardinality of a set S. The semantics of a causal model includes a specification of $\vec{c}$, the cardinality of the variables.

We also associate to each $a \in \texttt{nodes}(\mathcal{G})$ a random variable E_a, called an error variable, whose state space is denoted by $\mathcal{E}_a$. The variable X_a is determined by its associated error variable E_a and its causal parents $X_{\texttt{pa}_\mathcal{G}(a)}$ via a function $f_a : \mathcal{X}_{\texttt{pa}_\mathcal{G}(a)} \times \mathcal{E}_a \to \mathcal{X}_a$, that is,

$$X_a = f_a(X_{\texttt{pa}_\mathcal{G}(a)}, E_a) \tag{2.1}$$

Finally, to each error variable E_a, one associates a probability distribution $P(E_a)$ that stipulates how E_a is sampled. Thus, the inclusion of the error variable E_a allows for X_a to respond probabilistically to $X_{\texttt{pa}_\mathcal{G}(a)}$, even though f_a is a deterministic function.

The cardinality of the visible, latent and error variables, the functions, and the error-variable distributions are jointly termed the *parameters* of the causal model:

$$\texttt{par} = \{(|\mathcal{X}_a|, |\mathcal{E}_a|, f_a, P(E_a)) : a \in \texttt{nodes}(\mathcal{G})\} \tag{2.2}$$

A *causal hypothesis* $(\mathcal{G}, \texttt{par})$ is a pDAG $\mathcal{G}$ together with a choice of parameters for $\mathcal{G}$. A particular causal hypothesis realizes one particular probability distribution over the visible variables, which we will denote by $P^{(\mathcal{G},\texttt{par})}(X_{\texttt{vis}(\mathcal{G})})$.

The semantics of the causal model stipulates the scope of possibilities for par. For instance, the functional dependencies could all be linear and the error-variable distributions could all be uniform. It could also be the case that the scope of possibilities for par is unrestricted. When this is the case, we will say that the semantics is *classical unrestricted*. Note that there are still different options of classical unrestricted semantics, corresponding to different choices of cardinalities of the visible, latent and error variables.

In Chapters 2 to 4, we will be dealing only with classical unrestricted semantics. In general, one might be interested in the distributions over visible variables realizable by a

[6] As proven for example in Ref. [38], when the visible variables have finite cardinality, we can assume that the latent variables also have finite cardinality without loss of generality.

pDAG for different types of classical unrestricted semantics; for example, when the latent nodes are associated with classical random variables that have specific fixed cardinalities. In the causal inference literature, the case of most common interest is when the latent nodes are allowed to take arbitrary cardinalities. We now turn to the notion of realizability of observational data under this type of model.

Definition 4 (Set of probability distributions with cardinality vector $\vec{c}_{\text{vis}}$ realizable by a pDAG by a causal model with classical unrestricted semantics under arbitrary cardinality of the latent variables)**.** *Let $\mathcal{G}$ be a pDAG. Associate it to a causal model with classical unrestricted semantics where the visible variables $\{X_v : v \in \texttt{vis}(\mathcal{G})\}$ have cardinalities given by the vector $\vec{c}_{vis} \in \mathbb{N}^{|\texttt{vis}(\mathcal{G})|}$, and the latent variables are allowed to take arbitrary cardinalities.*

The set of probability distributions on the visible variables that are realizable under passive observation by $\mathcal{G}$ for this causal semantics, denoted $\mathcal{M}_{obs}(\mathcal{G}, \vec{c}_{vis})$, is the set of all probability distributions P over $X_{\texttt{vis}(\mathcal{G})}$ that can be obtained as

$$P(X_{\texttt{vis}(\mathcal{G})}) = \sum_{X_{\texttt{lat}(\mathcal{G})}} \sum_{E_{\texttt{nodes}(\mathcal{G})}} \prod_{a \in \texttt{nodes}(\mathcal{G})} \delta_{X_a, f_a(X_{\texttt{pa}_{\mathcal{G}}(a)}, E_a)} P(E_a) \tag{2.3}$$

for some choice of parameters, i.e., for some choice of the set $\{(|\mathcal{X}_a|, |\mathcal{E}_a|, f_a, P(E_a)) : a \in \texttt{nodes}(\mathcal{G})\}$.

We will refer to a distribution that is realizable by a pDAG under arbitrary cardinality of the latent variables simply as a *distribution that is realizable* by that pDAG.

If a given probability distribution over visible variables is not realizable in a certain causal structure, then this causal structure is *not* a valid causal explanation for the statistical data on these variables. Causal structures that can realize *all* the probability distributions over the visible variables are called *saturated*.

The set $\mathcal{M}_{\text{obs}}(\mathcal{G}, \vec{c}_{\text{vis}})$ for a pDAG $\mathcal{G}$ is defined with respect to a specific assignment of the cardinalities of the visible variables. It is useful to consider the tuple of possibilities for this assignment, and the corresponding tuple of probability distributions that are realizable for each possibility. Since this tuple contains all the information that can be extracted from the causal structure by passive observations, we call it the *observational profile* of that structure:

Definition 5 (Observational profile)**.** *Let $\mathcal{G}$ be a pDAG. The* observational profile *of $\mathcal{G}$, denoted $\mathcal{M}\text{-}\texttt{profile}(\mathcal{G})$, is the tuple*

$$\mathcal{M}\text{-}\texttt{profile}(\mathcal{G}) := \left(\mathcal{M}_{obs}(\mathcal{G}, \vec{c}_{vis}) : \vec{c}_{vis} \in \mathbb{N}^{|\texttt{vis}(\mathcal{G})|}\right), \tag{2.4}$$

that describes the set of realizable distributions over the visible variables of $\mathcal{G}$ for each choice $\vec{c}_{vis}$ of their cardinalities.

We refer here to a *tuple* rather than a set (and use parentheses rather than curly brackets) to highlight that the order in which the terms appear inside the observational profile is important. In other words, in the observational profile every set $\mathcal{M}_{\text{obs}}(\mathcal{G}, \vec{c}_{\text{vis}})$ is indexed by the vector $\vec{c}_{\text{vis}}$ of cardinalities of the visible variables. This allows us to define an inclusion relation among observational profiles:

Definition 6 (Element-wise inclusion of observational profiles)**.** *Let $\mathcal{G}$ and $\mathcal{G}'$ be pDAGs such that $\texttt{vis}(\mathcal{G}) = \texttt{vis}(\mathcal{G}')$. We say that the observational profile of $\mathcal{G}'$ is* element-wise included *in the observational profile of $\mathcal{G}$, denoted by $\mathcal{M}\texttt{-profile}(\mathcal{G}') \subseteq_{ew} \mathcal{M}\texttt{-profile}(\mathcal{G})$, if*

$$\mathcal{M}_{obs}(\mathcal{G}', \vec{c}_{vis}) \subseteq \mathcal{M}_{obs}(\mathcal{G}, \vec{c}_{vis}) \quad \forall \vec{c}_{vis} \in \mathbb{N}^{|\texttt{vis}(\mathcal{G})|}. \tag{2.5}$$

2.1.2 Observational Equivalence and Dominance

If it happens that *all* of the probability distributions that are observationally realizable by a pDAG $\mathcal{G}$ are also observationally realizable by a different pDAG $\mathcal{G}'$ and vice-versa, then it is impossible to know which one of these two pDAGs generated the statistical data obtained from passive observations. This idea is formalized by the definition of *observational equivalence* of pDAGs, which is a special case of *observational dominance*:[7]

Definition 7 (Observational dominance and equivalence of pDAGs)**.** *Let $\mathcal{G}$ and $\mathcal{G}'$ be pDAGs such that $\texttt{vis}(\mathcal{G}) = \texttt{vis}(\mathcal{G}')$. We say that $\mathcal{G}$* observationally dominates *$\mathcal{G}'$ when the set of realizable distributions of $\mathcal{G}$ includes the set of realizable distributions of $\mathcal{G}'$, regardless of the cardinalities of the visible variables. That is, when $\mathcal{M}\texttt{-profile}(\mathcal{G}') \subseteq_{ew} \mathcal{M}\texttt{-profile}(\mathcal{G})$. The observational dominance relation is denoted by $\mathcal{G} \succeq \mathcal{G}'$.*

We say that $\mathcal{G}$ is observationally equivalent *to $\mathcal{G}'$ when $\mathcal{M}\texttt{-profile}(\mathcal{G}') = \mathcal{M}\texttt{-profile}(\mathcal{G})$. The observational equivalence relation is denoted by $\mathcal{G} \cong \mathcal{G}'$. If $\mathcal{G} \succeq \mathcal{G}'$ but $\mathcal{G} \ncong \mathcal{G}'$, we say that $\mathcal{G}$* strictly dominates *$\mathcal{G}'$ and denote this relation as $\mathcal{G} \succ \mathcal{G}'$. If $\mathcal{G} \not\succeq \mathcal{G}'$ and $\mathcal{G}' \not\succeq \mathcal{G}$, we say that $\mathcal{G}$ and $\mathcal{G}'$ are* incomparable.

[7] The notion of observational dominance of pDAGs is similar to Pearl's notion of *preference* among latent structures (see Definition 2.3.3 of Ref. [57]), except that the preference order is opposite to the dominance order. That is, $\mathcal{G}$ is preferred to $\mathcal{G}'$ when $\mathcal{G}'$ observationally dominates $\mathcal{G}$. Also, the preference order is defined for a fixed cardinality of the visible variables, whereas the dominance order defined here involves a universal quantifier over these cardinalities.

Note that to prove observational equivalence, one must demonstrate that $\mathcal{M}\text{-}\texttt{profile}(\mathcal{G}') = \mathcal{M}\text{-}\texttt{profile}(\mathcal{G})$, which requires demonstrating equality for all possible cardinalities of the visible nodes. To prove observational *in*equivalence, on the other hand, it suffices to show that $\exists \vec{c}_{\text{vis}} \in \mathbb{N}^{\texttt{vis}(\mathcal{G})} : \mathcal{M}_{\text{obs}}(\mathcal{G}, \vec{c}_{\text{vis}}) \neq \mathcal{M}_{\text{obs}}(\mathcal{G}', \vec{c}_{\text{vis}})$, that is, it suffices to demonstrate inequality for just one particular assignment of cardinalities. An example of two pDAGs that are observationally equivalent to each other is given in Fig. 2.1.1; the observational equivalence of the three pDAGs there presented will be formally justified in Section 3.5.

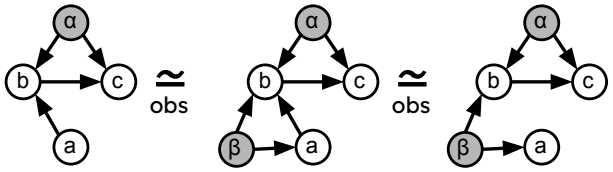


Figure 2.1.1: An example of pDAGs that are observationally equivalent. The first of this triple is generally known as the *instrumental scenario*.

Note that observational equivalence is *not* the same as Markov equivalence [79, 1]. Two pDAGs are Markov equivalent if their sets of conditional independence constraints are the same. For two pDAGs to be observationally equivalent, on the other hand, their *entire* set of causal-compatibility constraints (including conditional independence constraints, nested Markov constraints and inequality constraints) need to coincide. We will discuss these constraints in more detail in Chapter 3.

2.1.3 mDAGs

In this section, we recall the notion of an mDAG, defined in Ref. [33]. For each mDAG, we can associate a set of pDAGs. The pDAGs in such a set are always observationally equivalent to each other. Note, however, that observational equivalence does not imply equivalence of the mDAG structure: there are different mDAGs which are observationally equivalent.

The definition of an mDAG is motivated by two graphical operations defined in Ref. [33], here referred to as *exogenization of latent nodes* and *removal of redundant latent nodes*. Ref. [33] showed that these two graphical operations preserve the observational profile. That is, any two pDAGs related by such an operation are observationally equivalent. As we will see in Section 2.4, two pDAGs that are related by application of these operations

are not only indistinguishable relative to passive observations of the visible variables, but also relative to more informative interventional probing schemes, including those that are informationally complete in a sense that we will define further on.

Before defining these two operations in full generality, we can get a sense of their content by analyzing Fig. 2.1.2. In Fig. 2.1.2(a), we have a pDAG where the latent node α is *endogenous*, since it has parents β and c in the pDAG. It passes information from β and c to its children a and b. In Fig. 2.1.2(b), the latent node α does not have parents anymore, but its children a and b now have direct access to the information about β and c. The operation of transitioning from the pDAG of Fig. 2.1.2(a) to that of Fig. 2.1.2(b) is an instance of what we term "exogenization of latent nodes". It consists in making the latent variables of a pDAG *exogenous* (parentless) while connecting their parents directly to their children.

In Fig. 2.1.2(b), the latent node α establishes a common cause between a and b. On the other hand, the latent node β establishes a common cause between a, b and d. Since we allow the cardinalities of latent variables to be arbitrary (Def. 4), the latent node α can be absorbed into β, thus leading to Fig. 2.1.2(c). The operation of transitioning from the pDAG of Fig. 2.1.2(b) to that of Fig. 2.1.2(c) is an instance of what we term "removal of redundant latent nodes". It consists in deleting latent nodes such as α, whose set of children is a subset of the set of children of another latent node. The fact that these nodes only establish a latent confounder between nodes that already had a latent confounder is what justifies referring to them as redundant.

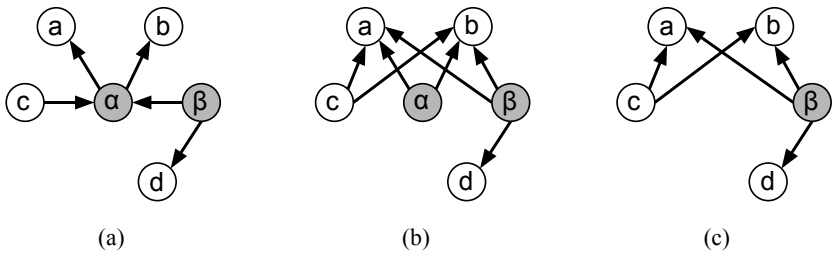


Figure 2.1.2: (a) A pDAG. (b) The pDAG obtained from (a) by exogenizing the latent node α. (c) The pDAG obtained from that of (b) by removing the latent node α, which is redundant to β since the children of α are a subset of the children of β. By Lemmas 1 and 2, these three pDAGs are observationally equivalent. The pDAG in (c) is RE-reduced.

The proof that these two operations always preserve observational equivalence was given in Lemmas 3.7 and 3.8 of Ref. [33], reproduced below:

Lemma 1 (Exogenize Latent Nodes [33, Lemma 3.7]). *Let $\mathcal{G}$ be a pDAG, and let* $\texttt{EndoLNodes}(\mathcal{G}) \subseteq \texttt{lat}(\mathcal{G})$ *be the set of latent nodes of $\mathcal{G}$ that are endogenous (i.e., they have one or more parents in $\mathcal{G}$). Construct the pDAG* $\texttt{Exog}(\mathcal{G})$ *as follows. For every* $u \in \texttt{EndoLNodes}(\mathcal{G})$*, start from $\mathcal{G}$ and: (i) delete all directed edges leading into u, (ii) add a directed edge from every parent of u to every child of u. The pDAG* $\texttt{Exog}(\mathcal{G})$ *so constructed is observationally equivalent to $\mathcal{G}$,* $\texttt{Exog}(\mathcal{G}) \cong \mathcal{G}$.

Lemma 2 (Remove Redundant Latent Nodes [33, Lemma 3.8]). *Let $\mathcal{G}$ be a pDAG where all latent nodes are exogenous. Let* $\texttt{RedundLnodes}(\mathcal{G}) \subset \texttt{lat}(\mathcal{G})$ *be a maximal subset of latent nodes such that their set of children is already common-cause-connected by another latent node. (Note that there is ambiguity in the choice, since when two latent nodes have the same set of children, one can take either to be the redundant one.) Formally, we define such a subset of latent nodes of $\mathcal{G}$, denoted* $\texttt{RedundLnodes}(\mathcal{G})$*, by the following two conditions: (i) For every* $u \in \texttt{RedundLnodes}(\mathcal{G})$ *there exists some distinct latent node* $v \in \texttt{lat}(\mathcal{G}) \setminus \texttt{RedundLnodes}(\mathcal{G})$ *such that* $\texttt{ch}(u) \subseteq \texttt{ch}(v)$*. (ii)* $\texttt{RedundLnodes}(\mathcal{G})$ *is not a strict subset of any other set of latent nodes that obeys (i). Let* $\texttt{RemoveRedund}(\mathcal{G})$ *be the pDAG constructed by removing from $\mathcal{G}$ every node in* $\texttt{RedundLnodes}(\mathcal{G})$*. The pDAG* $\texttt{RemoveRedund}(\mathcal{G})$ *so constructed is observationally equivalent to $\mathcal{G}$, i.e.,* $\texttt{RemoveRedund}(\mathcal{G}) \cong \mathcal{G}$.

For any pDAG $\mathcal{G}$, one can find another pDAG whose latent variables are all exogenous and non-redundant that is observationally equivalent to $\mathcal{G}$. Namely, it is the pDAG $\texttt{RemoveRedund} \circ \texttt{Exog}(\mathcal{G})$, which we will call the RE-*reduction* of $\mathcal{G}$. Accordingly, the map RE-reduce is defined as:

$$\texttt{RE-reduce} := \texttt{RemoveRedund} \circ \texttt{Exog} \tag{2.6}$$

It is easy to see that many different pDAGs have the same RE-reduction. A pDAG which is invariant under the RE-reduce map, i.e., such that $\mathcal{G} = \texttt{RE-reduce}(\mathcal{G})$, will be called a RE-*reduced* pDAG.

As noted in Ref. [33], this justifies the introduction of a new structure for studying realizable distributions for pDAGs: the *marginalized DAG* (mDAG). The notion of an mDAG makes use of the concept of a *simplicial complex*:

Definition 8 (Simplicial complex). *A simplicial complex over a finite set V is a set $\mathcal{B}$ of subsets of V such that*

- $\{v\} \in \mathcal{B}$ *for all* $v \in V$*;*
- *If* $A \subseteq B \subseteq V$ *and* $B \in \mathcal{B}$*, then* $A \in \mathcal{B}$.

The elements of $\mathcal{B}$ are called faces. The inclusion-maximal elements of a simplicial complex (the faces that are maximal in the order over faces induced by subset inclusion) are called facets.

An mDAG is a pair $\mathfrak{G} = (\mathcal{D}, \mathcal{B})$, where $\mathcal{D}$ is a DAG and $\mathcal{B}$ is a simplicial complex over the set of nodes of $\mathcal{D}$. The DAG $\mathcal{D}$ is called the *directed structure* of $\mathfrak{G}$. The nodes of $\mathcal{D}$ will also be referred to as the nodes of $\mathfrak{G}$. The edges of $\mathcal{D}$ will be referred to as the directed edges of $\mathfrak{G}$ and denoted $\texttt{DirectedEdges}(\mathfrak{G})$. Here, we will use the font $\mathfrak{G}$ to denote mDAGs, while the font $\mathcal{G}$ will continue to be used to denote DAGs and pDAGs.

With this notion in hand, we can introduce the map `ToFaces` that associates an mDAG to each pDAG:

Definition 9 (Map `ToFaces` from pDAGs to mDAGs)**.** *Let $\mathcal{G}$ be a pDAG, and let $\tilde{\mathcal{G}}$ be the* RE*-reduction of $\mathcal{G}$, i.e., $\tilde{\mathcal{G}} = \texttt{RE-reduce}(\mathcal{G})$. Recall that $\texttt{vis}(\tilde{\mathcal{G}}) = \texttt{vis}(\mathcal{G})$. The* mDAG associated with $\mathcal{G}$*, denoted by $\texttt{ToFaces}(\mathcal{G})$, is the pair $(\mathcal{D}, \mathcal{B})$ where*

- *$\mathcal{D}$ is a DAG such that $\texttt{nodes}(\mathcal{D}) = \texttt{vis}(\mathcal{G})$ and whose edges correspond to the edges between the visible nodes of $\tilde{\mathcal{G}}$.*
- *$\mathcal{B}$ is a simplicial complex over $\texttt{vis}(\mathcal{G})$. The facets of $\mathcal{B}$ are the maximal subsets of $\texttt{vis}(\mathcal{G})$ that are children of the same latent node in $\tilde{\mathcal{G}}$, i.e., for each $u \in \texttt{lat}(\tilde{\mathcal{G}})$, the set $\texttt{ch}_{\tilde{\mathcal{G}}}(u)$ is a facet of $\mathcal{B}$.*

The correspondence between a pDAG and its associated mDAG is exemplified in Fig. 2.1.3: each latent node u in the pDAG corresponds to a facet in the mDAG whose elements are the children of u in the pDAG. One can conceptualize the mDAG as a hypergraph with two types of edges, namely, the directed edges and a set of undirected hyperedges, where the latter represent the facets of the simplicial complex and are depicted by red loops.[8]

We now define the notion of *structural dominance* of one mDAG over another.

[8]Note that mDAGs are different from other hypergraph constructions, such as acyclic directed mixed graphs (ADMGs) [61], insofar as mDAGs capture information about *all* of the causal-compatibility constraints of the original pDAG, and not only its conditional independence constraints. For example, the two mDAGs of Fig. 2.1.4 are associated with the same ADMG, because they have the same set of conditional independence constraints (in this case, the empty set). Graphically, while the hyperedges of an ADMG include only two nodes each, the hyperedges of the mDAG (i.e., facets of the simplicial complex) can include any number of nodes.

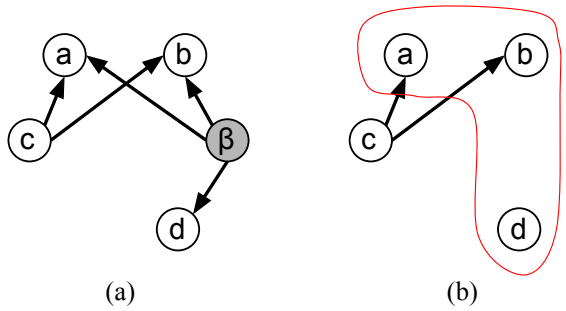


Figure 2.1.3: (a) A RE-reduced pDAG. (b) The mDAG associated with (a). This mDAG has simplicial complex $\mathcal{B} = \{\{a\}, \{b\}, \{c\}, \{d\}, \{a, b\}, \{a, d\}, \{b, d\}, \{a, b, d\}\}$. This is the mDAG associated with all of the pDAGs of Fig. 2.1.2. The inclusion-maximal element of $\mathcal{B}$ are indicated by the red loop on the mDAG.

Definition 10 (Structural Dominance of mDAGs)**.** *Let $\mathfrak{G}$ and $\mathfrak{G}'$ be two mDAGs with the same sets of nodes. $\mathfrak{G}$ is said to* structurally dominate *$\mathfrak{G}'$ if the following pair of conditions hold: (i) the directed structure of $\mathfrak{G}'$ can be obtained from the directed structure of $\mathfrak{G}$ by dropping edges, $\texttt{DirectedEdges}(\mathfrak{G}') \subseteq \texttt{DirectedEdges}(\mathfrak{G})$, and (ii) the simplicial complex of $\mathfrak{G}'$ can be obtained from the simplicial complex of $\mathfrak{G}$ by dropping faces, $\mathcal{B}' \subseteq \mathcal{B}$.*

Note that the definition of structural dominance requires a subset inclusion relation for the *faces*, not the *facets* of the simplicial complexes. For example, in Fig. 2.1.4, the set of facets of the simplicial complex of the mDAG 2.1.4(a) is not a subset of the set of facets of the simplicial complex of 2.1.4(b), but the set of *faces* do stand in a relation of subset inclusion to one another, so that mDAG 2.1.4(b) structurally dominates mDAG 2.1.4(a).

It is convenient to define a map that takes an mDAG to the RE-reduced pDAG that yields it. We will call this the *canonical pDAG* associated with the mDAG:

Definition 11 (Canonical pDAG)**.** *Let $\mathfrak{G} = (\mathcal{D}, \mathcal{B})$ be an mDAG. The map* can *is given by the following procedure: starting from $\mathcal{D}$, add one latent node l for each* facet *$A \in \mathcal{B}$ and add edges from l to each element of A so that $\texttt{ch}(l) = A$. The resulting pDAG, $\texttt{can}(\mathfrak{G})$, is the canonical pDAG associated with $\mathfrak{G}$.*

Consider the canonical pDAG associated with the mDAG of Fig. 2.1.4(a), which is represented in Fig. 2.1.5(a), and the canonical pDAG associated with the mDAG of Fig. 2.1.4(b), which is represented in Fig. 2.1.5(b). Now, construct the pDAG of Fig. 2.1.5(c) by starting from Fig. 2.1.5(b) and adding to it all of the latent nodes of Fig. 2.1.5(a), as well as their connections to visible nodes. Clearly, Fig. 2.1.5(c) and Fig. 2.1.5(b) are observationally

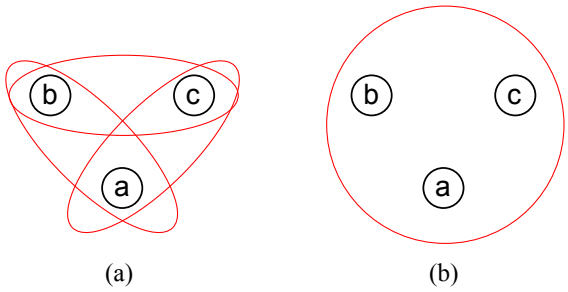


Figure 2.1.4: (a) mDAG with trivial directed structure and with simplicial complex $\mathcal{B} = \{\{a\},\{b\},\{c\},\{a,b\},\{a,c\},\{b,c\}\}$. (b) mDAG with trivial directed structure and with simplicial complex $\mathcal{B}' = \{\{a\},\{b\},\{c\},\{a,b\},\{a,c\},\{b,c\},\{a,b,c\}\}$.The facets (inclusion-maximal elements of $\mathcal{B}$ and $\mathcal{B}'$) are indicated by red loops.

equivalent by Lemma 2, because all of the added latent nodes are redundant. The pDAG of Fig. 2.1.5(a) is a subgraph (in the sense of Definition 2) of Fig. 2.1.5(c).

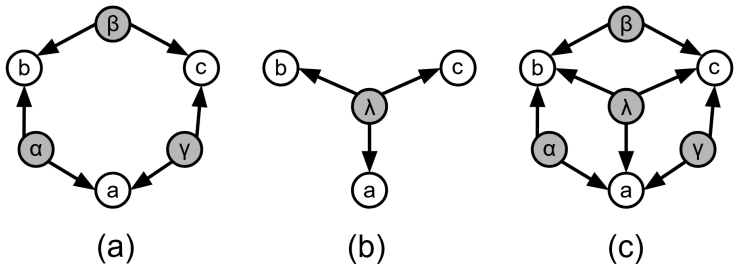


Figure 2.1.5: (a) Canonical pDAG of Fig. 2.1.4(a). (b) Canonical pDAG of Fig. 2.1.4(b). (c) pDAG obtained by adding the latent nodes of (a) to (b). This is not a canonical pDAG. By Lemma 2, (b) and (c) are observationally equivalent.

This observation is generic: Let $\mathfrak{G}$ and $\mathfrak{G}'$ be mDAGs such that $\mathfrak{G}'$ is structurally dominated by $\mathfrak{G}$. Construct the pDAG $\bar{\mathcal{G}}$ by starting from $\mathtt{can}(\mathfrak{G})$ and adding to it all of the latent nodes of $\mathtt{can}(\mathfrak{G}')$, as well as the corresponding edges connecting them to visible nodes. Then, $\mathtt{can}(\mathfrak{G}')$ is a subgraph of $\bar{\mathcal{G}}$.

Therefore, Lemma 2 together with the well-known fact that dropping edges from a pDAG results in an observationally dominated pDAG (see, for example, Proposition 3.3(b) in Ref. [33] or Theorem 26.1 in Ref. [47]) imply that:

Lemma 3 (Structural dominance ⇒ Observational dominance [33, 47])**.** *Let* $\mathfrak{G}$ *and* $\mathfrak{G}'$ *be two mDAGs with the same sets of nodes. If* $\mathfrak{G}$ *structurally dominates* $\mathfrak{G}'$*, then it also*

observationally dominates it, i.e., $\mathfrak{G} \succeq \mathfrak{G}'$.

In Sec. 2.4, it will be shown that the opposite implication does *not* hold. See, e.g., Figs. 2.4.1 and 2.4.2. Furthermore, it will be shown that structural dominance *also* implies dominance relative to informationally complete probing schemes.

The results and definitions that we presented for pDAGs are extended to mDAGs through their associated canonical pDAG. When there is a latent confounder between nodes a and b in the canonical pDAG, i.e., when the simplicial complex of the mDAG has at least one facet that includes nodes a and b, then we say that there is a *latent confounder* between a and b in the mDAG. When the directed structure of $\mathfrak{G}$ is trivial in the sense of containing no edges, then $\mathfrak{G}$ is said to be *directed-edge-free*. The mDAGs of Fig. 2.1.4 are examples of directed-edge-free mDAGs. Similarly, when the simplicial complex of $\mathfrak{G}$ is trivial in the sense of all its facets being singleton sets, then $\mathfrak{G}$ is said to be *latent-confounder-free*. Clearly, if the pDAG $\mathcal{G}$ is latent-confounder-free, then the mDAG $\texttt{ToFaces}(\mathcal{G})$ is latent-confounder-free. Finally, a distribution is realizable by an mDAG if it is realizable by its canonical pDAG, an mDAG observationally dominates another if the observational dominance relation holds for their canonical pDAGs, and so on. A final definition concerning mDAGs that will be useful is presented below:

Definition 12 (Subgraph of an mDAG over a subset of nodes)**.** *Let* $\mathfrak{G}$ *be an mDAG, and let* $S \subseteq \texttt{nodes}(\mathfrak{G})$ *be a subset of its nodes. The subgraph of* $\mathfrak{G}$ *over* S*, denoted* $\mathfrak{G}_S$*, is the mDAG obtained by starting from* $\mathfrak{G}$ *and deleting all of the nodes in* $\texttt{nodes}(\mathfrak{G}) \setminus S$.

Note that $\mathfrak{G}_S$ is the same as the mDAG associated with the pDAG $\texttt{can}(\mathfrak{G})_{\texttt{lat}(\texttt{can}(\mathfrak{G}))\cup S}$, that is, the subgraph of $\texttt{can}(\mathfrak{G})$ obtained by deleting all of the *visible* nodes that are not in S.

2.1.4 Restriction of scope to sets of mDAGs consistent with a fixed nodal ordering

Consider the following definition:

Definition 13 (Set of pDAGs consistent with a fixed nodal ordering)**.** *Consider a set* S *of pDAGs, all of which share the same set of visible nodes* $\{v_1, ..., v_n\}$*. We say that* S *is a* set of pDAGs consistent with a fixed nodal ordering *if there is an ordering of the nodes,* $(v_{\Pi(1)}, ..., v_{\Pi(n)})$ *for some permutation* Π*, such that every pDAG of* S *is such that* $v_{\Pi(j)}$ *is not an ancestor of* $v_{\Pi(i)}$ *for all* i, j *such that* $\Pi(i) < \Pi(j)$.

In the classification of causal structures made in Chapters 2 and 3, we will take the nodal ordering to be fixed. That is, we will only compare candidate causal structures for which the visible nodes are consistent with a fixed nodal ordering.

This stems from the assumption that we know how the visible variables are ordered in time. Suppose that two visible variables, X_a and X_b, are such that in each sample of the ensemble, the variable X_a is associated to a property of a system at one time and the variable X_b is associated to a property of a system (possibly the same system) at a *later* time. In this case, the only relevant candidate causal structures are those that are consistent with the fixed nodal ordering (a, b). That is, although we might be unsure of whether the underlying causal structure has an arrow $a \to b$ or not, we can nonetheless be sure that there is *no* arrow $b \to a$, under the assumption that there is no backwards-in-time causation.

There are several reasons why we restrict the set of causal structures appearing in our classification in this way. Clearly, if the variables of interest are temporally localized (and we do not need to take into account relativity theory, such as the possibility of space-like separation), then we should restrict our search to causal structures compatible with a fixed nodal ordering. The reason typically given for why fixed nodal ordering should *not* be presumed [57, Section 7.5.1] is that the variables of interest may fail to be temporally localized, referring instead to properties that persist over time. An individual's health and their exercise regime provides a good example. However, whenever the variables of interest are of this type, they can have a mutual influence on one another, and consequently one cannot restrict attention to directed graphs that are *acyclic* when considering the possibilities for the causal structure holding between them. Thus, if one is contemplating a causal discovery algorithm that returns only acyclic graphs (as we are doing here), then one should not apply it to variables that fail to be temporally localized. A second reason for restricting attention to variables that are temporally localized is that most probing schemes only make sense for such variables. In particular, it is difficult to make sense of the Observe&Do probing scheme described in Section 2.4 for a variable that refers to a property that persists over time.

One could still argue that, although the visible variables are temporally localized, for some reason we do not know the temporal order in which they were probed. However, one needs to have some knowledge about the probing scheme to be confident about its nature (for example, whether it consists purely of passive observations or it involves interventions). Situations where this is the case but one remains ignorant of when each variable was probed are likely to be the exception rather than the rule.

Therefore, in this work we will only consider causal structures consistent with a fixed

nodal ordering. Fig. 2.1.6 shows an example of two causal structures, a fork and a chain, that are consistent with the nodal ordering (a, b, c). Note that they are observationally *in*equivalent. (There exist pairs of causal structures, one of the "fork" type and the other of the "chain" type, that *are* observationally equivalent, but this requires the parent node in the fork to correspond to the middle node in the chain, and thus they would not be consistent with a fixed nodal ordering.)

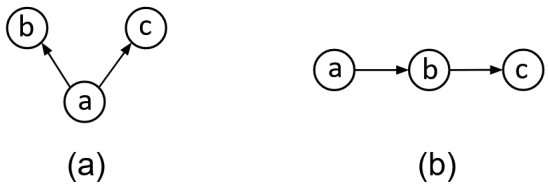


Figure 2.1.6: (a) Causal structure of the "fork" type that is consistent with the nodal ordering (a, b, c). (b) Causal structure of the "chain" type that is consistent with the nodal ordering (a, b, c). These two causal structures are *not* observationally equivalent.

2.1.5 3-pDAGs

To end this section, we will introduce an object that will be useful later on. Imagine a setup where the values of some of the variables are not sampled from a distribution that arises from natural causal mechanisms, but are instead set by an experimentalist. These variables can be conceptualized as *inputs*, whereas conventional variables (visible or latent) are outputs. One can think, for example, of experiments involving interventions, or simply of experiments where some variables are settings that we choose, such as in an experimental test of Bell inequalities [23, 9]. To describe the causal mechanism of such an experimental scenario, we need another structure, that allows for a third type of node. This structure will be called a *3-pDAG*:

Definition 14 (3-Partitioned DAG)**.** *A* 3-partitioned DAG (3-pDAG) $\mathscr{G}$ *is a DAG together with a partition of its nodes into* three *subsets: the set of* visible *nodes, denoted* $\texttt{vis}(\mathscr{G}) \subseteq \texttt{nodes}(\mathscr{G})$*, the set of* latent *nodes, denoted* $\texttt{lat}(\mathscr{G}) \subseteq \texttt{nodes}(\mathscr{G})$*, and the set of* input *nodes, denoted* $\texttt{Inodes}(\mathscr{G}) \subseteq \texttt{nodes}(\mathscr{G})$*. Input nodes are exogenous nodes.*

The visible and latent nodes have the same meaning as in pDAGs. The input nodes will be associated with *input variables*, the variables that are set by an experimenter rather than sampled from a distribution. As such, the 3-pDAG defines a conditional

probability distribution where the variables associated to input nodes are on the right of the conditional. In a 3-pDAG $\mathscr{G}$, the subset of parents of a node a that are input nodes will be denoted $\mathtt{Ipa}_{\mathscr{G}}(a)$. In the depictions of 3-pDAGs in this article, visible and latent nodes will continue to be represented by circles with white and grey backgrounds respectively, while input nodes will be represented as squares.

Similar to the notion of observational realizability of a probability distribution for pDAGs, we can define the *observational realizability of a conditional distribution by a 3-pDAG*. Here, the input variables always appear on the right of the conditional. A conditional probability distribution $P(X_{\mathtt{vis}(\mathscr{G})}|X_{\mathtt{Inodes}(\mathscr{G})})$ is said to be *observationally*[9] *realizable* by a 3-pDAG $\mathscr{G}$ if there exist parameters

$$\mathtt{par} = \{(f_a, P(E_a)) : a \in \mathtt{nodes}(\mathscr{G}) \setminus \mathtt{Inodes}(\mathscr{G})\} \tag{2.7}$$

such that, after setting Eq. (2.1) for all nodes $a \in \mathtt{vis}(\mathscr{G}) \cup \mathtt{lat}(\mathscr{G})$, one obtains a conditional probability distribution $Q(X_{\mathtt{vis}(\mathscr{G})\cup\mathtt{lat}(\mathscr{G})}|X_{\mathtt{Inodes}(\mathscr{G})})$, i.e.,

$$\begin{aligned} &Q(X_{\mathtt{vis}(\mathscr{G})\cup\mathtt{lat}(\mathscr{G})}|X_{\mathtt{Inodes}(\mathscr{G})}) \\ &= \prod_{a\in\mathtt{vis}(\mathscr{G})\cup\mathtt{lat}(\mathscr{G})} \sum_{e\in\mathcal{E}_a} \delta_{X_a, f_a(X_{\mathtt{pa}_{\mathscr{G}}(a)}, E_a=e)} P(E_a = e) \end{aligned} \tag{2.8}$$

whose marginal over $X_{\mathtt{vis}(\mathscr{G})}$ is $P(X_{\mathtt{vis}(\mathscr{G})}|X_{\mathtt{Inodes}(\mathscr{G})})$.

Suppose the visible variables of a 3-pDAG $\mathscr{G}$ have cardinalities $\vec{c}_{\mathtt{vis}(\mathscr{G})}$ and the input variables have cardinalities $\vec{c}_{\mathtt{Inodes}(\mathscr{G})}$. The set of conditional probability distributions of the visible variables given the input variables that are observationally realizable by the 3-pDAG $\mathscr{G}$ is denoted by

$$\mathcal{M}_{\text{obs}}(\mathscr{G}, \vec{c}_{\mathtt{vis}(\mathscr{G})}, \vec{c}_{\mathtt{Inodes}(\mathscr{G})}).$$

In analogy to the definition for pDAGs, we can define a notion of observational equivalence for 3-pDAGs. We will say that two 3-pDAGs are *observationally equivalent* if, for any choice of cardinalities for the visible variables, their set of observationally realizable conditional distributions are the same Similarly, a 3-pDAG $\mathscr{G}$ *observationally dominates* another 3-pDAG $\mathscr{G}'$ if, for any choice of cardinalities for the visible variables, its set of observationally realizable distributions is a superset of that of $\mathscr{G}'$. In short, the definition of observational dominance and equivalence for 3-pDAGs is obtained from Definition 7 by

[9]The use of the term "observational" here indicates that the data of interest, i.e., the conditional distributions $P(\mathtt{vis}(\mathscr{G})|\mathtt{Inodes}(\mathscr{G}))$, are obtained by a probing scheme where the *visible variables* of the 3-pDAG are passively observed.

replacing every instance of the pDAGs $\mathcal{G}, \mathcal{G}'$ by the 3-pDAGs $\mathscr{G}, \mathscr{G}'$, the realizable distributions $\mathcal{M}_{\text{obs}}(\mathcal{G}, \vec{c}_{\texttt{vis}(\mathcal{G})})$ by the realizable *conditional* distributions $\mathcal{M}_{\text{obs}}(\mathscr{G}, \vec{c}_{\texttt{vis}(\mathscr{G})}, \vec{c}_{\texttt{Inodes}(\mathscr{G})})$, and the universal quantification over the vector of cardinalities $\vec{c}_{\texttt{vis}(\mathcal{G})} \in \mathbb{N}^{|\texttt{vis}(\mathcal{G})|}$ by the universal quantification over both the vector of cardinalities $\vec{c}_{\texttt{vis}(\mathscr{G})} \in \mathbb{N}^{|\texttt{vis}(\mathscr{G})|}$ and the vector of cardinalities $\vec{c}_{\texttt{Inodes}(\mathscr{G})} \in \mathbb{N}^{|\texttt{Inodes}(\mathscr{G})|}$.

It turns out that one can easily leverage the characterization of observational dominance in pDAGs to obtain the characterization of observational dominance in 3-pDAGs.

Suppose $\mathscr{G}$ is a 3-pDAG. Now consider the operation of converting all of the input nodes of $\mathscr{G}$ to visible nodes, so that the result has only visible and latent nodes and consequently is a pDAG. Denote the map that achieves this conversion by `ConvertItoV`. Denote the image of the 3-pDAG $\mathscr{G}$ under the `ConvertItoV` map, which is a pDAG, by $\bar{\mathcal{G}}$:

$$\bar{\mathcal{G}} = \texttt{ConvertItoV}(\mathscr{G}).$$

Lemma 4. *Let $\mathscr{G}$ be a 3-pDAG and let $\bar{\mathcal{G}}$ be the pDAG one obtains from it by converting input nodes into visible nodes, i.e., $\bar{\mathcal{G}} = \texttt{ConvertItoV}(\mathscr{G})$. Let $\texttt{newVnodes}(\bar{\mathcal{G}})$ be the set of nodes that were input nodes in $\mathscr{G}$ but have become visible in $\bar{\mathcal{G}}$, and let $\texttt{oldVnodes}(\bar{\mathcal{G}})$ be the set of nodes that are visible in both $\mathscr{G}$ and $\bar{\mathcal{G}}$. That is, $\texttt{newVnodes}(\bar{\mathcal{G}}) = \texttt{Inodes}(\mathscr{G})$, $\texttt{oldVnodes}(\bar{\mathcal{G}}) = \texttt{vis}(\mathscr{G})$, and $\texttt{vis}(\bar{\mathcal{G}}) = \texttt{newVnodes}(\bar{\mathcal{G}}) \cup \texttt{oldVnodes}(\bar{\mathcal{G}})$.*

A conditional distribution $P(X_{\texttt{vis}(\mathscr{G})}|X_{\texttt{Inodes}(\mathscr{G})})$ is realizable by $\mathscr{G}$ if and only if the corresponding conditional distribution $P(X_{\texttt{oldVnodes}(\bar{\mathcal{G}})}|X_{\texttt{newVnodes}(\bar{\mathcal{G}})})$ is realizable by $\bar{\mathcal{G}}$.

Proof. If the conditional distribution $P(X_{\texttt{vis}(\mathscr{G})}|X_{\texttt{Inodes}(\mathscr{G})})$ is observationally realizable by $\mathscr{G}$, then for every product probability distribution $\prod_{a\in\texttt{newVnodes}(\bar{\mathcal{G}})} P(X_a)$, the joint distribution $P(X_{\texttt{vis}(\bar{\mathcal{G}})}) = P(X_{\texttt{oldVnodes}(\bar{\mathcal{G}})}|X_{\texttt{newVnodes}(\bar{\mathcal{G}})}) \prod_{a\in\texttt{newVnodes}(\bar{\mathcal{G}})} P(X_a)$ is observationally realizable by $\bar{\mathcal{G}}$. This follows from the fact that every node $a \in \texttt{newVnodes}(\bar{\mathcal{G}})$ is *exogenous* in $\bar{\mathcal{G}}$, so that the only restriction on the joint distribution over the collection of X_a is that it be a product distribution.

Similarly, if the joint distribution $P(X_{\texttt{vis}(\bar{\mathcal{G}})})$ is observationally realizable on $\bar{\mathcal{G}}$, then because the nodes in $\texttt{newVnodes}(\bar{\mathcal{G}})$ are all exogenous, it follows that the marginal distribution on $\texttt{newVnodes}(\bar{\mathcal{G}})$ factorizes, i.e., $P(X_{\texttt{newVnodes}(\bar{\mathcal{G}})}) = \prod_{a\in\texttt{newVnodes}(\bar{\mathcal{G}})} P(X_a)$. Consequently, $P(X_{\texttt{vis}(\bar{\mathcal{G}})})$ can be expressed as

$$\begin{aligned} P(X_{\texttt{vis}(\bar{\mathcal{G}})}) = P(X_{\texttt{oldVnodes}(\bar{\mathcal{G}})}|X_{\texttt{newVnodes}(\bar{\mathcal{G}})}) \\ \times \prod_{a\in\texttt{newVnodes}(\bar{\mathcal{G}})} P(X_a). \end{aligned} \tag{2.9}$$

Consider any conditional $P\big(X_{\texttt{oldVnodes}(\bar{\mathcal{G}})}\big|X_{\texttt{newVnodes}(\bar{\mathcal{G}})}\big)$ that can be obtained from a joint distribution $P\big(X_{\texttt{vis}(\bar{\mathcal{G}})}\big)$ that is observationally realizable in $\bar{\mathcal{G}}$ by dividing by some product distribution $\prod_{a\in\texttt{newVnodes}(\bar{\mathcal{G}})} P(X_a)$. Clearly, it can be used to define a conditional $P\big(X_{\texttt{Vnodes}(\mathscr{G})}\big|X_{\texttt{Inodes}(\mathscr{G})}\big)$ which is observationally realizable in $\mathscr{G}$. □

Finally, we define the notion of a *3-mDAG*, which stands to the notion of an mDAG in the same manner in which the notion of a 3-pDAG stands to the notion of a pDAG. That is, a 3-mDAG is defined similarly to an mDAG but where the set of nodes can include input nodes (defined in the text below Definition 14) in addition to visible nodes. These input nodes have no parents in the directed structure, nor can they be part of any nontrivial face of the simplicial complex. Structural dominance of 3-mDAGs can be defined precisely as it was for mDAGs (see Definition 10).

Lemma 5. *Let $\mathbb{G}$ and $\mathbb{G}'$ be two 3-mDAGs. If $\mathbb{G}$ structurally dominates $\mathbb{G}'$ then $\mathbb{G}$ observationally dominates $\mathbb{G}'$.*

Proof. Follows straightforwardly from Lemmas 3 and 4. □

Just as we defined the map $\texttt{ToFaces}(\cdot)$ from a pDAG to its associated mDAG in Definition 9, one can define the map that takes a 3-pDAG to its associated 3-mDAG in an analogous way. We will again denote this map by $\texttt{ToFaces}(\cdot)$.

2.2 Full-SWIGs

Suppose now that we have the experimental power to intervene on a variable X_a, in such a way that the value of X_a that arises from natural causal mechanisms is not necessarily the same as the value of X_a that will influence the descendants of the node a in the underlying causal structure. These two values will be kept in two new variables that we define, called $X_{a^\flat}$ and $X_{a^\sharp}$. The variable $X_{a^\flat}$ corresponds to the natural value of X_a, while the variable $X_{a^\sharp}$ corresponds to the value that will influence the descendants of a.

We can represent those variables in a 3-pDAG that accounts for the possibility of an intervention on a variable X_a. There, the variable $X_{a^\flat}$ will be associated with a visible node that has no children, and the variable $X_{a^\sharp}$ will be associated with an input node, whose information is sent to the nodes that were descendants of a in the original pDAG. An example is provided in Fig. 2.2.1.[10]

[10]Note that, if the nodes a and b of Fig. 2.2.1(a) are associated respectively with a treatment variable

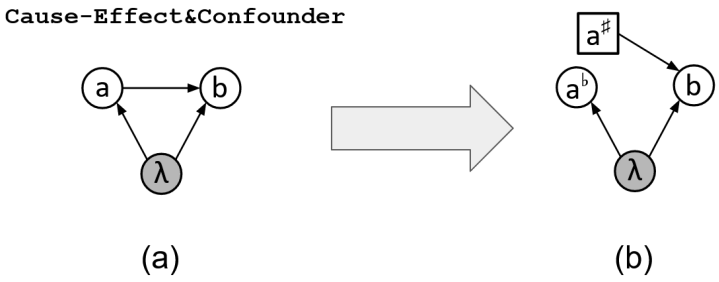


Figure 2.2.1: (a) The pDAG `Cause-Effect&Confounder`. (b) The SWIG that represents the possibility of intervening on the node a of `Cause-Effect&Confounder`.

A 3-pDAG that is obtained by splitting one or more visible nodes of a pDAG into $\flat$ and $\sharp$ versions, like the one in Fig. 2.2.1, is called a *single world intervention graph* (SWIG) [64] [11] We will denote the SWIG obtained from $\mathcal{G}$ by splitting all the nodes in a particular subset $A \subseteq \texttt{vis}(\mathcal{G})$ by $\text{SWIG}_A(\mathcal{G})$. The SWIG generated by a pDAG $\mathcal{G}$ by splitting *all* visible nodes, $\text{SWIG}_{\texttt{vis}(\mathcal{G})}(\mathcal{G})$, will be called a *full-SWIG*, and will be denoted by $\texttt{split}(\mathcal{G})$. Fig. 2.2.2 presents an example of a full-SWIG. Just as the $\flat$ and $\sharp$ versions of an individual node a will be represented by $a^\flat$ and $a^\sharp$ respectively, the set of $\flat$ and $\sharp$ versions of the nodes in a set S will be represented by $S^\flat$ and $S^\sharp$ respectively .

We will also define a map that takes an mDAG and returns the 3-mDAG where all of the visible nodes of the original mDAG are split into $\flat$ and $\sharp$ versions. As well as for the case of pDAGs, this map will be denoted by `split`. In other words, we are extending the domain of the map `split`: now it can take pDAGs to 3-pDAGs, or mDAGs to 3-mDAGs. An example of the application of the map `split` on an mDAG is given in Fig. 2.2.3, which

and a recovery variable, then the conditional distribution $P(X_{a^\flat} X_b | X_{a^\sharp})$ that is realized by the SWIG of Fig. 2.2.1(b) with the parameters of the causal hypothesis corresponds to the *Effect of the Treatment on the Treated (ETT)* [70].

[11] The concept of a SWIG can also be useful when formulated using frameworks for causal modelling wherein causal structures are represented by circuits, i.e., each variable is represented by a wire and each causal mechanism is represented by a gate [66, 54]. A latent variable is represented by a wire that is internal to the circuit. The act of passively observing a variable is represented by applying a copy operation on the variable, which in the circuit implies creating a new wire to represent the copy and letting this be an open wire that is an output of the circuit. Conversely, an input variable is simply represented by an open wire that is an input to the circuit. It follows that the circuit framework has a particularly simple manner of representing the node-splitting operation of a visible variable: cut the wire corresponding to the copy of the variable that feeds forward to other variables in the causal model, and marginalize over the open output wire one thereby creates. In this way, one preserves the open output wire that was already present and one creates a new open input wire.

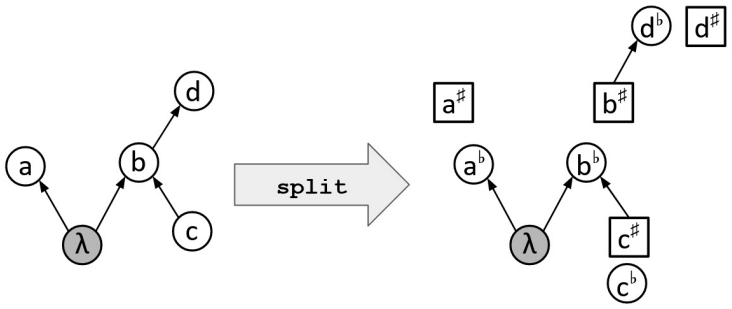


Figure 2.2.2: Example of the operation of splitting all visible nodes of a pDAG $\mathcal{G}$ to obtain the corresponding full-SWIG $\mathscr{G} = \texttt{split}(\mathcal{G})$.

starts from the mDAG that is associated with the pDAG of Fig. 2.2.2.

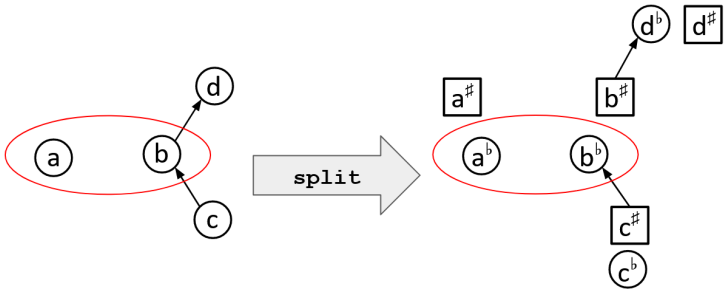


Figure 2.2.3: Example of the operation of splitting all nodes of an mDAG $\mathfrak{G}$ to obtain the corresponding 3-mDAG $\mathbb{G} = \texttt{split}(\mathfrak{G})$.

As it turns out, it does not matter whether we first split all the visible nodes of a pDAG $\mathcal{G}$ and then take the 3-mDAG of the resulting 3-pDAG $\texttt{split}(\mathcal{G})$, or if we first take the mDAG associated to $\mathcal{G}$ and then split all of the visible nodes. This is formalized in the following lemma.

Lemma 6. *The maps* `split` *and* `ToFaces` *commute. That is, for a pDAG* $\mathcal{G}$*, we have*

$$\begin{aligned}&\texttt{split}(\texttt{ToFaces}(\mathcal{G}))\\ &= \texttt{ToFaces}(\texttt{split}(\mathcal{G})).\end{aligned} \tag{2.10}$$

Proof. Given in Appendix A.1. □

For full-SWIGs, one asks about the realizability of a given conditional probability distribution

$$P\big(X_{\mathtt{vis}(\mathcal{G})^\flat} \big| X_{\mathtt{vis}(\mathcal{G})^\sharp}\big). \tag{2.11}$$

To answer this realizability question, we need to specify the possible choices of parameters on the full-SWIG, that is, the scope of causal hypotheses. Recall that any 3-pDAG has parameters as specified in Eq. (2.7), that is, functions and distributions over error variables for all the nodes that are visible. The scope of possible parameter values for the full-SWIG is presumed to be inherited from the pDAG from which the full-SWIG was defined. Specifically, a variable $X_{a^\flat}$ depends functionally on its parents and the local error variable $E_{a^\flat}$ in the same way that X_a depended on its parents and E_a, and the distribution over the error variable is the same. That is, for each causal hypothesis $(\mathcal{G}, \mathtt{par})$ where $\mathtt{par} = \{f_a, P(E_a) : a \in \mathtt{vis}(\mathcal{G})\}$, we have a corresponding causal hypothesis for the full-SWIG $\mathtt{split}(\mathcal{G})$, namely,

$$\mathtt{par}^{\mathtt{split}} = \{f_{a^\flat}, P(E_{a^\flat}) : a^\flat \in \mathtt{vis}(\mathcal{G})^\flat\} \tag{2.12}$$

where $f_{a^\flat} = f_a$ and $P(E_{a^\flat}) = P(E_a)$. Note that there are no error variables associated with the $\sharp$ nodes because these are input nodes.

As noted earlier, there are instances of observational dominance of pDAGs without structural dominance, that is, that the converse of Lemma 3 does not hold. By contrast, the observational dominance of full-SWIGs that have the same set of visible and input nodes *can* be fully characterized by structural dominance:

Lemma 7. *Let $\mathscr{G}$ and $\mathscr{G}'$ be a pair of full-SWIGs having the same set of visible and of input nodes, $\mathtt{vis}(\mathscr{G}) = \mathtt{vis}(\mathscr{G}')$ and $\mathtt{Inodes}(\mathscr{G}) = \mathtt{Inodes}(\mathscr{G}')$, and let $\mathbb{G}$ and $\mathbb{G}'$ denote the corresponding 3-mDAGs, i.e., $\mathbb{G} = \mathtt{ToFaces}(\mathscr{G})$ and $\mathbb{G}' = \mathtt{ToFaces}(\mathscr{G}')$.*

Then, the full-SWIG $\mathscr{G}$ observationally dominates the full-SWIG $\mathscr{G}'$ if and only if the 3-mDAG $\mathbb{G}$ structurally dominates the 3-mDAG $\mathbb{G}'$.

Proof. The "if" side follows from Lemma 5, so we only need to prove the "only if" side. We will proceed by proving the contrapositive. Suppose that $\mathscr{G}$ observationally dominates $\mathscr{G}'$ even though $\mathbb{G}$ does not structurally dominate $\mathbb{G}'$. The lack of structural dominance implies one of two possibilities:

1. There is at least one directed edge between an input node $a^\sharp$ and a visible node $b^\flat$, $a^\sharp \to b^\flat$, that is present in $\mathbb{G}'$ but not in $\mathbb{G}$, or

2. $\mathbb{G}$ has all the directed edges that are present in $\mathbb{G}'$, but there is at least one set $S^{\flat}$ of visible nodes which is a face of $\mathbb{G}'$ but not of $\mathbb{G}$.

These two conditions for 3-mDAGs respectively imply the following conditions for the corresponding 3-pDAGs:

1. There is at least input node $a^{\sharp}$ and one visible node $b^{\flat}$ such that $\mathscr{G}'$ presents a chain $a^{\sharp} \to m_1 \to ... \to m_n \to b^{\flat}$ where all of the mediary nodes $m_1, ..., m_n$ are latent nodes, and there is *no* such chain in $\mathscr{G}$, or

2. Item 1 does not hold, but there is at least one set $S^{\flat}$ of visible nodes that have a common latent ancestor λ in $\mathscr{G}'$, but no such latent ancestor exists in $\mathscr{G}$.

In Case 1, consider a conditional distribution where the value of $X_{b^{\flat}}$ depends on the value of $X_{a^{\sharp}}$, that is, there exist values $x_{a^{\sharp}}, x'_{a^{\sharp}} \in \mathcal{X}_{a^{\sharp}}$ such that

$$P\big(X_{b^{\flat}} \big| X_{a^{\sharp}} = x_{a^{\sharp}}\big) \neq P\big(X_{b^{\flat}} \big| X_{a^{\sharp}} = x'_{a^{\sharp}}\big). \tag{2.13}$$

Such a conditional distribution can clearly be realized by $\mathscr{G}'$ given the presence of a chain from $a^{\sharp}$ to $b^{\flat}$. However, it cannot be realized by $\mathscr{G}$. This is because there is no chain between $a^{\sharp}$ and $b^{\flat}$, and also no latent common cause of $a^{\sharp}$ and $b^{\flat}$. This can be seen as follows. Since $a^{\sharp}$ is a $\sharp$ node of the full-SWIG $\mathscr{G}$, it must be *parentless*, so there cannot be a common cause between $a^{\sharp}$ and $b^{\flat}$ nor a chain from $b^{\flat}$ to $a^{\sharp}$. By assumption, there is no chain from $a^{\sharp}$ to $b^{\flat}$ where all of the mediary nodes are latent; however, in a full-SWIG all of the nodes that are not latent are either parentless ($\sharp$ nodes) or childless ($\flat$ nodes), so chains with visible or input mediary nodes are also not allowed.

In case 2, consider a conditional distribution where all of the variables of $X_{S^{\flat}}$ are perfectly correlated; that is

$$P\big(X_{S^{\flat}} \big| X_{\texttt{Inodes}(\mathscr{G})}\big) = p[0, ..., 0]_{S^{\flat}} + (1-p)[1, ..., 1]_{S^{\flat}} \tag{2.14}$$

where $p \in [0, 1]$. In the right hand side, $p[0, ..., 0]_{S^{\flat}}$ indicates that with probability p all the variables in the set $X_{S^{\flat}}$ take the value 0, and $(1-p)[1, ..., 1]_{S^{\flat}}$ indicates that with probability $1-p$ all the variables in the set $X_{S^{\flat}}$ take the value 1. The probabilities p could depend on the value of $X_{\texttt{Inodes}(\mathscr{G})}$ that we are conditioning upon; all that matters is that all of the variables in $X_{S^{\flat}}$ are perfectly correlated.

Such a conditional distribution can clearly be realized by $\mathscr{G}'$, given the presence of a common cause λ between all of the nodes of $S^{\flat}$. However, it cannot be realized by $\mathscr{G}$.

Ref. [75, Example 2] says that, in a pDAG, a set of variables can only be perfectly correlated if all of the corresponding nodes share a common cause[12]. Via Lemma 4, this result can be straightforwardly extended to 3-pDAGs: a set of visible nodes of a 3-pDAG can only be perfectly correlated given the input nodes if they share a common cause that is latent or visible (an input node cannot act as the common cause that establishes perfect correlation, since all input variables are conditioned upon in the data). By assumption, the nodes of the set $S^{\flat}$ do not all share a latent common cause in $\mathscr{G}$. They also cannot share a visible common cause since, in full-SWIGs, visible nodes do not have children. Thus, $\mathscr{G}$ cannot realize the conditional distribution of Eq. (2.14).

Therefore, in both cases where $\mathbb{G}$ does not structurally dominate $\mathbb{G}'$, we have explicitly shown conditional probability distributions that are realizable by $\mathscr{G}'$ but not by $\mathscr{G}$.

□

2.3 Probing Schemes and Shadows

In this thesis, we study different probing schemes on visible variables with the goal of adjudicating between causal structures.

The most we can hope to learn about a causal hypothesis $(\mathcal{G}, \texttt{par})$ by interacting with the visible variables via different probing schemes[13] is expression (2.11), that is, the conditional distribution of the $\flat$ variables given the $\sharp$ variables generated by the causal hypothesis. When a probing scheme is able to provide us the whole conditional distribution $P(X_{\texttt{vis}(\mathcal{G})^{\flat}} | X_{\texttt{vis}(\mathcal{G})^{\sharp}})$, we say it is *informationally complete.* In Section 2.4, we will discuss the distinguishability of causal structures under access to informationally complete probing schemes.

After characterizing the informationally complete probing schemes, one can also ask what can be done with probing schemes that are *not* informationally complete. To describe them, we start by conceptualizing the conditional probability distribution $P(X_{\texttt{vis}(\mathcal{G})^{\flat}} | X_{\texttt{vis}(\mathcal{G})^{\sharp}})$ as a vector, where each component is given by a probability corresponding to a choice of

[12] Correlation could appear without a common cause if we were conditioning on a common effect. However, this is not the case: the nodes that are being conditioned upon are all input nodes, hence parentless, hence not candidates for a common effect of other nodes.

[13] In this paper we do *not* consider edge interventions, which are interventions where one sends a different value of the intervened variable to each of the children of its corresponding node. In other words, when a variable X_a is intervened upon, we will consider that all of the children of a receive information about the same variable $X_{a^{\sharp}}$.

values of $X_{\texttt{vis}(\mathcal{G})^\flat}$ and $X_{\texttt{vis}(\mathcal{G})^\sharp}$. A probing scheme that is not informationally complete will not give us the full conditional distribution $P(X_{\texttt{vis}(\mathcal{G})^\flat}|X_{\texttt{vis}(\mathcal{G})^\sharp})$, but only some of its components, or some functions of these components. The set of functions of the components of the full conditional distribution that can be obtained from a probing scheme is said to be the *shadow* of the full conditional that is revealed by this probing scheme.

To illustrate the notion of the shadow revealed by a probing scheme, we consider the example of the causal structure of Fig. 2.3.1 and two different probing schemes applied to it. To shorten our notation, here we will abbreviate the name of the pDAG `Cause-Effect&Confounder` to `CE&C`.

Suppose that a certain phenomenon is explained by the pDAG `CE&C` with a specific set of functional parameters $\texttt{par} = [f_a, f_b, P(X_\lambda)]$[14], where f_a and f_b are functions

$$f_a : \mathcal{X}_\lambda \to \mathcal{X}_a \tag{2.15}$$
$$f_b : \mathcal{X}_\lambda \times \mathcal{X}_a \to \mathcal{X}_b. \tag{2.16}$$

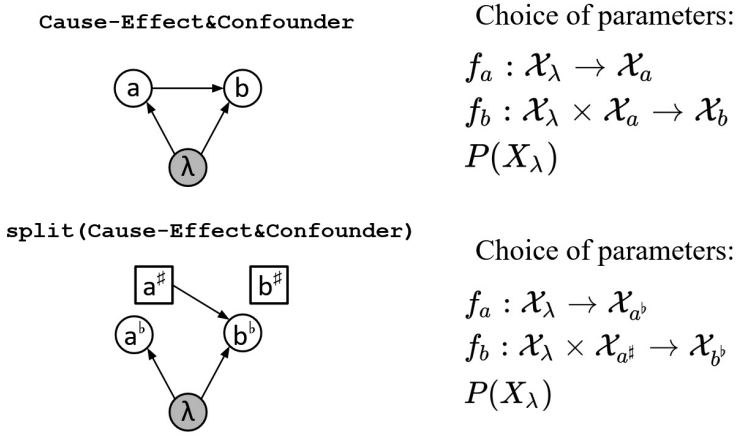


Figure 2.3.1: Example used to understand the notion of shadow of the full conditional distribution $P(X_{\texttt{vis}(\mathcal{G})^\flat}|X_{\texttt{vis}(\mathcal{G})^\sharp})$ that is revealed by a specific probing scheme.

The full-SWIG `split(CE&C)` is presented in the bottom part of Fig. 2.3.1. A choice of parameters for a full-SWIG corresponds to a choice of parameters for the original pDAG, because a variable $X_{a^\flat}$ in the full-SWIG reacts to its parents in the same way as X_a reacts to its parents in the original pDAG. The same choice of parameters `par`

[14]Strictly speaking, the `par` that we defined in Eq. (2.2) involves the distributions over error variables. However, here we have absorbed all of the error variables into the latent variable X_λ.

for this full-SWIG[15], as indicated in the figure, gives rise to the conditional distribution $P^{(\texttt{split(CE\&C)},\texttt{par})}(X_{a^\flat}X_{b^\flat}|X_{a^\sharp}X_{b^\sharp})$.

Consider first a probing scheme consisting of passive observations of X_a and X_b. It is easy to see that the observational probability distribution realized by the original pDAG `CE&C` under the choice of parameters `par` takes the same value as the conditional probability distribution realized by the full-SWIG `split(CE&C)` under the same choice of parameters `par` *when the values of the* $\sharp$ *variables coincide with the values of the corresponding* $\flat$ *variables*. That is,

$$P^{(\texttt{CE\&C},\texttt{par})}(X_a = x_a, X_b = x_b) = \\ P^{(\texttt{split(CE\&C)},\texttt{par})}(X_{a^\flat} = x_a, X_{b^\flat} = x_b | X_{a^\sharp} = x_a, X_{b^\sharp} = x_b). \tag{2.17}$$

Therefore, the shadow of the full conditional $P^{(\texttt{split(CE\&C)},\texttt{par})}(X_{a^\flat}X_{b^\flat}|X_{a^\sharp}X_{b^\sharp})$ that is revealed by passive observations of X_a and X_b is the set of components given by Eq. (2.17).

Consider now a different probing scheme, consisting of passive observation of X_b and a do-intervention on X_a, which is a procedure that forces X_a to take a specific value x_a'. The probability distribution over X_b obtained after performing this do-intervention on X_a is called a *do-conditional*, denoted by $P(X_b|\text{do}(X_a = x_a'))$.

The do-conditional obtained by applying this probing scheme on the causal hypothesis (`Cause-Effect&Confounder`, `par`), denoted by $P^{(\texttt{CE\&C},\texttt{par})}(X_b|\text{do}(X_a))$, equals the corresponding conditional distribution $P^{(\texttt{CE\&C},\texttt{par})}(X_{b^\flat}|X_{a^\sharp})$ realized by `split(CE&C)` under the same choice of parameters `par`. That is,

$$P^{(\texttt{CE\&C},\texttt{par})}(X_b = x_b|\text{do}(X_a = x_a')) = \\ \sum_{x_a} P^{(\texttt{split(CE\&C)},\texttt{par})}(X_{a^\flat} = x_a, X_{b^\flat} = x_b| \\ |X_{a^\sharp} = x_a', X_{b^\sharp} = x_b), \tag{2.18}$$

where the variables $X_{b^\flat}$ and $X_{b^\sharp}$ take the same value and we marginalize over the variable $X_{a^\flat}$. This equation gives the shadow of the full conditional $P^{(\texttt{split(CE\&C)},\texttt{par})}(X_{a^\flat}X_{b^\flat}|X_{a^\sharp}X_{b^\sharp})$ that is revealed by a passive observation of X_b together with a do-intervention that forces X_a to take the value x_a'.

[15]The error variables of $a^\flat$ and $b^\flat$ are similarly absorbed into λ, whose distribution is part of `par`. Note that $a^\sharp$ and $b^\sharp$, being input variables of a 3-pDAG, do not have associated error variables.

In summary, each probing scheme is associated with a *shadowing function*. This is the function that, given a conditional distribution $P(X_{\texttt{vis}(\mathcal{G})^\flat}|X_{\texttt{vis}(\mathcal{G})^\sharp})$, returns the shadow revealed by the probing scheme in question. In the example of Fig. 2.3.1, the shadowing function associated with passive observations of X_a and X_b corresponds to taking the components where $X_{a^\flat} = X_{a^\sharp}$ and $X_{b^\flat} = X_{b^\sharp}$. The shadowing function associated with passive observation of X_b and a do-intervention setting X_a to x'_a corresponds to taking the components where $X_{a^\sharp} = x'_a$ and $X_{b^\flat} = X_{b^\sharp}$, and marginalizing out $X_{a^\flat}$.

With this, we can now define a notion of realizability that extends to sets of data obtained from any probing scheme:

Definition 15 (Realizability of a Shadow)**.** *Let $\mathcal{G}$ be a pDAG, and let* par *be a choice of parameters for $\mathcal{G}$. Consider a probing scheme whose shadowing function is $\mathcal{F}_\mathcal{S}$, that is, a probing scheme that reveals the function $\mathcal{F}_\mathcal{S}$ of the full conditional distribution $P(X_{\texttt{vis}(\mathcal{G})^\flat}|X_{\texttt{vis}(\mathcal{G})^\sharp})$. The shadow $\mathcal{S}^{(\mathcal{G},\texttt{par})}$ obtained by applying this probing scheme on the causal hypothesis $(\mathcal{G},\texttt{par})$ is given by:*

$$\mathcal{S}^{(\mathcal{G},\texttt{par})} = \mathcal{F}_\mathcal{S}\left(P^{(\texttt{split}(\mathcal{G}),\texttt{par})}(X_{\texttt{vis}(\mathcal{G})^\flat}|X_{\texttt{vis}(\mathcal{G})^\sharp})\right). \tag{2.19}$$

Accordingly, a set $\mathcal{S}$ of data obtained from the probing scheme in question is said to be realizable by *$\mathcal{G}$ if there exists some choice of parameters* par *such that $\mathcal{S} = \mathcal{S}^{(\mathcal{G},\texttt{par})}$.*

We have previously defined $P^{(\mathcal{G},\texttt{par})}(X_{\texttt{vis}(\mathcal{G})})$ to be the observational distribution realized by a pDAG $\mathcal{G}$ under the choice of parameters par. In the definition above, we extend this superscript notation for shadows obtained from a particular probing scheme[16].

2.4 Equivalence and Dominance of causal structures under the Observe&Do Probing scheme

In this section, we will define and completely characterize the equivalence and dominance of causal structures when there is experimental access to an informationally complete probing scheme.

An example of an informationally complete probing scheme consists of what we call the *Observe&Do probing scheme*. An Observe&Do probe of a visible variable consists in the

[16]Note that we had already used this notation for the do-conditional of Eq. (2.18).

process of first recording the natural value of the variable, that will be indicated by the superscript $\flat$, and then forcing the variable to take a potentially different value, that will be indicated by the superscript $\sharp$. For concreteness, imagine the example of a drug trial, where the administration (or not) of a drug is encoded in a variable X_d, and the recovery (or not) is encoded in a variable X_r. The experimenter initially does not know if the drug causes recovery, but they definitely know that X_d comes temporally before X_r. In this case, the Observe&Do probe of the variable X_d consists of the following steps: first, the experimenter asks each subject if they would like to take the drug or not, and registers the value obtained in the variable $X_{d^\flat}$. Then, the experimenter intervenes on d, forcing $X_{d^\sharp}$ to take a value that is potentially different from the value of $X_{d^\flat}$. That is, they assign each subject to the drug or the placebo, with the assigned treatment potentially differing from the subject's expressed preference. After that, they will observe the value that the recovery variable exhibits (as a response to $X_{d^\sharp}$).

The Observe&Do *probing scheme*, which we will sometimes abbreviate to O&D probing scheme, is the name given to the probing scheme where *all* of the visible variables are Observe&Do probed. In the O&D probing scheme, implementation of the probings of the different variables follows the same temporal ordering as the variables themselves. That is, if X_a is temporally prior to X_b, then the Observe&Do probe of X_a is temporally prior to the Observe&Do probe of X_b. As we argued in Section 2.1.4, here we assume that the variables are probed in a fixed, known temporal order.

Suppose we have a pDAG $\mathcal{G}$ on which we perform the O&D probing scheme. The natural values of all the visible variables are stored in the set of variables $X_{\mathtt{vis}(\mathcal{G})^\flat}$, and the values that these variables are subsequently forced to take are stored in the set of variables $X_{\mathtt{vis}(\mathcal{G})^\sharp}$. The data obtained from this procedure takes the form $P\big(X_{\mathtt{vis}(\mathcal{G})^\flat} \big| X_{\mathtt{vis}(\mathcal{G})^\sharp}\big)$, which is exactly the conditional distribution of expression (2.11). Because the O&D probing scheme provides complete information about the conditional $P\big(X_{\mathtt{vis}(\mathcal{G})^\flat} \big| X_{\mathtt{vis}(\mathcal{G})^\sharp}\big)$, it provides the maximum information about the causal hypothesis that can be obtained by probing the visible nodes (assuming no edge interventions)). Therefore, it is an informationally complete probing scheme. While this section focuses on the Observe&Do probing scheme, its results are valid for any informationally complete probing scheme.

The set of conditional probability distributions $P\big(X_{\mathtt{vis}(\mathcal{G})^\flat} \big| X_{\mathtt{vis}(\mathcal{G})^\sharp}\big)$ for visible variables of cardinalities $\vec{c}_{\mathtt{vis}(\mathcal{G})}$ that are realizable by the Observe&Do probing scheme on the pDAG $\mathcal{G}$, will be denoted $\mathcal{M}_{\mathrm{O\&D}}\big(\mathcal{G}, \vec{c}_{\mathtt{vis}(\mathcal{G})}\big)$. Note that it corresponds exactly to the set of conditional distributions that are realizable by $\mathtt{split}\big(\mathcal{G}\big)$ for the same cardinalities, i.e., $\mathcal{M}_{\mathrm{O\&D}}\big(\mathcal{G}, \vec{c}_{\mathtt{vis}(\mathcal{G})}\big) = \mathcal{M}_{\mathrm{obs}}\big(\mathtt{split}\big(\mathcal{G}\big), \vec{c}_{\mathtt{vis}(\mathcal{G})}\big)$.

We can now define notions of dominance and equivalence relative to the Observe&Do

probing scheme.

Definition 16 (O&D dominance and O&D equivalence of pDAGs)**.** *Let $\mathcal{G}$ and $\mathcal{G}'$ be two pDAGs such that $\texttt{vis}(\mathcal{G}) = \texttt{vis}(\mathcal{G}')$. We say that $\mathcal{G}$* O&D-dominates *$\mathcal{G}'$ when the set of conditional probability distributions realizable by an O&D probing scheme on $\mathcal{G}$ includes the set of conditional probability distributions realizable by an O&D probing scheme on $\mathcal{G}'$, regardless of the assignment of cardinalities of the visible variables, i.e., when*

$$\begin{aligned} &\forall \vec{c}_{\texttt{vis}(\mathcal{G})} \in \mathbb{N}^{|\texttt{vis}(\mathcal{G})|} : \\ &\mathcal{M}_{O\&D}(\mathcal{G}', \vec{c}_{\texttt{vis}(\mathcal{G}')}) \subseteq \mathcal{M}_{O\&D}(\mathcal{G}, \vec{c}_{\texttt{vis}(\mathcal{G})}). \end{aligned} \tag{2.20}$$

We say that $\mathcal{G}$ is O&D-equivalent *to $\mathcal{G}'$ when their sets of O&D-realizable distributions are the same:*

$$\begin{aligned} &\forall \vec{c}_{\texttt{vis}(\mathcal{G})} \in \mathbb{N}^{|\texttt{vis}(\mathcal{G})|} : \\ &\mathcal{M}_{O\&D}(\mathcal{G}', \vec{c}_{\texttt{vis}(\mathcal{G}')}) = \mathcal{M}_{O\&D}(\mathcal{G}, \vec{c}_{\texttt{vis}(\mathcal{G})}). \end{aligned} \tag{2.21}$$

It is important to stress a point already mentioned in the introduction of this chapter: when we ask about equivalence or dominance under a given probing scheme, we start from two candidate causal structures and ask if *there exists* some set of data coming from that probing scheme that, if obtained, would tell us that we have one causal structure and not the other. Importantly, even if two pDAGs $\mathcal{G}$ and $\mathcal{G}'$ are inequivalent under certain probing scheme, it is still possible that the data obtained from a particular experiment that applies this probing scheme does not allow us to tell if the underlying causal explanation is $\mathcal{G}$ or $\mathcal{G}'$. In other words, the fact that two pDAGs $\mathcal{G}$ and $\mathcal{G}'$ are inequivalent under a certain probing scheme means that their sets of realizable data are *not the same*, and not that these sets have an empty overlap.

A conditional probability distribution $P\big(X_{\texttt{vis}(\mathcal{G})^\flat} \big| X_{\texttt{vis}(\mathcal{G})^\sharp}\big)$ is said to be *O&D-realizable* by a pDAG $\mathcal{G}$ if it is *observationally realizable* by the full-SWIG $\texttt{split}(\mathcal{G})$. This fact implies the following lemma:

Lemma 8. *Let $\mathcal{G}$ and $\mathcal{G}'$ be two pDAGs with the same set of visible nodes. $\mathcal{G}$ O&D-dominates $\mathcal{G}'$ if and only if the full-SWIG $\texttt{split}(\mathcal{G})$ observationally dominates the full-SWIG $\texttt{split}(\mathcal{G}')$.*

This lemma shows that it suffices to characterize the dominance structure of full-SWIGs under passive observations in order to understand the dominance structure of pDAGs under

an O&D probing scheme. In this way, we reduce questions about the unconventional notion of O&D dominance to questions about the better-studied notion of observational dominance.

With this lemma in hand, we can now present one of the main results of this article. It is that O&D dominance of pDAGs is completely characterized by structural dominance of the corresponding mDAGs:

Theorem 1. *Let $\mathcal{G}$ and $\mathcal{G}'$ be two pDAGs where* $\texttt{vis}(\mathcal{G})=\texttt{vis}(\mathcal{G}')$*, and let $\mathfrak{G}$ and $\mathfrak{G}'$ be the corresponding mDAGs, i.e.,* $\mathfrak{G} \equiv \texttt{ToFaces}(\mathcal{G})$ *and* $\mathfrak{G}' \equiv \texttt{ToFaces}(\mathcal{G}')$*. The pDAG $\mathcal{G}$ O&D-dominates the pDAG $\mathcal{G}'$ if and only if the mDAG $\mathfrak{G}$ structurally dominates the mDAG $\mathfrak{G}'$.*

Proof. Together, Lemmas 7 and 8 imply that $\mathcal{G}$ O&D-dominates $\mathcal{G}'$ if and only if the 3-mDAG $\texttt{ToFaces}(\texttt{split}(\mathcal{G}))$ structurally dominates $\texttt{ToFaces}(\texttt{split}(\mathcal{G}'))$. With Lemma 6, we can commute `ToFaces` with `split`. Therefore, we have that $\mathcal{G}$ O&D-dominates $\mathcal{G}'$ if and only if $\texttt{split}(\texttt{ToFaces}(\mathcal{G})) = \texttt{split}(\mathfrak{G})$ structurally dominates $\texttt{split}(\texttt{ToFaces}(\mathcal{G}')) = \texttt{split}(\mathfrak{G}')$.

It remains therefore to prove that $\mathfrak{G}$ structurally dominates $\mathfrak{G}'$ if and only if $\texttt{split}(\mathfrak{G})$ structurally dominates $\texttt{split}(\mathfrak{G}')$. This follows from the fact that, for any two mDAGs $\mathfrak{G}$ and $\mathfrak{G}'$, there is an edge $a \to b$ present in $\mathfrak{G}$ but not present in $\mathfrak{G}'$ if and only if the corresponding edge $a^\sharp \to b^\flat$ is present in $\texttt{split}(\mathfrak{G})$ but not present in $\texttt{split}(\mathfrak{G}')$, and there is a facet S present in $\mathfrak{G}$ but not present in $\mathfrak{G}'$ if and only if the corresponding facet $S^\flat$ is present in $\texttt{split}(\mathfrak{G})$ but not present in $\texttt{split}(\mathfrak{G}')$. □

In particular, this result implies that pDAGs that correspond to different mDAGs are always Observe&Do inequivalent. Note that, as stressed below Definition 16, this *does not* imply that a given set of Observe&Do data uniquely determines the underlying mDAG that generated it.

By using structural dominance of mDAGs, we can construct a pre-order of pDAGs where the equivalence classes are given by the sets of all pDAGs that are associated with the same mDAG. We can convey information about this pre-order by instead presenting the corresponding partial order between equivalence classes, represented by the mDAGs. This partial order over mDAGs will be called the *structural partial order*. As an example, Fig. 2.4.1 shows the structural partial order of mDAGs with two nodes consistent with a fixed nodal ordering. By Theorem 1, the structural partial order captures the partial order over causal structures induced by the O&D dominance relation.

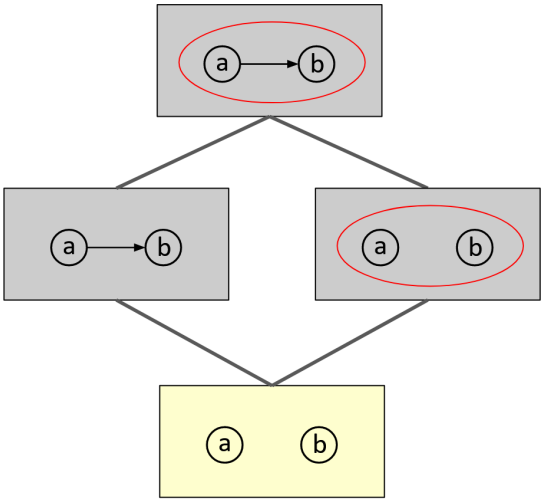


Figure 2.4.1: Structural partial order for the set of two-node mDAGs that are consistent with a fixed nodal ordering. This also expresses the partial order of equivalence classes under the O&D relation. The background colour of each mDAG indicates its observational equivalence class.

Fig. 2.4.2 shows the partial order that holds among equivalence classes of mDAGs of two nodes under the observational dominance relation. As can be seen in this figure, there are only two observational equivalence classes of two-node mDAGs: the one represented in gray can realize all probability distributions over two variables, and the one represented in white can only realize distributions where the two variables are independent. By comparing Fig. 2.4.1 with Fig. 2.4.2, we see that the partial order under O&D dominance is distinct from the partial order under observational dominance. In particular, this implies that there are cases where one pDAG observationally dominates another without structurally dominating it, that is, the converse of Lemma 3 does not hold. For instance, for the pDAG where a and b are cause-effect related and the pDAG where they are confounded, we have observational dominance relations holding in both directions, but no structural dominance in either direction.

2.5 Equivalence and Dominance under all patterns of do-interventions

Theorem 1 provides a complete characterization of what can be inferred about the underlying causal structure if there is access to an informationally complete probing scheme, such

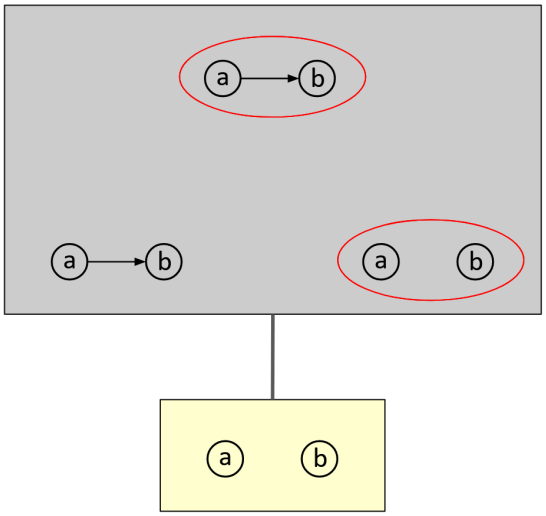


Figure 2.4.2: All two-node mDAGs consistent with a fixed nodal ordering, organized into the partial order of equivalence classes under the observational dominance relation.

as the Observe&Do probing scheme. We have shown that the full conditional distribution $P\big(X_{\mathtt{vis}(\mathcal{G})^\flat} \big| X_{\mathtt{vis}(\mathcal{G})^\sharp}\big)$ allows one to identify the pDAG up to its mDAG equivalence class. The question arises, then, whether strictly less information, in the form of certain shadows of the full conditional distribution, might still be sufficient for identifying the pDAG up to its mDAG equivalence class, that is, whether knowing the full conditional probability distribution $P\big(X_{\mathtt{vis}(\mathcal{G})^\flat} \big| X_{\mathtt{vis}(\mathcal{G})^\sharp}\big)$ is in fact *not necessary* for achieving such an identification. We will see shortly that this is in fact the case. In this section, we will investigate a probing scheme that is *not* informationally complete, but nevertheless dominance relative to it is *also* completely characterized by structural dominance of mDAGs.

Consider the case where, for each visible variable, the experimentalist implements a passive observation or a do-intervention, but never both, that is, the case where there is a disjunction of the two possibilities rather than the conjunction: Observe *or* Do, rather than Observe *and* Do.

Such a probing scheme is of practical signficance because in many real-world experiments that involve an intervention on one or more variables, the intervention is of the purely Do variety rather than of the Observe&Do variety. For instance, in most blind drug trials, where a given subject receives, at random, the drug or a placebo as the treatment, the experimentalist *does not* trouble themselves to ask the subject what their preference regarding taking the drug would have been were they to have been given the choice. In other words, the value that the treatment variable would take if it were determined by

the natural causal mechanisms (rather than fixed by intervention) is *not* recorded. A do-intervention is in fact the most common type of intervention considered in the causal inference literature.

We will refer to a specification of the set of visible nodes of a causal structure $\mathcal{G}$ that are subject to a do-intervention as a *do-pattern*, with the understanding that the complementary set of visible nodes is passively observed. The probing scheme wherein the subset of visible nodes that are subject to a do-intervention is $A \subseteq \texttt{vis}(\mathcal{G})$ will be referred to as the *A-pattern Observe-or-Do probing scheme*.

The data obtained from implementing the A-pattern Observe-or-Do probing scheme for some $A \subseteq \texttt{vis}(\mathcal{G})$ takes the following form:

$$\{P(X_{(\texttt{vis}(\mathcal{G})\setminus A)^\flat} | X_{A^\sharp} = x'_A) \mid x'_A \in \mathcal{X}_A\}, \tag{2.22}$$

where the natural value of a visible variable $X_v \in X_{\texttt{vis}(\mathcal{G})}$ is stored in the variable $X_{v^\flat}$, and the values that the variables X_A are forced to take are stored in the set of variables $X_{A^\sharp}$.

Note that the Observe&Do probing scheme (studied in Section 2.4) provides data where, for a node $a \in \texttt{vis}(\mathcal{G})$, the variable $X_{a^\flat}$ appears to the left side of the conditional *at the same time* that the variable $X_{a^\sharp}$ appears to the right side of the conditional. By contrast, this *does not happen* in the A-pattern Observe-or-Do probing scheme for some $A \subseteq \texttt{vis}(\mathcal{G})$. That is, for a node $a \in \texttt{vis}(\mathcal{G})$, in expression (2.22) there is *either* $X_{a^\flat}$ to the left side of the conditional *or* $X_{a^\sharp}$ to the right side of the conditional.

As discussed in Section 2.3, the expression (2.22) captures the same information as the standard notion of a "do-conditional", which is traditionally denoted by $P(X_{\texttt{vis}(\mathcal{G})\setminus A} | \text{do}(X_A))$. We will call it the "$A$-pattern do-conditional".

The shadowing function associated with the A-pattern Observe-or-Do probing scheme consists of taking the components of $P(X_{\texttt{vis}(\mathcal{G})^\flat} | X_{\texttt{vis}(\mathcal{G})^\sharp})$ where $X_{(\texttt{vis}(\mathcal{G})\setminus A)^\flat} = X_{(\texttt{vis}(\mathcal{G})\setminus A)^\sharp}$ and marginalizing out $X_{A^\flat}$. That is, defining $\bar{A} \equiv \texttt{vis}(\mathcal{G}) \setminus A$, the output of the shadowing function is the set

$$\left\{ \sum_{x_{A^\flat}} P(X_{A^\flat} = x_{A^\flat}, X_{\bar{A}^\flat} | X_{A^\sharp} = x'_A, X_{\bar{A}^\sharp}) \;\middle|\; x'_A \in \mathcal{X}_A, \right.$$
$$\left. P(X_{A^\flat}, X_{\bar{A}^\flat} = x_{\bar{A}^\flat} | X_{A^\sharp}, X_{\bar{A}^\sharp} = x_{\bar{A}^\sharp}) = 0 \text{ if } x_{\bar{A}^\flat} \neq x_{\bar{A}^\sharp} \right\} \tag{2.23}$$

This shadowing function gives the definition of realizability of an A-pattern do-conditional by a pDAG through Def. 15.

We will be primarily interested in a probing scheme wherein one collects statistical data from *all* possible patterns of do-interventions, that is, wherein one partitions the ensemble of samples and implements the A-pattern Observe-or-Do probing scheme for *every* subset $A \subseteq \mathtt{vis}(\mathcal{G})$. We refer to this as the *all-patterns Observe-or-Do* probing scheme. We will also make some brief remarks about *single-pattern* Observe-or-Do probing schemes, where one only has access to the data for a single subset $A \subseteq \mathtt{vis}(\mathcal{G})$.

In this section, we will show that even though the all-patterns Observe-or-Do probing scheme is generally not informationally complete, dominance relative to it is completely characterized by structural dominance of mDAGs. A claim found in Ref. [33, Section 7] seems, at first glance, to contradict this statement. So we endeavour here to explain how the result described there fits into the discussion in this article.

Ref. [33] discussed the problem of distinguishability of different pDAGs under (what we are here calling) single-pattern Observe-or-Do probing schemes, that is, access to the statistical data obtained from do-interventions on a single subset A of visible nodes. As noted in Ref. [33], a single-pattern Observe-or-Do probing scheme generally *cannot* distinguish pDAGs associated to different mDAGs, no matter which pattern one chooses. Although Ref. [33] considered what could be inferred from any pattern of do-interventions, it crucially did not allow that the data for one pattern could be compared with the data from another pattern, unlike the all-patterns probing scheme. This, ultimately, accounts for the differences between the results described earlier in this section and the claim of Ref. [33].

To make the distinction between the single-pattern and all-patterns Observe-or-Do probing schemes more explicit, here we discuss the example of Fig.16 of Ref. [33], which we reproduce in Fig. 2.5.1.

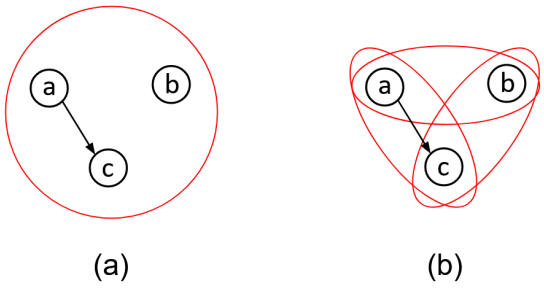


Figure 2.5.1: mDAGs that appear in Fig. 16 of Ref [33].

Individually, all of the distributions obtained from passive observations or from different do-patterns in the mDAG of Fig. 2.5.1(a) can also be realized by the mDAG of Fig. 2.5.1(b), which led Ref. [33] to classify these mDAGs as indistinguishable under each of these probing

schemes. However, further inspection reveals that there are sets of data obtained from passive observations and do-interventions that can be *jointly* realized by Fig. 2.5.1(a) but not Fig. 2.5.1(b). The following is an example: under passive observation, the three variables are perfectly correlated; under a do-intervention on variable X_a, the variables associated X_b and X_c remain perfectly correlated. In Fig. 2.5.1(a), one can jointly realize this behaviour if all variables copy the latent variable that acts as a three-way common cause (note that the arrow from a to c plays no role). In Fig. 2.5.1(b), the *only* way one can realize the behaviour for the passive observation case is by having the variables X_a and X_b copy the latent variable that acts upon both of them, and having variable X_c copy variable X_a. But this choice of parameters does not realize the behaviour for the do-intervention case because the intervention on X_a eliminates the correlation between X_a and X_b and thus also the correlation between X_b and X_c.

If one imagines having the experimental capability to implement *any* pattern of do-interventions, as Ref. [33] clearly did, there is no reason not to consider what can be inferred from comparing the statistical data obtained from these different patterns. That is, if one has access to the single-pattern probing scheme for every pattern, then there is no reason not to implement each of these on a subset of samples, and hence to implement the all-patterns probing scheme. This is an oversight of the analysis presented in Ref. [33].

In conclusion, to implement the all-patterns Observe-or-Do probing scheme it is *not* enough to show that every distribution from each do-pattern that is realizable by $\mathcal{G}$ is also realizable by $\mathcal{G}'$. Rather, it is necessary to show that all the *combinations* of distributions from different do-patterns that are realizable by $\mathcal{G}$ under one given choice of parameters $\texttt{par}$ are also realizable by $\mathcal{G}'$ under one single choice $\texttt{par}'$ of parameters.

To help us consider the relation between data obtained from different do-patterns, we now introduce the terminology of *joint realizability* to indicate that a set of data obtained from different do-patterns is realizable by a pDAG under the same choice of parameters.

Definition 17 (Joint Realizability of data from Different Do-Patterns)**.** *Let $\mathcal{G}$ be a pDAG, let $A_i \subseteq \texttt{vis}(\mathcal{G})$ be a subset of visible nodes of $\mathcal{G}$ and let $\mathcal{S} = \{A_i : i \in \{1, \ldots, N\}\}$ be a set of such subsets. Let $\mathcal{Q}_i(x'_{A_i})$ be the data obtained by performing a do-intervention on the variables associated to nodes in A_i, namely fixing $X_{A_i} = x'_{A_i}$, and passive observation on the rest,*

$$\mathcal{Q}_i(x'_{A_i}) = Q(X_{\texttt{vis}(\mathcal{G})\setminus A_i} | X_{A_i} = x'_{A_i}) \tag{2.24}$$

Let $\mathcal{Q}$ be the union, over all values of i and of $x'_{A_i}p$, of the $\mathcal{Q}_i(x'_{A_i})$,

$$\mathcal{Q} = \{\mathcal{Q}_i(x'_{A_i}) | x'_{A_i} \in \mathcal{X}_{A_i}, i \in \{1, \ldots, N\}\} \tag{2.25}$$

We say that the set of data $\mathcal{Q}$ is jointly realizable *by the pDAG $\mathcal{G}$ if there exists a choice of parameters* par *in $\mathcal{G}$ such that*

$$\begin{aligned}&\forall i \in \{1, ..., n\}, \forall x'_{A_i} \in \mathcal{X}_{A_i} :\\ &P^{(\mathcal{G},\texttt{par})}(X_{\texttt{vis}(\mathcal{G})\setminus A_i} | do(X_{A_i} = x'_{A_i})) =\\ &= Q(X_{\texttt{vis}(\mathcal{G})\setminus A_i} | X_{A_i} = x'_{A_i}) \ .\end{aligned} \tag{2.26}$$

Note that $\mathcal{Q}$ can include data obtained from passive observations: this is the case if, for some value of i, $A_i = \varnothing$. The set of all such sets of data $\mathcal{Q}$ that are jointly realizable by a pDAG $\mathcal{G}$ and exhibit the cardinalities $\vec{c}_{\texttt{vis}(\mathcal{G})}$ for the visible variables will be denoted by $\mathcal{M}_{\text{All O-or-D}}(\mathcal{G}, \vec{c}_{\texttt{vis}(\mathcal{G})})$, where All O-or-D stands for "all-patterns Observe-or-Do".

Now we can finally give our definitions of dominance and equivalence relative to the all-patterns Observe-or-Do probing scheme:

Definition 18 (All-patterns O-or-D dominance and equivalence of pDAGs)**.** *Let $\mathcal{G}$ and $\mathcal{G}'$ be two pDAGs such that* $\texttt{vis}(\mathcal{G}) = \texttt{vis}(\mathcal{G}')$*. We say that $\mathcal{G}$* all-patterns O-or-D dominates *$\mathcal{G}'$ when, regardless of the assignment of cardinalities of the visible variables, all of the sets of data obtained from the all-patterns probing scheme that are jointly realizable by $\mathcal{G}'$ are also jointly realizable by $\mathcal{G}$. That is, when*

$$\begin{aligned}&\forall \vec{c}_{\texttt{vis}(\mathcal{G})} \in \mathbb{N}^{|\texttt{vis}(\mathcal{G})|} :\\ &\mathcal{M}_{All\ O\text{-}or\text{-}D}(\mathcal{G}', \vec{c}_{\texttt{vis}(\mathcal{G}')}) \subseteq \mathcal{M}_{All\ O\text{-}or\text{-}D}(\mathcal{G}, \vec{c}_{\texttt{vis}(\mathcal{G})})\end{aligned} \tag{2.27}$$

We say that $\mathcal{G}$ is all-patterns O-or-D equivalent *to $\mathcal{G}'$ when dominance holds in both directions.*

In other words, $\mathcal{G}$ all-patterns O-or-D dominates $\mathcal{G}'$ if for all choices $\texttt{par}'$ of parameters of $\mathcal{G}'$ there exists at least one choice par of parameters of $\mathcal{G}$ such that, for *any* choice of set of visible variables $X_A \in X_{\texttt{vis}(\mathcal{G})}$ to intervene upon and of values to set these variables to, the realized do-conditional (or joint distribution, if $A = \varnothing$) is the same:

$$\begin{aligned}&\forall A \in \texttt{vis}(\mathcal{G}), \forall x'_A \in \mathcal{X}_A :\\ &P^{(\mathcal{G},\texttt{par})}(X_{\texttt{vis}(\mathcal{G})} \setminus X_A | \text{do}(X_A = x'_A)) =\\ &= P^{(\mathcal{G}',\texttt{par}')}(X_{\texttt{vis}(\mathcal{G}')} \setminus X_A | \text{do}(X_A = x'_A)).\end{aligned} \tag{2.28}$$

Now, we prove our second main result: even though the all-patterns Observe-and-Do probing scheme is not informationally complete, all-patterns O-or-D dominance is *also* characterized by structural dominance of mDAGs.

Theorem 2. *Let $\mathcal{G}$ and $\mathcal{G}'$ be two pDAGs and let $\mathfrak{G}$ and $\mathfrak{G}'$ be the corresponding mDAGs, i.e., $\mathfrak{G} \equiv \texttt{ToFaces}(\mathcal{G})$ and $\mathfrak{G}' \equiv \texttt{ToFaces}(\mathcal{G}')$. $\mathcal{G}$ all-patterns O-or-D dominates $\mathcal{G}'$ if and only if $\mathfrak{G}$ structurally dominates $\mathfrak{G}'$.*

Proof. The "if" side follows from Theorem 1. This is so because, as we saw with the example of Fig. 2.3.1, passive observations and do-interventions reveal different shadows of the conditional probability distribution $P(X_{\texttt{vis}(\mathcal{G})^\flat} | X_{\texttt{vis}(\mathcal{G})^\sharp})$. If $\mathfrak{G}$ structurally dominates $\mathfrak{G}'$, then Theorem 1 says that even with access to the full *distribution* $P(X_{\texttt{vis}(\mathcal{G})^\flat} | X_{\texttt{vis}(\mathcal{G})^\sharp})$ it is impossible to find data that is realizable by $\mathcal{G}'$ but not by $\mathcal{G}$; therefore, the same is certainly true when one only has access to *shadows* of this distribution.

To prove the "only if" side, we will proceed by proving its contrapositive. Assume that $\mathfrak{G}$ *does not* structurally dominate $\mathfrak{G}'$. The lack of structural dominance implies one of two possibilities:

1. There is at least one directed edge between visible nodes a and b, $a \to b$, that is present in $\mathfrak{G}'$ but not in $\mathfrak{G}$, or

2. $\mathfrak{G}$ has all the directed edges that are present in $\mathfrak{G}'$, but there is at least one set S of visible nodes which is a face of $\mathfrak{G}'$ but not of $\mathfrak{G}$.

These two conditions for mDAGs respectively imply the following conditions for the corresponding pDAGs:

1. There is at least one pair of visible nodes a and b such that $\mathcal{G}'$ presents a chain $a \to m_1 \to ... \to m_n \to b$ where all of the mediary nodes $m_1, ..., m_n$ are latent nodes, and there is *no* such chain in $\mathcal{G}$, or

2. Item 1 does not hold, but there is at least one set S of visible nodes that have a common latent ancestor λ in $\mathcal{G}'$, but no such latent ancestor exists in $\mathcal{G}$.

For case 1, note that all the chains that connect a to b in $\mathcal{G}$ must have at least one mediary node m that is *visible*. Let the set of all such visible mediary nodes be called M. Now, if A is the set of nodes subject to a do-intervention in a given do-pattern, we will compare the case where $A = \varnothing$ with the case where $A = \{a\}$. That is, we consider two do-patterns: (i) passive observations on all the visible variables; and (ii) do-intervention on

a and passive observations on all other visible variables. Suppose that the data obtained respectively from these two do-patterns have the marginals:

$$\text{(i)} \qquad P(X_b|X_M) \tag{2.29a}$$

$$\text{(ii)} \qquad P(X_b|X_M, \text{do}(X_a = 0)) \neq P(X_b|X_M). \tag{2.29b}$$

This set of data is jointly realizable by $\mathcal{G}'$, since in this pDAG the variable X_b can have a direct functional dependence on X_a, in a way that the marginal over X_b changes when X_a is forced to take the value 0. However, this set of data is *not* jointly realizable by $\mathcal{G}$: after we condition on all of the variables X_M associated with mediary nodes M, all of the directed chains between a and b become blocked paths. Furthermore, none of the nodes in M can be common children of a and b. Therefore, all of the unblocked paths between a and b are paths that have incoming arrows to a; these causal connections are suspended when we perform do-interventions on a. Therefore, $\mathcal{G}$ *does not* all-patterns O-or-D dominates $\mathcal{G}'$[17].

For case 2, consider the do-patterns $A = \varnothing$ and, for each element $T \in 2^S$ of the powerset of S, $A = \texttt{vis}(\mathcal{G}) \setminus T$. That is, consider (i) passive observations on all the visible variables; and (ii) for each element $T \in 2^S$ of the powerset of S, a do-intervention that fixes a variable to 0 on all of the visible variables except for the ones in X_T, and a passive observation of X_T. Suppose that the data obtained from these do-patterns is:

$$\text{(i)} \qquad P(X_S) = p[0, ..., 0] + (1 - p)[1, ..., 1] \tag{2.30a}$$

$$\text{(ii)} \qquad \forall T \in 2^S,$$
$$Q(X_T = 0|\text{do}(X_{\texttt{vis}(\mathcal{G})\setminus T} = 0_{\texttt{vis}(\mathcal{G})\setminus T})) = p \tag{2.30b}$$

Where $0_{\texttt{vis}(\mathcal{G})\setminus T}$ denotes an assignment of value 0 to each variable X_a where $a \in \texttt{vis}(\mathcal{G}) \setminus T$, and the expression in the right hand side of Eq. (2.30a) means that with p probability all of the variables of X_S take value 0, and with $1 - p$ probability all of them take the value 1. That is, all of the variables in X_S are perfectly correlated in the data obtained from passive observations, and the marginal over each subset of the variables of X_S does not

[17] Alternatively, case 1 can be solved by the following observation, made in Ref. [33, Proposition 7.5]: do-conditionals where

$$\exists x_a, x'_a \text{ such that}$$
$$P(X_b|\text{do}(X_a = x_a)) \neq P(X_b|\text{do}(X_a = x'_a)),$$

that is, where the observed value of X_b depends on the value to which X_a was set, are jointly realizable by $\mathcal{G}'$ but not by $\mathcal{G}$.

change by performing a do-intervention that forces all of the other visible variables to take the value 0.

As it turns out, the set of data of Eqs. (2.30a) and (2.30b) is jointly realizable by $\mathcal{G}'$, but not by $\mathcal{G}$. It is easy to see why it is jointly realizable by $\mathcal{G}'$: there, all of the variables in X_S can copy the latent common cause X_λ that they share. The proof that this set of data is *not* jointly realizable by $\mathcal{G}$ makes use of the fact that $\texttt{split}(\mathcal{G})$ cannot realize the conditional distribution of Eq. (2.14), together with a technique due to Pedro Lauand that is described in Appendix A.2.

Therefore, in both cases we have explicitly presented a set of data obtained from different do-patterns that is jointly realizable by $\mathcal{G}'$ but not by $\mathcal{G}$. This shows that, if $\mathfrak{G} = \texttt{ToFaces}(\mathcal{G})$ does not structurally dominate $\mathfrak{G}' = \texttt{ToFaces}(\mathcal{G}')$, then $\mathcal{G}$ does not all-patterns O-or-D dominate $\mathcal{G}'$. □

2.5.1 One-Value do-interventions

In the proof of Theorem 2, Eqs. (2.29b) and (2.30b) only involve interventions that set a variable to *one* of its possible values, here called 0. This shows that, to distinguish two pDAGs that correspond to the same mDAG, it is sufficient to perform do-interventions that set the variables to only one value. We will call those *one-value do-interventions*, and the probing scheme that consists of implementing one-value do-interventions on all possible do-patterns will be called the *all-patterns Observe-or-1Do probing scheme*. To see why the restriction to the all-patterns Observe-or-1Do probing scheme can be of interest, imagine an experiment where we are allowed to force the subjects to quit smoking, but we cannot ethically force them to start smoking. In this case, we can intervene on the experiment to force the variable "smoker" to take the value 0, but not to take the value 1.

We define dominance and equivalence relative to this probing scheme, termed *all-patterns O-or-1D dominance and equivalence*, in an analogous way to Def. 18. That is, we say that $\mathcal{G}$ *all-patterns O-or-1D dominates* $\mathcal{G}'$ when, regardless of the assignment of cardinalities of the visible variables, all of the sets of data that are realizable by $\mathcal{G}'$ in the all-patterns Observe-or-1Do probing scheme are also realizable by $\mathcal{G}$. The observation of the previous paragraph implies that all-patterns O-or-1D dominance is *also* characterized by the structural dominance of mDAGs:

Theorem 3. *Let $\mathcal{G}$ and $\mathcal{G}'$ be two pDAGs and let $\mathfrak{G}$ and $\mathfrak{G}'$ be the corresponding mDAGs, i.e., $\mathfrak{G} \equiv \texttt{ToFaces}(\mathcal{G})$ and $\mathfrak{G}' \equiv \texttt{ToFaces}(\mathcal{G}')$. $\mathcal{G}$ all-patterns O-or-1D dominates $\mathcal{G}'$ if and only if $\mathfrak{G}$ structurally dominates $\mathfrak{G}'$.*

Proof. The "if" side follows from Theorem 2. For the "only if" side, because our proof of Theorem 2 only made use of do-interventions that force the variables to take the value 0, it serves as a proof for this case as well. □

2.6 Structural Partial Order for mDAGs with three or four visible nodes

Theorems 1, 2 and 3 respectively say that the partial order of equivalence classes of causal structures under the relations of O&D dominance, all-patterns O-or-D dominance and all-patterns O-or-1D dominance all correspond exactly to the partial order induced by structural dominance of mDAGs, i.e., the structural partial order. Finding this partial order for mDAGs with any number of nodes is straightforward. It is nonetheless pedagogically useful to see some concrete examples. The first example was already presented in Fig. 2.4.1: the structural partial order for two-node mDAGs. We will here present the structural partial orders for three-node and four-node mDAGs.

The structural order for the full set of n-node mDAGs consistent with a fixed nodal ordering can be understood in terms of the structural order holding among two subsets thereof: those that are confounder-free and those that are directed-edge-free.

First, we note that one can think of the full set as the Cartesian product of these two subsets. The set of directed-edge-free n-node mDAGs describes the set of all possible simplicial complexes that can hold among n nodes, while the set of confounder-free n-node mDAGs describes the set of all possible directed-edge structures that can hold for these. The set of all mDAGs is simply the set of all possible combinations of a simplicial complex with a directed-edge structure. For the simplest possible example, of two-node mDAGs, the structural partial order for the subset of two-node mDAGs that are confounder-free is presented in Fig. 2.6.1(a), and the structural partial order for the subset of two-node mDAGs that are directed-edge-free is presented in Fig. 2.6.1(b). The full structural partial order in this case was presented in Fig. 2.4.1.

Second, we note that a dominance relation that holds between two mDAGs that are directly connected in the partial order is due either to dropping a face from the simplicial complex or from dropping an edge from the directed-edge structure. It is useful to organize the full set of n-node mDAGs consistent with a fixed nodal ordering into "islands" where the mDAGs within an island have the same simplicial complex and differ only in their directed-edge structure. For the simple case of two-node mDAGs, these islands are given by the large rectangular boxes of Fig. 2.6.2. The partial order holding among the elements

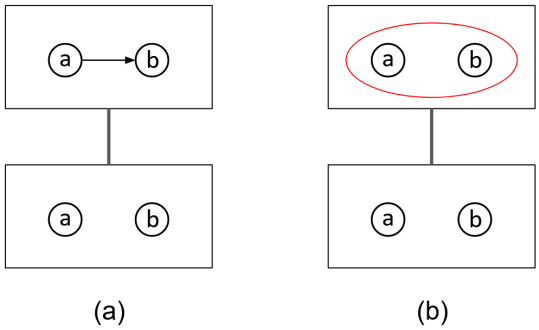


Figure 2.6.1: (a) Structural partial order for the set of two-node counfounder-free mDAGs that are consistent with a fixed nodal ordering. (b) Structural partial order that holds between all of the two-node directed-edge-free mDAGs. Note that a directed-edge-free mDAG is always consistent with any nodal ordering.

of each island is the one induced by dropping edges, which is simply the partial order of the confounder-free mDAGs. Meanwhile, if one takes the collection of mDAGs that have the same directed-edge structure but different simplicial complexes (containing one mDAG from each island), the partial order holding among these is the one induced by dropping faces, which is simply the partial order of the directed-edge-free mDAGs. These two sets of partial order relations exhaust the set of relations holding among the mDAGs.

When depicting the full structural partial order, we explicitly depict the partial order relations that hold between the mDAGs *within* each island, but we use a shorthand notation for the partial order relations that hold between mDAGs from different islands (those induced by dropping of faces of the simplicial complex). Rather than drawing the face-dropping partial order explicitly for each collection of mDAGs that have the same directed-edge structure, one puts a box around each island and one draws the face-dropping partial order among the islands. This is explicit in Fig. 2.6.2, which is our shorthand depiction of the full structural order of Fig. 2.4.1. Within each island of Fig. 2.6.2, the partial order is that of Fig. 2.6.1(a). The partial order between the islands is that of Fig. 2.6.1(b), and represents the partial order between the collection of mDAGs that have the same directed-edge structure (one from each island).

To obtain the structural partial order that holds between the three-node mDAGs that are consistent with a fixed nodal ordering, we can combine the structural partial order that holds between confounder-free three-node mDAGs consistent with a fixed nodal ordering, shown in Fig. 2.6.3(a), with the structural partial order that holds between directed-edge-free three-node mDAGs consistent with a fixed nodal ordering, shown in Fig. 2.6.3(b).

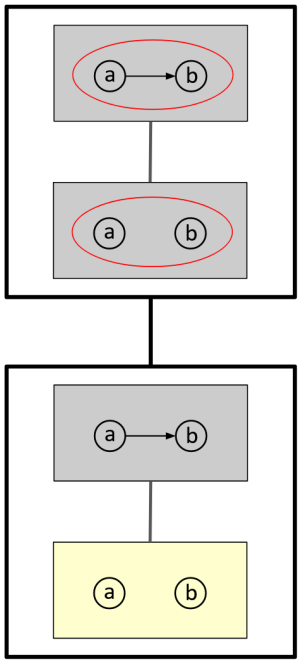


Figure 2.6.2: Compressed way of expressing the structural partial order that holds between two-node mDAGs. The decompressed structural partial order was given in Fig. 2.4.1.

The resulting partial order is presented in Fig. 2.6.4. Note that because the subset of confounder-free mDAGs is of cardinality 8 and the subset of directed-edge-free mDAGs is of cardinality 9, the full set, which can be obtained from their Cartesian product, is of cardinality 72.

Although complicated, Fig. 2.6.4 is useful as it gives us a picture of *all* the structural dominance relations that hold between mDAGs of three nodes that are consistent with a fixed nodal ordering, and thus all of the O&D, all-patterns O-or-D and all-patterns O-or-1D dominance relations as well. Sameness of background colour and of number in Fig. 2.6.4 represents membership in the same observational equivalence class, i.e., two mDAGs have the same background colour and the same number if they are observationally equivalent.[18] This is yet another example of the difference between observational equivalence and equivalence when there is access to interventional probing schemes, under the three notions defined here.

We turn now to the case of the four-node mDAGs that are consistent with a fixed nodal

[18] Such observational equivalences will be studied and justified in Chapter 3.

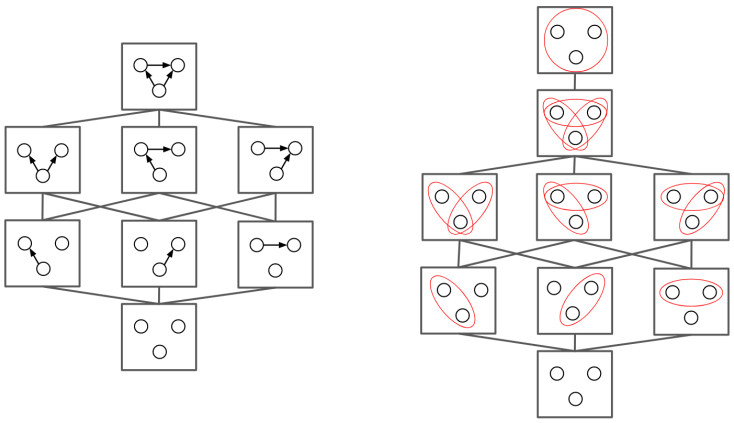


Figure 2.6.3: (a) Structural partial order for the set of three-node confounder-free mDAGs that are consistent with a fixed nodal ordering. (b) Structural partial order for the set of three-node directed-edge-free mDAGs. Note that a directed-edge-free mDAG is always consistent with any nodal ordering.

ordering. The structural partial order of the subset that is confounder-free is presented in Fig 2.6.5, and the structural partial order of the subset that is directed-edge-free is presented in Fig. 2.6.6. The full set of four-node mDAGs can be obtained by taking the Cartesian product of these two subsets. However, because the confounder-free subset has cardinality 64 and the directed-edge-free subset has cardinality 113, the full set has cardinality 7232 and it is thus too cumbersome to depict it explicitly. Nonetheless, one can easily infer the shape of the structural partial order of the full set by imagining combining Figs. 2.6.5 and 2.6.6 with the procedure described above.

2.7 Conclusion

The findings of this chapter are summarized in Fig. 2.7.1. As discussed in Section 2.3, each probing scheme (without edge interventions) that can be implemented on the visible variables $X_{\mathtt{vis}(\mathcal{G})}$ reveals a *shadow* of the full conditional probability distribution $P\big(X^{\flat}_{\mathtt{vis}(\mathcal{G})}\big|X^{\sharp}_{\mathtt{vis}(\mathcal{G})}\big)$. By applying the shadowing function associated with a probing scheme to the set of all possible conditional probability distributions realizable by a causal structure, we obtain the set of possible shadows realizable by the probing scheme.

We can define an equivalence relation among probing schemes based on whether they

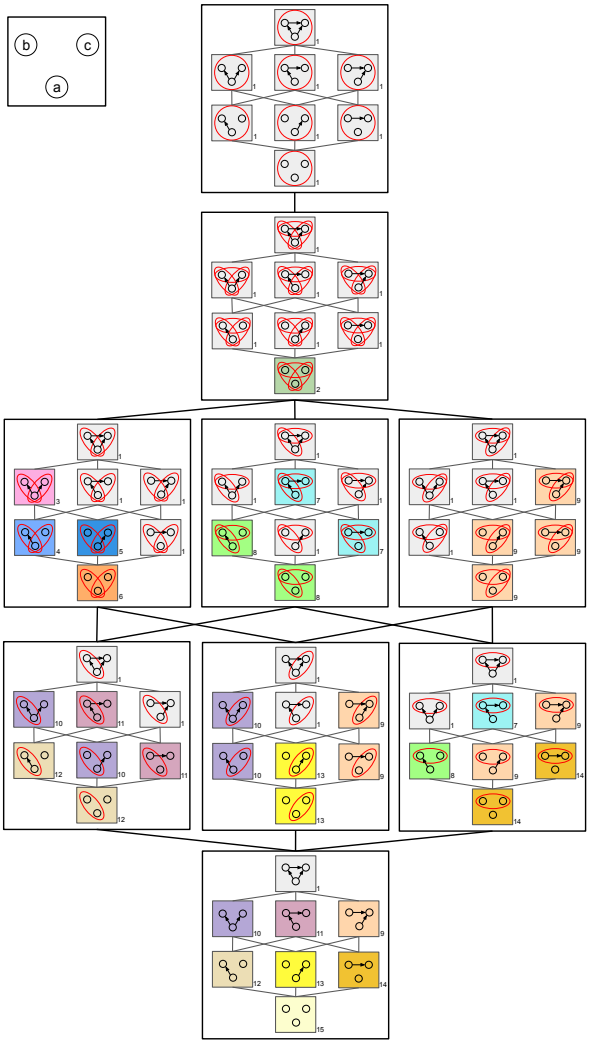


Figure 2.6.4: Structural partial order for the set of three-node mDAGs consistent with a fixed nodal ordering. The line between a pair of large boxes (islands) is to be interpreted as every mDAG within the higher island dominating the *corresponding* mDAG within the lower island, i.e., the mDAG that has the same directed-edge structure. Two mDAGs of this picture have the same background colour and the same number if they are *observationally* equivalent.

reveal the same information about the causal hypothesis. More precisely, two probing schemes are judged to be in the same equivalence class if they are associated to the same shadowing function. For example, all of the *informationally complete* probing

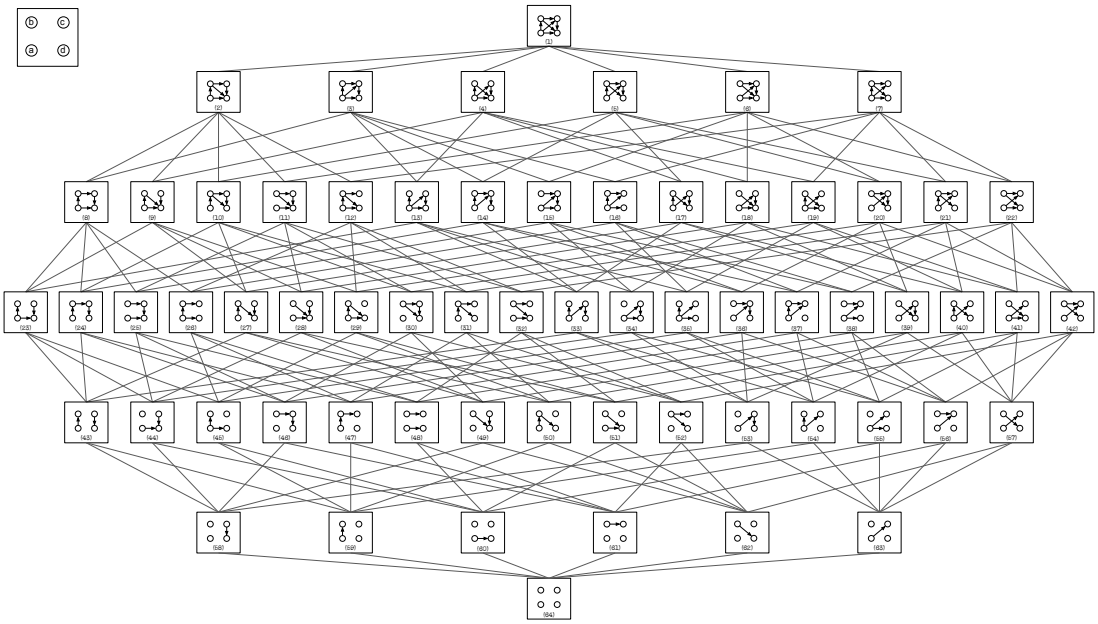


Figure 2.6.5: Structural partial for the set of four-node confounder-free mDAGs consistent with a fixed nodal ordering.

schemes, that is, the ones that allow us to infer the full conditional probability distribution $P\big(X_{\mathtt{vis}(\mathcal{G})^\flat} \big| X_{\mathtt{vis}(\mathcal{G})^\sharp}\big)$ (so that the shadowing function is the identity function), are inside the same equivalence class.

Moreover, we can define a partial order over the equivalence classes of probing schemes using the ordering relation induced by set inclusion of the images of their shadowing functions. Thus, one probing scheme is above another in the partial order if the image of the first's shadowing function contains all of the information that is included in the image of the second's shadowing function; in other words, if it reveals strictly more information about the full conditional probability distribution $P\big(X_{\mathtt{vis}(\mathcal{G})^\flat} \big| X_{\mathtt{vis}(\mathcal{G})^\sharp}\big)$.

Fig. 2.7.1 presents the partial order of the equivalence classes of probing schemes that we studied here.

In the box at the top of Fig. 2.7.1, we have probing schemes that are informationally complete. In this case, the problem of equivalence and dominance of pDAGs under the given probing scheme is completely characterized by the mDAG structure, and by the structural dominance of mDAGs respectively. In particular, we noted that an example of an informationally complete probing scheme is the Observe&Do probing scheme, wherein,

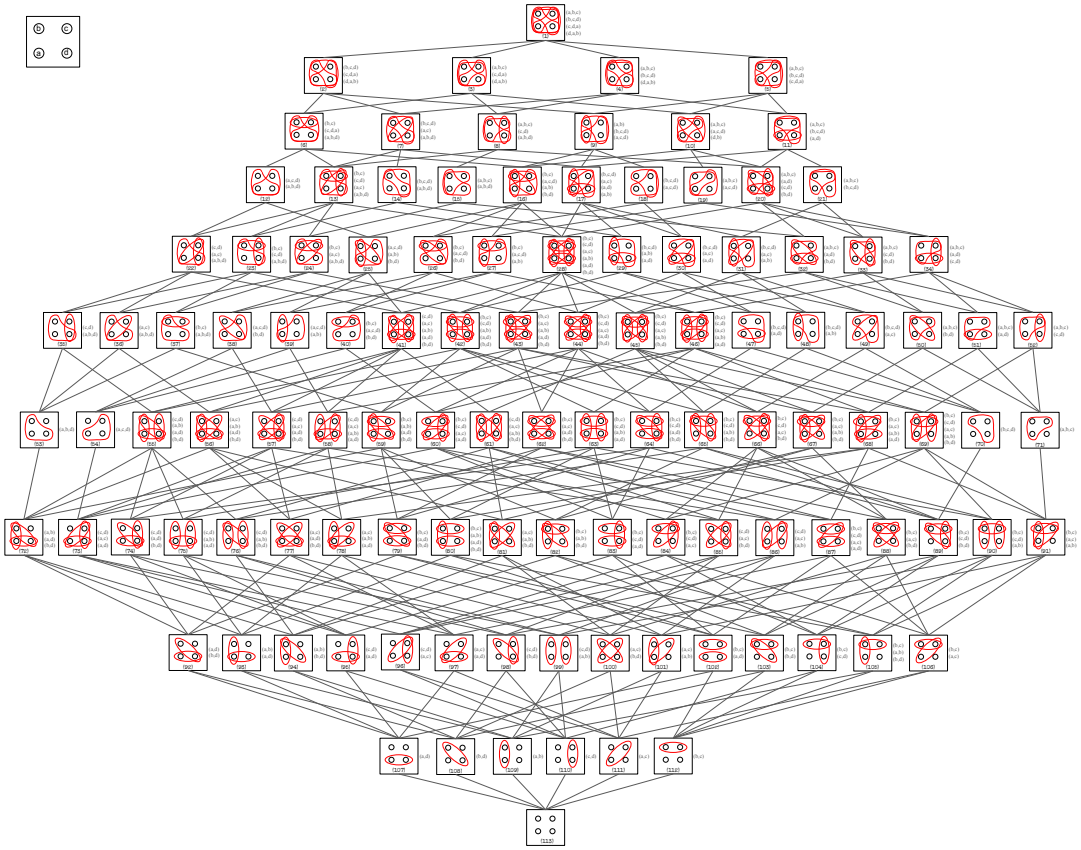


Figure 2.6.6: Structural partial order that holds between all four-node directed-edge-free mDAGs. Next to each mDAG, we indicate the set of facets of its simplicial complex. Note that a directed-edge-free mDAG is always consistent with any nodal ordering.

for each visible variable, one observes its natural value and subsequently performs a do-intervention upon it. Theorem 1 demonstrates that dominance relative to such a probing scheme corresponds to structural dominance of the mDAGs.

Fig. 2.7.1 also summarizes what one can infer about the causal structure when one has access to various probing schemes that are *not* informationally complete. The box at the second highest level of the partial order of Fig. 2.7.1 corresponds to the all-patterns Observe-or-Do probing scheme, wherein, for each subset of visible nodes (which we here call a *do-pattern*), one implements a do-intervention on the variables in the subset and passive observation on the rest. In Theorem 2 we show that, even though this probing scheme is not informationally complete, it is informative enough to be able to distinguish between different mDAGs. More specifically, the problem of equivalence and dominance of pDAGs relative to the all-patterns Observe-or-Do probing scheme is also completely characterized by the mDAG structure and by the structural dominance of mDAGs respectively.

Finally, in Theorem 3, we strengthen this result by showing that it is not necessary to have the ability to perform do-interventions that set the visible variables to each of their possible values; in fact, implementing *one-value* do-interventions (wherein a variable can be set to just one of its possible values) is sufficient to distinguish different mDAGs. We refer to this probing scheme as the *all-patterns Observe-or-1-value-Do* probing scheme. The mDAG structure also completely characterizes equivalence and dominance relative to this probing scheme.

The lowest level in the hierarchy of probing schemes presented in Fig. 2.7.1 is passive observation of all visible nodes. Equivalence and dominance of pDAGs relative to passive observations is *not* characterized by structural dominance of mDAGs. That is, one can find pairs of pDAGs that are observationally equivalent in spite of being associated to different mDAGs [33]. Equivalence and dominance of mDAGs relative to passive observations will be studied in Chapter 3.

We end with some open problems.

The first open problem is whether pDAGs associated to the same mDAG are still indistinguishable when one has access to edge interventions, that is, interventions that send a different value of the intervened node to each of its children. It is worth noting how one can conceptualize an edge intervention as a probing scheme on a visible variable. It suffices to imagine that a variable is first copied and then the different copies are what causally influence each of the original variable's children. One then intervenes differently on each of these copies. Is there any probing scheme *with* edge interventions that can distinguish between two pDAGs that are associated with the same mDAG? We believe the answer is no.

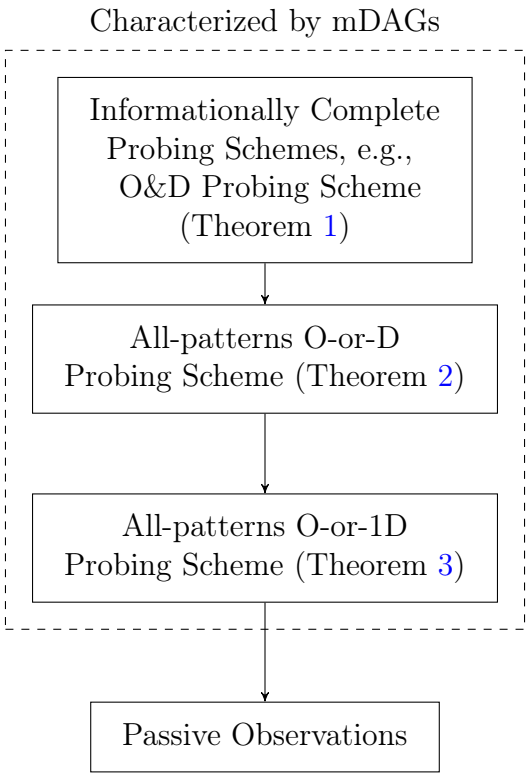


Figure 2.7.1: Partial order of equivalence classes of the different probing schemes studied in this work, where the relation is inclusion of the image of the shadowing function of each probing scheme. The informationally complete probing schemes, that reveal the entirety of this conditional distribution $P(X^{\flat}_{\mathtt{vis}(\mathcal{G})}|X^{\sharp}_{\mathtt{vis}(\mathcal{G})})$, form an equivalence class. In the O&D Probing Scheme, for each visible variable one can first observe its natural value and then force it to take any desired value. In the all-patterns O-or-D probing scheme, for each visible variable one can either passively observe it or force it to take any desired value, without previously observing the natural value of the variable. In the all-patterns O-or-1D probing scheme, for each visible variable one can either passively observe it or force it to take *one* fixed value, without previously observing the natural value of the variable.

A second open problem to be explored goes in the other direction: how much can a probing scheme be weakened and *still* induce a dominance order of realizable distributions that is characterized by structural dominance of the associated mDAG structure? Of the probing schemes we studied, the weakest one that satisfies this is the all-patterns Observe-or-1Do probing scheme. Are there even weaker probing schemes that also satisfy this condition? Also, might the condition be satisfied by other probing schemes that are not

informationally complete and strictly incomparable (in the partial order) to the all-patterns Observe-or-Do scheme or the all-patterns Observe-or-1Do scheme?

Finally, all of the results presented in this chapter pertain to causal structures where all the nodes are associated with classical random variables. When the latent nodes of a pDAG are associated with *quantum systems*, however, it is known that even passive observations of the visible variables can distinguish some pDAGs that are associated to *the same* mDAG. This was first noted in Ref. [81], through an example that will be reproduced in Chapter 5. Consequently, an open question of interest to quantum physicists is: if not the mDAG, then what characterizes equivalence and dominance of causal structures under an informationally complete probing scheme on the visible variables when the latent nodes are quantum?

Chapter 3

Observational distinguishability of causal structures in the presence of latent nodes

In the previous chapter, we studied the distinguishability of causal structures under different interventional probing schemes. While interventions yield strong causal conclusions, they can be prohibitively expensive (e.g., a drug trial), unethical (e.g., deciding whether smoking causes lung cancer), or simply impossible (e.g., in observational astronomy). For this reason, it is useful to understand what causal conclusions can be drawn from purely observational data, that is, from the probability distribution over the visible variables.

The set of probability distributions over the visible variables obtained from passive observations that can be causally explained by a given causal structure generally satisfy nontrivial constraints. We refer to these constraints as *causal-compatibility constraints*. While the causal-compatibility constraints of a latent-confounder-free causal structure always take the form of equalities, a causal structure that has latent variables can also exhibit causal-compatibility constraints in the form of inequalities. We will refer to these two types of constraints as *equality constraints* and *inequality constraints* respectively. A complete solution to the question of which probability distributions can be realized by a causal structure requires a characterization of all causal-compatibility constraints for each causal structure, which is a very difficult problem in general. This is because the inequality constraints are hard to find.

A more straightforward project is to simply *classify* the causal structures. This task requires determining when a pair of such causal structures can realize the *same* set of

distributions over the visible variables, that is, when are they observationally equivalent. We can then classify these causal structures into *observational equivalence classes*. More generally, one wishes to know, for each pair of causal structures, whether the set of distributions realizable by one contains the set of distributions realizable by the other, that is, whether the first observationally dominates the second. This dominance relation defines a partial order among the observational equivalence classes of causal structures with a fixed number of visible nodes, which we term the *observational partial order*.

In the previous chapter, we have shown that the mDAG structure captures *everything* one can in principle infer about the causal structure from any (potentially interventional) probing scheme that interacts with the visible variables. Therefore, it is clear that two pDAGs that correspond to the same mDAG will also not be distinguishable when one only has access to passive observations. Because of that, here we will start from the set of mDAGs with a fixed number of visible nodes, organize them into observational equivalence classes and then organize these classes into the observational partial order. The simplest example of the observational partial order, the one for two-node mDAGs consistent with a fixed nodal ordering, was presented in Figure 2.4.2. The determination of the observational partial order is the main aim of this chapter.

Of course, a determination of *all* of the causal-compatibility constraints for each mDAG would be *sufficient* for determining the observational partial order, but it is not *necessary*, and this is why the project of identifying the observational partial order is easier than the project of achieving a full characterization of the causal-compatibility constraints for each mDAG. The project of determining the observational partial order, however, has applications in causal discovery: given knowledge of when two causal structures are observationally equivalent, if we find a causal-compatibility constraint for one of them, we know it applies also to the other. In particular, a classification into observational equivalence classes is helpful for attesting the presence or absence of inequality constraints: if a causal structure is known to be observationally equivalent to another one that is latent-confounder-free, then we know that it cannot present inequality constraints.

Characterizing the observational partial order is also significant for identifying the best causal explanations of observational data. It is standard to argue that if two causal structures can both realize a probability distribution, one should prefer the structure that has less explanatory power, that is, the one that is lower in the observational dominance partial order. (This is sometimes motivated by the fact that this structure is more falsifiable.). Identifying the set of causal structures that provide the best explanation of a given distribution, therefore, requires identifying the lowest elements in the observational dominance order that can realize the distribution. In this way, a knowledge of the observational partial order permits a narrowing-down of the scope of possibilities for a good causal explanation

of the data. As we will see in Section 3.2.3, the observational partial order is particularly significant when one seeks to take into account finite-sample effects in causal discovery.

This chapter focuses on rules for proving that two causal structures are observationally equivalent, inequivalent, or that one dominates (or not) the other. We gather the known such rules from the literature, determine logical relations among these, and present new ones (Propositions 7 and 8). With these rules, we aim to determine the resulting observational partial order. Sec. 3.2.4 will explain the precise procedure that we follow to do so. For the reasons we presented in Section 2.1.4, our observational partial orders will only include causal structures compatible with a fixed nodal ordering.

As mentioned, meeting the challenge of identifying the observational equivalence classes also contributes towards identifying pDAGs that present *inequality* constraints among their causal-compatibility constraints. Such pDAGs will be called *nonalgebraic*, for reasons that we explain in Section 3.1. The study of inequality constraints is of particular importance for quantum physicists who are interested in causal inference, because such constraints are analogues of Bell inequalities [8], and it is only through their violation that one can witness the need for quantumness of a causal model [39, 78, 50, 22]. Motivated by this, there has been previous work on the problem of identifying when a causal structure will present inequality constraints [47, 59, 48].

In the realm of classical causal inference, the most common constraint-based causal discovery algorithms (such as IC and PC) [79, 74] leverage only conditional independence relations (which are *equality* constraints) in trying to adjudicate between the different observational equivalence classes. These algorithms can be supplemented by techniques for identifying the full set of equality constraints, including nested Markov constraints [68]. However, any algorithm that is restricted to using equality constraints necessarily lacks discriminatory power relative to algorithms that appeal to inequality constraints.

An important question is: how significant are causal discovery techniques that go beyond conditional independence constraints? For every causal structure whose causal-compatibility constraints are not simply conditional independence relations, it is known [35] that there are nontrivial inequality constraints. Consequently, the first question one might ask is how *rare* inequality constraints are; if they are too rare, one might conclude that not much is lost by leveraging only conditional independence relations in one's causal discovery algorithm.

When we look at data obtained from passive observation of the visible nodes, the only opportunity for distinguishing two causal structures is if the set of probability distributions realizable by one is different from the set realizable by the other. For example, there is no opportunity for observationally distinguishing two distinct causal structures that are both

in the saturating class, since the realizable set of distributions is the full set of possible distributions over the visible variables in both cases. In particular, given data obtained from passive observation of the visible nodes, one does not have the ability to make any discrimination that is more fine-grained than the observational equivalence classes.[1] It follows that in order to get a sense of the significance of inequality constraints in making discriminations among causal structures, one can count the fraction of observational equivalence classes that imply inequality constraints.

It is worth noting that we are *not* seeking to determine whether nonalgebraic causal structures are *physically* ubiquitous in a given context, in the sense of arising with greater frequency. To do so, one would require a prior probability distribution over the set of possible causal structures, which will be context-dependent and generally difficult to estimate. Rather, we are here focused on the question: what is everything that can in principle be inferred about the causal structure from passive observation of the visible variables? Such inferences proceed by ruling out every causal structure for which there is a causal-compatibility constraint that is violated by the observed distribution. All of the causal structures in an observational equivalence class are ruled out (or ruled in) together. Consequently, when seeking to determine the significance of inequality constraints in such inferences, it is natural to determine how prevalent they are *among the observational equivalence classes*.

Here, we consider the question of what proportion of observational equivalence classes are nonalgebraic for sets of causal structures consistent with a fixed nodal ordering that have two, three and four visible nodes.

For causal structures consistent with a fixed nodal ordering that have two nodes, there are only two observational equivalence classes, and both are algebraic. For causal structures consistent with a fixed nodal ordering that have three nodes, we find that there are fifteen observational equivalence classes, of which five are nonalgebraic. Thus, the fraction of nonalgebraic observational equivalence classes is $1/3$. For causal structures consistent with a fixed nodal ordering that have four nodes, we find that this fraction is at least 0.852 (the uncertainty about this number comes from the fact that we do not yet have a full

[1] The discriminations one can make may generally also be more *coarse-grained* than the observational equivalence classes. Consider again data obtained from passive observation of the visible nodes. Without making further assumptions (such as appealing to the principle of faithfulness), this data will not uniquely pick out an observational equivalence class. For one, there might be two or more observational equivalence classes that are *incomparable* within the partial order but for which the corresponding sets of realizable distributions have a nontrivial intersection, and the distribution of interest might lie in this intersection. For another, for every class that can realize the distribution, the set of classes in its upward closure within the observational partial order can *also* realize the distribution.

classification of pDAGs with four visible variables into observational equivalence classes).

The trend that is emerging is that the fraction of nonalgebraic observational equivalence classes increases with the number of visible variables[2]. This suggests that as one increases the number of visible variables in the causal structures being considered, inequality constraints become generic among the observational equivalence classes.

Ref. [48] already identified which causal structures (i.e., which mDAGs) with 2, 3 or 4 visible nodes are nonalgebraic. However, only if one sorts these causal structures into observational equivalence classes can one determine what fraction of the observational equivalence classes are nonalgebraic.

In short, our results show that causal discovery algorithms that leverage only conditional independence constraints have a poor discriminatory power relative to algorithms that go beyond these for the case of four visible nodes, and this is likely to also be true for higher numbers of visible nodes.

Our result about the abundance of nonalgebraic observational equivalence classes has particular relevance for the study of quantum causal models: it indicates that the majority of observational equivalence classes of causal structures have the potential to show a quantum-classical gap. We will come back to the topic of quantum-classical gaps in Chapter 5.

The chapter is structured as follows: after an exposition of the relevant background concepts in Section 3.1, we will present our objectives and our results for the case of three visible variables in Section 3.2. In particular, Section 3.2.2 presents and discusses the complete observational partial order of causal structures of three visible nodes consistent with a fixed nodal ordering. We will then present the known rules to show observational dominance and nondominance in Sections 3.3 and 3.4 (including two new dominance-proving rules in Sections 3.3.2 and 3.3.2). In light of that, Section 3.5 will justify the results for causal structures with three visible nodes that were presented without justification in Section 3.2.2. Section 3.6 then presents some partial results about the case of causal structures of four visible nodes consistent with a fixed nodal ordering, and we end the chapter with concluding remarks in Section 3.7.

[2]Recent work by Evans [35] shows that DAGs that have nontrivial Nested Markov constraints always have nontrivial inequality constraints as well. That finding, however, is insufficient to establish the claim that the proportion of observational equivalence classes that have inequalities increases with the number of visible nodes.

3.1 Preliminaries

Most of the preliminary notions needed for this chapter were already introduced in the previous chapter. Here, we add to those a few extra concepts.

Recall the definitions of $\mathcal{M}_{\text{obs}}(\mathcal{G}, \vec{c}_{\text{vis}})$ and $\mathcal{M}\text{-}\texttt{profile}(\mathcal{G})$, from Definitions 4 and 5 respectively. The mathematical constraints on the set of realizable probability distributions of a given causal structure can be sorted into trivial and nontrivial types. The trivial type merely follows from nonnegativity of probabilities and normalization. The nontrivial types of constraint are those that are implied by the causal structure, and are termed *causal-compatibility constraints*. Using the observational profile, as we will see now, it is possible to classify pDAGs based on the types of causal-compatibility constraints that they present.

The set $\mathcal{M}_{\text{obs}}(\mathcal{G}, \vec{c}_{\text{vis}})$ is subject to both trivial and nontrivial constraints. To focus on the latter, we define $\tilde{\mathcal{M}}(\mathcal{G}, \vec{c}_{\text{vis}})$ as the set of distributions over the visible variables of $\mathcal{G}$ with cardinalities $\vec{c}_{\text{vis}}$ obeying *only* the nontrivial constraints. Note that, geometrically, $\mathcal{M}_{\text{obs}}(\mathcal{G}, \vec{c}_{\text{vis}})$ can be recovered from $\tilde{\mathcal{M}}(\mathcal{G}, \vec{c}_{\text{vis}})$ simply by restricting the latter to the simplex defined by the nonnegativity and normalization constraints. Out of the trivial constraints, normalization is a linear equality and nonnegativity is a linear inequality. In Ref. [65], it is proven that for classical causal models, all nontrivial causal-compatibility constraints also take the form of equalities or inequalities, that are polynomial on the probabilities. If all such constraints are equalities, then the set $\tilde{\mathcal{M}}(\mathcal{G}, \vec{c}_{\text{vis}})$ is an *algebraic set* (also termed an algebraic variety), [3] It is convenient to use the mathematical notion of algebraicness of sets to define a dichotomy among pDAGs, for which we will re-use the "algebraic" terminology.

Definition 19 (Algebraicness)**.** *A pDAG $\mathcal{G}$ will be said to be algebraic if, for every element $\mathcal{M}_{obs}(\mathcal{G}, \vec{c}_{vis})$ of $\mathcal{M}\text{-}\texttt{profile}(\mathcal{G})$, all nontrivial causal-compatibility constraints are equality constraints, or, equivalently, if the set $\tilde{\mathcal{M}}(\mathcal{G}, \vec{c}_{vis})$ (wherein one has dropped the trivial constraints of nonnegativity and normalization relative to $\mathcal{M}_{obs}(\mathcal{G}, \vec{c}_{vis})$) is an algebraic set. Otherwise, the pDAG $\mathcal{G}$ is said to be nonalgebraic.*

As established by Ref. [43], the only causal-compatibility constraints that are implied by a *latent-confounder-free* pDAG $\mathcal{G}$ are conditional independence relations, which are equality constraints that are valid regardless of the choice of cardinalities. Hence, all latent-confounder-free pDAGs are algebraic. One example of a nonalgebraic pDAG is the

[3] Note that the set $\mathcal{M}_{\text{obs}}(\mathcal{G}, \vec{c})$, unlike $\tilde{\mathcal{M}}(\mathcal{G}, \vec{c})$, is *never* an algebraic set simply because the nonnegativity constraints are inequality constraints.

causal structure appearing in Bell's theorem [8], because this structure implies causal-compatibility constraints that take the form of inequalities, namely, Bell inequalities [83].

Recall that if two pDAGs are observationally equivalent (Definition 7), they impose the same set of causal-compatibility constraints on the probability distributions over the visible variables. Therefore, it is clear that all of the pDAGs that are observationally equivalent to an algebraic pDAG are also algebraic, and all of the pDAGs that are observationally equivalent to a nonalgebraic pDAG are also nonalgebraic. This shows that the classification of pDAGs under observational equivalence and dominance is highly pertinent to the question of which pDAGs are algebraic and which are nonalgebraic.

The terminology of algebraic/nonalgebraic causal structures was introduced in Ref. [48]. Note, however, that we have here been more explicit than in Ref. [48] about the fact that this distinction relies on the observational profile of a pDAG, that is, the tuple obtained by looking at all possible cardinalities of the visible variables. This emphasis on cardinality is important because the presence of certain inequality constraints might depend on the choice of cardinalities of the visible variables. For example, the Bonet inequalities [11] for the instrumental scenario (first pDAG of Fig. 2.1.1) are constraints on distributions where the instrumental variable (associated to node a in Fig. 2.1.1) has cardinality at least 3. On the other hand, Pearl's instrumental inequalities [56] for the same pDAG have to be satisfied even when all variables have binary cardinalities.

Our definition of algebraicness is such that a pDAG only needs to present inequality constraints for *one* choice of cardinalities of the visible variables to be considered nonalgebraic. This captures the potential that a causal structure has of presenting inequality constraints in *some* scenario, instead of being a definition that is only applicable to a specific causal discovery scenario (in which case the cardinality of the visible variables $\vec{c}_{\text{vis}}$ would be known, and we would not need to discuss the entire observational profile). As of now, however, we do not know of any example of a pDAG that is nonalgebraic while also being free of inequality constraints when all visible variables are binary, i.e., have cardinality 2. To highlight this open problem, we conjecture that such examples do not exist.

Conjecture 1. *If a pDAG is nonalgebraic, this implies inequality constraints even when the cardinalities of all the visible variables have their smallest possible nontrivial values, i.e.,* $\vec{c}_{vis} = (2, 2, \dots)$.

In Section 3.6.3, we will present possible candidates for counterexamples to Conjecture 1: the causal structures of Fig. 3.6.2(a)-(d) are known to present inequality constraints at cardinalities $\vec{c}_{\text{vis}} = (3, 2, 2, 2)$, but we still do not know if they present inequality constraints at cardinalities $\vec{c}_{\text{vis}} = (2, 2, 2, 2)$.

Unlike inequality constraints, the equality constraints implied by a causal structure do not depend on the cardinality of the visible variables. Furthermore, it is much easier to obtain equality constraints than inequality constraints. The simplest type of equality constraints, namely the conditional independence relations, can be obtained from a graphical criterion called the *d-separation criterion* [80, 42]. Given a pDAG $\mathcal{G}$ and three disjoint sets of visible nodes $A, B, C \subseteq \texttt{nodes}(\mathcal{G})$, the criterion returns whether A and B are "d-separated" by C or not. The affirmative case is denoted by $A \perp\!\!\!\perp_{\text{d}} B|C$. As shown in Ref. [80] and reproduced in Theorem 4 below, a d-separation relation of a pDAG implies that the realizable distributions have to satisfy an associated conditonal independence relation. In a certain distribution P, the variable X_A is said to be independent of the variable X_B after conditioning on the variable X_C if:

$$P(X_A X_B|X_C) = P(X_A|X_C)P(X_B|X_C). \tag{3.1}$$

Such a conditional independence relation is denoted by $X_A \perp\!\!\!\perp X_B|X_C$. Note that it has to be satisfied regardless of the cardinality of X_A, X_B and X_C.

Theorem 4 (d-separation and conditional independence [80])**.** *Let $\mathcal{G}$ be a pDAG, and associate it to a classical unrestricted semantics under arbitrary cardinalities of the latent variables. Let A, B and C be three disjoint subsets of the set of visible nodes, $\texttt{vis}(\mathcal{G})$. Then:*

1. *If P is a probability distribution over $X_{\texttt{vis}(\mathcal{G})}$ realizable by $\mathcal{G}$, then $A \perp\!\!\!\perp_d B|C$ in $\mathcal{G}$ implies that $X_A \perp\!\!\!\perp X_B|X_C$ in P.*
2. *If $X_A \perp\!\!\!\perp X_B|X_C$ for every probability distribution over $X_{\texttt{vis}(\mathcal{G})}$ that is realizable by $\mathcal{G}$, then $A \perp\!\!\!\perp_d B|C$.*

3.2 Objectives and Methods

3.2.1 Special Cases: Latent-confounder-free and Directed-edge-free mDAGs

In this section, we discuss two special cases whose observational partial order is completely solved for any number of visible nodes: the case of latent-confounder-free mDAGs and the case of directed-edge-free mDAGs.

Most of the causal inference literature studying observational equivalence discusses only latent-confounder-free mDAGs, or, equivalently, only latent-confounder-free pDAGs [79, 2, 21]. That literature, however, classifies the set of such mDAGs for *all* nodal orderings of the visible nodes rather than a fixed one. When we restrict our attention to one nodal ordering of the visible nodes, the organization of latent-confounder-free mDAGs into observational equivalence classes is trivial: for a given nodal ordering of the nodes, if two latent-confounder-free mDAGs are different then they are observationally inequivalent. This fact seems to be well-known (see, e.g., below Theorem 1.2.8 in Ref. [57]). For pedagogical purposes, we provide a proof in Appendix B.1. Here, we adapt this proof to show something more, that finding the observational *partial order* is also simple when one is restricted to the latent-confounder-free mDAGs for a fixed nodal ordering of the nodes:

Proposition 1 (Consequence of Appendix B.1)**.** *Let* $\mathfrak{G}$ *and* $\mathfrak{G}'$ *be two latent-confounder-free mDAGs that are consistent with the same nodal ordering. Then,* $\mathfrak{G}$ *observationally dominates* $\mathfrak{G}'$ *if and only if it structurally dominates it.*

Proof. The "if" side follows from Lemma 3. To prove the "only if" side, assume that $\mathfrak{G}$ *does not* structurally dominate $\mathfrak{G}'$. Since they are latent-confounder-free, this implies that there is an arrow between two nodes that is present in $\mathfrak{G}'$ but not in $\mathfrak{G}$. In Appendix B.1, we showed that this implies that there is a d-separation relation that is implied by $\mathfrak{G}$ but not by $\mathfrak{G}'$. By point 2 of Theorem 4, this implies that there is a probability distribution P realizable by $\mathfrak{G}'$ that violates the conditional independence relation that is implied by this d-separation relation. By point 1 of Theorem 4, P is not realizable by $\mathfrak{G}$. Therefore, $\mathfrak{G}$ *does not* observationally dominate $\mathfrak{G}'$. □

Therefore, the observational partial order of the set of latent-confounder-free mDAGs with a fixed nodal ordering is simply the partial order defined by structural dominance. The case of for 3-node and 4-node mDAGs of this type are shown in Figures 7.3(a) and 7.5 of Ref. [3].

For the case of directed-edge-free mDAGs, the concept of nodal ordering is not relevant, since none of the visible nodes is an ancestor of any other visible node. In Proposition 6.8 of Ref. [33], it was shown that two different directed-edge-free mDAGs are always observationally inequivalent. Here, we extend this idea to show that observational dominance for directed-edge-free mDAGs is determined by structural dominance:

Proposition 2 (Consequence of Ref. [75, Example 2])**.** *Let* $\mathfrak{G}$ *and* $\mathfrak{G}'$ *be two directed-edge-free mDAGs. Then,* $\mathfrak{G}$ *observationally dominates* $\mathfrak{G}'$ *if and only if it structurally dominates it.*

Proof. The "if" side follows from Lemma 3. To prove the "only if" side, assume that $\mathfrak{G}$ *does not* structurally dominate $\mathfrak{G}'$. Since they are directed-edge-free, this implies that there is a face that is present in the simplicial complex of $\mathfrak{G}'$ but not in the simplicial complex of $\mathfrak{G}$. That is, there is a set S of visible nodes in $\texttt{can}(\mathfrak{G}')$ that shares a latent common cause, while it does not share a latent common cause in $\texttt{can}(\mathfrak{G})$. Since there are no arrows between visible nodes, this implies that the nodes in S share a common ancestor in $\texttt{can}(\mathfrak{G}')$ but not in $\texttt{can}(\mathfrak{G})$.

In Example 2 of Ref. [75], it was shown that a distribution where all the variables $\{X_1, ..., X_n\}$ are perfectly correlated, i.e., where the only points with nonzero weight are those where $X_1 = X_2 = ... = X_n$, is only realizable by causal structures where all the nodes corresponding to such variables share a common ancestor.[4] Therefore, $\texttt{can}(\mathfrak{G}')$ can explain perfect correlation between the variables of X_S, while $\texttt{can}(\mathfrak{G})$ cannot. Thus, $\mathfrak{G}$ *does not* observationally dominate $\mathfrak{G}'$. □

Therefore, the observational partial order of directed-edge-free mDAGs is the partial order defined by structural dominance. The case of for 3-node and 4-node mDAGs of this type are shown in Figures 7.3(b) and 7.6 of Ref. [3].

3.2.2 Full Solution of Observational Partial Order for 3-node mDAGs

Now, let us turn to the problem of classifying the entire set of mDAGs with a given number of nodes and a fixed nodal ordering. As we will see, this problem is still open for a general number of nodes: we do not yet know the necessary and sufficient conditions for two mDAGs to be observationally equivalent or to observationally dominate each other in the case of an arbitrary number of nodes.

For mDAGs with two nodes, of course, this problem is completely solved but the solution is rather trivial, as we saw in Figure 2.4.2. For mDAGs with three nodes, the problem is also completely solved, but the solution is more interesting in this case. In this section, we will describe this order; the proofs of the ordering relations are postponed to Section 3.5.

There are 72 3-node mDAGs consistent with a fixed nodal ordering. In Fig. 3.2.1, we group these into observational equivalence classes, and present the observational dominance relations that hold among the classes. In order to avoid having to explicitly label the nodes

[4] Ref. [75] does not consider causal structures with latent variables. However, this result is easily generalizable to that case, because a pDAG $\mathcal{G}$ with latent variables can realize perfect correlation between the visible variables $\{X_1, ..., X_n\}$ if and only if the latent-confounder-free pDAG obtained by transforming all of the latent nodes of $\mathcal{G}$ into visible nodes can also realize it.

throughout the figure, we adopt a simplifying convention wherein the position of the node in a certain mDAG indicates its label: in each mDAG the bottom node is a, the top-left node is b and the top-right node is c. All of the mDAGs of this figure follow the nodal ordering (a, b, c). In this figure, the highest elements in the observational partial order appear at the top of the diagram and the lowest elements appear at the bottom, and if one class dominates another with no classes falling between them in the observational partial order, we draw a line from the former to the latter.

3.2.3 Significance of observational partial order for causal discovery

We recall the problem of causal discovery: the input data is a sample from a probability distribution over visible variables obtained from passive observations, and the objective is to infer which causal structures can provide an explanation of this data. In general, it is not possible to single out the pDAG or even the mDAG that generated the data, because many different causal structures might be able to realize it. Nonetheless, one can seek to answer the question: what is *everything* that can be learned about the underlying causal structure of a given phenomenon when we only have access to passive observations? We here explain how determining the observational partial order is useful for this problem.

We begin with a limiting case of the causal discovery problem, namely, one wherein the sample size is infinite rather than finite. Only in this case do the relative frequencies computed from the sample reflect the true probability distribution, whereas in the case of a finite sample, the relative frequencies merely yield an *estimate* of the true probability distribution. In other words, we begin by considering the version of the causal discovery problem wherein the input is the true probability distribution.

The most basic respect in which the observational partial order is relevant to this problem is the following. If one determines that a distribution is not realizable by a particular mDAG $\mathfrak{G}$, then one can conclude that it is also not realizable by any of the mDAGs that are *equivalent to or below* $\mathfrak{G}$ in the observational partial order. Similarly, if one determines that a given distribution *is* realizable by a particular mDAG $\mathfrak{G}'$, then one can conclude that it is also realizable by any of the mDAGs that are *equivalent to or above* $\mathfrak{G}'$ in the observational partial order.

It is useful here to make use of the notions of *upward closure* and *downward closure* of a set of elements in a partial order. If S denotes a set of elements of a partial order, then the upward closure of S, denoted $\mathrm{UC}(S)$, includes every x such that $x \succeq y$ for some $y \in S$.

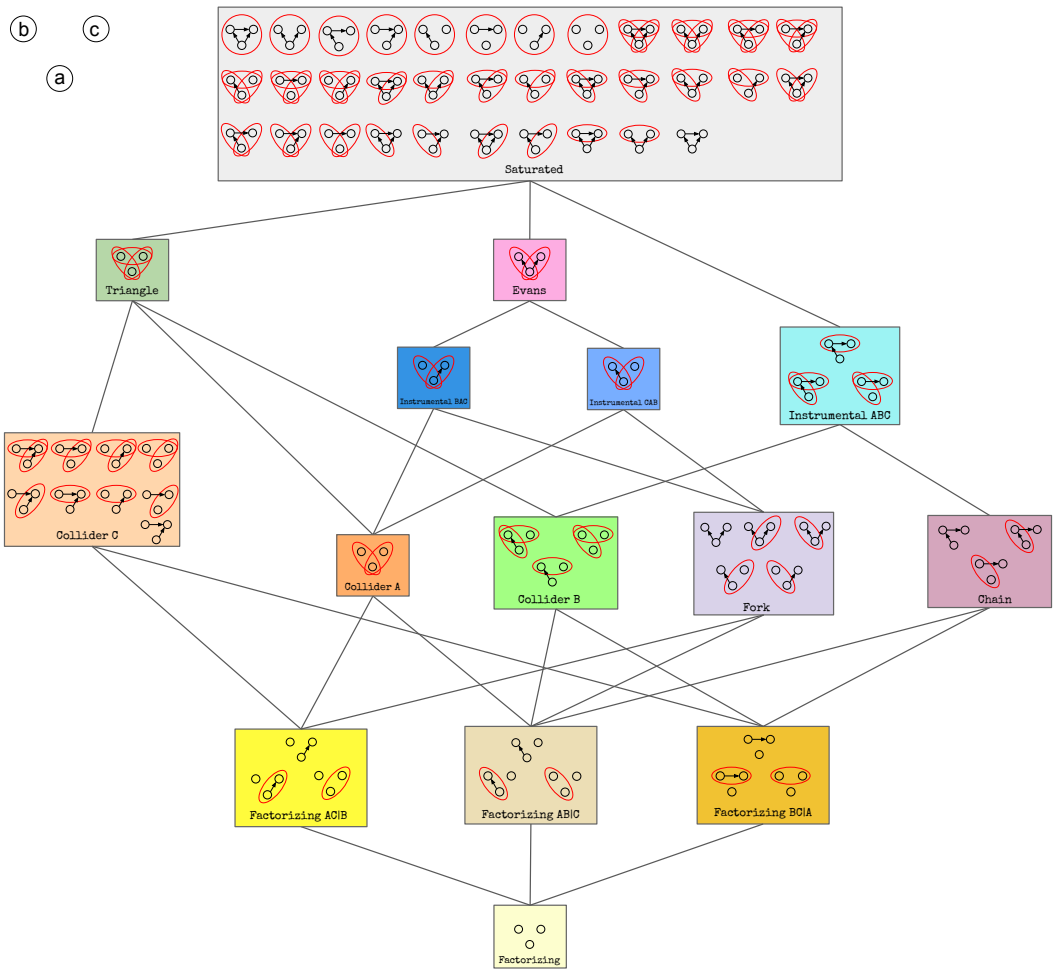


Figure 3.2.1: The 72 mDAGs with three visible nodes with nodal ordering (a, b, c) organized into observational equivalence classes and the observational dominance relations that hold between these classes. In each mDAG of this figure, the bottom node is a, the top-left node is b and the top-right node is c.

Similarly, the downward closure of S, denoted $\mathrm{DC}(S)$, includes every x such that $x \preceq y$ for some $y \in S$.

Using these notions, if a distribution is known to be realizable by all of the mDAGs of a set S, then one can infer that it is also realizable by all of the mDAGs of $\mathrm{UC}(S)$. Similarly, if a distribution is known to *not* be realizable by any mDAG in a set S', then one can infer that it is also not realizable by any mDAG in $\mathrm{DC}(S')$. If the union of $\mathrm{UC}(S)$ and

$\mathrm{DC}(S')$ is not the full set of mDAGs, then the remaining mDAGs are those for which one remains uncertain whether they can realize the distribution or not. Note that to specify what one knows about realizability, it suffices to specify the least elements of $\mathrm{UC}(S)$ and the greatest elements of $\mathrm{DC}(S')$. In general, each of these will be a set of incomparable elements.

The observational partial order of 3-node mDAGs, depicted in Fig. 3.2.1, helps to illustrate the idea. For instance, it could be known that a probability distribution over three variables is realizable by `Collider C` but not by `Instrumental CAB` or `Instrumental ABC`. In this case, one can infer that it is realizable by the upward closure of `Collider C`, namely, `Collider C`, `Triangle` and `Saturated`, and one can infer that it is not realizable by the downward closure of the set consisting of `Instrumental CAB` and `Instrumental ABC`. Meanwhile it remains unknown whether it is realizable by any mDAG that is in neither of these sets, namely, `Instrumental BAC` and `Evans`.

In the case where we learned everything there is to learn about the causal structure from passive observations, we have that the upward closure $\mathrm{UC}(S)$ of a set S of mDAGs *can* realize our probability distribution, and that all of the mDAGs outside of $\mathrm{UC}(S)$ *cannot* realize it. In general, therefore, the set of causal structures that could in principle realize a given probability distribution over the visible variables (in a passive observation scheme) constitutes the elements of the upward closure (within the partial order) of a set S of observational equivalence classes of mDAGs

Without appealing to interventional experiments or to any model selection principle such as faithfulness (which will be discussed below), the most we can say about a given observational distribution is that it can be realized by all of the mDAGs in a set of the form $\mathrm{UC}(S)$.

This discussion assumed the limit of infinite sample size, but the same ideas extend to the case of finite sample size. Using standard techniques of model selection, we can sometimes conclude that a causal structure $\mathfrak{G}$ *underfits* the finite sample of data (e.g., its chi-squared error is too large). In this case, so too will all the causal structures that are below it in the observational partial order. The reason is as follows. For $\mathfrak{G}$ to *underfit* the finite sample means that the finite sample is improbable relative to all of the probability distributions realizable by $\mathfrak{G}$, i.e., that it is unlikely that the finite sample came from one of the probability distributions realizable by $\mathfrak{G}$. But for all of the causal structures below $\mathfrak{G}$ in the observational partial order, the set of probability distributions that they can realize is a *strict subset* of those that can be realized by $\mathfrak{G}$, and therefore the finite sample must be improbable relative to all these distributions as well. By a similar reasoning, if an mDAG *does not* underfit the data, then all of the mDAGs that are above it in the observational

partial order will also not underfit the data.

In summary, if a set S' of mDAGs underfits the finite sample of data, we discard all of the mDAGs in $\mathrm{DC}(S')$ as candidate causal structures to explain the data. If a set S of mDAGs does not underfit the data, then all elements of $\mathrm{UC}(S)$ are candidate causal structures.

Using model-selection principles to adjudicate between possible candidates

We established that the upward closure of a set of mDAGs in the observational partial order is the most we can learn about the underlying causal structure when we only have access to passive observations. Now, we will discuss how it is possible to make finer-grained estimates of the causal structure by appealing to certain model-selection principles. We begin with the case of an infinite sample of data.

Among all of the mDAGs that *can* realize a distribution, one might wonder whether any one of them provides as good of a causal explanation as any other. It is generally thought that this is not the case and that one should prefer causal structures that are *lower* in the observational partial order.

For instance, in Pearl's textbook on causality [57], he argues for this on the grounds that such causal models are more falsifiable.[5] Indeed, Pearl does not introduce any language for referring to observational dominance besides the language of preference. (Pearl's order relation of preference is opposite to the order relation of observational dominance in the sense that $\mathfrak{G}$ is preferred to $\mathfrak{G}'$ just in case $\mathfrak{G}'$ observationally dominates $\mathfrak{G}$.)

According to this proposal, the causal structures providing the best explanation of the distribution are the *least elements* in the upward-closed set of classes that can realize it. This is referred to as the principle of *minimality* in Ref. [57] (see Definition 2.3.4).

It is worth noting that preferring causal models that satisfy the principle of *faithfulness* [74] (referred to as *stability* in Ref. [57] and *no fine-tuning* in Ref. [83]) is also an instance of preferring classes lower in the observational partial order.

In the case of a finite sample of data, it is also possible to adjudicate between a set of hypotheses about the causal structure each of which *do not* underfit the data. It suffices to use a standard technique of statistical model selection: the train-and-test methodology. In the train-and-test methodology, the observational data is divided into two sets, the first

[5] For instance, on p. 46 of Ref. [57], he states: "such theories are more constraining and thus more falsifiable; they provide the scientist with less opportunities to overfit the data 'hindsightedly' and therefore command greater credibility if a fit is found."

one being the *training set* and the second one being the *test set*. For each of the candidate causal structures, we find the parameters that provide the best fit to the training set. One then compares the predictions of this best-fit model to the test set to compute the test error. One can then compare the test error of different causal structures to compare how well the best-fit model of each causal structure generalizes to new data, that is, to compare their predictive power.

Consider a comparison of two causal structures, $\mathfrak{G}$ and $\mathfrak{G}'$. If $\mathfrak{G}'$ gives a higher test error than $\mathfrak{G}$ but a lower training error, then this suggests that the causal model associated to $\mathfrak{G}'$ is mistaking statistical fluctuations that appear in the training data for a *stable signal*, i.e., a relationship that would reappear in fresh samples.

For example, imagine that two variables X_a and X_b are associated with nodes a and b that are d-separated given the empty set in $\mathfrak{G}$, but are connected by an arrow $a \to b$ in $\mathfrak{G}'$. Suppose that $\mathfrak{G}$ is the ground truth. In this case, one would find that the variables X_a and X_b are statistically independent in the limit of an infinite sample of data. In the finite sample constituting the training data, however, statistical fluctuations may be such that these variables exhibit some statistical dependence. While all models on $\mathfrak{G}$ are forced to account for this dependence as a fluctuation, models on $\mathfrak{G}'$ might identify it as a signature of a causal influence of a on b. This in turn might make the latter models perform worse on the test data, because the test data has different fluctuations from the training data. When this happens, we say that $\mathfrak{G}'$ *overfits* the data relative to $\mathfrak{G}$.[6] This train-and-test methodology was applied in Ref. [27] to adjudicate between different causal explanations of a finite sample of data in a quantum experiment that violated Bell inequalities.

We now turn to the significance of the observational partial order for adjudicating between different candidate causal structures that *do not* underfit the finite sample of data.

Suppose $\mathfrak{G}'$ is above $\mathfrak{G}$ is the observational partial order. If $\mathfrak{G}'$ overfits the data relative to $\mathfrak{G}$, then everything *above* $\mathfrak{G}'$ in the order *also* overfits the data relative to $\mathfrak{G}$. The reason is that the extra expressive power of $\mathfrak{G}'$ relative to $\mathfrak{G}$ (i..e., the additional probability distributions realizable by $\mathfrak{G}'$) is what must lead $\mathfrak{G}'$ to mistake statistical fluctuations in the training data for stable signals that would reappear in a fresh sample. But this extra expressive power is also present in all of the structures above $\mathfrak{G}'$ in the observational partial order, and so these must be led to make the same mistake.

Note that if $\mathfrak{G}$ fits (i.e., does not underfit) the finite sample of data and $\mathfrak{G}'$ is a causal

[6] This is similar to the way in which the error of a polynomial fit to training data can be made to decrease by increasing the order of the polynomial, but because this low error is achieved by treating statistical fluctuations as real features, the fit will tend to have a large error when predicting the test data.

structure higher in the observational order, then $\mathfrak{G}'$ does not necessarily overfit the data relative to $\mathfrak{G}$ simply by virtue of having additional expressive power. This is because the extra expressive power may be of a type that does not lead $\mathfrak{G}'$ to mistake statistical fluctuations for a stable signal. In this case, $\mathfrak{G}$ and $\mathfrak{G}'$ will have roughly *equal* test errors. This point was emphasized in Ref. [26], in response to a contrary claim (that one model being more expressive than another generically *did* cause the first to overfit the data relative to the second.) Nonetheless, it is worth considering the question of the circumstances in which one causal structure *does* tend to overfit the data relative to a causal structure that is directly below it in the observational partial order.

One such circumstance is if the lower causal structure implies an equality constraint on the realizable distributions that does not hold in the higher causal structure. Suppose that data is generated from the lower causal structure. Because it is a finite sample of data, the equality constraint will generically not be satisfied exactly—there will be statistical fluctuations away from its exact satisfaction. Because the higher causal structure can realize distributions that violate the equality constraint, it can fall into the trap of mistaking such fluctuations for stable signals.

Consider, for example, the pair of observational equivalence classes `Collider C` and `Triangle`. The class `Collider C` presents the d-separation relation $a \perp_d b$, so the set of distributions on three variables that are realizable by causal structures in that class satisfy $X_a \perp\!\!\!\perp X_b$, that is, the equality constraint $P(X_aX_b) = P(X_a)P(X_b)$. This forces the probability distributions to lie on an algebraic variety that has a smaller dimension than the full space of such distributions. The set of distributions realizable by the class `Triangle`, on the other hand, is *not* confined to this variety. In fact, since `Triangle` exhibits no equality constraints, its set of realizable distributions has the same dimension as the full space of distributions on three variables. Now, suppose the finite sample of data is generated by a causal structure in the class `Collider C`. Due to statistical fluctuation, the relative frequencies will generically describe a point slightly outside of this variety. A causal structure in the `Triangle` class can fit this finite sample of data to a probability distribution that is outside of this variety, and thereby interpret the statistical fluctuation as a stable signal. (Clearly, the `Saturated` class is *also* prone to such overfitting, given that the set of distributions realizable by it are also not confined to this variety). An example of this type, but involving a pair of 4-node mDAGs, is discussed in Ref. [27] where such a train-and-test methodology is implemented on data from a quantum experiment.

A second circumstance in which one causal structure may overfit a finite sample of data relative to a causal structure that is directly below it in the observational partial order is when the two are distinguished by an inequality constraint. An inequality constraint does not lower the dimension of the space of realizable distributions, but it creates a boundary

below which the distribution is realizable, and above which it is not. In this case, overfitting can arise when the causal structure generating the data implies an inequality constraint, but statistical fluctuations are taking the finite-sample relative frequencies outside the boundary.

Suppose, for instance, that the `Instrumental BAC` class fits the data well, while the `Evans` class overfits it relative to `Instrumental BAC`. Because `Instrumental BAC` only has inequality constraints, it must be that statistical fluctuations away from their satisfaction in the finite sample of data are being mistaken by `Evans` for a stable signal, and leading to a high test error.

Of course, the data generated must be very close to the boundary described by the inequality constraint in order for statistical fluctuations to lead to a violation of this constraint in the finite-sample relative frequencies. Consequently, we do not expect this circumstance to arise for generic parameter values (most of which lead to distributions that lie far from the boundary). The possibility of this type of overfitting is discussed further in Appendix A.2 of Ref. [26].

Attesting to the presence of inequality constraints

Another reason that classifying mDAGs into observational equivalence classes is useful is to identify which ones are nonalgebraic. In Ref. [35] it was proven that an mDAG is algebraic *if and only if* it is observationally equivalent to some latent-confounder-free mDAG. Note that this result does not require the latent-confounder-free mDAG in question to be in the set of mDAGs consistent with a fixed nodal ordering that we have been considering in our observational partial order. In Fig. 3.2.1, for example, the class `Collider B` does not include any latent-confounder-free mDAG, but it is nevertheless algebraic. This is so because its elements are observationally equivalent to the latent-confounder-free mDAG $a \to b \leftarrow c$, that is not shown in the figure because it is not consistent with the fixed nodal ordering (a, b, c). To be able to easily refer to this idea later on, we formulate it as a proposition:

Proposition 3 (Trivial consequence of Ref. [35, Theorem 1.1])**.** *An observational equivalence class of mDAGs consistent with a fixed nodal ordering is algebraic if and only if its mDAGs are observationally equivalent to a latent-confounder-free mDAG, which might not be consistent with the fixed nodal ordering.*

Using Proposition 3 to analyze the classes of Fig. 3.2.1, we can see that among equivalence classes of 3-node mDAGs consistent with a fixed nodal ordering, the only nonalgebraic

ones are `Triangle`, `Evans`, `Instrumental BAC`, `Instrumental CAB` and `Instrumental ABC`. (The nonalgebraicness of all of these mDAGs was in fact already known [39, 56, 33]).

3.2.4 How to Express Partial Information about the Observational Equivalence Partition

In Fig. 3.2.1, we presented the complete observational partial order of 3-node mDAGs consistent with a fixed nodal ordering. This was only possible because, as will be made explicit in Section 3.5, the set of rules to show observational dominance and nondominance that are currently known happen to be sufficient to completely determine the observational partial order in this case. However, in general there is not yet a set of rules that is proven to be necessary and sufficient. As we will see, there are sets of 4-node mDAGs consistent with a fixed nodal ordering whose observational equivalence remains uncertain.

The first step towards obtaining an observational partial order is to find the observational equivalence classes. These classes establish a partition of the set of all n-node mDAGs consistent with a fixed nodal ordering, which we will call the *observational equivalence partition*. An element of a partition is called a *block*. This section will give the means to express partial information about the observational equivalence partition, for cases where we cannot present its full characterization.

Fig. 3.2.2 illustrates how we express partial information about the observational equivalence partition. In this figure, the set of red points represent the set of n-node mDAGs consistent with a fixed nodal ordering, and the green dashed loops represent the observational equivalence partition. Partial information about the observational equivalence partition can be given in the form of two other partitions, which we term the *proven-equivalence partition*, where two mDAGs belong to *the same* block of the partition when they are proven to be equivalent, and the *proven-inequivalence partition*, where two mDAGs belong to *different* blocks of the partition when they are proven to be inequivalent. Examples are depicted in Fig. 3.2.2 by the pink and gray loops respectively.

Each of the latter partitions can be built up in a step-wise fashion. For the proven-equivalence partition, one begins with a full partition of the mDAGs consistent with a fixed nodal ordering into singleton sets. When one proves that two mDAGs are observationally equivalent, one can merge the two blocks containing these mDAGs into a single block. The successive application of different rules for proving observational dominance (which will be presented in Section 3.3) reveals more and more observational equivalences, thereby merging blocks of the proven-equivalence partition and lowering the total number of such blocks. This number is an upper bound on the number of observational equivalence classes.

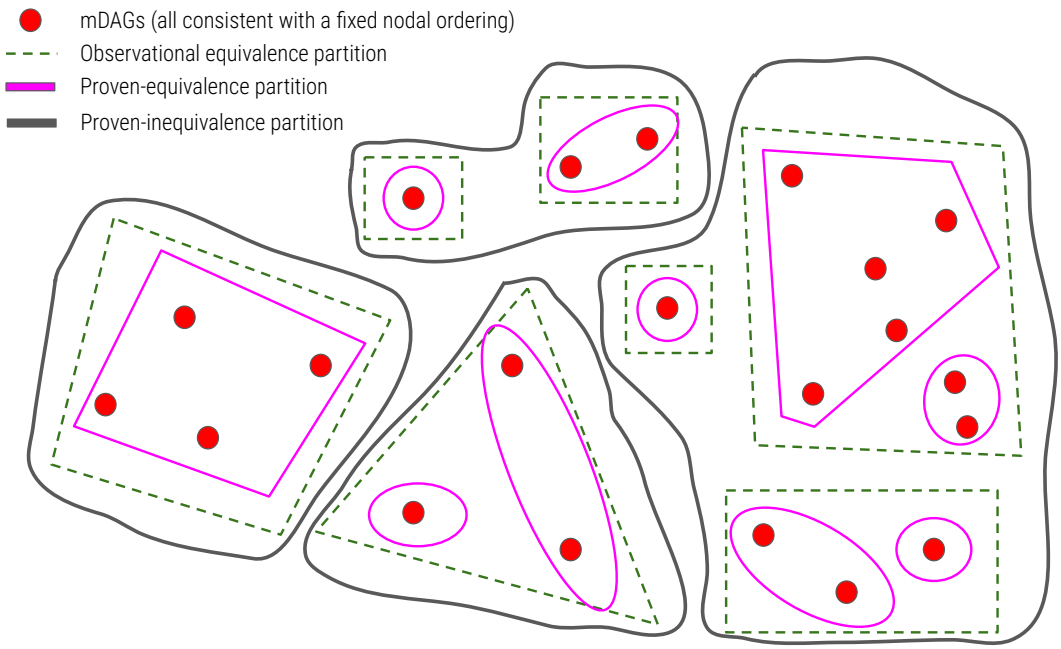


Figure 3.2.2: The proven-equivalence partition (pink loops) is found by applying the equivalence rules of Section 3.3. The proven-inequivalence partition (gray loops) is found by the inequivalence rules of Section 3.4. The final goal is to find the observational equivalence partition (green dashed loops).

For a given set R of dominance-proving rules, the proven-equivalence partition one can obtain from this set is referred to as the proven-equivalence partition *induced by R*.

For the proven-inequivalence partition, one begins with a partition having all mDAGs in a single block. When one proves that two mDAGs are observationally inequivalent, one proves that they must be in different blocks of the proven-inequivalence partition. In fact, one thereby proves that the blocks of the proven-equivalence partition to which each of these mDAGs belong are observationally inequivalent. In other words, the blocks of the proven-inequivalence partition are always coarse-grainings of the blocks of the proven-equivalence partition.

Successive application of different rules for proving observational nondominance (which will be presented in Section 3.4) reveals more and more observational inequivalences, thus splitting blocks of the proven-inequivalence partition and raising the total number of such blocks. This number is a lower bound on the number of observational equivalence classes. For a given set R of nondominance-proving rules, the proven-inequivalence partition one

can obtain from this set is referred to as the proven-inequivalence partition *induced by R*.

As we will see, the comparison of the set of d-separation relations of each mDAG is a nondominance-proving rule. The blocks of the proven-inequivalence partition induced by the comparison of d-separation relations are what are conventionally referred to as the Markov equivalence classes [1].

If a given set of mDAGs is a block of *both* the proven-equivalence partition and the proven-inequivalence partition, then it is a block of the observational equivalence partition. The leftmost block of Fig. 3.2.2 illustrates this circumstance. This is the method we use to characterize what we can prove about the observational equivalence classes of 3-node and 4-node mDAGs consistent with a fixed nodal ordering.

To obtain the proven-equivalence partition from a set of dominance-proving rules, we start by constructing a partially ordered set of mDAGs with the partial order induced by such rules. This will be called the *proven-dominance partial order*. The proven-equivalence partition is then determined by taking pairs of mDAGs such that each dominates the other in the proven-dominance partial order to be part of the same block.

3.3 Rules for Establishing Observational Dominance

In this section, we will present the known rules that provide conditions under which one mDAG is known to observationally dominate another. In Fig. 3.3.1 presents a diagram with all of the dominance-proving rules that we will use here. The rules that are indicated in thick green rectangles are our original theoretical results, namely, two generalizations of a rule presented in Ref. [33].

The first dominance-proving rule was already given in Lemma 3: structural dominance implies observational dominance.

3.3.1 HLP Edge-Adding Rule

The dominance-proving rule we present in this section is Theorem 26.4 of the work of Henson, Lal and Pusey[47], that we are going to call the *HLP Edge-Adding Rule*:

Proposition 4 (HLP Edge-Adding Rule (HLP) [47])**.** *Let* $\mathfrak{G} = \{\mathcal{D}, \mathcal{B}\}$ *be an mDAG, and let* x *and* y *be two of its nodes. Let* $\mathfrak{G}'$ *be the mDAG obtained from* $\mathfrak{G}$ *by adding a directed edge* $x \to y$.

Suppose that:

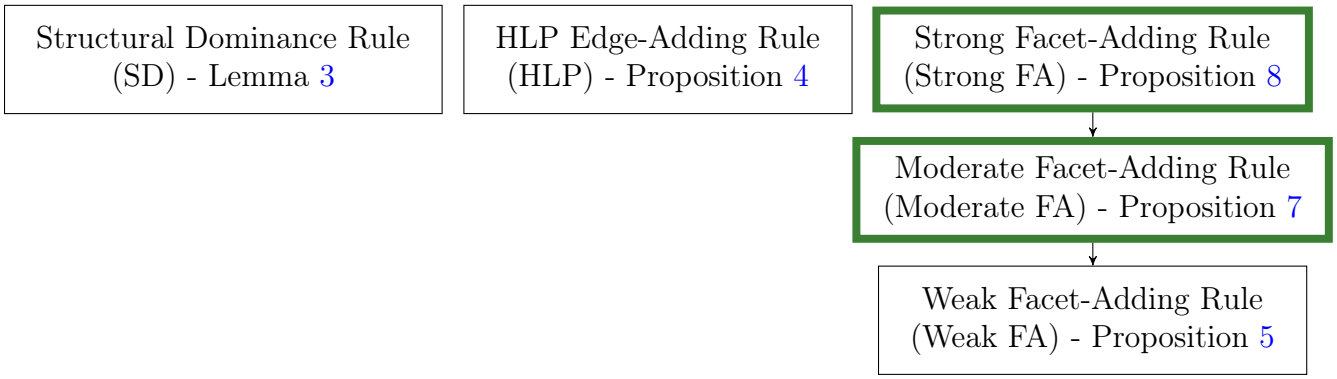


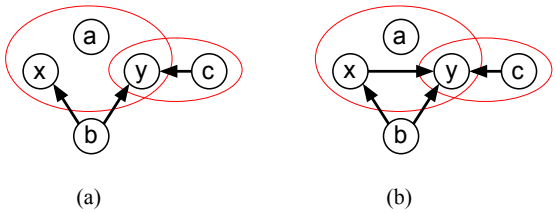
Figure 3.3.1: Classification of the different dominance-proving rules that we use. Classification of the different nondominance-proving rules. An arrow from one rule to another indicates that the latter is strictly less powerful than the former (i.e., the latter is subsumed in the former). The rules indicated in thick green rectangles are original results of this chapter.

1. $\mathtt{pa}_{\mathcal{D}}(x) \subseteq \mathtt{pa}_{\mathcal{D}}(y)$,
2. *Whenever* $x \in B$ *for a facet* $B \in \mathcal{B}$, *then also* $y \in B$.

In this case, $\mathfrak{G}$ *observationally dominates* $\mathfrak{G}'$, *i.e.,* $\mathfrak{G} \succeq \mathfrak{G}'$.

Note that, since $\mathfrak{G}'$ *structurally dominates* $\mathfrak{G}$, *by Lemma 3 we know that* $\mathfrak{G}'$ *observationally dominates* $\mathfrak{G}$, *i.e.,* $\mathfrak{G}' \succeq \mathfrak{G}$. *Therefore,* $\mathfrak{G}$ *and* $\mathfrak{G}$ *are observationally equivalent, i.e.,* $\mathfrak{G} \cong \mathfrak{G}'$.

For example, from the HLP Edge-Adding Rule it follows that the two mDAGs of Fig. 3.3.2 are observationally equivalent.



(a) (b)

Figure 3.3.2: Two mDAGs that can be proven to be observationally equivalent by the HLP Edge-Adding Rule [47].

Note that this dominance-proving rule establishes observational *equivalence*, because

the conditions for its applicability already imply observational dominance in the other direction. This will be also the case for the next dominance-proving rules.

Furthermore, note that the HLP Edge-Adding rule can establish observational dominance between two mDAGs that are *not* consistent with the same nodal ordering: if nodes x and y have the same set of parents and belong to the same facets, we can first use this rule in reverse to remove the arrow $x \to y$, and then use the rule to add the arrow $y \to x$. Because of that, it is conceivable that there exist observational dominances between mDAGs $\mathfrak{G}$ and $\mathfrak{G}'$ consistent with the same nodal ordering that can only be discovered by first showing that $\mathfrak{G}$ dominates some mDAG that is *not* consistent with the same nodal ordering, and then showing that this mDAG dominates $\mathfrak{G}'$. Therefore, when constructing the observational partial order of mDAGs that follow a fixed nodal ordering, we first implement the dominance-proving rules of this section for the set of *all* mDAGs, and then restrict our presentation to those that are consistent with a fixed nodal ordering.

3.3.2 Facet-Adding Rules

In this section, we will present three more dominance-proving rules. As we will explain, the first rule is obtained from a result of Ref. [33], and the second and third rules are generalizations of the first.

Weak Facet-Adding

Proposition 5 (Weak Facet-Adding (Weak FA) - Consequence of Ref. [33, Proposition 6.1])**.** *Let $\mathfrak{G} = \{\mathcal{D}, \mathcal{B}\}$ be an mDAG whose simplicial complex $\mathcal{B}$ contains two disjoint faces C and D. If $B = C \cup D$ is not already present in the simplicial complex of $\mathfrak{G}$, let $\mathfrak{G}'$ be the mDAG obtained by starting from $\mathfrak{G}$ and adding B and all of the faces contained in B to its simplicial complex.*

Suppose that:

1. $\mathtt{pa}_{\mathcal{D}}(C) \cup C \subseteq \mathtt{pa}_{\mathcal{D}}(d)$ *for each* $d \in D$,
2. *Every face that contains any* $c \in C$ *is a subset of* C.

In this case, $\mathfrak{G}$ observationally dominates $\mathfrak{G}'$, i.e., $\mathfrak{G} \succeq \mathfrak{G}'$.

Note that, since $\mathfrak{G}'$ structurally dominates $\mathfrak{G}$, by Lemma 3 we know that $\mathfrak{G}'$ observationally dominates $\mathfrak{G}$, i.e., $\mathfrak{G}' \succeq \mathfrak{G}$. Therefore, $\mathfrak{G}$ and $\mathfrak{G}'$ are observationally equivalent, i.e., $\mathfrak{G} \cong \mathfrak{G}'$.

Note that condition 2 implies that C is not only a face, but also a facet of the simplicial complex of $\mathfrak{G}$. The face D, on the other hand, does not need to be a facet. We call this rule "weak" because it will be generalized in the next two sections.

As an example, the Weak Facet-Adding Rule implies that the two mDAGs of Fig. 3.3.3 are observationally equivalent. An intuition for the validity of this rule is discussed in Section 3.5.1.

The Weak Facet-Adding Rule is a special case of Proposition 6.1 of Ref. [33], that we are going to call "Evans' Rule". We paraphrase it below.

Proposition 6 (Evans' Rule [33, Proposition 6.1])**.** *Let $\mathfrak{G} = \{\mathcal{D}, \mathcal{B}\}$ be an mDAG whose simplicial complex $\mathcal{B}$ contains two disjoint faces C and D. Further suppose that:*

1. $\mathtt{pa}_{\mathcal{D}}(C) \cup C \subseteq \mathtt{pa}_{\mathcal{D}}(d)$ *for each* $d \in D$.
2. *Every face that contains any* $c \in C$ *is a subset of* C.

If $B = C \cup D$ is not already present in the simplicial complex of $\mathfrak{G}$, let $\mathfrak{G}'$ be the mDAG obtained by starting from $\mathfrak{G}$, adding B to its simplicial complex and deleting all of the directed edges $c \to d$ for every $c \in C$, $d \in D$. In this case, $\mathfrak{G}$ is observationally equivalent to $\mathfrak{G}'$, i.e., $\mathfrak{G} \cong \mathfrak{G}'$.

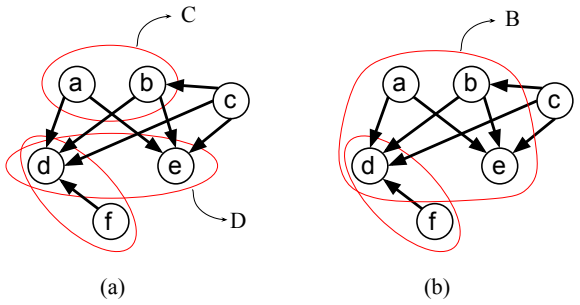


Figure 3.3.3: Two mDAGs that can be proven observationally equivalent by the Weak Facet-Adding rule. Here, the sets B, C and D from the statement of Proposition 5 are respectively $B = \{a, b, d, e\}$, $C = \{a, b\}$ and $D = \{d, e\}$.

It is easy to see that the Weak Facet-Adding Rule (Proposition 5) is a special case of Evans' Rule (Proposition 6): in Weak Facet-Adding we only add a facet, we do not delete the directed edges $c \to d$. However, as we will see below, the HLP Edge-Adding Rule can be applied together with Weak Facet-Adding to obtain Evans' Rule. An example of this is given in Fig. 3.3.4.

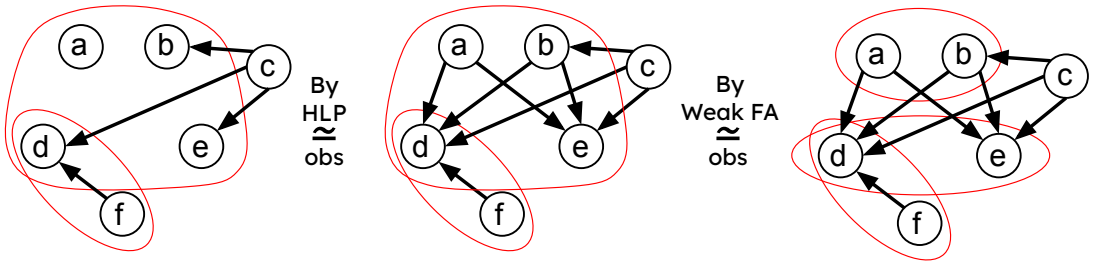


Figure 3.3.4: The first and third mDAG of this figure can be shown observationally equivalent by Evans' rule, as well as by the HLP Edge-Adding rule together with the Weak Facet-Adding rule. This exemplifies Lemma 9.

Lemma 9 (Weak FA ∧ HLP⇒Evans)**.** *The application of Weak Facet-Adding (Proposition 5) together with the HLP Edge-Adding Rule (Proposition 4) subsumes Evans's Rule (Proposition 6).*

Proof. Let $\mathfrak{G} = \{\mathcal{D}, \mathcal{B}\}$ be an mDAG that satisfies the conditions of Propositions 5 and 6. By Weak Facet-Adding (Proposition 5), we know that $\mathfrak{G}$ is observationally equivalent to an mDAG $\mathfrak{G}'$ where the face $B = C \cup D$ is added. This mDAG $\mathfrak{G}'$ is such that, for each $c \in C$ and $d \in D$:

1. $\mathtt{pa}_{\mathcal{D}}(c) \subseteq \mathtt{pa}_{\mathcal{D}}(d)$.
2. The only facet that includes c is B. This facet also includes d.

That is, the conditions of the HLP Edge-Adding Rule are satisfied. This rule implies that the mDAG $\mathfrak{G}''$ obtained by starting with $\mathfrak{G}'$ and removing all of the arrows $c \to d$ for $c \in C$ and $d \in D$ is observationally equivalent to $\mathfrak{G}'$, i.e., $\mathfrak{G}'' \cong \mathfrak{G}'$.

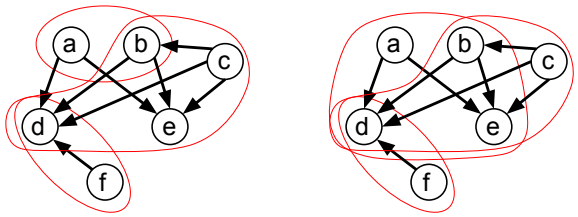
This mDAG $\mathfrak{G}''$ is the one that is obtained directly from $\mathfrak{G}$ by application of Evans's Rule. Therefore, the joint application of the Weak Facet-Adding Rule and the HLP Edge-Adding Rule can subsume Evans's Rule. □

Hence, we can separate the part of Evans's Rule that adds a facet, that is included in the Weak Facet-Adding Rule, from the part that removes the directed edges $c \to d$, that is taken care of by the HLP Edge-Adding Rule. Here, we will use HLP and Weak FA in proving dominances. Lemma 9 implies that doing so, we obtain any conclusions that could have been obtained by applying Evans' Rule.

Moderate Facet-Adding

Our first supplement to existing dominance-proving rules is an extension of Weak Facet-Adding, wherein condition 2 is relaxed. We will call it *Moderate Facet-Adding* rule. Fig. 3.3.5 presents an observational equivalence that can be shown by Moderate Facet-Adding but not by Weak Facet-Adding.

In the next subsection, the Moderate Facet-Adding rule will be further generalized, when we present the *Strong Facet-Adding* (Strong FA) rule. Even though the Moderate Facet-Adding rule is an intermediate result between Weak FA and Strong FA, we opted to present it because, as can be seen from its proof, it is a very simple extension of the Weak FA rule. As we will see, the statement and proof of the Strong FA rule are less intuitive.



(a) (b)

Figure 3.3.5: Two mDAGs that can be proven observationally equivalent by Moderate Facet-Adding, but not by Weak Facet-Adding. Here, the sets B, C and D from the statement of Proposition 5 are respectively $B = \{a, b, d, e\}$, $C = \{a, b\}$ and $D = \{d, e\}$.

Proposition 7 (Moderate Facet-Adding (Moderate FA))**.** *Let $\mathfrak{G} = \{\mathcal{D}, \mathcal{B}\}$ be an mDAG whose simplicial complex $\mathcal{B}$ contains two disjoint faces C and D. If $B = C \cup D$ is not already present in the simplicial complex of $\mathfrak{G}$, let $\mathfrak{G}'$ be the mDAG obtained by starting from $\mathfrak{G}$ and adding B and all of the faces contained in B to its simplicial complex.*

Suppose that:

1. $\mathtt{pa}_{\mathcal{D}}(C) \cup C \subseteq \mathtt{pa}_{\mathcal{D}}(d)$ *for each* $d \in D$,
2. *If for a facet* $B' \in \mathcal{B}$ *such that* $B' \notin C$ *we have* $c \in B'$ *for some* $c \in C$, *then* $D \subset B'$.

In this case, $\mathfrak{G}$ observationally dominates $\mathfrak{G}'$, i.e., $\mathfrak{G} \succeq \mathfrak{G}'$.

Note that, since $\mathfrak{G}'$ structurally dominates $\mathfrak{G}$, by Lemma 3 we know that $\mathfrak{G}'$ observationally dominates $\mathfrak{G}$, i.e., $\mathfrak{G}' \succeq \mathfrak{G}$. Therefore, $\mathfrak{G}$ and $\mathfrak{G}'$ are observationally equivalent, i.e., $\mathfrak{G} \cong \mathfrak{G}'$.

Proof. Construct the mDAG $\tilde{\mathfrak{G}}$ by starting from $\mathfrak{G}$ and i) for each *facet* $\tilde{B} \in \mathcal{B}$ such that $\tilde{B} \notin C$ and $c \in \tilde{B}$ for some $c \in C$, adding a parentless visible node $\tilde{b}$ that is a parent of every node in $\tilde{B}$, and ii) removing every *face* $B' \in \mathcal{B}$ such that $B' \notin C$ and $c \in B'$ for some $c \in C$. For example, starting from the mDAG of Fig. 3.3.5(a), where $C = \{a, b\}$, the two steps above result in the mDAG of Fig. 3.3.6, where g is the visible node added in step i).

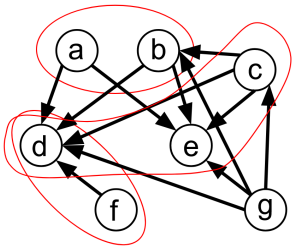


Figure 3.3.6: mDAG obtained by starting from Fig. 3.3.5(a) and performing the steps i) and ii) described in the proof of Proposition 7.

The mDAG $\tilde{\mathfrak{G}}$ satisfies the assumptions of Weak Facet-Adding. Therefore, we know that $\tilde{\mathfrak{G}}$ is observationally equivalent to $\tilde{\mathfrak{G}'}$, the mDAG obtained by adding the face $B = C \cup D$ to it.

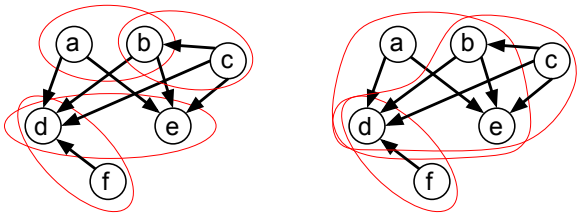
If two pDAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are observationally equivalent, then the two pDAGs $\mathcal{G}_3$ and $\mathcal{G}_4$ obtained by starting from $\mathcal{G}_1$ and $\mathcal{G}_2$ respectively and transforming the same set of visible nodes into latent nodes also should be observationally equivalent. With this idea, we can show the observational equivalence between $\mathfrak{G}$ and $\mathfrak{G}'$: by starting from $\texttt{can}(\tilde{\mathfrak{G}})$ and transforming the visible nodes added in step i) into latent nodes, we obtain a pDAG whose corresponding mDAG is $\mathfrak{G}$. By starting from $\texttt{can}(\tilde{\mathfrak{G}'})$ and transforming the same visible nodes into latent, we obtain a pDAG whose corresponding mDAG is $\mathfrak{G}'$.

□

Note the difference in condition 2 for this proposition and for Propositions 5 and 6: here, nodes of the set C are allowed to belong to faces that are not included in C, as long as these faces also include the entire set D.

Strong Facet-Adding

In Fig. 3.3.7, we have two mDAGs that cannot be shown equivalent by the combination of the HLP Edge-Adding Rule and the Moderate Facet-Adding Rule.



(a) (b)

Figure 3.3.7: Two mDAGs that can be shown observationally equivalent by the Strong Facet-Adding rule.

As it turns out, however, these mDAGs *are* observationally equivalent. This can be shown by the following proposition, which is our second supplement to existing dominance-proving rules:

Proposition 8 (Strong Facet-Adding (Strong FA)). *Let* $\mathfrak{G} = \{\mathcal{D}, \mathcal{B}\}$ *be an mDAG whose simplicial complex contains a sequence of faces* $C_i \in \mathcal{B}$ *for* $i = 1, ..., n$*, that do not need to be disjoint. Suppose further that the mDAG* $\mathfrak{G}$ *also contains a face* $D \in \mathcal{B}$ *that is disjoint to every* C_i*,* $i = 1, ..., n$*. Let* $\mathfrak{G}'$ *be the mDAG obtained by starting from* $\mathfrak{G}$ *and, if not already present, adding the sets* $B_i = C_i \cup D$*,* $i = 1, ..., n$*, together with all of the faces they contain, to its simplicial complex.*

Suppose that:

1. $\mathtt{pa}_{\mathcal{D}}(C_i) \cup C_i \subseteq \mathtt{pa}_{\mathcal{D}}(d)$ *for each* $d \in D$*,* $i = 1, ..., n$*.*
2. *If for a facet* $B' \in \mathcal{B}$ *such that* $B' \nsubseteq C_i$ *for all* $i = 1, ..., n$ *we have* $c \in B'$ *for some* $c \in \cup_{i=1,...,n} C_i$*, then* $D \subseteq B'$*.*

In this case, $\mathfrak{G}$ *observationally dominates* $\mathfrak{G}'$*, i.e.,* $\mathfrak{G} \succeq \mathfrak{G}'$*.*

Note that, since $\mathfrak{G}'$ *structurally dominates* $\mathfrak{G}$*, by Lemma 3 we know that* $\mathfrak{G}'$ *observationally dominates* $\mathfrak{G}$*, i.e.,* $\mathfrak{G}' \succeq \mathfrak{G}$*. Therefore,* $\mathfrak{G}$ *and* $\mathfrak{G}'$ *are observationally equivalent, i.e.,* $\mathfrak{G} \cong \mathfrak{G}'$*.*

The proof of this proposition is presented in Appendix B.2. The general intuition is given in Section 3.5.1, in the discussion of the observational equivalence between the second and third mDAGs of Fig. 3.5.1.

This result shows conditions under which multiple facets can be merged simultaneously. In the example of Fig. 3.3.7, we have $D = \{1, 2\}$, $C_1 = \{3, 4\}$, $C_2 = \{4, 5\}$, $B_1 = \{1, 2, 3, 4\}$ and $B_2 = \{1, 2, 4, 5\}$.

The Structural Dominance Rule (Lemma 3), the HLP Edge-Adding Rule and the Strong Facet-Adding Rule are together the strongest set of dominance-proving rules currently known.

As we saw, when combined with Lemma 3, all of the dominance-proving rules that we presented in this section establish observational *equivalence*. That is, every time that one of these rules can be used to show that an mDAG $\mathfrak{G}$ observationally dominates another mDAG $\mathfrak{G}'$, the conditions for applicability of the rule imply that $\mathfrak{G}'$ structurally dominates $\mathfrak{G}$.

Therefore, the only known dominance-proving rule that can potentially establish a strict observational dominance (i.e., observational dominance without observational equivalence) is Lemma 3. This leads us to the following conjecture, that says that Lemma 3 is indeed the *only* way through which strict observational dominance can be established:

Conjecture 2. *If an mDAG $\mathfrak{G}$ strictly observationally dominates an mDAG $\mathfrak{G}'$, i.e., $\mathfrak{G} \succ \mathfrak{G}'$, then there exists a sequence of mDAGs $\mathfrak{G} = \mathfrak{G}_1, \mathfrak{G}_2, ..., \mathfrak{G}_n = \mathfrak{G}'$ such that each pair $\mathfrak{G}_i$, $\mathfrak{G}_{i+1}$ is such that either:*

(i) $\mathfrak{G}_i$ is observationally equivalent to $\mathfrak{G}_{i+1}$, or

(ii) $\mathfrak{G}_i$ structurally *dominates* $\mathfrak{G}_{i+1}$.

While this is true for the case of 3-node mDAGs, it is still not known whether this is true in general.

To illustrate the need for a sequence in this conjecture, Fig. 3.3.8 shows an example of latent-confounder-free mDAGs $\mathfrak{G}$ and $\mathfrak{G}'$ (that are *not* consistent with the same nodal ordering). Using the dominance-proving rules described in this section together with the nondominance-proving rules described in the next section, we can check that *none* of the mDAGs that are observationally equivalent to $\mathfrak{G}$ structurally dominate any of the mDAGs that are observationally equivalent to $\mathfrak{G}'$. Using the sequence $\mathfrak{G} = \mathfrak{G}_1, \mathfrak{G}_2, \mathfrak{G}_3, \mathfrak{G}_4 = \mathfrak{G}'$, however, we can see that observational equivalence together with structural dominance is sufficient to show the observational dominance of $\mathfrak{G}$ over $\mathfrak{G}'$. In fact, the version of Conjecture 2 where both $\mathfrak{G}$ and $\mathfrak{G}'$ are latent-confounder-free is known to be true, as a consequence of Ref. [21].

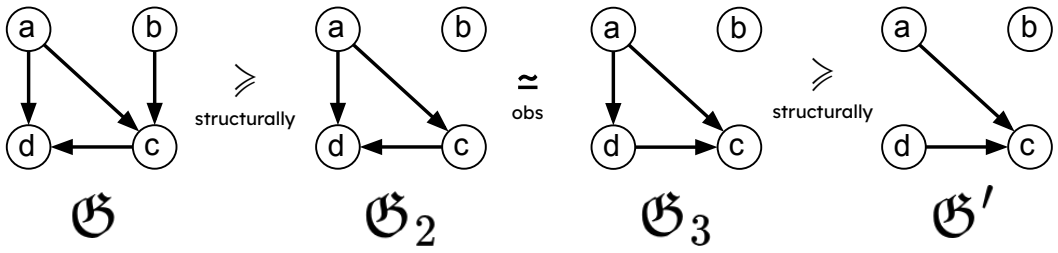


Figure 3.3.8: Example of the sequence of mDAGs that appears in Conjecture 2. One can see that $\mathfrak{G}_2$ and $\mathfrak{G}_3$ are observationally equivalent by applying the HLP Edge-Adding rule twice: first to remove the arrow $c \to d$ and second to add the arrow $d \to c$. Conjecture 2 says that every strict observational dominance can be obtained by a sequence like this one.

3.4 Rules for Establishing Observational Nondominance

In this section, we are going to present all the known rules for establishing that one mDAG does *not* observationally dominate another. In particular, observational nondominance implies observational inequivalence, so the successive application of a set of nondominance-proving rules will give us a proven-inequivalence partition induced by that set of rules (gray loops of Fig. 3.2.2).

In Fig. 3.4.1, we present a diagram with all the nondominance-proving rules that will be presented in this section. One rule may be subsumed by another if all of the nondominances that can be shown by the first can also be shown by the second. We represent this in Fig. 3.4.1 as arrows from the stronger to the weaker rule. However, it is still of value to discuss the weaker rules, as they are in general computationally cheaper to apply.

We believe that the comparison of the sets of nested Markov constraints [68] presented by each mDAG constitutes a rule for establishing observational nondominance. However, the literature still does not present an explicit construction of a distribution that is realizable by one mDAG but not by the other when their sets of nested Markov constraints differ.[7] In any case, we believe that it is unlikely that this comparison would show observational nondominances that our other rules do not already show. Therefore, we have opted to not include this rule here.

We will organize our discussion of the nondominance-proving rules starting from the

[7] According to private communication with Robin Evans, however, such a distribution exists and is to appear in future work.

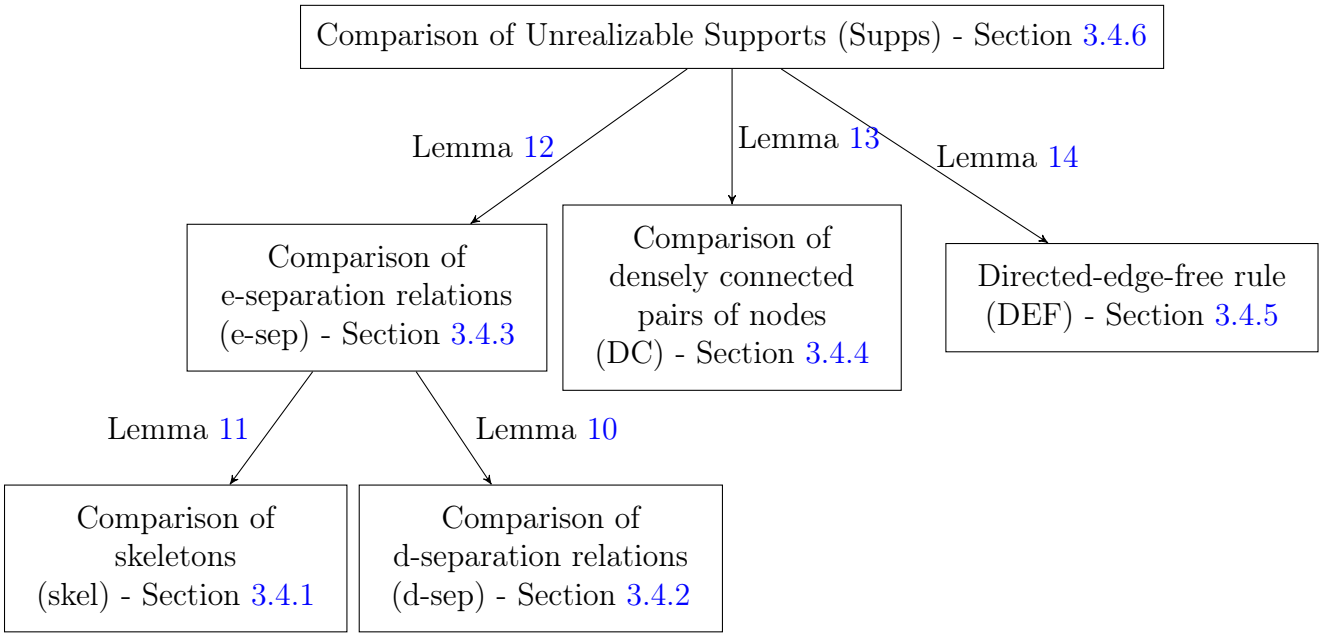


Figure 3.4.1: Classification of the different nondominance-proving rules. An arrow from one rule to another indicates that the latter is strictly less powerful than the former (i.e., the latter is subsumed in the former), and the Lemma that proves it is indicated on the arrow. We defined a shorthand name for each rule, shown inside the parentheses.

computationally cheaper ones (i.e., we proceed bottom-up in the diagram of Fig. 3.4.1).

3.4.1 Skeletons

The first nondominance rule is the comparison of skeletons of mDAGs, that was introduced in Refs. [32, 33]. For that, we first define the *skeleton* of an mDAG, which is exemplified in Fig. 3.4.2.

Definition 20 (Skeleton)**.** *Let $\mathfrak{G} = (\mathcal{D}, \mathcal{B})$ be an mDAG. We define the* skeleton *of $\mathfrak{G}$ by the undirected graph with the same nodes as $\mathcal{D}$ and with an edge between nodes u and w whenever there is a directed edge between them in $\mathcal{D}$ or when $u, w \in B$ for some $B \in \mathcal{B}$.*

Below, we present a corollary of Theorem 4.2 of Ref. [32], which is a generalization of Proposition 6.5 of Ref. [33]:

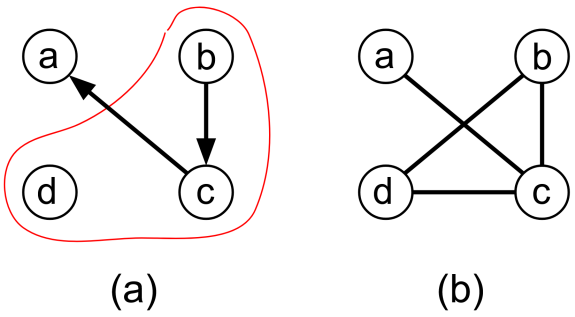


Figure 3.4.2: (a) An mDAG; and (b) Its skeleton.

Proposition 9 (Comparison of skeletons - Consequence of Refs. [32, 33])**.** *Let* $\mathfrak{G}$ *and* $\mathfrak{G}'$ *be two mDAGs such that* $\texttt{nodes}(\mathfrak{G}) = \texttt{nodes}(\mathfrak{G}')$*. If there exist nodes* $x, y \in \texttt{nodes}(\mathfrak{G})$ *such that the undirected edge between* x *and* y *is present in the skeleton of* $\mathfrak{G}'$ *but not in the skeleton of* $\mathfrak{G}$*, then* $\mathfrak{G}$ does not *observationally dominate* $\mathfrak{G}'$*, i.e.,* $\mathfrak{G} \not\succeq \mathfrak{G}'$*.*

As a special case of this rule, we know that two mDAGs are observationally inequivalent whenever their skeletons are different.

3.4.2 d-separation relations

As discussed in Section 2.1.1, all the conditional independence constraints that an mDAG imposes on its compatible probability distributions can be obtained from the d-separation relations of that mDAG. We can compare the sets of d-separation relations among the visible variables as a way to show observational nondominance between two causal structures:

Proposition 10 (Comparison of d-separation relations - Consequence of Theorem 4)**.** *Let* $\mathfrak{G}$ *and* $\mathfrak{G}'$ *be two mDAGs such that* $\texttt{nodes}(\mathfrak{G}) = \texttt{nodes}(\mathfrak{G}')$*. If there is a d-separation relation that is presented by* $\mathfrak{G}$ *but not by* $\mathfrak{G}'$*, then* $\mathfrak{G}$ does not *observationally dominate* $\mathfrak{G}'$*, i.e.,* $\mathfrak{G} \not\succeq \mathfrak{G}'$*.*

Proof. We will make a proof by contradiction. Assume that $\mathfrak{G}$ observationally dominates $\mathfrak{G}'$, but there is a d-separation relation $x \perp\!\!\!\perp_{\mathrm{d}} y | z$ for $x, y, z \in \texttt{nodes}(\mathfrak{G})$ that is presented by $\mathfrak{G}$ but not by $\mathfrak{G}'$.

By the first point of Theorem 4, if a probability distribution P is realizable by $\mathfrak{G}$, then P is such that $X_x \perp\!\!\!\perp X_y | X_z$ in P. Because of the assumed observational dominance, *every* distribution realizable by $\mathfrak{G}'$ is also realizable by $\mathfrak{G}$, and thus has to satisfy $X_x \perp\!\!\!\perp X_y | X_z$.

By the second point of Theorem 4, this would imply that $x \perp\!\!\!\perp_{\mathrm{d}} y | z$ in $\mathfrak{G}'$, which is a contradiction. □

Therefore, we can use the comparison of d-separation relations to classify the mDAGs into another proven-inequivalence partition.

As it turns out, the comparison of skeletons rule and the comparison of d-separation relations rule are *not* redundant to each other as rules to show observational nondominance. This can already be seen by looking at 3-node mDAGs; Figs. 3.4.3 and 3.4.4 show explicit examples.

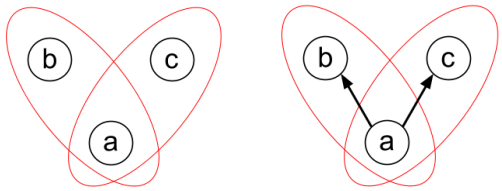


Figure 3.4.3: Example of mDAGs that have the same skeleton, but different sets of d-separation relations. The mDAG (a) presents the d-separation relation $b \perp\!\!\!\perp_{\mathrm{d}} c$, while the mDAG (b) does not. In Fig. 3.2.1, the mDAG (a) is in the `Collider A` observational equivalence class, while (b) is in the `Evans` observational equivalence class.

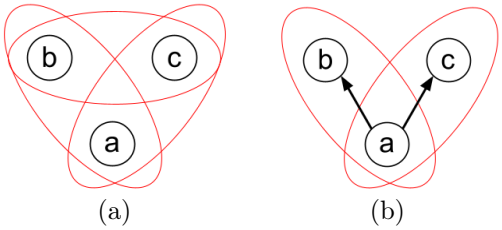


Figure 3.4.4: Example of mDAGs that have the same set of d-separation relations (i.e., no d-separation relations), but different skeletons. In Fig. 3.2.1, the mDAG (a) is in the `Triangle` observational equivalence class, while (b) is in the `Evans` observational equivalence class.

3.4.3 e-separation relations

In Ref. [32], an extension of d-separation was introduced, called *e-separation*.While the d-separation method yields all of the conditional independence constraints that a causal

structure implies, the e-separation method gives *some* of the inequality constraints that it implies. These are made explicit in Ref. [37] (note that *not all* inequality constraints can be obtained from e-separation).

Definition 21 (e-separation)**.** *Let* $\mathfrak{G}$ *be an mDAG, and let* $A, B, C, D \subseteq \texttt{nodes}(\mathfrak{G})$ *be four sets of nodes of* $\mathfrak{G}$*. We say that* A *and* B *are e-separated by* C *after deletion of* D*, denoted* $A \perp\!\!\!\perp_e B | C \neg D$*, if* $A \perp\!\!\!\perp_d B | C$ *in* $\mathfrak{G}_{\texttt{nodes}(\mathfrak{G}) \setminus D}$*, the subgraph of* $\mathfrak{G}$ *obtained by deleting the nodes in the set* D*.*[8]

Clearly, d-separation is a special case of e-separation, where the set of nodes being deleted is the empty set, i.e., $D = \varnothing$. As a consequence of Ref. [37], the comparison of e-separation relations is a nondominance-proving rule:

Proposition 11 (Comparison of e-separation relations - Consequence of Ref. [37])**.** *Let* $\mathfrak{G}$ *and* $\mathfrak{G}'$ *be two mDAGs such that* $\texttt{nodes}(\mathfrak{G}) = \texttt{nodes}(\mathfrak{G}')$*. If there is an e-separation relation that is presented by* $\mathfrak{G}$ *but not by* $\mathfrak{G}'$*, then* $\mathfrak{G}$ does not *observationally dominate* $\mathfrak{G}'$*, i.e.,* $\mathfrak{G} \not\geq \mathfrak{G}'$*.*

Proof. As mentioned, Ref. [37] shows that the e-separation relation yields an inequality constraint implied by $\mathfrak{G}$. In Proposition 12 in the Appendix E of Ref. [37], it is shown that if $\mathfrak{G}'$ *does not* present the e-separation relation in question, then *there exists* some probability distribution realizable by $\mathfrak{G}'$ that violates the associated inequality (in the case where the nodes in D are each associated to a discrete state space). Therefore, $\mathfrak{G}$ does not observationally dominate $\mathfrak{G}'$. □

The comparison of e-separation relations shows itself to be a better nondominance-proving rule than either comparison of skeletons or comparison of d-separation relations. This is made explicit by Lemmas 10 and 11.

Lemma 10 (The comparison of e-separation relations subsumes The comparison of d-separation relations)**.** *Let* $\mathfrak{G}$ *and* $\mathfrak{G}'$ *be two mDAGs such that* $\mathfrak{G}$ *can be shown to* not *observationally dominate* $\mathfrak{G}'$ *by comparison of d-separation relations. Then,* $\mathfrak{G}$ *can also be shown to not observationally dominate* $\mathfrak{G}'$ *by comparison of e-separation relations.*

[8] A reader familiar with the concept of *dormant independence* [69] might wonder what is the difference between these two concepts. While a dormant independence is a type of Nested Markov constraint, which are *equality* constraints, the presence of an e-separation relation leads to the existence of an *inequality* constraint. Graphically, the difference is that dormant independencies need to be such that $P(X_A | X_B X_C \text{do}(X_D))$ is identifiable, while for e-separation we can check any combination of nodes being independent given the deletion of other nodes.

Proof. This follows trivially from the fact that d-separation is a special case of e-separation (when $D = \varnothing$ in Definition 21). □

Lemma 11 (The comparison of e-separation relations subsumes the comparison of skeletons)**.** *Let $\mathfrak{G}$ and $\mathfrak{G}'$ be two mDAGs such that $\mathfrak{G}$ can be shown to* not *observationally dominate $\mathfrak{G}'$ by comparison of skeletons. Then, $\mathfrak{G}$ can also be shown to not observationally dominate $\mathfrak{G}'$ by comparison of e-separation relations.*

Proof. Suppose that the nodes u and v are connected by an edge in the skeleton of $\mathfrak{G}'$ but not in the skeleton of $\mathfrak{G}$. This means that, in $\mathfrak{G}$, there is neither a directed edge nor a confounder between u and v. It is then clear that, if we delete all other nodes of $\mathfrak{G}$, u and v are going to be d-separated by the empty set. In $\mathfrak{G}'$, on the other hand, u and v are either connected by a directed edge or are part of the same facet. Thus, if we delete all the other nodes, u and v are not going to be d-separated by the empty set. Therefore, $\mathfrak{G}$ has the e-separation relation $u \perp\!\!\!\perp_e v | \neg(\texttt{nodes}(\mathfrak{G}) \setminus \{u, v\})$, while $\mathfrak{G}'$ *does not* have it. Thus, the comparison of e-separation relations can show that $\mathfrak{G}$ that does not observationally dominate $\mathfrak{G}'$. □

3.4.4 Pairs of Densely Connected Nodes

In this section, we show a rule for observational nondominance derived from the main result of Ref. [34]. This rule is based on the concept of *densely connected nodes*, first defined in Ref. [67]. This concept makes use of the preliminary definitions of *districts* of an mDAG and the *closure* of a set of nodes. The concept of districts, first introduced in Ref. [77], is reproduced below.

Definition 22 (Districts)**.** *Let $\mathfrak{G} = \{\mathcal{D}, \mathcal{B}\}$ be an mDAG. Let D be a set of nodes of $\mathfrak{G}$ that are connected to each other through a set of facets of $\mathcal{B}$; i.e., if $v, w \in D$, then there is a sequence of nodes $(u_0, u_1, u_2, ..., u_{n-1}, u_n)$ with $u_0 = v$ and $u_n = w$ such that for all i, $\{u_i, u_{i+1}\} \subseteq B_i \in \mathcal{B}$. If D is inclusion-maximal (i.e., maximal in the order induced by subset inclusion), we say that it is a* district*.*

If v is a node of $\mathfrak{G}$, we are going to denote the district of $\mathfrak{G}$ to which v belongs by $dis_{\mathfrak{G}}(v)$. Similarly, for a set of nodes S, we have $dis_{\mathfrak{G}}(S) = \cup_{v \in S} dis_{\mathfrak{G}}(v)$.

In the mDAG of Fig. 3.4.5, for example, we have three districts: $\{a, b, c, h\}$, $\{d\}$ and $\{e, f, g\}$.

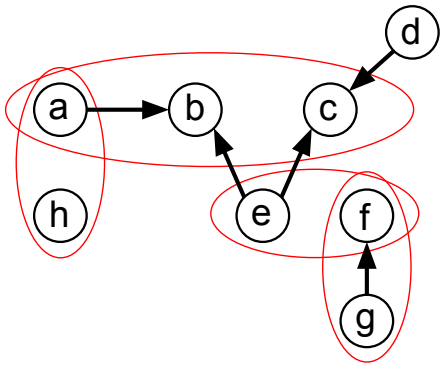


Figure 3.4.5: The districts of this mDAG are $\{a, b, c, h\}$, $\{d\}$ and $\{e, f, g\}$.

The second preliminary definition we need is the *closure* of a set of nodes, first introduced in Ref. [67]. This is just a step towards defining densely connected pairs of nodes, which is the important definition for us.

Definition 23 (Closure of a set of nodes)**.** *Let* $\mathfrak{G} = \{\mathcal{D}, \mathcal{B}\}$ *be an mDAG, and let* $A \subseteq \mathtt{nodes}(\mathfrak{G})$*. Set* $A^{(0)} \equiv \mathtt{nodes}(\mathfrak{G})$ *and construct a sequence of sets of nodes by applying alternately the following rules:*

$$A^{(i+1)} = dis_{\mathfrak{G}_{A^{(i)}}}(A) \quad , \qquad A^{(i+2)} = \mathtt{an}_{\mathcal{D}_{A^{(i+1)}}}(A)$$

Where $dis_{\mathfrak{G}}(S)$ *is the district of* S *in* $\mathfrak{G}$*, and* $\mathtt{an}_{\mathcal{D}}(S)$ *is the set of ancestors of the nodes of* S *in the directed structure (which includes the nodes of* S *themselves). As usual,* $\mathfrak{G}_S$ *is the subgraph of* $\mathfrak{G}$ *over nodes* S*, and* $\mathcal{D}_S$ *is the subgraph of* $\mathcal{D}$ *over nodes* S*. Each one of these operations removes nodes from the set* $A^{(i)}$ *or keeps it the same, the case which indicates the termination of the process. When the process terminates, we call the resulting set the closure of* A *and denote it* $\langle A \rangle_{\mathfrak{G}}$*.*

Let us go through an example of this definition. Say that we want to know what is the closure of the set $A = \{b, c, g\}$ in the mDAG $\mathfrak{G}$ of Fig. 3.4.5. The set $A^{(0)}$ is the entire set of nodes of the mDAG, so we have that $\mathfrak{G}_{A^{(0)}} = \mathfrak{G}$. This mDAG is reproduced in Fig. 3.4.6(a). It implies:

$$A^{(1)} = \mathrm{dis}_{\mathfrak{G}}(A) = \{a, b, c, e, f, g\}.$$

Therefore, $\mathfrak{G}_{A^{(1)}}$ is the sub-mDAG of $\mathfrak{G}$ shown in Fig. 3.4.6(b). This in turn gives us

$$A^{(2)} = \mathtt{an}_{\mathcal{D}_{A^{(1)}}}(A) = \{a, b, c, e, g\},$$

where $\mathcal{D}_{A^{(1)}}$ is the directed structure of $\mathfrak{G}_{A^{(1)}}$. This implies that $\mathfrak{G}_{A^{(2)}}$ is the mDAG of Fig. 3.4.6(c), which gives:

$$A^{(3)} = \mathrm{dis}_{\mathfrak{G}_{A^{(2)}}}(A) = \{a, b, c, g\}.$$

Then, $\mathfrak{G}_{A^{(3)}}$ is the mDAG of Fig. 3.4.6(d). Therefore:

$$A^{(4)} = \mathtt{an}_{\mathcal{D}_{A^{(3)}}}(A) = \{a, b, c, g\}.$$

The sets $A^{(3)}$ and $A^{(4)}$ are equal, which indicates the termination of the process. Therefore, in this example $\langle A \rangle_{\mathfrak{G}} = \{a, b, c, g\}$.

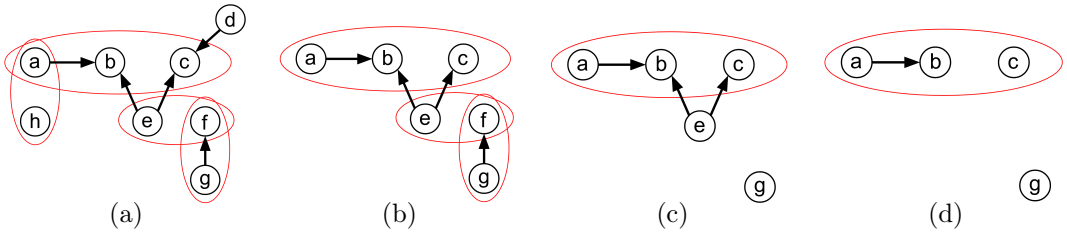


Figure 3.4.6: Steps used to find the closure of the set $A = \{b, c, g\}$ in the mDAG of Fig. 3.4.5.

Now we can finally present the definition of dense connectedness between nodes, taken from Ref. [67]:

Definition 24 (Densely connected pair of nodes)**.** *Let $\mathfrak{G} = \{\mathcal{D}, \mathcal{B}\}$ be an mDAG. A pair of nodes $v \neq w \in \mathtt{nodes}(\mathfrak{G})$ is said to be* densely connected *if any of the following conditions are satisfied:*

- $v \in \mathtt{pa}_{\mathcal{D}}(\langle w \rangle_{\mathfrak{G}})$.
- $w \in \mathtt{pa}_{\mathcal{D}}(\langle v \rangle_{\mathfrak{G}})$.
- $\langle \{v, w\} \rangle_{\mathfrak{G}}$ *is a bidirected-connected set.*

where we say that a set of nodes S is bidirected-connected if it forms a district of the mDAG $\mathfrak{G}_S$, the mDAG obtained by deleting all of the nodes of $\mathfrak{G}$ that are not in S. In other words, the set S is bidirected-connected if every vertex of S can be reached from every other using a path of nodes, all in S, that are connected by facets of $\mathcal{B}$.

For example, one can check that the nodes b and h in the mDAG $\mathfrak{G}$ of Fig. 3.4.5 are densely connected in $\mathfrak{G}$, because $\langle\{b,h\}\rangle_{\mathfrak{G}} = \{a,b,h\}$, which is a bidirected-connected set in $\mathfrak{G}$. On the other hand, b and g are not densely connected, because b is not a parent of $\langle\{g\}\rangle_{\mathfrak{G}} = \{g\}$, g is not a parent of $\langle\{b\}\rangle_{\mathfrak{G}} = \{a,b\}$ and $\langle\{b,g\}\rangle_{\mathfrak{G}} = \{a,b,g\}$ is not bidirected-connected.

Due to the main result of Ref. [34], the comparison of pairs of densely connected nodes is a nondominance-proving rule:

Proposition 12 (Comparison of densely connected pairs of nodes - Consequence of Ref. [34, Theorem 6.4])**.** *Let* $\mathfrak{G}$ *and* $\mathfrak{G}'$ *be two mDAGs such that* $\texttt{nodes}(\mathfrak{G}) = \texttt{nodes}(\mathfrak{G}')$*. If there is an pair of nodes* $v, w \in \texttt{nodes}(\mathfrak{G})$ *that is densely connected in* $\mathfrak{G}'$ *but not in* $\mathfrak{G}$*, then* $\mathfrak{G}$ does not *observationally dominate* $\mathfrak{G}'$*, i.e.,* $\mathfrak{G} \not\succeq \mathfrak{G}'$*.*

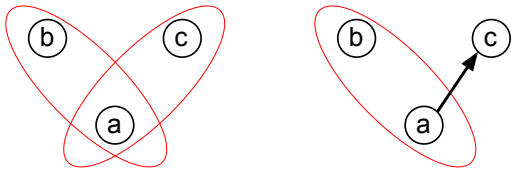
Proof. Ref. [34] showed that the specific probability distribution where X_v and X_w are perfectly correlated and independent of all other visible variables (which are uniformly distributed) is realizable by a causal structure *if and only if* the nodes v and w are densely connected in that causal structure. Therefore, this distribution is realizable by $\mathfrak{G}'$ but not by $\mathfrak{G}$, which implies that $\mathfrak{G}$ does not observationally dominate $\mathfrak{G}'$. □

The comparison of densely connected pairs of nodes is not redundant to the comparison of e-separation relations; both techniques can show observational inequivalences that the other one cannot. Examples of mDAGs that can be shown to be observationally inequivalent by comparison of e-separation relations but not by pairs of densely connected nodes are found already in the three-node case; an explicit example is shown in Fig. 3.4.7. On the other hand, there are no three-node mDAGs that can be shown observationally inequivalent by comparison of densely connected pairs of nodes but not by comparison of e-separation relations. Among four-node mDAGs, however, such examples appear. One such example is shown in Fig. 3.4.8.

3.4.5 Directed-edge-free rule

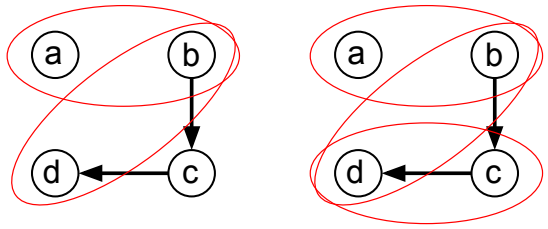
In this section, we discuss a rule for showing observational nondominance that is based on Proposition 2:

Proposition 13 (Directed-edge-free rule - Consequence of Ref. [75, Example 2])**.** *Let* $\mathfrak{G}$ *and* $\mathfrak{G}'$ *be two mDAGs. Suppose* $\mathfrak{G}$ *is observationally equivalent to the directed-edge-free mDAG* $\tilde{\mathfrak{G}}$ *and* $\mathfrak{G}'$ *is observationally equivalent to the directed-edge-free mDAG* $\tilde{\mathfrak{G}}'$*. Then,* $\mathfrak{G}$ *observationally dominates* $\mathfrak{G}'$ *if and only if* $\tilde{\mathfrak{G}}$ *structurally dominates* $\tilde{\mathfrak{G}}'$*.*



(a) (b)

Figure 3.4.7: Example of a pair of mDAGs that have the same pairs of densely connected nodes, but different sets of e-separation relations (which in this case are simply d-separation relations). In Fig. 3.2.1, the mDAG (a) is in the `Collider A` observational equivalence class, while (b) is in the `Fork` observational equivalence class.



(a) (b)

Figure 3.4.8: Example of a pair of mDAGs that have the same sets of e-separation relations, but different pairs of densely connected nodes. The e-separation relations of both of these mDAGs are $a \perp\!\!\!\perp_{\mathrm{d}} c|\neg b$, $a \perp\!\!\!\perp_{\mathrm{d}} c|\neg bd$, $a \perp\!\!\!\perp_{\mathrm{d}} c|b$, $a \perp\!\!\!\perp_{\mathrm{d}} c|b\neg d$, $a \perp\!\!\!\perp_{\mathrm{d}} c|d\neg b$, $a \perp\!\!\!\perp_{\mathrm{d}} d|\neg b$, $a \perp\!\!\!\perp_{\mathrm{d}} d|\neg bc$, $a \perp\!\!\!\perp_{\mathrm{d}} d|\neg c$ and $a \perp\!\!\!\perp_{\mathrm{d}} d|c\neg b$. Meanwhile, the pair of nodes (a, d) is densely connected in (b) but not in (a).

Proof. This is a straightforward corollary of Proposition 2. □

Note that this rule has a difference from the previous rules shown in this section. The other rules assign a signature to each mDAG, such as their skeleton or their set of d-separation relations, and then compare this signature. Because of that, the other rules can be applied to any pair of mDAGs, even though sometimes they give inconclusive results (for example when neither set of d-separation relations is contained in the other). Proposition 13, on the other hand, can only be applied to compare mDAGs that are equivalent to a directed-edge-free mDAG. This makes the implications of this rule for the proven-inequivalence partition a bit more subtle than before.

This is explained in Fig. 3.4.9. There, we *can* split the proven-inequivalence block (gray

loop) on the left into four, because all the blocks of the proven-equivalence partition (pink loops) inside it contain a directed-edge-free mDAG, so Proposition 2 implies that they are all observationally inequivalent. The proven-inequivalence block on the right, on the other hand, contains a distinguished proven-equivalence block that does not contain a directed-edge-free mDAG. Therefore, the proven-inequivalence block on the right still *cannot* be split, because the distinguished proven-equivalence block inside it cannot be compared to the others by Proposition 2.

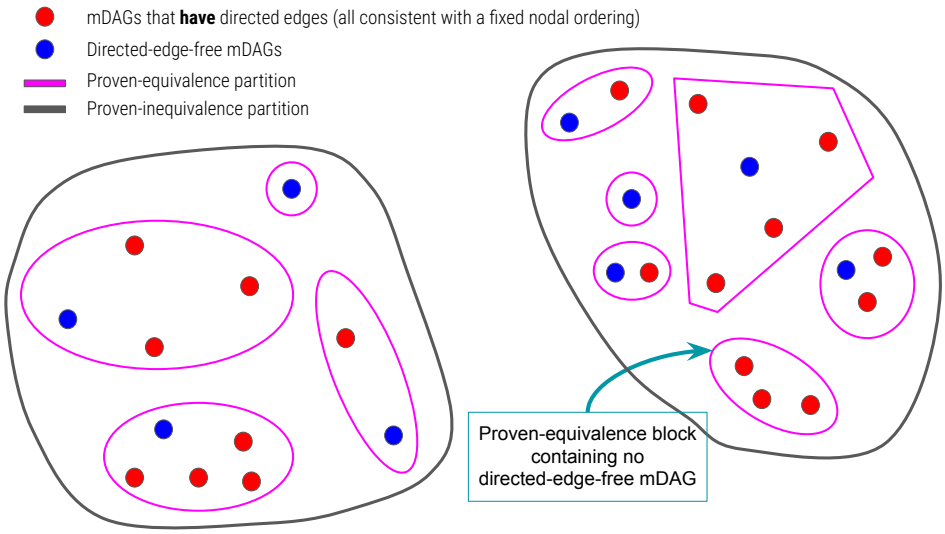


Figure 3.4.9: The directed-edge-free mDAGs are represented by blue dots, while the other mDAGs are red dots. Suppose that nondominance-proving rules other than the directed-edge-free rule have been used to obtain the proven-inequivalence partition depicted. The directed-edge-free rule says that the proven-inequivalence block on the left should split into the four proven-equivalence blocks that live inside it. On the other hand, the proven-inequivalence block on the right includes one distinguished proven-equivalence block that does not contain any directed-edge-free mDAG. Therefore, the directed-edge-free rule cannot be used to split the proven-inequivalence block on the right.

The directed-edge-free rule is *not* redundant to the comparison of e-separation relations nor to the comparison of densely connected pairs of nodes; it can show nondominances that the other two rules cannot, and vice versa. The fact that the mDAG of Fig. 2.1.4(a) does not observationally dominate the mDAG of Fig. 2.1.4(b) can be shown by the directed-edge-free rule, but not by the comparison of e-separation relations or the comparison of densely connected pairs. The fact that the mDAGs of Fig. 3.4.7 do not observationally dominate

each other (i.e., they are incomparable) can be shown by comparison of e-separation relations, but not by the directed-edge-free rule (nor by the comparison of densely connected pairs of nodes, as we have seen). The directed-edge-free rule cannot be applied here because, as we can see from Fig. 3.2.1, the mDAG of Fig. 3.4.7(b) is not observationally equivalent to any directed-edge-free mDAG. The fact that the mDAG of Fig. 3.4.8(a) does not observationally dominate the mDAG of Fig. 3.4.8(b) can be shown by the comparison of densely connected pairs of nodes, but not by the directed-edge-free rule (nor by the comparison of e-separation relations, as we have seen). The directed-edge-free rule cannot be applied here because those two mDAGs are not equivalent to any directed-edge-free mDAG. To prove this we note that, as shown in Example 2 of Ref. [75], perfect correlation between a set of nodes is realizable if and only if these nodes share a common ancestor in the causal structure (that can be visible or latent). This implies that both mDAGs of Fig. 3.4.8 are able to realize perfect correlation between the variables X_a, X_b, X_c and X_d. Therefore, if Fig. 3.4.8(b) were observationally equivalent to a directed-edge-free mDAG, it should be the saturated mDAG, where all four nodes are in the same facet (and thus all four nodes share a common ancestor in the corresponding canonical pDAG). This cannot be the case, however, since Fig. 3.4.8(b) is *not* saturated (this is easy to see, for example, by noting that nodes a and d are not connected in its skeleton).

3.4.6 Comparison of unrealizable supports

In this section, we will present a rule to establish observational nondominance based on the work of Fraser [38]. This rule deals with the problem of *possibilistic causal realizability*, which we will define shortly.

An important *caveat* on this method is that it can only be applied when the visible variables have finite cardinalities. As already stated, this is not a problem for our purposes: to show observational inequivalence, it suffices to show that there is a difference in the set of realizable distributions for one specific assignment of cardinalities.

The problem that motivates the work of this chapter is the problem of probabilistic causal realizability (Definition 4), which asks whether a probability distribution is realizable by a causal structure or not. In this section, we turn our attention to a different, strictly weaker, question: the *possibilistic* causal realizability problem. To define it, let us first define the support of a probability distribution:

Definition 25 (Support)**.** *Let $P : \mathcal{X} \to [0, 1]$ be a probability distribution. The elements of $\mathcal{X}$ are called "events". The support of P is the set of events that are deemed to be possible,*

in the sense of having nonzero probability:

$$\mathrm{Supp}(P) = \{\omega \in \mathcal{X} | P(\omega) > 0\}$$

A set of events S is said to be a realizable support of the pDAG $\mathcal{G}$ if there is *some* distribution P that is realizable by $\mathcal{G}$ such that $\mathrm{Supp}(P) = S$. Let $\mathcal{S}(\mathcal{G}, \vec{c}_{\mathrm{vis}})$ denote the set of all sets of events that are realizable by the pDAG $\mathcal{G}$ for cardinalities of the visible variables given by $\vec{c}_{\mathrm{vis}}$. Ref. [38] presented an algorithm to find the set $\mathcal{S}(\mathcal{G}, \vec{c}_{\mathrm{vis}})$. We refer to it here as *Fraser's algorithm*. This algorithm is explained in Appendix B.3.

Just as we defined the observational profile $\mathcal{M}\text{-}\mathtt{profile}(\mathcal{G})$ of a pDAG $\mathcal{G}$, we can define $\mathcal{G}$'s *possibilistic observational profile* by:

$$\mathcal{S}\text{-}\mathtt{profile}(\mathcal{G}) \coloneqq \left(\mathcal{S}(\mathcal{G}, \vec{c}_{\mathrm{vis}}) : \vec{c}_{\mathrm{vis}} \in \mathbb{N}^{|\mathtt{vis}(\mathcal{G})|}\right), \tag{3.2}$$

where again the parentheses indicate that this object behaves as a tuple: each $\mathcal{S}(\mathcal{G}, \vec{c}_{\mathrm{vis}})$ is indexed by $\vec{c}_{\mathrm{vis}}$.

If the support of a probability distribution P is *not* realizable by $\mathcal{G}$, then clearly P is not realizable by $\mathcal{G}$. Thus, if there is some assignment of cardinalities for which a pDAG $\mathcal{G}$ cannot realize a support that is realizable by another pDAG $\mathcal{G}'$, then $\mathcal{G}$ cannot observationally dominate $\mathcal{G}'$. In other words, we have:

$$\begin{gathered} \mathcal{S}\text{-}\mathtt{profile}(\mathcal{G}) \not\subseteq_{\mathrm{ew}} \mathcal{S}\text{-}\mathtt{profile}(\mathcal{G}') \\ \Rightarrow \\ \mathcal{M}\text{-}\mathtt{profile}(\mathcal{G}) \not\subseteq_{\mathrm{ew}} \mathcal{M}\text{-}\mathtt{profile}(\mathcal{G}') \end{gathered} \tag{3.3}$$

where $\not\subseteq_{\mathrm{ew}}$ denotes the failure of element-wise inclusion, which is defined for the possibilistic observational profile in an analogous way to which it was defined for the observational profile in Definition 6. From this implication, we know that the comparison of unrealizable supports gives us a nondominance-proving rule:

Proposition 14 (Comparison of unrealizable supports)**.** *Let* $\mathfrak{G}$ *and* $\mathfrak{G}'$ *be two mDAGs such that* $\mathtt{nodes}(\mathfrak{G}) = \mathtt{nodes}(\mathfrak{G}')$*. If there is a support that is realizable by* $\mathfrak{G}'$ *but not by* $\mathfrak{G}$*, then* $\mathfrak{G}$ does not *observationally dominate* $\mathfrak{G}'$*, i.e.,* $\mathfrak{G} \not\trianglerighteq \mathfrak{G}'$*.*

Note that it is unclear whether the converse implication to that of (3.3) holds. Thus, we formalize it as a conjecture:

Conjecture 3. *If it is possible to find a* probability distribution *that is realizable by the pDAG* $\mathcal{G}$ *but not by the pDAG* $\mathcal{G}'$*, then there is a* support *that is realizable by* $\mathcal{G}$ *but not by* $\mathcal{G}'$*. In other words, the converse of Eq.* (3.3) *holds.*

We will discuss potential prospects to tackle this open question in Section 3.6.3.

Subsumption of other nondominance-proving rules

In Lemmas 10 and 11, we showed that the comparison of e-separation relations subsumes the comparison of skeletons and d-separation relations as rules to show observational nondominance. In this section, we discuss how the comparison of unrealizable supports subsumes other rules.

Since Fraser's algorithm for unrealizable supports is a brute force search, one wants to apply as many of the weaker but computationally cheaper rules as possible before implementing the comparison of unrealizable supports. However, it is possible to show that all of the cheaper rules — comparison of e-separation relations, comparison of densely connected pairs of nodes, and the directed-edge-free rule — are subsumed by the comparison of unrealizable supports. In fact, to subsume the cheaper rules it is sufficient to look at supports at the level of binary visible variables; that is, whenever a nondominance can be shown by one of the cheaper rules, one can find a support with $\vec{c}_{\text{vis}} = (2, 2, 2, ...)$ that also shows said nondominance.

Lemma 12 (The comparison of unrealizable supports subsumes the comparison of e-separation relations)**.** *Let $\mathfrak{G}$ and $\mathfrak{G}'$ be two mDAGs that have different sets of e-separation relations. Then, it is possible to find a support which is realizable by one of these mDAGs but not by the other. In fact, it is possible to find such a support even when all of the visible variables are binary, i.e., $\vec{c}_{vis} = (2, 2, 2, \dots)$.*

The proof is given in Appendix B.4. The idea behind this proof is that the set of e-separation relations of a given mDAG might already imply that some supports are unrealizable. As an example of this, imagine a causal structure that presents the d-separation relation $a \perp\!\!\!\perp_{\text{d}} b$. Suppose that S is a support over binary variables X_a and X_b that does not contain any events involving $\{X_a = 0, X_b = 1\}$ or $\{X_a = 1, X_b = 0\}$, but it does contain events involving $\{X_a = 0, X_b = 0\}$ and events involving $\{X_a = 1, X_b = 1\}$. This implies that in any probability distribution P that has the support S, i.e., where $\text{Supp}(P) = S$, the variables X_a and X_b are correlated. This is incompatible with $a \perp\!\!\!\perp_{\text{d}} b$, thus the support S is ruled out by $a \perp\!\!\!\perp_{\text{d}} b$. To save time in the computation, we only apply Fraser's algorithm for supports that are not already ruled out by the e-separation relations of the mDAG in question.

Lemma 13 (The comparison of unrealizable supports subsumes the comparison of densely connected pairs of nodes)**.** *Let $\mathfrak{G}$ and $\mathfrak{G}'$ be two mDAGs such that $\mathfrak{G}'$ can be shown to* not *observationally dominate $\mathfrak{G}$ by comparison of densely connected pairs of nodes. Then, $\mathfrak{G}'$ can also be shown to not observationally dominate $\mathfrak{G}$ by comparison of unrealizable*

supports. In fact, it is possible to do so even when we are only comparing unrealizable supports where all of the visible variables are binary, i.e., $\vec{c}_{vis} = (2, 2, 2, \dots)$.

The proof is given in Appendix B.4. The idea behind this proof is as follows: as already discussed, Ref. [34] says that the specific distribution where all of the visible variables are binary and X_v and X_w are perfectly correlated while the other variables are all mutually independent and uniformly distributed is realizable by an mDAG if and only if v and w are densely connected in this mDAG. Appendix B.4 then shows that not only this specific probability distribution, but also its *support* is realizable by an mDAG if and only if v and w are densely connected in this mDAG. More specifically, this support is the one where X_v and X_w are both 0 in half of the events and both 1 in the other half, while the other visible variables take all possible combinations of 0 and 1. The explicit construction of such a support for the case of four visible nodes is given in Eq. (B.8).

Lemma 14 (The comparison of unrealizable supports subsumes the directed-edge-free rule)**.** *Let* $\mathfrak{G}$ *and* $\mathfrak{G}'$ *be two mDAGs such that* $\mathfrak{G}'$ *can be shown to* not *observationally dominate* $\mathfrak{G}$ *by the directed-edge-free rule. Then,* $\mathfrak{G}'$ *can also be shown to not observationally dominate* $\mathfrak{G}$ *by comparison of unrealizable supports. In fact, it is possible to do so even when we are only comparing unrealizable supports where all of the visible variables are binary, i.e.,* $\vec{c}_{vis} = (2, 2, 2, \dots)$*, and the supports are constituted by only two events.*

The proof of this lemma is given in Appendix B.4, where an explicit construction of the support that is realizable by $\mathfrak{G}$ but not $\mathfrak{G}'$ is provided.

To conclude this section, note that unlike the case of dominance-proving rules, for the application of nondominance-proving rules we *do not* need to go outside of the set of mDAGs consistent with a fixed nodal ordering. This is because, while the dominance-proving rules work by creating a sequence of mDAGs such that each one dominates the next, most of the nondominance-proving rules work by assigning a certain signature to each mDAG (such as its set of unrealizable supports or d-separation relations, for example), and comparing the signatures of two mDAGs. Therefore, such rules would not benefit by the inclusion of mDAGs that are not consistent with the fixed nodal ordering. As we will see, the only nondominance-proving rule that does not work in this way is the directed-edge-free rule, but this is not an issue because all of the directed-edge-free mDAGs are consistent with any nodal ordering.

3.5 Deriving the Observational Partial Order for 3-node mDAGs

Now that we presented our rules for proving observational dominance and nondominance between mDAGs, we will go back to the results for 3-node mDAGs consistent with a fixed nodal ordering, presented in Section 3.2.2, and explain how they are derived.

3.5.1 Observational Dominances

Take, for example, the mDAGs in the `Instrumental ABC` class of Fig. 3.2.1, that are reproduced in Fig. 3.5.1. As we will see now, the rules discussed in Section 3.3 establish that these three mDAGs are observationally equivalent to each other. Here, we will discuss the intuition behind those rules for the specific case at hand.

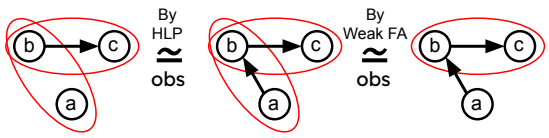


Figure 3.5.1: Three observationally equivalent mDAGs, that lie in the `Instrumental ABC` class.

In the first mDAG of Fig. 3.5.1, the node b already has access to all of the parents of a. Since the local randomness of a can also be absorbed by the latent parent that it shares with b, the node b already has access to all information about a. Therefore, this mDAG is just as powerful as another one where an edge between a and b is added, namely the second mDAG of Fig. 3.5.1. This is the application of the HLP Edge-Adding Rule (Proposition 4) to show the observational equivalence between the two first mDAGs of Fig. 3.5.1.

It is clear that the second mDAG of Fig. 3.5.1 observationally dominates the third mDAG of Fig. 3.5.1, because it structurally dominates it (Lemma 3). Therefore, we only need to show dominance in the other way to prove equivalence; this proof will be explained using the pDAGs of Fig. 3.5.2. Note that the first pDAG of Fig. 3.5.2 corresponds to the second mDAG of Fig. 3.5.1, and the third pDAG of Fig. 3.5.2 corresponds to the third mDAG of Fig. 3.5.1. So, we need to show that the first pDAG is observationally dominated by the third pDAG in Fig. 3.5.2.

In the first pDAG of Fig. 3.5.2, the node a is influenced by the latent common cause β that it shares with the node b. As such, learning the value of X_a teaches us something about

the latent common cause β via Bayesian updating. As will be formalized in Lemma 22 of Appendix B.2, this is at most as powerful as a situation where a acts as a setting for β. This implies that the first pDAG of Fig. 3.5.2 is observationally dominated by the second pDAG of Fig. 3.5.2.

Via exogenization of latent nodes, the second pDAG of Fig. 3.5.2 is observationally equivalent to the third pDAG of Fig. 3.5.2. Thus, we conclude that the third mDAG of Fig. 3.5.1 observationally dominates the second mDAG of Fig. 3.5.1, which shows that they are observationally equivalent. This is the implementation of Weak Facet-Adding (Proposition 5) to show the observational equivalence of these two mDAGs.

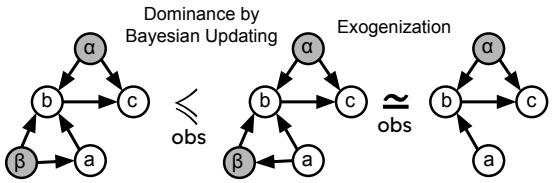


Figure 3.5.2: Intermediary steps to show observational equivalence between the second and the third mDAGs of Fig. 3.5.1.

In the same fashion, the Structural Dominance rule, the HLP Edge-Adding Rule and the Weak Facet-Adding rule can be used to show that the partition of the 72 3-node mDAGs consistent with a fixed nodal ordering into the 15 blocks that are depicted in Fig. 3.2.1 is a proven-equivalence partition, that is, all of the mDAGs inside the same set are observationally equivalent to each other. Moreover, these rules can show that all of the lines that connect one set to another are proven observational dominances.

It is worth noting that even the Evans' Proposition (Proposition 6), which is subsumed by HLP+Weak Facet-Adding (Lemma 9), is sufficient to show all of the observational equivalences and dominances of Fig. 3.2.1.

What is left to be justified is that no observational dominance between 3-node mDAGs consistent with a fixed nodal ordering is missing from Fig. 3.2.1[9].

[9] In Figure 13 of Ref. [33], representatives of 8 blocks of a proven-equivalence partition of 3-node mDAGs obtained by Evans's Rule were shown. That proven-equivalence partition is different than the one we present here for two reasons: first, the organization of Ref. [33] is not focused only on mDAGs consistent with a fixed nodal ordering. Second, that organization only counts mDAGs up to permutation of nodes.

3.5.2 Observational Nondominances

As we will show now, for the case of 3-node mDAGs the proven-dominance partial order established in the previous section corresponds to the observational partial order. That is, every time that one 3-node mDAG $\mathfrak{G}$ cannot be shown to observationally dominate another 3-node mDAG $\mathfrak{G}'$ by the conjunction of the Structural Dominance rule, the HLP Edge-Adding Rule and the Weak Facet-Adding rule, then it is possible to prove that $\mathfrak{G}$ *does not* observationally dominate $\mathfrak{G}'$.

We will do so by applying the nondominance-proving rules of Fig. 3.4.1 starting from the computationally cheaper rules and progressing to more expensive ones. Therefore, the first rule we apply is the comparison of skeletons.

The blocks of the proven-inequivalence partition induced by the comparison of skeletons rule are depicted by the loops of Fig. 3.5.3. Each cut arrow ($\nrightarrow$) between two blocks indicates a proven nondominance relation between the mDAGs of those blocks. For example, because the skeleton of `Factorizing AC|B` is a subgraph of the skeleton of `Collider C`,the former cannot observationally dominate the latter. The skeleton of `Factorizing AC|B` is also a subgraph of the skeleton that is common to `Evans`, `Instrumental BAC`, `Instrumental CAB`, `Collider A`, and `Fork`. This implies observational nondominances `Factorizing AC|B` over each of these. In Fig. 3.5.3 we represent all of these proven nondominances by one single cut arrow from `Factorizing AC|B` to the block of the proven-inequivalence partition induced by comparison of skeletons that contains `Evans`, `Instrumental BAC`, `Instrumental CAB`, `Collider A`, and `Fork`. Because nondominance is not a transitive property, we must explicitly represent *all* of the proven nondominances in Fig. 3.5.3. For example, there is a cut arrow into the `Saturated` block from every other block of the proven-inequivalence partition.

As we can see in Fig. 3.5.3, the comparison of skeletons is sufficient to show that the following blocks of the proven-equivalence partition are observational equivalence classes: `Factorizing`, `Factorizing AC|B`, `Factorizing AB|C`, `Factorizing B|CA` and `Collider C`. Furthermore, it shows that in all of the instances where an observational dominance involving one of these five classes could not be proven by the conjunction of the structural dominance rule, the HLP Edge-Adding rule and the Weak Facet-Adding rule, then there is a proven observational nondominance. Therefore, the part of the observational partial order concerning these five classes is solved by the comparison of skeletons.

Next, we proceed to applying the comparison of d-separation relations to the remaining mDAGs. The proven-equivalence blocks `Saturated`, `Triangle`, `Evans`, `Instrumental BAC`, `Instrumental CAB` and `Instrumental ABC` do not present any d-separation relations.

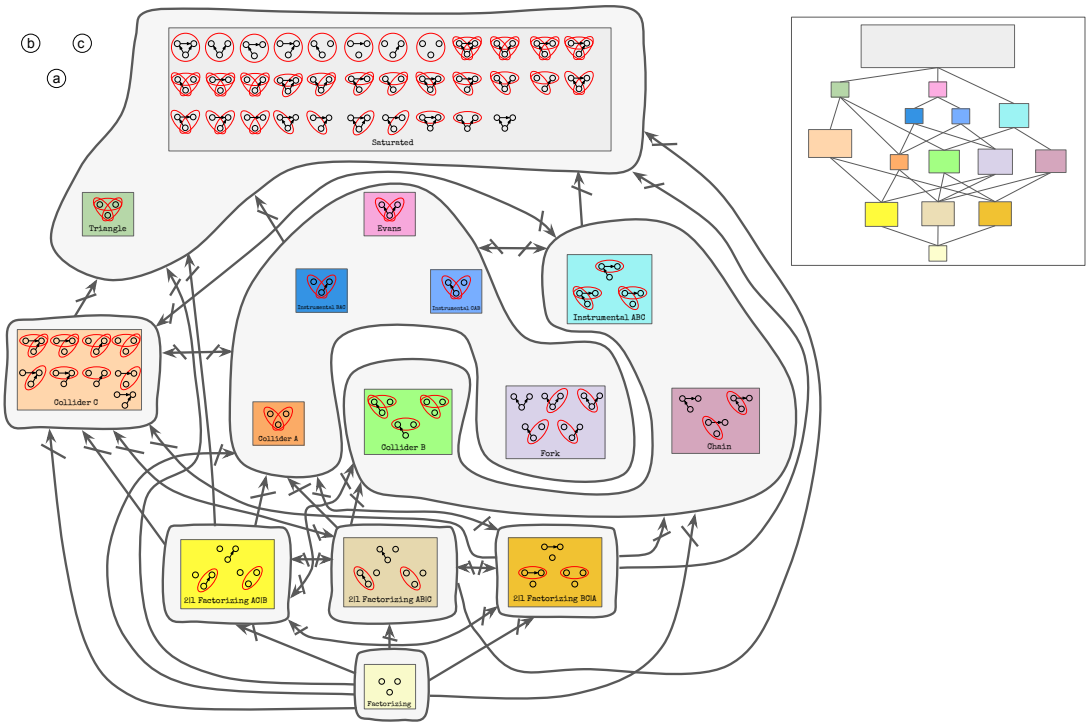


Figure 3.5.3: Representation of the information obtained by applying the comparison of skeletons to the proven-equivalence partition of 3-node mDAGs. In this figure, the loops represent blocks of the proven-inequivalence partition and the cut arrows ($\nrightarrow$) represent proven nondominances. Double-headed cut arrows mean that there is proven nondominance in both directions, i.e., proven incomparability. Compare this to the set of observational *dominances* proven by Structural Dominance, the HLP Edge-Adding rule and Weak Facet-Adding, reproduced in the miniature on the top right.

On the other hand, `Collider A` presents the d-separation relation $b \perp\!\!\!\perp_{\mathrm{d}} c$, `Collider B` presents the d-separation relation $a \perp\!\!\!\perp_{\mathrm{d}} c$, `Fork` presents the d-separation relation $b \perp\!\!\!\perp_{\mathrm{d}} c|a$ and `Chain` presents the d-separation relation $a \perp\!\!\!\perp_{\mathrm{d}} c|b$.

From this, we are able to show more nondominances, and in particular split some of the blocks of the proven-inequivalence partition induced by comparison of skeletons. When we present the resulting nondominances in Fig. 3.5.4, to make it easier to visualize, we do not include the classes that were already completely classified with the comparison of skeletons alone (i.e., `Factorizing`, `Factorizing AC|B`, `Factorizing AB|C`, `Factorizing B|CA` and `Collider C`).

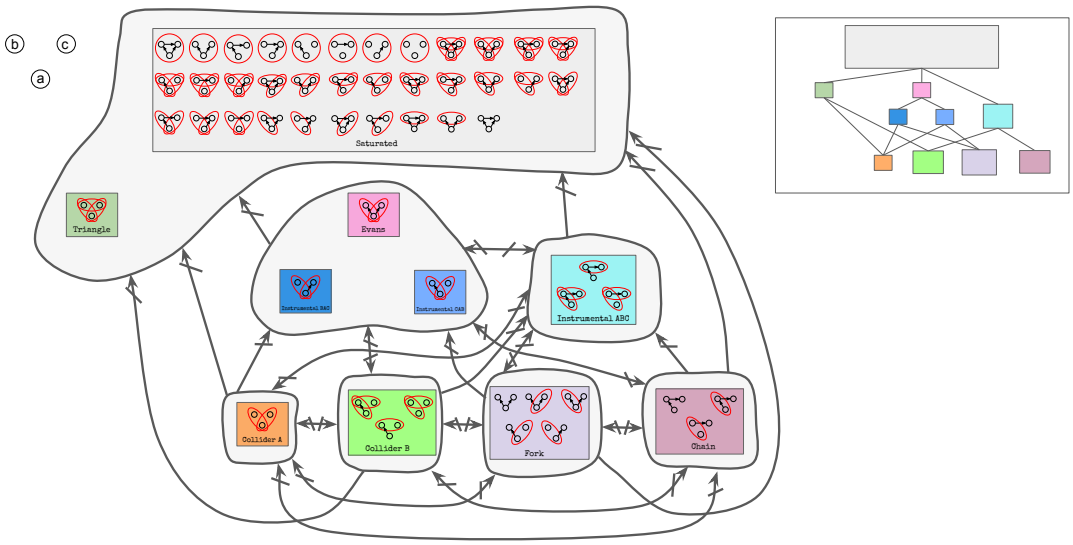


Figure 3.5.4: Representation of the information obtained by applying the comparison of skeletons and the comparison of d-separation relations to the proven-equivalence partition of 3-node mDAGs. Here, we only show the blocks of the proven-equivalence partition that were not already shown to be observational equivalence classes by the comparison of skeletons alone. The loops represent blocks of the proven-inequivalence partition and the cut arrows ($\nrightarrow$) represent proven nondominances. Double-headed cut arrows mean that there is proven nondominance on both directions, i.e., proven incomparability. Compare this to the set of observational *dominances* among these mDAGs proven by Structural Dominance, the HLP Edge-Adding rule and Weak Facet-Adding, reproduced in the miniature on the top right.

As we can see in Fig. 3.5.4, by applying the comparison of d-separation relations to the blocks of the proven-inequivalence partition induced by the comparison of skeletons, we further learn that `Instrumental ABC`, `Collider A`, `Collider B`, `Fork` and `Chain` are observational equivalence classes. Furthermore, we learn that in all of the instances where an observational dominance involving one of these five classes could not be proven by the conjunction of the structural dominance rule, the HLP Edge-Adding rule and the Weak Facet-Adding rule, then there is a proven observational nondominance. Therefore, the part of the observational partial order concerning these five classes is solved by comparison of skeletons together with comparison of d-separation relations.

Next, we attempt to apply the comparison of e-separation relations and the comparison of densely connected pairs to the remaining mDAGs. This, however, does not show any nondominance that was not already known. Therefore, we move to applying the directed-

edge-free rule, and this shows one extra nondominance: `Triangle` does not dominate `Saturated`. With this, we learn that `Triangle` is an observational equivalence class.

At this point, we have the proven-inequivalence partition induced by all nondominance rules except for comparison of unrealizable supports. It is this last rule that establishes the remaining nondominances embodied in the observational partial order described in Fig. 3.2.1.

We start by applying the comparison of unrealizable supports on binary variables constituted of two events. Consider the set constituted by the events

$$\{X_a = 0, X_b = 0, X_c = 0\},$$
$$\{X_a = 1, X_b = 1, X_c = 1\}.$$

As a support, this set expresses perfect correlation between X_a, X_b and X_c. This support is realizable by all of `Evans`, `Instrumental ABC`, `Instrumental BAC`, `Instrumental CAB`, `Fork` and `Chain`, but it is *not* realizable by `Triangle`. Therefore, `Triangle` does not observationally dominate the aforementioned proven-equivalence blocks. With this, we learn that the observational dominances involving `Triangle` that can be proven using Structural Dominance, HLP Edge-Adding and Weak Facet-Adding are indeed all of the observational dominance relations involving `Triangle`.

When we apply the comparison of unrealizable supports on binary variables constituted of three events, we can show more nondominances. Let S be the set constituted by the events

$$\{X_a = 0, X_b = 0, X_c = 0\},$$
$$\{X_a = 0, X_b = 1, X_c = 1\},$$
$$\{X_a = 1, X_b = 0, X_c = 0\}.$$

The set S is a support realizable by `Evans`, but unrealizable by `Instrumental BAC` and `Instrumental CAB`. This was found by Fraser's algorithm, but here we give a direct proof. First, an explicit choice of parameters can show that there are distributions with support S that are realizable by `Evans`. Say that the common cause between a and c is λ, and the common cause between a and b is γ. Take each of X_λ and X_γ to be binary and take the distribution over them to be full support. Then, let $X_a = X_\lambda \oplus X_\gamma$, where $\oplus$ means sum modulus 2. Furthermore, let $X_b = X_\lambda \cdot (X_a \oplus 1)$ and $X_c = X_\gamma \cdot (X_a \oplus 1)$. From these

functions, we have:

$$(X_\lambda, X_\gamma) = (0,0) \Rightarrow X_a = 0, X_b = 0, X_c = 0$$
$$(X_\lambda, X_\gamma) = (1,1) \Rightarrow X_a = 0, X_b = 1, X_c = 1$$
$$(X_\lambda, X_\gamma) = (0,1) \Rightarrow X_a = 1, X_b = 0, X_c = 0$$
$$(X_\lambda, X_\gamma) = (1,0) \Rightarrow X_a = 1, X_b = 0, X_c = 0.$$

That is, choices of parameters of this type give rise to distributions that have support S.

Note that this construction is possible because in `Evans` both b and c have direct access to a: they need to know when to copy the hidden variable and when to set their value to 0. Therefore, this specific construction is not possible in `Instrumental BAC` or `Instrumental CAB`. More than that, any distribution with support S violates the two versions of Pearl's instrumental inequalities [56] that are respectively associated with `Instrumental BAC` and `Instrumental CAB`, and thus is not realizable by these mDAGs. Therefore, this explicitly shows us that `Instrumental BAC` and `Instrumental CAB` do not dominate `Evans`. We infer that `Evans` is an observational equivalence class.

Finally, we apply the comparison of unrealizable supports on binary variables constituted of four events. Let S' be the set constituted by the events

$$\{X_a = 1, X_b = 0, X_c = 0\},$$
$$\{X_a = 0, X_b = 0, X_c = 1\},$$
$$\{X_a = 0, X_b = 1, X_c = 1\},$$
$$\{X_a = 0, X_b = 0, X_c = 0\}.$$

The set S' is a support realizable by `Instrumental CAB`, but unrealizable by `Instrumental BAC`. This was found by Fraser's algorithm, but here we give a direct proof. First, an explicit choice of parameters can show that there are distributions with support S' that are realizable by `Instrumental CAB`. Let the common cause between a and b be λ, and the common cause between a and c be γ. Take each of X_λ and X_γ to be binary and take the distribution over them to be full support. Then, let $X_a = X_\lambda \cdot X_\gamma$, $X_b = X_\lambda \cdot (X_a \oplus 1)$ and $X_c = X_\gamma \oplus 1$. Choices of parameters of this type give rise to distributions that have support S'. On the other hand, it is not hard to see that the any distribution with support S' violates the version of Pearl's instrumental inequalities [56] that are associated with `Instrumental BAC`. To find a support that is realizable by `Instrumental BAC` but unrealizable by `Instrumental CAB`, we only need to invert b and c in S'.

Methods Applied	Cardinality of the induced proven-equivalence partition
None	72
SD + HLP	44
SD + HLP + Weak FA	15

Table 3.1: Number of blocks of the proven-equivalence partition for 3-node mDAGs consistent with a fixed nodal ordering obtained by applying various combinations of the dominance-proving rules of Section 3.3. The row "None" gives the total number of 3-node mDAGs consistent with a fixed nodal ordering. These numbers give upper bounds on the number of observational equivalence classes.

Together, all of the arguments presented above close the problem of finding the observational partial order of observational equivalence classes of 3-node mDAGs consistent with a fixed nodal ordering, showing that it is indeed the one shown in Fig. 3.2.1.

Tables 3.1 and 3.2 give a selective summary of the steps shown in this section: Table 3.1 presents the number of elements of the proven-equivalence partition of 3-node mDAGs consistent with a fixed nodal ordering induced by different combinations of the dominance-proving rules, and Table 3.2 presents the number of elements of the proven-inequivalence partition of 3-node mDAGs consistent with a fixed nodal ordering induced by different combinations of the nondominance-proving rules.

3.6 Partial Solution of Observational Partial Order for 4-node mDAGs

Based on the case of three visible nodes, one might wonder whether Structural Dominance, HLP Edge-Adding and Weak Facet-Adding, that is, the dominance-proving rules that were known prior to this chapter, are sufficient to find the observational partial order for any number of visible nodes. However, the case of four visible nodes shows that this expectation is not borne out: Table 3.3 shows that the number of elements of the proven-equivalence partition, which is an upper bound on the number of observational equivalence classes, gets smaller upon successive applications of the equivalence rules of Section 3.3. Thus, our two extensions (Moderate and Strong Facet-Adding) are indeed necessary to establish certain dominances and equivalences.

Table 3.4 shows the number of elements of the proven-inequivalence partition that is obtained after applying each one of the nondominance-proving rules of Section 3.4 to

Methods Applied	Cardinality of the induced proven-inequivalence partition
skel	8
d-sep+skel	12
e-sep	12
DC + e-sep	12
DEF + DC + e-sep	13
DEF + DC + e-sep + Supps up to 2 events	13
DEF + DC + e-sep + Supps up to 3 events	14
DEF + DC + e-sep + Supps up to 4 events	15

Table 3.2: Number of blocks of the proven-inequivalence partition for 3-node mDAGs consistent with a fixed nodal ordering obtained by applying various combinations of the nondominance-proving rules of Section 3.4. These numbers give lower bounds on the final number of equivalence classes. The fact that the last line reaches 15, which is the upper bound obtained from the proven-equivalence partition, shows that we have identified the observational equivalence partition for 3-node mDAGs consistent with a fixed nodal ordering.

mDAGs consistent with a fixed nodal ordering. In other words, Table 3.4 is the analogue of Table 3.2 for the case of four visible nodes. As we can see, we have not solved all ambiguity about the observational equivalence partition for 4-node mDAGs consistent with a fixed nodal ordering: the best upper bound on the number of observational equivalence classes that we have obtained is 1444, while the best lower bound is 1256.

It is still possible that the reason we did not solve the problem is that we did not use the full power of the comparison of unrealizable supports: we only compared supports for binary variables and only up to 8 events. In Section 3.6.3, we will study one specific block of the proven-inequivalence partition, and show that there are still cases where whether observational equivalence holds or not is unknown even after making all of the possible comparisons of unrealizable supports while assuming that all visible variables are binary, i.e., for $\vec{c}_{\text{vis}} = (2, 2, 2, 2)$. We will then discuss the possibility of comparing unrealizable supports under the assumption that one or more of the visible variables have higher cardinality, even though this becomes increasingly impractical and computationally expensive.

Before doing so, we will look into the set of observational equivalence classes that were completely characterized by our analysis.

Methods Applied	Cardinality of the induced proven-equivalence partition
None	7296
SD + HLP	4417
SD + HLP + Weak FA	1481
SD + HLP + Moderate FA	1466
SD + HLP + Strong FA	1444

Table 3.3: Cardinalities of the proven-equivalence partitions induced by various combinations of the dominance-proving rules of Section 3.3 for the set of 4-node mDAGs consistent with a fixed nodal ordering. The row "None" gives the total number of 4-node mDAGs consistent with a fixed nodal ordering. These cardinalities give successively tighter upper bounds on the number of observational equivalence classes.

Methods Applied	Cardinality of the induced proven-inequivalence partition
None	1
skel	64
d-sep+skel	259
e-sep	326
DC + e-sep	334
DEF + DC + e-sep	350
DEF + DC + e-sep + Supps up to 2 events	447
DEF + DC + e-sep + Supps up to 3 events	595
DEF + DC + e-sep + Supps up to 4 events	1054
DEF + DC + e-sep + Supps up to 5 events	1153
DEF + DC + e-sep + Supps up to 6 events	1243
DEF + DC + e-sep + Supps up to 7 events	1253
DEF + DC + e-sep + Supps up to 8 events	1253

Table 3.4: Cardinalities of the proven-inequivalence partitions induced by various combinations of the nondominance-proving rules of Section 3.4 for the set of 4-node mDAGs consistent with a fixed nodal ordering. These cardinalities give successively tighter lower bounds on the number of observational equivalence classes. All the support comparison applied assumed all the visible variables to be binary, i.e., $\vec{c}_{\text{vis}} = (2, 2, 2, 2)$.

Methods Applied	Nº of Identified Observational Equivalence Classes
None	0
skel	15
d-sep	114
d-sep+skel	152
e-sep	174
DC + e-sep	186
DEF + DC + e-sep	218
DEF + DC + e-sep + Supps up to 2 events	298
DEF + DC + e-sep + Supps up to 3 events	378
DEF + DC + e-sep + Supps up to 4 events	859
DEF + DC + e-sep + Supps up to 5 events	990
DEF + DC + e-sep + Supps up to 6 events	1136
DEF + DC + e-sep + Supps up to 7 events	1156
DEF + DC + e-sep + Supps up to 8 events	1156

Table 3.5: Number of identified classes of 4-node mDAGs consistent with a fixed nodal ordering obtained by applying various combinations of the nondominance-proving rules of Section 3.4. These are sets of mDAGs that are both blocks of the proven-equivalence partition and of the proven-inequivalence partition.

3.6.1 Completely Identified Classes

When a block of the proven-equivalence partition is also a block of the proven-inequivalence partition (a pink loop coincides with a gray loop in Fig. 3.2.2), we infer that it is a block of the observational equivalence partition (green dashed line in Fig. 3.2.2). In that case, we say that the observational equivalence class is "identified".

For the case of 3-node mDAGs consistent with a fixed nodal ordering, as discussed, *all* of the classes are identified when we apply the nondominance rules of Section 3.4, where the comparison of unrealizable supports is applied for visible variables that are all binary and up to 4 events. For the case of four visible nodes, on the other hand, even the comparison of unrealizable supports for binary variables up to 8 events cannot identify all classes. However, some equivalence classes *can* be identified. Table 3.5 presents the number of classes that are identified by various combinations of nondominance-proving rules.

In particular, with this analysis we reproduced one of the results of Ref. [83]: the observational equivalence class of the Bell mDAG (Fig. 3.6.1(a)) consists of the 9 mDAGs

whose associated pDAGs are consistent with a fixed nodal ordering depicted in Figure 24 of Ref. [83]. The Bell equivalence class is identified by d-separation alone, meaning that the particular pattern of d-separation relations presented by this observational equivalence class is not presented by any other class. In other words, there is *no* mDAG that possesses the same pattern of d-separation relations as the Bell mDAG while being inequivalent to it.

The main point of Ref. [83] is that if one has a probability distribution that violates Bell inequalities (which are the inequality constraints imposed by the Bell mDAG) but satisfies the no-signalling relations (which are the conditional independence constraints imposed by the Bell mDAG), then it is only possible to find a causal explanation for this distribution in the framework of *classical* causal models if one allows for *fine-tuning*. Fine-tuning, which is another name for the notion of *unfaithfulness* [74], is the property that a probability distribution exhibits conditional independence relations that are not entailed by the d-separation relations of the causal structure that realized this distribution. Those extra conditional independence relations arise due to the specific choice of parameters that is used to realize the distribution in question, and not due to the structure of the DAG. As it turns out, the conclusion that was drawn in Ref. [83] for the Bell causal structure can be extended to any observational equivalence class of causal structures that is nonalgebraic and is identified by d-separation alone. For such a class of causal structures, any distribution that satisfies all and only the conditional independence relations implied by the set of d-separation relations of that class, but that violates the inequality constraints of the class necessarily cannot be explained within the framework of classical causal models without fine-tuning.

Fig. 3.6.1(b) shows an example of a causal structure (distinct from the Bell causal structure) whose equivalence class is identified by d-separation alone: the `Square` mDAG. The argument above says that any distribution whose set of conditional independence relations is exactly $X_a \perp\!\!\!\perp X_c$ and $X_d \perp\!\!\!\perp X_b$ but that violates the inequality constraints implied by the `Square` mDAG does not admit of any non-fine-tuned causal explanation in the framework of classical causal models.

Now, we go back to one of our motivations discussed in the Introduction: the question of the significance of analyzing inequality constraints for causal discovery.

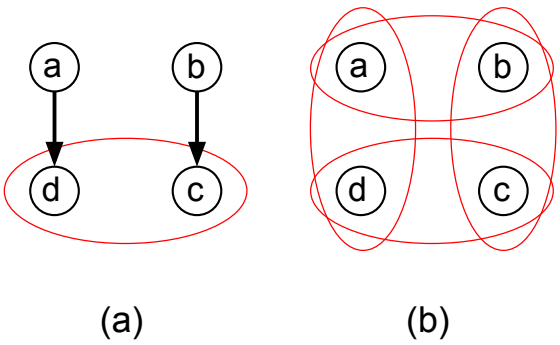


Figure 3.6.1: (a) Bell mDAG. (b) `Square` mDAG. Both of these belong to equivalence classes that are completely identified by d-separation, meaning that there is no mDAG that is observationally inequivalent to them while still imposing the same conditional independence constraints on the realizable distributions.

3.6.2 Importance of Inequality Constraints for Causal Discovery

Nonalgebraicness is Ubiquitous

As discussed in Section 3.2.2, in the set of 3-node mDAGs consistent with a fixed nodal ordering, there are 15 observational equivalence classes and 5 of them are nonalgebraic. Therefore, for the case of three visible variables, the fraction of nonalgebraic classes among all classes of mDAGs consistent with a fixed nodal ordering is $1/3$. As we will see in this section, the fraction of nonalgebraic classes gets much bigger in the case of 4-node mDAGs consistent with a fixed nodal ordering.

Since we still have not obtained a complete characterization of the observational equivalence partition of 4-node mDAGs consistent with a fixed nodal ordering, the best we can do is to find a lower bound on the fraction of nonalgebraic equivalence classes. We will explain two methods for obtaining such a lower bound. We then pick the best lower bound out of the two results. Method 1 consists in obtaining a *lower bound on the number of nonalgebraic classes*, and dividing it by our best *upper bound* on the total number of observational equivalence classes (which is currently 1444). The resulting number f_1 is a lower bound on the fraction of nonalgebraic classes. Method 2 consists in obtaining an *upper bound on the number of algebraic classes*, and dividing it by our best *lower bound* on the total number of observational equivalence classes (which is 1253). The resulting number f_2 is such that $1 - f_2$ is a lower bound on the fraction of nonalgebraic classes.

We first implement method 1. An observational equivalence class of mDAGs consistent with a fixed nodal ordering is identified as nonalgebraic, according to Proposition 3, if its

mDAGs are not observationally equivalent to *any* latent-confounder-free mDAG, including ones that are not consistent with the fixed nodal ordering. We thus look at the proven-inequivalence partition and, using the nondominance-proving rules, find all of the blocks thereof whose mDAGs are proven to be inequivalent to every latent-confounder-free mDAG (including ones that are not consistent with the chosen nodal ordering). The number of blocks of this type is a lower bound on the number of nonalgebraic observational equivalence classes. Applying this method on the proven-inequivalence partition induced by all the nondominance rules of Section 3.4 when comparison of unrealizable supports is made up to 8 events, we obtain the lower bound of 1137 on the number of nonalgebraic observational equivalence classes. Therefore, the lower bound on the fraction of nonalgebraic classes obtained by method 1 is $f_1 = 1137/1444 \approx 78.7\%$.

Now we implement method 2, where we want to find an *upper bound* on the number of algebraic classes. Again according to Proposition 3, an observational equivalence class of mDAGs consistent with a fixed nodal ordering is algebraic only if its mDAGs are observationally equivalent to *some* latent-confounder-free mDAG (not necessarily one that is consistent with the fixed nodal ordering). This implies that the *total number* of observationally inequivalent latent-confounder-free mDAGs (not restricting to a set that is consistent with a fixed nodal ordering) gives us an upper bound on the number of algebraic observational equivalence classes in the set of mDAGs that are consistent with a fixed nodal ordering. It is easy to find this number, since the observational profile of latent-confounder-free mDAGs is completely characterized by d-separation. Doing this, we find that the number of algebraic classes of 4-node mDAGs in the set of mDAGs that are consistent with a fixed nodal ordering is *at most* 185. To obtain the maximum fraction of algebraic mDAGs consistent with a fixed nodal ordering, we divide this number by our best lower bound on the total number of equivalence classes, thus obtaining $f_2 = 185/1253 \approx 14.8\%$. Therefore, the lower bound on the fraction of nonalgebraic classes obtained by method 2 is $1 - f_2 \approx 85.2\%$. The lower bound obtained by method 1 is worse than the lower bound obtained by method 2, so we will use the latter.

To conclude, the rate of incidence of inequality constraints among observational equivalence classes of 4-node mDAGs consistent with a fixed nodal ordering is *at least* 85.2%. This high rate and the fact that it increases with the number of nodes suggests that inequality constraints are not rare; on the contrary, they seem to be generic. This is of special relevance to quantum physicists interested in causal inference, as it shows that there is much potential for finding new observational equivalence classes of causal structures that exhibit a quantum-classical gap.

Singling out equivalence classes under the assumption of faithfulness via equality and inequality constraints

As discussed in Section 3.2.3, without assuming any model-selection principles, the most we can learn about the underlying causal structure starting from one observational distribution is the upward closure (in the observational partial order) of a given set of observational equivalence classes. In this section, we discuss the case where we assume faithfulness to get a finer-grained estimate of the underlying causal structure, and consider the importance of inequality constraints for causal discovery in this case.

Under the assumption of faithfulness, we only consider candidate causal structures that impose all and only the conditional independence relations that appear in our observed probability distribution. In general, this does not single out only one observational equivalence class: there could be multiple equivalence classes whose set of conditional independence constraints correspond exactly to the relations that appear in our data, but that have different sets of nested Markov constraints or inequality constraints. In this case, we would have to test our distribution against these extra constraints (and our distribution could turn out to be realizable by more than one of these classes).

However, there are special cases where one *can* single out one observational equivalence class under the assumption of faithfulness without having to check nested Markov constraints or inequality constraints: when the exact set of d-separation relations of that observational equivalence class does not appear in any other observational equivalence class. When this happens, using the language of Section 3.6.1, the observational equivalence class is identified by the comparison of d-separation relations. We will say that such classes are *identified by conditional independence relations alone*; in Fig. 3.6.1, we gave examples of two mDAGs, Bell and Square, that belong to equivalence classes of this type. If our observed probability distribution satisfies all and only the conditional independence relations that are imposed by an equivalence class that is identified by conditional independence relations alone, then that class provides the only possible faithful causal explanation to our observed data. If the conditional independence relations of the set of data do not correspond to those of any equivalence class that is identified by conditional independence relations alone, then there is remaining ambiguity about the observational equivalence class that explains the data, even when assuming faithfulness.

For the case of 2-node mDAGs there are no inequality constraints, so every observational equivalence class can be identified by conditional independence relations alone. For the case of 3-node mDAGs consistent with a fixed nodal ordering, there are 9 equivalence classes that can be identified by d-separation alone. The remaining classes, namely, `Triangle`, `Evans`, `Instrumental ABC`, `Instrumental BAC`, `Instrumental CAB` and `Saturated`, are

all devoid of nontrivial d-separation relations. Therefore, for this case the fraction of causal structures that can be singled out by the analysis of conditional independence constraints alone is $3/5$. For the case of 4-node mDAGs consistent with a fixed nodal ordering, Table 3.5 shows that only 114 equivalence classes are identified by conditional independence relations alone. This is less than 10% of our best lower bound on the total number of equivalence classes. Therefore, in most cases it will *not* be possible to discriminate between observationally inequivalent classes of causal structures if one only avails oneself of conditional independence relations.

Therefore, our analysis shows that as well as the fraction of nonalgebraic classes increasing with the number of nodes for the cases of 2, 3 and 4 nodes, the fraction of all observational equivalence classes that can be singled out by analysis of conditional independence relations alone drastically decreases with the number of nodes for the cases of 2, 3 and 4 nodes.

When the set of conditional independence relations satisfied by our data corresponds to the set of d-separation relations of a class that is identified by conditional independence relations alone, we might still want to check the data against the inequality constraints of the class. If the data does not satisfy these inequality constraints, then, as already discussed in Section 3.6.1, there are no classical causal structures that can faithfully generate this data.

3.6.3 mDAGs that can Realize Every Support on Binary Variables

In this section, we will discuss a particular set of 4-node mDAGs consistent with a fixed nodal ordering: those that have no unrealizable supports on binary variables, that is, no unrealizable supports when $\vec{c}_{\text{vis}} = (2, 2, 2, 2)$. In other words, an mDAG $\mathfrak{G}$ of this set, whose canonical pDAG is $\mathcal{G}$, is such that $\mathcal{S}(\mathcal{G}, \vec{c}_{\text{vis}} = (2, 2, 2, 2))$ includes all possible sets of events over four binary variables. This set of mDAGs includes all of the mDAGs that are proved to be saturated (i.e., the saturated proven-equivalence block), which consists of 2134 mDAGs, together with the extra 7 mDAGs depicted in Fig. 3.6.2. Each one of these 7 mDAGs belongs to a different block of the proven-equivalence partition, and each of these blocks includes just one mDAG.

As it turns out, this set of mDAGs is a block of the proven-inequivalence partition induced by all the nondominance rules of Section 3.4 when comparison of unrealizable supports is made on $\vec{c}_{\text{vis}} = (2, 2, 2, 2)$ and up to 4 events. As one can check, comparing unrealizable supports on $\vec{c}_{\text{vis}} = (2, 2, 2, 2)$ that have larger numbers of events does not help,

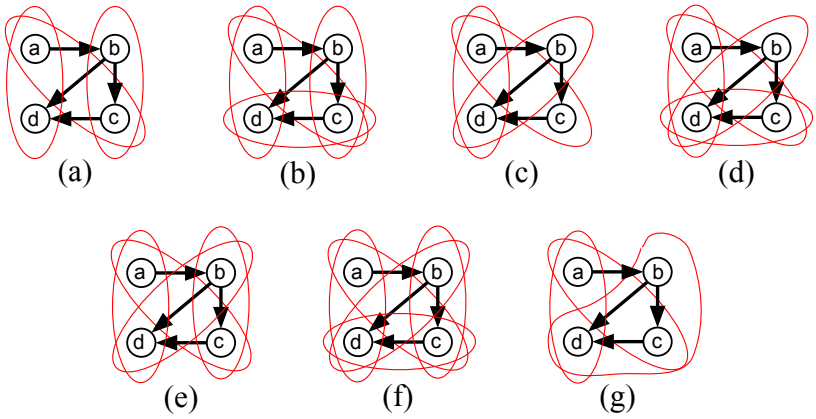


Figure 3.6.2: Seven mDAGs that do not appear in the same block as the saturated mDAGs in the proven-equivalence partition induced by the rules described in Section 3.3, but that nevertheless can realize any support on binary variables. mDAGs (a)-(d) are shown to *not* be saturated via analysis of supports on variables of higher cardinality. It is still unknown whether mDAGs (e)-(g) are observationally inequivalent to the saturated mDAGs.

as these mDAGs do not have any unrealizable supports on $\vec{c}_{\text{vis}} = (2, 2, 2, 2)$ even up to the maximum number of $2^4 = 16$ events.

If we go beyond $\vec{c}_{\text{vis}} = (2, 2, 2, 2)$, we can show that the mDAGs of Fig. 3.6.2(a)-(d) are *not* saturated. Specifically, for $\vec{c}_{\text{vis}} = (3, 2, 2, 2)$ they cannot realize the support constituted by the following events:

$$\{X_a = 0, X_b = 0, X_c = 0, X_d = 0\},$$
$$\{X_a = 0, X_b = 0, X_c = 1, X_d = 0\},$$
$$\{X_a = 0, X_b = 1, X_c = 0, X_d = 0\},$$
$$\{X_a = 1, X_b = 0, X_c = 0, X_d = 0\},$$
$$\{X_a = 1, X_b = 1, X_c = 0, X_d = 0\},$$
$$\{X_a = 2, X_b = 0, X_c = 0, X_d = 1\},$$
$$\{X_a = 2, X_b = 1, X_c = 1, X_d = 0\}.$$

This was already noted in Eq. (18) of Ref. [48]. In fact, the mDAGs of Fig. 3.6.2(a)-(d) correspond to Table II of Ref. [48].

It follows that the mDAGs of Fig. 3.6.2(a)-(d) are not saturated. Furthermore, by the comparison of unrealizable supports, we can also see that the support on $\vec{c}_{\text{vis}} = (3, 2, 2, 2)$

constituted by the following events cannot be realized by Fig. 3.6.2(c) and Fig. 3.6.2(d), but can be realized by Fig. 3.6.2(a) and Fig. 3.6.2(b):

$$\{X_a = 0, X_b = 0, X_c = 0, X_d = 0\},$$
$$\{X_a = 0, X_b = 0, X_c = 1, X_d = 0\},$$
$$\{X_a = 0, X_b = 1, X_c = 0, X_d = 0\},$$
$$\{X_a = 1, X_b = 0, X_c = 0, X_d = 1\},$$
$$\{X_a = 1, X_b = 1, X_c = 1, X_d = 0\},$$
$$\{X_a = 2, X_b = 0, X_c = 1, X_d = 1\},$$
$$\{X_a = 2, X_b = 1, X_c = 1, X_d = 1\}.$$

Therefore, we learn that the mDAGs of Fig. 3.6.2(a)-(b) are inequivalent to the mDAGs of Fig. 3.6.2(c)-(d), which gives us the proven-inequivalence blocks depicted in Fig. 3.6.3. This is the most we can learn about inequivalences between these mDAGs by checking supports on $\vec{c}_{\text{vis}} = (3, 2, 2, 2)$ up to 7 events.

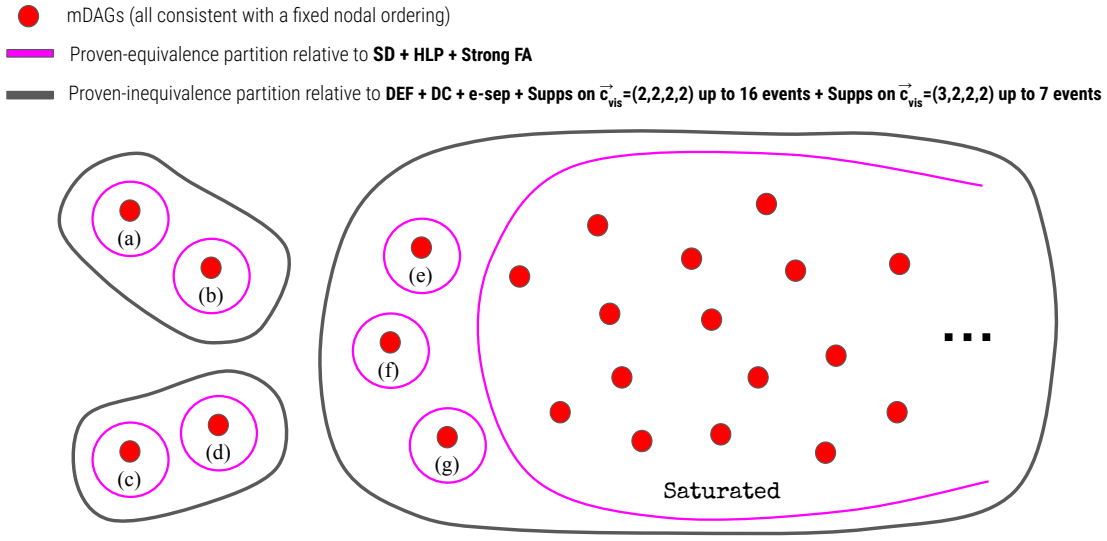


Figure 3.6.3: Diagrammatic representation of the proven-equivalence and proven-inequivalence partitions of 4-node mDAGs consistent with a fixed nodal ordering that are induced by the rules indicated in the labels of the figure. The saturated proven-equivalence block includes 2134 mDAGs consistent with a fixed nodal ordering. The letters (a)-(g) in this diagram refer to the corresponding mDAGs of Fig. 3.6.2.

Regarding mDAGs 3.6.2(e)-3.6.2(g), there are three possibilities. The first one is that, if we go to even higher cardinality of the visible variables, we can prove inequivalence by comparison of unrealizable supports. This option would only use theoretical tools that we already have at hand, but it would require significant computational power. The second option is that these mDAGs are actually saturated for all cardinalities of the visible variables, even if the dominance-proving rules that we have at hand cannot prove this. In this case, we would potentially be guided to new dominance-proving rules. The third possibility is that these mDAGs are *not* saturated, potentially even for binary cardinalities of the visible variables, but that this cannot be deduced from the comparison of unrealizable supports. That is, it could be the case that they have unrealizable *distributions* over binary variables, even if there is no unrealizable *support* over binary variables. If we find out that this is the case, then we would have found an example of pDAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ such that $\mathcal{M}_{\text{obs}}(\mathcal{G}_1, \vec{c}_{\text{vis}} = (2,2,2,2)) \neq \mathcal{M}_{\text{obs}}(\mathcal{G}_2, \vec{c}_{\text{vis}} = (2,2,2,2))$, while $\mathcal{S}(\mathcal{G}_1, \vec{c}_{\text{vis}} = (2,2,2,2)) = \mathcal{S}(\mathcal{G}_2, \vec{c}_{\text{vis}} = (2,2,2,2))$. This would imply that the converse implication to that of Eq. (3.3) *does not* hold, that is, that Conjecture 3 is false.

Note that this third case could also happen with mDAGs 3.6.2(a)-3.6.2(d): since they can realize any support on on $\vec{c}_{\text{vis}} = (2,2,2,2)$, if we find that they *are not* saturated on $\vec{c}_{\text{vis}} = (2,2,2,2)$ we would disprove Conjecture 3. On the other hand, note that if we find that they *are* saturated on $\vec{c}_{\text{vis}} = (2,2,2,2)$, we would disprove Conjecture 1, since we already know that these mDAGs are nonalgebraic.

It is worth noting that, by Lemma 3, the mDAG 3.6.2(a) is dominated by 3.6.2(b) and 3.6.2(e)-(g), while the mDAG 3.6.2(c) is dominated by 3.6.2(d) and 3.6.2(e)-(g). Therefore, if 3.6.2(a) and 3.6.2(c) are saturated for $\vec{c}_{\text{vis}} = (2,2,2,2)$, all of the others are as well.

3.7 Conclusion

In this chapter, we have gathered all the currently known techniques to prove observational dominance or nondominance between causal structures, extended the scope of these techniques, proved logical implications among them, and applied them to causal structures with three and four visible nodes.

For the case of three visible variables, Fig.13 of Ref. [33] had already presented representatives of blocks of a proven-equivalence partition. However, in Ref. [33] it was not explicitly shown that all of these were observational equivalence classes (i.e., that they were all observationally inequivalent), and the observational partial order was not studied. Here, we were able to deduce the observational equivalence classes and observational partial order for 3-node mDAGs.

For the case of four visible variables, we showed that the total number of observational equivalence classes for 4-node mDAGs consistent with a fixed nodal ordering is between 1253 and 1444. Furthermore, 1156 observational equivalence classes are already identified.

Classifying the classical causal structures into observational equivalence classes and establishing the observational partial order among these classes is important for several reasons. For one, it drastically lowers the number of possible causal explanations to be considered for a set of data relative to the total number of mDAGs. For another, the dominance order can be used to simplify the task of model selection (as discussed in Section 3.2.3). It also helps in assessing how many classes present non-trivial inequality constraints.

Inequality constraints seem to be generic in the sense of the number of observational equivalence classes having inequality constraints growing with the number of visible variables: it is $1/3$ for 3-node mDAGs consistent with a fixed nodal ordering, and at least 85.2% for 4-node mDAGs consistent with a fixed nodal ordering. We also argued against relying only on conditional independence relations for causal discovery on the basis that these constraints have a poor distinguishing power for observational equivalence classes: for 3-node mDAGs consistent with a fixed nodal ordering, $3/5$ of the observational equivalence classes can be identified by conditional independence relations alone; for 4-node mDAGs consistent with a fixed nodal ordering, however, this fraction drops to less than 10%. These arguments highlight the importance of inequality constraints and nested Markov constraints for the task of classical causal discovery.[10]

From the point of view of quantum causality, this result is also relevant because only the causal structures that exhibit inequality constraints have the potential of presenting quantum-classical gaps, that is, of being such that the distributions realizable by quantum semantics are a superset of the distributions realizable by classical semantics. All of the nonalgebraic mDAGs that we identified here were already identified in Ref. [48], but our classification shows that the *majority* of observational equivalence classes have the potential of presenting quantum-classical gaps. This shows that the property of exhibiting a quantum-classical gap might not be the exception, but rather the norm. In Chapter 5 we will argue in favor of this conclusion, by showing that the majority of nonalgebraic 4-node mDAGs indeed present a quantum-classical gap.

[10] We acknowledge that in real-life causal discovery scenarios, with many different variables involved, the analysis of conditional independence constraints is already difficult, and the analysis of inequality constraints may be more difficult still. However, our hope is that by recognizing the foundational importance of nested Markov constraints and inequality constraints, the community might put more emphasis towards developing techniques to study them.

Another result from this chapter that is relevant to quantum physicists who are interested in causal inference is an identification of the set of all 4-node mDAGs consistent with a fixed nodal ordering that admit a fine-tuning theorem similar to the one that holds for Bell's causal structure [83]. That is, if the set of conditional independence relations of an observed distribution matches exactly the set of d-separation relations of the mDAG but the distribution *violates* one of the inequality constraints of the mDAG, then this distribution *does not* admit any non-fine-tuned causal explanation in the classical framework for causal modelling. The mDAGs for which such an argument can be made are the ones that belong to observational equivalence classes identified by d-separation alone, one example being the `Square` structure of Fig. 3.6.1(b). This result is of interest for quantum physicists, because many times quantum theory predicts such violations.

To finish the classification of structures with four visible nodes, there are two options: (i) we continue the comparison of unrealizable supports (Section 3.4.6) but with higher cardinalities of the visible variables, that might complete the classification but will be computationally expensive. (ii) we find new dominance-proving or nondominance-proving rules that help us finish the classification. Note that these two options are not mutually exclusive: we might find a new nondominance-proving rule that helps us complete the classification, at the same time that a comparison of unrealizable supports at high enough cardinality would in principle also solve it.

Another open problem that this chapter has generated is the question of whether possibilistic equivalence implies probabilistic equivalence, as expressed in Conjecture 3. In other words, if two causal structures have the same set of unrealizable supports for a certain assignment of cardinalities of the visible variables, does this mean that they have the same set of realizable distributions for those cardinalities? This problem was not solved here, but a possible route of investigation was outlined: all the mDAGs of Fig. 3.6.2 can realize every support on binary variables, even if there is currently no rule that shows that 3.6.2(a)-(g) are observationally equivalent to the saturated class for binary variables. Therefore, if one can show that any of these mDAGs *cannot* realize every probability distribution on binary variables, one has found a counter-example to the conjecture that possibilistic equivalence implies probabilistic equivalence.

Data Availability

The code developed for this chapter is available at https://github.com/eliewolfe/mDAG-analysis

Chapter 4

Interventional and observational distinguishability of causal structures in the presence of latent and selected nodes

As we discussed in the introduction to this thesis, in general the correlation between two variables X_a and X_b can be due to three different types of causal mechanisms:

a) **Direct causal influence**, where X_a exerts an influence on X_b or vice-versa.

b) **Latent confounding**, where a common cause X_m induces correlation between X_a and X_b.

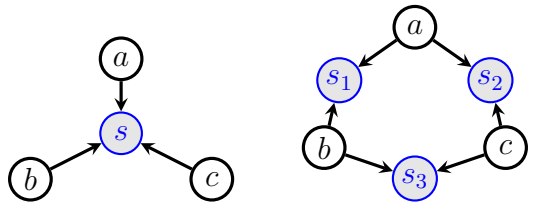
c) **Selection effects**, where conditioning on a variable X_s that is influenced by both X_a and X_b creates correlation (e.g. when we restrict the study sample).

For example, if we observe a correlation between obesity and mortality in a sample of hospital patients, this could be a) because obesity worsens heart disease, and/or b) because there is a health condition that causes weight loss, while also shortening lifespan, and/or c) because we are not taking data from the general population, but only from the hospital population. While obesity is positively correlated with mortality in the general population, among the hospital population they could be negatively correlated, because when a person in the hospital is underweight this can be due to serious disease.

So far, in this thesis we have considered only causal explanations of the two first types, direct causal influence and latent confounding. In this chapter, we will extend the analysis of the previous chapters to take selection effects into account. More specifically, we will introduce a third type of node, the *selected node*, such that the associated *selected variable* is conditioned to take one specific value (that we conventionally call 0). We will assume that *all* of the data obtained is always conditioned on the fact that the selected variables take the value 0.

We will define a structure called S-pDAG, composed by a DAG together with a partition of its nodes into visible, marginalized (previously called latent) and selected nodes. Through this chapter, visible nodes will be represented in black with a white background, marginalized nodes will be represented in red with a grey background and selected nodes will be represented in blue with a grey background. Figure 4.0.1 shows examples of S-pDAGs with visible and selected nodes, and Figure 4.0.2 shows examples of S-pDAGs with visible, marginalized and selected nodes.

There has been previous work on organizing S-pDAGs into equivalence classes whose members share the same set of *conditional independence constraints*, such as the work in maximal ancestral graphs (MAGs) [62]. As we have seen, however, only looking at conditional independence constraints is not sufficient to determine observational equivalence classes. The S-pDAGs of Figure 4.0.1 are yet another example of that: none of them exhibit any conditional independence constraints, but they are nevertheless observationally inequivalent; Figure 4.0.1a can realize the selected distribution $P(X_a = 1, X_b = 0, X_c = 0 \mid X_s = 0) = P(X_a = 0, X_b = 1, X_c = 0 \mid X_s = 0) = P(X_a = 0, X_b = 0, X_c = 1 \mid X_s = 0) = \frac{1}{3}$, while Figure 4.0.1b cannot realize the corresponding distribution (obtained from this one by substituting $X_s = 0$ on the right side of the conditionals by $X_{s_1} = X_{s_2} = X_{s_3} = 0$). Thus, MAGs clump together S-pDAGs that are not even observationally equivalent, let alone Observe&Do equivalent.



(a) An S-pDAG (b) Another S-pDAG

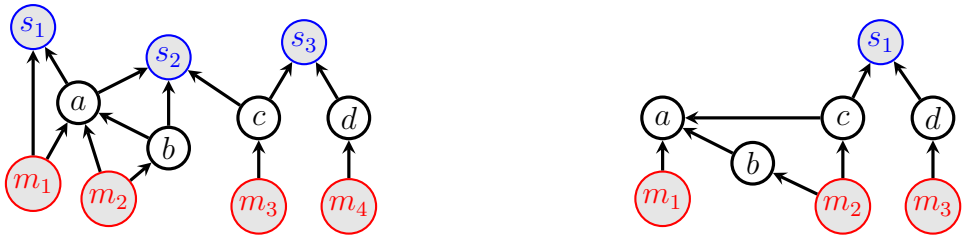
Figure 4.0.1: Two selected S-pDAGs that are interventionally distinguishable but correspond to the same MAG.

Therefore, previous literature lacks an analogue of the mDAG that faithfully represents what can be learned about the causal structure from experimental data in the presence of *both* marginalization and selection.

This chapter fills this gap by introducing the *selected-latent projection*, a map that

takes an S-pDAG with visible, marginalized and selected nodes and maps it to a *selected-marginalized directed graph (smDG)*, a new data structure that generalizes the mDAG. Analogously to what was done in Chapter 2 for the case without selection, here we show that two S-pDAGs are Observe&Do equivalent if and only if they project to the same smDG.

After that, as a secondary goal, we will explore alternative probing schemes for distinguishing different smDGs, such as passive observation or more limited interventional probing schemes. In particular, we will provide a series of rules to show that two smDGs are observationally equivalent, that is, indistinguishable by passive observations.



(a) An S-pDAG

(b) An alternative S-pDAG

Figure 4.0.2: Can these S-pDAGs produce different data on a, b, c, d after marginalization of the m_i nodes and selection of the s_i nodes? Spoiler: as we will see, they can be distinguished by interventions but not by passive observations.

4.1 Preliminaries

Our definition of a pDAG (Definition 3) allows for two types of nodes: visible and latent. In this chapter, we need three types of nodes, because some of the nodes will be selected upon. Furthermore, we will *not* use the terminology "latent node" in this chapter, as we understand it may cause confusion.[1] Instead, we will divide the nodes of a DAG into visible nodes, marginalized nodes (which play the same role as the latent nodes of previous chapters), and selected nodes. We define a new graphical structure accordingly:

Definition 26 (S-pDAG)**.** *An S-partitioned pDAG (S-pDAG) $\mathcal{G}$ is a DAG together with a partition of its nodes into three subsets: the set of visible nodes, denoted* $\texttt{vis}(\mathcal{G})$*, the set of marginalized nodes, denoted* $\texttt{lat}(\mathcal{G})$ *and the set of selected nodes, denoted* $\texttt{sel}(\mathcal{G})$*,*

[1] Since some could interpret the term "latent node" as a synonym of marginalized node, while others might interpret it as "non-visible node", which would also include selected nodes.

where $\texttt{vis}(\mathcal{G}) \cup \texttt{lat}(\mathcal{G}) \cup \texttt{sel}(\mathcal{G}) = \texttt{nodes}(\mathcal{G})$ *and* $\texttt{vis}(\mathcal{G}) \cap \texttt{lat}(\mathcal{G}) = \texttt{vis}(\mathcal{G}) \cap \texttt{sel}(\mathcal{G}) = \texttt{lat}(\mathcal{G}) \cap \texttt{sel}(\mathcal{G}) = \varnothing$.

The variables associated with visible, marginalized and selected nodes of an S-pDAG are respectively also called visible, marginalized and selected. We will refer to the union of marginalized *and* selected variables as *non-visible variables*.

When discussing the realizability of a probability distribution by a causal structure in the previous chapters, we were always implicitly assuming that every visible node has a marginalized parent, that is, that every node can depend stochastically on its parents. This was present for example in the fact that an error variable (Equation (2.1)) is always present. In this chapter, however, we will allow for the possibility of causal structures where we know that a visible variable is a *deterministic* function of its parents, with no local error variable. This corresponds to the notion of deterministic variables from Geiger et al. [44], and we use it because it makes our findings more general: the more common case where all variables are stochastic is the special case where each visible variable has a marginalized parent.[2]

Now, we will define the set of observational data realizable by an S-pDAG. In this chapter, such set will be called the *selected-marginal observational model* of the S-pDAG. Note that this is the analogue of the $\mathcal{M}-\texttt{profile}$ of the previous chapters, but in this chapter we will not focus on the subtleties regarding the cardinalities of the visible variables that we focused on when defining $\mathcal{M}-\texttt{profile}$. Instead, the selected-marginal observational model will be a set that includes *all* of the distributions realizable by this S-pDAG, regardless of the cardinality of the visible variables.

In the selected data, each selected variable X_s, $s \in S$, is fixed to one value which we conventionally call 0 [6]. We assume that there is *no* access to data obtained by selecting upon other values, such as $P(X_V \mid X_s = 1)$.

Definition 27 (Selected-Marginal Observational Model)**.** *Let $\mathcal{D}$ be an S-pDAG with $\texttt{vis}(\mathcal{D}) = V$, $\texttt{lat}(\mathcal{D}) = M$ and $\texttt{sel}(\mathcal{D}) = S$. A distribution $P(X_V | X_S = \vec{0})$ is said to be* realizable *by $\mathcal{D}$ if there exists a joint distribution $P(X_{V \dot\cup M \dot\cup S})$ that is realizable by the DAG $\mathcal{D}$ when all nodes are visible and that recovers $P(X_V | X_S = \vec{0})$ by marginalizing out X_M and conditioning on $X_S = \vec{0}$.*

We now extend the notion of observational equivalence and dominance for S-pDAGs.

[2] To translate a "stochastic variable S-pDAG" into our notation, just add a marginalized parent to each visible node. To translate back, simply remove every marginalized vertex that is parentless and has a single (visible) child.

Definition 28 (Observational Equivalence and Dominance)**.** *Let $\mathcal{D}$ and $\mathcal{D}'$ be two S-pDAGs with $\texttt{vis}(\mathcal{D}) = \texttt{vis}(\mathcal{D}') = V$, $\texttt{lat}(\mathcal{D}) = M$, $\texttt{lat}(\mathcal{D}') = M'$, $\texttt{sel}(\mathcal{D}) = S$ and $\texttt{sel}(\mathcal{D}') = S'$. We say that $\mathcal{D}$ and $\mathcal{D}'$ are* observationally equivalent *if their SMO models are identical, i.e. if $\mathcal{M}(\mathcal{D}, V \mid S) = \mathcal{M}(\mathcal{D}', V \mid S')$. More generally, $\mathcal{D}'$* observationally dominates *$\mathcal{D}$ if $\mathcal{M}(\mathcal{D}, V \mid S) \subseteq \mathcal{M}(\mathcal{D}', V \mid S')$.*

4.1.1 Observe&Do in the presence of selection

In this section, we extend the definitions pertaining the Observe&Do probing scheme studied in Chapter 2 for the case with selection effects. We will also extend the definition to allow for soft interventions—ones that set each variable to a distribution over values, rather than to just one value. We do that because we want the strongest possible probing scheme.[3] This implies that when we present conditions under which two S-pDAGs are indistinguishable even under access to the Observe&Do probing scheme with soft interventions (Section 4.2), such conditions will also imply that the two S-pDAGs in question are indistinguishable under more commonly studied weaker probing schemes.

Now, we will formally define the set of data obtained when the experimenter performs the Observe&Do probing scheme with soft interventions, that is, if they can prepare the intervened variable $X_{v^\sharp}$ in a generic probability distribution $Q(X_{v^\sharp})$. In the interest of clarity, let us first define the selected marginal interventional (SMI) distribution obtained from one specific soft intervention and one specific choice of parameters.

Definition 29 (SMI Distribution)**.** *Let $\mathcal{D}$ be an S-pDAG with $\texttt{vis}(\mathcal{D}) = V$, $\texttt{lat}(\mathcal{D}) = M$ and $\texttt{sel}(\mathcal{D}) = S$. Furthermore, let $Q(X_{V^\sharp})$ be the distribution defining a soft intervention. A distribution $P_{Q(X_{V^\sharp})}(X_{V^\sharp}, X_{V^\flat})$ is said to be a* SMI distribution *of $\mathcal{D}$ if it can be written as*

$$P_{Q(X_{V^\sharp})}(X_{V^\sharp}, X_{V^\flat} \mid X_S = \vec{0}) = \frac{\int_{\mathfrak{X}_M} P_{Q(X_{V^\sharp})}(X_{V^\sharp}, X_{V^\flat}, X_M, X_S = \vec{0})\, dx_M}{\int_{\mathfrak{X}_{V^\sharp, V^\flat, M}} P_{Q(X_{V^\sharp})}(X_{V^\sharp}, X_{V^\flat}, X_M, X_S = \vec{0})\, d(x_{V^\sharp}, x_{V^\flat}, x_M)}$$

$$\text{for} \quad P_{Q(X_{V^\sharp})}(X_{V^\flat}, X_{V^\sharp}, X_M, X_S) = Q(X_{V^\sharp}) \prod_{y \in V^\flat \dot\cup M \dot\cup S} P(X_y \mid X_{\mathrm{pa}^\sharp(y)}), \tag{4.1}$$

[3] In the case without selection, the Observe&Do probing scheme with hard interventions is just as powerful as the Observe&Do probing scheme with soft interventions. However, as we will see below Definition 30, in the case with selection the soft interventions give us more power: they can help us detect the presence of selection effects even when there is only one visible variable.

where $\vec{0}$ *is the zero vector,* $\mathrm{pa}^\sharp(y) := \{z^\sharp \mid z \in \mathrm{pa}(y) \cap V\} \cup (\mathrm{pa}(y) \setminus V)$ *and* $\mathfrak{X}_A$ *denotes the set of all joint probability distributions of the variables in a set* A*. The SMI distribution* $P_{Q(X_{V^\sharp})}(X_{V^\sharp}, X_{V^\flat} \mid X_S = \vec{0})$ *represents data obtained from the Observe&Do probing scheme when the interventions force the visible variables to take the distribution* $Q(X_{V^\sharp})$*.*

Sometimes in this chapter, the parameter $P(X_y|X_{\mathtt{pa}^\sharp(y)})$ will be referred to as the *kernel* associated with node y.

Note that flat variables are directly influenced by their sharp visible parents and ordinary non-visible parents, whereas the sharp variables are set by the intervened distribution $Q(X_{V^\sharp})$.

By considering all possible choices of parameters $P(X_y|X_{\mathtt{pa}^\sharp(y)})$ and all possible soft interventions, we define the totality of Observe&Do data that can come from one S-pDAG:

Definition 30 (SMI Model)**.** *For an S-pDAG* $\mathcal{D}$ *with* $\mathtt{vis}(\mathcal{D}) = V$*,* $\mathtt{lat}(\mathcal{D}) = M$ *and* $\mathtt{sel}(\mathcal{D}) = S$*, consider the set:*

$$P_* = \left\{\left(Q(X_{V^\sharp}), P_{Q(X_{V^\sharp})}(X_{V^\sharp}, X_{V^\flat} \mid X_S = \vec{0})\right) : Q(X_{V^\sharp}) \in \mathcal{P}(\mathfrak{X}_V)\right\}. \tag{4.2}$$

We say that P_* *is a* set of selected-marginal interventional (SMI) pairs compatible with $\mathcal{D}$ *if*

- *it contains one element associated with each distribution* $Q(X_{V^\sharp}) \in \mathcal{P}(\mathfrak{X}_V)$*,*
- *each distribution* $P_{Q(X_{V^\sharp})}(X_{V^\sharp}, X_{V^\flat} \mid X_S = \vec{0})$ *is constructed according to Equation (4.1), using a choice of kernels* $\{P(X_w \mid X_{\mathtt{pa}_\mathcal{D}(w)}), w \in V \dot\cup M \dot\cup S\}$ *that is fixed across the elements of* P_**, and where the kernels associated with visible variables are deterministic,*
- *the choice of deterministic kernels* $\{P(X_w \mid X_{\mathtt{pa}_\mathcal{D}(w)}), w \in V \dot\cup M \dot\cup S\}$ *associated with* P_* *is such that, for all* $Q(X_{V^\sharp}) \in \mathcal{P}(\mathfrak{X}_V)$*,* $P_{Q(X_{V^\sharp})}(X_S = \vec{0}) > 0$*.*

The selected-marginal interventional (SMI) model *of the S-pDAG* $\mathcal{D}$*, denoted* $\mathcal{C}(\mathcal{D}, V \mid S)$*, is the collection of all sets* P_* *of SMI pairs compatible with* $\mathcal{D}$*.*

To recap, the SMI model of $\mathcal{D}$ is a collection of sets of pairs, one set P_* for each specific choice of parameters $P(X_w|X_{\mathtt{pa}(w)})$ for the S-pDAG. Each element of P_* contains a soft

intervention $Q(X_{V^\sharp})$ paired with its SMI distribution $P_{Q(X_{V^\sharp})}(X_{V^\sharp,V^\flat} \mid X_S = \vec{0})$ — P_* has one pair for each possible soft intervention. Overall, the SMI model describes the visible consequences of all soft interventions given all choices of parameters.

We now highlight four aspects of the SMI model definition.

Firstly, why does P_* contain pairs of distributions $\big(Q(X_{V^\sharp}), P_{Q(X_{V^\sharp})}(X_{V^\sharp} \mid X_S = \vec{0})\big)$ rather than only distributions $P_{Q(X_{V^\sharp})}(X_{V^\sharp} \mid X_S = \vec{0})$? Given a soft intervention, the distribution over the intervened variables can change with selection: in general $P(X_{V^\sharp} \mid X_S = 0) \neq Q(X_{V^\sharp})$. In our formalism, both distributions are made available, which in practical terms means we are assuming that the experimenter can remember what experiment they have performed. For example, in a graph $v \to s$ where $X_S = X_V$, if the experimenter performs the intervention that sets $Q(x_{v^\sharp})$ to a uniform distribution over two possible values, they will obtain the pair $\big(Q(X_{v^\sharp} = 0) = Q(X_{v^\sharp} = 1) = \frac{1}{2}, P(X_{v^\sharp} = 0 \mid X_s = 0) = 1\big)$. The fact that $P(X_{V^\sharp} \mid X_S = 0) \neq Q(X_{V^\sharp})$ shows us that, even with only one visible variable, it is already possible to detect the presence of selection effects.

Secondly, notice that hard interventions—those that fix X_V to a single value—are included in P_* as the case where $Q(X_{V^\sharp})$ is a point distribution; that is, $Q(X_{V^\sharp})$ is an indicator function $\delta(X_V = x_V)$ that is equal to 1 if $X_V = x_V$ is true, and 0 otherwise.

Thirdly, the effect of interventions to arbitrary subsets of nodes can always be recovered from P_*. In a graph with nodes V, if we want know the effect of intervening $\mathtt{do}(X_A = x_A)$ where $A \subsetneq V$ then we can choose $Q_{V^\sharp}$ to intervene $\mathtt{do}(X_A = x_A)$ and assign a uniform random distribution to $X_{V \setminus A}$, and then condition on $X_{Z^\sharp}$ equalling its natural value. That is, the interventional distribution $P(X_Z = x_Z \mid \mathtt{do}(X_A = x_A))$ is equivalent to $P_{Q_{V^\sharp}}(X_{Z^\flat} = x_Z \mid X_{A^\sharp} = x_A, X_{Z^\sharp} = x_Z)$. This is the same as the discussion we made in Section 2.3.

Fourthly, our definition uses selection, so in some models, we will obtain a selected intervened distribution $P_{Q(X_{V^\sharp})}(X_{V^\flat}, X_{V^\sharp} \mid X_s = 0)$ that is undefined. Conceptually, this represents a situation where the experimenter performs an intervention, and then gets back no data. For instance, perhaps one wants to learn the relationship between having a bank account and employment status but only those with a bank account possess a phone line. So, an experimenter forces some subjects to go without a bank account, and attempts a phone survey about employment status, thus obtaining no data. This scenario makes intuitive sense, so from a formal perspective, whenever an intervened distribution $P_{Q(X_{V^\sharp})}(X_{V^\flat}, X_{V^\sharp} \mid X_s = 0)$ is undefined then we omit it from P_*.

Similar to observational equivalence, we can define interventional equivalence under access to the Observe&Do probing scheme as follows.

Definition 31 (Observe&Do Equivalence and Dominance)**.** *Let $\mathcal{D}$ and $\mathcal{D}'$ be two S-pDAGs with $\texttt{vis}(\mathcal{D}) = \texttt{vis}(\mathcal{D}') = V$, $\texttt{lat}(\mathcal{D}) = M$, $\texttt{lat}(\mathcal{D}') = M'$, $\texttt{sel}(\mathcal{D}) = S$ and $\texttt{sel}(\mathcal{D}') = S'$. We say that $\mathcal{D}$ and $\mathcal{D}'$ are* Observe&Do equivalent *if their SMI models are identical, i.e. if $\mathcal{C}(\mathcal{D}, V \mid S) = \mathcal{C}(\mathcal{D}', V \mid S')$, and that $\mathcal{D}'$* Observe&Do dominates *$\mathcal{D}$ if $\mathcal{C}(\mathcal{D}, V \mid S) \subseteq \mathcal{C}(\mathcal{D}', V \mid S')$.*

Clearly, any two S-pDAGs that are Observe&Do equivalent are also observationally equivalent. However, the converse is not true.

4.2 Transformations that maintain Observe&Do equivalence in the presence of selection

In this section we will present a series of transformations that, when applied on an S-pDAG $\mathcal{D}$, yield another S-pDAG that is Observe&Do equivalent to $\mathcal{D}$. Lemmas 1 and 2 give special cases of these transformations.

The first transformation we discuss is *exogenization*, that was already discussed in Lemma 1. The second is called *terminalization*, because it makes all selected nodes *terminal* (i.e. childless). The following proposition defines these operations and establishes that they do not change the SMI model.

Proposition 15. *Let $\mathcal{D}$ be an S-pDAG with $\texttt{vis}(\mathcal{D}) = V$, $\texttt{lat}(\mathcal{D}) = M$ and $\texttt{sel}(\mathcal{D}) = S$ where ...*

a) $m \in M$. Let $\mathfrak{r}_m(\mathcal{D})$ be the S-pDAG obtained from $\mathcal{D}$ by exogenizing *m, that is, by i) adding an edge $l \to k$ from every $l \in \texttt{pa}_{\mathcal{D}}(m)$ to every $k \in \texttt{ch}_{\mathcal{D}}(m)$, and ii) deleting the edges $l \to m$ for $l \in \texttt{pa}_{\mathcal{D}}(m)$. Then, $\mathcal{D}$ and $\mathfrak{r}_m(\mathcal{D})$ are Observe&Do equivalent; that is, $\mathcal{C}(\mathcal{D}, V \mid S) = \mathcal{C}(\mathfrak{r}_m(\mathcal{D}), V \mid S)$.*

b) $s \in S$. Let $\mathcal{D}_{\underline{s}}$ be the graph obtained from $\mathcal{D}$ by removing edges outgoing from s. Then, $\mathcal{D}$ and $\mathcal{D}_{\underline{s}}$ are Observe&Do equivalent; that is, $\mathcal{C}(\mathcal{D}, V \mid S) = \mathcal{C}(\mathcal{D}_{\underline{s}}, V \mid S)$.

Note that the order in which these operations are applied does not matter. Therefore, we can define a transformation, denoted by `Exog`*, that acts on an S-pDAG by sequentially exogenizing all of the marginalized nodes of an S-pDAG. Similarly, we can define a transformation, denoted by* `Term`*, that acts on an S-pDAG by terminalizing all of the selected nodes of an S-pDAG.*

The proof of Proposition 15 is in Section C.1.1. Exogenization is already known to preserve Observe&Do equivalence when there are no selected variables, as described in Chapter 2. We merely prove that after subsequent selection, the models remain equivalent. For selection, the proof idea is that after edges from s to its children are removed, those children can behave just as they would when $X_s = 0$, and then we will recover the same selected distribution.

Once `Exog` and `Term` have been performed, we will obtain an S-pDAG $\texttt{Exog} \circ \texttt{Term}(\mathcal{D})$, wherein the marginalized nodes are parentless and the selected nodes are childless. It follows that the marginalized and selected nodes in the resulting pDAG are not neighbours of any nodes of their own type. It is still possible, however, to have marginalized nodes that neighbour selected nodes.

Now, we proceed to the next two operations that preserve Observe&Do equivalence. They consist of merging marginalized nodes that share a common selected child, and merging selected nodes that share a common marginalized parent.

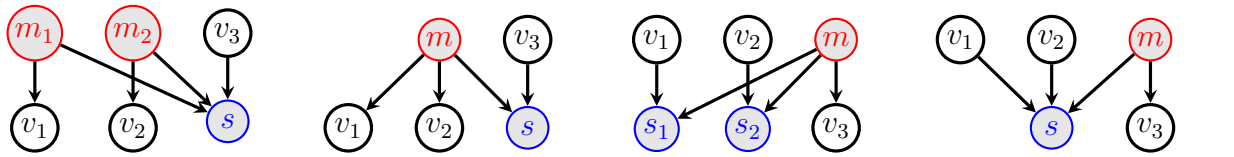
Proposition 16. *Let $\mathcal{D}$ be an S-pDAG with $\texttt{vis}(\mathcal{D}) = V$, $\texttt{lat}(\mathcal{D}) = M$ and $\texttt{sel}(\mathcal{D}) = S$ where $\mathcal{D} = \texttt{Exog} \circ \texttt{Term}(\mathcal{D})$ and...*

a) Let $m_1, m_2 \in M$, and $\texttt{ch}_{\mathcal{D}}(m_1) \cap \texttt{ch}_{\mathcal{D}}(m_2) \cap S \neq \varnothing$. Obtain the modified S-pDAG $\tilde{\mathcal{D}}$ by merging m_1 and m_2 into one node m with $\texttt{ch}_{\tilde{\mathcal{D}}}(m) = \texttt{ch}_{\mathcal{D}}(m_1) \cup \texttt{ch}_{\mathcal{D}}(m_2)$. Then, $\mathcal{D}$ and $\tilde{\mathcal{D}}$ are Observe&Do equivalent; that is, $\mathcal{C}(\mathcal{D}, V \mid S) = \mathcal{C}(\tilde{\mathcal{D}}, V \mid S)$.

b) Let $s_1, s_2 \in S$, and $\texttt{pa}_{\mathcal{D}}(s_1) \cap \texttt{pa}_{\mathcal{D}}(s_2) \cap M \neq \varnothing$. Obtain the modified S-pDAG $\tilde{\mathcal{D}}$ by merging s_1 and s_2 into one node s with $\texttt{pa}_{\tilde{\mathcal{D}}}(s) = \texttt{pa}_{\mathcal{D}}(s_1) \cup \texttt{pa}_{\mathcal{D}}(s_2)$. Then, $\mathcal{D}$ and $\tilde{\mathcal{D}}$ are Observe&Do equivalent; that is, $\mathcal{C}(\mathcal{D}, V \mid S) = \mathcal{C}(\tilde{\mathcal{D}}, V \mid (\{s\} \cup S) \setminus \{s_1, s_2\})$.

The order in which these operations are applied does not matter. Therefore, we can define a transformation, denoted by `MergeM`*, that acts on an S-DAG by merging all of the marginalized nodes that can be merged by a). Similarly, we can define a transformation, denoted by* `MergeS`*, that acts on an S-pDAG by merging all of the selected nodes that can be merged by b).*

For example, this implies that since the marginalized nodes m_1 and m_2 in Figure 4.2.1a share the selected child s, the S-pDAG in Figure 4.2.1b where the marginalized variables are merged is Observe&Do equivalent to the original. Since nodes s_1 and s_2 in Figure 4.2.1c share the parent m, when we merge them, the resulting graph Figure 4.2.1d, then it is also equivalent to its original.



(a) An S-pDAG $\mathcal{D}_1$. (b) The S-pDAG $\texttt{MergeM}(\mathcal{D}_1)$. (c) An S-pDAG $\mathcal{D}_2$. (d) The S-pDAG $\texttt{MergeS}(\mathcal{D}_2)$.

Figure 4.2.1: Example of the application of Proposition 16.

The proof of Proposition 16 is in Section C.1.2. For Proposition 16a), to prove that $\tilde{\mathcal{D}}$ can realize any data that is generated by $\mathcal{D}$, our approach is to let X_m equal the Cartesian product of X_{m_1} and X_{m_2}. Conversely, to prove that $\mathcal{D}$ can realize any data that is realizable by $\tilde{\mathcal{D}}$, we let both variables X_{m_1} and X_{m_2} of $\mathcal{D}$ sample over the whole domain of X_m in $\tilde{\mathcal{D}}$, and take their shared selected child to be 0 precisely when $X_{m_1} = X_{m_2}$. For Proposition 16b), to prove that $\tilde{\mathcal{D}}$ can realize any data that is generated by $\mathcal{D}$, we will let X_s equal Cartesian product of X_{s_1} and X_{s_2}. Conversely, to show that $\mathcal{D}$ can realize any data generated by $\tilde{\mathcal{D}}$, we will let X_m (the shared parent of s_1, s_2) be the Cartesian product of two variables, one sampled uniformly over the domain of $X_{\texttt{pa}_{\mathcal{D}}(s_1)}$ and the other sampled uniformly over the domain of $X_{\texttt{pa}_{\mathcal{D}}(s_2)}$. Furthermore, we take s_1 and s_2 to equal 0 precisely when the two variables that compose X_m equal $X_{\texttt{pa}_{\mathcal{D}}(s_1)}$ and $X_{\texttt{pa}_{\mathcal{D}}(s_2)}$, respectively. Then, under selection, X_{s_1} and X_{s_2} will behave just as if they had direct access to $\texttt{pa}_{\mathcal{D}}(s_1) \cup \texttt{pa}_{\mathcal{D}}(s_2)$.

After applying Propositions 15 and 16, we obtain the S-pDAG $\texttt{MergeS} \circ \texttt{MergeM} \circ \texttt{Exog} \circ \texttt{Term}(\mathcal{D})$ where each marginalized or selected variable has at most one non-visible neighbour, which must be of the opposite type. However, each non-visible variable may still have multiple visible neighbours.

The next operation that preserves Observe&Do equivalence deals with arrows $m \to s$ where the marginalized node m has multiple visible children and/or the selected node s has multiple visible parents, such as in Figure 4.2.2a. This operation transforms the S-pDAG into one where every marginalized node either has only visible children or it has exactly one selected child and one visible child, and every selected node either has only visible parents or it has exactly one marginalized parent and one visible parent, such as in Figure 4.2.2b.

Proposition 17. *Let $\mathcal{D}$ be an S-pDAG with $\texttt{vis}(\mathcal{D}) = V$, $\texttt{lat}(\mathcal{D}) = M$ and $\texttt{sel}(\mathcal{D}) = S$ such that $\mathcal{D} = \texttt{MergeS}\circ\texttt{MergeM}\circ\texttt{Exog}\circ\texttt{Term}(\mathcal{D})$, and $\mathcal{D}$ contains the edge $m \to s$, where $m \in M$ and $s \in S$. Let V_s denote all visible nodes which are parents of s, i.e. $V_s \coloneqq V \cap \texttt{pa}_{\mathcal{D}}(s)$, and similarly let V_m denote all visible nodes which are children of m, i.e. $V_m \coloneqq V \cap \texttt{ch}_{\mathcal{D}}(m)$. If either V_s or V_m are non-singleton sets (that is, if m has more than one visible child or*

s has more than one visible parent), then we perform the following operations to create a new S-pDAG $\tilde{\mathcal{D}}$*:*

1. *For every pairing of visible nodes* $\{a, b\}$ *with* $a \in V_s$ *and* $b \in V_m$ *(i.e. for every path* $a \to s \leftarrow m \to b$*) add the new selected node* s_{ab} *and the new marginalized node* m_{ab} *along with the three new edges* $a \to s_{ab} \leftarrow m_{ab} \to b$*.*

   Note that we consider a *to be part of a valid pairing if and only if* $a \in V_m \cup V_s$*. Consequently, if there is a path* $a \to s \leftarrow m \to a$*,then we must still add the path* $a \to s_{aa} \leftarrow m_{aa} \to a$*.*

2. *Delete the edge* $m \to s$*.*

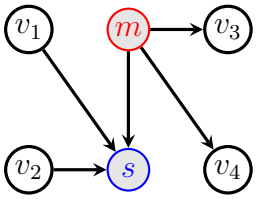
Then, $\mathcal{D}$ *and* $\tilde{\mathcal{D}}$ *are Observe&Do equivalent; that is,* $\mathcal{C}(\mathcal{D}, V \mid S) = \mathcal{C}(\tilde{\mathcal{D}}, V \mid S \cup S')$ *where* S' *denotes the set of added selected nodes.*

The order of application of these operations on different edges $m \to s$ *does not matter. Therefore, we can define a transformation, denoted by* `SplitM->S`*, that acts on an S-pDAG by applying the steps above to each edge* $m \to s$ *until every edge* $m \to s$ *satisfies that* m *has only one visible child and* s *has only one visible parent.*

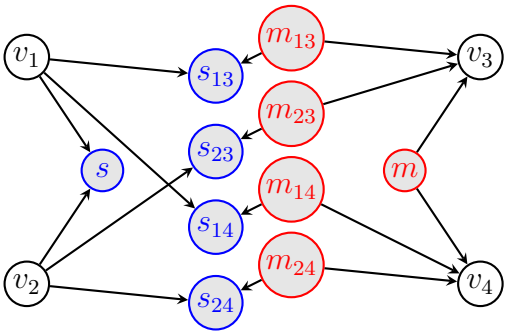
The proof of Proposition 17 is presented in Section C.1.3. The intuition is that all that the paths in Figure 4.2.2a that go through $m \to s$ do is to allow children of m to depend arbitrarily on parents of s. So, we need a way to assign the children of m in Figure 4.2.2b a distribution $P(X_{v_3}, X_{v_4} \mid X_{v_1}, X_{v_2}, X_S = \vec{0})$ that matches the same distribution in Figure 4.2.2a. To achieve this, we will sample each $X_{m_{ij}}$ uniformly over the domain of X_{v_i} and let each $X_{s_{ij}}$ equal zero precisely when $X_{m_{ij}} = X_{v_i}$. Then, under selection, each child variable v_j essentially has access to the assignment of each parent v_i. We can then use X_m to ensure that X_{v_3} and X_{v_4} have the right distribution given any joint assignment x_{v_1}, x_{v_2}. Conversely, we can reproduce the behaviour of m_{13}, m_{14} in $\mathcal{D}$ by having the variable X_m be a Cartesian product of $X_{m_{13}}$ and $X_{m_{14}}$.

Figure 4.2.3 provides a second example of the application of Proposition 17, where the nodes v_1 and v_2 are simultaneously parents of s and children of m.

In the resulting S-pDAG $\texttt{SplitM->S} \circ \texttt{MergeS} \circ \texttt{MergeM} \circ \texttt{Exog} \circ \texttt{Term}(\mathcal{D})$, whenever two non-visible variables are neighbours, each has precisely one visible neighbour. Essentially, our S-pDAG now consists of: 1) marginalized nodes with exclusively visible children, 2) selected nodes with exclusively visible parents, 3) directed edges $a \to b$ between visible nodes, and 4) paths $a \to s_{ab} \leftarrow m_{ab} \to b$ between visible nodes a, b, which will be called *special edges* between a and b.

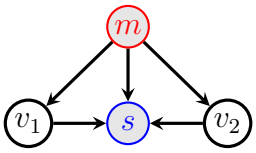


(a) An S-pDAG $\mathcal{D}$.

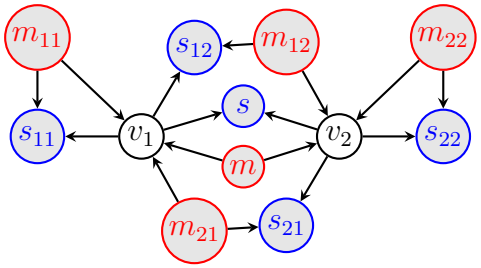


(b) The S-pDAG `SplitM->S`$(\mathcal{D})$.

Figure 4.2.2: First example of the application of Proposition 17.



(a) An S-pDAG $\mathcal{D}$.



(b) The S-pDAG `SplitM->S`$(\mathcal{D})$, which visually resembles the Pokémon Magnemite.

Figure 4.2.3: Second example of the application of Proposition 17.

After selection, both directed edges and special edges allow X_b to vary in response to an intervention to X_a. This is the idea behind the next operation that preserves Observe&Do equivalence: replacing directed edges $a \to b$ by special edges $a \to s_{ab} \leftarrow m_{ab} \to b$ whenever a has a selected child and b has a marginalized parent.

Proposition 18. *Let $\mathcal{D}$ be an S-pDAG with* `vis`$(\mathcal{D}) = V$, `lat`$(\mathcal{D}) = M$ *and* `sel`$(\mathcal{D}) = S$ *such that* $\mathcal{D} =$ `SplitM->S`$\circ$`MergeS`$\circ$`MergeM`$\circ$`Exog`$\circ$`Term`$(\mathcal{D})$, *and let $a \to b$ be an edge of $\mathcal{D}$ where the node a has at least one selected child and the node b has at least one marginalized parent. Let the S-pDAG $\mathcal{D}'$ be constructed by replacing $a \to b$ with a special edge; that is, with a path $a \to s_{ab} \leftarrow m_{ab} \to b$ where s_{ab} is a selected node and m_{ab} is a marginalized node. Then, $\mathcal{D}$ and $\mathcal{D}'$ are Observe&Do equivalent; that is, $\mathcal{C}(\mathcal{D}, V \mid S) = \mathcal{C}(\mathcal{D}', V \mid S \cup \{s_{ab}\})$.*

The order in which the directed edges are transformed into special edges does not matter. Therefore, we can define a transformation, denoted by `ToSpecial`, *that acts on an S-pDAG by replacing every directed edge between visible nodes with a special edge, whenever the*

parent has a selected child and the child has a marginalized parent.

The proof of Proposition 18 is presented in Section C.1.4. The intuition is that to produce a distribution from $\mathcal{D}$ on $\mathcal{D}'$, we will set $X_{s_{ab}}$ to equal 0 if and only if X_a equals $X_{m_{ab}}$, and set X_b to depend on $X_{m_{ab}}$ in $\mathcal{D}'$ in the same way that it depends on X_a in $\mathcal{D}$. Conversely, to reproduce a distribution from $\mathcal{D}'$ on $\mathcal{D}$, we can obtain $P'(X_b \mid X_{s_{ab}} = 0, X_{\mathtt{pa}_{\mathcal{D}}(b)})$ from $\mathcal{D}'$ and use this as the kernel for X_b in $\mathcal{D}$.

At first sight it may be unclear why we choose to replace the simple directed edges $a \to b$ with the more complicated-looking special edges. We do that because in many cases the inverse transformation, that is replacing special edges with regular directed edges, would induce a cycle. Therefore, since our aim is to conflate these two indistinguishable patterns to obtain a data structure that captures all Observe&Do equivalences, we must use Proposition 18 to transform directed edges into special edges wherever possible.

Our final two operations that preserve Observe&Do equivalence consist in removing *redundant* marginalized and selected nodes, which are marginalized nodes whose children are a subset of the children of another marginalized node or selected nodes whose parents are a subset of the parents of another selected node. An example is shown in Figure 4.2.4. The removal of redundant marginalized nodes in the case without selection effects was already proven to preserve observational equivalence [33], and Observe&Do equivalence (Chapter 2).

Proposition 19. *Let $\mathcal{D}$ be an S-pDAG with $\mathtt{vis}(\mathcal{D}) = V$, $\mathtt{lat}(\mathcal{D}) = M$ and $\mathtt{sel}(\mathcal{D}) = S$ such that $\mathcal{D} = \mathtt{ToSpecial} \circ \mathtt{SplitM->S} \circ \mathtt{MergeS} \circ \mathtt{MergeM} \circ \mathtt{Exog} \circ \mathtt{Term}(\mathcal{D})$, and...*

a) There are $m_1, m_2 \in M$ such that $\mathtt{ch}(m_1) \subseteq \mathtt{ch}(m_2)$. Let $\mathcal{D}_{-m_1}$ be the subgraph of $\mathcal{D}$ on nodes other than m_1. Then, $\mathcal{D}$ and $\mathcal{D}_{-m_1}$ are Observe&Do equivalent, so $\mathcal{C}(\mathcal{D}, V \mid S) = \mathcal{C}(\mathcal{D}_{-m_1}, V \mid S)$.

b) There are $s_1, s_2 \in S$ such that $\mathtt{pa}_{\mathcal{D}}(s_1) \subseteq \mathtt{pa}(s_2)$. Let $\mathcal{D}_{-s_1}$ be the subgraph of $\mathcal{D}$ on nodes other than s_1. Then, $\mathcal{D}$ and $\mathcal{D}_{-s_1}$ are Observe&Do equivalent, so $\mathcal{C}(\mathcal{D}, V \mid S) = \mathcal{C}(\mathcal{D}_{-s_1}, V \mid S \setminus \{s_1\})$.

The order in which redundant marginalized and selected nodes are removed does not matter. Therefore, we can define a transformation, denoted by `RmvRedM`*, that acts on an S-pDAG by applying a) until no redundant marginalized nodes remain. Similarly, we can define a transformation, denoted by* `RmvRedS`*, that acts on an S-pDAG by applying b) until no redundant selected nodes remain.*

The proof of Proposition 19 is presented in Section C.1.5. Part (a) follows from the fact that $\mathcal{D}$ and $\mathcal{D}_{-m_1}$ have the same Observe&Do model, by Theorem 1. For part (b), the idea is that we can take the variable X_{s_2} in $\mathcal{D}_{-s_1}$ to equal zero when $X_{s_1} = 0$ and $X_{s_2} = 0$ in $\mathcal{D}$.



(a) An S-pDAG $\mathcal{D}$.

(b) The S-pDAG $\texttt{RmvRedS} \circ \texttt{RmvRedM}(\mathcal{D})$.

Figure 4.2.4: Example of the application of Proposition 19.

After applying all of the operations of Propositions 15 to 19 on an initial S-pDAG, we obtain a resulting S-pDAG $\mathcal{D}$ that is such that

$$\mathcal{D} = \texttt{RmvRedS} \circ \texttt{RmvRedM} \circ \texttt{ToSpecial} \circ \texttt{SplitM->S} \circ \texttt{MergeS} \circ \texttt{MergeM} \circ \texttt{Exog} \circ \texttt{Term}(\mathcal{D}).$$

When this condition is satisfied, we will call $\mathcal{D}$ a *canonical S-pDAG*. In a canonical S-pDAG, the marginalized (or selected) variables associated with nodes that only have visible neighbours will be called the *marginalized (or selected) variables of the faces*; the reason for this will be evident from Definition 34.

The canonical S-pDAG is the analogue of the construction `RE-reduce` from Eq. (2.6) for the case where there are selected nodes.

4.3 Selected-marginalized directed graphs (smDGs)

In this section, we will formulate a data structure that already encodes all of the regularities of the canonical S-pDAG; this is a generalization of the mDAG [33] for the case where there are selection effects. The main idea is to view the canonical S-pDAG as a set of visible nodes connected by three kinds of relationships: latent confounding, selection effects, and paths of the form $a \to b$ or $a \to s' \leftarrow m' \to b$. To represent latent confounding and selection effects, our new data structure will use two *independence systems*:

Definition 32 (Independence System)**.** *An* independence system $\mathcal{B}$ *over a finite set* V *is a collection of subsets of* V *that contains the empty set* $\varnothing \in \mathcal{B}$*, and such that the system is closed under taking subsets: if* $B \in \mathcal{B}$ *and* $A \subseteq B$ *then* $A \in \mathcal{B}$*. We call* $A \in \mathcal{B}$ *a* maximal face *of* $\mathcal{B}$ *if it is not a strict subset of any other set in* $\mathcal{B}$*.*

We will use an independence system $\mathcal{L}$ to represent latent confounding, and another independence system $\mathcal{S}$ to represent selection effects. Each subset $V_m \in \mathcal{L}$ corresponds to the visible children of a marginalized node m in the canonical S-pDAG. An independence system neatly captures the fact that `RmvRedM` has removed all marginalized variables whose children are a subset of another marginalized variable's children. The original definition of mDAGs, that we presented in Section 2.1.3, uses a *simplicial complex*, which is an independence system that contains every singleton set, i.e. such that $\{v\} \in \mathcal{B}$ for all $v \in V$. Our reason for relaxing this assumption is that $\{v\} \notin \mathcal{L}$ represents the case where X_v must be a deterministic function of its visible parents.

Each subset $V_s \in \mathcal{S}$ corresponds to the visible parents of one selected node s of the canonical S-pDAG. Similar to the case of latent confounding, nested selected variables have been removed by `RmvRedS`, making an independence system a natural representation of this structure. The presence or absence of a node $\{v\} \in \mathcal{S}$ in the selected independence system is meaningful, as it can be detected with soft interventions.

Recall that Proposition 18 has shown that whenever a has selected children and b has marginalized parents, a path $a \to b$ can be replaced by a path $a \to s \leftarrow m \to b$ while preserving Observe&Do equivalence. In our new data structure, both types of paths $a \to b$ and $a \to s \leftarrow m \to b$ will be represented by a directed edge from a to b, and the presence or absence of a selected face that includes a and a marginalized face that includes b is what will tell us what is the type of path that appears in the corresponding canonical S-pDAG. With this, note that this directed graph in the new data structure need not be acyclic. Our new data structure is defined as:

Definition 33 (smDG)**.** *A* selected-marginalized directed graph *(smDG)* $\mathcal{G}$ *is a quadruple* $(V, \mathcal{E}, \mathcal{L}, \mathcal{S})$ *where* $(V, \mathcal{E})$ *is a directed graph called the* directed structure*,* $\mathcal{L}$ *is the* marginalized independence system*, and* $\mathcal{S}$ *is the* selected independence system*. (The directed structure is allowed to contain cycles and even self-loops* $v \to v$*.)*

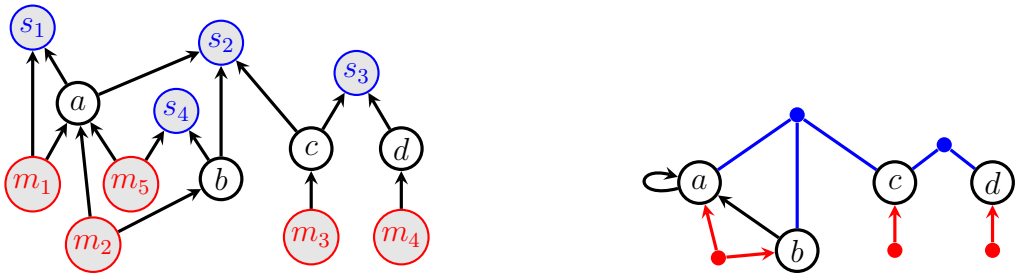
It will be useful to fix some notation for an smDG $\mathcal{G} = (V, \mathcal{E}, \mathcal{L}, \mathcal{S})$. The parents of a node b, denoted $\mathtt{pa}_{\mathcal{G}}(b)$, are the set of nodes such that $a \to b$ is in $\mathcal{E}$. This includes both the cases where the canonical graph contains the path $a \to b$ or a path $a \to s \leftarrow m \to b$. The subgraph $\mathcal{G}_{V'} = (V', \mathcal{E}', \mathcal{L}', \mathcal{S}')$ is the smDG with sets of edges $\mathcal{E}'$, marginalized faces $\mathcal{L}'$, and selected faces $\mathcal{S}'$ in $\mathcal{G}$ that contain only those nodes V'.

When we draw smDGs in this chapter, the graph $(V, \mathcal{E})$ will have ordinary black edges, the maximal faces of $\mathcal{L}$ will be represented by red hyperedges with arrowheads, and the maximal faces of $\mathcal{S}$ will be represented by blue hyperedges without arrowheads. When a maximal marginal (or selected) face contains only one node v, we will illustrate it using a red dot incoming to v (or a blue dot with an undirected edge to v), as shown in Figure 4.3.1b. If the directed structure of an smDG is acyclic, all of the nodes are in some face of the marginalized independence system,[4] and the selected independence system contains only the empty set, then the smDG is simply an mDAG [33].

We connect a canonical S-pDAG to an smDG via the operation of *selected-latent projection*, named in analogy to the latent projection operation [80, 33]. This definition will use the notion of a closure $\mathrm{cl}(A)$ of a set of nodes A, defined as the set of all subsets of A.

Definition 34 (Selected-latent projection)**.** *Let $\mathcal{D}^{can}$ be a canonical S-pDAG with $\mathtt{vis}(\mathcal{D}) = V$, $\mathtt{lat}(\mathcal{D}) = M$ and $\mathtt{sel}(\mathcal{D}) = S$. Its selected-latent projection (SLP), denoted $\mathfrak{p}(\mathcal{D}^{can}, V \mid S)$, is the smDG $(V, \mathcal{E}, \mathcal{L}, \mathcal{S})$ where the directed structure $(V, \mathcal{E})$ is the induced subgraph $\mathcal{D}^{can}_V$ (according to Definition 2 of induced subgraph) with an extra added edge $a \to b$ for each path $a \to s \leftarrow m \to b$ where $a, b \in V, m \in M, s \in S$, the marginalized independence system $\mathcal{L}$ is the union $\bigcup_{m \in M} \mathrm{cl}(\mathtt{ch}_{\mathcal{D}^{can}}(m) \cap V)$, and the selected independence system $\mathcal{S}$ is the union $\bigcup_{s \in S} \mathrm{cl}(\mathtt{pa}_{\mathcal{D}^{can}}(s) \cap V)$.*

For a non-canonical S-pDAG $\mathcal{D}$ with $\mathtt{vis}(\mathcal{D}) = V$, $\mathtt{lat}(\mathcal{D}) = M$ and $\mathtt{sel}(\mathcal{D}) = S$, we define its selected-latent projection $\mathfrak{p}(\mathcal{D}, V \mid S)$ as the SLP of $\mathcal{D}^{can}$, where $\mathcal{D}^{can}$ is the canonical S-pDAG obtained from $\mathcal{D}$ via the rules described in Section 4.2.



(a) A canonical S-pDAG

(b) The selected latent projection of (a)

Figure 4.3.1: Under marginalization of m_1, m_2, m_3, m_4, m_5 and selection of s_1, s_2, s_3, s_4, we obtain the selected-latent projection shown in (b).

[4] This requirement is included because, as already discussed, Evans [33] uses a simplicial complex instead of an independence system to define mDAGs.

For instance, the selected-latent projection of both the non-canonical S-pDAG of Figure 4.0.2a and the canonical S-pDAG of Figure 4.3.1a (that are related by `ToSpecial`) is the smDG of Figure 4.3.1b. It is also possible to define the canonical graph associated with any smDG as follows.

Definition 35 (Canonical graph associated with an smDG)**.** *The* canonical graph *associated with an smDG* $\mathcal{G} = (V, \mathcal{E}, \mathcal{L}, \mathcal{S})$ *is the graph* $\texttt{can}(\mathcal{G})$ *obtained by starting with the graph* $(V, \varnothing)$ *and,*

1. *for each edge* $a \to b \in \mathcal{E}$ *where* a *is an element of some* $V_s \in \mathcal{S}$ *and* b *is an element of some* $V_m \in \mathcal{L}$*, add a marginalized node* m_{ab}*, a selected node* s_{ab} *and a path* $a \to s_{ab} \leftarrow m_{ab} \to b$*;*
2. *for each other edge* $a \to b \in \mathcal{E}$*, add an edge* $a \to b$*;*
3. *for each maximal face* $V_m \in \mathcal{L}$*, add a marginalized node* m *and edges from* m *to each node in* V_m*;*
4. *for each maximal face* V_s *in* $\mathcal{S}$*, add a selected node* s *and edges from each node in* V_s *to* s*.*

In the case where $\texttt{can}(\mathcal{G})$ *is acyclic, we call it the canonical S-pDAG associated with* $\mathcal{G}$*.*

In the case where the canonical graph $\texttt{can}(\mathcal{G})$ is acyclic, the canonical S-pDAG associated with $\mathcal{G}$ can be mapped back to $\mathcal{G}$ via the selected-latent projection. That is, in this case Definitions 34 and 35 are the inverse of each other.

Not every smDG represents a realistic data-generating process however, because the canonical graph is not guaranteed to be acyclic. For example, in the two node smDG $\mathcal{G} : a \to b \to a$, the canonical graph $\texttt{can}(\mathcal{G})$ is equal to $\mathcal{G}$ itself, and thus it is cyclic.

When an smDG does have a canonical S-pDAG, that is, when it can be lifted to an S-pDAG via the process described in Definition 35, we call it *liftable*:

Definition 36 (Liftable smDG)**.** *An smDG* $\mathcal{G}$ *is said to be* liftable *if its canonical graph is acyclic.*

An smDG will *not* be liftable when its directed edges contain a cycle that has no edge that satisfies condition 1 of Definition 35:

Proposition 20. *An smDG* $\mathcal{G} = (V, \mathcal{E}, \mathcal{L}, \mathcal{S})$ *is liftable if and only if for every cycle in* $\mathcal{G}$*, at least one edge* $a \to b$ *in that cycle has* $a \in \bigcup \mathcal{S}$ *and* $b \in \bigcup \mathcal{L}$*.*

Proof. In the canonical graph, marginalized nodes lack parents and selected nodes lack children, so cycles can only arise from edges created at step (1) of Definition 35. Thus a cycle will arise in the canonical graph precisely when there is a cycle in the smDG, and no edge $a \to b$ in that cycle satisfies the conditions for step (1), i.e. $a \notin \bigcup \mathcal{L}$ or $b \notin \bigcup \mathcal{S}$. □

4.4 Interventional inequivalence of smDGs

4.4.1 Distinct smDGs are Observe&Do distinguishable

So far, we have established that S-pDAGs with the same selected-latent projection cannot be distinguished by the Observe&Do probing scheme. We now prove that all S-pDAGs with different selected-latent projections are Observe&Do distinguishable. These results jointly imply that the set of liftable smDGs (as can be identified by Proposition 20) are isomorphic to the equivalence classes of S-pDAGs that share the same SMI model.

That pDAGs with different mDAGs are Observe&Do distinguishable was established in Chapter 2. To generalize this result to smDGs, we will need to consider not just how to distinguish a relationship $a \to b$ with a relationship where a and b are in the same marginalized face, but also how to distinguish each of these from a scenario where a and b are in the same selected face.

Theorem 5 (Same smDG $\iff$ Same SMI model)**.** *The S-pDAGs $\mathcal{D}$ and $\mathcal{D}'$ with $\texttt{vis}(\mathcal{D}) = \texttt{vis}(\mathcal{D}') = V$, $\texttt{lat}(\mathcal{D}) = M$, $\texttt{lat}(\mathcal{D}') = M'$, $\texttt{sel}(\mathcal{D}) = S$ and $\texttt{sel}(\mathcal{D}') = S'$ are Observe&Do equivalent, i.e. $\mathcal{C}(\mathcal{D}, V \mid S) = \mathcal{C}(\mathcal{D}', V \mid S')$, if and only if they have the same selected-latent projection, i.e. $\mathfrak{p}(\mathcal{D}, V \mid S) = \mathfrak{p}(\mathcal{D}', V \mid S')$.*

The proof is detailed in Section C.2 but we summarize our approach here. A pair of SLPs can differ in four ways: (1) different self-loops (i.e. edges from a node to itself), (2) other differences in their directed structures, (3) different marginalized independence systems, (4) different selected independence systems. For case (1), we show that if an SLP contains a self-loop (v, v), i.e. if the canonical S-pDAG contains a path $v \to s \leftarrow m \to v$, and we ignore all of the other variables, say by setting them to constant values, and we simultaneously observe $X_{v^\flat}$ and intervene on $X_{v^\sharp}$, then the value of $X_{v^\flat}$ might depend on the value of $X_{v^\sharp}$.[5] For case (2), we show that if an smDG contains a directed edge $a \to b$, then X_b can statistically depend on X_a when X_a is intervened even while the distribution

[5] This might seem paradoxical at first sight, but here is a scenario where this could occur: suppose that a person's income (m) affects both their likelihood of having a phone (s) and of having a bank account

of X_a is unchanged by an intervention to X_b. For case (3), we show that only with a marginalized face can we see a conditional dependence between a set of variables that disappears whenever all but one of the variables is intervened upon. Finally, for case (4), we show that only with a selected face we can see a conditional dependence between a set of variables that persists whenever all but one of its variables is intervened upon.

Note that the fact that two smDGs are always Observe&Do distinguishable *does not* imply that, starting from a specific set of Observe&Do data (i.e. a set P_* of SMI pairs), one can always identify what was the smDG that generated it. This is due to the same reasoning stressed below Definition 16: the fact that two smDGs are always Observe&Do distinguishable only implies that the SMI models of two different smDGs are *never the same*, and not that they never overlap.

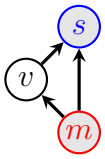
4.4.2 Distinct smDGs are not always All-patterns Observe-Or-Do distinguishable

So far, we have chosen the most powerful probing scheme possible to make sure the smDG structure captures all important differences between causal structures. Under weaker probing schemes, however, some pairs of smDGs may be indistinguishable.

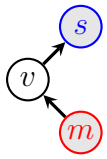
One kind of weaker probing scheme is the All-patterns Observe-Or-Do probing scheme described in Section 2.5, where each visible variable is allowed to be observed or intervened upon, but not both at once. Here, we need a slight extension of the definition of this probing scheme, to take into account selection effects and to allow for soft interventions. Such formal definition is thus provided in Section C.3.

As it turns out, if there is only access to the All-patterns Observe-Or-Do probing scheme, some pairs of smDGs *cannot* be distinguished. Figure 4.4.1 provides an example: the two smDGs in that figure cannot be distinguished even though they have distinct smDGs.

(v). Further imagine that having a bank account is helpful in getting a phone plan, so it increases one's likelihood having a phone. The causal structure of such a scenario is given by $v \leftarrow m \rightarrow s$ and $v \rightarrow s$. An experimenter asks what fraction of the population has a bank account, via a phone survey, so it is subjected to selection effects on s. Suppose that before conducting this survey, the experimentalist opens a bank account for some of the subjects ($v^\sharp$). Some low-income people will now gain access to a phone, and will therefore appear in the survey data. So, if the survey asks "Did you have a bank account *before* the experiment started?" ($v^\flat$), we will see a higher percentage of the answer "no" than what we would have seen without the intervention—something that cannot happen if the self-loop is not present.



(a) An S-pDAG where v has a self-loop



(b) An S-pDAG where $m \to s$ is removed

Figure 4.4.1: Two canonical S-pDAGs whose selected-latent projections differ in the self-loop $v \to v$.

The selected-latent projections of Figure 4.4.1a and Figure 4.4.1b have identical marginal and selected independence systems (which just contain the empty set and the set $\{v\}$). The SLPs differ, however, in their directed structure: in the SLP of Figure 4.4.1a, the directed structure contains the self-loop $v \to v$, which is not present in the SLP of Figure 4.4.1b. In Proposition 31 of Section C.3, we establish that these S-pDAGs *cannot* be distinguished by the All-patterns Observe-or-Do probing scheme. This means that a pair of S-pDAGs is sometimes distinguishable by observing the intervened and non-intervened distributions jointly, even if they cannot be distinguished by observing the same distributions separately.

Note that this is an important difference from the case without selection: as we have seen in Chapter 2, when two causal structures without selected nodes are associated with different mDAGs, they are distinguishable even by the weaker Observe-or-1Do probing scheme, where a variable cannot be simultaneously observed and intervened upon *and* we are restricted to hard interventions that set variables to the value 0.

4.5 Rules to prove observational equivalence of smDGs

Up to this point, in this chapter we have studied how causal structures with selection effects can be distinguished when one has access to interventions. In particular, our main focus was the very powerful Observe&Do probing scheme (in all of the chapter except for Section 4.4.2). In this section, we go to the other extreme, investigating the weakest probing scheme: passive observation of the visible variables. As argued in Chapter 3, this probing scheme is of particular relevance, since many times it is not possible to implement interventions.

While different liftable graphs can always be distinguished by the Observe&Do probing scheme, there are smDGs that are observationally equivalent. As such, it is useful to define

observational equivalence and dominance between liftable smDGs. This is defined via their canonical S-pDAGs:

Definition 37 (Observational Equivalence of smDGs)**.** *Let $\mathcal{G}$ and $\mathcal{G}'$ be two liftable smDGs that have the same set of nodes V. We say that $\mathcal{G}$ and $\mathcal{G}'$ are* observationally equivalent *if their canonical S-pDAGs are observationally equivalent; that is, if $\mathcal{M}(\texttt{can}(\mathcal{G}), V \mid S) = \mathcal{M}(\texttt{can}(\mathcal{G}'), V \mid S')$, where S is the set of selected nodes in $\texttt{can}(\mathcal{G})$ and S' is the set of selected nodes in $\texttt{can}(\mathcal{G}')$.*

The problem of establishing observational equivalence classes of causal structures is complicated, and it is not fully solved even in the case without selection effects, as we have seen in Chapter 3. Therefore, in this section we do not aim to completely solve this problem for the case with selection effects. Instead, we will simply present a list of rules that give sufficient—but not necessary—conditions for two smDGs to be observationally equivalent.

4.5.1 Adding marginalized faces

Our first rule for proving observational equivalence (but not Observe&Do equivalence) of liftable smDGs gives conditions under which a marginalized face can be *added* without changing the selected-marginal observational model. In many cases, this added marginalized face will enable the application of our subsequent rule, Proposition 23, where this would not be otherwise possible.

Proposition 21. *Let $\mathcal{G} = (V, \mathcal{E}, \mathcal{L}, \mathcal{S})$ be a liftable smDG that includes a maximal selected face $V_s \in \mathcal{S}$. Assume that:*

- *V_s includes all its parents, i.e. $\texttt{pa}_{\mathcal{G}}(V_s) \subseteq V_s$;*
- *every node $v \in V_s$ belongs to some marginalized face of $\mathcal{L}$; and*
- *every marginalized face $V_m \in \mathcal{L}$ that intersects with V_s is subsumed by V_s, i.e. $V_s \cap V_m \neq \varnothing \implies V_m \subseteq V_s$.*

Let $\mathcal{G}' = (V, \mathcal{E}, \mathcal{L}', \mathcal{S})$ be the smDG constructed from $\mathcal{G}$ by adding $\mathrm{cl}(V_s)$ to the marginalized independence system, i.e. $\mathcal{L}' = \mathcal{L} \cup \mathrm{cl}(V_s)$. Then, $\mathcal{G}$ and $\mathcal{G}'$ are observationally equivalent.

Figure 4.5.1 gives an example of the application of Proposition 21. The idea behind this proposition is as follows: a selected node s can induce correlation between its ancestors, which is mediated by the parents of s. For example, the S-pDAG $a \to b \to s \leftarrow c$ can explain correlations between X_a and X_c if we condition on $X_s = 0$, but *not* if we further condition on X_b. In the special case where the ancestors of a selected node s coincide with its parents, the presence of s can induce *any* correlation between its ancestors, without any mediation. That is, it is at least as powerful as a marginalized node m pointing to all of the ancestors of s. The proof of this proposition is supplied in Section C.4.1.



(a) (b)

Figure 4.5.1: Two smDGs that can be shown observationally equivalent by Proposition 21.

4.5.2 Removing special edges

Our second rule for proving observational equivalence shows situations in which we can remove special edges. Specifically, whenever the start and endpoint of a special edge are in the same marginal face, it will be possible to remove that special edge while preserving observational equivalence.

Proposition 22. *Let $\mathcal{G} = (V, \mathcal{E}, \mathcal{L}, \mathcal{S})$ be a liftable smDG. Suppose that $\mathcal{G}$ contains an edge $a \to b$ such that a is an element of some $V_s \in \mathcal{S}$ and b is an element of some $V_m \in \mathcal{L}$. Further suppose that a and b belong to the same face of the marginalized independence system, that is, that a is also an element of V_m. Then, $\mathcal{G}$ is observationally equivalent to the smDG $\mathcal{G}'$ obtained by starting from $\mathcal{G}$ and removing the edge $a \to b$.*

The proof is provided in Section C.4.2. All of the edges $a \to b$ described correspond to special edges $a \to s \leftarrow m \to b$ in the canonical S-pDAG, where s is selected and m is marginalized. The intuition is that if a and b share a common marginalized parent m', selecting on s can establish (through m') all of the correlation between a and b (and the correlation between b and the ancestors of a) that could have been established through the special edge.

From this, we can extract the following corollary:

Corollary 1. *Let $\mathcal{G}$ be a liftable smDG containing the self-loop $a \to a$. Let $\mathcal{G}'$ be the smDG constructed by removing this self-loop. Then, $\mathcal{G}$ and $\mathcal{G}'$ are observationally equivalent.*

Proof. In any liftable smDG with a self-loop $a \to a$, the node a is included in both a marginalized face and a selected face. Therefore, this is the special case of Proposition 22 where a is the same node as b. □

We showed in Section 4.4.2 a particular smDG where the self-loop $v \to v$ was undetectable by the All-patterns Observe-Or-Do probing scheme. Corollary 1 now establishes that at least as far as passive observation goes, such self-loops can *never* be detected.

With this proposition, we can show for example that the smDG of Figure 4.5.1b is observationally equivalent to the smDG of Figure 4.5.2.

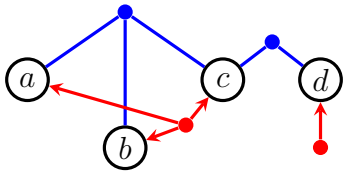


Figure 4.5.2: An smDG that can be shown observationally equivalent to Figure 4.5.1b by Proposition 22 (and Corollary 1).

4.5.3 Removing selected faces

We say that a collider $a \to c \leftarrow b$ is *shielded* if the parents a and b are connected via a directed edge, either $a \to b$ or $b \to a$. In an S-pDAG without marginalized nodes, if a selected node has no unshielded colliders among its ancestors, that is if "moralization" has been performed, then the selected node may be removed without altering the selected distribution [51, Chapter 3]. In the presence of marginalization, however, it is possible to remove a selected node in a wider range of circumstances. Instead of necessarily requiring that the parents a, b of a collider be connected by a directed edge, it also suffices that they share a common marginalized parent.

In this section, we will formulate a condition on smDGs where, based on the idea described above, a maximal selected face can be removed. Note that this condition is *not* simply that the canonical S-pDAG of the smDG is such that ancestors of the corresponding selected node s obey the condition above; that is, include no colliders $a \to c \leftarrow b$ where a and b are not connected via a directed edge nor a marginalized common cause. This

is because there are cases where the canonical S-pDAG does not obey this condition, but some non-canonical S-pDAG that corresponds to the same smDG does. For example, the S-pDAG $s \leftarrow a \rightarrow s' \leftarrow m' \rightarrow b \rightarrow s$, where a points to b with a special edge and both a and b are parents of a selected node s, is canonical. However, in this S-pDAG the nodes a and b are not connected via a directed edge nor share a common marginalized parent, so apparently one cannot remove the node s. However, if one looks at the non-canonical S-pDAG $s \leftarrow a \rightarrow b \rightarrow s$, which is Observe&Do equivalent to $s \leftarrow a \rightarrow s' \leftarrow m' \rightarrow b \rightarrow s$ by Proposition 18, one concludes that s can be removed.

Proposition 23. *Let $\mathcal{G} = (V, \mathcal{E}, \mathcal{L}, \mathcal{S})$ be a liftable smDG, let $V_s \in \mathcal{S}$ be one of its maximal selected faces, and let $\mathcal{G}_{\mathtt{an}(V_s)}$ be the subgraph of $\mathcal{G}$ over the set V_s and all of its ancestors. Suppose that:*

a) There are no cycles in $\mathcal{G}_{\mathtt{an}(V_s)}$.

b) In $\mathcal{G}_{\mathtt{an}(V_s)}$, nodes that share a child or are in the same selected face are neighbours or are in the same marginalized face.

c) All of the maximal marginalized faces of $\mathcal{L}$ that include nodes of $\mathtt{an}_{\mathcal{G}}(V_s)$ are disjoint.

d) For each maximal marginalized face $V_m \in \mathcal{L}$ that includes nodes of $\mathtt{an}_{\mathcal{G}}(V_s)$, the parents of V_m that are not in V_m are shared by all the nodes of V_m, i.e. $\mathtt{pa}_{\mathcal{G}}(V_m) \setminus V_m \subseteq \bigcap_{v \in V_m} \mathtt{pa}_{\mathcal{G}}(v)$.

Construct the independence system $\mathcal{S}'$ by starting from $\mathcal{S}$ and then deleting V_s and all of its subsets. Define the smDG $\mathcal{G}' = (V, \mathcal{E}, \mathcal{L}, \mathcal{S}')$. Then, $\mathcal{G}$ and $\mathcal{G}'$ are observationally equivalent.

When one applies Proposition 22 before Proposition 23, the condition a), which prohibits cycles in $\mathcal{G}_{\mathtt{an}(V_s)}$, is equivalent to the weaker requirement that every cycle of $\mathcal{G}_{\mathtt{an}(V_s)}$ has at least one edge $a \rightarrow b$ in the cycle such that $a \in \bigcup \mathcal{S}$ and both a and b are in the same marginal face, that is, $a, b \in V_m$ for some $V_m \in \mathcal{L}$. This is because all such cycles will be broken by Proposition 22.

The proof of Proposition 23 is in Section C.4.3. The general idea is that a selected node establishes correlation between its ancestors, but if all colliders in the ancestry of a selected node are connected via a directed edge or a marginalized common cause, then any such correlations can be reproduced without using the selected node. The conditions of

Proposition 23 guarantee that this is the case in *some* S-pDAG that has $\mathcal{G}$ as its selected-latent projection.

With Proposition 23, we can show that the smDG of Figure 4.5.2 is observationally equivalent to the smDG of Figure 4.5.3. Note that here it was important that we had previously applied Proposition 21 to Figure 4.5.1a: if the marginalized face $\{a, b, c\}$ was not present, the condition b) of Proposition 23 would not be satisfied for $V_s = \{a, b, c\}$. Therefore, by making the subgraph $\mathcal{G}_{\mathtt{an}_{\mathcal{G}}(V_s)}$ saturated, Proposition 21 enabled the application of Proposition 23. Note, however, that Proposition 23 can also be applied in some cases where $\mathcal{G}_{\mathtt{an}_{\mathcal{G}}(V_s)}$ is not saturated; Figure 4.5.4 is an example.

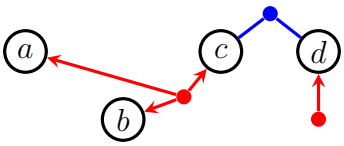


Figure 4.5.3: An smDG that can be shown observationally equivalent to Figure 4.5.2 by Proposition 23.



Figure 4.5.4: Two smDGs that can be shown observationally equivalent by Proposition 23, for $V_s = \{c, d\}$. This is an example where Proposition 23 can be applied even though $\mathcal{G}_{\mathtt{an}_{\mathcal{G}}(V_s)}$ is not saturated.

4.5.4 Leveraging mDAG equivalence for smDG equivalence

Now, we present our last technique for showing observational equivalence between smDGs: leveraging the known rules to prove observational equivalence between mDAGs. To do so, we will take the canonical S-pDAGs of our smDGs, and treat their selected nodes as visible. If the resulting pDAGs are observationally equivalent, then so are the original smDGs.

Proposition 24. *Let* $\mathcal{G} = (V, \mathcal{E}, \mathcal{L}, \mathcal{S})$ *and* $\mathcal{G}' = (V, \mathcal{E}', \mathcal{L}', \mathcal{S})$ *be liftable smDGs without deterministic nodes; that is, where every node is included in some marginalized face*[6]*. Construct the mDAGs* $\mathcal{G}_{mDAG} = (\tilde{V}, \tilde{\mathcal{E}}, \mathcal{L}, \varnothing)$ *and* $\mathcal{G}'_{mDAG} = (\tilde{V}, \tilde{\mathcal{E}}', \mathcal{L}', \varnothing)$ *by transforming every selected node of* $\texttt{can}(\mathcal{G})$ *and* $\texttt{can}(\mathcal{G}')$ *into a visible node, and taking the latent projection thereof. If the mDAGs* $\mathcal{G}_{mDAG}$ *and* $\mathcal{G}'_{mDAG}$ *are observationally equivalent, then the smDGs* $\mathcal{G}$ *and* $\mathcal{G}'$ *are also observationally equivalent.*

Proof. Since $\mathcal{G}_{\text{mDAG}}$ and $\mathcal{G}'_{\text{mDAG}}$ are observationally equivalent, $\texttt{can}(\mathcal{G}_{\text{mDAG}})$ and $\texttt{can}(\mathcal{G}'_{\text{mDAG}})$ have the same marginal observational model. The selected-marginal observational model of $\texttt{can}(\mathcal{G})$ is obtained by conditioning the distributions of the marginal observational model of $\texttt{can}(\mathcal{G}_{\text{mDAG}})$ on an assignment of values to the selected variables. The same is true for $\texttt{can}(\mathcal{G}')$ and $\texttt{can}(\mathcal{G}'_{\text{mDAG}})$. Since the set of selected nodes of $\texttt{can}(\mathcal{G})$ and $\texttt{can}(\mathcal{G}')$ is the same, their selected-marginal observational models are also equal. □

For example, this proposition together with the rule for proving observational equivalence of mDAGs presented in Proposition 5 of [33] can show the observational equivalence between the smDGs of Figure 4.5.3 and Figure 4.5.5.

The steps shown in this section answer the question posed in the example of Figure 4.0.2: since the SLP of Figure 4.0.2a is Figure 4.5.1 and the SLP of Figure 4.0.2b is Figure 4.5.5, we learn that those two S-pDAGs are observationally equivalent. (Since they correspond to different smDGs, we also know that they are Observe&Do inequivalent.)

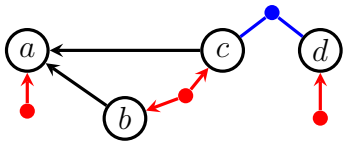


Figure 4.5.5: An smDG that can be shown to be observationally equivalent to Figure 4.5.3 by Proposition 24 together with Proposition 5 of [33]. That is, we treat all of the selected nodes of the canonical S-pDAG as visible, apply Proposition 5 and then take the selected-latent projection.

The complete set of currently known rules for proving observational equivalence between mDAGs was compiled in Chapter 3. Via Proposition 24, all of these rules may be employed to establish observational equivalence between smDGs.

[6] We require each node to have marginalized parents because visible variables in an mDAG are never required to be a deterministic function of their visible parents, and so the mDAG equivalence rules as currently stated can only be applied in this case. As soon as the mDAG is generalized to have deterministic variables, and there are rules for their equivalence, then Proposition 24 can be trivially extended to accommodate their equivalences.

4.6 Conclusion

The first contribution of this chapter is to completely characterize all of the information about the causal structure that can be obtained from the Observe&Do interventional probing scheme, which is as powerful as possible, in the presence of selection and marginalization. We have established that the SMI model is invariant to several kinds of graphical transformations (Section 4.2), which can involve merging and then splitting non-visible components of the graph. This series of changes maps an arbitrary S-pDAG to a canonical S-pDAG, and ultimately, to the smDG, a cyclic graph with marginal and selected independence systems, which more naturally represents the structure under marginalization and selection. We have further established that two S-pDAGs have an identical selected marginal interventional model if and only if their smDGs are the same.

Our second contribution is to investigate weaker probing schemes. In the All-patterns Observe-Or-Do probing scheme, a pair of distinct smDGs are sometimes indistinguishable. This is an interesting distinction from the case without selection effects, where all mDAGs are All-patterns Observe-Or-Do distinguishable (Section 2.5). For the case where only passive observation is allowed, we have presented rules for removing or replacing selected variables, and have discussed how known rules to show observational equivalence between mDAGs may be applied to the case of smDGs.

We now highlight four possible extensions of the work presented in this chapter.

Firstly, the selected-latent projection is complete for Observe&Do equivalence, but we still do not have a complete characterization of observational equivalence between smDGs. In fact, even in the case without selection the observational equivalence classes are not completely characterized for mDAGs with four or more visible variables (Chapter 3). In Section 4.5 we have presented some rules for turning smDGs into observationally equivalent mDAGs, such as removing self-loops, and removing selected faces in some cases. However, we expect that significant further work is required, firstly to exhaust the cases where an smDG can be transformed into an observationally equivalent mDAG, and secondly to exhaust the equivalences between smDGs that have no observationally equivalent mDAG at all.

Secondly, we also do not have a complete characterization of equivalence for the case of weaker interventional probing schemes, such as for the All-patterns Observe-or-Do probing scheme. This is unlike the case without selection, where both Observe&Do equivalence and Observe-or-Do equivalence is completely characterized by the mDAG structure (Chapter 2).

Thirdly, we have treated the case where a variable is *selected* to a single value (that we called 0) so that under selection, their children would be unaffected by them, which is why

we were able to remove outgoing edges from selected nodes without altering the selected model. More generally, we could consider scenarios where variables are *conditioned* to a set of values, and where outgoing edges from selected variables would have to be considered.

Finally, although our Observe&Do interventional scheme is highly general, it excludes the edge interventions of Shpitser and Tchetgen Tchetgen [71], and further work could seek to evaluate the circumstances in which edge interventions can distinguish a pair of S-pDAGs.

Chapter 5

Finding new causal structures that present quantum-classical gaps

Up to this point, in this thesis we have only studied *classical* causal inference, that is, we have only considered causal models whose semantics is classical. In this chapter, we will discuss causal models whose semantics sometimes allows for marginalized nodes to be associated with quantum systems. Visible and selected nodes, however, will still be associated with classical variables. Since most of this chapter deals only with pDAGs (and not S-pDAGs), we will go back to using the terminology "latent" to refer to the marginalized nodes.

When the semantics of a causal model allows for quantum latent nodes, *realizability* of a probability distribution is not given by Equation (2.3) anymore. Instead, now a probability distribution over the visible variables is *realizable by a pDAG $\mathcal{G}$ under a semantics $\mathscr{T}$* if it can be written as:

$$P(X_{\mathtt{vis}(\mathcal{G})}) = \sum_{X_{\mathtt{Clat}(\mathcal{G})}} \prod_{a\in A} P(X_a|X_{\mathtt{pa}_{\mathcal{G}}(a)}) \operatorname{Tr}\left[\bigotimes_{b\in B} \Pi_{\mathcal{H}_b^{\text{in}}}^{X_b|X_{\mathtt{Cpa}_{\mathcal{G}}(b)}} \bigotimes_{u\in U} \mathcal{E}_{\mathcal{H}_u^{\text{out}}|\mathcal{H}_u^{\text{in}}}^{X_{\mathtt{Cpa}_{\mathcal{G}}(u)}} \left(\bigotimes_{l\in L} \rho_{\mathcal{H}_l^{\text{out}}}^{X_{\mathtt{pa}_{\mathcal{G}}(l)}}\right)\right] \tag{5.1}$$

where, relative to $\mathscr{T}$, A is the set of classical nodes that only have classical parents, B is the set of classical nodes that have quantum parents, L is the set of quantum nodes that only have classical parents, U is the set of quantum nodes that have quantum parents, $\mathtt{Clat}(\mathcal{G})$ is the set of all classical latent nodes of $\mathcal{G}$, and $\mathtt{Cpa}_{\mathcal{G}}(v)$ is the set of classical parents of a node v. Furthermore:

- $\Pi_{\mathcal{H}_b^{\text{in}}}^{X_b|X_{\mathtt{Cpa}_{\mathcal{G}}(b)}}$ is a POVM element applied by b on the system it receives from its quantum

parents, $\mathcal{H}_b^{\text{in}}$, that is indexed by the outcome X_b and is chosen by the value of the classical parents $X_{\texttt{Cpa}_{\mathcal{G}}(b)}$,

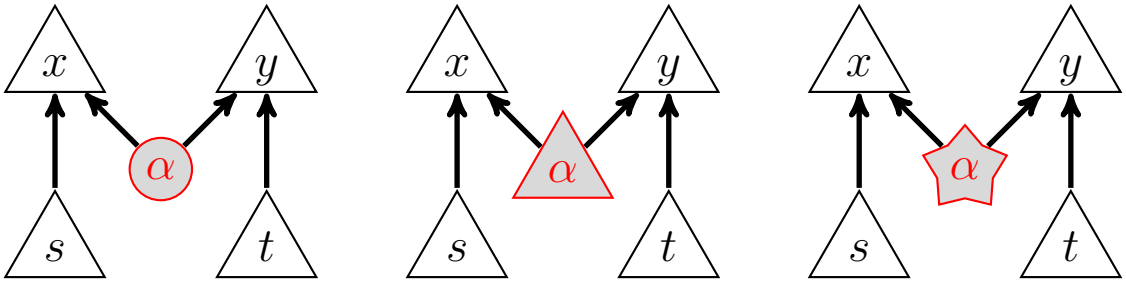
- $\mathcal{E}_{\mathcal{H}_u^{\text{out}}|\mathcal{H}_u^{\text{in}}}^{X_{\texttt{Cpa}_{\mathcal{G}}(u)}}$ is a quantum channel applied by u on the system it receives from its quantum parents, $\mathcal{H}_u^{\text{in}}$, with output in the system $\mathcal{H}_u^{\text{out}}$ that it will send to its children, that is chosen by the value of the classical parents $X_{\texttt{Cpa}_{\mathcal{G}}(u)}$, and

- $\rho_{\mathcal{H}_l^{\text{out}}}^{X_{\texttt{pa}_{\mathcal{G}}(l)}}$ is a quantum state in which the system $\mathcal{H}_l^{\text{out}}$ associated with the node l is prepared, that is chosen by the value of the parents $X_{\texttt{pa}_{\mathcal{G}}(l)}$ (remember that all of the parents of l are classical).

For the definition of realizability by an S-pDAG, one simply divides the joint distribution $P\big(X_{\texttt{vis}(\mathcal{G})}, X_{\texttt{sel}(\mathcal{G})} = \vec{0}\big)$ obtained as above by the marginal $P\big(X_{\texttt{sel}(\mathcal{G})} = \vec{0}\big)$. Definitions of observational equivalence and dominance are extended accordingly.

To be able to compare cases with different semantics more easily, in this chapter we will use a different notation for depicting causal structures. More specifically, we will introduce notation to depict a causal structure endowed with a choice of semantics: classical nodes will be represented by triangles, quantum nodes will be represented by stars and nodes whose theory was not yet specified will be represented by circles. As before, visible nodes will be represented with black border and white background, and latent nodes will be represented with red border and gray background. To exemplify this new notation, Figure 5.0.1 presents three versions of the Bell pDAG. Since we always assume that visible and selected nodes are classical, we always depict them as triangles. Figure 5.0.1(a) shows the Bell pDAG with an *unspecified* semantics, Figure 5.0.1(b) shows the Bell pDAG endowed with a *classical* semantics, which is the term we use when all of the latent nodes are classical, and Figure 5.0.1(c) shows the Bell pDAG endowed with a *quantum* semantics, which is the term we use when all of the latent nodes are quantum. The set of distributions with cardinalities $\vec{c}_{\text{vis}}$ of the visible variables that are *classically* realizable by a pDAG $\mathcal{G}$, which was denoted by $\mathcal{M}_{\text{obs}}(\mathcal{G}, \vec{c}_{\text{vis}})$ in Chapters 2 and 3, will be denoted by $\mathcal{M}_C(\mathcal{G}, \vec{c}_{\text{vis}})$ here. The set of distributions with cardinalities $\vec{c}_{\text{vis}}$ of the visible variables that are *quantumly* realizable by a pDAG $\mathcal{G}$ will be denoted by $\mathcal{M}_Q(\mathcal{G}, \vec{c}_{\text{vis}})$. In general, a pDAG could be endowed with a semantics $\mathscr{T}$ where some of the latent nodes might be quantum and others classical. In this case, the set of distributions with cardinalities $\vec{c}_{\text{vis}}$ of the visible variables that are realizable by $\mathcal{G}$ under $\mathscr{T}$ will be denoted by $\mathcal{M}_{\mathscr{T}}(\mathcal{G}, \vec{c}_{\text{vis}})$.

In this chapter we are interested in finding new causal structures that present a quantum-classical gap (QC gap). A causal structure presents a QC gap if it can realize strictly more



(a) (b) (c)

Figure 5.0.1: (a) Bell pDAG with an unspecified semantics. (b) Bell pDAG with a classical semantics. (c) Bell pDAG with a quantum semantics.

probability distributions under a quantum semantics than under a classical semantics. As discussed before, a pre-requisite for a causal structure to present a QC gap is it being nonalgebraic. Thus, the work of Chapter 3 is relevant in our search for new causal structures that present QC gaps: the causal structures that were identified as algebraic in Chapter 3 can automatically be filtered out and don't need to be inspected. Now, we will first describe how we will choose candidate causal structures for this search for QC gaps, and then we will describe the idea behind the techniques we developed for doing so.

We start by noting one important fact: two causal structures that correspond to the same mDAG are *not necessarily* observationally equivalent in the quantum case. This was first noted in Ref. [81], and we reproduce the argument below.

The two pDAGs of Figure 5.0.2 are associated with the same mDAG, and thus are classically observationally equivalent. Furthermore, as shown in Chapter 2, in the classical case they are also equivalent under access to informationally complete probing schemes, such as Observe&Do. However, in the quantum case they are not even observationally equivalent. To see that, consider a probability distribution where

$$\begin{cases} P(X_a, X_b | X_x, X_y, X_s = 0) \text{ maximally violates the CHSH inequality} \\ P(X_a, X_c | X_x, X_z, X_s = 1) \text{ maximally violates the CHSH inequality} \end{cases}, \tag{5.2}$$

that is, a distribution where Alice and Bob achieve the Tsirelson bound when $X_s = 0$, and Alice and Charlie achieve the Tsirelson bound when $X_s = 1$.

This distribution is quantumly *realizable* by the pDAG of Figure 5.0.2(a), with the following choice of parameters: first, the node α is associated with a preparation of two

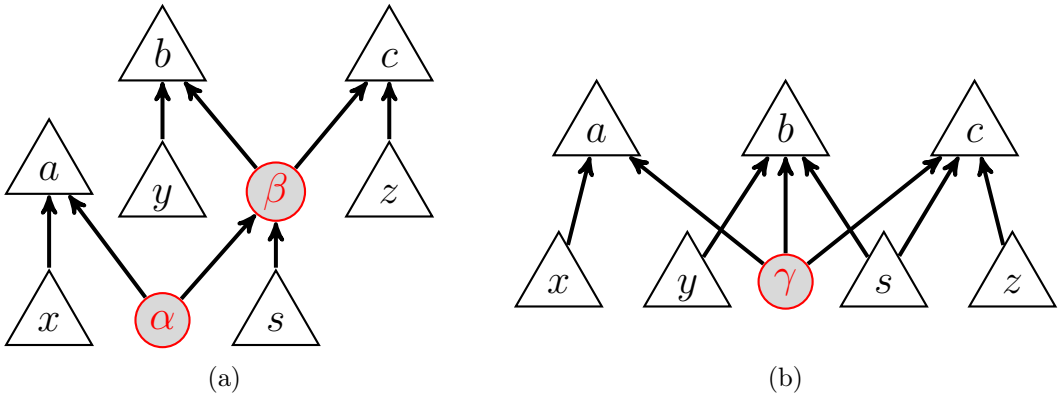


Figure 5.0.2: Two pDAGs that correspond to the same mDAG, and thus are Observe&Do equivalent in the classical case, but that are observationally *inequivalent* in the quantum case. Example taken from Ref. [81].

qubits in a maximally entangled state. Then, it sends one of these qubits to a and the other to β. If $X_s = 0$, β sends the qubit it received from α to b, and sends another qubit that is not entangled with the others to c. If $X_s = 1$, β sends the qubit it received from α to c, and sends another qubit that is not entangled with the others to b. The nodes a, b and c react to their parents in the same way they would in a standard Bell experiment that achieves the Tsirelson bound.

This distribution, however, is *not* quantumly realizable by the pDAG of Figure 5.0.2(b). Due to self-testing, we know that for Alice and Bob to achieve the Tsirelson bound in some rounds, we need them to share a maximally entangled state in those rounds, and the same for Alice and Charlie. However, in Figure 5.0.2(b) the quantum latent node γ does not have access to the setting s, so it cannot decide to which child to send the quantum systems based on the value of X_s (like the node β did in Figure 5.0.2(a)). Instead, γ would have to prepare a tripartite quantum state whose marginal state sent to a and b and whose marginal state sent to a and c are both maximally entangled, which is impossible by monogamy of entanglement. To read more about causal structures that are classically observationally equivalent but quantumly observationally inequivalent, see Ref. [17].

Therefore, we know that the *quantum* observational partial order is going to be very different from the *classical* observational partial order studied in Chapter 3. In particular, note that if we know that one pDAG $\mathcal{G}$ has a QC gap, this does not necessarily imply that another pDAG $\mathcal{G}'$ that is *classically* observationally equivalent to $\mathcal{G}$ also has a QC gap: it is conceivable that quantumly, $\mathcal{G}'$ is weaker than $\mathcal{G}$ in such a way as to close the gap. At the moment we do not have any explicit examples of this phenomena, but this possibility

implies that learning that one causal structure from a classical observational equivalence class has a QC gap does not allow us to conclude that every causal structure in the class also presents a QC gap. Therefore, we will not start our search for new QC gaps from a list of representatives of the observational equivalence classes found in Chapter 3. We *will*, however, start our search for new QC gaps from a list of mDAGs, as we justify now.

Take again the example of Figure 5.0.2. In this figure, the pDAG of 5.0.2(a) quantumly observationally *dominates* the pDAG of 5.0.2(b). To see that, let $P(X_aX_bX_c|X_xX_yX_zX_s)$ be a distribution realizable by 5.0.2(b). The realization of this distribution requires γ to be associated with a tripartite Hilbert space $\mathcal{H}_a \otimes \mathcal{H}_b \otimes \mathcal{H}_c$ in the state ρ_{abc}. In 5.0.2(a), the same state can be shared between the nodes a, b and c, by simply being prepared in α. Then, α sends $\mathcal{H}_a$ to a and $\mathcal{H}_b \otimes \mathcal{H}_c$ to β, which in turn sends $\mathcal{H}_b$ to b and $\mathcal{H}_c$ to c. Furthermore, the information about s can also be shared to b and c through β. Therefore, $P(X_aX_bX_c|X_xX_yX_zX_s)$ can also be realized in 5.0.2(a). The converse is not true, as we have seen from the example of Equation (5.2).

In Figure 5.0.2, the pDAG of 5.0.2(b) is `RE`-reduced, while the pDAG of 5.0.2(a) is not. As it turns out, the fact described above is generic: a `RE`-reduced pDAG is always quantumly observationally dominated by any other pDAG that is associated with the same mDAG. Classically, we know that these pDAGs are observationally equivalent. Together, these facts imply that, if we prove that the canonical pDAG associated with an mDAG (Definition 11) presents a QC gap, then *all* of the pDAGs associated with that mDAG also present a QC gap. Therefore, when searching for new causal structures that present QC gaps, we will focus our attention on canonical pDAGs associated with mDAGs. With this strategy we might miss some QC gaps (i.e., there could in principle be a pDAG $\mathcal{G}$ that presents a QC gap while `RE-reduce`$(\mathcal{G})$ does not), but we will at least be sure that the QC gaps we show can be extended to all of the pDAGs associated with the same mDAG. Therefore, we will start our search from a list of mDAGs, and in the end we will compare the total number of mDAGs known to be nonalgebraic with the number of mDAGs we learned to present a QC gap. It is known that *all* of the nonalgebraic 3-node mDAGs *do* present a QC gap (and thus we are not missing any pDAG with three visible nodes that has a QC gap): those are the mDAGs that appear in the classes `Triangle`, `Evans`, `Instrumental BAC`, `Instrumental CAB` and `Instrumental ABC` of Figure 3.2.1.

Now, we describe the idea behind the techniques we will use to prove new QC gaps among 4-node mDAGs. The strategy is to start from a pDAG $\mathcal{G}_{\text{new}}$ for which it is not known whether a QC gap exists, apply to it a series of transformations that *do not increase* the QC gap, and then reach a different pDAG $\mathcal{G}_{\text{pig}}$ for which a QC gap is *known*, such as the Bell, Instrumental [18, 78] or Triangle [39] scenarios. This allows us to conclude that $\mathcal{G}_{\text{new}}$ also has a QC gap. Since we are piggybacking on the knowledge of QC gaps in certain

causal structures to prove QC gaps in new causal structures, we will call these techniques *piggybacks*. In Section 5.1, we will explain our piggybacks through simple examples, that already encode the main ideas of each of them. In Section 5.2, we will present formal results that generalize these ideas. As we will see in the simple examples, at times our proofs of new QC gaps make use of *observational dominance* results that are valid for quantum semantics. All of the observational dominance results of Chapters 3 and 4 were only proven for classical semantics; as such, in Section 5.3 we present a list of observational dominance rules that are valid in the quantum case. Apart from being useful in the project of finding new QC gaps, these results are interesting in themselves, as generalizations of some of the results of Chapters 3 and 4.

In Section 5.4, we start from the list of all nonalgebraic mDAGs of four nodes compatible with a fixed nodal ordering and show that the *majority* of them present QC gaps: out of a total number of 2807 such mDAGs, now we know that 2587 present QC gaps. Previously, only 48 of these mDAGs were known to present QC gaps: the Bell mDAG and the other mDAGs in the same classical observational equivalence class as the Bell mDAG (Section 3.6.1).

5.1 Examples of application of the piggybacks

In this section, we will carefully go through examples that illustrate all of the main ideas in our techniques to prove the existence of a QC gap in new causal structures.

5.1.1 The Point Distribution Piggyback: Bell-with-mediary Scenario

We start by showing that the pDAG $\mathcal{G}_{\text{BellMed}}$ of Figure 5.1.1, which we call *Bell-with-mediary*, has a QC gap. The technique we will use to do so is called the *point-distribution piggyback*.

At first glance one could doubt whether $\mathcal{G}_{\text{BellMed}}$ presents a QC gap or not, because it allows for causal influence between s and y. However, note that such influence passes through a *visible* mediary node, m.

The point distribution piggyback will consist in only looking at probability distributions where X_m is fixed to a point distribution, that is, where the cardinality of X_m is one. Since our definitions say that a causal structure exhibits a QC gap if it exhibits a QC gap for

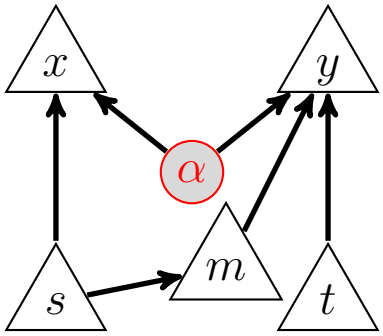


Figure 5.1.1: The Bell-with-mediary pDAG $\mathcal{G}_{\text{BellMed}}$. The nodes x, y, s, t and m are visible, and the node α is latent.

some choice of cardinalities for the visible nodes, looking at the case of cardinality one is perfectly valid for showing a QC gap. We will adopt the convention that point distributions set the variable to the value 0, and a point distribution on X_m will be represented by $\delta_{X_m,0}$.

A variable that is fixed to a point distribution is not affected by any of its causal parents. Likewise, it cannot be used to causally explain the correlation between any other variables. Therefore, a distribution where X_m has cardinality one is realizable by $\mathcal{G}_{\text{BellMed}}$ under some assignment of theory for the latent node α *if and only if* it is also realizable by the pDAG obtained by deleting all incoming and outgoing edges from m, i.e., the Bell pDAG plus a disconnected node m, under the same assignment of theory for α. In other words:

$$P(X_xX_yX_sX_tX_m) \in \mathcal{M}_{\mathscr{T}}(\mathcal{G}_{\text{BellMed}}, (2,2,2,2,1))$$
$$\Updownarrow$$
$$P(X_xX_yX_sX_tX_m) = P(X_xX_yX_sX_t)\delta_{X_m,0} \text{ for } P(X_xX_yX_sX_t) \in \mathcal{M}_{\mathscr{T}}(\mathcal{G}_{\text{Bell}}, (2,2,2,2,1)).$$

Since we know that there is a QC gap in the Bell pDAG, the reasoning above implies that there is also a QC gap in the Bell-with-mediary pDAG: for any distribution $P(X_xX_yX_sX_t)$ that is quantumly but not classically realizable by the Bell pDAG, the distribution $P(X_xX_yX_sX_tX_m) = P(X_xX_yX_sX_t)\delta_{X_m,0}$ is quantumly but not classically realizable by the Bell-with-mediary pDAG.

5.1.2 The Marginalization Piggyback

In this section we will exemplify the *marginalization piggyback*. The idea behind this technique is to turn one of the visible nodes of a pDAG into a classical latent node, and

analyze whether the resulting pDAG has a QC gap.

We will give three examples of this piggyback, in increasing order of complexity. This difference in complexity comes from the fact that sometimes we can simply use classical causal inference techniques to reduce the resulting pDAG to one where a QC gap is known, but sometimes we also need to appeal to quantum arguments such as the teleportation protocol [10].

Bell-with-broken-leg Scenario

Here, we will show that there is a QC gap in the pDAG $\mathcal{G}_{\text{BrokenLeg}}$ of Figure 5.1.2, termed the *Bell-with-broken-leg pDAG*. We will piggyback on the QC gap of the Bell pDAG.

The QC gap in $\mathcal{G}_{\text{BrokenLeg}}$ can be shown in a few different ways. The most straightforward way is by the point distribution piggyback: if we fix c to a point distribution, the remaining pDAG is the Bell pDAG. A second way is by turning the node c into a latent node, also ending up with the Bell pDAG. A third way is by turning the node a into a latent node, as we describe now.

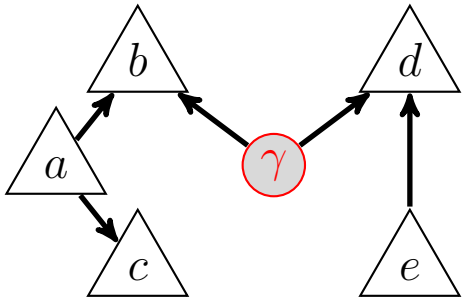


Figure 5.1.2: The Bell-with-broken-leg pDAG $\mathcal{G}_{\text{BrokenLeg}}$.

If the node a in the Bell-with-broken-leg pDAG is turned into a latent node, we obtain the pDAG $\mathcal{G}_{\text{EqToBell1}}$ of Figure 5.1.3. Therefore, if

$$P_C(X_aX_bX_cX_dX_e) \in \mathcal{M}_C(\mathcal{G}_{\text{BrokenLeg}}, (n, 2, 2, 2, 2))$$

is a distribution that is classically realizable by $\mathcal{G}_{\text{BrokenLeg}}$, its marginal over X_b, X_c, X_d and X_e must be classically realizable by $\mathcal{G}_{\text{EqToBell1}}$:

$$P_C(X_bX_cX_dX_e) \in \mathcal{M}_C(\mathcal{G}_{\text{EqToBell1}}, (2, 2, 2, 2)), \tag{5.3}$$

where

$$P_C(X_bX_cX_dX_e) = \sum_{X_a} P_C(X_aX_bX_cX_dX_e). \tag{5.4}$$

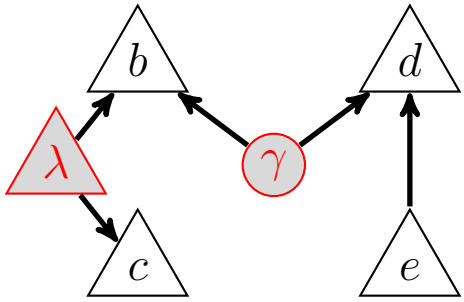


Figure 5.1.3: The pDAG $\mathcal{G}_{\text{EqToBell1}}$, obtained from $\mathcal{G}_{\text{BrokenLeg}}$ by turning a into a classical latent node.

As noted in Section 3.6.1, $\mathcal{G}_{\text{EqToBell1}}$ is classically observationally equivalent to the Bell pDAG. Furthermore, when the latent node γ is quantum and the latent node λ is classical, $\mathcal{G}_{\text{EqToBell1}}$ can reproduce any distribution that is quantumly realizable by the Bell pDAG. This can be done simply by setting X_λ to the distribution that X_c would take in the Bell pDAG, having X_c copy X_λ, and having X_b depend on X_λ in the same way that it would depend on X_c in the Bell pDAG.

Thus, there exist distributions $P_Q(X_bX_cX_dX_e)$ that are realizable by $\mathcal{G}_{\text{EqToBell1}}$ when λ is classical and γ is quantum, but not when both latent nodes are classical. Since the latent node λ is classical, such a distribution is the marginal over X_b, X_c, X_d, X_e of a joint distribution $P_Q(X_\lambda X_bX_cX_dX_e)$ that is quantumly realizable by $\mathcal{G}_{\text{BrokenLeg}}$, identifying λ with a. Therefore, there is a QC gap in $\mathcal{G}_{\text{BrokenLeg}}$.

Charlie Brown Tree Scenario

Now, we will show that there is a QC gap in the pDAG $\mathcal{G}_{\text{Charlie}}$ of Figure 5.1.4, termed the *Charlie Brown Tree pDAG*. This gap appears even when the latent node μ remains classical, if γ goes quantum. We will piggyback on the QC gap of the Instrumental pDAG $\mathcal{G}_{\text{Instr}}$ (Figure 5.1.5).

The marginalization piggyback can show that there is a QC gap in the Charlie Brown Tree pDAG when the cardinalities are $(c_a, c_b, c_c, c_d) = (3, 2, 2, 2)$.

If the node d in the Charlie Brown Tree pDAG is turned into a latent node, we obtain the pDAG $\mathcal{G}_{\text{EqToInstr1}}$ of Figure 5.1.6. Therefore, if $P_C(X_aX_bX_cX_d) \in \mathcal{M}_C(\mathcal{G}_{\text{Charlie}}, (3, 2, 2, 2))$

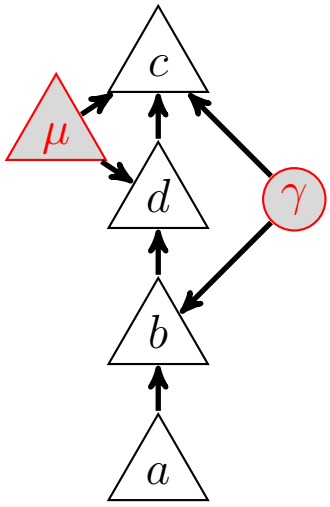


Figure 5.1.4: The Charlie Brown Tree pDAG $\mathcal{G}_{\text{Charlie}}$. Here we look at semantics where μ is classical, because those are enough to show the QC gap.

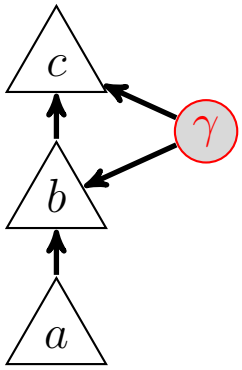


Figure 5.1.5: The Instrumental pDAG $\mathcal{G}_{\text{Instr}}$.

is a distribution that is classically realizable by $\mathcal{G}_{\text{Charlie}}$, its marginal over X_a, X_b and X_c must be classically realizable by $\mathcal{G}_{\text{EqToInstr1}}$.

It is easy to see that $\mathcal{G}_{\text{EqToInstr1}}$ is observationally equivalent to $\mathcal{G}_{\text{Instr}}$ both when all latent nodes are classical *and* when γ is quantum and the other two are classical. This is because the latent node λ and all its parents are classical, so we can apply exogenization to it.

We know that the Instrumental scenario has a QC gap when the cardinalities are $(c_a, c_b, c_c) = (3, 2, 2)$ [18, 78]. Therefore, we conclude that there exists at least one dis-

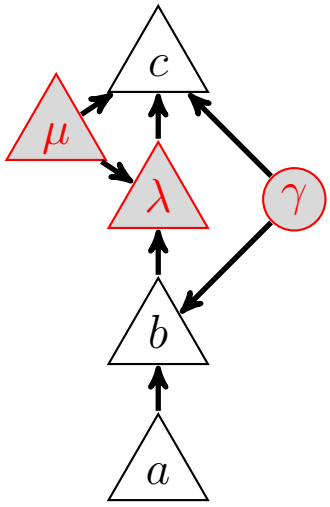


Figure 5.1.6: The pDAG $\mathcal{G}_{\text{EqToInstr1}}$, obtained from $\mathcal{G}_{\text{Charlie}}$ by turning d into a classical latent node.

tribution $P_Q(X_aX_bX_c)$ over variables with cardinalities $(c_a, c_b, c_c) = (3, 2, 2)$ that is not classically realizable by $\mathcal{G}_{\text{EqToInstr1}}$, but is realizable by $\mathcal{G}_{\text{EqToInstr1}}$ when the latent node γ is quantum and the other two are classical.

Since the latent nodes λ and μ are classical, this distribution is the marginal over X_a, X_b, X_c of a joint distribution $P_Q(X_aX_bX_cX_\mu X_\lambda)$. Marginalizing over X_μ and identifying λ with d, we can conclude that there exists one distribution $P_Q(X_aX_bX_cX_d)$ that is realizable by the Charlie Brown Tree pDAG when γ is quantum, but whose marginal $P_Q(X_aX_bX_c)$ is *not* classically realizable by $\mathcal{G}_{\text{EqToInstr1}}$, and thus $P_Q(X_aX_bX_cX_d)$ is *not* classically realizable by the Charlie Brown Tree pDAG.

Pike Scenario

Now, we will show that there is a QC gap in the pDAG $\mathcal{G}_{\text{Pike}}$ of Figure 5.1.7, termed the *Pike pDAG*. We will piggyback on the QC gap of the Instrumental pDAG $\mathcal{G}_{\text{Instr}}$ (Figure 5.1.5). As we will see, this example is a bit more complicated than the previous ones.

We start by transforming the visible node d into a classical latent node, obtaining the pDAG $\mathcal{G}_{\text{EqToInstr2}}$ of Figure 5.1.8. It is easy to see that this pDAG is classically observationally equivalent to the Instrumental pDAG, by exogenization.

As we will show now, when the latent nodes α and β are quantum and the latent

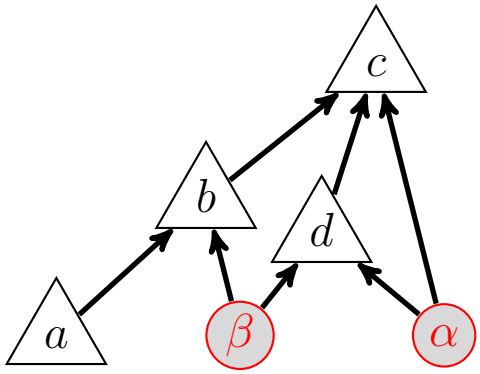


Figure 5.1.7: The Pike pDAG $\mathcal{G}_{\text{Pike}}$.

node λ is classical, $\mathcal{G}_{\text{EqToInstr2}}$ can realize every probability distribution that is quantumly realizable by the Instrumental pDAG.

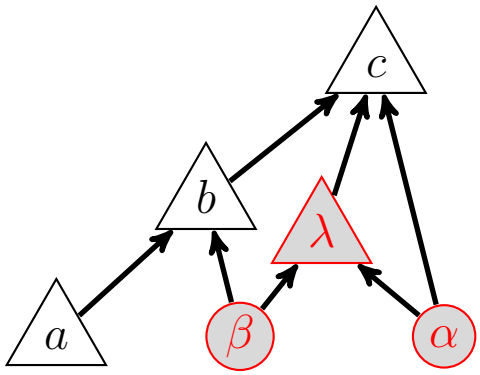


Figure 5.1.8: The pDAG $\mathcal{G}_{\text{EqToInstr2}}$, obtained from $\mathcal{G}_{\text{Pike}}$ by turning d into a classical latent node.

In the Instrumental pDAG, a quantumly realizable distribution $P_Q(X_a X_b X_c)$ is generated by assigning a specific quantum state $\sigma \in \mathcal{H}_B \otimes \mathcal{H}_C$ to the latent node, then sending system $\mathcal{H}_B$ to the node b and system $\mathcal{H}_C$ to the node c, that subsequently make measurements to generate the values of X_b and X_c.

We show that any shared state σ between b and c can be reproduced in $\mathcal{G}_{\text{EqToInstr2}}$ when λ is classical but α and β are quantum. We start by assigning the Hilbert Space $\mathcal{H}_\alpha = \mathcal{H}_{C1} \otimes \mathcal{H}_{C2}$ to α and the Hilbert space $\mathcal{H}_\beta = \mathcal{H}_B \otimes \mathcal{H}_{C3}$ to β, where $\dim(\mathcal{H}_{C1}) = \dim(\mathcal{H}_{C2}) = \dim(\mathcal{H}_{C3}) = \dim(\mathcal{H}_C)$. The system $\mathcal{H}_{C1}$ will be sent to c, the system $\mathcal{H}_B$ will be sent to b, and the systems $\mathcal{H}_{C2}$ and $\mathcal{H}_{C3}$ will be sent to λ.

Choose the system $\mathcal{H}_\beta$ to be in the state σ, and the system $\mathcal{H}_\alpha$ to be in the following

maximally entangled state:

$$|\Phi^+\rangle_\alpha = \frac{1}{\sqrt{\dim(\mathcal{H}_C)}} \sum_{i=1}^{\dim(\mathcal{H}_C)-1} |i\rangle_{C1} \otimes |i\rangle_{C2} \tag{5.5}$$

for choices of basis $\{|i\rangle\}_{\dim(\mathcal{H}_C)}$ for $\mathcal{H}_{C1}$ and $\mathcal{H}_{C2}$.

Now, λ and c can perform the teleportation protocol. That is, following Ref. [10], the node λ performs a generalized Bell basis measurement on the system $\mathcal{H}_{C2} \otimes \mathcal{H}_{C3}$, communicates its outcome to the node c, and c performs the appropriate unitary transformation to $\mathcal{H}_{C1}$. With this, the initial state of the system $\mathcal{H}_{C3}$ is now the state that describes the system $\mathcal{H}_{C1}$. Since the system $\mathcal{H}_\beta = \mathcal{H}_B \otimes \mathcal{H}_{C3}$ was initially in the state σ, now the system $\mathcal{H}_B \otimes \mathcal{H}_{C1}$ is in the state σ.

After this, nodes b and c only need to perform on $\mathcal{H}_B$ and $\mathcal{H}_{C1}$ the same measurements they would have performed on $\mathcal{H}_B$ and $\mathcal{H}_C$ to realize $P_Q(X_a X_b X_c)$ in the Instrumental pDAG, and the distribution $P_Q(X_a X_b X_c)$ is realized in $\mathcal{G}_{\text{EqToInstr2}}$. Therefore, the pDAG $\mathcal{G}_{\text{EqToInstr2}}$ exhibits a QC gap. By the same logic as the one presented in the other examples of the marginalization piggyback, then, the Pike pDAG also exhibits a QC gap.

5.1.3 The Fritz Piggyback: Triangle Scenario

In this section, we will review the proof of a QC gap in the Triangle scenario (DAG $\mathcal{G}_{\text{Tri}}$ of Figure 5.1.9) that is presented in Ref. [39]. Due to Ref. [39], we term the technique used here as the *Fritz piggyback*.

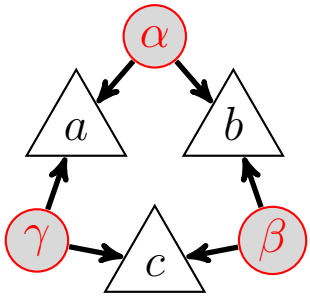


Figure 5.1.9: The Triangle pDAG $\mathcal{G}_{\text{Tri}}$. The nodes a, b and c are visible, and the nodes α, β and γ are latent.

This proof will show that there exist distributions $P(X_a X_b X_c)$ in the gap *when the variables a, b and c have cardinality* 4. When the variables have cardinality 2 or 3, such distributions also exist, but are proven by different techniques [14, 28].

If we assume that a, b and c have cardinality 4, we can view each one as a pair of variables, each one with cardinality 2:

$$a = (a_1, a_2) \tag{5.6}$$
$$b = (b_1, b_2) \tag{5.7}$$
$$c = (c_1, c_2) \tag{5.8}$$

A distribution $P(X_a X_b X_c)$ over cardinalities $\vec{c}_{\text{vis}} = (4, 4, 4)$ is *classically* realizable by the Triangle pDAG if and only if it is classically realizable by the pDAG $\mathcal{G}_{\text{SplitTri}}$ of Figure 5.1.10 when looked at as a distribution $P(X_{a_1} X_{a_2} X_{b_1} X_{b_2} X_{c_1} X_{c_2})$ over cardinalities $\vec{c}_{\text{vis}} = (2, 2, 2, 2, 2, 2)$. When splitting a variable in two, we need to make sure that the two resulting nodes share at least one latent parent. This is so because in the original pDAG these two variables can be arbitrarily correlated (for example, a_1 and a_2 are both inside the node a, so $P(X_{a_1} X_{a_2})$ can exhibit a lot of correlation between X_{a_1} and X_{a_2}), so after splitting the two variables we still want them to share any correlation possible.

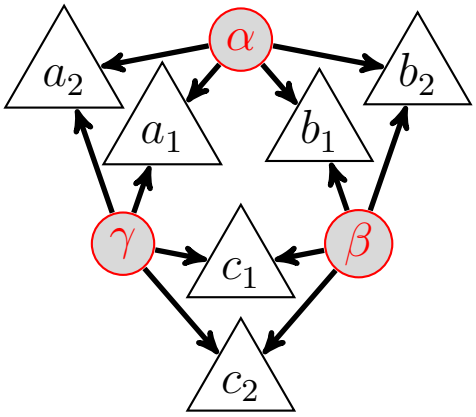


Figure 5.1.10: DAG $\mathcal{G}_{\text{SplitTri}}$ obtained by splitting each visible variable of the Triangle pDAG into two variables of smaller cardinality.

The set of all probability distributions $P(X_{a_1} X_{a_2} X_{b_1} X_{b_2} X_{c_1} X_{c_2})$ over binary variables that are classically realizable by $\mathcal{G}_{\text{SplitTri}}$ is $\mathcal{M}_C(\mathcal{G}_{\text{SplitTri}}, (2, 2, 2, 2, 2, 2))$. Let

$$\mathcal{S}_C \subseteq \mathcal{M}_C(\mathcal{G}_{\text{SplitTri}}, (2, 2, 2, 2, 2, 2))$$

be the subset of those distributions that also obey $X_{a_2} = X_{c_2}$ (with statistical dependence)[1]

[1] That is, we want distributions where X_{a_2} and X_{c_2} are equal and *can take more than one value*. If a probability distribution is such that $P(X_{a_2} = X_{c_2} = 0) = 1$, these two variables are statistically independent, and thus this distribution is not part of $\mathcal{S}_C$. We refer to the property of two variables being equal and statistically dependent as *perfect correlation*.

and $X_{b_1} = X_{c_1}$ (with statistical dependence).

Let $\tilde{P}_{\mathcal{S}_C}(X_{a_1}X_{a_2}X_{b_1}X_{b_2}X_{c_1}X_{c_2}) \in \mathcal{S}_C$ be a probability distribution that belongs to $\mathcal{S}_C$. We know that the underlying distribution $\tilde{P}_{\mathcal{S}_C}(X_{a_1}X_{a_2}X_{b_1}X_{b_2}X_{c_1}X_{c_2}X_\alpha X_\beta X_\gamma)$ that gives rise to $\tilde{P}_{\mathcal{S}_C}(X_{a_1}X_{a_2}X_{b_1}X_{b_2}X_{c_1}X_{c_2})$ by marginalization has to be such that $\tilde{P}_{\mathcal{S}_C}(X_{c_2}|X_\alpha X_\gamma) = \tilde{P}_{\mathcal{S}_C}(X_{c_2}|X_\gamma)$, because $\mathcal{G}_{\text{SplitTri}}$ exhibits the d-separation relation $c_2 \perp \alpha|\gamma$. Furthermore, since $\tilde{P}_{\mathcal{S}_C}(X_{a_1}X_{a_2}X_{b_1}X_{b_2}X_{c_1}X_{c_2})$ is an element of $\mathcal{S}_C$, we have that X_{a_2} and X_{c_2} are perfectly correlated in this distribution. Therefore, if c_2 is independent of α when conditioned on γ, so is a_2, that is, we conclude $\tilde{P}_{\mathcal{S}_C}(X_{a_2}|X_\alpha X_\gamma) = \tilde{P}_{\mathcal{S}_C}(X_{a_2}|X_\gamma)$. That is, the parameter of the causal model associated with a_2 does not depend on α; therefore, this distribution is also realizable by a pDAG obtained from $\mathcal{G}_{\text{SplitTri}}$ by dropping the arrow $\alpha \to a_2$.

We can use analogous arguments to show that $\tilde{P}_{\mathcal{S}_C}(X_{a_1}X_{a_2}X_{b_1}X_{b_2}X_{c_1}X_{c_2})$ is realizable by the pDAG $\mathcal{G}_{\text{DropSplitTri}}$ of Figure 5.1.11, where the arrows $\alpha \to a_2$, $\gamma \to c_1$, $\alpha \to b_1$ and $\beta \to c_2$ are dropped. Therefore, we conclude that

$$\mathcal{S}_C \subseteq \mathcal{M}_C(\mathcal{G}_{\text{DropSplitTri}}, (2,2,2,2,2,2)). \tag{5.9}$$

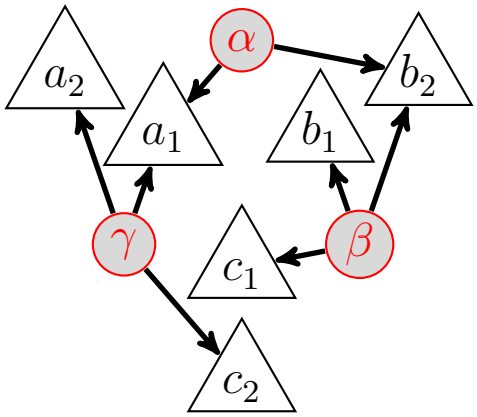


Figure 5.1.11: DAG $\mathcal{G}_{\text{DropSplitTri}}$ obtained by dropping the arrows $\alpha \to a_2$, $\gamma \to c_1$, $\alpha \to b_1$ and $\beta \to c_2$ from $\mathcal{G}_{\text{SplitTri}}$. This pDAG can classically realize all of the distributions of $\mathcal{S}_C$.

The pDAG obtained by starting from $\mathcal{G}_{\text{DropSplitTri}}$ and marginalizing out b_1 and a_2, that is represented in Figure 5.1.12 and will be called $\mathcal{G}_{\text{EqToBell2}}$, is classically observationally equivalent to the Bell pDAG where c_1 and c_2 are settings and a_1 and b_2 are outcomes, as noted in Section 3.6.1. Therefore, we conclude that every distribution that belongs to $\mathcal{S}_C$, that is, every distribution classically realizable by $\mathcal{G}_{\text{SplitTri}}$ that obeys $X_{a_2} = X_{c_2}$ and $X_{b_1} = X_{c_1}$, is such that its marginal $P(X_{a_1}X_{a_2}X_{b_1}X_{c_2})$ is classically realizable by the Bell pDAG.

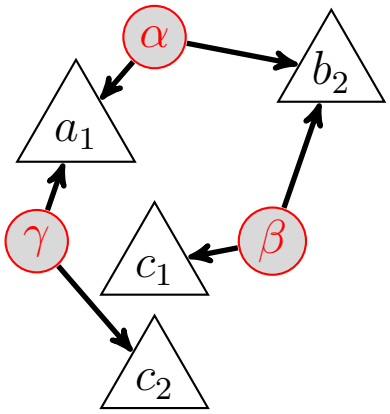


Figure 5.1.12: pDAG $\mathcal{G}_{\text{EqToBell2}}$ obtained by deleting b_1 and a_2 from $\mathcal{G}_{\text{DropSplitTri}}$.

Now, let us look at what can be realized by $\mathcal{G}_{\text{SplitTri}}$ *when the latent node* α *is quantum.*

When α is quantum, $\mathcal{G}_{\text{EqToBell2}}$ is observationally equivalent to the Bell pDAG with a quantum latent node. Therefore, by taking one distribution $P_Q(X_{a_1}X_{b_2}X_{c_1}X_{c_2})$ that violates the CHSH inequality and extending it to a distribution $P_Q(X_{a_1}X_{a_2}X_{b_1}X_{b_2}X_{c_1}X_{c_2})$ by choosing $X_{a_2} = X_{c_2}$ and $X_{b_1} = X_{c_1}$, we obtain a distribution that is realizable by $\mathcal{G}_{\text{DropSplitTri}}$ when α is quantum. Clearly, this distribution is also realizable by $\mathcal{G}_{\text{SplitTri}}$ when α is quantum, since this pDAG includes $\mathcal{G}_{\text{DropSplitTri}}$ as a subgraph.

The distribution $P_Q(X_{a_1}X_{a_2}X_{b_1}X_{b_2}X_{c_1}X_{c_2})$ is *not* classically realizable by $\mathcal{G}_{\text{SplitTri}}$, otherwise it would be an element of $\mathcal{S}_C$ and we already proved that the marginal $P(X_{a_1}X_{b_2}X_{c_1}X_{c_2})$ of every element of $\mathcal{S}_C$ is classically realizable by the Bell pDAG. Therefore, the distribution $P_Q(X_{a_1}X_{a_2}X_{b_1}X_{b_2}X_{c_1}X_{c_2})$ is realizable by $\mathcal{G}_{\text{SplitTri}}$ when α is quantum, but not when all of the latent nodes are classical. Therefore, $\mathcal{G}_{\text{SplitTri}}$ exhibits a QC gap.

As we already argued, the sets of distributions that are classically realizable by $\mathcal{G}_{\text{SplitTri}}$ and by the Triangle pDAG are the same (after making the association of Eqs. (5.6), (5.7) and (5.8)). We can also see that the set of distributions $P(X_{a_1}X_{a_2}X_{b_1}X_{b_2}X_{c_1}X_{c_2})$ over cardinalities $\vec{c}_{\text{vis}} = (2,2,2,2,2,2)$ that are quantumly realizable by $\mathcal{G}_{\text{SplitTri}}$ is *contained* inside the set of distributions $P(X_aX_bX_c)$ over cardinalities $\vec{c}_{\text{vis}} = (4,4,4)$ that are quantumly realizable by the Triangle pDAG, because the Triangle pDAG can reproduce at least all of the protocols that can be realized in $\mathcal{G}_{\text{SplitTri}}$. Therefore, the fact that $\mathcal{G}_{\text{SplitTri}}$ has a QC gap implies that the Triangle pDAG also has a QC gap.

We thus conclude the proof that the Triangle scenario has a QC gap when all of its visible nodes have cardinality 4. Note that we only need *one* of the latent nodes of the Triangle pDAG to go quantum, namely, the node α, to see this QC gap.

5.1.4 The Interruption Piggyback: Ghost Scenario

In this section, we will prove that there is a QC gap in the pDAG $\mathcal{G}_{\text{Ghost}}$ of Fig. 5.1.13, termed the *Ghost scenario*. We will first point out that a QC gap in this scenario can be proven in two ways via techniques already exemplified, namely the Fritz piggyback and the marginalization piggyback. In the first way, we piggyback on the QC gap of the Bell pDAG, and in the second way we piggyback on the QC gap of the Triangle pDAG. Then, we will show a third way to prove that there is a QC gap in $\mathcal{G}_{\text{Ghost}}$, via the *interruption piggyback*. In this third way, we piggyback on the QC gap of the *Evans scenario* (Figure 5.1.14) [33, 50], which is also called *unrelated confounders scenario*.

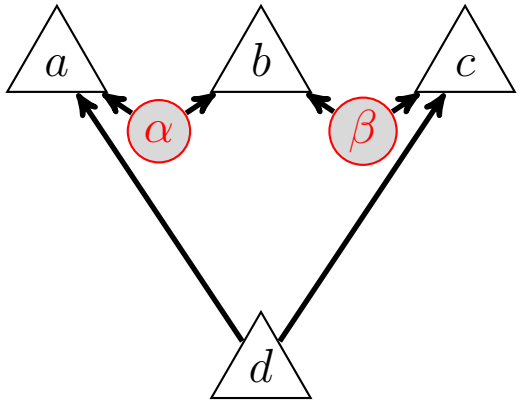


Figure 5.1.13: The Ghost pDAG $\mathcal{G}_{\text{Ghost}}$.

To apply the Fritz piggyback to $\mathcal{G}_{\text{Ghost}}$, we can split b into b_1 and b_2, look at the subset of realizable distributions where $b_2 = c$, and then marginalize out b_2 to obtain the Bell pDAG. This would show that there is a distribution $P(X_a X_b X_c X_d)$ over variables with cardinalities $(c_a, c_b, c_c, c_d) = (2, 4, 2, 2)$ that is realizable by the Ghost pDAG when α is quantum but not realizable when α is classical.

To apply the marginalization piggyback to $\mathcal{G}_{\text{Ghost}}$, we could marginalize out the node d to obtain the Triangle pDAG (Figure 5.1.9). Since it is known that there exist distributions $P(X_a X_b X_c)$ that are not classically realizable by the Triangle pDAG but are realizable when the latent node α is quantum and the other two are classical [39, 14, 28], we can conclude that there are distributions $P(X_a X_b X_c X_d)$ that are realizable by the Ghost pDAG when α is quantum but not realizable when α is classical.

As promised, now we use the Ghost pDAG to exemplify the interruption piggyback. This is a technique that, as noted in Ref. [78], can be used to piggyback on the QC gap of

the Instrumental pDAG to show that there is a QC gap in the Bell pDAG. Because this is a somewhat unnatural order (since historically the gap in the Bell pDAG was proven before the gap in the Instrumental pDAG), here we opt to instead exemplify this technique by piggybacking on the QC gap of the Evans pDAG to show that there is a QC gap in the Ghost pDAG.

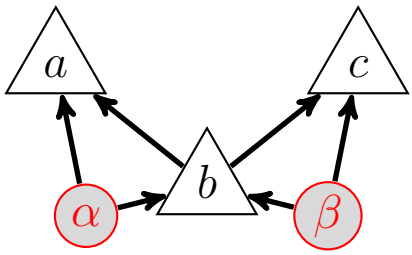


Figure 5.1.14: The Evans pDAG $\mathcal{G}_{\text{Evans}}$.

To start, note that the Ghost pDAG can be obtained from the Evans pDAG by splitting the node b into one node that has no children (b in $\mathcal{G}_{\text{Ghost}}$) and one node that has no parents (d in $\mathcal{G}_{\text{Ghost}}$). We refer to the node-splitting process as *interruption*. Therefore, we say that $\mathcal{G}_{\text{Ghost}}$ can be obtained from $\mathcal{G}_{\text{Evans}}$ by interruption [81].

Let $P^{\text{Ghost}}(X_a X_b X_c X_d) \in \mathcal{M}_C(\mathcal{G}_{\text{Ghost}}, (2,2,2,2))$ be a distribution that is classically realizable by the Ghost pDAG. By looking at the specific elements of this probability distribution where the value assignment of X_b is the same as the value assignment of X_d, we can define a distribution P^{Evans} that is classically realizable by the Evans pDAG via:

$$P^{\text{Evans}}(X_a = x_a, X_b = x_b, X_c = x_c) := \frac{P^{\text{Ghost}}(X_a = x_a, X_b = x_b, X_c = x_c, X_d = x_b)}{\sum_{x_b^*} P^{\text{Ghost}}(X_a = x_a, X_b = x_b^*, X_c = x_c, X_d = x_b^*)}. \tag{5.10}$$

In fact, by the interruption lemma of Ref. [81], a probability distribution P^{Evans} is realizable by $\mathcal{G}_{\text{Evans}}$ under some assignment of theories for the nodes α and β *if and only if* there exists some distribution P^{Ghost} that is realizable by $\mathcal{G}_{\text{Ghost}}$ under the same assignment of theories for α and β, and that reproduces P^{Evans} when restricted to the elements where $X_b = X_d$.

It is known that there is a QC gap in the Evans pDAG [50] when the cardinalities are $(c_a, c_b, c_c) = (2,2,2)$. Therefore, there exists at least one probability distribution P_Q^{Evans} over binary variables that is quantumly but not classically realizable by $\mathcal{G}_{\text{Evans}}$. Therefore,

there exists one distribution P_Q^{Ghost} over binary variables that is quantumly realizable by $\mathcal{G}_{\text{Ghost}}$ and that reproduces P^{Evans} when restricted to the elements where $X_b = X_d$, but there exists *no* distribution that is classically realizable by $\mathcal{G}_{\text{Ghost}}$ and that reproduces P^{Evans} when restricted to the elements where $X_b = X_d$. Therefore, P_Q^{Ghost} is *not* classically realizable by $\mathcal{G}_{\text{Ghost}}$, thus proving that there is a QC gap in the Ghost pDAG when the cardinalities are $(c_a, c_b, c_c) = (2, 2, 2)$.

To summarize, we proved the QC gap of the Ghost pDAG in three ways: i) using the Fritz piggyback on the QC gap of the Bell pDAG to prove a gap when the cardinalities are $(c_a, c_b, c_c, c_d) = (2, 4, 2, 2)$; ii) using the marginalization piggyback on the QC gap of the Triangle pDAG to prove a gap when the cardinalities are $(c_a, c_b, c_c, c_d) = (3, 3, 3, 2)$; and iii) using the interruption piggyback on the QC gap of the Evans pDAG to prove a gap when the cardinalities are $(c_a, c_b, c_c, c_d) = (2, 2, 2, 2)$. Even though the last method is better in the sense that it proves a gap even when the visible nodes are all binary, it is worse in the sense that it requires *both* latent nodes to go quantum, while the two first methods show a gap even when only one of the latent nodes goes quantum.

5.1.5 The Postselection Piggyback: Flag Scenario

In this section, we will show that there is a QC gap in the pDAG $\mathcal{G}_{\text{Flag}}$ of Figure 5.1.15, termed the *Flag scenario*. We will piggyback on the QC gap of the Instrumental pDAG $\mathcal{G}_{\text{Instr}}$ (Figure 5.1.5).

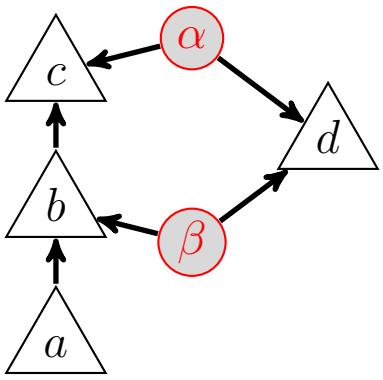


Figure 5.1.15: The Flag pDAG $\mathcal{G}_{\text{Flag}}$.

The technique we will use here will be called the *postselection piggyback*. It can show that there is a QC gap in the Flag scenario when the cardinalities are $(c_a, c_b, c_c, c_d) = (3, 2, 2, 2)$.

The idea behind the postselection piggyback is that conditioning on the value of a node induces correlation between its ancestors. Furthermore, as we will see in more detail in Lemma 18 of Section 5.3, if the node being conditioned upon has no ancestors that are not parents, then postselection on that node is equivalent to having a latent common cause between its parents.

In the case of the Flag pDAG, the node d has no grandparents that are not parents. As a consequence of Propositions 21 and 23[2], every distribution $P_C(X_aX_bX_cX_d) \in \mathcal{M}_C(\mathcal{G}_{\text{Flag}}, (3,2,2,2))$ that is classically realizable by the Flag pDAG has to be such that

$$P_C(X_aX_bX_c|X_d = 0) \in \mathcal{M}_C(\mathcal{G}_{\text{EqToInstr3}}, (3,2,2)), \tag{5.11}$$

where $\mathcal{G}_{\text{EqToInstr3}}$ is the pDAG of Figure 5.1.16, obtained from $\mathcal{G}_{\text{Flag}}$ by adding a latent node pointing to all of the parents of d and then deleting d.

It is easy to see that $\mathcal{G}_{\text{EqToInstr3}}$ is classically observationally equivalent to the Instrumental scenario, by exogenization. Therefore, every distribution $P_C(X_aX_bX_cX_d)$ that is classically realizable by the Flag pDAG needs to be such that $P_C(X_aX_bX_c|X_d = 0)$ is classically realizable by the Instrumental pDAG.

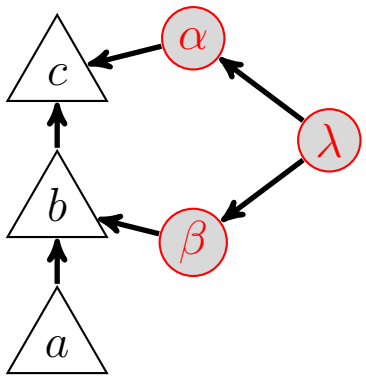


Figure 5.1.16: The pDAG $\mathcal{G}_{\text{EqToInstr3}}$, obtained from $\mathcal{G}_{\text{Flag}}$ by adding a latent parent of the parents of d and then deleting d.

Now, consider a distribution $P_Q(X_aX_bX_c)$ in the QC gap of the Instrumental scenario, that is, a distribution that is quantumly but not classically realizable by the Instrumental scenario. This distribution is generated by assigning a specific quantum state $\sigma \in \mathcal{H}_B \otimes \mathcal{H}_C$ to the latent node of the Instrumental scenario, then sending system $\mathcal{H}_B$ to the node b and

[2]And also as a special case of the converse of Lemma 18. We do not prove this converse here, but we believe it to be true.

system $\mathcal{H}_C$ to the node c, that subsequently make measurements to generate the values of X_b and X_c.

We will now show that this can be reproduced in $\mathcal{G}_{\text{Flag}}$ by postselection on d when both latent nodes α and β are quantum. Choose the quantum system $\mathcal{H}_\alpha = \mathcal{H}_{\alpha_1} \otimes \mathcal{H}_{\alpha_2}$ associated with node α to have dimension $\dim(\mathcal{H}_C)^2$ and the quantum system $\mathcal{H}_\beta = \mathcal{H}_{\beta_1} \otimes \mathcal{H}_{\beta_2}$ associated with node β to have dimension $\dim(\mathcal{H}_B)^2$. The system $\mathcal{H}_{\alpha_1}$, of dimension $\dim(\mathcal{H}_C)$, will be sent to the node c. The system $\mathcal{H}_{\beta_1}$, of dimension $\dim(\mathcal{H}_B)$, will be sent to the node b. The systems $\mathcal{H}_{\beta_2}$ and $\mathcal{H}_{\alpha_2}$, of dimensions $\dim(\mathcal{H}_B)$ and $\dim(\mathcal{H}_C)$ respectively, will be sent to the node d. Choose the systems $\mathcal{H}_\alpha$ and $\mathcal{H}_\beta$ to be in the following maximally entangled states:

$$|\Phi^+\rangle_\alpha = \frac{1}{\sqrt{\dim(\mathcal{H}_C)}} \sum_{i=0}^{\dim(\mathcal{H}_C)-1} |i\rangle_{\alpha_1} \otimes |i\rangle_{\alpha_2}, \tag{5.12}$$

$$|\Phi^+\rangle_\beta = \frac{1}{\sqrt{\dim(\mathcal{H}_B)}} \sum_{j=0}^{\dim(\mathcal{H}_B)-1} |j\rangle_{\beta_1} \otimes |j\rangle_{\beta_2}, \tag{5.13}$$

for choices of basis $\{|i\rangle\}_{\dim(\mathcal{H}_C)}$ for $\mathcal{H}_{\alpha_1}$ and $\mathcal{H}_{\alpha_2}$, and $\{|j\rangle\}_{\dim(\mathcal{H}_B)}$ for $\mathcal{H}_{\beta_1}$ and $\mathcal{H}_{\beta_2}$.

Then, choose the node d to perform a measurement that includes the POVM element $\sigma^\intercal$, where the transpose is defined relative to the bases $\{|i\rangle\}_{\dim(\mathcal{H}_C)}$ and $\{|j\rangle\}_{\dim(\mathcal{H}_B)}$. This is possible because the Hilbert space of the system that is received by d, $\mathcal{H}_{\beta_2} \otimes \mathcal{H}_{\alpha_2}$, is the same as $\mathcal{H}_B \otimes \mathcal{H}_C$. Finally, choose $X_d = 0$ if and only if the measurement gives the outcome associated with the POVM element $\sigma^\intercal$.

With these choices of parameters in $\mathcal{G}_{\text{Flag}}$, the probability of obtaining $X_d = 0$ is

$$\begin{aligned} P(X_d = 0) &= \operatorname{Tr}\left[\left(|\Phi^+\rangle_{\alpha_1\alpha_2}\langle\Phi^+| \otimes |\Phi^+\rangle_{\beta_1\beta_2}\langle\Phi^+| \right) \left(\mathbb{I}_{\beta_1\alpha_1} \otimes \sigma^\intercal_{\beta_2\alpha_2} \right) \right] \\ &= \frac{1}{\dim(\mathcal{H}_B)\dim(\mathcal{H}_C)} \operatorname{Tr}\left(\sigma^\intercal_{\beta_2\alpha_2} \right) = \frac{1}{\dim(\mathcal{H}_B)\dim(\mathcal{H}_C)}, \end{aligned} \tag{5.14}$$

and the updated state $\rho_{\beta_1\alpha_1} \in \mathcal{H}_{\beta_1} \otimes \mathcal{H}_{\beta_2}$ that is sent to b and c after conditioning on $X_d = 0$ is:

$$\begin{aligned} &\rho_{\beta_1\alpha_1} = \\ &\frac{1}{P(X_d = 0)} \operatorname{Tr}_{\beta_2\alpha_2}\left[\left(|\Phi^+\rangle_{\alpha_1\alpha_2}\langle\Phi^+| \otimes |\Phi^+\rangle_{\beta_1\beta_2}\langle\Phi^+| \right) \right. \\ &\qquad \left. \left(\mathbb{I}_{\beta_1\alpha_1} \otimes \sigma^\intercal_{\beta_2\alpha_2} \right) \right]. \end{aligned} \tag{5.15}$$

Now, we use the well-known identity

$$\mathrm{Tr}_B\left[\left(|\Phi^+\rangle_{AB}\langle\Phi^+|\right)(\mathbb{I}_A \otimes X_B)\right] = \frac{1}{d}X^\intercal, \tag{5.16}$$

where $d = \dim(A) = \dim(B)$. Applying this identity twice, Equation (5.15) becomes:

$$\rho_{\beta_1\alpha_1} = \sigma. \tag{5.17}$$

Therefore, with entanglement swapping and with the appropriate choices of states and measurements in $\mathcal{G}_{\text{Flag}}$, it is possible to simulate any state σ shared between b and c. Then, b and c just need to make the same measurements as in the realization of $P_Q(X_aX_bX_c)$ in the Instrumental scenario, and we can obtain a distribution $\tilde{P}_Q(X_aX_bX_cX_d) \in \mathcal{M}_Q(\mathcal{G}_{\text{Flag}}, (3,2,2,2))$ such that

$$\tilde{P}_Q(X_aX_bX_c|X_d = 0) = P_Q(X_aX_bX_c). \tag{5.18}$$

The distribution $\tilde{P}_Q(X_aX_bX_cX_d)$ is not classically realizable by the Flag scenario, because otherwise $\tilde{P}_Q(X_aX_bX_c|X_d = 0)$ would be classically realizable by the Instrumental scenario (Equation (5.11)).

Note that while the QC gaps that we have shown for the Triangle and the Ghost scenarios only need one of the latent nodes to go quantum, the QC gap that we have just proven for the Flag scenario requires both latent nodes to go quantum.

5.2 Formalization of the piggybacks

Now, we will formulate the generic conditions under which the piggybacks exemplified in the previous section can be applied. Here we only present the statements of these generic results; examples were already provided, and all of the proofs are found in Appendix D.1.

Proposition 25 (Point Distribution Piggyback)**.** *Let $\mathcal{G}_{new}$ be a pDAG, and let $\mathcal{G}_{pig}$ be the pDAG obtained by deleting a visible node $v \in \mathtt{vis}(\mathcal{G}_{new})$ from $\mathcal{G}_{new}$. If $\mathcal{G}_{pig}$ has a QC gap, then $\mathcal{G}_{new}$ also has a QC gap.*

Proposition 26 (Marginalization Piggyback)**.** *Let $\mathcal{G}_{new}$ be a pDAG, and let $\mathcal{G}_{pig}$ be the pDAG obtained by transforming a visible node $d \in \mathtt{vis}(\mathcal{G}_{new})$ from $\mathcal{G}_{new}$ into a latent node. That is, $d \in \mathtt{lat}(\mathcal{G}_{pig})$. If there is a gap between $\mathcal{G}_{pig}$ when all latent nodes are classical and $\mathcal{G}_{pig}$ when d is classical but all of the other latent nodes are quantum, then $\mathcal{G}_{new}$ has a QC gap.*

Note that this formulation of the marginalization piggyback includes both cases like the Pike scenario, where teleportation was necessary, and cases like the Bell-with-broken-leg scenario and Charlie Brown Tree scenario, where it was not. This is because this difference lies in the step of checking whether there is a gap between $\mathcal{G}_{\text{pig}}$ when all latent nodes are classical and $\mathcal{G}_{\text{pig}}$ when λ is classical but all of the other latent nodes are quantum. This step will be taken care of in Section 5.3.

Proposition 27 (Fritz Piggyback)**.** *Let $\mathcal{G}_{new}$ be a pDAG, and let $a, b \in \texttt{vis}(\mathcal{G}_{new})$ be two visible nodes of this pDAG such that either a is a parent of b or a and b share a common latent parent (or both). Let $p \in \texttt{pa}_{\mathcal{G}_{new}}(b)$ be a parent of b (not necessarily visible) such that $\mathcal{G}_{new}$ exhibits the d-separation relation $a \perp p | \texttt{pa}_{\mathcal{G}_{new}}(b) \setminus \{p\}$. Let $\mathcal{G}_{pig}$ be the pDAG obtained by adding two visible nodes a' and b' to $\mathcal{G}_{new}$ such that $\texttt{ch}_{\mathcal{G}_{pig}}(a') = \texttt{ch}_{\mathcal{G}_{new}}(a)$, $\texttt{pa}_{\mathcal{G}_{pig}}(a') = \texttt{pa}_{\mathcal{G}_{new}}(a)$, $\texttt{ch}_{\mathcal{G}_{pig}}(b') = \texttt{ch}_{\mathcal{G}_{new}}(b)$, $\texttt{pa}_{\mathcal{G}_{pig}}(b') = \texttt{pa}_{\mathcal{G}_{new}}(b) \setminus \{p\}$, adding a common latent parent between a and a' if it is not already present and adding a common latent parent between b and b' if it is not already present. If $\mathcal{G}_{pig}$ has a QC gap that includes distributions where $X_{a'}$ and $X_{b'}$ are perfectly correlated, then $\mathcal{G}_{new}$ also has a QC gap.*

The application of the Fritz piggyback can be tricky in practice: to conclude that $\mathcal{G}_{\text{new}}$ has a QC gap, we need to guarantee that $\mathcal{G}_{\text{pig}}$ has a QC gap *that includes distributions where $X_{a'}$ and $X_{b'}$ are perfectly correlated*. But when we algorithmically apply our piggybacks in Section 5.4, our starting point will be the knowledge that a small set of pDAGs have QC gaps, without further information about the nature of these QC gaps. There are, however, cases where this starting point still allows us to conclude that $\mathcal{G}_{\text{pig}}$ has a QC gap that satisfies the perfect correlation requirement. One example is given by the proof of the the QC Gap of the Triangle pDAG from Section 5.1.3.

In Section 5.1.3, we implicitly used the Fritz piggyback twice, as we will make explicit now. Starting from $\mathcal{G}_{\text{Tri}}$ of Fig. 5.1.9 and noting that it has the d-separation relation $b \perp \gamma | \beta$, Proposition 27 constructs a pDAG $\mathcal{G}_{\text{pig}}$ by adding to it one node b_1 that has parents α and β, and one node c_1 that only has the parent β. Then, noting that $\mathcal{G}_{\text{pig}}$ has the d-separation relation $a \perp \beta | \gamma$, Proposition 27 constructs a pDAG $\mathcal{G}'_{\text{pig}}$ by adding to it one node a_2 that has parents α and γ, and one node c_2 that only has the parent γ. If $\mathcal{G}'_{\text{pig}}$ has a QC gap that includes distributions where X_{b_1} and X_{c_1} are perfectly correlated at the same time that X_{a_2} and X_{c_2} are perfectly correlated, then $\mathcal{G}_{\text{Tri}}$ also has a QC gap.

To see that this is indeed the case, note that the nodes b_1 and a_2 do not have any children in $\mathcal{G}'_{\text{pig}}$. Therefore, if we marginalize out the variables X_{b_1} and X_{a_2}, the corresponding nodes can be removed from the pDAG (because latent nodes that do not have any children can be removed). After also removing the node c (either by marginalizing the corresponding

variable or setting it to a point distribution), we obtain the pDAG $\mathcal{G}_{\text{EqToBell2}}$ of Fig. 5.1.12, which is known to have a QC gap. We can then just choose one distribution in this gap and extend it to include X_{b_1} and X_{a_2} via the perfect correlation conditions (and also extend it to include X_c in any way we desire), and we will obtain one distribution that lives in the QC gap of $\mathcal{G}'_{\text{pig}}$ and satisfies the perfectly correlation conditions required.

The important property that allowed us to reach this conclusion was the fact that b_1 and a_2 do not have any children. If the variable that plays the role of a in Proposition 27 has children, then the node a' cannot be removed from the graph by transforming it into a latent node, so it might not be possible to conclude that $\mathcal{G}_{\text{pig}}$ has a QC gap that includes distributions where $X_{a'}$ and $X_{b'}$ are perfectly correlated.

Proposition 28 (Interruption Piggyback)**.** *Let $\mathcal{G}_{new}$ be a pDAG, let $c \in \texttt{vis}(\mathcal{G}_{new})$ be a visible node of $\mathcal{G}_{new}$ that has no parents and $p \in \texttt{vis}(\mathcal{G}_{new})$ be a visible node of $\mathcal{G}_{new}$ that has no children, that are such that p is not an ancestor of c in $\mathcal{G}_{new}$. Let $\mathcal{G}_{pig}$ be the pDAG obtained by deleting c and p from $\mathcal{G}_{new}$, and adding a visible node $b \in \texttt{vis}(\mathcal{G}_{pig})$ such that $\texttt{pa}_{\mathcal{G}_{pig}}(b) = \texttt{pa}_{\mathcal{G}_{new}}(c)$ and $\texttt{ch}_{\mathcal{G}_{pig}}(b) = \texttt{ch}_{\mathcal{G}_{new}}(p)$. If $\mathcal{G}_{pig}$ has a QC gap, then $\mathcal{G}_{new}$ also has a QC gap.*

Proposition 29 (Postselection Piggyback)**.** *Let $\mathcal{G}_{new}$ be a pDAG, and let $\mathcal{G}_{pig}$ be the S-pDAG obtained by transforming a visible node v from $\mathcal{G}_{new}$ into a selected node. If $\mathcal{G}_{pig}$ has a QC gap, then $\mathcal{G}_{new}$ also has a QC gap.*

Note that this formulation of the Postselection piggyback assumes one already knows that the S-pDAG $\mathcal{G}_{\text{pig}}$ has a QC gap. When showing the QC gap of the Flag scenario in Section 5.1.5, we had to prove it. This step will be discussed in generality in Section 5.3.

5.3 Observational dominance in the quantum case

After applying one of the piggybacks, we need to check whether the resulting graph $\mathcal{G}_{\text{pig}}$ (that is either a pDAG or an S-pDAG) has a known QC gap or not. As we saw in Section 5.1, many times the graph obtained after piggybacking is *not* immediately one of the pDAGs with known quantum classical gaps. For example, after applying the Fritz piggyback on the Triangle scenario in Section 5.1.3, we obtain the pDAG $\mathcal{G}_{\text{EqToBell2}}$ of Figure 5.1.12, which at first sight is not one of the pDAGs with a known QC gap. One then needs to go through the extra step of showing that $\mathcal{G}_{\text{EqToBell2}}$ is classically observationally equivalent to the Bell pDAG, and that when α is quantum and the other two latent nodes are classical it is observationally equivalent to the quantum Bell pDAG.

The step of showing the *classical* observational equivalence of $\mathcal{G}_{\text{pig}}$ with a pDAG $\mathcal{G}_{\text{known}}$ that has a known QC gap can be done via observational equivalence rules from classical causal inference. In the case where $\mathcal{G}_{\text{pig}}$ is a pDAG, all of the known such rules are compiled in Chapter 3. In the case where $\mathcal{G}_{\text{pig}}$ is an S-pDAG, Chapter 4 has shown a series of observational equivalence rules. As we already stressed, however, in general those rules are *not* valid to show observational equivalence when one or more of the latent nodes is quantum.

In reality, for our purposes we do not need to be able to show observational *equivalence* of $\mathcal{G}_{\text{pig}}$ and $\mathcal{G}_{\text{known}}$ in the quantum case; instead, observational *dominance* of $\mathcal{G}_{\text{pig}}$ over $\mathcal{G}_{\text{known}}$ suffices. For example, in the discussion of Section 5.1.2 about the Pike scenario, we only showed that $\mathcal{G}_{\text{EqToInstr2}}$ when λ is classical and α and β are quantum can realize all of the distributions that are quantumly realizable by the Instrumental scenario, and not the other way around.[3] This is sufficient because we only need to know that $\mathcal{G}_{\text{pig}}$ has a QC gap.

Thus, in this section we will present observational dominance rules that are valid when one or more of the latent nodes of a pDAG or S-pDAG are quantum. Some of these rules also show observational equivalence, but not all of them. Since for us these are just steps to help showing more QC gaps, we will present all of these rules as lemmas. However, they are interesting in themselves, as generalizations of some of the results of Chapters 3 and 4.

The first rule we will use is a slight generalization of the HLP edge-adding rule from Section 3.3.1:

Lemma 15 (HLP edge-adding rule for classical parents)**.** *Let $\mathcal{G}$ be a pDAG, and let $a, b \in \texttt{vis}(\mathcal{G})$ be two of its visible nodes. Let $\mathcal{G}'$ be the pDAG obtained by adding a directed edge $a \to b$ to $\mathcal{G}$. Suppose that $\texttt{pa}_{\mathcal{G}}(a) \subseteq \texttt{pa}_{\mathcal{G}}(b)$, and that a and b share at least one latent parent. Then, the pDAG $\mathcal{G}$ under a semantics where all of the parents of a are classical is observationally equivalent to the pDAG $\mathcal{G}'$ under the same semantics.*

Proof. Since all of the parents of a are classical, the variable X_a can be obtained from a deterministic function $X_a = f(X_{\texttt{pa}_{\mathcal{G}}(a)}, E_a)$, where E_a is an error variable. This error variable can be absorbed into one of the latent parents that a and b share. Therefore, b has access to $\texttt{pa}_{\mathcal{G}}(a)$ and E_a, and thus it can infer the value of X_a even without direct access to the node a. □

This rule was used when showing the QC gap of the Bell-with-broken-leg scenario in Section 5.1.2, to prove that the pDAG $\mathcal{G}_{\text{EqToBell1}}$ of Fig. 5.1.3 with quantum γ and classical

[3] Although the other way around also holds in this specific case, this is not relevant for our proof.

λ can realize any distribution that is quantumly realizable by the Bell pDAG. It was also used when showing the QC gap of the Triangle scenario in Section 5.1.3, to prove that the pDAG $\mathcal{G}_{\text{EqToBell2}}$ of Fig. 5.1.12 with quantum α and classical β and γ can realize any distribution that is quantumly realizable by the Bell pDAG.

The next rule we will use was already noted in Ref. [17]:

Lemma 16 (Exogenization for classical parents)**.** *Let $\mathcal{G}$ be a pDAG, and let $\lambda \in \texttt{lat}(\mathcal{G})$ be one of its latent nodes. Let $\mathcal{G}'$ be the pDAG obtained by exogenizing λ. The pDAG $\mathcal{G}$ under a semantics where all of the parents of λ are classical is observationally equivalent to the pDAG $\mathcal{G}'$ under the same semantics.*

Proof. Taken from Ref. [17]. □

This rule was used when showing the QC gap of the Charlie Brown Tree scenario in Section 5.1.2, to show that $\mathcal{G}_{\text{EqToInstr1}}$ when γ is quantum and μ and λ are classical is observationally equivalent to the quantum Instrumental pDAG.

Lemma 17 (Observational equivalence via teleportation)**.** *Let $\mathcal{G}$ be a pDAG, and let $\lambda \in \texttt{lat}(\mathcal{G})$ be one of its latent nodes. Associate to $\mathcal{G}$ a semantics $\mathscr{T}$ where λ is classical. Let $\texttt{Qpa}_{\mathcal{G}}(\lambda)$ denote the set of quantum parents of λ in $\mathcal{G}$ under $\mathscr{T}$, and $\texttt{Cpa}_{\mathcal{G}}(\lambda)$ denote the set of classical parents of λ in $\mathcal{G}$ under $\mathscr{T}$. Let $T = \texttt{ch}_{\mathcal{G}}(\lambda) \cap \texttt{ch}_{\mathcal{G}}(\texttt{Qpa}_{\mathcal{G}}(\lambda))$ denote the set of children of λ that share a quantum parent with λ.*

Let $\mathcal{G}'$ be the pDAG obtained by

1. *Adding one arrow from every node in $\texttt{Qpa}_{\mathcal{G}}(\lambda)$ to every node in T, and*
2. *Adding one arrow from every node in $\texttt{Cpa}_{\mathcal{G}}(\lambda)$ to every node in $\texttt{ch}_{\mathcal{G}}(\lambda)$.*

Then, under $\mathscr{T}$, $\mathcal{G}$ is observationally equivalent to $\mathcal{G}'$.

Proof. Given in Appendix D.2. □

This rule was used when proving the QC gap of the Pike scenario in Section 5.1.2, to show that $\mathcal{G}_{\text{EqToInstr2}}$ when λ is classical and α and β are quantum can realize all of the distributions that are quantumly realizable by the Instrumental pDAG.

Now, we present a rule that involves deleting one selected node from an S-pDAG, and adding new latent nodes. We will present it as an observational dominance rule, because this is the only direction needed here. However, we believe that the converse is also true, i.e., we believe this is an observational *equivalence* result, in which case it would be a generalization of Propositions 21 and 23.

Lemma 18 (Observational dominance via entanglement swapping)**.** *Let $\mathcal{G}$ be an S-pDAG, and let $s \in \mathtt{sel}(\mathcal{G})$ be one of its selected nodes. Associate to $\mathcal{G}$ a semantics $\mathscr{T}$, and let $\mathtt{Qpa}_{\mathcal{G}}(s)$ denote the set of quantum parents of s in $\mathscr{T}$.*

Let $\mathcal{G}'$ be the S-pDAG obtained by

1. *Adding one latent node α that has no parents and is a parent of all of the nodes in $\mathtt{Qpa}_{\mathcal{G}}(s)$,*
2. *Adding one latent node β that has no parents and is a parent of all of the nodes in $\mathtt{pa}_{\mathcal{G}}(s)$, and*
3. *Deleting s.*

Suppose that the node s has no ancestors that are not its parents in $\mathcal{G}$. Then, $\mathcal{G}$ under $\mathscr{T}$ observationally dominates $\mathcal{G}'$ when α is quantum, β is classical and the remaining latent nodes follow $\mathscr{T}$.

Proof. Given in Appendix D.2. □

This rule was used when showing the QC gap of the Flag scenario in Section 5.1.5, to show that the S-pDAG obtained by starting from the Flag pDAG and transforming d into a selected node quantumly observationally dominates the pDAG $\mathcal{G}_{\text{EqToInstr3}}$.

The next rule was also noted in Ref. [17]:

Lemma 19 (Intermediate latent node that has only one child or only one parent)**.** *Let $\mathcal{G}$ be a pDAG, and let $\lambda \in \mathtt{lat}(\mathcal{G})$ be one of its latent nodes. Let $\mathcal{G}'$ be the pDAG obtained by exogenizing λ. If λ has only one child or only one parent in $\mathcal{G}$, then $\mathcal{G}$ and $\mathcal{G}'$ are observationally equivalent under any semantics.*

Proof. Taken from Ref. [17]. □

This rule was used when showing the QC gap of the Flag scenario in Section 5.1.5, to show that the pDAG $\mathcal{G}_{\text{EqToInstr3}}$ is quantumly observationally equivalent to the Instrumental pDAG.

Lastly, we present the rule mentioned in the introduction to this chapter, that justifies our choice of starting from a set of mDAGs when starting the search for new QC gaps:

Lemma 20. *Let $\mathcal{G}$ be a* RE*-reduced pDAG, and let $\mathcal{G}'$ be another pDAG that is associated with the same mDAG as $\mathcal{G}$, i.e., such that* $\texttt{ToFaces}(\mathcal{G}) = \texttt{ToFaces}(\mathcal{G}')$*. Then, $\mathcal{G}'$ quantumly observationally dominates $\mathcal{G}$.*

Proof. RemoveRedund only removes nodes from a pDAG, so it is clear that $\mathcal{G}'$ observationally dominates $\texttt{RemoveRedund}(\mathcal{G}')$. Let $\alpha \in \texttt{lat}(\mathcal{G}')$ be a latent node that *has parents* in $\mathcal{G}'$. Take a distribution that is quantumly realizable by $\texttt{Exog}(\mathcal{G}')$. In the parameters that realize this distribution, certain classical variables or quantum systems are sent directly from $\texttt{pa}_{\mathcal{G}'}(\alpha)$ to $\texttt{ch}_{\mathcal{G}'}(\alpha)$. To reproduce this in $\mathcal{G}'$, it suffices to send these classical variables or quantum systems to $\texttt{ch}_{\mathcal{G}'}(\alpha)$ indirectly, by passing through α. Therefore, $\mathcal{G}'$ observationally dominates $\texttt{Exog}(\mathcal{G}')$. Together, these two facts imply that $\mathcal{G}$ observationally dominates $\mathcal{G} = \texttt{RemoveRedund} \circ \texttt{Exog}(\mathcal{G}')$. □

5.4 New QC gaps in causal structures with four visible nodes

In this section we describe how we used our piggybacks algorithmically to, starting from a small number of causal structures that have known QC gaps (Bell, Triangle, Instrumental and Evans), show the existence of QC gaps in a large number of causal structures with four visible nodes. As argued in the introduction to this chapter and formally justified through Lemma 20, every time we prove that the canonical pDAG of an mDAG has a QC gap, this implies that all of the other pDAGs associated with that same mDAG also have a QC gap. When this happens, we will say that the mDAG itself has a QC gap.

As shown in the first line of Table 3.3, the total number of 4-node mDAGs consistent with a fixed nodal ordering is 7296. This is a large number, but we do not need to inspect all of these mDAGs in our search: those that are proven to be algebraic can be filtered out, since they cannot present a QC gap. As it turns out, 4489 of these mDAGs are known to be classically observationally equivalent to some latent-confounder-free mDAG, and thus are algebraic. Therefore, there are 2807 4-node mDAGs consistent with a fixed nodal ordering that are *not* proven to be algebraic, and as such still have the potential to present a QC gap. Ref. [48] proves that almost all of these mDAGs *are* in fact nonalgebraic; this reference only leaves three mDAGs as unresolved, which are exactly the mDAGs (e)-(g) of Figure 3.6.2. Because we still do not know if these three are algebraic or not, we will continue referring to our list of mDAGs of interest as *not proved to be algebraic*. We *already* know that 48 of the 4-node mDAGs consistent with a fixed nodal ordering that

are not proved to be algebraic *have* QC gaps: these are the mDAGs in the same classical observational equivalence classes as the Bell mDAG (Section 3.6.1).[4] Therefore, we do not need to check those.

Note that the point distribution piggyback (Proposition 25), the interruption piggyback (Proposition 28), the marginalization piggyback (Proposition 26) and the postselection piggyback (Proposition 29) always prove the presence of a QC gap on a pDAG $\mathcal{G}_{\text{new}}$ by piggybacking on a pDAG $\mathcal{G}_{\text{pig}}$ that has a number of visible nodes *smaller* than the number of visible nodes of $\mathcal{G}_{\text{new}}$. The Fritz piggyback (Proposition 27), on the other hand, can use a pDAG $\mathcal{G}_{\text{pig}}$ that has a larger number of visible nodes to prove a QC gap on a pDAG $\mathcal{G}_{\text{new}}$ that has a smaller number of visible nodes: for example, it piggybacks on the Bell pDAG to prove that the Triangle pDAG has a QC gap.

Because of that, our algorithm broadly works as follows. For each candidate 4-node mDAG, we first apply each of the four first piggybacks (point distribution, interruption, marginalization and postselection) in all of the possible ways (for example, for the point distribution piggyback, we generate one different resulting mDAG for each visible node that could be deleted). In each step, if the resulting mDAG is in our list of 3-node mDAGs for which it is known that there is a QC gap, then the original candidate mDAG also has a QC gap. This list consists in mDAGs obtained from all of the possible relabellings of the nodes of the Instrumental, Triangle and Evans scenarios. If the candidate mDAG is still not proven to have a QC gap, we proceed to applying the Fritz piggyback. We then check if any of the mDAGs resulting from applying Fritz is in our *larger* list of mDAGs for which it is known that there is a QC gap: this list now also includes mDAGs obtained from all of the possible relabellings of the nodes of all of the 4-node mDAGs that were proven to have QC gaps in the previous step (plus the mDAGs obtained from all of the possible relabellings of the nodes of the mDAGs that are observationally equivalent to the Bell mDAG, which we did not check in the previous step because we already knew they have a QC gap).[5]

[4] As noted before, the fact that they are in the same classical observational equivalence class as an mDAG that has a QC gap does not imply that they have QC gaps. However, for the special case of the mDAGs in the classical observational equivalence class of Bell, it is known that all present QC gaps.

[5] Note that the Fritz piggyback as formalized in Proposition 27 yields a $\mathcal{G}_{\text{pig}}$ whose number of nodes is *larger* than the number of nodes of $\mathcal{G}_{\text{new}}$, i.e., in this case it would yield mDAGs with more than 4 nodes. To then obtain mDAGs with 3 or 4 nodes and be able to compare them to the list of mDAGs for which it is known that there is a QC gap, we must subsequently apply another piggyback. For example, in the proof of the QC gap of the Triangle pDAG in Section 5.1.3, we marginalized out b_1 and a_2. It is in general not so trivial to apply those subsequent piggybacks: for example, as discussed below Proposition 27, marginalizing out a variable X_a that is perfectly correlated to another variable can give problems if the node a has children. This was taken into account in our implementation.

With this implementation, we obtain the numerical results presented in Table 5.1. There, we see that most of the 4-node mDAGs that *can* present a QC gap (by virtue of not yet being proven to be algebraic) in fact *do* present a QC gap!

Piggybacks	Known QC gaps	Remaining mDAGs *not* provably-algebraic
None	48	2759
PD	2424	383
PD+Interr	2428	379
PD+Interr+Marg	2517	290
PD+Interr+Marg+Post	2582	225
PD+Interr+Marg+Post+Fritz	2587	220

Table 5.1: Results obtained from applying the techniques developed in this chapter to 4-node mDAGs consistent with a fixed nodal ordering. The first column indicates the piggybacks used: PD is the point distribution piggyback (Proposition 25), Interr is the interruption piggyback (Proposition 28), Marg is the marginalization piggyback (Proposition 26), Post is the postselection piggyback (Proposition 29) and Fritz is the Fritz piggyback (Proposition 27). The second column indicates the number of 4-node mDAGs consistent with a fixed nodal ordering that are known to have a QC gap up to that point, and the third column indicates the number of remaining 4-node mDAGs consistent with a fixed nodal ordering that were not proven to be algebraic. The sum of the two numbers on a line is always 2807, which is the total number of 4-node mDAGs consistent with a fixed nodal ordering that are not proven to be algebraic.

As we see in this table, most of the mDAGs are already shown to have a QC gap by the point distribution piggyback, that is, they can be reduced to one of the Triangle, Instrumental or Evans scenarios by the simple deletion of one visible node. One might feel that this operation is too trivial, and that causal structures whose QC gap is *not* shown by the point distribution piggyback are more interesting. One example of an mDAG that is proven to have a QC gap by the Fritz piggyback (and none of the previous piggybacks) is the Tetrahedron scenario, which is the mDAG over nodes a, b, c, d that has no directed edges between nodes and whose simplicial complex has facets $\{a, b, c\}$, $\{a, b, d\}$, $\{a, c, d\}$, and $\{b, c, d\}$.

Remember that here we looked at nonalgebraic mDAGs *compatible with a fixed nodal ordering*. Many of those mDAGs might be obtained from other mDAGs in the same list by relabelling of the nodes. For example, if you look at the only mDAG of the class `Instrumental BAC` of Figure 3.2.1, you see that there is one mDAG that is equal to it up to relabelling of the nodes in the class `Instrumental CAB`, and another one in the class `Instrumental ABC`. To have a clearer idea of how many different causal structures are

now known to have QC gaps, it might be interesting to look at how many QC gaps are now known *up to relabelling of the nodes*. The 2807 4-node mDAGs consistent with a fixed nodal ordering that are not proved to be algebraic organize into 996 groups of mDAGs that are equal to each other up to relabelling of the nodes. The 48 4-node mDAGs consistent with a fixed nodal ordering that were *initially* known to present a QC gap organize into 6 groups of mDAGs that are equal to each other up to relabelling of the nodes. These are shown in Figure 5.4.1, where each such group is represented by an mDAG without node labels. The 2587 4-node mDAGs consistent with a fixed nodal ordering that are now known to present a QC gap organize into 926 groups of mDAGs that are equal to each other up to relabelling of the nodes. The 220 remaining mDAGs organize into 70 groups of mDAGs that are equal to each other up to relabelling of the nodes.

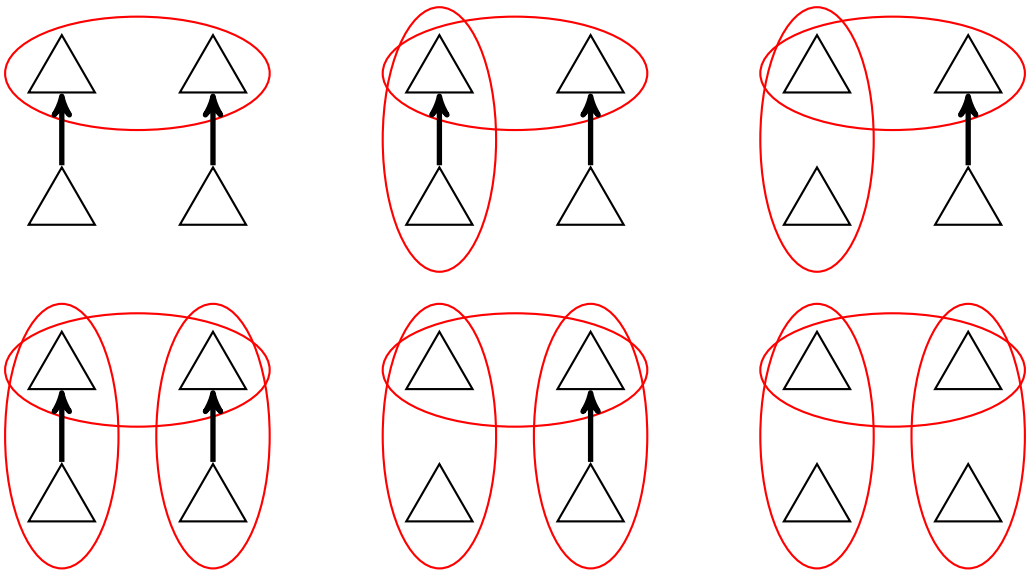


Figure 5.4.1: Before this work, these were all of the 4-node mDAGs (up to relabelling of the nodes) that were known to have a QC gap. Now, the number of 4-node mDAGs (up to relabelling of the nodes) that are known to have a QC gap is 926.

5.5 Conclusion

In this chapter, we presented a series of techniques to show that a causal structure has a QC gap, by piggybacking on the fact that another causal structure is known to have a QC

gap. Applying those techniques to 4-node mDAGs that are not proven to be algebraic, we see that the *majority* of them do present a QC gap. Thus, QC gaps seem to be a prevalent property of causal structures, and *most* of the causal structures that are known to present inequality constraints also present QC gaps.

The most obvious follow-up question to this work is the status of the remaining 4-node mDAGs that are not proven to be algebraic and also not proven to present QC gaps. Those are 220 mDAGs consistent with a fixed nodal ordering, that organize into 70 groups of mDAGs that are equal to each other up to relabelling of the nodes.[6] While we were working on this project, Ref. [49] proved that one of these mDAGs (and of course all of the mDAGs that are equal to it by relabelling of the nodes) does present a QC gap.

It is unknown whether there are causal structures that have inequality constraints but do not have a QC gap for any cardinality of the visible variables. The Instrumental scenario (Figure 2.1.1) is an example where there are inequality constraints for all choices of cardinalities for the visible variables, but where there is no QC gap if all of the visible variables are binary. It *does*, however, present a QC gap when X_a has cardinality 3 (and X_b and X_c are binary). For 3-node mDAGs (Figure 3.2.1), we know that all of the causal structures that present inequalities also present a QC gap. If one wants to search for a nonalgebraic 4-node mDAG that does not have a QC gap, one must look at the 220 mDAGs that were left unresolved here (minus the cases solved by Ref. [49]).

If such a thing exists, another question one can ask is whether there exists any pDAG $\mathcal{G}$ that has a QC gap even though `RE-reduce`$(\mathcal{G})$ does not have a QC gap. In this case, `RE-reduce`$(\mathcal{G})$ would in particular be a nonalgebraic pDAG that does not have a QC gap. As explained before, this is logically conceivable: there are cases where $\mathcal{G}$ and `RE-reduce`$(\mathcal{G})$ are quantumly observationally inequivalent, such as the example presented in Figure 5.0.2. However, we do not currently have any explicit example of this phenomena. If these do exist, then the number of causal structures with 4 visible nodes that present QC gaps might be even larger than indicated here, and this could have interesting consequences for the study of intermediate latent nodes [17].

Another follow-up project one could pursue is the further development of observational dominance rules such as the ones presented in Section 5.3, with the goal of constructing the observational partial order in the quantum case. Apart from sharing the same motivations presented in Chapter 3, this would also help future inquiries about which causal structures present a QC gap: for example, if two causal structures are shown to be observationally equivalent both classically *and* quantumly, then knowing that one has a QC gap clearly implies that the other also has a QC gap. More generally, comparing the classical

[6] As already mentioned, the majority of these are known to be nonalgebraic [48].

and quantum observational partial orders could teach us fundamental lessons about the difference between classical and quantum causality.

Our work shows that QC gaps are common in the landscape of causal structures, but it does not give details about the nature of these gaps. A fundamental question that can be posed about this work is then what do we learn from these new QC gaps. That is, will there be conceptual questions or practical tasks where knowing that one of these new causal structures exhibits a QC gap leads to better conclusions than those already implied by the previously known QC gaps? Since all of our QC gaps were proven by *piggybacking* on QC gaps that were already known, one might question this. In other words, one might question whether these QC gaps capture properties that are intrinsic to the new causal structures, or if they are just "old QC gaps in disguise". Steps towards formalizing what this notion even means were taken in Ref. [22], but the relationship between those proposed definitions and the new QC gaps that we showed here will be left for future investigation.

Data Availability

The algorithm implemented for this chapter is available at the file `proving_QC_Gaps.py` of https://github.com/eliewolfe/mDAG-analysis

References

[1] R. Ayesha Ali, Thomas S. Richardson, Peter Spirtes, and Jiji Zhang. Towards characterizing markov equivalence classes for directed acyclic graphs with latent variables. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, UAI'05, page 10–17, Arlington, Virginia, USA, 2005. AUAI Press.

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

[3] Marina Maciel Ansanelli, Elie Wolfe, and Robert W. Spekkens. Everything that can be learned about a causal structure with latent variables by observational and interventional probing schemes, 2024.

[4] Marina Maciel Ansanelli, Elie Wolfe, and Robert W. Spekkens. The observational partial order of causal structures with latent variables, 2025.

[5] Marina Maciel Ansanelli, Beata Zjawin, David Schmid, Yìlè Yīng, John H. Selby, Ciarán M. Gilligan-Lee, Ana Belén Sainz, and Robert W. Spekkens. The resource theory of causal influence and knowledge of causal influence, 2025.

[6] Angelos P. Armen and Robin J. Evans. Towards characterising Bayesian network models under selection. *arXiv preprint arXiv:1811.05530*, 2018.

[7] Jonathan Barrett, Robin Lorenz, and Ognyan Oreshkov. Quantum causal models, 2019.

[8] J. S. Bell. On the Einstein Podolsky Rosen paradox. *Physics Physique Fizika*, 1:195–200, Nov 1964.

[9] John Bell. Free variables and local causality. *Epistemological Letters*, 15, 1977.

[10] Charles H. Bennett, Gilles Brassard, Claude Crépeau, Richard Jozsa, Asher Peres, and William K. Wootters. Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels. *Phys. Rev. Lett.*, 70:1895–1899, Mar 1993.

[11] Blai Bonet. Instrumentality tests revisited. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, page 48–55, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[12] Kok-Wei Bong, Aníbal Utreras-Alarcón, Farzad Ghafari, Yeong-Cherng Liang, Nora Tischler, Eric G. Cavalcanti, Geoff J. Pryde, and Howard M. Wiseman. A Strong No-Go Theorem on the Wigner's Friend Paradox. *Nature Physics*, 16(12):1199–1205, 2020.

[13] Stephan Bongers, Patrick Forré, Jonas Peters, and Joris M Mooij. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):2885–2915, 2021.

[14] Sadra Boreiri, Antoine Girardin, Bora Ulu, Patryk Lipka-Bartosik, Nicolas Brunner, and Pavel Sekatski. Towards a minimal example of quantum nonlocality without inputs. *Phys. Rev. A*, 107:062413, Jun 2023.

[15] Cyril Branciard, Denis Rosset, Nicolas Gisin, and Stefano Pironio. Bilocal versus nonbilocal correlations in entanglement-swapping experiments. *Phys. Rev. A*, 85:032119, Mar 2012.

[16] Ryan Carey, Marina Maciel Ansanelli, Elie Wolfe, and Robin J. Evans. Distinguishability of causal structures under latent confounding and selection, 2025.

[17] Daniel Centeno and Elie Wolfe. On the Significance of Intermediate Latents: Distinguishing Quantum Causal Scenarios with Indistinguishable Classical Analogs, 2024.

[18] Rafael Chaves, Gonzalo Carvacho, Iris Agresti, Valerio Di Giulio, Leandro Aolita, Sandro Giacomini, and Fabio Sciarrino. Quantum violation of an instrumental test. *Nature Physics*, 14(3):291–296, December 2017.

[19] Rafael Chaves, Christian Majenz, and David Gross. Information–theoretic implications of quantum causal structures. *Nature Communications*, 6(1):5766, January 2015.

[20] Rafael Chaves, George Moreno, Emanuele Polino, Davide Poderini, Iris Agresti, Alessia Suprano, Mariana R. Barros, Gonzalo Carvacho, Elie Wolfe, Askery Canabarro, Robert W. Spekkens, and Fabio Sciarrino. Causal networks and freedom of choice in Bell's theorem, 2021.

[21] David Maxwell Chickering. Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, 3(null):507–554, March 2003.

[22] Maria Ciudad-Alañón, Emanuel-Cristian Boghiu, Paolo Abiuso, and Elie Wolfe. Escaping the shadow of bell's theorem in network nonlocality, 2024.

[23] John F. Clauser and Michael A. Horne. Experimental consequences of objective local theories. *Phys. Rev. D*, 10:526–535, Jul 1974.

[24] Xavier Coiteux-Roy, Elie Wolfe, and Marc-Olivier Renou. No bipartite-nonlocal causal theory can explain nature's correlations. *Phys. Rev. Lett.*, 127:200401, Nov 2021.

[25] Fabio Costa and Sally Shrapnel. Quantum causal modelling. *New Journal of Physics*, 18(6):063032, jun 2016.

[26] Patrick Daley, Kevin J Resch, and Robert W Spekkens. Reply to " comment on 'experimentally adjudicating between different causal accounts of bell-inequality violations via statistical model selection'". *arXiv preprint arXiv:2412.02829*, 2024.

[27] Patrick J. Daley, Kevin J. Resch, and Robert W. Spekkens. Experimentally adjudicating between different causal accounts of bell-inequality violations via statistical model selection. *Phys. Rev. A*, 105:042220, Apr 2022.

[28] Erwan Don, Jessica Bavaresco, Patryk Lipka-Bartosik, Nicolas Gisin, Nicolas Brunner, and Alejandro Pozas-Kerstjens. The minimal example of quantum network bell nonlocality, 2026.

[29] Frederick Eberhardt. Almost optimal intervention sets for causal discovery. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, UAI'08, page 161–168, Arlington, Virginia, USA, 2008. AUAI Press.

[30] Frederick Eberhardt, Clark Glymour, and Richard Scheines. On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. In *UAI 05, Proceedings of the 21st Conference in Uncertainty in Artificial Intelligence, July 26-29 2005, Edinburgh, Scotland*, pages 178–184. AUAI Press, 2005.

[31] Robin Evans. Margins of discrete Bayesian networks. *Annals of Statistics*, 46, 01 2015.

[32] Robin J. Evans. Graphical Methods for Inequality Constraints in Marginalized DAGs. *2012 IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6, 2012.

[33] Robin J. Evans. Graphs for Margins of Bayesian Networks. *Scandinavian Journal of Statistics*, 43(3):625–648, 2016.
Note: Whenever we cited Propositions and Lemmas from this reference, we mentioned the numbering of the arXiv version, that is different from the numbering of the published version. This choice was made because the arXiv version may be more accessible to the reader.

[34] Robin J. Evans. Dependency in DAG models with hidden variables. In *Conference on Uncertainty in Artificial Intelligence*, 2021.

[35] Robin J. Evans. Latent free equivalent mDAGs. *Algebraic Statistics*, 14(1):3–16, November 2023.

[36] Robin J. Evans. Latent-free equivalent mDAGs. *Algebraic Statistics*, 14:3–16, 2023.

[37] Noam Finkelstein, Beata Zjawin, Elie Wolfe, Ilya Shpitser, and Robert W. Spekkens. Entropic inequality constraints from e-separation relations in directed acyclic graphs with hidden variables. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 1045–1055. PMLR, 27–30 Jul 2021.

[38] Thomas Fraser. A Combinatorial Solution to Causal Compatibility. *Journal of Causal Inference*, 8:22–53, 07 2020.

[39] Tobias Fritz. Beyond Bell's theorem: correlation scenarios. *New Journal of Physics*, 14(10):103001, oct 2012.

[40] Morten Frydenberg. The chain graph markov property. *Scandinavian Journal of Statistics*, 17:333–353, 1990.

[41] Mariami Gachechiladze, Nikolai Miklin, and Rafael Chaves. Quantifying causal influences in the presence of a quantum common cause. *Phys. Rev. Lett.*, 125:230401, Dec 2020.

[42] D. Geiger. Toward the formalization of informational dependencies. 1988.

[43] Dan Geiger and Judea Pearl. On the Logic of Causal Models. *Proc. 4th Conf. UAI*, pages 136—147, 1988.

[44] Dan Geiger, Thomas Verma, and Judea Pearl. Identifying independence in Bayesian networks. *Networks*, 20(5):507–534, 1990.

[45] Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13(79):2409–2464, 2012.

[46] James J. Heckman. Randomization and social policy evaluation. In *Evaluations: Welfare and Training Programs*. Harvard University Press, Cambridge, MA, 1992.

[47] Joe Henson, Raymond Lal, and Matthew F. Pusey. Theory-independent limits on correlations from generalized Bayesian networks. 16(11):113043, 2014. Publisher: IOP Publishing.

[48] Shashaank Khanna, Marina Maciel Ansanelli, Matthew F. Pusey, and Elie Wolfe. Classifying causal structures: Ascertaining when classical correlations are constrained by inequalities. *Phys. Rev. Res.*, 6:023038, Apr 2024.

[49] Shashaank Khanna, Matthew Pusey, and Roger Colbeck. Closing the problem of which causal structures of up to six total nodes have a classical-quantum gap, 2025.

[50] Pedro Lauand, Davide Poderini, Rafael Rabelo, and Rafael Chaves. Quantum non-classicality in the simplest causal network, 2024.

[51] Steffen L Lauritzen. *Graphical Models*. Clarendon Press, 1996.

[52] Steffen L Lauritzen. Generating mixed hierarchical interaction models by selection. Technical report, Aalborg Universitetsforlag, 1998.

[53] Steffen L. Lauritzen, AP Dawid, BN Larsen, and HG Leimer. Independence properties of directed Markov fields. *Networks*, 20(5):491–505, 1990.

[54] Robin Lorenz and Sean Tull. Causal models in string diagrams. 2023.

[55] Miguel Navascués, Elie Wolfe, Denis Rosset, and Alejandro Pozas-Kerstjens. Genuine network multipartite entanglement. *Phys. Rev. Lett.*, 125:240505, Dec 2020.

[56] Judea Pearl. On the Testability of Causal Models with Latent and Instrumental Variables. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, page 435–443, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[57] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition, 2009.

[58] Judea Pearl and Dana Mackenzie. *The Book of Why*. Basic Books, New York, NY, 2018.

[59] Jacques Pienaar. Which causal structures might support a quantum–classical gap? *New Journal of Physics*, 19(4):043021, apr 2017.

[60] Davide Poderini, Ranieri Nery, George Moreno, Santiago Zamora, Pedro Lauand, and Rafael Chaves. Observational-interventional bell inequalities, 2024.

[61] Thomas Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30(1):145–157, 2003.

[62] Thomas Richardson and Peter Spirtes. Ancestral graph Markov models. *The Annals of Statistics*, 30(4):962–1030, 2002.

[63] Thomas S. Richardson, Robin J. Evans, James M. Robins, and Ilya Shpitser. Nested Markov Properties for Acyclic Directed Mixed Graphs. *arXiv e-prints*, page arXiv:1701.06686, January 2017.

[64] Thomas S Richardson and James M Robins. Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128(30), 2013.

[65] Denis Rosset, Nicolas Gisin, and Elie Wolfe. Universal bound on the cardinality of local hidden variables in networks. *Quantum Info. Comput.*, 18(11–12):910–926, September 2018.

[66] David Schmid, John H. Selby, and Robert W. Spekkens. Unscrambling the omelette of causation and inference: The framework of causal-inferential theories. 2020.

[67] Ilya Shpitser, Robin Evans, and Thomas Richardson. Acyclic linear sems obey the nested markov property. *Uncertainty in artificial intelligence : proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2018, 08 2018.

[68] Ilya Shpitser, Robin J. Evans, Thomas S. Richardson, and James M. Robins. Introduction to Nested Markov Models. *Behaviormetrika*, 41:3–39, 01 2014.

[69] Ilya Shpitser and Judea Pearl. Dormant independence. In *AAAI-08/IAAI-08 Proceedings - 23rd AAAI Conference on Artificial Intelligence and the 20th Innovative Applications of Artificial Intelligence Conference*, Proceedings of the National Conference on Artificial Intelligence, pages 1081–1087, December 2008.

[70] Ilya Shpitser and Judea Pearl. Effects of treatment on the treated: Identification and generalization. In Jeff A. Bilmes and Andrew Y. Ng, editors, *UAI 2009, Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, June 18-21, 2009*, pages 514–521. AUAI Press, 2009.

[71] Ilya Shpitser and Eric Tchetgen Tchetgen. Causal inference with a graphical hierarchy of interventions. *Annals of statistics*, 44(6):2433, 2016.

[72] Isaac D. Smith, Elie Wolfe, and Robert W. Spekkens. Fully Quantum Inflation: Quantum Marginal Problem Constraints in the Service of Causal Inference. *PRX Quantum*, 7:010351, Mar 2026.

[73] Robert W. Spekkens. The ontological identity of empirical indiscernibles: Leibniz's methodological principle and its significance in the work of Einstein. 8 2019.

[74] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2001.

[75] Bastian Steudel and Nihat Ay. Information-theoretic Inference of Common Ancestors. *Entropy*, 17(4):2304–2327, 2015.

[76] Stephen Thornton. Karl Popper. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2021 edition, 2021.

[77] Jin Tian and Judea Pearl. A general identification condition for causal effects. In *AAAI/IAAI*, 2002.

[78] Thomas Van Himbeeck, Jonatan Bohr Brask, Stefano Pironio, Ravishankar Ramanathan, Ana Belén Sainz, and Elie Wolfe. Quantum violations in the Instrumental scenario and their relations to the Bell scenario. *Quantum*, 3:186, September 2019.

[79] Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, UAI '90, page 255–270, USA, 1990. Elsevier Science Inc.

[80] Tom Verma and Judea Pearl. Causal networks: Semantics and expressiveness. *Proceedings of the Fourth Workshop on Uncertainty in Artificial Intelligence*, 4, 03 2013.

[81] Elie Wolfe, Alejandro Pozas-Kerstjens, Matan Grinberg, Denis Rosset, Antonio Acín, and Miguel Navascués. Quantum Inflation: A General Approach to Quantum Causal Compatibility. *Phys. Rev. X*, 11:021043, May 2021.

[82] Elie Wolfe, Robert W. Spekkens, and Tobias Fritz. The Inflation Technique for Causal Inference with Latent Variables. *Journal of Causal Inference*, 7(2):20170020, 2019.

[83] Christopher Wood and Robert Spekkens. The lesson of causal discovery algorithms for quantum correlations: Causal explanations of Bell-inequality violations require fine-tuning. *New Journal of Physics*, 17, 08 2012.

[84] Yìlè Yīng, Marina Maciel Ansanelli, Andrea Di Biagio, Elie Wolfe, David Schmid, and Eric Gama Cavalcanti. Relating Wigner's Friend Scenarios to Nonclassical Causal Compatibility, Monogamy Relations, and Fine Tuning. *Quantum*, 8:1485, September 2024.

APPENDICES

Appendix A

Appendices for Chapter 2

A.1 Proof of Lemma 6

In this appendix we will prove Lemma 6, which asserts that the maps `split` and `RE−reduce` commute, as depicted in the commutative diagram of Fig. A.1.1.

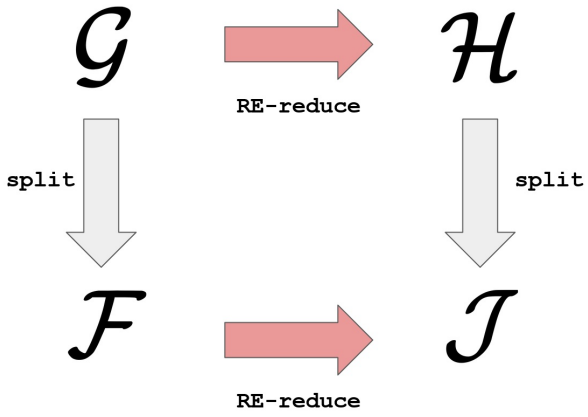


Figure A.1.1: Commutative diagram corresponding to the statement of Lemma 6.

Applying the `RE-reduce` map can be broken into a sequence of steps, where one exogenizes or removes only *one* latent node in each step. Therefore, it is enough to show that exogenizing *one* latent node commutes with the split-node map, as well as removing *one* redundant latent node commutes with the split-node map. That is, we just need to prove

that

$$\texttt{split}(\texttt{exog}(\mathcal{G}, u)) = \texttt{exog}(\texttt{split}(\mathcal{G}), u), \text{ and} \tag{A.1}$$
$$\texttt{split}(\texttt{remove}(\mathcal{G}, v)) = \texttt{remove}(\texttt{split}(\mathcal{G}), v) \tag{A.2}$$

for $u, v \in \texttt{lat}(\mathcal{G})$ and where v is parentless and redundant, i.e., there is another latent node $w \in \texttt{lat}(\mathcal{G})$ whose set of children is a superset of the set of children of v.

Eqs. (A.1) and (A.2) are respectively proven by the commuting diagrams presented in Figs. A.1.2 and A.1.3. In those diagrams, elongated ellipses represent a *set* of nodes (either visible or latent), and rectangles represent a set of input nodes.

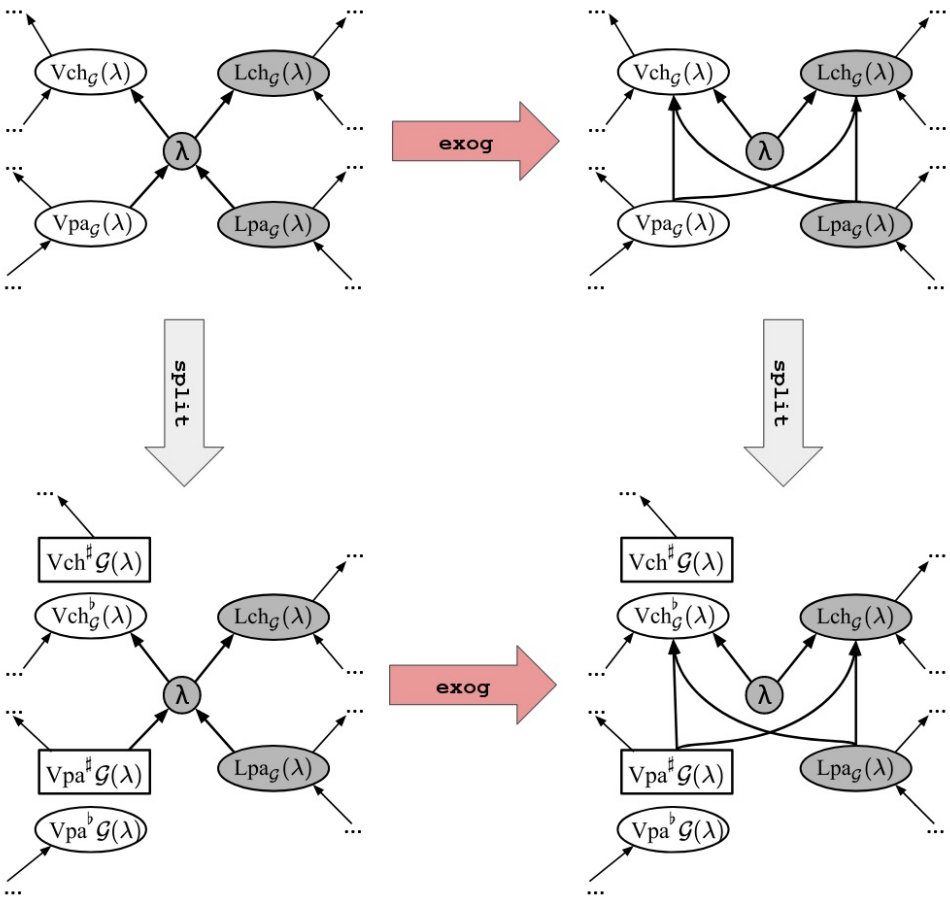


Figure A.1.2: Exogenizing a latent node commutes with the split-node map.

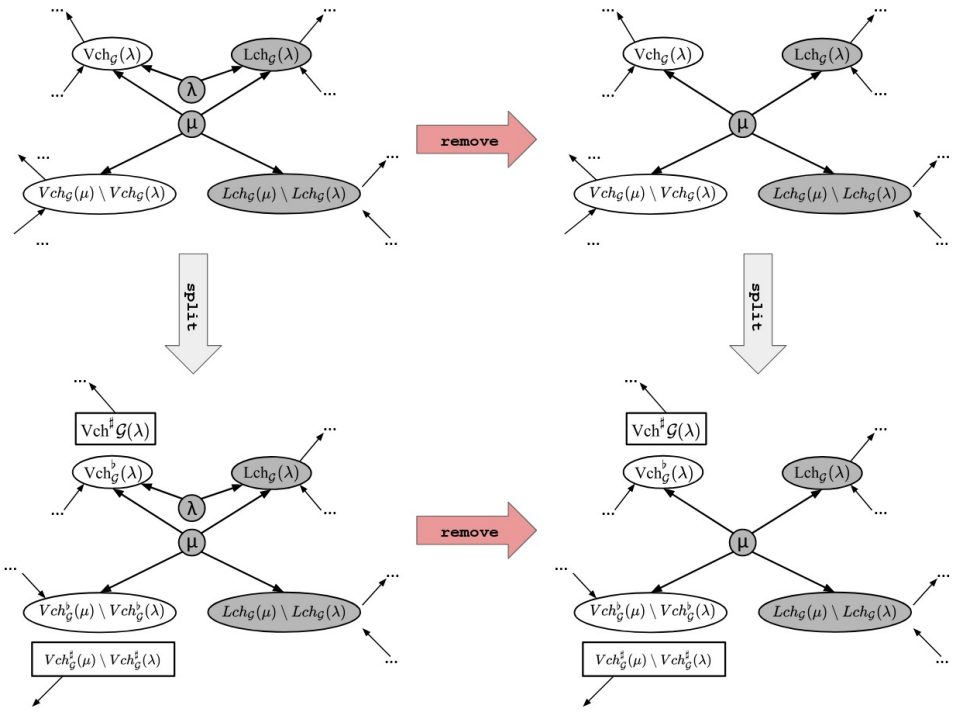


Figure A.1.3: Removing a redundant latent node commutes with the split-node map.

A.2 Proof of the Case 2 of Theorem 2

In this appendix, we will complete the proof of Theorem 2 by showing that in case 2, where there is at least one set S of nodes which is a face of $\mathfrak{G}' = \texttt{ToFaces}(\mathcal{G}')$ but not of $\mathfrak{G} = \texttt{ToFaces}(\mathcal{G})$, the pDAG $\mathcal{G}$ does not all-patterns O-or-D dominate the pDAG $\mathcal{G}'$.

If the set S does not have one common latent ancestor in $\mathcal{G}$, then the corresponding set $S^\flat$ also does not have one common latent ancestor in the full-SWIG $\mathscr{G} = \texttt{split}(\mathcal{G})$. Thus, in the proof of Lemma 7 we have shown that the conditional probability distribution of Eq. (2.14) cannot be realized by the full-SWIG $\mathscr{G}$. We reproduce this conditional distribution below:

$$P(X_{S^\flat}|X_{\texttt{vis}(\mathcal{G})^\sharp}) = p[0,...,0]_{S^\flat} + (1-p)[1,...,1]_{S^\flat} \tag{A.3}$$

In Theorem 2, for each visible variable our experimentalist can only perform a passive observation *or* a do-intervention. However, in the case of *binary* distributions it is possi-

ble to employ a technique [1] to find the full conditional distribution $P(X_{\texttt{vis}(\mathcal{G})^\flat} | X_{\texttt{vis}(\mathcal{G})^\sharp})$ from data that can be obtained from the all-patterns Observe-or-Do probing scheme. To understand it, imagine a simple case with two variables X_a and X_b. The most we can hope to learn from the causal hypothesis is given by the full conditional distribution $P(X_{a^\flat}, X_{b^\flat} | X_{a^\sharp}, X_{b^\sharp})$, which is given by a set of probabilities

$$P\left(X_{a^\flat} = x_{a^\flat}, X_{b^\flat} = x_{b^\flat} \middle| X_{a^\sharp} = x_{a^\sharp}, X_{b^\sharp} = x_{b^\sharp}\right). \tag{A.4}$$

Since we assume that the distribution is binary, the values $x_{a^\flat}$, $x_{b^\flat}$, $x_{a^\sharp}$ and $x_{b^\sharp}$ are either 0 or 1. We break the possibilities of values into four cases, where $\oplus$ indicates addition modulo 2:

(i) $x_{a^\flat} = x_{a^\sharp}$ and $x_{b^\flat} = x_{b^\sharp}$

In this case, the intervention does not change anything relative to a simple passive observation. Therefore, expression (A.4) reduces to the joint distribution obtained from passive observations on this causal hypothesis:

$$P\left(X_{a^\flat} = x_{a^\sharp}, X_{b^\flat} = x_{b^\sharp}\right). \tag{A.5}$$

(ii) $x_{a^\flat} = x_{a^\sharp}$ and $x_{b^\flat} = x_{b^\sharp} \oplus 1$

In this case, we use the fact that X_b is binary and the normalization of probabilities to reduce expression (A.4) to:

$$P(X_{a^\flat} = x_{a^\sharp} | X_{b^\sharp} = x_{b^\sharp}) - P\left(X_{a^\flat} = x_{a^\sharp}, X_{b^\flat} = x_{b^\sharp}\right) \tag{A.6}$$

(iii) $x_{a^\flat} = x_{a^\sharp} \oplus 1$ and $x_{b^\flat} = x_{b^\sharp}$

By symmetry to the previous case, in this case expression (A.4) becomes:

$$P(X_{b^\flat} = x_{b^\sharp} | X_{a^\sharp} = x_{a^\sharp}) - P\left(X_{a^\flat} = x_{a^\sharp}, X_{b^\flat} = x_{b^\sharp}\right) \tag{A.7}$$

(iv) $x_{a^\flat} = x_{a^\sharp} \oplus 1$ and $x_{b^\flat} = x_{b^\sharp} \oplus 1$

For this case, we will use the same technique two times. By using that $X_{b^\flat}$ is binary, expression (A.4) becomes:

$$P(X_{a^\flat} = x_{a^\sharp} \oplus 1 | X_{a^\sharp} = x_{a^\sharp}, X_{b^\sharp} = x_{b^\sharp}) - P\left(X_{a^\flat} = x_{a^\sharp} \oplus 1, X_{b^\flat} = x_{b^\sharp} \middle| X_{a^\sharp} = x_{a^\sharp}\right) \tag{A.8}$$

[1] The technique is due to Pedro Lauand, and was provided to us via private communication.

And by using that $X_{a^\flat}$ is also binary, this then becomes:

$$1 - P\big(X_{a^\flat} = x_{a^\sharp} \big| X_{b^\sharp} = x_{b^\sharp}\big) - P\left(X_{b^\flat} = x_{b^\sharp} \middle| X_{a^\sharp} = x_{a^\sharp}\right) + P\left(X_{a^\flat} = x_{a^\sharp}, X_{b^\flat} = x_{b^\sharp}\right). \quad \text{(A.9)}$$

Note that expressions (A.5)- (A.9) all correspond to data that can be obtained from the all-patterns Observe-or-Do probing scheme: in those expressions, for each visible variable, *either* its $\flat$ version appears to the left of the conditional *or* its $\sharp$ version appears to the right of the conditional (but not both at the same time). Therefore, the conditional distribution of expression (A.4) is in one-to-one correspondence to a set of data obtainable from the all-patterns Observe-or-Do probing scheme.

Now, suppose that the re-prepared values of X_a and X_b are 0, and that the conditional distribution of expression (A.4) presents perfect correlation between the natural values of X_a and X_b; that is, for some $p \in [0, 1]$:

$$\begin{aligned} &P\left(X_{a^\flat} = 0, X_{b^\flat} = 0 \middle| X_{a^\sharp} = 0, X_{b^\sharp} = 0\right) = p \\ &P\left(X_{a^\flat} = 1, X_{b^\flat} = 1 \middle| X_{a^\sharp} = 0, X_{b^\sharp} = 0\right) = 1 - p \end{aligned} \quad \text{(A.10)}$$

This is a special case of Eq. (A.3) when there are only two visible nodes and both of them are inside the set S. For this conditional distribution, the reductions (A.5), (A.6) and (A.7) of expression (A.4) give:

$$P\big(X_{a^\flat} = 0, X_{b^\flat} = 0\big) = p \quad \text{(A.11)}$$

$$P\big(X_{a^\flat} = 0 \big| X_{b^\sharp} = 0\big) = P\big(X_{a^\flat} = 0, X_{b^\flat} = 0\big) \quad \text{(A.12)}$$

$$P\big(X_{b^\flat} = 0 \big| X_{a^\sharp} = 0\big) = P\big(X_{a^\flat} = 0, X_{b^\flat} = 0\big) \quad \text{(A.13)}$$

Therefore, if our experimentalist is interested in attesting that the full conditional distribution is the one of Eq. (A.10), instead of performing an informationally complete probing scheme they can simply perform the all-patterns Observe-or-Do probing scheme to check that Eqs. (A.11), (A.12) and (A.13) are true. We did not write the equation correspondent to Eq. (A.9) here, because this would be redundant to Eqs. (A.11), (A.12) and (A.13) by normalization of probabilities.

The argument we have made before Eq. (A.3) implies that, in the special case where $S = \texttt{vis}(\mathcal{G}) = \{a, b\}$, Eqs. (A.11), (A.12) and (A.13) are not jointly realizable by $\mathcal{G}$. It is easy to see that, in this special case, the set of data described by Eqs. (2.30a) and (2.30b) (where passive observations show perfect correlation between the variables X_a and X_b) obeys Eqs. (A.11), (A.12) and (A.13). Thus, at least when $S = \texttt{vis}(\mathcal{G}) = \{a, b\}$, Eqs. (2.30a) and (2.30b) are *not* jointly realizable by $\mathcal{G}$.

For the generic case, the proof is very similar. We will suppose that the set of data described by Eqs. (2.30a) and (2.30b) is jointly realizable by $\mathcal{G}$, and we will show that this implies that the full conditional distribution of Eq. (A.3) must be realizable by $\mathcal{G}$, which is a contradiction.

We start by noting that, for the evaluation of the conditional distribution where all of the variables in $X_{S^\flat}$ take value 0, we have:

$$P\left(X_{S^\flat} = 0_{S^\flat} \middle| X_{\mathtt{vis}(\mathcal{G})^\sharp} = 0_{\mathtt{vis}(\mathcal{G})^\sharp}\right) = P\left(X_{S^\flat} = 0_{S^\flat} \middle| X_{\mathtt{vis}(\mathcal{G})^\sharp \setminus S^\sharp} = 0_{\mathtt{vis}(\mathcal{G})^\sharp \setminus S^\sharp}\right), \tag{A.14}$$

which can be obtained by the all-patterns Observe-or-Do probing scheme. The equality above holds due to a logic similar to the one that lead to Eq. (A.5). In this case, Eqs. (2.30a) and (2.30b) say that this expression takes the value p, which is exactly what is given by the corresponding evaluation of the full conditional distribution of Eq. (A.3).

Eq. (A.3) says that evaluations of the full conditional distribution where some of the variables in $X_{S^\flat}$ take the value 0 while others take the value 1 must be zero. Starting with the assumption that the set of data obtained from the all-patterns Observe-or-Do probing scheme is given by Eqs. (2.30a) and (2.30b), we will now prove that all of these evaluations are indeed zero, again in accordance with Eq. (A.3). As an example, below we show the first of these evaluations, where we enumerate the nodes of the set S as $S = \{s_1, ..., s_n\}$:

$$\begin{aligned}
&P\left(X_{s_1^\flat} = 1, X_{s_2^\flat} = ... = X_{s_n^\flat} = 0 \middle| X_{\mathtt{vis}(\mathcal{G})^\sharp} = 0_{\mathtt{vis}(\mathcal{G})^\sharp}\right) \\
&= P\left(X_{s_2^\flat} = ... = X_{s_n^\flat} = 0 \middle| X_{\mathtt{vis}(\mathcal{G})^\sharp} = 0_{\mathtt{vis}(\mathcal{G})^\sharp}\right) - P\left(X_{S^\flat} = 0_{S^\flat} \middle| X_{\mathtt{vis}(\mathcal{G})^\sharp} = 0_{\mathtt{vis}(\mathcal{G})^\sharp}\right) \qquad \text{(A.15)} \\
&= P\left(X_{s_2^\flat} = ... = X_{s_n^\flat} = 0 \middle| X_{s_1^\sharp} = 0, X_{\mathtt{vis}(\mathcal{G})^\sharp \setminus S^\sharp} = 0_{\mathtt{vis}(\mathcal{G})^\sharp \setminus S^\sharp}\right) - P\left(X_{S^\flat} = 0_{S^\flat} \middle| X_{\mathtt{vis}(\mathcal{G})^\sharp \setminus S^\sharp} = 0_{\mathtt{vis}(\mathcal{G})^\sharp \setminus S^\sharp}\right),
\end{aligned}$$

where we applied the technique described in the simple example above. In the data given by Eqs. (2.30a) and (2.30b), both terms of the last line of Eq. (A.15) are equal to p, and thus Eq. (A.15) is equal to zero.

In the generic case (an evaluation where the first i elements of $X_{S^\flat}$ take the value 1, while the last $n-i$ take the value 0), we will recursively apply the same technique. Applying it one time, we obtain:

$$\begin{aligned}
&P\left(X_{s_1^\flat} = ... = X_{s_i^\flat} = 1, X_{s_{i+1}^\flat} = ... = X_{s_n^\flat} = 0 \middle| X_{\mathtt{vis}(\mathcal{G})^\sharp} = 0_{\mathtt{vis}(\mathcal{G})^\sharp}\right) \\
&= P\left(X_{s_1^\flat} = ... = X_{s_{i-1}^\flat} = 1, X_{s_{i+1}^\flat} = ... = X_{s_n^\flat} = 0 \middle| X_{\mathtt{vis}(\mathcal{G})^\sharp} = 0_{\mathtt{vis}(\mathcal{G})^\sharp}\right) \qquad \text{(A.16)} \\
&- P\left(X_{s_1^\flat} = ... = X_{s_{i-1}^\flat} = 1, X_{s_i^\flat} = ... = X_{s_n^\flat} = 0 \middle| X_{\mathtt{vis}(\mathcal{G})^\sharp} = 0_{\mathtt{vis}(\mathcal{G})^\sharp}\right) \qquad \text{(A.17)}
\end{aligned}$$

In the first term of the right hand side, given in line (A.16), the variable $X_{s_i^\flat}$ does not appear. In the second term, given in line (A.17), this variable appears as $X_{s_i^\flat} = 0$.

Therefore, this first application of the technique got rid of any terms with $X_{s_i^\flat} = 1$. Note that when we recursively apply this technique for all variables, both of the terms (A.16) and (A.17) will be decomposed in many sub-terms that do not have any variable being equal to 1. However, the sub-terms in which (A.16) will be decomposed are equal to the sub-terms in which (A.17) will be decomposed, except for the fact that in the case of (A.17) there will be $X_{s_i^\flat} = 0$ appearing in each of them, while in the case of (A.16) there will not. In other words, a sub-term in which (A.16) will be decomposed is of the form

$$P\left(X_{T^\flat} = 0 \middle| X_{\texttt{vis}(\mathcal{G})^\sharp \setminus T^\sharp} = 0_{\texttt{vis}(\mathcal{G})^\sharp \setminus T^\sharp}\right), \tag{A.18}$$

where $T \subseteq S$ is a subset of S that does not include s_i. For each of these, there is a sub-term in which (A.17) will be decomposed that is of the form

$$P\left(X_{T^\flat} = 0, X_{s_i^\flat} = 0 \middle| X_{\texttt{vis}(\mathcal{G})^\sharp \setminus \{T^\sharp \cup \{s_i^\sharp\}\}} = 0_{\texttt{vis}(\mathcal{G})^\sharp \setminus \{T^\sharp \cup \{s_i^\sharp\}\}}\right). \tag{A.19}$$

In both sub-terms (A.18) and (A.19), we already used the fact that do-interventions that force the variable to take the value 0 do not do anything if the variable already naturally takes the value 0 (the same fact that was used in Eq. (A.14)).

In both expressions (A.18) and (A.19), there is no variable whose $\flat$ version appears to the left of the conditional at the same time that the $\sharp$ version appears to the right of the conditional. Therefore, they are both obtainable from the all-patterns Observe-or-Do probing scheme. We can thus look at Eqs. (2.30a) and (2.30b) to see that both sub-terms (A.18) and (A.19) are equal to p, which implies that they will cancel each other when the term (A.17) is subtracted from the term (A.16). This implies that

$$P\left(X_{s_1^\flat} = ... = X_{s_i^\flat} = 1, X_{s_{i+1}^\flat} = ... = X_{s_n^\flat} = 0 \middle| X_{\texttt{vis}(\mathcal{G})^\sharp} = 0_{\texttt{vis}(\mathcal{G})^\sharp}\right) = 0, \tag{A.20}$$

which is in accordance with Eq. (A.3).

Therefore, we proved that when all of the variables in $X_{S^\flat}$ evaluate to 0 (Eq. (A.14)), Eqs. (2.30a) and (2.30b) imply that $P(\texttt{vis}(\mathcal{G})^\flat | \texttt{vis}(\mathcal{G})^\sharp)$ is equal to p, and when some of the variables evaluate to 0 and others to 1 (Eq. (A.20)), Eqs. (2.30a) and (2.30b) imply that $P(\texttt{vis}(\mathcal{G})^\flat | \texttt{vis}(\mathcal{G})^\sharp)$ is equal to 0. In other words, Eqs. (2.30a) and (2.30b) imply in the conditional distribution of Eq. (A.3), which we already proved to be non-realizable by $\mathscr{G} = \texttt{split}(\mathcal{G})$. This contradiction implies that Eqs. (2.30a) and (2.30b) are not jointly realizable by $\mathcal{G}$.

Appendix B

Appendices for Chapter 3

B.1 Proof that latent-confounder-free pDAGs consistent with a fixed nodal ordering are observationally inequivalent

Here, we prove that all latent-confounder-free pDAGs consistent with a fixed nodal ordering are observationally inequivalent.

In this proof of observational inequivalence, we will explicitly show a d-separation relation that is presented by one of the pDAGs in question but not by the other. Given a pDAG $\mathcal{G}$ and three disjoint sets of visible nodes $A, B, C \in \texttt{vis}(\mathcal{G})$, the *d-separation criterion* [80, 42] returns whether A and B are "d-separated" by C or not. The affirmative case is denoted by $A \perp B|C$. As shown in Ref. [80], a d-separation relation of a DAG implies that the realizable distributions have to satisfy an associated *conditonal independence constraint*. In a certain distribution P, the variable X_A is said to be independent of the variable X_B after conditioning on the variable X_C if the following is valid:

$$P(X_A X_B|X_C) = P(X_A|X_C)P(X_B|X_C). \tag{B.1}$$

Importantly, Ref. [80] also showed that, if a pDAG $\mathcal{G}$ *does not* present such d-separation relation, i.e. if $A \not\perp B|C$ in $\mathcal{G}$, then there *is* some distribution that is realizable by $\mathcal{G}$ and does not obey the constraint of Eq. (B.1). That is, the d-separation criterion is necessary and sufficient to give us the set of conditional independence constraints imposed by a pDAG on its realizable distributions.

The following Lemma is a consequence of the observations above:

Lemma 21. *Let $\mathcal{G}$ and $\mathcal{G}'$ be two pDAGs that have the same set of visible nodes,* $\texttt{vis}(\mathcal{G}) = \texttt{vis}(\mathcal{G}')$. *If they present different sets of d-separation relations, then they are observationally inequivalent.*

Proof. Without loss of generality, let $A \perp B|C$ be a d-separation relation that is presented by $\mathcal{G}'$ but not in $\mathcal{G}$. Then, by the first result of Ref. [80] mentioned above, all of the distributions realizable by $\mathcal{G}'$ have to satisfy the constraint of Eq. (B.1). By the second result of Ref. [80] mentioned above, there *exists* some distribution that is realizable by $\mathcal{G}$ and does not satisfy this constraint. Therefore, there exists at least one probability distribution obtained from passive observations that is realizable by $\mathcal{G}$ but not by $\mathcal{G}'$, implying that they are observationally inequivalent. □

Now, we can proceed to the proof that two latent-confounder-free pDAGs consistent with a fixed nodal ordering are always observationally inequivalent. Let $\mathcal{G}$ and $\mathcal{G}'$ be two latent-confounder-free pDAGs (i.e., two DAGs) with n nodes, and let $\left\{a^{(i)}\right\}_{i=1,...,n}$ be the set of their nodes such that $i < j$ if and only if $a^{(i)}$ comes before $a^{(j)}$ in the nodal ordering. That is, the nodal ordering is $\left(a^{(1)}, a^{(2)}, ..., a^{(n)}\right)$. Assume that both $\mathcal{G}$ and $\mathcal{G}'$ are consistent with this nodal ordering.

Since $\mathcal{G}$ and $\mathcal{G}'$ are not equal, there must be at least one arrow that is present in one of them but not the other. Without loss of generality, assume that $\mathcal{G}$ possesses the arrow $a^{(i)} \to a^{(j)}$, while $\mathcal{G}'$ does not. Note that this necessarily implies that $i < j$, since we assumed that $\mathcal{G}$ and $\mathcal{G}'$ are consistent with the nodal ordering $\left(a^{(1)}, a^{(2)}, ..., a^{(n)}\right)$. Now, we show that the following d-separation relation is presented by $\mathcal{G}'$ but not $\mathcal{G}$:

$$a^{(i)} \perp a^{(j)} | a^{(1)}, ..., a^{(i-1)}, a^{(i+1)}, ..., a^{(j-1)} \tag{B.2}$$

It is clear that $\mathcal{G}$ *does not* present this d-separation relation, since it presents a direct arrow $a^{(i)} \to a^{(j)}$. To see that $\mathcal{G}'$ presents this d-separation relation note that, since $\mathcal{G}'$ is latent-confounder-free and consistent with the fixed nodal ordering $\left(a^{(1)}, a^{(2)}, ..., a^{(n)}\right)$, any node that is a common cause between $a^{(i)}$ and $a^{(j)}$ *or* a mediary in a chain between $a^{(i)}$ and $a^{(j)}$ is necessarily in the set $\left\{a^{(1)}, ..., a^{(i-1)}, a^{(i+1)}, ..., a^{(j-1)}\right\}$. Therefore, all such paths between $a^{(i)}$ and $a^{(j)}$ are blocked by conditioning on this set. Furthermore, conditioning on this set does not open any new paths via colliders, because the nodal ordering says that the set $\left\{a^{(1)}, ..., a^{(i-1)}, a^{(i+1)}, ..., a^{(j-1)}\right\}$ does not include any children of $a^{(j)}$.

Therefore, the d-separation relation of expression (B.2) is presented by $\mathcal{G}'$ but not by $\mathcal{G}$. Together with Lemma 21, this finishes our proof.

B.2 Proof of Strong Facet-Adding

In this appendix we will prove the Strong Facet-Adding proposition, which is reproduced below.

Proposition 8 (Strong Facet-Adding (Strong FA)). *Let $\mathfrak{G} = \{\mathcal{D}, \mathcal{B}\}$ be an mDAG whose simplicial complex contains a sequence of faces $C_i \in \mathcal{B}$ for $i = 1, ..., n$, that do not need to be disjoint. Suppose further that the mDAG $\mathfrak{G}$ also contains a face $D \in \mathcal{B}$ that is disjoint to every C_i, $i = 1, ..., n$. Let $\mathfrak{G}'$ be the mDAG obtained by starting from $\mathfrak{G}$ and, if not already present, adding the sets $B_i = C_i \cup D$, $i = 1, ..., n$, together with all of the faces they contain, to its simplicial complex.*

Suppose that:

1. $\mathtt{pa}_{\mathcal{D}}(C_i) \cup C_i \subseteq \mathtt{pa}_{\mathcal{D}}(d)$ *for each* $d \in D$, $i = 1, ..., n$.
2. *If for a facet* $B' \in \mathcal{B}$ *such that* $B' \nsubseteq C_i$ *for all* $i = 1, ..., n$ *we have* $c \in B'$ *for some* $c \in \cup_{i=1,...,n} C_i$, *then* $D \subseteq B'$.

In this case, $\mathfrak{G}$ observationally dominates $\mathfrak{G}'$, i.e., $\mathfrak{G} \succeq \mathfrak{G}'$.

Note that, since $\mathfrak{G}'$ structurally dominates $\mathfrak{G}$, by Lemma 3 we know that $\mathfrak{G}'$ observationally dominates $\mathfrak{G}$, i.e., $\mathfrak{G}' \succeq \mathfrak{G}$. Therefore, $\mathfrak{G}$ and $\mathfrak{G}'$ are observationally equivalent, i.e., $\mathfrak{G} \cong \mathfrak{G}'$.

To facilitate the proof, in this appendix we will use the language of pDAGs instead of that of mDAGs. Note that, as a special case, this is an alternative proof of the original Evans's Proposition (different from the proof presented in [33]).

We start with Lemma 22, that generalizes the idea that a causal structure where the nodes of the set C and their parents are settings for a latent variable is at least as powerful as a causal structure where the nodes of the set C are causally influenced by that latent variable. An example of the application of Lemma 22 is presented in Figure B.2.1.

Lemma 22 (Dominance by Bayesian Updating). *Let $\mathcal{G}_1$ be a pDAG that has a set of parentless latent variables $\{l_i^{CD}\}_{i=1,...,n}$, and let the children of l_i^{CD} be $ch_{\mathcal{G}_1}(l_i) = C_i \cup D$, such that $\cup_i C_i$ and D are disjoint. All the nodes in $\{l_i^{CD}\}_{i=1,...,n}$ share the children D, and the sets C_i, $i = 1, ..., n$ do not have to be disjoint from each other. Further suppose that none of the nodes in D is an ancestor of any node $c \in \cup_i C_i$.*

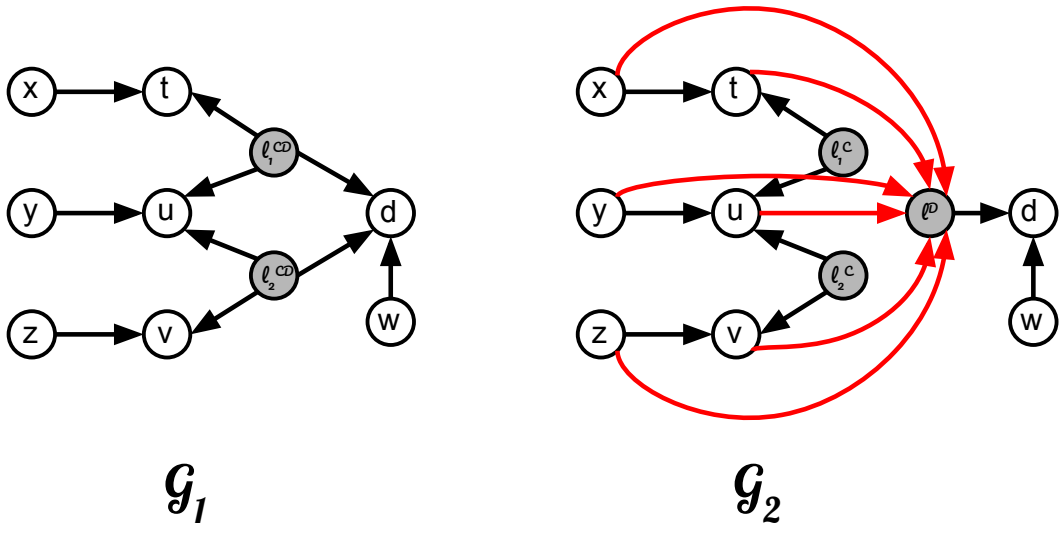


Figure B.2.1: By Lemma 22, we have $\mathcal{G}_2 \succeq \mathcal{G}_1$. Here, $C_1 = \{t, u\}$, $C_2 = \{u, v\}$, and $D = \{d\}$. Note that $\mathcal{G}_1$ does *not* observationally dominate $\mathcal{G}_2$: for example, $\mathcal{G}_1$ has the d-separation relation $x \perp\!\!\!\perp_{\mathrm{d}} d$, that is not presented by $\mathcal{G}_2$. Edges that point into a latent node are represented in red.

Let $\mathcal{G}_2$ be another pDAG that is constructed from $\mathcal{G}_1$ by deleting all the nodes $\{l_i^{CD}\}_{i=1,\dots,n}$, adding a latent node l^D whose children are $ch_{\mathcal{G}_2}(l_D) = D$ and whose parents are $pa_{\mathcal{G}_2}(l_D) = \cup_i (C_i \cup pa_{\mathcal{G}_1}(C_i)) \setminus \{l_i^{CD}\}_{i=1,\dots,n}$ and adding latent common causes l_i^C, $i = 1, \dots, n$ that have no parents and are such that $ch_{\mathcal{G}_2}(l_i^C) = C_i$.

Then, $\mathcal{G}_2$ observationally dominates $\mathcal{G}_1$, i.e., $\mathcal{G}_2 \succeq \mathcal{G}_1$.

Proof. For simplicity of notation, let $C \equiv \cup_i C_i$, $L^{CD} \equiv \{l_i^{CD}\}_{i=1,\dots,n}$ and $L^C \equiv \{l_i^C\}_{i=1,\dots,n}$. Also define $V \equiv \mathtt{vis}(\mathcal{G}_1) = \mathtt{vis}(\mathcal{G}_2)$.

The idea of this lemma is that, due to D not being part of the ancestors of C, one can assess the values of X_C before assessing the values of X_D. The knowledge of X_C in $\mathcal{G}_1$ will thus teach us something about the shared common causes $X_{L^{CD}}$ through Bayesian updating. As it turns out, this will be at most as powerful as $\mathcal{G}_2$, where C and its parents act as settings for the latent variable l^D.

We will show this by explicitly writing down the Markov realizability conditions for each one of the pDAGs. The Markov realizability conditions are equivalent to Eq. (2.3), but written in terms of conditional probability distributions instead of in terms of functional

dependences and error variables. It might be helpful to follow the proof below while looking at the example of Figure B.2.1.

To further simplify the notation, let $V^* = V \setminus (C \cup D \cup (\mathtt{an}_{\mathcal{G}_1}(C) \cap V))$ be the set of visible nodes of $\mathcal{G}_1$ and $\mathcal{G}_2$ except for C, D and the visible ancestors of C, and let $L^* = \mathtt{lat}(\mathcal{G}_1) \setminus (\mathtt{an}_{\mathcal{G}_1}(C) \cap \mathtt{lat}(\mathcal{G}_1))$ be the set of latent nodes of $\mathcal{G}_1$ except for the latent ancestors of C. Also let $\mathrm{pa}^*_{\mathcal{G}_1}(A) = \mathrm{pa}_{\mathcal{G}_1}(A) \setminus L^{CD}$ be the set of parents of a set of nodes A in $\mathcal{G}_1$ except for L^{CD} and $\mathtt{an}^*_{\mathcal{G}_1}(A) = \mathtt{an}_{\mathcal{G}_1}(A) \setminus L^{CD}$ be the set of ancestors of a set of nodes A in $\mathcal{G}_1$ except for L^{CD}. A probability distribution over the visible variables that is realizable by $\mathcal{G}_1$ can be factorized as:

$$\begin{aligned} &P_{\mathcal{G}_1}(X_V) = \\ &\sum_{X_{\mathtt{lat}(\mathcal{G}_1)}} P_{\mathcal{G}_1}(X_D | X_{L^{CD}} X_{\mathrm{pa}^*_{\mathcal{G}_1}(D)}) \cdot \\ &\cdot P_{\mathcal{G}_1}(X_C X_{\mathtt{an}^*_{\mathcal{G}_1}(C)} X_{L^{CD}}) \cdot \\ &\cdot P_{\mathcal{G}_1}(X_{V^*L^*} | X_C X_D X_{\mathtt{an}^*_{\mathcal{G}_1}(C)}) \end{aligned} \tag{B.3}$$

Note that Eq. (B.3) is a coarse-grained version of the equation that determines realizability by $\mathcal{G}_1$: the probability distributions realizable by $\mathcal{G}_1$ obey even more specific constraints than the ones presented by Eq. (B.3), because for example not all elements of D are necessarily children of every element of $\mathrm{pa}_{\mathcal{G}_1}(D)$.

Note that the term $P_{\mathcal{G}_1}(X_C X_{\mathtt{an}^*_{\mathcal{G}_1}(C)} X_{L^{CD}})$ can be written as:

$$\begin{aligned} &P_{\mathcal{G}_1}(X_C X_{\mathtt{an}^*_{\mathcal{G}_1}(C)} X_{L^{CD}}) = \\ &P_{\mathcal{G}_1}(X_C X_{\mathtt{an}^*_{\mathcal{G}_1}(C)}) P_{\mathcal{G}_1}(X_{L^{CD}} | X_C X_{\mathtt{an}^*_{\mathcal{G}_1}(C)}). \end{aligned} \tag{B.4}$$

This equality is true by Bayes's theorem, that is, it does not make use of the structure of $\mathcal{G}_1$.

Note also that L^{CD} is d-separated from the ancestors of C that are not parents of C when we condition on C and its other parents, i.e., the following conditional independence has to hold:

$$P_{\mathcal{G}_1}(X_{L^{CD}} | X_C X_{\mathtt{an}^*_{\mathcal{G}_1}(C)}) = P_{\mathcal{G}_1}(X_{L^{CD}} | X_C X_{\mathtt{pa}^*_{\mathcal{G}_1}(C)}). \tag{B.5}$$

Substituting Eqs. (B.4) and (B.5) into Eq. (B.3), we get:

$$
\begin{aligned}
&P_{\mathcal{G}_1}(X_V) = \\
&\sum_{X_{\mathtt{lat}(\mathcal{G}_1)}} P_{\mathcal{G}_1}(X_D | X_{L^{CD}} X_{\mathrm{pa}^*_{\mathcal{G}_1}(D)}) \cdot \\
&\cdot P_{\mathcal{G}_1}(X_C X_{\mathtt{an}^*_{\mathcal{G}_1}(C)}) P_{\mathcal{G}_1}(X_{L^{CD}} | X_C X_{\mathtt{pa}^*_{\mathcal{G}_1}(C)}) \cdot \\
&\cdot P_{\mathcal{G}_1}(X_{V^* L^*} | X_C X_D X_{\mathtt{an}^*_{\mathcal{G}_1}(C)})
\end{aligned}
\tag{B.6}
$$

To show that $\mathcal{G}_2$ can simulate $\mathcal{G}_1$, we need to show that all the probability distributions that can be factorized as in Eq. (B.3) can also be factorized according to the structure of $\mathcal{G}_2$. To write down the Markov realizability condition for $\mathcal{G}_2$, first note that $L^* = \mathtt{lat}(\mathcal{G}_2) \setminus ((\mathtt{an}_{\mathcal{G}_2}(C) \cap \mathtt{lat}(\mathcal{G}_2)) \cup \{l^D\})$, and that $\mathrm{pa}^*_{\mathcal{G}_1}(A) = \mathrm{pa}_{\mathcal{G}_2}(A) \setminus (L^C \cup \{l^D\})$ and $\mathrm{an}^*_{\mathcal{G}_1}(A) = \mathrm{an}_{\mathcal{G}_2}(A) \setminus (L^C \cup \{l^D\})$. Furthermore, define $L^C_c \subseteq L^C$ as the set of elements of L^C that are parents of $c \in C$. For example, in Figure B.2.1(b), $L^C_t = \{l^C_1\}$ and $L^C_u = \{l^C_1, l^C_2\}$. With this, for $\mathcal{G}_2$ we have:

$$
\begin{aligned}
&P_{\mathcal{G}_2}(X_V) = \\
&\sum_{X_{\mathtt{lat}(\mathcal{G}_2)}} P_{\mathcal{G}_2}(X_D | X_{l^D} X_{\mathrm{pa}^*_{\mathcal{G}_1}(D)}) \cdot \\
&\cdot P_{\mathcal{G}_2}(X_C X_{\mathtt{an}^*_{\mathcal{G}_1}(C)} X_{L^C}) P_{\mathcal{G}_2}(X_{l^D} | X_C X_{\mathtt{pa}^*_{\mathcal{G}_1}(C)}) \cdot \\
&\cdot P_{\mathcal{G}_2}(X_{V^* L^*} | X_C X_D X_{\mathtt{an}^*_{\mathcal{G}_1}(C)}) = \\
&= \sum_{X_{\mathtt{lat}(\mathcal{G}_2)} \setminus X_{L^C}} P_{\mathcal{G}_2}(X_D | X_{l^D} X_{\mathrm{pa}^*_{\mathcal{G}_1}(D)}) \cdot \\
&\cdot P_{\mathcal{G}_2}(X_C X_{\mathtt{an}^*_{\mathcal{G}_1}(C)}) P_{\mathcal{G}_2}(X_{l^D} | X_C X_{\mathtt{pa}^*_{\mathcal{G}_1}(C)}) \cdot \\
&\cdot P_{\mathcal{G}_2}(X_{V^* L^*} | X_C X_D X_{\mathtt{an}^*_{\mathcal{G}_1}(C)})
\end{aligned}
\tag{B.7}
$$

where in the second equality we used the fact that the latent variables X_{L^C} only appeared in one term, so they can be easily marginalized out.

If we lump all the variables $X_{L^{CD}}$ of $\mathcal{G}_1$ together in one variable X_{l^D}, we can see that Eqs. (B.6) and (B.7) take exactly the same form.

Therefore, $\mathcal{G}_2$ can realize all of the distributions realizable by $\mathcal{G}_1$. Note that the inverse is not true; Fig. B.2.1 presents a counter-example.

□

With this Lemma at hand, we can now prove Proposition 8. The following is a rephrasing of this proposition in pDAG language, where $\mathcal{G}_1$ corresponds to $\mathfrak{G}'$ of Proposition 8 and $\mathcal{G}_3$ corresponds to $\mathfrak{G}$ of Proposition 8. That is, the proposition below describes the inverse of the transformation described in Proposition 8.

Proposition 30 (Strong Facet-Adding - pDAG language)**.** *Let $\mathcal{G}_1$ be a pDAG containing a sequence of latent nodes $\{l_i^{CD}\}_{i=1,\dots,n}$ that share the children $D \subseteq \cap_{i=1,\dots,n} ch_{\mathcal{G}_1}(l_i^{CD})$. Denote $C_i \equiv ch_{\mathcal{G}_1}(l_i^{CD}) \setminus D$.*

To simplify the notation, we will use $C \equiv \cup_i C_i$ and $L^{CD} \equiv \{l_i^{CD}\}_{i=1,\dots,n}$

If the following condition holds:

1. *$pa_{\mathcal{G}_1}(C) \cup C \subseteq pa_{\mathcal{G}_1}(d)$ for each $d \in D$. This has to hold for both visible and latent parents.*

Then, $\mathcal{G}_1$ is observationally equivalent to the pDAG $\mathcal{G}_3$ defined by starting from $\mathcal{G}_1$, removing the nodes L^{CD} and adding the new latent nodes $\{l_i^C\}_{i=1,\dots,n}$ and l_D, whose children are respectively $ch_{\mathcal{G}_3}(l_i^C) = C^i$ for $i = 1, \dots, n$ and $ch_{\mathcal{G}_3}(l_D) = D$.

Proof. It might be helpful to look at Figure B.2.2 while following this proof.

By structural dominance (Lemma 3), it is easy to see that $\mathcal{G}_1 \succeq \mathcal{G}_3$.

By Lemma 22, the pDAG $\mathcal{G}_1$ here defined is dominated by a pDAG $\mathcal{G}_2$ where the nodes $\{l_i^{CD}\}_{i=1,\dots,n}$ are substituted by parentless latent nodes $\{l_i^C\}_{i=1,\dots,n}$ that point to each of the C_i, and a node l^D that points to the set D and has parents $pa_{\mathcal{G}_2}(l_D) = \cup_i (C_i \cup pa_{\mathcal{G}_1}(C_i)) \setminus \{l_i^{CD}\}_{i=1,\dots,n}$.

The next step is to exogenize the latent node l^D, using Lemma 1. Since $pa_{\mathcal{G}_1}(C) \cup C \subseteq pa_{\mathcal{G}_1}(d)$ for each $d \in D$, when we exogenize l^D there are no new edges added; only edges removed (the edges between $\mathrm{pa}_{\mathcal{G}_2}(l^D)$ and l^D). With this process, we reach a pDAG $\mathcal{G}_3 \cong \mathcal{G}_2$. This is the pDAG $\mathcal{G}_3$ that appears in the statement of the proposition.

Thus, we have that $\mathcal{G}_3 \cong \mathcal{G}_2 \succeq \mathcal{G}_1$, and we already knew that $\mathcal{G}_1 \succeq \mathcal{G}_3$. Therefore, we conclude that $\mathcal{G}_1 \cong \mathcal{G}_3$. □

B.3 Fraser's Algorithm for Feasible Supports

In this section, we present the algorithm of Ref. [38] for finding $\mathcal{S}(\mathcal{G}, \vec{c}_{\text{vis}})$, the set of supports with cardinalities $\vec{c}_{\text{vis}}$ of the visible variables that are realizable by a pDAG $\mathcal{G}$.

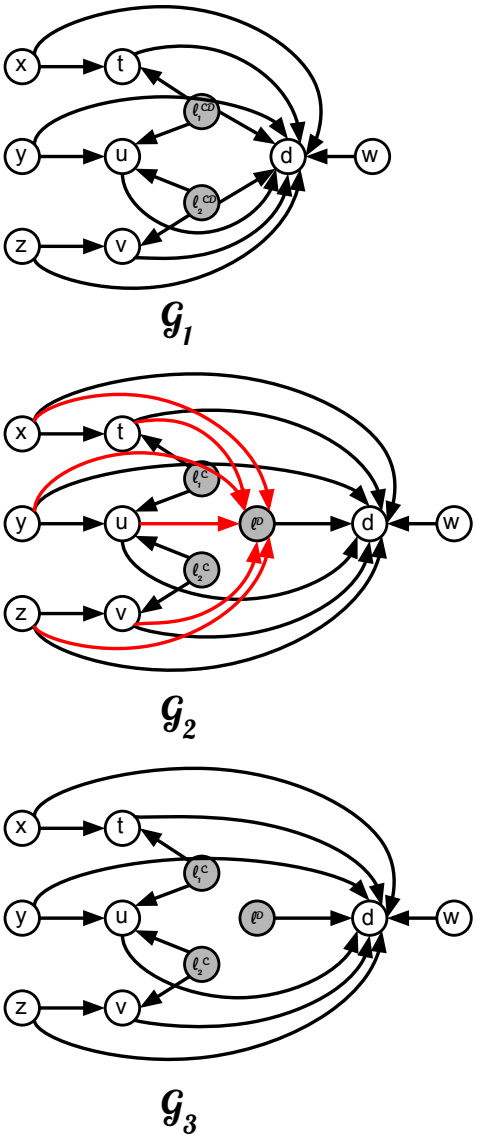


Figure B.2.2: The pDAGs $\mathcal{G}_1$ and $\mathcal{G}_3$ can be shown observationally equivalent by Proposition 30: Lemma 22 says that $\mathcal{G}_2 \succeq \mathcal{G}_1$, and Exogenization (Lemma 1) says that $\mathcal{G}_2 \cong \mathcal{G}_3$. We can see that $\mathcal{G}_1 \succeq \mathcal{G}_3$ by structural dominance (Lemma 3). Edges that point into a latent node are represented in red.

This algorithm consists of enumerating all the possible responses that a visible variable can have to its parents, given that the cardinalities of the visible variables and the latent

variables are fixed.

In Ref. [38], it is shown that whenever the visible variables have finite cardinalities, we can assume the latent variables to have finite cardinalities without loss of generality. Furthermore, for a given number of events in the support, one gets an upper bound on the cardinality of the latent variables that has to be considered. This upper bound will be discussed after we describe the algorithm.

Given a pDAG $\mathcal{G}$:

1. Fix the cardinalities of every latent variable to be k and the cardinality of each visible variable v to be c_v.

2. For each visible variable, enumerate all possible response functions it can have to the valuations of its parents.

   For example, if a visible v has only one parent and this parent is latent, it has $(c_v)^k$ possible response functions; we could have that v reacts with the each one of its c_v possible values for each one of the k possible valuations of its parent.

   In general, if the product of the cardinalities of the parents of v is $K(v)$, then we have $(c_v)^{K(v)}$ possibilities of response function for v.

3. For each possibility of response functions of all the visible variables of $\mathcal{G}$, we compute the support: the set of visible events that occur under that response function, for some valuation of the latent variables.

The question that is left is how to choose k, the cardinality of latent variables. First we note that, if a support containing s events is realizable by $\mathcal{G}$, it will certainly be found with the algorithm for all $k \geq s$. Thus, s is an upper bound on the values of k that we need to consider.

Assuming that all the visible variables have the same cardinality c, and considering that the total number of visible variables is $|\mathtt{vis}(\mathcal{G})|$, the maximum number of events that we can have is $c^{|\mathtt{vis}(\mathcal{G})|}$. So, to be certain that we found all compatible supports, we should choose $k = c^{|\mathtt{vis}(\mathcal{G})|}$.

Note that if $\mathcal{S}(\mathcal{G}, \vec{c}_{\text{vis}} = (2,2,2,2)) \neq \mathcal{S}(\mathcal{G}', \vec{c}_{\text{vis}} = (2,2,2,2))$, i.e., if the set of realizable *binary* supports of $\mathcal{G}$ and $\mathcal{G}'$ are different, then the sets of realizable supports will also be different when we go to higher cardinalities. This is so because we can simply gather events together (for example joining events x and y in $z = x \wedge y$) to go back to the binary case, where the difference is proven.

B.4 Proofs that comparison of unrealizable supports subsumes graphical methods as rules to show observational nondominance

Lemma 12 (The comparison of unrealizable supports subsumes the comparison of e-separation relations)**.** *Let* $\mathfrak{G}$ *and* $\mathfrak{G}'$ *be two mDAGs that have different sets of e-separation relations. Then, it is possible to find a support which is realizable by one of these mDAGs but not by the other. In fact, it is possible to find such a support even when all of the visible variables are binary, i.e.,* $\vec{c}_{vis} = (2, 2, 2, \dots)$.

Proof. The proof of Theorem 48 of Ref. [48] starts from the assumption that $\mathfrak{G}$ presents an e-separation relation that is not presented by $\mathfrak{G}'$ and shows that there must be a set of events over binary visible variables that is a realizable support of $\mathfrak{G}'$ but not of $\mathfrak{G}$. This implies that the fact that $\mathfrak{G}$ does not observationally dominates $\mathfrak{G}'$ can be shown by the comparison of unrealizable supports. □

Lemma 13 (The comparison of unrealizable supports subsumes the comparison of densely connected pairs of nodes)**.** *Let* $\mathfrak{G}$ *and* $\mathfrak{G}'$ *be two mDAGs such that* $\mathfrak{G}'$ *can be shown to* not *observationally dominate* $\mathfrak{G}$ *by comparison of densely connected pairs of nodes. Then,* $\mathfrak{G}'$ *can also be shown to not observationally dominate* $\mathfrak{G}$ *by comparison of unrealizable supports. In fact, it is possible to do so even when we are only comparing unrealizable supports where all of the visible variables are binary, i.e.,* $\vec{c}_{vis} = (2, 2, 2, \dots)$.

Proof. Assume that the nodes v and w are densely connected in $\mathfrak{G} = \{\mathcal{D}, \mathcal{B}\}$ but not in $\mathfrak{G}' = \{\mathcal{D}', \mathcal{B}'\}$.

Let S be a set of $2^{|\mathtt{vis}(\mathfrak{G})|-1}$ events such that in half of the events both X_v and X_w are 0 and in the other half both are 1, while every possible combination of 0's and 1's occurs for all of the other visible variables. For example, for the case when there are four visible

nodes x, y, v, w, S contains the events

$$\begin{aligned}
&\{X_x = 0, X_y = 0, X_v = 0, X_w = 0\},\\
&\{X_x = 1, X_y = 0, X_v = 0, X_w = 0\},\\
&\{X_x = 0, X_y = 1, X_v = 0, X_w = 0\},\\
&\{X_x = 1, X_y = 1, X_v = 0, X_w = 0\},\\
&\{X_x = 0, X_y = 0, X_v = 1, X_w = 1\},\\
&\{X_x = 0, X_y = 1, X_v = 1, X_w = 1\},\\
&\{X_x = 1, X_y = 0, X_v = 1, X_w = 1\},\\
&\{X_x = 1, X_y = 1, X_v = 1, X_w = 1\}.
\end{aligned} \tag{B.8}$$

The result of Ref. [34] says that a distribution P where X_v and X_w are perfectly correlated while all of the other visible variables are mutually independent and uniformly distributed is realizable by $\mathfrak{G}$ if v and w are densely connected in $\mathfrak{G}$. Therefore, S is a realizable support of $\mathfrak{G}$.

On the other hand, in $\mathfrak{G}'$, the nodes v and w are not densely connected. As pointed out in the proof of Theorem 6.4 of Ref. [34], this implies that $\mathfrak{G}'$ presents a nested Markov constraint between X_v and X_w. In the binary case, such a constraint takes one of two forms [67]:

$$\begin{aligned}
&P(X_v = x_v | X_A, X_w = 0, \mathrm{do}(X_B)) =\\
&P(X_v = x_v | X_A, X_w = 1, \mathrm{do}(X_B)); \text{ or}
\end{aligned} \tag{B.9}$$

$$\begin{aligned}
&P(X_v = x_v | X_A, \mathrm{do}(X_B, X_w = 0)) =\\
&P(X_v = x_v | X_A, \mathrm{do}(X_B, X_w = 1)),
\end{aligned} \tag{B.10}$$

for sets $A \subseteq \mathtt{nodes}(\mathfrak{G})$ and $B \subseteq \mathtt{nodes}(\mathfrak{G})$ that do not include v nor w, and where the do-conditionals that appear in these constraints are identifiable from observational data.

Regardless of whether it is identifiable or not, every component of a do-conditional is always larger than or equal to the corresponding component of the respective joint observational distribution. More specifically, $P(X_a | \mathrm{do}(X_b)) \geq P(X_a, X_b)$. Therefore, since the support $\mathcal{S}$ includes all events where $X_v = X_w$, then if P has support $\mathcal{S}$ we must have:

$$P(X_v = 0 | X_A, X_w = 0, \mathrm{do}(X_B)) > 0 \tag{B.11}$$

$$P(X_v = 0 | X_A, \mathrm{do}(X_B, X_w = 0)) > 0. \tag{B.12}$$

On the other hand, we know that the do-conditionals of the forms that appear in Eqs.(B.9) and (B.10), when identifiable, are obtained by the joint probability distribution over all visible variables divided by marginal conditional probability distributions, *none of which* include both X_v and X_w at the same time [57]. Since P has support $\mathcal{S}$, these marginal conditional probability distributions will always take non-zero values. The joint distribution over all visible variables, on the other hand, is zero when $X_v \neq X_w$. When the nested constraint between v and w is of the form of Eq. (B.9), the do-conditionals that appear there are identifiable, and thus

$$P\big(X_v = 0 \big| X_A, X_w = 1, \mathrm{do}\big(X_B\big)\big) = 0. \tag{B.13}$$

Together with Eq. (B.11), this shows that Eq. (B.9) cannot be satisfied by a distribution with support $\mathcal{S}$. When the nested constraint between v and w is of the form of Eq. (B.10), the do-conditionals that appear there are identifiable, and thus

$$P\big(X_v = 0 \big| X_A, \mathrm{do}\big(X_B, X_w = 1\big)\big) = 0. \tag{B.14}$$

Together with Eq. (B.12), this shows that Eq. (B.10) cannot be satisfied by a distribution with support $\mathcal{S}$.

Thus, we find a binary support that is realizable by $\mathfrak{G}$ but not by $\mathfrak{G}'$, showing that the comparison of supports subsumes the comparison of densely connected pairs as a rule to show observational inequivalence.

□

Lemma 14 (The comparison of unrealizable supports subsumes the directed-edge-free rule)**.** *Let $\mathfrak{G}$ and $\mathfrak{G}'$ be two mDAGs such that $\mathfrak{G}'$ can be shown to* not *observationally dominate $\mathfrak{G}$ by the directed-edge-free rule. Then, $\mathfrak{G}'$ can also be shown to not observationally dominate $\mathfrak{G}$ by comparison of unrealizable supports. In fact, it is possible to do so even when we are only comparing unrealizable supports where all of the visible variables are binary, i.e., $\vec{c}_{vis} = (2, 2, 2, \dots)$, and the supports are constituted by only two events.*

Proof. By the assumption that $\mathfrak{G}$ and $\mathfrak{G}'$ can be shown inequivalent by the directed-edge-free rule, it follows that $\mathfrak{G}$ is observationally equivalent to a directed-edge-free mDAG $\mathfrak{G}_{\mathrm{DEF}}$, and $\mathfrak{G}'$ is observationally equivalent to a directed-edge-free mDAG $\mathfrak{G}'_{\mathrm{DEF}}$. Both $\mathfrak{G}_{\mathrm{DEF}}$ and $\mathfrak{G}'_{\mathrm{DEF}}$ have a trivial directed structure. Since they are different, their simplicial complexes must be different. Suppose that the simplicial complex of $\mathfrak{G}$ has a facet B that is not a face of the simplicial complex of $\mathfrak{G}'$. This means that the nodes in B share a joint common cause in $\mathtt{can}(\mathfrak{G})$, but not in $\mathtt{can}(\mathfrak{G}')$.

In Example 2 of Ref. [75], it was shown that a distribution where all the variables $X_1, ..., X_n$ are perfectly correlated (i.e., where all the weight is on elements of the event space where $X_1 = X_2 = ... = X_n$) is only realizable by causal structures where all the nodes corresponding to such variables share a common ancestor. Therefore, in our case, $\mathfrak{G}$ can realize perfect correlation between the node variables in B while $\mathfrak{G}'$ cannot. Taking the case of distributions over binary variables, this implies that there are supports consisting of two events that are realizable by $\mathfrak{G}$ but not by $\mathfrak{G}'$; namely, supports where in one of the events all of the variables associated with nodes of B take the value 0 (and the remaining variables can take any value), and in the other event all of the variables associated with nodes of B take the value 1 (and the remaining variables can take any value).

□

Appendix C

Appendices for Chapter 4

Many of the proofs presented in this appendix involve integrals, because the presentation of Ref. [16] allowed for continuous random variables. To understand these in terms of discrete random variables only, it suffices to looks at those integrals as sums.

C.1 Proofs that the transformations of Section 4.2 maintain Observe&Do equivalence

C.1.1 Proof of Proposition 15

Before establishing interventional *equivalence*, we begin with a useful lemma that shows interventional *dominance*.

Lemma 23 (Interventional dominance of subgraphs)**.** *Let $\mathcal{D}$ and $\mathcal{D}'$ be two S-pDAGs with $\texttt{vis}(\mathcal{D}) = \texttt{vis}(\mathcal{D}') = V$, $\texttt{lat}(\mathcal{D}) = M$, $\texttt{lat}(\mathcal{D}') = M'$, $\texttt{sel}(\mathcal{D}) = S$ and $\texttt{sel}(\mathcal{D}') = S'$. If $\mathcal{D}$ is a subgraph of $\mathcal{D}'$, and $M \subseteq M'$, $S \subseteq S'$, then $\mathcal{D}'$ Observe&Do dominates $\mathcal{D}$. That is, $\mathcal{C}(\mathcal{D}, V \mid S) \subseteq \mathcal{C}(\mathcal{D}', V \mid S')$.*

Proof. Let P_* be a set of pairs in the SMI model of $\mathcal{D}$ produced by some kernels $P(X_y \mid X_{\texttt{pa}(y)})$. Consider the same kernels in $\mathcal{D}'$, but set the additional variables $(M' \setminus M)$ and $(S' \setminus S)$ to constant values. Since intervention is only possible on visible variables and the kernels are equal, we will recover P_*, meaning that $P_* \in \mathcal{C}(\mathcal{D}', V \mid S')$. This is true for every $P_* \in \mathcal{C}(\mathcal{D}, V \mid S)$ so $\mathcal{C}(\mathcal{D}, V \mid S) \subseteq \mathcal{C}(\mathcal{D}', V \mid S')$. □

We now restate Proposition 15.

Proposition 15. *Let $\mathcal{D}$ be an S-pDAG with $\texttt{vis}(\mathcal{D}) = V$, $\texttt{lat}(\mathcal{D}) = M$ and $\texttt{sel}(\mathcal{D}) = S$ where ...*

a) $m \in M$. Let $\mathfrak{r}_m(\mathcal{D})$ be the S-pDAG obtained from $\mathcal{D}$ by exogenizing *m, that is, by i) adding an edge $l \to k$ from every $l \in \texttt{pa}_{\mathcal{D}}(m)$ to every $k \in \texttt{ch}_{\mathcal{D}}(m)$, and ii) deleting the edges $l \to m$ for $l \in \texttt{pa}_{\mathcal{D}}(m)$. Then, $\mathcal{D}$ and $\mathfrak{r}_m(\mathcal{D})$ are Observe&Do equivalent; that is, $\mathcal{C}(\mathcal{D}, V \mid S) = \mathcal{C}(\mathfrak{r}_m(\mathcal{D}), V \mid S)$.*

b) $s \in S$. Let $\mathcal{D}_{\underline{s}}$ be the graph obtained from $\mathcal{D}$ by removing edges outgoing from s. Then, $\mathcal{D}$ and $\mathcal{D}_{\underline{s}}$ are Observe&Do equivalent; that is, $\mathcal{C}(\mathcal{D}, V \mid S) = \mathcal{C}(\mathcal{D}_{\underline{s}}, V \mid S)$.

Note that the order in which these operations are applied does not matter. Therefore, we can define a transformation, denoted by Exog*, that acts on an S-pDAG by sequentially exogenizing all of the marginalized nodes of an S-pDAG. Similarly, we can define a transformation, denoted by* Term*, that acts on an S-pDAG by terminalizing all of the selected nodes of an S-pDAG.*

Proof. **Proof of a)** This result would be immediate from Theorem 1 of [3] except that our smDGs have deterministic variables, so we instead construct a proof from first principles.

As shorthand, we will denote:

- the exogenized graph by $\tilde{\mathcal{D}} := \mathfrak{r}_m(\mathcal{D})$,
- the parents in $\mathcal{D}$ of a vertex z by $R_z := \texttt{pa}_{\mathcal{D}}(z)$,
- the sharp parents in $\mathcal{D}$ of a vertex z by $R_z^\sharp := \texttt{pa}^\sharp_{\mathcal{D}}(z)$,
- the parents in $\tilde{\mathcal{D}}$ by $\tilde{R}_z := \texttt{pa}_{\tilde{\mathcal{D}}}(z)$.

➔ **Proof of $\mathcal{C}(\mathcal{D}, V \mid S) \subseteq \mathcal{C}(\tilde{\mathcal{D}}, V \mid S)$:**

Let $P_* \in \mathcal{C}(\mathcal{D}, V \mid S)$ be a set of SMI pairs compatible with $\mathcal{D}$, and let $P(X_y \mid X_{R_y})$ for $y \in V \dot\cup M \dot\cup S$ be the kernels associated with P_*. In $\tilde{\mathcal{D}}$, the vertex m does not have parents. We will define kernels in $\tilde{\mathcal{D}}$ in such a way that the variable X_m encodes a library of values, one for each assignment to the variables $X_{\texttt{pa}_{\mathcal{D}}(m)}$, and the children of m look only at the element of this library that correspond to the actual value of X_{R_m}:

- $P'(X_m = \mathbf{x}_m) = \prod_{\alpha \in \mathfrak{X}_{R_m}} P(X_m = x_m^\alpha \mid X_{R_m} = \alpha)$, where each $\mathbf{x}_m = \times_{\alpha \in \mathfrak{X}_{R_m}} x_m^\alpha$. That is, X_m is a Cartesian product of variables, one for each assignment α to the parents R_m of m, where $P'(X_m^\alpha) = P(X_m \mid X_{R_m} = \alpha)$.
- $P'(X_w \mid \mathbf{x}_m, X_{R_w \setminus m}) = P(X_w \mid X_m = x_m^\alpha, x_{R_w \setminus m})$ for each $w \in \mathtt{ch}_{\mathcal{D}}(m)$ and $\mathbf{x}_m$
- $P'(X_w \mid X_{R_w}) = P(X_w \mid X_{R_w})$ otherwise.

Let $P'_* \in \mathcal{C}(\tilde{\mathcal{D}}, V \mid S)$ be the set of SMI pairs realizable by $\tilde{\mathcal{D}}$ with the choice of parameters above, and $Y := (V^\flat \dot\cup M \dot\cup S) \setminus (\{m\} \cup \mathtt{ch}_{\mathcal{D}}(m))$. Every pair $\big(Q(X_{V^\sharp}),\ P'_{Q(X_{V^\sharp})}(X_{V^\sharp \dot\cup V^\flat} \mid X_s = 0)\big) \in P'_*$ must obey (4.1). Therefore, the joint distribution is:

$$
\begin{aligned}
&P'_{Q(X_{V^\sharp})}(x_{V^\sharp \dot\cup V^\flat}, x_M, x_S) \\
&= Q(x_{V^\sharp}) \prod_{\alpha \in \mathfrak{X}_{R_m}} P(X_m = x_m^\alpha \mid X_{R_m^\sharp} = \alpha) \prod_{w \in \mathtt{ch}(m)} P(x_w \mid X_m = x_m^\alpha, x_{R_w \setminus m}) \prod_{y \in Y} P(x_y \mid x_{R_y^\sharp})
\end{aligned}
$$

After marginalizing over $\mathbf{x}_m$, the only term that does not go to 1 is the term corresponding to the actual value of $x_{R_m^\sharp}$. That is:

$$
\begin{aligned}
&P'_{Q(X_{V^\sharp})}(x_{V^\sharp \dot\cup V^\flat}, x_{M \setminus \{m\}}, x_S) \\
&= Q(x_{V^\sharp}) \int_{x_{R_m} \in \mathfrak{X}_{R_m}} P(X_m = x_m^{x_{R_m}} \mid X_{R_m^\sharp} = x_{R_m}) \prod_{w \in \mathtt{ch}(m)} P(x_w \mid X_m = x_m^{R_m^\sharp}, x_{R_w^\sharp \setminus m}) \prod_{y \in Y} P(x_y \mid x_{R_y^\sharp})\, dx_{R_m^\sharp} \\
&= P_{Q(X_{V^\sharp})}(x_{V^\sharp \dot\cup V^\flat}, x_{M \setminus \{m\}}, x_S).
\end{aligned}
$$

Since the joint distribution is the same, the probabilities of $X_S = \vec{0}$ will also be the same, i.e. $P'_{Q(X_{V^\sharp})}(X_S = \vec{0}) = P_{Q(X_{V^\sharp})}(X_S = \vec{0})$. Thus, by marginalizing $X_{M \setminus \{m\}}$ and selecting on $X_S = \vec{0}$, we get:

$$
P'_{Q(X_{V^\sharp})}(X_{V^\sharp \dot\cup V^\flat} \mid X_S = \vec{0}) = P_{Q(X_{V^\sharp})}(X_{V^\sharp \dot\cup V^\flat} \mid X_S = \vec{0}).
$$

Therefore, our choice of parameters gives $P'_* = P_*$. This shows that the SMI model of $\tilde{\mathcal{D}}$ contains every element of the SMI model of $\mathcal{D}$; that is, $\mathcal{C}(\mathcal{D}, V \mid S) \subseteq \mathcal{C}(\tilde{\mathcal{D}}, V \mid S)$.

➔ **Proof of $\mathcal{C}(\mathcal{D}, V \mid S) \supseteq \mathcal{C}(\tilde{\mathcal{D}}, V \mid S)$:**

Let $P'_* \in \mathcal{C}(\tilde{\mathcal{D}}, V \mid S)$ be a set of SMI pairs realizable by $\tilde{\mathcal{D}}$, and let $P'(X_y \mid X_{R_y}), y \in V \dot\cup M \dot\cup S$ be the kernels associated with P'_*. Define the following kernels in $\mathcal{D}$:

- for each $\bar{x}_m, \bar{x}_m^1$, $P(X_m = (\bar{x}_m, \bar{x}_m^1) \mid X_{R_m}) = P'(X_m = \bar{x}_m)\mathbf{1}_{\{\bar{x}_1 = X_{R_m}\}}$, where the indicator variable $\mathbf{1}_{\{Y=Z\}}$ is equal to 1 when $Y = Z$, and 0 otherwise. That is, X_m contains two variables, one that behaves as X_m in $\tilde{\mathcal{D}}$ and other that records the values of the parents of m.
- $P(X_w \mid X_m = (\bar{x}_m, \bar{x}_m^1), x_{R_w \setminus m}) = P'(X_w \mid X_m = \bar{x}_m, X_{R_m} = \bar{x}_1, x_{R_w \setminus (R_m \cup \{m\})})$ for $w \in \mathtt{ch}_{\mathcal{D}}(m)$.
- $P(X_w \mid X_{R_w}) = P'(X_w \mid X_{R_w})$ otherwise.

Letting $Y := (V^{\flat} \dot{\cup} M \dot{\cup} S) \setminus (\{m\} \cup \mathtt{ch}_{\mathcal{D}}(m))$, the joint distribution realized by this choice of kernels is:

$$
\begin{aligned}
&P_{Q(X_{V^\sharp})}(x_{V^\sharp}, x_{V^\flat}, x_{M\setminus\{m\}}, x_S) \\
&= \prod_{y \in Y} \int_{\bar{x}_m} \int_{\bar{x}_1} P'(x_y \mid x_{\tilde{R}_y^\sharp}) P'(X_m = \bar{x}_m)\, \mathbf{1}_{\{\bar{x}_1 = x_{R_m^\sharp}\}}\, Q(x_{V^\sharp}) \\
&\qquad\qquad \times \prod_{w \in \mathtt{ch}_{\mathcal{D}}(m)} P'\big(x_w \,\big|\, X_m = \bar{x}_m, X_{R_m} = \bar{x}_1, x_{R_w^\sharp \setminus (R_m^\sharp \cup \{m\})}\big)\, d\bar{x}_m\, d\bar{x}_1 \\
&= P'_{Q(X_{V^\sharp})}(x_{V^\sharp}, x_{V^\flat}, x_{M\setminus\{m\}}, x_S)
\end{aligned}
$$

By similar reasoning as in the previous case, this equality implies that the entire set of SMI pairs P'_* is realizable by $\mathcal{D}$, thus showing that $\mathcal{C}(\mathcal{D}, V \mid S) \supseteq \mathcal{C}(\tilde{\mathcal{D}}, V \mid S)$.

Proof of b)

➔ **Proof of $\mathcal{C}(\mathcal{D}, V \mid S) \supseteq \mathcal{C}(\mathcal{D}_{\underline{s}}, V \mid S)$:** Follows from Lemma 23, since $\mathcal{D}_{\underline{s}}$ is a subgraph of $\mathcal{D}$ with the same set of marginalized and selected vertices.

➔ **Proof of $\mathcal{C}(\mathcal{D}, V \mid S) \subseteq \mathcal{C}(\mathcal{D}_{\underline{s}}, V \mid S)$:** The idea of the proof is that when an edge $s \to y$ is removed, we can simply set the variable associated with the child y to the distribution that it would take in $\mathcal{D}$ when $X_S = \vec{0}$.

We will once again use R_z as shorthand for $\mathtt{pa}_{\mathcal{D}}(z)$. We will let $\underline{R}_z$ denote the parents of z in $\mathcal{D}_{\underline{s}}$, and $R_z^\sharp$ the sharp parents of z in $\mathcal{D}$, and $\underline{R}_z^\sharp$ the sharp parents of z in $\mathcal{D}_{\underline{s}}$. Let $P_* \in \mathcal{C}(\mathcal{D}, V \mid S)$ be a set of SMI pairs compatible with $\mathcal{D}$, and let $P(X_y \mid X_{R_y}), y \in V \dot{\cup} M \dot{\cup} S$ be the kernels associated with P_*. Consider now the following choice of kernels for $\mathcal{D}_{\underline{s}}$:

$$
y \in W \setminus \{s\} \Rightarrow P'(X_y \mid X_{\underline{R}_y}) = \begin{cases} P(X_y \mid X_{R_y \setminus \{s\}}, X_s = 0) & y \in \mathtt{ch}_{\mathcal{D}}(s) \\ P(X_y \mid X_{R_y}) & \text{otherwise.} \end{cases} \tag{C.1}
$$

Let $P'_* \in \mathcal{C}(\mathcal{D}_{\underline{s}}, V \mid S)$ be the set of SMI pairs compatible with $\mathcal{D}_{\underline{s}}$ that is obtained from the choice of parameters above. Every pair $\big(Q(X_{V^\sharp}),\ P'_{Q(X_{V^\sharp})}(X_{V^\sharp \dot\cup V^\flat} \mid X_s = 0)\big) \in P'_*$ must obey (4.1). Therefore, letting $Y := V^\flat \dot\cup M \dot\cup S \setminus \{s\}$, the joint distribution is:

$$
\begin{aligned}
&P'_{Q(X_{V^\sharp})}(X_{V^\sharp \dot\cup V^\flat}, X_M, X_{S\setminus s}, X_s = 0) \\
&= Q(X_{V^\sharp})\, P'(X_s = 0 \mid X_{\underline{R}_s^\sharp}) \prod_{y \in Y} P'(X_y \mid X_{\underline{R}_y^\sharp}) \\
&= Q(X_{V^\sharp})\, P(X_s = 0 \mid X_{R_s^\sharp}) \prod_{y \in Y \setminus \mathtt{ch}_{\mathcal{D}}(s)} P(X_y \mid X_{R_y^\sharp}) \prod_{y \in Y \cap \mathtt{ch}_{\mathcal{D}}(s)} P(X_y \mid X_{R_y^\sharp \setminus \{s\}}, X_s = 0) \\
&= P_{Q(X_{V^\sharp})}(X_{V^\sharp \dot\cup V^\flat}, X_M, X_{S\setminus s}, X_s = 0).
\end{aligned}
$$

Since the joint distribution is the same, the probabilities of $X_S = \vec{0}$ will also be the same, i.e. $P'_{Q(X_{V^\sharp})}(X_S = \vec{0}) = P_{Q(X_{V^\sharp})}(X_S = \vec{0})$. Thus, by marginalizing X_M and selecting on $X_{S\setminus\{s\}} = \vec{0}$, we obtain:

$$
P'_{Q(X_{V^\sharp})}(X_{V^\sharp \dot\cup V^\flat} \mid X_S = \vec{0}) = P_{Q(X_{V^\sharp})}(X_{V^\sharp \dot\cup V^\flat} \mid X_S = \vec{0}).
$$

Therefore, the choice of parameters (C.1) gives $P'_* = P_*$. This shows that the SMI model of $\mathcal{D}_{\underline{s}}$ contains every element of the SMI model of $\mathcal{D}$; that is, $\mathcal{C}(\mathcal{D}, V \mid S) \subseteq \mathcal{C}(\mathcal{D}_{\underline{s}}, V \mid S)$. □

C.1.2 Proof of Proposition 16

Proposition 16. *Let $\mathcal{D}$ be an S-pDAG with $\mathtt{vis}(\mathcal{D}) = V$, $\mathtt{lat}(\mathcal{D}) = M$ and $\mathtt{sel}(\mathcal{D}) = S$ where $\mathcal{D} = \mathtt{Exog} \circ \mathtt{Term}(\mathcal{D})$ and...*

a) Let $m_1, m_2 \in M$, and $\mathtt{ch}_{\mathcal{D}}(m_1) \cap \mathtt{ch}_{\mathcal{D}}(m_2) \cap S \neq \varnothing$. Obtain the modified S-pDAG $\tilde{\mathcal{D}}$ by merging m_1 and m_2 into one node m with $\mathtt{ch}_{\tilde{\mathcal{D}}}(m) = \mathtt{ch}_{\mathcal{D}}(m_1) \cup \mathtt{ch}_{\mathcal{D}}(m_2)$. Then, $\mathcal{D}$ and $\tilde{\mathcal{D}}$ are Observe&Do equivalent; that is, $\mathcal{C}(\mathcal{D}, V \mid S) = \mathcal{C}(\tilde{\mathcal{D}}, V \mid S)$.

b) Let $s_1, s_2 \in S$, and $\mathtt{pa}_{\mathcal{D}}(s_1) \cap \mathtt{pa}_{\mathcal{D}}(s_2) \cap M \neq \varnothing$. Obtain the modified S-pDAG $\tilde{\mathcal{D}}$ by merging s_1 and s_2 into one node s with $\mathtt{pa}_{\tilde{\mathcal{D}}}(s) = \mathtt{pa}_{\mathcal{D}}(s_1) \cup \mathtt{pa}_{\mathcal{D}}(s_2)$. Then, $\mathcal{D}$ and $\tilde{\mathcal{D}}$ are Observe&Do equivalent; that is, $\mathcal{C}(\mathcal{D}, V \mid S) = \mathcal{C}(\tilde{\mathcal{D}}, V \mid (\{s\} \cup S) \setminus \{s_1, s_2\})$.

The order in which these operations are applied does not matter. Therefore, we can define a transformation, denoted by `MergeM`*, that acts on an S-DAG by merging all of the marginalized nodes that can be merged by a). Similarly, we can define a transformation, denoted by* `MergeS`*, that acts on an S-pDAG by merging all of the selected nodes that can be merged by b).*

We will begin by proving part (a). To show that every SMI pair compatible with $\mathcal{D}$ is also compatible with $\tilde{\mathcal{D}}$ (where m_1 and m_2 are merged), our approach will be to define the variable X_m in $\tilde{\mathcal{D}}$ as the Cartesian product of variables X_{m_1} and X_{m_2} from $\mathcal{D}$. To demonstrate the converse, the idea will be that even though m_1 and m_2 are separate in $\mathcal{D}$, we can use the common selected child to make sure that the variables X_{m_1} and X_{m_2} are always equal, so that all their children can treat them as though they were a single variable.

Proof of Proposition 16 part (a). As shorthand, let $R_z := \mathtt{pa}_{\mathcal{D}}(z)$ and $\tilde{R}_z := \mathtt{pa}_{\tilde{\mathcal{D}}}(z)$. For sharp variables, let $R_z^\sharp := \mathtt{pa}_{\mathcal{D}}^\sharp(z)$ and $\tilde{R}_z^\sharp := \mathtt{pa}_{\tilde{\mathcal{D}}}^\sharp(z)$.

➔ **Proof that $\mathcal{C}(\mathcal{D}, V \mid S) \subseteq \mathcal{C}(\tilde{\mathcal{D}}, V \mid S)$:**

To reproduce a set of SMI pairs P_* that is realizable by $\mathcal{D}$ in $\tilde{\mathcal{D}}$, we simply define X_m in $\tilde{\mathcal{D}}$ as the Cartesian product of X_{m_1} and X_{m_2}. The distribution over these variables is chosen to be the same as it is in $\mathcal{D}$. It is not hard to see that this choice allows for P_* to be realized in $\tilde{\mathcal{D}}$.

➔ **Proof that $\mathcal{C}(\mathcal{D}, V \mid S) \supseteq \mathcal{C}(\tilde{\mathcal{D}}, V \mid S)$:**

Define $\mathbf{m} = \{m_1, m_2\}$.

Let $P'_* \in \mathcal{C}(\tilde{\mathcal{D}}, V \mid S)$ be an arbitrary set of SMI pairs compatible with $\tilde{\mathcal{D}}$, and let $P'(X_y \mid X_{\tilde{R}_y}), y \in V \dot\cup M' \dot\cup S$ be the kernels associated with P'_*. Consider now the following choice of kernels for $\mathcal{D}$:

- $P(X_{m_1} = x_m) = P'(X_m = x_m)$.
- $P(X_{m_2} = x_m) \sim \mathrm{Unif}(\mathfrak{X}_m)$, i.e. a uniform distribution over the domain of X_m.
- $P(X_s = 0 \mid X_{m_1} = x_{m_1}, X_{m_2} = x_{m_2}, X_{R_s \setminus \mathbf{m}}) = P'(X_s = 0 \mid X_m = x_{m_1}, X_{\tilde{R}_s \setminus m}) \mathbf{1}_{\{x_{m_1} = x_{m_2}\}}$ for every $x_{\mathbf{m}}$ and for an arbitrarily chosen $s \in S \cap \mathtt{ch}(m_1) \cap \mathtt{ch}(m_2)$, where the indicator variable $\mathbf{1}_{\{Y=Z\}}$ is equal to 1 when $Y = Z$, and 0 otherwise.
- $P(X_y \mid X_{m_1} = x_{m_1}, X_{R_y \setminus m_1}) = P'(X_y \mid X_m = x_{m_1}, X_{\tilde{R}_y \setminus \{m\}})$ for every $x_{\mathbf{m}}$ and $y \in \mathtt{ch}_{\mathcal{D}}(m_1) \setminus \{s\}$. Note that if y is also a child of m_2 in $\mathcal{D}$, this choice of parameters says that y ignores the parent m_2.
- $P(X_y \mid X_{m_2} = x_{m_2}, X_{R_y \setminus m_2}) = P'(X_y \mid X_m = x_{m_2}, X_{\tilde{R}_y \setminus \{m\}})$ for every $x_{\mathbf{m}}$ and $y \in \mathtt{ch}_{\mathcal{D}}(m_2) \setminus \mathtt{ch}_{\mathcal{D}}(m_1)$.
- $P(X_y \mid X_{R_y}) = P'(X_y \mid X_{\tilde{R}_y})$ for $y \in V \dot\cup M \dot\cup S \setminus (\mathbf{m} \cup \mathtt{ch}_{\mathcal{D}}(\mathbf{m}) \cup \{s\})$.

In this construction, m_1 carries the value of the merged variable, while m_2 is only an auxiliary copy sampled independently. The role of the selected child s is to filter to the event $X_{m_1} = X_{m_2}$, so after conditioning all children behave as though they were receiving a single parent m. Thus, given $X_s = 0$, it will follow that $X_{m_1} = X_{m_2}$.

Let $P_* \in \mathcal{C}(\mathcal{D}, V \mid S)$ be the set of SMI pairs obtained by the kernels above. We will now prove that any pair in P_* is also in $\mathcal{C}(\mathcal{D}, V \mid S)$. The idea will be to focus on the factors that involve m_1, m_2, s; selecting on $X_s = 0$ will imply that the assignment $x_{m_1} = x_{m_2}$. Thus, after integrating and conditioning, these factors will behave the same as the factors of m, s in $\mathcal{C}$. Secondly, the factors corresponding to children of m_1, m_2 will, when $s = 0$, behave just as if their parent was m in $\mathcal{C}$. And thirdly, the factors in $Y := (V^\flat \dot\cup M \dot\cup S) \setminus \{m_1, m_2, s\}$, the ones that do not involve m_1, m_2, s at all, will be unchanged. Every pair $\big(Q(X_{V^\sharp}), P_{Q(X_{V^\sharp})}(X_{V^\sharp \dot\cup V^\flat} \mid X_S = \vec{0})\big) \in P_*$ must obey (4.1); therefore, the joint distribution over all variables of $\mathcal{D}$ except for X_{m_1} and X_{m_2} must satisfy, for any assignments $x_{V^\sharp \dot\cup V^\flat \dot\cup (M \setminus \{m_1,m_2\}) \dot\cup S \setminus s}$:

$$
\begin{aligned}
&P_{Q(X_{V^\sharp})}(x_{V^\sharp \dot\cup V^\flat \dot\cup (M\setminus\{m_1,m_2\}) \dot\cup S \setminus s}, X_s = 0) \\
&= \int_{x_{m_1}} \int_{x_{m_2}} Q(x_{V^\sharp}) P(x_{m_1}) P(x_{m_2}) P(X_s = 0 | x_{m_1}, x_{m_2}, x_{R^\sharp{}_s \setminus \mathbf{m}}) \prod_{y \in Y} P(x_y | x_{R^\sharp{}_y}) dx_{m_2} dx_{m_1} \\
&= c \int_{x_{m_1}} \int_{x_{m_2}} Q(x_{V^\sharp}) P'(X_m = x_{m_1}) P'(X_s = 0 | X_m = x_{m_1}, x_{\tilde{R}^\sharp_s \setminus \{m\}}) \mathbf{1}_{\{x_{m_1} = x_{m_2}\}} \prod_{y \in Y \setminus \mathtt{ch}_{\tilde{\mathcal{D}}}(m)} P'(x_y | x_{R^\sharp{}_y}) \\
&\quad \prod_{y \in \mathtt{ch}_{\mathcal{D}}(m_1) \setminus s} P'(x_y \mid X_m = x_{m_1}, x_{\tilde{R}^\sharp_y \setminus \{m\}}) \prod_{y \in \mathtt{ch}_{\mathcal{D}}(m_2) \setminus \mathtt{ch}_{\mathcal{D}}(m_1)} P'(x_y \mid X_m = x_{m_2}, x_{\tilde{R}^\sharp_y \setminus \{m\}}) \, dx_{m_2} \, dx_{m_1} \\
&= c \int_{x_m} Q(x_{V^\sharp}) \, P'(x_m) \, P'(X_s = 0 \mid x_{\tilde{R}^\sharp_s}) \prod_{y \in Y \setminus \mathtt{ch}_{\tilde{\mathcal{D}}}(m)} P'(x_y \mid x_{\tilde{R}^\sharp_y}) \prod_{y \in Y \cap \mathtt{ch}_{\tilde{\mathcal{D}}}(m)} P'(x_y \mid x_m, x_{\tilde{R}^\sharp_y \setminus m}) \, dx_m \\
&= c P'_{Q(X_{V^\sharp})}(x_{V^\sharp \dot\cup V^\flat \dot\cup (M\setminus\{m_1,m_2\}) \dot\cup S \setminus s}, X_s = 0)
\end{aligned}
$$

where the proportionality constant $c \in \mathbb{R}$ comes from the uniform distribution on X_{m_2}.

One consequence of this equation is that $P_{Q(X_{V^\sharp})}(X_S = \vec{0}) = c P'_{Q(X_{V^\sharp})}(X_S = \vec{0})$. This implies that:

$$
P_{Q(X_{V^\sharp})}(X_{V^\sharp \dot\cup V^\flat} \mid X_S = \vec{0}) = \frac{c P'_{Q(X_{V^\sharp})}(X_{V^\sharp \dot\cup V^\flat}, X_S = \vec{0})}{c P'_{Q(X_{V^\sharp})}(X_S = \vec{0})} = P'_{Q(X_{V^\sharp})}(X_{V^\sharp \dot\cup V^\flat} \mid X_S = \vec{0}).
$$

Thus, every SMI pair of $\mathcal{C}(\tilde{\mathcal{D}}, V \mid S)$ is in $\mathcal{C}(\mathcal{D}, V \mid S)$. □

We will now prove part (b). To prove that every SMI pair compatible with $\mathcal{D}$ is also compatible with $\tilde{\mathcal{D}}$ (where s_1 and s_2 are merged), we will choose the variable X_s in $\tilde{\mathcal{D}}$ so that its domain contains all of the information from X_{s_1} and X_{s_2}. To show the converse, the idea is to use m to uniformly sample over the joint domain of all parents of s_1 and s_2, and for s_1 and s_2 to select the event where m contains the values of all parents of s_1 and s_2. Then, s_1 can depend on X_m in $\mathcal{D}$ just as s depends on X_m in $\tilde{\mathcal{D}}$.

Proof of Proposition 16 part (b). ➔ **Proof that** $\mathcal{C}(\mathcal{D}, V \mid S) \subseteq \mathcal{C}(\tilde{\mathcal{D}}, V \mid S \cup \{s\} \setminus \{s_1, s_2\})$**:**

To reproduce a set of SMI pairs P_* that is realizable by $\mathcal{D}$ in $\tilde{\mathcal{D}}$, we simply define X_s in $\tilde{\mathcal{D}}$ as the Cartesian product of X_{s_1} and X_{s_2}. The distribution over these variables is chosen to be the same as it is in $\mathcal{D}$. It is not hard to see that this choice allows for P_* to be realized in $\tilde{\mathcal{D}}$.

➔ **Proof that** $\mathcal{C}(\mathcal{D}, V \mid S) \supseteq \mathcal{C}(\tilde{\mathcal{D}}, V \mid S \cup \{s\} \setminus \{s_1, s_2\})$**:**

As shorthand, let $R_z := \mathtt{pa}_{\mathcal{D}}(z)$ and $\tilde{R}_z := \mathtt{pa}_{\tilde{\mathcal{D}}}(z)$. For sharp variables, let $R_z^\sharp := \mathtt{pa}_{\mathcal{D}}^\sharp(z)$ and $\tilde{R}_z^\sharp := \mathtt{pa}_{\tilde{\mathcal{D}}}^\sharp(z)$. Let $\mathbf{s} := \{s_1, s_2\}$, and let $S' := S \cup \{s\} \setminus \mathbf{s}$ be the set of selected vertices of $\tilde{\mathcal{D}}$.

First let $m \in R_{s_1} \cap R_{s_2} \cap M$ be a marginalized vertex that is a parent of both s_1 and s_2 in $\mathcal{D}$. In $\tilde{\mathcal{D}}$, by assumption, m is a parent of s.

Let $P'_* \in \mathcal{C}(\tilde{\mathcal{D}}, V \mid S')$ be a set of SMI pairs compatible with $\tilde{\mathcal{D}}$, and let $P'(X_y \mid X_{R_y}), y \in V \dot\cup M \dot\cup S'$ be the kernels associated with P'_*. Consider now the following choice of kernels for $\mathcal{D}$:

- Let the variable X_m associated with m in $\mathcal{D}$ equal a Cartesian product of independent variables $\bar{X}_m \times \bar{X}_1$, where $\bar{X}_m$ is distributed like X_m in $\tilde{\mathcal{D}}$ and $\bar{X}_1 \sim \mathrm{Unif}(\mathfrak{X}_{R_{s_1} \setminus R_{s_2}})$. So, $P(X_m = \langle \bar{x}_m, \bar{x}_1 \rangle) \propto P'(X_m = \bar{x}_m)$ for any $\bar{x}_m, \bar{x}_1$.
- $P(X_{s_1} = 0 \mid X_m = \langle \bar{x}_m, \bar{x}_1 \rangle, X_{R_{s_1} \setminus m} = x_{R_{s_1} \setminus \{m\}}) = \mathbf{1}_{\{\bar{x}_1 = x_{R_{s_1} \setminus R_{s_2}}\}}$ for each $x_{R_{s_1} \setminus m}, \bar{x}_m, \bar{x}_1$. That is, selecting on $X_{s_1} = 0$ establishes that $\bar{x}_1$ equals $x_{R_{s_1} \setminus R_{s_2}}$.
- $P(X_{s_2} = 0 \mid X_m = \langle \bar{x}_m, \bar{x}_1 \rangle, X_{R_{s_2} \setminus \{m\}} = x_{R_{s_2} \setminus \{m\}})$
  $= P'(X_s = 0 \mid X_m = \bar{x}_m, X_{R_{s_1} \setminus R_{s_2}} = \bar{x}_1, X_{R_{s_2} \setminus \{m\}} = x_{R_{s_2} \setminus \{m\}})$ for each $x_{R_{s_2} \setminus m}$, $\bar{x}_m$, $\bar{x}_1$. That is, X_{s_2} depends on $\bar{X}_m$, $\bar{X}_1$ and $X_{R_{s_2} \setminus \{m\}}$ similarly to how X_s depends on X_m, $X_{R_{s_1} \setminus R_{s_1}}$ and $X_{R_{s_2} \setminus \{m\}}$ in $\tilde{\mathcal{D}}$.
- $P(X_y \mid X_m = \langle \bar{x}_m, \bar{x}_1 \rangle, X_{R_y \setminus \{m\}}) = P'(X_y \mid X_m = \bar{x}_m, X_{R_y \setminus \{m\}})$ for each $\bar{x}_m, \bar{x}_1$ for $y \in \mathtt{ch}_{\mathcal{D}}(m) \setminus \mathbf{s}$.

- $P(X_y \mid X_{R_y}) = P'(X_y \mid X_{R_y})$ for $y \notin \mathtt{ch}_{\mathcal{D}}(m) \cup \{m\}$.

The next calculation shows that, before normalization, the joint distribution under selection in $\mathcal{D}$ agrees with the corresponding selected joint distribution in $\tilde{\mathcal{D}}$ up to a constant factor coming from the uniform choice of X_{m_2}. Let $P_* \in \mathcal{C}(\mathcal{D}, V \mid S)$ be the set of SMI pairs compatible with $\mathcal{D}$ that is obtained from these kernels. Every given element $\big(Q(X_{V^\sharp}), P_{Q(X_{V^\sharp})}(X_{V^\sharp \dot\cup V^\flat} \mid X_S = \vec{0})\big) \in P_*$ must obey (4.1); therefore, letting $Y = V^\flat \dot\cup (M \setminus m) \dot\cup (S \setminus \mathbf{s})$, the joint distribution $P_{Q(X_{V^\sharp})}(x_{V^\sharp \dot\cup V^\flat \dot\cup M \dot\cup (S\setminus s)} \dot\cup X_{s_1} = 0, X_{s_2} = 0)$ must, for any assignment $x_{V^\sharp \dot\cup V^\flat \dot\cup (M\setminus m) \dot\cup (S\setminus s)}$, satisfy:

$$
\begin{aligned}
&P_{Q(X_{V^\sharp})}(x_{V^\sharp \dot\cup V^\flat \dot\cup (M\setminus m) \dot\cup (S\setminus s)}, X_{s_1} = X_{s_2} = 0) \\
&= \int_{x_m} Q(X_{V^\sharp})\, P(x_m)\, P(X_{s_1} = 0 \mid X_{R^\sharp{}_{s_1}})\, P(X_{s_2} = 0 \mid X_{R^\sharp{}_{s_2}}) \prod_{y \in Y} P(X_y \mid X_{R^\sharp{}_y})\, dx_m \\
&= c \int_{\bar{x}_m, \bar{x}_1} Q(x_{V^\sharp})\, P'(X_m = \bar{x}_m)\, \mathbf{1}_{\{\bar{x}_1 = x_{R^\sharp{}_{s_1} \setminus R^\sharp{}_{s_2}}\}} \\
&\qquad \times P'(X_s = 0 \mid X_m = \bar{x}_m, X_{R^\sharp{}_{s_1} \setminus R^\sharp{}_{s_2}} = \bar{x}_1, X_{R^\sharp{}_{s_2} \setminus \{m\}} = x_{R_{s_2} \setminus \{m\}}) \qquad \text{(C.2)} \\
&\qquad \times \prod_{y \in Y \setminus \mathtt{ch}_{\mathcal{D}}(m)} P'(x_y \mid x_{R^\sharp{}_y}) \prod_{y \in Y \cap \mathtt{ch}_{\mathcal{D}}(m)} P'(x_y \mid X_m = \bar{x}_m, x_{R^\sharp{}_y \setminus \{m\}})\, d\bar{x}_m\, d\bar{x}_1 \\
&= c \int_{x_m} P'(x_m) \prod_{y \in Y} P'(x_y \mid x_{R^\sharp{}_y})\, P'(X_s = 0 \mid x_{\tilde{R}^\sharp_s})\, Q(x_{V^\sharp}) dx_m \\
&= c P'_{Q(X_{V^\sharp})}(x_{V^\sharp \dot\cup V^\flat \dot\cup (M\setminus m) \dot\cup (S\setminus s)}, X_s = 0) \qquad \text{(C.3)}
\end{aligned}
$$

where the proportionality constant $c \in \mathbb{R}$ comes from the uniform distribution on $\bar{X}_1$. The key step is that the indicator $\mathbf{1}_{\{x_{m_1} = x_{m_2}\}}$ collapses the double integral to a single integral over the common value x_m. Once this happens, the remaining factors are exactly those of $\tilde{\mathcal{D}}$, with only the constant c left over.

By marginalizing $X_{V^\sharp \dot\cup V^\flat}, X_{M\setminus m}$ and conditioning $X_{S\setminus s} = 0$, we obtain

$$P_{Q(X_{V^\sharp})}(X_S = \vec{0}) = c P'_{Q(X_{V^\sharp})}(X_{S'} = \vec{0}).$$

Together with (C.3), this implies that:

$$P_{Q(X_{V^\sharp})}(X_{V^\sharp \dot\cup V^\flat} \mid X_S = \vec{0}) = P'_{Q(X_{V^\sharp})}(X_{V^\sharp \dot\cup V^\flat} \mid X_{S'} = \vec{0}).$$

Thus $P_* = P'_*$, so the SMI of $\mathcal{D}$ dominates the SMI of $\tilde{\mathcal{D}}$. □

C.1.3 Proof of Proposition 17

We begin by restating the result.

Proposition 17. *Let $\mathcal{D}$ be an S-pDAG with $\texttt{vis}(\mathcal{D}) = V$, $\texttt{lat}(\mathcal{D}) = M$ and $\texttt{sel}(\mathcal{D}) = S$ such that $\mathcal{D} = \texttt{MergeS}\circ\texttt{MergeM}\circ\texttt{Exog}\circ\texttt{Term}(\mathcal{D})$, and $\mathcal{D}$ contains the edge $m \to s$, where $m \in M$ and $s \in S$. Let V_s denote all visible nodes which are parents of s, i.e. $V_s \coloneqq V \cap \texttt{pa}_{\mathcal{D}}(s)$, and similarly let V_m denote all visible nodes which are children of m, i.e. $V_m \coloneqq V \cap \texttt{ch}_{\mathcal{D}}(m)$. If either V_s or V_m are non-singleton sets (that is, if m has more than one visible child or s has more than one visible parent), then we perform the following operations to create a new S-pDAG $\tilde{\mathcal{D}}$:*

1. *For every pairing of visible nodes $\{a, b\}$ with $a \in V_s$ and $b \in V_m$ (i.e. for every path $a \to s \leftarrow m \to b$) add the new selected node s_{ab} and the new marginalized node m_{ab} along with the three new edges $a \to s_{ab} \leftarrow m_{ab} \to b$.*

   Note that we consider a to be part of a valid pairing if and only if $a \in V_m \cup V_s$. Consequently, if there is a path $a \to s \leftarrow m \to a$,then we must still add the path $a \to s_{aa} \leftarrow m_{aa} \to a$.

2. *Delete the edge $m \to s$.*

Then, $\mathcal{D}$ and $\tilde{\mathcal{D}}$ are Observe&Do equivalent; that is, $\mathcal{C}(\mathcal{D}, V \mid S) = \mathcal{C}(\tilde{\mathcal{D}}, V \mid S \cup S')$ where S' denotes the set of added selected nodes.

The order of application of these operations on different edges $m \to s$ does not matter. Therefore, we can define a transformation, denoted by $\texttt{SplitM->S}$, that acts on an S-pDAG by applying the steps above to each edge $m \to s$ until every edge $m \to s$ satisfies that m has only one visible child and s has only one visible parent.

For the $\supseteq$ direction, it is just a matter of having the m and s variables in $\mathcal{D}$ be the cross-product of all of the marginalized and selected variables in $\tilde{\mathcal{D}}$ respectively.

For the $\subseteq$ direction, at a high level, the added special edges in $\tilde{\mathcal{D}}$ transmit the realized values of X_{V_s} to the vertices in V_m, while the new variable X_m stores a copy of the original m-distribution for each possible value of X_{V_s}. After conditioning on the added selection variables, each $b \in V_m$ accesses exactly the copy corresponding to the realized value of X_{V_s}, so the same selected interventional distributions are recovered.

Proof. ➔ **Proof that $\mathcal{C}(\mathcal{D}, V \mid S) \supseteq \mathcal{C}(\tilde{\mathcal{D}}, V \mid S \cup S')$.**

To reproduce a set of SMI pairs P'_* that is realizable by $\tilde{\mathcal{D}}$ in $\mathcal{D}$, we simply define X_m in $\mathcal{D}$ as the Cartesian product of the variable X_m in $\tilde{\mathcal{D}}$ and every variable $X_{m_{ab}}$, and define X_s in $\mathcal{D}$ as the Cartesian product of the variable X_s in $\tilde{\mathcal{D}}$ and every variable $X_{s_{ab}}$. The distribution over these variables is chosen to be the same as it is in $\tilde{\mathcal{D}}$. It is not hard to see that this choice allows for P'_* to be realized in $\mathcal{D}$.

➔ **Proof that $\mathcal{C}(\mathcal{D}, V \mid S) \subseteq \mathcal{C}(\tilde{\mathcal{D}}, V \mid S \cup S')$.**

As shorthand, we will let $R_z := \mathtt{pa}_{\mathcal{D}}(z)$ and $\tilde{R}_z := \mathtt{pa}_{\tilde{\mathcal{D}}}(z)$ for a vertex z. For sharp variables, let $R_z^\sharp := \mathtt{pa}_{\mathcal{D}}^\sharp(z)$ and $\tilde{R}_z^\sharp := \mathtt{pa}_{\tilde{\mathcal{D}}}^\sharp(z)$.

Let $P_* \in \mathcal{C}(\mathcal{D}, V \mid S)$ be an arbitrary set of SMI pairs compatible with $\mathcal{D}$, with associated kernels $P(X_y \mid X_{R_y}), y \in V \dot\cup M \dot\cup S$. We will show that P_* is also realizable by $\tilde{\mathcal{D}}$. The idea is as follows: the kernels associated with the marginalized and selected variables $X_{m_{ab}}$ and $X_{s_{ab}}$ for each $a \in V_s, b \in V_m$ will be chosen in such a way that b has access to the value of X_a. Furthermore, the vertex m in $\tilde{\mathcal{D}}$ will be associated with a library of variables $X_m^{x_{V_s}}$, one for each assignment of values x_{V_s} for the variables in X_{V_s}. Each variable $X_m^{x_{V_s}}$ will follow the distribution of X_m in $\mathcal{D}$ conditioned on $X_s = 0$ and on $X_{V_s} = x_{V_s}$. Since the vertices $b \in V_m$ have access to the value of every variable X_{V_s}, they can look at the library provided by m and pick the variable $X_m^{x_{V_s}}$ that corresponds to the values that the variables X_{V_s} actually take.

Formally, we choose the kernels on $\tilde{\mathcal{D}}$ to be:

- For each $a \in V_s, b \in V_m$:
  - $X_{m_{ab}} \sim \mathrm{Unif}(\mathfrak{X}_a)$,
  - $P'(X_{s_{ab}} = 0 \mid X_{m_{ab}}, X_a) = \mathbf{1}_{\{X_{m_{ab}} = X_a\}}$.
- $P'(X_m = \times_{x_{V_s} \in \mathfrak{X}_{V_s}} x_m^{x_{V_s}}) = \prod_{x_{V_s} \in \mathfrak{X}_{V_s}} P(X_m = x_m^{x_{V_s}} \mid X_{V_s} = x_{V_s}, X_s = 0)$ for each assignment $\times_{x_{V_s} \in \mathfrak{X}_{V_s}} x_m^{x_{V_s}}$, i.e. X_m is a Cartesian product of variables $X_m^{x_{V_s}}$ where $P'(X_m^{x_{V_s}}) = P(X_m \mid X_{V_s}, X_s = 0)$.
- $P'(X_s \mid X_{V_s}) = P(X_s \mid X_{V_s}) = \int_{\mathfrak{X}_m} P(X_s \mid X_{V_s}, x_m)\, P(x_m)\, dx_m$.
- For each $b \in V_m$, let $m_{V_s b}$ be the set of added marginalized parents $\{m_{ab} \mid a \in V_s\}$. Then:
  $P'(X_b \mid X_m = x_m, X_{m_{V_s}} = x_{m_{V_s b}}, X_{\tilde{R}_b \setminus \{m, m_{V_s b}\}}) = P\big(X_b \,\big|\, X_m = x_m^{x_{m_{V_s b}}}, X_{\mathtt{pa}_{\mathcal{D}}(b) \setminus \{m\}}\big)$, where $x_m = \times_{x_{V_s} \in \mathfrak{X}_{V_s}} x_m^{x_{V_s}}$.

- $P'(X_y \mid X_{\tilde{R}_y}) = P(X_y \mid X_{R_y})$ for $y \in (V \setminus V_m) \dot\cup (M \setminus m) \dot\cup (S \setminus s)$.

Define $Y = V^\flat \dot\cup (M \setminus m) \dot\cup (S \setminus s)$, and $m_{V_s V_m} = \{m_{V_s b} \mid b \in V_m\}$ as the set of all added marginalized vertices. Furthermore, define the shorthand notation $\mathbf{x}_m = \times_{x_{V_s} \in \mathfrak{X}_{V_s}} x_m^{x_{V_s}}$. Let $P'_* \in \mathcal{C}(\tilde{\mathcal{D}}, V \mid S \cup S')$ be the set of SMI pairs compatible with $\tilde{\mathcal{D}}$ that is obtained from the choice of parameters above. Every given element $\big(Q(X_{V^\sharp}), P'_{Q(X_{V^\sharp})}(X_V \mid X_S = \vec{0})\big) \in P'_*$ has to obey (4.1); therefore, the joint distribution over all variables of $\tilde{\mathcal{D}}$ except for X_m and $X_{m_{V_s V_m}}$ has to be:

$$
\begin{aligned}
&P'_{Q(X_{V^\sharp})}\big(X_{V^\sharp \dot\cup V^\flat \dot\cup (M\setminus m)\dot\cup(S\setminus s)} = x_{V^\sharp \dot\cup V^\flat \dot\cup (M\setminus m)\dot\cup(S\setminus s)}, X_{S' \dot\cup s} = \vec{0}\big) \\
&= \int_{\mathbf{x}_m} \int_{x_{m_{V_s V_m}}} Q(x_{V^\sharp}) P'(X_s = 0 \mid x_{V_s{}^\sharp})\, P'(X_m = x_m) \Big\{ \prod_{b \in V_s} P'\big(x_b \mid x_m, x_{m_{V_s b}}, x_{\tilde{R}_b^\sharp \setminus \{m, m_{V_s b}\}}\big) \\
&\qquad \times \prod_{a \in V_m} P'(x_{m_{ab}}) P'(X_{s_{ab}} = 0 \mid x_{m_{ab}}, x_{V_s{}^\sharp}) \Big\} \prod_{y \in Y \setminus V_m} P'(x_y \mid x_{\tilde{R}_y^\sharp})\, dx_{m_{V_s V_m}}\, d\mathbf{x}_m \\
&= c \int_{\mathbf{x}_m} \int_{x_{m_{V_s V_m}}} Q(x_{V^\sharp}) \prod_{a,b \in V_s \times V_m} \mathbf{1}_{\{x_{m_{ab}} = x_{a^\sharp}\}} P(X_s = 0 | x_{V_s{}^\sharp}) \prod_{x'_{V_s} \in \mathfrak{X}_{V_s}} P(X_m = x_m^{x'_{V_s}} | X_{V_s^\sharp} = x'_{V_s}, X_s = 0) \\
&\qquad \times P\big(x_b \,\big|\, X_m = x_m^{x_{m_{V_s b}}}, x_{R^\sharp{}_b \setminus \{m\}}\big) \prod_{y \in Y \setminus V_m} P(x_y \mid x_{R^\sharp{}_y})\, dx_{m_{V_s V_m}}\, d\mathbf{x}_m \\
&= c \int_{x_m^{x_{V_s^\sharp}}} Q(x_{V^\sharp}) P(X_s = 0 \mid x_{V_s{}^\sharp}) P\big(X_m = x_m^{x_{V_s^\sharp}} \,\big|\, x_{V_s{}^\sharp}, X_s = 0\big) \\
&\qquad \times \int_{\mathbf{x}_m \setminus \{x_m^{x_{V_s^\sharp}}\}} \prod_{x'_{V_s} \neq x_{V_s^\sharp}} P(X_m = x_m^{x'_{V_s}} \mid X_{V_s{}^\sharp} = x'_{V_s}, X_s = 0)\, d\big(\mathbf{x}_m \setminus \{x_m^{x_{V_s^\sharp}}\}\big) \\
&\qquad \times \prod_{b \in V_m} P\big(x_b \,\big|\, X_m = x_m^{x_{V_s^\sharp}}, x_{R^\sharp{}_b \setminus \{m\}}\big) \prod_{y \in Y \setminus V_m} P(x_y \mid x_{R^\sharp{}_y})\, dx_m^{x_{V_s^\sharp}} \\
&= c \int_{x_m^{x_{V_s^\sharp}}} Q(x_{V^\sharp}) P(X_m = x_m^{x_{V_s^\sharp}}, X_s = 0 \mid x_{V_s{}^\sharp}) \prod_{y \in Y} P(x_y \mid x_{R^\sharp{}_y})\, dx_m^{x_{V_s^\sharp}} \\
&= c P_{Q(X_{V^\sharp})}\big(x_{V^\sharp \dot\cup V^\flat \dot\cup (M\setminus m)\dot\cup(S\setminus s)}, X_s = 0\big)
\end{aligned}
$$

where the proportionality constant $c \in \mathbb{R}$ comes from the uniform distribution on $X_{m_{V_s V_m}}$, and the penultimate step removes the integral over all parts of $\mathbf{x}_m$ other than $x_m^{x_{V_s^\sharp}}$, because they are independent of all other variables, and so are marginalized out.

By marginalizing M and conditioning upon $X_S = \vec{0}$ we will cancel out c and recover

the marginal selected distribution from $\mathcal{D}$, and so any SMI pairs compatible with $\mathcal{D}$ can be reproduced in $\tilde{\mathcal{D}}$. □

C.1.4 Proof of Proposition 18

We begin by restating the proposition.

Proposition 18. *Let $\mathcal{D}$ be an S-pDAG with $\texttt{vis}(\mathcal{D}) = V$, $\texttt{lat}(\mathcal{D}) = M$ and $\texttt{sel}(\mathcal{D}) = S$ such that $\mathcal{D} = \texttt{SplitM->S}\circ\texttt{MergeS}\circ\texttt{MergeM}\circ\texttt{Exog}\circ\texttt{Term}(\mathcal{D})$, and let $a \to b$ be an edge of $\mathcal{D}$ where the node a has at least one selected child and the node b has at least one marginalized parent. Let the S-pDAG $\mathcal{D}'$ be constructed by replacing $a \to b$ with a special edge; that is, with a path $a \to s_{ab} \leftarrow m_{ab} \to b$ where s_{ab} is a selected node and m_{ab} is a marginalized node. Then, $\mathcal{D}$ and $\mathcal{D}'$ are Observe&Do equivalent; that is, $\mathcal{C}(\mathcal{D}, V \mid S) = \mathcal{C}(\mathcal{D}', V \mid S \cup \{s_{ab}\})$.*

The order in which the directed edges are transformed into special edges does not matter. Therefore, we can define a transformation, denoted by `ToSpecial`*, that acts on an S-pDAG by replacing every directed edge between visible nodes with a special edge, whenever the parent has a selected child and the child has a marginalized parent.*

Proof. ➔ **Proof that $\mathcal{C}(\mathcal{D}, V \mid S) \subseteq \mathcal{C}(\mathcal{D}', V \mid S \cup \{s_{ab}\})$.**

The idea is that we can choose the distribution on $X_{m_{ab}}$ and $X_{s_{ab}}$ in such a way that, after selecting on $X_{s_{ab}} = 0$, b has access to the value of X_a.

Let $P_* \in \mathcal{C}(\mathcal{D}, V \mid S)$ be an arbitrary set of SMI pairs compatible with $\mathcal{D}$, with associated kernels $P(X_y \mid X_{\texttt{pa}_{\mathcal{D}}(y)}), y \in V \dot\cup M \dot\cup S$. Now, we define a set of kernels on $\mathcal{D}'$.

- $X_{m_{ab}} \sim \text{Unif}(\mathfrak{X}_a)$.
- $P'(X_{s_{ab}} = 0 \mid X_{m_{ab}}, X_a) = \mathbf{1}_{\{X_{m_{ab}} = X_a\}}$.
- $P'(X_b \mid X_{\texttt{pa}_{\mathcal{D}'}(b)} = x_{\texttt{pa}_{\mathcal{D}'}(b)}) = P(X_b \mid X_a = x_{m_{ab}}, X_{\texttt{pa}_{\mathcal{D}}(b)\setminus a} = x_{\texttt{pa}_{\mathcal{D}}(b)\setminus a})$. Note that b has a marginalized parent in $\mathcal{D}'$, so its distribution is allowed to be stochastic given its parent's values.
- $P'(X_y \mid X_{\texttt{pa}_{\mathcal{D}'}(y)}) = P(X_y \mid X_{\texttt{pa}_{\mathcal{D}}(y)})$ for $y \in (V \setminus b) \dot\cup M \dot\cup S$.

Then, letting $Y = (V^\flat \setminus v'^\flat) \cup M \dot\cup S$, this choice of kernels gives the joint distribution:

$$
\begin{aligned}
&P'_{Q(X_{V^\sharp})}(x_{V^\sharp \dot\cup V^\flat \dot\cup M \dot\cup S}, X_{s_{ab}} = 0) \\
&= \int_{x_{m_{ab}}} Q(x_{V^\sharp})\, P'(x_{m_{ab}})\, P'(X_{s_{ab}} = 0 \mid x_{a^\sharp}, x_{m_{ab}})\, P'(x_b \mid x_{\mathtt{pa}^\sharp_{\mathcal{D}'}(b)}) \prod_{y \in Y} P'(x_y \mid x_{\mathtt{pa}^\sharp_{\mathcal{D}'}(y)})\, dx_{m_{ab}} \\
&= c \int_{x_{m_{ab}}} Q(x_{V^\sharp})\, \mathbf{1}_{\{x_{m_{ab}} = x_{a^\sharp}\}}\, P(x_b \mid X_{a^\sharp} = x_{m_{ab}}, x_{\mathtt{pa}^\sharp_{\mathcal{D}}(b) \setminus a^\sharp}) \prod_{y \in Y} P(x_y \mid x_{\mathtt{pa}^\sharp_{\mathcal{D}}(y)})\, dx_{m_{ab}} \\
&= c\, Q(x_{V^\sharp})\, P(x_b \mid X_{a^\sharp} = x_{a^\sharp}, x_{\mathtt{pa}^\sharp_{\mathcal{D}}(b) \setminus a^\sharp}) \prod_{y \in Y} P(x_y \mid x_{\mathtt{pa}^\sharp_{\mathcal{D}}(y)}) \\
&= c P_{Q(x_{V^\sharp})}(x_{V^\sharp \dot\cup V^\flat \dot\cup M \dot\cup S})
\end{aligned}
$$

where the proportionality constant $c \in \mathbb{R}$ comes from the uniform distribution on $X_{m_{ab}}$.

From this, we conclude that $P_{Q(X_{V^\sharp})}(X_{V^\sharp \dot\cup V^\flat} \mid X_{S \dot\cup s_{ab}} = \vec{0}) = P'_{Q(X_{V^\sharp})}(X_{V^\sharp \dot\cup V^\flat} \mid X_S = \vec{0})$, thus showing that the set of SMI pairs P_* can be reproduced by $\mathcal{D}'$.

➔ **Proof that** $\mathcal{C}(\mathcal{D}, V \mid S) \supseteq \mathcal{C}(\mathcal{D}', V \mid S \cup \{s_{ab}\})$

To help with this proof, define the S-pDAG $\mathcal{D}''$ obtained by starting from $\mathcal{D}'$ and adding an edge $a \to b$. That is, $\mathcal{D}''$ has *both* a regular edge and a special edge between a and b. Clearly, $\mathcal{D}''$ interventionally dominates $\mathcal{D}'$. Now, we will show that $\mathcal{D}$ interventionally dominates $\mathcal{D}''$, which implies that $\mathcal{D}$ interventionally dominates $\mathcal{D}'$.

Let $P''_* \in \mathcal{C}(\mathcal{D}'', V \mid S \cup \{s_{ab}\})$ be an arbitrary set of SMI pairs compatible with $\mathcal{D}''$, with associated kernels $P''(X_y \mid X_{\mathtt{pa}_{\mathcal{D}''}(y)})$, $y \in V \dot\cup M \dot\cup S \cup \{s_{ab}\}$. To reproduce the same set of SMI pairs in $\mathcal{D}$, define the following kernels:

- $P(X_s \mid X_{\mathtt{pa}_{\mathcal{D}}(s)}) = P''(X_s \mid X_{\mathtt{pa}_{\mathcal{D}}(s)}) P''(X_{s_{ab}} = 0 \mid X_a)$, where s is any arbitrarily selected child of a in $\mathcal{D}$.
- $P(X_b \mid X_{\mathtt{pa}_{\mathcal{D}}(b)}) = P''(X_b \mid X_{\mathtt{pa}_{\mathcal{D}''}(b) \setminus \{m_{ab}\}}, X_{s_{ab}} = 0)$, and
- $P(X_y \mid X_{\mathtt{pa}_{\mathcal{D}}(y)}) = P''(X_y \mid X_{\mathtt{pa}_{\mathcal{D}''}(y)})$ for $y \in V \dot\cup M \dot\cup S \setminus \{s, b\}$,

Expanding $P''(X_b \mid X_{\mathtt{pa}_{\mathcal{D}''}(b)\setminus\{m_{ab}\}}, X_{s_{ab}} = 0)$, we get:

$$
\begin{aligned}
&P''(X_b \mid X_{\mathtt{pa}_{\mathcal{D}''}(b)\setminus\{m_{ab}\}}, X_{s_{ab}} = 0) \\
&= \int_{x_{m_{ab}}} P''(X_b, x_{m_{ab}} \mid X_{\mathtt{pa}_{\mathcal{D}''}(b)\setminus\{m_{ab}\}}, X_{s_{ab}} = 0)\, dm_{ab} \\
&= \int_{x_{m_{ab}}} P''(X_b \mid x_{m_{ab}}, X_{\mathtt{pa}_{\mathcal{D}''}(b)\setminus\{m_{ab}\}}, X_{s_{ab}} = 0)\, P''(x_{m_{ab}} \mid X_{\mathtt{pa}_{\mathcal{D}''}(b)\setminus\{m_{ab}\}}, X_{s_{ab}} = 0)\, dm_{ab} \\
&= \int_{x_{m_{ab}}} P''(X_b \mid x_{m_{ab}}, X_{\mathtt{pa}_{\mathcal{D}''}(b)\setminus\{m_{ab}\}})\, P''(x_{m_{ab}} \mid X_a, X_{s_{ab}} = 0)\, dm_{ab} \\
&= \int_{x_{m_{ab}}} P''(X_b \mid x_{m_{ab}}, X_{\mathtt{pa}_{\mathcal{D}''}(b)\setminus\{m_{ab}\}})\, P''(x_{m_{ab}} \mid X_a) \frac{P''(X_{s_{ab}} = 0 \mid x_{m_{ab}}, X_a)}{P''(X_{s_{ab}} = 0 \mid X_a)}\, dm_{ab}
\end{aligned}
$$

where in the third equality we used $m_{ab} \perp_{\mathcal{D}''} \mathtt{pa}_{\mathcal{D}''}(b) \setminus \{a, m_{ab}\} \mid a, s_{ab}$ and $b \perp_{\mathcal{D}''} s \mid \mathtt{pa}_{\mathcal{D}''}(b)$, and in the last equality we used Bayes' theorem.

Letting $Y := V^\flat \dot\cup M \dot\cup S \setminus \{s, b\}$, this choice of kernels for $\mathcal{D}$ gives the following joint distribution:

$$
\begin{aligned}
&P_{Q(X_{V^\sharp})}(x_{V^\sharp}, x_{V^\flat}, x_M, X_S, X_{s_{ab}} = 0) \\
&= Q(x_{V^\sharp}) P''(X_s \mid X_{\mathtt{pa}^\sharp_{\mathcal{D}}(s)}) P''(X_{s_{ab}} = 0 \mid X_{a^\sharp}) P''(X_b \mid X_{\mathtt{pa}^\sharp_{\mathcal{D}''}(b)\setminus\{m_{ab}\}}, X_{s_{ab}} = 0) \\
&\qquad \times \prod_{y \in Y} P(x_y \mid x_{\mathtt{pa}^\sharp(y)}) \\
&= Q(x_{V^\sharp})\, P''(X_s \mid X_{\mathtt{pa}^\sharp_{\mathcal{D}}(s)}) \prod_{y \in Y} P(x_y \mid x_{\mathtt{pa}^\sharp(y)}) \\
&\qquad \times \int_{x_{m_{ab}}} P''(X_b \mid x_{m_{ab}}, X_{\mathtt{pa}_{\mathcal{D}''}(b)\setminus\{m_{ab}\}})\, P''(x_{m_{ab}} \mid X_a) P''(X_{s_{ab}} = 0 \mid x_{m_{ab}}, X_a)\, dm_{ab} \\
&= P''_{Q(X_{V^\sharp})}(x_{V^\sharp}, x_{V^\flat}, x_M, X_{s_{ab}} = 0, X_{S\setminus s_{ab}}, X_{s_{ab}} = 0).
\end{aligned}
$$

By marginalizing X_M and conditioning on $X_S = \vec{0}$, we then obtain $P_{Q(X_{V^\sharp})}(X_{V^\sharp \dot\cup V^\flat} \mid X_S = \vec{0}) = P''_{Q(X_{V^\sharp})}(X_{V^\sharp \dot\cup V^\flat} \mid X_{S \dot\cup \{s_{ab}\}} = \vec{0})$, proving the result. □

C.1.5 Proof of Proposition 19

Proposition 19. *Let $\mathcal{D}$ be an S-pDAG with $\mathtt{vis}(\mathcal{D}) = V$, $\mathtt{lat}(\mathcal{D}) = M$ and $\mathtt{sel}(\mathcal{D}) = S$ such that $\mathcal{D} = \mathtt{ToSpecial} \circ \mathtt{SplitM\text{-}>S} \circ \mathtt{MergeS} \circ \mathtt{MergeM} \circ \mathtt{Exog} \circ \mathtt{Term}(\mathcal{D})$, and...*

a) There are $m_1, m_2 \in M$ such that $\mathtt{ch}(m_1) \subseteq \mathtt{ch}(m_2)$. Let $\mathcal{D}_{-m_1}$ be the subgraph of $\mathcal{D}$ on nodes other than m_1. Then, $\mathcal{D}$ and $\mathcal{D}_{-m_1}$ are Observe&Do equivalent, so $\mathcal{C}(\mathcal{D}, V \mid S) = \mathcal{C}(\mathcal{D}_{-m_1}, V \mid S)$.

b) There are $s_1, s_2 \in S$ such that $\mathtt{pa}_{\mathcal{D}}(s_1) \subseteq \mathtt{pa}(s_2)$. Let $\mathcal{D}_{-s_1}$ be the subgraph of $\mathcal{D}$ on nodes other than s_1. Then, $\mathcal{D}$ and $\mathcal{D}_{-s_1}$ are Observe&Do equivalent, so $\mathcal{C}(\mathcal{D}, V \mid S) = \mathcal{C}(\mathcal{D}_{-s_1}, V \mid S \setminus \{s_1\})$.

The order in which redundant marginalized and selected nodes are removed does not matter. Therefore, we can define a transformation, denoted by `RmvRedM`*, that acts on an S-pDAG by applying a) until no redundant marginalized nodes remain. Similarly, we can define a transformation, denoted by* `RmvRedS`*, that acts on an S-pDAG by applying b) until no redundant selected nodes remain.*

Proof. **Proof of a).** This result would be immediate from Theorem 1 of [3] except that our smDGs have deterministic variables, so we will prove it from first principles.

Proof that $\mathcal{C}(\mathcal{D}, V \mid S) \supseteq \mathcal{C}(\mathcal{D}_{-m_1}, V \mid S)$: Follows from Lemma 23 and the fact that $\mathcal{D}_{-m_1}$ is a subgraph of $\mathcal{D}$.

Proof that $\mathcal{C}(\mathcal{D}, V \mid S) \subseteq \mathcal{C}(\mathcal{D}_{-m_1}, V \mid S)$: Let $P_* \in \mathcal{C}(\mathcal{D}, V \mid S)$ be a set of SMI pairs compatible with $\mathcal{D}$ and let $P(X_y \mid X_{\mathtt{pa}_{\mathcal{D}}(y)})$, $y \in V \dot\cup M \dot\cup S$ be the kernels associated with P_*. Then, consider the following choice of kernels for $\mathcal{D}_{-m_1}$.

- Let X_m in $\mathcal{D}$ be a Cartesian product of independent variables $\bar{X}_{m_1} \times \bar{X}_{m_2}$, where each $\bar{X}_{m_i}$ is distributed in $\mathcal{D}_{-m_1}$ like X_{m_i} in $\mathcal{D}$.
- Each child of m_i then depends on $\bar{X}_{m_i}$ in place of X_{m_i}, for $i = 1, 2$.
- Every variable $y \notin \{m_1, m_2\} \cup \mathtt{ch}_{\mathcal{D}}(m_1)$ has the same kernel in $\mathcal{D}_{-m_1}$ as in $\mathcal{D}$.

It is straightforward to verify that the selected interventional distributions then are identical in $\mathcal{D}_{-m_1}$ and $\mathcal{D}$.

Proof of b). ➔ Proof that $\mathcal{C}(\mathcal{D}, V \mid S) \supseteq \mathcal{C}(\mathcal{D}_{-s_1}, V \mid S \setminus \{s_1\})$. Follows from lemma 23, since $\mathcal{D}'$ is a subgraph of $\mathcal{D}$ with the same set of marginalized vertices and with selected vertices $S \setminus \{s_1\} \subseteq S$.

➔ Proof that $\mathcal{C}(\mathcal{D}, V \mid S) \subseteq \mathcal{C}(\mathcal{D}_{-s_1}, V \mid S \setminus \{s_1\})$. Let $P_* \in \mathcal{C}(\mathcal{D}, V \mid S)$ be a set of SMI pairs compatible with $\mathcal{D}$, and let $P(X_y \mid X_{\mathtt{pa}_{\mathcal{D}}(y)})$, $y \in W$ be the kernels associated with P_*. Consider now the following choice of kernels for $\mathcal{D}_{-s_1}$:

- $P'(X_{s_2} = 0 \mid X_{\mathtt{pa}_{\mathcal{D}_{-s_1}}(s_2)}) = P(X_{s_2} = 0 \mid X_{\mathtt{pa}_{\mathcal{D}}(s_2)})\, P(X_{s_1} = 0 \mid X_{\mathtt{pa}_{\mathcal{D}}(s_1)})$; this is possible because $\mathtt{pa}_{\mathcal{D}}(s_1) \subseteq \mathtt{pa}_{\mathcal{D}}(s_2)$.
- $P'(X_y \mid X_{\mathtt{pa}_{\mathcal{D}_{-s_1}}(y)}) = P(X_y \mid X_{\mathtt{pa}_{\mathcal{D}}(y)}) \quad \forall y \in V \dot\cup M \dot\cup S \setminus \{s_1, s_2\}$.

Then:

$$
\begin{aligned}
&P'_{Q(X_{V^\sharp})}(X_{V^\sharp \dot\cup V^\flat \dot\cup M \dot\cup (S\setminus\{s_1,s_2\})}, X_{s_2} = 0) \\
&= \prod_{y \in V^\flat \dot\cup M \dot\cup S \setminus \{s_1,s_2\}} Q(X_{V^\sharp})\, P'\big(X_y \,\big|\, X_{\mathtt{pa}^\sharp_{\mathcal{D}_{-s_1}}(y)}\big)\, P'\big(X_{s_2} = 0 \,\big|\, X_{\mathtt{pa}^\sharp_{\mathcal{D}_{-s_1}}(s_2)}\big) \\
&= \prod_{y \in V^\flat \dot\cup M \dot\cup S \setminus \{s_1,s_2\}} Q(X_{V^\sharp})\, P\big(X_y \,\big|\, X_{\mathtt{pa}^\sharp_{\mathcal{D}}(y)}\big)\, P\big(X_{s_2} = 0 \,\big|\, X_{\mathtt{pa}^\sharp_{\mathcal{D}}(s_2)}\big)\, P\big(X_{s_1} = 0 \,\big|\, X_{\mathtt{pa}^\sharp_{\mathcal{D}}(s_1)}\big) \\
&= P_{Q(X_{V^\sharp})}(X_{V^\sharp \dot\cup V^\flat \dot\cup M \dot\cup (S\setminus s_1)}, X_{s_1} = 0, X_{s_2} = 0).
\end{aligned}
$$

From this we obtain:

$$
P'_{Q(X_{V^\sharp})}(X_{V_\sharp}, X_{V^\flat} \mid X_{S\setminus\{s_1\}} = 0) = P_{Q(X_{V^\sharp})}(X_{V^\sharp \dot\cup V^\flat} \mid X_S = \vec{0}).
$$

So $P'_* = P_*$, thus every set of SMI pairs compatible with $\mathcal{D}$ can be reproduced by $\mathcal{D}_{-s_1}$. □

C.2 Proof that different SLP implies interventional inequivalence (Theorem 5)

Throughout this section, we will use the following shorthand: for respective assignments $(x_{a_1}, ..., x_{a_N})$ to variables $X_{a_1}, ..., X_{a_N}$, we will write $[x_{a_1} x_{a_2} ... x_{a_N}]_{a_1 ... a_N}$, e.g. $[01]_{ab}$ is an alternative notation for the assignment $X_a = 0, X_b = 1$.

Here, we will prove that if two S-pDAGs have different selected-latent projections, then they can be distinguished by the Observe&Do probing scheme. We begin with the case where the selected-latent projections differ in their directed structure.

Lemma 24 (Different self-loop implies distinguishable)**.** *Let $\mathcal{D}$ be a S-pDAG with vertices partitioned as (V, M, S) and $\mathcal{D}'$ a S-pDAG with vertices partitioned as (V, M', S'). If the selected latent projection $\mathfrak{p}(\mathcal{D}, V \mid S)$ contains a self-loop edge (v, v) that is not in $\mathfrak{p}(\mathcal{D}', V \mid S')$ then $\mathcal{D}$ is not interventionally dominated by $\mathcal{D}'$; that is, $\mathcal{C}(\mathcal{D}, V \mid S) \nsubseteq \mathcal{C}(\mathcal{D}', V \mid S')$.*

Proof. Our approach will be to construct a model for $\mathcal{D}$ that realizes a set of SMI pairs that cannot be reproduced by $\mathcal{D}'$. In our model, the variables X_s and X_v only take two possible values each, while the variable X_m takes four possible values. Let $X_m = (X_m^1, X_m^2)$, where the variables X_m^1 and X_m^2 are each binary.

The canonical S-pDAG of $\mathfrak{p}(\mathcal{D}, V \mid S)$ contains a path $v \to s \leftarrow m \to v$. So let:

- $X_m^1 \sim \text{Bernoulli}(\frac{1}{2})$.
- $X_m^2 \sim \text{Bernoulli}(\frac{1}{2})$.
- $X_v = 0$ if and only if $X_m^1 = X_m^2 = 0$. Otherwise, $X_v = 1$.
- $X_s = 1$ if and only if $X_m^1 = X_v = 0$. Otherwise, $X_s = 0$.
- Every variable in $X_{V \dot\cup M \dot\cup S \setminus \{m,v,s\}}$ ignores its parents, and takes the constant value 0.

The observational distribution generated by this model gives $P(X_v = 0 \,|\, X_s = 0) = 0$, while all other visible variables are fixed to the value 0. If we perform an intervention that sets $X_{v^\sharp}$ to the value 0, denoted $\mathtt{do}(X_{v^\sharp} = 0)$, then the resulting distribution over the natural variable $X_{v^\flat}$ gives $P_{\mathtt{do}(X_{v^\sharp}=0)}(X_{v^\flat} = 0 \mid X_s = 0) = 0$, while all of the other visible variables are still fixed to the value 0. If we perform an intervention that sets $X_{v^\sharp}$ to the value 1, denoted $\mathtt{do}(X_{v^\sharp} = 1)$, then the resulting distribution over the natural variable $X_{v^\flat}$ gives $P_{\mathtt{do}(X_{v^\sharp}=1)}(X_{v^\flat} = 0 \mid X_s = 0) = 1/4$, while all of the other visible variables are still fixed to the value 0. This is an example where the selected distribution over $v^\flat$ depends on the value of $v^\sharp$.[1]

Now, we will prove that this set of data cannot be reproduced by $\mathcal{D}'$. In the canonical S-pDAG of $\mathfrak{p}(\mathcal{D}', V \mid S')$, we will allow for the possibility that v still has selected children and marginalized parents, but there are no edges from marginalized parents of v to selected children of v. Let $\mathtt{ch}(v)_{S'} := \mathtt{ch}(v) \cap S'$ denote the set of all selected children of v. When we perform an intervention $Q(X_{V^\sharp})$ that sets all of the variables in $X_{V^\sharp \setminus \{v^\sharp\}}$ to the value 0 and does something arbitrary to $v^\sharp$, the resulting data gives:

$$
\begin{aligned}
&P_{Q(X_{V^\sharp})}(X_{v^\sharp} X_{v^\flat} X_{M'} X_{\mathtt{ch}(v)_{S'}} = \vec{0}) \\
&= Q(X_{v^\sharp})\, P(X_{v^\flat} \mid X_{\mathtt{pa}^\sharp_{\mathcal{D}'}(v) \cap V} = \vec{0}, X_{\mathtt{pa}_{\mathcal{D}'}(v) \cap M'})\, P(X_{M'}) \\
&\qquad \times P(X_{\mathtt{ch}(v)_{S'}} = \vec{0} \,\big|\, X_{v^\sharp}, X_{\mathtt{pa}^\sharp_{\mathcal{D}'}(\mathtt{ch}(v)_{S'}) \cap V \setminus \{v\}} = \vec{0}, X_{\mathtt{pa}_{\mathcal{D}'}(\mathtt{ch}(v)_{S'}) \cap M'}).
\end{aligned}
$$

[1] For a concrete illustrative scenario, see footnote 5.

Since none of the variables of $\mathtt{ch}(v)_{S'}$ share any marginalized parent with v, by marginalizing M' we get:

$$\begin{aligned}&P_{Q(X_{V^\sharp})}(X_{v^\sharp}X_{v^\flat}, X_{\mathtt{ch}(v)_{S'}} = \vec{0})\\ &= Q(X_{v^\sharp})\, P(X_{v^\flat} \mid X_{\mathtt{pa}^\sharp_{\mathcal{D}'}(v)\cap V} = \vec{0})\, P(X_{\mathtt{ch}(v)_{S'}} = \vec{0} \mid X_{v^\sharp}, X_{\mathtt{pa}^\sharp_{\mathcal{D}'}(\mathtt{ch}(v)_{S'})\cap V\setminus\{v\}} = \vec{0}).\end{aligned}$$

This implies:

$$P_{Q(X_{V^\sharp})}(X_{v^\flat} \mid X_{\mathtt{ch}(v)_{S'}} = \vec{0}) = P_{Q(X_{V^\sharp})}(X_{v^\flat}) = P(X_{v^\flat} \mid X_{\mathtt{pa}^\sharp_{\mathcal{D}'}(v)\cap V} = \vec{0}),$$

which is independent of what is the specific intervention $Q(X_{v^\sharp})$ being performed. Therefore, $\mathcal{D}'$ cannot reproduce the set of data described, because there the selected distribution over $v^\flat$ depends on $v^\sharp$. □

Lemma 25 (Different directed edge implies distinguishable)**.** *Let $\mathcal{D}$ be a S-pDAG with vertices $V \dot\cup M \dot\cup S$ and $\mathcal{D}'$ a S-pDAG with vertices $V \dot\cup M' \dot\cup S'$. If the selected latent projections $\mathcal{G} = \mathfrak{p}(\mathcal{D}, V \mid S)$ and $\mathcal{G}' = \mathfrak{p}(\mathcal{D}', V \mid S')$ contain the same self-loops but $\mathcal{G}$ contains an edge $a \to b$ that is not in $\mathcal{G}'$ then $\mathcal{D}$ is not interventionally dominated by $\mathcal{D}'$; that is, $\mathcal{C}(\mathcal{D}, V \mid S) \notin \mathcal{C}(\mathcal{D}', V \mid S')$.*

Proof. We will specify a set of SMI pairs in $\mathcal{C}(\mathcal{D}, V \mid S)$ that is not in $\mathcal{C}(\mathcal{D}', V \mid S)$. This distribution will produce the following data:

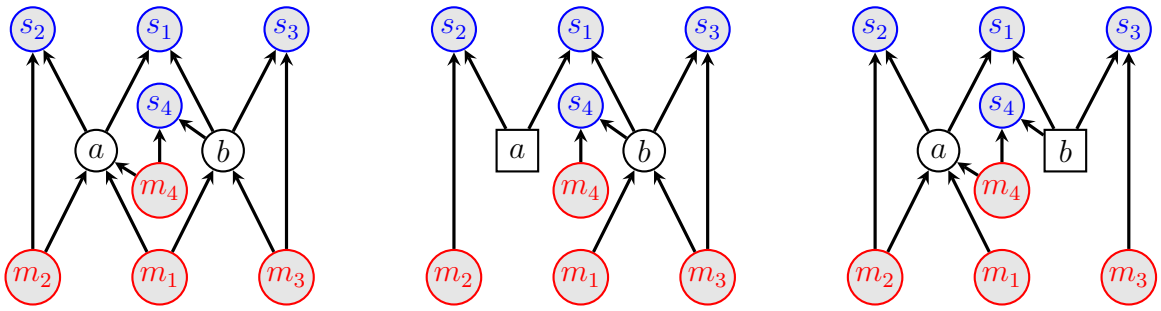
$$P(X_{V\setminus\{a\}} \mid X_S = \vec{0}, \mathtt{do}(X_a = x_a)) = [x_a]_b \cdot [0]_{V\setminus\{a,b\}} \text{ for } x_a \in \{0,1\} \tag{C.4a}$$

$$P(X_{V\setminus\{b\}} \mid X_S = \vec{0}, \mathtt{do}(X_b = 1)) = [0]_{V\setminus\{b\}}. \tag{C.4b}$$

That is, when we intervene X_a to 0 or 1, then X_b copies it, but when we intervene X_b to 1, then X_a is 0. Meanwhile, every other visible variable is fixed to a point distribution on 0. This data is clearly jointly realizable by $\mathtt{can}(\mathcal{G})$: if the edge $a \to b$ in $\mathcal{G}$ corresponds to a regular edge $a \to b$ in the canonical S-pDAG, then we can satisfy both conditions above by choosing the kernels $P(X_a \mid X_{\mathtt{pa}(a)}) = [0]_a$ and $P(X_b \mid X_{\mathtt{pa}(b)}) = X_a$. If the edge $a \to b$ in $\mathcal{G}$ corresponds to a special edge $a \to s_{ab} \leftarrow m_{ab} \to b$ in the canonical S-pDAG, then we can satisfy both conditions above by choosing the kernels $P(X_a \mid X_{\mathtt{pa}(a)}) = [0]_a$, $P(X_b \mid X_{\mathtt{pa}(b)}) = X_{m_{ab}}$, $P(X_{m_{ab}}) \sim \mathrm{Unif}(\mathfrak{X}_a)$, and $X_{s_{ab}} = 0$ if and only if $X_a = X_{m_{ab}}$.

We will now prove that the set of data of (C.4a) and (C.4b) is *not* jointly realizable by $\mathtt{can}(\mathcal{G}')$.

First notice that, since all of the other visible variables are fixed to a point distribution, the only causal paths between a and b that are open in $\texttt{can}(\mathcal{G}')$ are through marginalized parents and/or selected children. Therefore, it is sufficient to look at the subgraph over a, b, their marginalized parents and their selected children. If the SMI distributions over X_a and X_b of (C.4) are not jointly realizable by this subgraph, then (C.4) cannot be realized by the full graph. Figure C.2.1a shows the best case scenario for this subgraph, where a and b share both a selected child and a marginalized parent, there is a path $b \to s_4 \leftarrow m_4 \to a$, and both variables contain self-loops. As promised, the smDG associated with the S-pDAG of Figure C.2.1a contains no edge $a \to b$.



(a) Non-intervened graph (b) Intervention on a (c) Intervention on b

Figure C.2.1: (a) Canonical S-pDAG of the subgraph of $\mathcal{G}'$ over a and b. (b) Graph that represents the data realized by (a) after an intervention on a (when we omit the natural value $X_{a^\flat}$). (c) Graph that represents the data realized by (a) after an intervention on b (when we omit the natural value $X_{b^\flat}$).

An interventional distribution $P(X_b \mid X_S = \vec{0}, \texttt{do}(X_a = x_a)$ is realizable by Figure C.2.1a if it can be factorized as in Figure C.2.1b:

$$P(X_b \mid X_S = \vec{0}, \texttt{do}(x_a)) = \frac{P(X_b, X_S = \vec{0} \mid \texttt{do}(x_a))}{P(X_S = \vec{0} \mid \texttt{do}(x_a))}$$
$$= \frac{P(X_b, X_{s_3} = 0)\, P(X_{s_1} = 0 \mid x_a, X_b)\, P(X_{s_2} = 0 \mid x_a)\, P(X_{s_4} = 0 \mid X_b)}{P(X_S = \vec{0} \mid \texttt{do}(x_a))}. \tag{C.5}$$

We know from (C.4a) that $P(X_b = 1 \mid X_S = \vec{0}, \texttt{do}(X_a = 1)) > 0$, so from (C.5), it follows that: i) $P(X_b = 1, X_{s_3} = 0) > 0$, and ii) $P(X_{s_4} = 0 \mid X_b = 1) > 0$. Furthermore, from (C.4a), $P(X_b = 0 \mid X_S = \vec{0}, \texttt{do}(X_a = 0)) > 0$, so from (C.5), it follows that: iii) $P(X_{s_2} = 0 \mid x_a = 0) > 0$.

From (C.4a), we also have $P(X_b = 1 \,|\, X_S = \vec{0}, \texttt{do}(X_a = 0)) = 0$. Together with facts (i-iii), (C.5) then implies that we must have $P(X_{s_1} = 0 \,|\, X_a = 0, X_b = 1) = 0$.

Note that $P(X_{s_1} = 0 \,|\, X_a = 0, X_b = 1)$ is also going to appear as one of the kernels of the decomposition of a probability $P(X_a = 0 \,|\, X_S = \vec{0}, \texttt{do}(X_b = 1)$ that is realizable by Figure C.2.1a. Therefore, we conclude that $P(X_a = 0 \,|\, X_S = \vec{0}, \texttt{do}(X_b = 1) = 0$, thus contradicting (C.4b). We conclude that the observation (C.4) is *not* jointly realizable by $\texttt{can}(\mathcal{G}')$. □

So far, we have proven that two liftable smDGs that differ in their directed structure can necessarily be distinguished by the Observe&Do probing scheme. Now, we will look at liftable smDGs that share the same directed structure, but are different in their marginalized or selected independence system.

Lemma 26 (Different marginalized independence system implies distinguishable)**.** *Let $\mathcal{D}$ be a S-pDAG with vertices $V \dot\cup M \dot\cup S$ and $\mathcal{D}'$ a S-pDAG with vertices $V \dot\cup M' \dot\cup S'$. If the selected latent projections $\mathcal{G} = \mathfrak{p}(\mathcal{D}, V \mid S)$ and $\mathcal{G}' = \mathfrak{p}(\mathcal{D}', V \mid S')$ have the same directed structure, but the marginal independence system of $\mathcal{G}$ contains a maximal face V_m that is not in $\mathcal{G}'$, then $\mathcal{D}$ is not interventionally dominated by $\mathcal{D}'$; that is, $\mathcal{C}(\mathcal{D}, V \mid S) \nsubseteq \mathcal{C}(\mathcal{D}', V \mid S')$.*

Proof. Since V_m is a maximal face in $\mathcal{G}$ but not $\mathcal{G}'$, the vertices of V_m have a shared marginalized parent in $\texttt{can}(\mathcal{G})$, but not in $\texttt{can}(\mathcal{G}')$.

We consider two cases: (a) $V_m = \{a\}$, and (b) $|V_m| > 1$. In case (a), $\texttt{can}(\mathcal{G})$ has a pattern $m \to a$ that is not present in $\texttt{can}(\mathcal{G}')$. In $\mathcal{G}$, it is clearly possible to observe $X_a \mid X_S = 0 \sim \text{Bernoulli}(\frac{1}{2})$. In $\mathcal{G}'$, a cannot be in any element of the marginal independence system, so X_a is a deterministic function of its parents. Thus, X_a cannot be Bernoulli if other variables are held constant, proving the result for this case.

The rest of our proof covers case (b).

We obtain data in the following ways: (i) passive observation of all visible variables; and (ii) for each element $T \in 2^{V_m}$ of the powerset of V_m, a do-intervention on X_T, and passive observation on the remaining visible variables. Suppose that the data obtained from these interventions is:

$$\text{(i)} \quad P(X_V \mid X_{S_{\text{DAG}}} = 0) = \big\}p[0, \ldots, 0]_{V_m} + (1-p)[1, \ldots, 1]_{V_m}\big\}\,[0]_{V \setminus V_m} \tag{C.6a}$$

$$\text{(ii)} \quad P(X_{V \setminus T} \,|\, X_{S_{\text{DAG}}} = \vec{0}, \texttt{do}(X_T)) = \big\{p[0, \ldots, 0]_{V_m \setminus T} + (1-p)[1, \ldots, 1]_{V_m \setminus T}\big\}\,[0]_{V \setminus V_m} \quad \forall T \in 2^{V_m} \tag{C.6b}$$

where $p \in (0,1)$ and S_{DAG} is the set of selected vertices of the S-pDAG in consideration, be it $\texttt{can}(\mathcal{G})$ or $\texttt{can}(\mathcal{G}')$. The two equations above say that all of the variables in X_{V_m} are perfectly correlated in the data obtained from passive observations, and performing do-interventions on a subset $X_T \subseteq X_{V_m}$ does not change the marginals over the remaining variables of V_m. Meanwhile, all of the visible variables in $V \setminus V_m$ are fixed to a point distribution on 0. This data is jointly realizable by $\texttt{can}(\mathcal{G})$: if the correlation between the variables of X_{V_m} that is presented in (C.6a) is established through the shared marginalized parent, then interventions will not change the marginal distributions. We will now show that this set of data is *not* jointly realizable by $\texttt{can}(\mathcal{G}')$.

This proof will use and extend the proof of a similar result [3, Lemma 8].

From (C.6b), when we perform do-interventions on every vertex except for one, i.e. $T = V_m \setminus \{t\}$, the remaining vertex t has full support. Therefore, for any $x_t \in \mathfrak{X}_t$ and $x_T \in \mathfrak{X}_T$:

$$\begin{aligned} 0 &< P(X_t = x_t \mid X_{S'} = \vec{0}, \texttt{do}(X_T = x_T)) \\ &= \frac{P(X_{S'} = \vec{0} \mid X_t = x_t, \texttt{do}(X_T = x_T))\, P(X_t = x_t \mid \texttt{do}(X_T = x_T))}{P(X_{S'} = \vec{0} \mid \texttt{do}(X_T))} \end{aligned} \tag{C.7}$$

where we have used Bayes' theorem. This implies that the selected variables equal $\vec{0}$ with strictly positive probability for any possible assignment to X_{V_m}.

Bayes' theorem also tells us that

$$P(X_{V_m}, X_{V \setminus V_m} = 0 \mid X_{S_{\text{DAG}}} = \vec{0}) = P(X_{V_m} \mid X_{S_{\text{DAG}}} = \vec{0}) = \frac{P(X_{S_{\text{DAG}}} = \vec{0} \mid X_{V_m})}{P(X_{S_{\text{DAG}}} = \vec{0})} P(X_{V_m}). \tag{C.8}$$

Also, (C.6a) implies $P(X_{S_{\text{DAG}}} = \vec{0}) > 0$, and the reasoning after (C.7) implies $P(X_{S_{\text{DAG}} = \vec{0}} \mid X_{V_m}) > 0$. Therefore:

$$P(X_{V_m}, X_{V \setminus V_m} = 0 \mid X_{S_{\text{DAG}}} = \vec{0}) > 0 \iff P(X_{V_m}) > 0. \tag{C.9}$$

This implies that, when the interventional distributions follow (C.6b), perfect correlation between the variables of X_m in the observational distribution after selection (C.6a) is possible if and only if perfect correlation between the variables of X_m in the observational distribution is possible without selection.

Therefore, we can now appeal to the same argument from Lemma 8 of [3]: without selection, the only way for a set of vertices to be perfectly correlated is if they share a common ancestor [75], which will be the source of the perfect correlation. In this case, the

common ancestor cannot be in $V \setminus V_m$, because all such variables are fixed to 0. Furthermore, it also cannot be in V_m, otherwise the remaining variables of V_m could not continue being perfectly correlated after a do-intervention on this common ancestor, contradicting (C.6b). Therefore, such a common ancestor must be a common marginalized parent, which is *not* present in $\texttt{can}(\mathcal{G}')$. □

Now, we prove that smDGs that only differ in their selected independece system are Observe&Do distinguishable.

Lemma 27 (Different selected independence system implies distinguishable)**.** *Let $\mathcal{D}$ be a S-pDAG with vertices $V \dot\cup M \dot\cup S$ and $\mathcal{D}'$ a S-pDAG with vertices $V \dot\cup M' \dot\cup S'$. If the selected latent projections $\mathcal{G} = \mathfrak{p}(\mathcal{D}, V \mid S)$ and $\mathcal{G}' = \mathfrak{p}(\mathcal{D}', V \mid S')$ have the same directed structure and marginal independence system but the selected independence system of $\mathcal{G}$ contains a maximal face V_s that is not in $\mathcal{G}'$, then $\mathcal{D}$ is not interventionally dominated by $\mathcal{D}'$; that is, $\mathcal{C}(\mathcal{D}, V \mid S) \nsubseteq \mathcal{C}(\mathcal{D}', V \mid S')$*

Proof. By assumption, the vertices of V_s have a shared selected child in $\texttt{can}(\mathcal{G})$, but not in $\texttt{can}(\mathcal{G}')$.

We consider two cases: (a) $V_s = \{b\}$, and (b) $|V_s| > 1$. In case (a), $\texttt{can}(\mathcal{G})$ has a pattern $b \to s$ that is not present in $\texttt{can}(\mathcal{G}')$. In $\texttt{can}(\mathcal{G})$, set every variable in the graph to a fixed value, set $X_s = X_b$, and apply the soft intervention $Q(X_{b^\sharp}) \sim \text{Bernoulli}(\frac{1}{2})$. We will see that the intervened distribution $P(X_{b^\sharp} \mid X_s = 0)$ is a point distribution on 0, and thus $P(X_{b^\sharp} \mid X_s = 0) \neq Q(X_{b^\sharp})$, something that cannot happen in a model on $\texttt{can}(\mathcal{G})$.

The rest of our proof addresses case (b).

We obtain data from an intervention $Q_{\text{un}}(X_{V^\sharp})$ that sets each variable of $V_s^\sharp$ to an independent uniform distribution $\text{Bernoulli}(\frac{1}{2})$, and each variable of $V^\sharp \setminus V_s^\sharp$ to the fixed value 0. Let the resulting distribution be:

$$P_{Q_{\text{un}}(X_{V^\sharp})}(X_{V^\sharp} \mid X_{S_{\text{DAG}}} = \vec{0}) = \left(\frac{1}{2^{|V_s|-1}} \sum_{x_{V_s}} (1 - XOR(x_{V_s}))[x_{V_s}]_{V_s^\sharp} \right) [0]_{V^\sharp \setminus V_s^\sharp} \tag{C.10}$$

where $XOR(x_{V_s})$ is the parity of the sum of values in x_{V_s} (0 if even and 1 if odd), and S_{DAG} is again the set of selected vertices of the S-pDAG in consideration. In (C.10), all of the visible variables that are not in X_{V_s} are set to a point distribution on 0, and the distribution over X_{V_s} has support on all events with even parity. For example, when V_s has three vertices, (C.10) is $P_{Q_{\text{un}}(X_{V^\sharp})}(X_{V^\sharp} \mid X_{S_{\text{DAG}}} = \vec{0}) = \frac{1}{4}([000]_{V_s^\sharp} + [110]_{V_s^\sharp} + [101]_{V_s^\sharp} +$

$[011]_{V_s^\sharp})[0]_{V^\sharp \setminus V_s^\sharp}$. This data is realizable by $\texttt{can}(\mathcal{G})$, because the selected vertex $s \in S$ that is a shared child of V_s can be such that $X_S = \vec{0}$ if and only if the parity of the sum of variables X_{V_s} is even. We now show that this data is *not* realizable by $\texttt{can}(\mathcal{G}')$.

Note that, from (C.10), every strict subset $X_A \subset X_{V_s}$ has full support in $P_{Q_{\text{un}}(X_{V^\sharp})}(X_{V^\sharp} \mid X_{S_{\text{DAG}}} = \vec{0})$. If (C.10) is realizable by $\texttt{can}(\mathcal{G}')$, this implies that every individual selected vertex $s \in S_{\texttt{can}(\mathcal{G}')}$ is such that

$$P(X_s = 0 \mid X_{A^\sharp} = x_A) > 0 \quad \forall x_A \in \mathfrak{X}_A, \forall A \subset V_s. \tag{C.11}$$

By assumption, none of the selected vertices $s \in S_{\texttt{can}(\mathcal{G}')}$ is a child of every vertex of V_s. Thus, the following is a special case of the equation above:

$$P(X_s = 0 \mid X_{\texttt{pa}^\sharp(s) \cap V_s^\sharp} = x_{\texttt{pa}(s) \cap V_s}) > 0 \quad \forall x_{\texttt{pa}(s) \cap V_s} \in \mathfrak{X}_{\texttt{pa}(s) \cap V_s}. \tag{C.12}$$

Now, consider the remaining variables of $V_s^\sharp$, i.e. $X_{V_s^\sharp \setminus \texttt{pa}^\sharp(s)}$. We will show that in this distribution, these variables are conditionally independent of X_s given $X_{\texttt{pa}^\sharp(s) \cap V_s^\sharp}$. Let $S_F \subseteq S_{\texttt{can}(\mathcal{G}')}$ be the subset of selected vertices of $\texttt{can}(\mathcal{G}')$ that correspond to *facets* of $\mathcal{G}'$, that is, selected vertices that have only visible parents. The remaining selected vertices $S_{\texttt{can}(\mathcal{G}')} \setminus S_F$ are inside special edges, and have one marginalized parent and one visible parent. (i) When $s \in S_F$: if there is a chain with a visible mediary vertexbetween a variable of $X_{V_s^\sharp \setminus \texttt{pa}^\sharp(s)}$ and s or a fork with a visible common cause between a variable of $X_{V_s^\sharp \setminus \texttt{pa}^\sharp(s)}$ and s, then this visible mediary or common cause is either in $\texttt{pa}^\sharp(s) \cap V_s^\sharp$, and the path is blocked, or it is in $V^\sharp \setminus V_s^\sharp$, and it cannot establish correlation because all of the variables of $X_{V^\sharp \setminus V_s^\sharp}$ are fixed to the value 0. (ii) When $s \in S_{\texttt{can}(\mathcal{G}')} \setminus S_F$: in this case, s is part of a special edge $a \to s \leftarrow m \to b$. Since we intervene on all variables of $X_{V_s^\sharp}$ and all of the remaining visible variables are fixed to 0, the causal connection $m \to b$ is broken regardless of whether b is in V_s or not. Therefore, s behaves as a selected vertexthat has only one parent (a), and thus this case reduces back to case (i).

This conditional independence, together with the fact that $X_{V^\sharp \setminus V_s^\sharp}$ are fixed to the value 0, implies that:

$$P(X_s = 0 \mid X_{V_s^\sharp} = x_{V_s}, X_{V^\sharp \setminus V_s^\sharp} = \vec{0}) > 0 \qquad \forall x_{V_s} \in \mathfrak{X}_{V_s}. \tag{C.13}$$

Then:

$$
\begin{aligned}
&P_{Q_{\text{un}}}\big(X_{V_s^\sharp} = x_{V_s}, X_{V^\sharp \setminus V_s^\sharp} = \vec{0}, X_{S_{\text{can}(\mathcal{G}')}} = \vec{0}\big) \\
&= Q_{\text{un}}(X_{V\sharp}) \prod_{s \in S_{\text{can}(\mathcal{G}')}} P\big(X_s \,\big|\, X_{V_s^\sharp} = x_{V_s}, X_{V^\sharp \setminus V_s^\sharp} = \vec{0}\big) \\
&= \frac{1}{2^{|V_s|-1}} \prod_{s \in S_{\text{can}(\mathcal{G}')}} P\big(X_s \,\big|\, X_{V_s^\sharp} = x_{V_s}, X_{V^\sharp \setminus V_s^\sharp} = \vec{0}\big) && (Q_{\text{un}} \text{ is uniform}) \\
&> 0 \text{ for all } x_{V_s}. && \text{by (C.13)}
\end{aligned}
$$

But this implies that $P_{Q_{\text{un}}}\big(X_{V_s^\sharp} = x_{V_s}, X_{V^\sharp \setminus V_s^\sharp} = \vec{0} \mid X_{S_{\text{can}(\mathcal{G}')}} = \vec{0}\big) > 0$ for all assignments to x_{V_s}, which contradicts (C.10), proving the result. □

Note that the proofs of lemmas 25 to 27 do *not* simultaneously involve an intervention on a vertex and an observation of the natural value of the variable associated with the same vertex; that is, the expressions (C.4), (C.6), and (C.10) never simultaneously involve the $\flat$ and $\sharp$ version of a variable. Therefore, these proofs also show distinguishability of causal structures under a weaker probing scheme that will be defined in section C.3, called *Observe-or-Do probing scheme*. The same *does not* hold for lemma 24.

We are now in a position to prove the overall result below.

Theorem 5 (Same smDG ⟺ Same SMI model)**.** *The S-pDAGs $\mathcal{D}$ and $\mathcal{D}'$ with $\texttt{vis}(\mathcal{D}) = \texttt{vis}(\mathcal{D}') = V$, $\texttt{lat}(\mathcal{D}) = M$, $\texttt{lat}(\mathcal{D}') = M'$, $\texttt{sel}(\mathcal{D}) = S$ and $\texttt{sel}(\mathcal{D}') = S'$ are Observe&Do equivalent, i.e. $\mathcal{C}(\mathcal{D}, V \mid S) = \mathcal{C}(\mathcal{D}', V \mid S')$, if and only if they have the same selected-latent projection, i.e. $\mathfrak{p}(\mathcal{D}, V \mid S) = \mathfrak{p}(\mathcal{D}', V \mid S')$.*

Proof. The "if" side follows from Propositions 15 to 19. The "only if" direction follows from Lemmas 24 to 27, because differences in the self-loops, directed edges, marginal structure, and selected structure exhaust all of the possible ways that two smDGs with the same vertices can differ. □

C.3 A weaker probing scheme: Observe-or-do

To begin with, we define the Observe-or-do probing scheme, generalizing Equation (2.22) to include selection effects.

Definition 38 (SMI-Observe-Or-Do model)**.** *Let $\mathcal{D}$ be a S-pDAG with vertices partitioned into $W = V \dot\cup M \dot\cup S$. Consider the set of pairs $P_* = \left\{\left(Q(X_Z), P_{Q(X_Z)}(X_V \mid X_S = \vec{0})\right) : Z \subseteq V, Q(X_Z) \in \mathcal{P}(\mathfrak{X}_Z)\right\}$ where $\mathcal{P}(\mathfrak{X}_Z)$ is the set of probability measures on the domain of X_Z.*

We say that P_* is a set of SMI-Observe-Or-Do pairs compatible with $\mathcal{D}$ if each $\left(Q(X_Z), P_{Q(X_Z)}(X_V \mid X_S = \vec{0})\right) \in P_*$ satisfies:

$$P_{Q(X_Z)}(X_V \mid X_S = \vec{0}) = \frac{\int_{x_M} P_{Q(X_Z)}(X_V, X_M, X_S = \vec{0})\, dx_M}{\int_{x_V}\int_{x_M} P_{Q(X_Z)}(X_V, X_M, X_S = \vec{0})\, dx_M\, dx_V}$$
$$\text{for} \quad P_{Q(X_Z)}(X_V, X_M, X_S) = Q(X_Z) \prod_{y \in V \setminus Z} P(X_y \mid X_{\mathtt{pa}(y)}) \prod_{y \in M \dot\cup S} P(X_y \mid X_{\mathtt{pa}(y)}).$$

The collection of all sets of SMI-Observe-Or-Do pairs is called the SMI-Observe-Or-Do model.

Note that in this definition each visible vertex is *either* intervened upon (in which case it belongs to Z) *or* observed (in which case it belongs to $V \setminus Z$). Contrast this with definitions 29 and 30 for the Observe&Do probing scheme.

Proposition 31. *The S-pDAGs Figure 4.4.1a and Figure 4.4.1b are not distinguishable by the Observe-or-do probing scheme.*

Proof. Let $\mathcal{D}$ be the graph $m \to v \to s$; $m \to s$ from Figure 4.4.1a, and $\mathcal{D}'$ be the graph $m \to v \to s$ from Figure 4.4.1b. $\mathcal{D}'$ is a subgraph of $\mathcal{D}$, so clearly the SMI-Obs-Or-Do model of $\mathcal{D}$ dominates that of $\mathcal{D}'$. For the remainder of the proof, we will establish that the opposite is also true.

Let P_* be a set of SMI-Obs-Or-Do pairs in the SMI-Obs-Or-do model of $\mathcal{D}$, that is generated from kernels $P(X_s \mid X_m, X_v)$, $P(X_m)$ and $P(X_v \mid X_m)$. Make the following choice of kernels for $\mathcal{D}'$:

- $P'(X_m) = P(X_m)$,
- $P'(X_s = 0 \mid X_v) = \int_{x_m} P(X_s = 0 \mid x_m, X_v)\, P(x_m)\, dx_m$,
- $P'(X_v \mid X_m)$
$$= \begin{cases} c\frac{P(X_v, X_s=0 \mid X_m)}{P'(X_s=0 \mid X_v)} = c\frac{P(X_v \mid X_m) P(X_s=0 \mid X_m, X_v)}{\int_{x_m} P(X_s=0 \mid x_m, X_v)\, P(x_m)\, dx_m} & \text{if } P'(X_s = 0 \mid X_v) \neq 0 \\ 0 & \text{if } P'(X_s = 0 \mid X_v) = 0, \end{cases}$$

where the constant of proportionality c is chosen so that $\int_{x_v} P'(x_v \mid X_m)\, dx_v = 1$.

Now, we will show that this choice of kernels can reproduce P_* in $\mathcal{D}'$. First, we show that it gives rise to all of the correct pairs $\big(Q(X_Z), P_{Q(X_Z)}(X_V \mid X_S = \vec{0})\big)$ when $Z = \{v\}$, that is, when we perform an intervention $Q(X_v)$ on v.

$$\begin{aligned} P'_{Q(X_v)}(X_v \mid X_s = 0) &= \frac{P'_{Q(X_v)}(X_v, X_s = 0)}{P'_{Q(X_v)}(X_s = 0)} = \frac{Q(X_v)\, P'(X_s = 0 \mid X_v)}{\int_{x'_v} Q(x'_v)\, P'(X_s = 0 \mid X_v = x'_v)\, dx'_v} \\ &= \frac{Q(X_v) \int_{x_m} P(X_s = 0 \mid x_m, X_v)\, P(x_m)\, dx_m}{\int_{x'_v} Q(x'_v) \int_{x'_m} P(X_s = 0 \mid x'_m, x'_v)\, P(x'_m)\, dx'_m\, dx'_v} \\ &= \frac{P_{Q(X_v)}(X_v, X_s = 0)}{P_{Q(X_v)}(X_s = 0)} = P_{Q(X_v)}(X_v \mid X_s = 0). \end{aligned}$$

Therefore, our choice of kernels for $\mathcal{D}'$ can reproduce all of the pairs $\big(Q(X_Z), P_{Q(X_Z)}(X_V \mid X_S = \vec{0})\big) \in P_*$ where $Z = \{v\}$. Now, we look at the pair where $Z = \varnothing$, that is, when we make a passive observation of v.

$$\begin{aligned} P'(X_v \mid X_s = 0) &= \frac{P'(X_v, X_s = 0)}{P'(X_s = 0)} = \frac{\int_{x_m} P'(X_s = 0 \mid X_v)\, P'(X_v \mid x_m)\, P(x_m)\, dx_m}{\int_{x'_v} \int_{x'_m} P'(X_s = 0 \mid x'_v)\, P'(x'_v \mid x'_m)\, P(x'_m)\, dx'_m\, dx'_v} \\ &= \frac{P'(X_s = 0 \mid X_v) \frac{P(X_v, X_s = 0)}{P'(X_s = 0 \mid X_v)}}{\int_{x'_v} P'(X_s = 0 \mid X_v = x'_v) \frac{P(x'_v, X_s = 0)}{P'(X_s = 0 \mid X_v = x'_v)}\, dx'_v} = \frac{P(X_v, X_s = 0)}{\int_{x'_v} P(x'_v, X_s = 0)\, dx'_v} \\ &= \frac{P(X_v, X_s = 0)}{P(X_s = 0)} = P(x_v \mid X_s = 0). \end{aligned}$$

Therefore, our choice of kernels for $\mathcal{D}'$ reproduces the entire set of SMI pairs P_*. □

C.4 Proofs of rules to show observational equivalence

C.4.1 Proofs for section 4.5.1

We begin by restating Proposition 21.

Proposition 21. *Let $\mathcal{G} = (V, \mathcal{E}, \mathcal{L}, \mathcal{S})$ be a liftable smDG that includes a maximal selected face $V_s \in \mathcal{S}$. Assume that:*

- *V_s includes all its parents, i.e.* $\mathtt{pa}_{\mathcal{G}}(V_s) \subseteq V_s$*;*
- *every node* $v \in V_s$ *belongs to some marginalized face of* $\mathcal{L}$*; and*
- *every marginalized face* $V_m \in \mathcal{L}$ *that intersects with* V_s *is subsumed by* V_s*, i.e.* $V_s \cap V_m \neq \varnothing \implies V_m \subseteq V_s$*.*

Let $\mathcal{G}' = (V, \mathcal{E}, \mathcal{L}', \mathcal{S})$ *be the smDG constructed from* $\mathcal{G}$ *by adding* $\mathrm{cl}(V_s)$ *to the marginalized independence system, i.e.* $\mathcal{L}' = \mathcal{L} \cup \mathrm{cl}(V_s)$*. Then,* $\mathcal{G}$ *and* $\mathcal{G}'$ *are observationally equivalent.*

Proof. This proof will use the idea of *district factorization* [31]. For the definition of districts and district factorization, we refer the reader to that paper.

Since every marginalized face $V_m \in \mathcal{L}$ such that $V_s \cap V_m \neq \varnothing$ is also such that $V_m \subseteq V_s$, the vertices of V_s are *not* part of the same district as any vertex outside of V_s in $\mathcal{G}$. Thus, using district factorization together the fact that $\mathtt{pa}_{\mathcal{G}}(V_s) \subseteq V_s$, we obtain:

$$P(X_V \mid X_S = 0) = P(X_{V_s} \mid X_S = 0)\, P(X_{V \setminus V_s} \mid X_{V_s}, X_S = 0). \tag{C.14}$$

By assumption, all of the vertices $v \in V_s$ have access to local randomness (because each one belongs to some marginalized face in $\mathcal{G}$). Together with the fact that all of them share one selected child $s \in S$, this implies that the S-pDAG $\mathtt{can}(\mathcal{G})$ can realize *any* conditional distribution $P(X_{V_s} \mid X_S = 0)$, by simply setting the variables in X_{V_s} to be uniformly distributed and choosing the selection condition on X_s according to the desired $P(X_{V_s} \mid X_S = 0)$.

In $\mathcal{G}'$, it is also true that the vertices of V_s are not in the same district as any vertex outside of V_s. Therefore, distributions realizable by $\mathcal{G}'$ can also be factorized as in (C.14). Furthermore, the only difference between $\mathtt{can}(\mathcal{G})$ and $\mathtt{can}(\mathcal{G}')$ is the addition of a common marginalized parent between the vertices of V_s, a modification that can only possibly affect the term $P(X_{V_s} \mid X_S = 0)$ of (C.14). However, as we saw, $\mathtt{can}(\mathcal{G})$ itself can already realize *any* conditional distribution $P(X_{V_s} \mid X_S = 0)$; therefore, the addition of this extra marginalized vertex cannot change the set of realizable distributions.

□

C.4.2 Proofs for Section 4.5.2

In this subsection, we present the proof of Proposition 22.

Proposition 22. *Let $\mathcal{G} = (V, \mathcal{E}, \mathcal{L}, \mathcal{S})$ be a liftable smDG. Suppose that $\mathcal{G}$ contains an edge $a \to b$ such that a is an element of some $V_s \in \mathcal{S}$ and b is an element of some $V_m \in \mathcal{L}$. Further suppose that a and b belong to the same face of the marginalized independence system, that is, that a is also an element of V_m. Then, $\mathcal{G}$ is observationally equivalent to the smDG $\mathcal{G}'$ obtained by starting from $\mathcal{G}$ and removing the edge $a \to b$.*

Proof. Let $\mathcal{D} = \texttt{can}(\mathcal{G})$ be the canonical S-pDAG associated with $\mathcal{G}$. Then, if its vertices are partitioned as (V, M, S), $\mathcal{D}$ contains a path $a \to s \leftarrow m \to b$ where $\texttt{pa}_{\mathcal{D}}(m) = \texttt{ch}_{\mathcal{D}}(s) = \varnothing$, $\texttt{pa}_{\mathcal{D}}(s) = \{a, m\}$ and $\texttt{ch}_{\mathcal{D}}(m) = \{s, b\}$. Furthermore, a and b share a common parent m', and $a, b \in V$, $s \in S$, and $m, m' \in M$. Let $\mathcal{D}'$ be the S-pDAG obtained by removing the vertex m from $\mathcal{D}$. The SLP of $\mathcal{D}'$ is $\mathcal{G}'$.

Clearly $\mathcal{D}$ can reproduce any distribution from $\mathcal{D}'$ since $\mathcal{D}$ is a subgraph of $\mathcal{D}$, so what remains is to prove that the converse is also true.

Let $P(X_V | X_S = 0)$ be an arbitrary observational distribution that is realizable by $\mathcal{D}$, with associated kernels $P(X_y \mid X_{\texttt{pa}_{\mathcal{D}}(y)}), y \in V \dot\cup M \dot\cup S$.

Now, we will define a set of kernels P' on the S-pDAG $\mathcal{D}'$ as follows:

- The vertex m' in $\mathcal{D}'$ contains two variables: $X_{m'} = (\bar{X}_{m'}, X_m)$. They are distributed as $X_{m'}$ and X_m from $\mathcal{D}$, i.e., $P'(X_{m'} = (\bar{x}_{m'}, x_m)) = P(X_{m'} = \bar{x}_{m'}) P(X_m = x_m)$.
- The variable X_b responds to $\bar{X}_{m'}$ and X_m as it used to respond to $X_{m'}$ and X_m in $\mathcal{D}$, i.e., $P'\left(X_b | X_{m'} = (\bar{x}_{m'}, x_m), X_{\texttt{pa}_{\mathcal{D}}(b) \setminus \{m', m\}}\right) = P\left(X_b | X_{m'} = \bar{x}_{m'}, X_m = x_m, X_{\texttt{pa}_{\mathcal{D}}(b) \setminus \{m', m\}}\right)$.
- The variable X_s responds to X_a as $P'(X_s = 0 | X_a) = P(X_s = 0 | X_a)$.
- The variable X_a is such that:

$$P'\left(X_a | X_{m'} = (\bar{x}_{m'}, x_m), X_{\texttt{pa}_{\mathcal{D}}(a) \setminus \{m'\}}\right) = \begin{cases} \frac{P(X_s = 0 | X_a, X_m = x_m)}{P(X_s = 0 | X_a)} P\left(X_a | X_{m'} = \bar{x}_{m'}, X_{\texttt{pa}_{\mathcal{D}}(a) \setminus \{m'\}}\right) & \text{if } P(X_s = 0 | X_a) \neq 0 \\ 0 & \text{if } P(X_s = 0 | X_a) = 0 \end{cases}.$$

- The remaining variables respond to their parents in the same way as they did in $\mathcal{D}$.

With these kernels, the distribution $P'(X_V, X_S = 0)$ realized by $\mathcal{D}'$ is:

$$
\begin{aligned}
&P'(X_V, X_S = 0) \\
&= \int_{x_{m'}} P'(X_a|X_{m'}, X_{\mathtt{pa}_{\mathcal{D}}(a)\setminus\{m'\}})P'(X_b|X_{m'}, X_{\mathtt{pa}_{\mathcal{D}}(b)\setminus\{m',m\}})P'(X_s = 0|X_a)P(X_{m'})\cdot \\
&\cdot \prod_{y\in V\cup M\setminus\{a,b,m',m\}} P'(X_y|X_{\mathtt{pa}_{\mathcal{D}}(y)}) \prod_{w\in S\setminus\{s\}} P'(X_w = 0|X_{\mathtt{pa}_{\mathcal{D}}(w)})dx_{m'} \qquad \text{(C.15)} \\
&= \int_{\bar{x}_{m'},x_m} P(X_s = 0|X_a, X_m)P(X_a|X_{m'}, X_{\mathtt{pa}_{\mathcal{D}}(a)\setminus\{m'\}})P(X_b|X_{m'}, X_m, X_{\mathtt{pa}_{\mathcal{D}}(b)\setminus\{m',m\}}) \\
&\cdot P(X_{m'})P(X_m) \prod_{y\in V\cup M\setminus\{a,b,m',m\}} P(X_y|X_{\mathtt{pa}_{\mathcal{D}}(y)}) \prod_{w\in S\setminus\{s\}} P(X_w = 0|X_{\mathtt{pa}_{\mathcal{D}}(w)})d\bar{x}_{m'}dx_m \qquad \text{(C.16)} \\
&= P(X_V, X_S = 0). \qquad \text{(C.17)}
\end{aligned}
$$

Therefore, every observational distribution $P(X_V, X_S = 0)$ realizable by $\mathcal{D}$ can be reproduced by $\mathcal{D}'$, and thus $\mathcal{D}$ and $\mathcal{D}'$ are observationally equivalent. □

C.4.3 Proofs for section 4.5.3

In this subsection, we will prove Proposition 23, that says that when an smDG $\mathcal{G}$ satisfies certain conditions, one can split a selected face into many singleton selected faces without any change to the SMO model:

Proposition 23. *Let $\mathcal{G} = (V, \mathcal{E}, \mathcal{L}, \mathcal{S})$ be a liftable smDG, let $V_s \in \mathcal{S}$ be one of its maximal selected faces, and let $\mathcal{G}_{\mathtt{an}(V_s)}$ be the subgraph of $\mathcal{G}$ over the set V_s and all of its ancestors. Suppose that:*

a) There are no cycles in $\mathcal{G}_{\mathtt{an}(V_s)}$.

b) In $\mathcal{G}_{\mathtt{an}(V_s)}$, nodes that share a child or are in the same selected face are neighbours or are in the same marginalized face.

c) All of the maximal marginalized faces of $\mathcal{L}$ that include nodes of $\mathtt{an}_{\mathcal{G}}(V_s)$ are disjoint.

d) For each maximal marginalized face $V_m \in \mathcal{L}$ that includes nodes of $\mathtt{an}_{\mathcal{G}}(V_s)$, the parents of V_m that are not in V_m are shared by all the nodes of V_m, i.e. $\mathtt{pa}_{\mathcal{G}}(V_m)\setminus V_m \subseteq \bigcap_{v\in V_m} \mathtt{pa}_{\mathcal{G}}(v)$.

Construct the independence system $\mathcal{S}'$ by starting from $\mathcal{S}$ and then deleting V_s and all of its subsets. Define the smDG $\mathcal{G}' = (V, \mathcal{E}, \mathcal{L}, \mathcal{S}')$. Then, $\mathcal{G}$ and $\mathcal{G}'$ are observationally equivalent.

We will make this proof in steps, presenting several lemmas first.

Lemma 28. *Let $\mathcal{D}$ be a S-pDAG with vertices partitioned as (V, M, S), and let $s \in S$ be a selected vertex of $\mathcal{D}$. If the set of vertices $\mathtt{an}_{\mathcal{D}}(s)$ (including s itself) contains no unshielded colliders, then $\mathcal{D}$ is observationally equivalent to the S-pDAG $\mathcal{D}_{-s}$ obtained by starting from $\mathcal{D}$ and deleting s.*

Proof. Let $P(X_V \mid X_s = 0)$ be a distribution that is realizable by $\mathcal{D}$, and let $v, v' \in V$ be two vertices such that $X_{v'}$ is conditionally *dependent* of X_v in P given the parents of v and selection on s. That is, $P(X_v \mid X_{\mathtt{pa}_{\mathcal{D}}(v)}, X_{v'}, X_s = 0) \neq P(X_v \mid X_{\mathtt{pa}_{\mathcal{D}}(v)}, X_s = 0)$. Then, by the d-separation criterion, there exists a path $\pi : v \dashrightarrow v'$ that is active in $\mathcal{D}$ given $\mathtt{pa}_{\mathcal{D}}(v)$ and s.

Suppose that the path π contains colliders. All of the colliders $v_1 \to z \leftarrow v_2$ of π need to be such that z is an ancestor of either $\mathtt{pa}_{\mathcal{D}}(v)$ or s, otherwise π would not be active given these vertices. The collider that is closest to v cannot be an ancestor of $\mathtt{pa}_{\mathcal{D}}(v)$ (or it would induce a cycle), so it is an ancestor of s. This implies that $v \in \mathtt{an}_{\mathcal{D}}(s)$, which in turn implies that *every* collider in π must be in $\mathtt{an}_{\mathcal{D}}(s)$.

By assumption, there are no unshielded colliders in $\mathtt{an}_{\mathcal{D}}(s)$. Therefore, if $v_1 \to z \leftarrow v_2$ is a collider in π, then $\mathcal{D}$ includes either the arrow $v_1 \to v_2$ or $v_1 \leftarrow v_2$. Consider now the path π' that is the same as π up until v_1, but then it goes directly to v_2 (without passing through z). This path is active in $\mathcal{D}$ given $\mathtt{pa}_{\mathcal{D}}(v) \cup \{s\}$: its non-collider vertices are a subset of those in π, and if π' has a collider in either v_1 or v_2, this is not a problem because v_1 and v_2 are both ancestors of s (since they are parents of z).

Since π has finite length, by repeating this process until we have no colliders, we will eventually obtain a path with no colliders, that is still active given $\mathtt{pa}(v) \cup \{s\}$. Since the path is active given $\mathtt{pa}(v) \cup \{s\}$, and it has no colliders, it can only be of the form $v \dashrightarrow v'$.

That is, v' is a descendant of v in $\mathcal{D}$. Since the node s does not have children, v' is also a descendant of v in $\mathcal{D}_{-s}$. So, if we define the distribution $P'(X_V) := P(X_V \mid X_s = 0)$, we have that $P'(X_V \mid X_{\mathtt{pa}(v)}, X_{v'}) \neq P(X_v \mid X_{\mathtt{pa}(v)})$ implies that $v' \in \mathtt{des}_{\mathcal{D}_{-s}}(v)$. This is equivalent to saying that $P'(X_V)$ is realizable by $\mathcal{D}_{-s}$, which establishes the observational equivalence of $\mathcal{D}$ and $\mathcal{D}_{-s}$.

□

Lemma 29. *Let $\mathcal{G}$ be an smDG that satisfies the conditions of Proposition 23.*

Let $\tilde{\mathcal{D}}$ be constructed from $\texttt{can}(\mathcal{G})$ by performing the following transformations:

A) Identify every path in $\texttt{can}(\mathcal{G})$ of the form $a \to s' \leftarrow m' \to b$ where $a, b \in V$, a and b do not *share a marginalized parent, $s' \in S$ and $m' \in M$. For each such path, delete the edge $m' \to s'$ and add the edge $a \to b$.*

B) Prior to the transformation, m' had precisely two children, $\{s', b\}$. Now m' has only one child, namely, b. If b has any other marginalized parents, remove the marginalized vertex m' per `RmvRedM`.

C) For every remaining path $a \to s \leftarrow m \to b$ with $s \in S$ and $m \in M$, where a and b do *share a marginalized parent, delete the vertex m.*

Then $\tilde{\mathcal{D}}$ is a S-pDAG, and it is observationally equivalent to $\texttt{can}(\mathcal{G})$.

Proof. Condition a) of Proposition 23 says that $\mathcal{G}$ contains no cycles, so step A) does not create cycles. Due to Proposition 18, we know that step A) preserves Observe&Do equivalence. Due to Proposition 22, we know that step B) preserves observational equivalence. □

Lemma 30. *Let $\mathcal{G}$ be an smDG that satisfies the conditions of Proposition 23, and let $\tilde{\mathcal{D}}$ be defined as in Lemma 29. Then, $\mathcal{M}(\tilde{\mathcal{D}}, V \mid S) = \mathcal{M}(\tilde{\mathcal{D}}_{-s}, V \mid S \setminus \{s\})$.*

Before proving Lemma 30, let us spell out what the conditions of Proposition 23 will imply about the structure of $\tilde{\mathcal{D}}$.

Lemma 31. *Let $\mathcal{G}$ be an smDG that satisfies the conditions of Proposition 23 and define $\tilde{\mathcal{D}}$ as in Lemma 30.*

Let M' be the marginalized vertices of $\texttt{an}_{\tilde{\mathcal{D}}}(s)$ that have only visible children (i.e. none of the vertices of M' are part of a special edge).

Then, $\tilde{\mathcal{D}}$ will satisfy the following:

1. *It is invariant under the operations* `SplitM->S`, `MergeS`, `MergeM`, `Exog`, `Term` *and* `RmvRedM`.
2. *a') $\tilde{\mathcal{D}}$ contains no special edges.*

b') *Visible vertices in* $\texttt{an}(s)$ *that share a child or are in the same selected face in* $\mathcal{G}_{\texttt{an}(s)}$ *are neighbours or share a marginalized parent in* $\tilde{\mathcal{D}}$.

c') *Letting* $V'_m := \texttt{ch}_{\tilde{\mathcal{D}}}(m)$, *the sets* $\{V'_m\}_{m \in M'}$ *are disjoint.*

d') *For every* $m \in M'$, *the parents of* V'_m *that are not in* V'_m *are parents shared by every vertex in* V'_m, *i.e.* $\texttt{pa}_{\mathcal{G}}(V'_m) \setminus V'_m \subseteq \bigcap_{v \in V'_m} \texttt{pa}_{\mathcal{G}}(v)$.

Proof. Condition 1 holds because removing some edges $m' \to s'$ in the canonical graph will yield a graph that is still invariant under `SplitM->S`, `MergeS`, `Exog`, `Term`. Then, step B) maintains these properties while also ensuring invariance under `RmvRedM`.

Proof that $\tilde{\mathcal{D}}$ satisfies Condition 2a'): Straightforward from the definition of $\tilde{\mathcal{D}}$.

Proof that $\tilde{\mathcal{D}}$ satisfies Condition 2b'): If any two visible vertices share a child or are in the same selected face in $\mathcal{G}_{\texttt{an}(V_s)}$ then by b), they will be neighbours in $\mathcal{G}$ or they will share a marginalized parent (which will be an element of M') in both $\texttt{can}(\mathcal{G})$ and $\tilde{\mathcal{D}}$, thereby satisfying 2b').

Proof that $\tilde{\mathcal{D}}$ satisfies Condition 2c'): $\texttt{can}(\mathcal{G})$ clearly satisfies Condition 2c'), and $\tilde{\mathcal{D}}$ only adds a single marginalized parent to visible variables that lack any marginalized parent.

Proof that $\tilde{\mathcal{D}}$ satisfies Condition 2d'): By Proposition 23's Condition d), all of the parents of V'_m that are not in V'_m are parents of all of the vertices of V'_m in the smDG $\mathcal{G}$. The arrows between these parents and the vertices of V'_m could represent either normal edges or special edges in $\texttt{can}(\mathcal{G})$. When constructing $\tilde{D}$, however, all of these will turn into normal edges: because of Condition c) of Proposition 23, we know that the parents of V'_m that are not in V'_m do not share a marginalized parent with any vertex of V'_m in $\texttt{can}(\mathcal{G})$. Therefore, condition A) of lemma 30 will transform all of those edges that are special into normal edges of $\tilde{D}$. With this, for every $m \in M$, the parents $\texttt{pa}_{\mathcal{G}}(\texttt{ch}_{\texttt{can}(\mathcal{G})}(m)) \setminus \texttt{ch}_{\texttt{can}(\mathcal{G})}(m)$ are shared parents of every $v \in \texttt{ch}_{\texttt{can}(\mathcal{G})}(m)$ by Proposition 23's Condition d). In $\tilde{\mathcal{D}}$ visible variables have the same marginalized parents as in $\texttt{can}(\mathcal{G})$, except that some visible variables that lack a marginalized parent in $\texttt{can}(\mathcal{G})$ acquire one, so we also have that the parents $\texttt{pa}_{\mathcal{G}}(\texttt{ch}_{\mathcal{D}_{-s}}(m))$ are shared parents of every $v \in \texttt{ch}_{\mathcal{D}_{-s}}(m)$.

□

We can now prove that there exists an ordering over the marginalized variables M' with certain properties, that will ultimately allow us to factorize the model according to the districts of the graph [31].

Lemma 32. *Let $\mathcal{G}$ be an smDG that satisfies the conditions of Proposition 23 and define $\tilde{\mathcal{D}}$ as in Lemma 29. Define M' and V'_m as per Lemma 31. Then, $\tilde{\mathcal{D}}$ is consistent with a topological ordering of the vertices of $\tilde{\mathcal{D}} \cap \mathtt{an}_{\tilde{\mathcal{D}}}(s) \setminus \{s\}$ that takes the form:*

$$m_i \prec v^{i,1} \prec \dots v^{i,l} \prec m_j \dots$$

where:

- *$m_i, m_j, \dots \in M'$ are the marginalized vertices with only visible children.*
- *$v^{i,1} \dots, v^{i,l}$ are the visible children of m_i;*

Proof. We will assume that there is no such ordering, and prove a contradiction.

If there is no topological ordering, then note that the marginalized variables have no parents, so there must be a cycle in the visible variables. Specifically, there must exist a cyclic sequence of distinct marginalized variables $m_1, \dots, m_N = m_1$ in M' such that for each i there are vertices $v_i \in V'_{m_i}$ and $v'_{i-1} \in V'_{m_{i-1}}$ with $v'_{i-1} \in \mathtt{an}_{\mathcal{G}}(v_i)$. By Condition 2d'), if a vertex that is not in V'_{m_i} is a parent of some $v_i \in V'_{m_i}$, then it is a parent of every vertex in V'_{m_i}. This would imply that $\tilde{\mathcal{D}}$ contains a cycle $v'_{i-1} \dashrightarrow v'_i \dashrightarrow v'_{i+1} \dashrightarrow \dots \dashrightarrow v'_{i-1}$, which contradicts the assumption that $\tilde{\mathcal{D}}$ is a S-pDAG. □

Lemma 33. *Let $\mathcal{G}$ be an smDG that satisfies the conditions of Proposition 23 and define $\tilde{\mathcal{D}}$ as in Lemma 29. Considering the topological ordering defined in Lemma 32, start from $\tilde{\mathcal{D}}$ and:*

(ii) (Fully connect each district) For every $v^{i,l} \prec v^{i,l'}$, add $v^{i,l} \to v^{i,l'}$.

(ii) (Parents outside a district) For every $v^{i,l} \prec m^{i'}$, $i \neq i'$, where $\mathtt{ch}(v^{i,l}) \cap \mathtt{ch}(m^{i'}) \neq \varnothing$, add $v^{i,l} \to m^{i'}$.

Then, the resulting graph $\tilde{\mathcal{D}}^{shd}$ is a S-pDAG which is observationally equivalent to $\tilde{\mathcal{D}}$. Furthermore, $\mathtt{an}_{\tilde{\mathcal{D}}^{shd}}(s)$ has no unshielded colliders.

Proof. Since the transformations (i)–(ii) only add an arrow $a \to b$ in cases where $a \prec b$ in the topological ordering of Lemma 32, the resulting graph is a S-pDAG.

To show that operation (i) preserves observational equivalence, we will use district factorization [31]. We consider a factorization of the distribution by districts (the set of

children of a marginalized variable m^i), and wherever a visible or selected variable z lacks marginalized parents, we assign it to its own district $\{z\}$. Even though [31] only considers mDAGs, the idea can be extended to the case when there are selected faces by simply treating selected vertices as visible. Consider a district $V'_{m^i} = \mathtt{ch}_{\tilde{\mathcal{D}}}(m^i)$ belonging to a marginalized variable m^i. By Lemma 31, all of the visible parents of a vertex in this district are shared between all of the vertices of the district. Therefore, the district subgraph composed by V'_{m^i} and their parents (where the vertices of $\mathtt{pa}_{\tilde{\mathcal{D}}}(V'_{m^i}) \setminus V'_{m^i}$ are fixed vertices, following Definition 2.9 of [31]) is *saturated*, in the sense that it can realize *any* kernel $P\big(X_{V'_{m^i}} \big| X_{\mathtt{pa}_{\tilde{\mathcal{D}}}(V'_{m^i}) \setminus V'_{m^i}}\big)$. Because of that, adding edges inside this district does not increase the set of observational distributions that are realizable by the causal structure, thus showing that operation (i) preserves observational equivalence.

Operation (ii) also preserves observational equivalence, as a consequence of Proposition 15(a).

To see that $\mathtt{an}_{\tilde{\mathcal{D}}^{\text{shd}}}(s)$ has no unshielded colliders in $\tilde{\mathcal{D}}^{\text{shd}}$, note that:

- Due to Lemma 31(b'), two visible vertices of $\mathtt{an}_{\tilde{\mathcal{D}}^{\text{shd}}}(s)$ that share a child are either already connected by an edge in $\tilde{\mathcal{D}}$, or are children of the same m^i, in which case they become connected by an edge in $\tilde{\mathcal{D}}^{\text{shd}}$ due to operation (i). Therefore, all of the colliders of $\mathtt{an}_{\tilde{\mathcal{D}}^{\text{shd}}}(s)$ that take the form $v \to \cdot \leftarrow v'$ for visible v, v' are shielded in $\tilde{\mathcal{D}}^{\text{shd}}$.
- Since there are no special edges in $\tilde{\mathcal{D}}$, there are no colliders of the form $v \to s \leftarrow m$ for visible v, selected s and marginalized m.
- Since there are no special edges in $\tilde{\mathcal{D}}$ and because of Lemma 31c'), there are no colliders of the form $m \to v \leftarrow m'$ for marginalized m and m'.
- Since there are no special edges, colliders of the form $v \to v' \leftarrow m$ for v, v' visible and m marginalized can only arise when m only has visible children. Such colliders are either shielded by an edge $m \to v$, in case v and v' belong to the same district, or by an edge $v \to m$ that is added by operation (ii), in case v and v' belong to different districts.

□

Now, we can prove Lemma 30:

Proof of Lemma 30. Define $\tilde{\mathcal{D}}$ as in Lemma 29, and $\tilde{\mathcal{D}}^{\text{shd}}$ as in Lemma 33. By Lemma 28, we know that the S-pDAG $\tilde{\mathcal{D}}^{\text{shd}}_{-s}$ is observationally equivalent to $\tilde{\mathcal{D}}^{\text{shd}}$, which is in turn observationally equivalent to $\tilde{\mathcal{D}}$.

We can then perform the inverse of operations (i-ii) (Lemma 33) on $\tilde{\mathcal{D}}^{\text{shd}}_{-s}$ to obtain the observationally equivalent S-pDAG $\tilde{\mathcal{D}}_{-s}$, which is the subgraph of $\tilde{\mathcal{D}}$ on variables other than s. Thus, $\mathcal{M}(\tilde{\mathcal{D}}, V \mid S) = \mathcal{M}(\tilde{\mathcal{D}}_{-s}, V \mid S \setminus \{s\})$. □

With Lemma 30, we can finally prove Proposition 23, which we restate below.

Proposition 23. *Let $\mathcal{G} = (V, \mathcal{E}, \mathcal{L}, \mathcal{S})$ be a liftable smDG, let $V_s \in \mathcal{S}$ be one of its maximal selected faces, and let $\mathcal{G}_{\mathtt{an}(V_s)}$ be the subgraph of $\mathcal{G}$ over the set V_s and all of its ancestors. Suppose that:*

a) There are no cycles in $\mathcal{G}_{\mathtt{an}(V_s)}$.

b) In $\mathcal{G}_{\mathtt{an}(V_s)}$, nodes that share a child or are in the same selected face are neighbours or are in the same marginalized face.

c) All of the maximal marginalized faces of $\mathcal{L}$ that include nodes of $\mathtt{an}_{\mathcal{G}}(V_s)$ are disjoint.

d) For each maximal marginalized face $V_m \in \mathcal{L}$ that includes nodes of $\mathtt{an}_{\mathcal{G}}(V_s)$, the parents of V_m that are not in V_m are shared by all the nodes of V_m, i.e. $\mathtt{pa}_{\mathcal{G}}(V_m) \setminus V_m \subseteq \bigcap_{v \in V_m} \mathtt{pa}_{\mathcal{G}}(v)$.

Construct the independence system $\mathcal{S}'$ by starting from $\mathcal{S}$ and then deleting V_s and all of its subsets. Define the smDG $\mathcal{G}' = (V, \mathcal{E}, \mathcal{L}, \mathcal{S}')$. Then, $\mathcal{G}$ and $\mathcal{G}'$ are observationally equivalent.

Proof. Clearly, $\mathcal{G}$ observationally dominates $\mathcal{G}'$. We proceed to prove the converse.

Due to Lemma 29, we know that $\mathtt{can}(\mathcal{G})$ is observationally equivalent to $\tilde{\mathcal{D}}$. Due to Lemma 30, we know that $\tilde{\mathcal{D}}$ is observationally equivalent to $\tilde{\mathcal{D}}_{-s}$. In $\tilde{\mathcal{D}}_{-s}$, we identify every edge $a \to b$ that was created in step A) of Lemma 30. All of these were originally special edges $a \to s' \leftarrow m' \to b$ in $\mathtt{can}(\mathcal{G})$. Then, we construct a new graph $\mathcal{D}'$ by starting from $\tilde{\mathcal{D}}_{-s}$ and reverting this step, that is, by transforming these normal edges back into special edges. Due to Proposition 18, we know that $\tilde{\mathcal{D}}_{-s}$ and $\mathcal{D}'$ are Observe&Do equivalent. Now, we identify every special edge that was broken in step B) of of Lemma 30. Construct the S-pDAG $\mathcal{D}''$ by starting from $\mathcal{D}'$ and reverting this step, recovering those special edges. Due to Proposition 22, $\mathcal{D}''$ and $\mathcal{D}'$ are observationally equivalent. Finally, note that $\mathcal{D}'' = \mathtt{can}(\mathcal{G}')$. This completes the proof that $\mathcal{G}$ and $\mathcal{G}'$ are observationally equivalent. □

Appendix D

Appendices for Chapter 5

D.1 Proofs of Piggybacks

Proposition 25 (Point Distribution Piggyback)**.** *Let $\mathcal{G}_{new}$ be a pDAG, and let $\mathcal{G}_{pig}$ be the pDAG obtained by deleting a visible node $v \in \mathtt{vis}(\mathcal{G}_{new})$ from $\mathcal{G}_{new}$. If $\mathcal{G}_{pig}$ has a QC gap, then $\mathcal{G}_{new}$ also has a QC gap.*

Proof. Let $P(X_{\mathtt{vis}(\mathcal{G}_{\text{pig}})})$ be a probability distribution that is realizable by $\mathcal{G}_{\text{pig}}$ when its latent nodes are quantum but not when its latent nodes are classical. Define the probability distribution $P'(X_{\mathtt{vis}(\mathcal{G}_{\text{new}})})$ as

$$\begin{aligned} &P'(X_{\mathtt{vis}(\mathcal{G}_{\text{new}})} = x_{\mathtt{vis}(\mathcal{G}_{\text{new}})}) = P(X_{\mathtt{vis}(\mathcal{G}_{\text{pig}})} = x_{\mathtt{vis}(\mathcal{G}_{\text{new}})\setminus\{v\}}) && \text{when } x_v = \bar{x}_v, \\ &P'(X_{\mathtt{vis}(\mathcal{G}_{\text{new}})} = x_{\mathtt{vis}(\mathcal{G}_{\text{new}})}) = 0 && \text{when } x_v \neq \bar{x}_v. \end{aligned} \tag{D.1}$$

That is, P' sets X_v to a point distribution in the value $\bar{x}_v$.

It is straightforward to see that $P'(X_{\mathtt{vis}(\mathcal{G}_{\text{new}})})$ is quantumly realizable by $\mathcal{G}_{\text{new}}$: for the nodes $\mathtt{vis}(\mathcal{G}_{\text{pig}})$ we can use the same parameters that were used to quantumly realize P in $\mathcal{G}_{\text{pig}}$, and for the node v we can set X_v to a point distribution (independent of all its parents).

Now, assume by contradiction that P' is classically realizable by $\mathcal{G}_{\text{new}}$. If that was the case, the parameters used to realize it are such that all of the children of v effectively do not depend on X_v, since X_v only takes one possible value. Therefore, these same parameters could be used in $\mathcal{G}_{\text{pig}}$ to classically realize P', which would be a contradiction.

Therefore, $\mathcal{G}_{\text{new}}$ has a QC gap, and $P'(X_{\mathtt{vis}(\mathcal{G}_{\text{new}})})$ is a distribution in this gap.

□

Proposition 26 (Marginalization Piggyback)**.** *Let $\mathcal{G}_{new}$ be a pDAG, and let $\mathcal{G}_{pig}$ be the pDAG obtained by transforming a visible node $d \in \mathtt{vis}(\mathcal{G}_{new})$ from $\mathcal{G}_{new}$ into a latent node. That is, $d \in \mathtt{lat}(\mathcal{G}_{pig})$. If there is a gap between $\mathcal{G}_{pig}$ when all latent nodes are classical and $\mathcal{G}_{pig}$ when d is classical but all of the other latent nodes are quantum, then $\mathcal{G}_{new}$ has a QC gap.*

Proof. Let $P(X_{\mathtt{vis}(\mathcal{G}_{\text{pig}})})$ be a probability distribution that is realizable by $\mathcal{G}_{\text{pig}}$ when d is classical and $\mathtt{lat}(\mathcal{G}_{\text{pig}}) \setminus \{d\}$ are quantum, but that is *not* realizable by $\mathcal{G}_{\text{pig}}$ when all of its latent nodes are classical.

Since $P(X_{\mathtt{vis}(\mathcal{G}_{\text{pig}})})$ is realizable by $\mathcal{G}_{\text{pig}}$ when d is classical but the other latent nodes are quantum, it must be the marginal over $X_{\mathtt{vis}(\mathcal{G}_{\text{pig}})}$ of a joint distribution $P'(X_{\mathtt{vis}(\mathcal{G}_{\text{pig}})}, X_d)$ that is quantumly realizable by the pDAG obtained by transforming d into a visible node, i.e., by $\mathcal{G}_{\text{new}}$.

The distribution P' is such that

$$\sum_{X_d} P'(X_{\mathtt{vis}(\mathcal{G}_{\text{new}})}) = \sum_{X_d} P'(X_{\mathtt{vis}(\mathcal{G}_{\text{pig}})}, X_d) = P(X_{\mathtt{vis}(\mathcal{G}_{\text{pig}})}). \tag{D.2}$$

The marginalization of Eq. (D.2) is the one that appears in the definition of realizability by $\mathcal{G}_{\text{pig}}$ when d is a classical latent node. Therefore, if $P'(X_{\mathtt{vis}(\mathcal{G}_{\text{new}})})$ were classically realizable by $\mathcal{G}_{\text{new}}$, then $P(X_{\mathtt{vis}(\mathcal{G}_{\text{pig}})})$ would be classically realizable by $\mathcal{G}_{\text{pig}}$. We thus conclude that $P'(X_{\mathtt{vis}(\mathcal{G}_{\text{new}})})$ *is not* classically realizable by $\mathcal{G}_{\text{new}}$. Therefore, $\mathcal{G}_{\text{new}}$ has a QC gap, and $P'(X_{\mathtt{vis}(\mathcal{G}_{\text{new}})})$ is a distribution in this gap. □

Proposition 27 (Fritz Piggyback)**.** *Let $\mathcal{G}_{new}$ be a pDAG, and let $a, b \in \mathtt{vis}(\mathcal{G}_{new})$ be two visible nodes of this pDAG such that either a is a parent of b or a and b share a common latent parent (or both). Let $p \in \mathtt{pa}_{\mathcal{G}_{new}}(b)$ be a parent of b (not necessarily visible) such that $\mathcal{G}_{new}$ exhibits the d-separation relation $a \perp p | \mathtt{pa}_{\mathcal{G}_{new}}(b) \setminus \{p\}$. Let $\mathcal{G}_{pig}$ be the pDAG obtained by adding two visible nodes a' and b' to $\mathcal{G}_{new}$ such that $\mathtt{ch}_{\mathcal{G}_{pig}}(a') = \mathtt{ch}_{\mathcal{G}_{new}}(a)$, $\mathtt{pa}_{\mathcal{G}_{pig}}(a') = \mathtt{pa}_{\mathcal{G}_{new}}(a)$, $\mathtt{ch}_{\mathcal{G}_{pig}}(b') = \mathtt{ch}_{\mathcal{G}_{new}}(b)$, $\mathtt{pa}_{\mathcal{G}_{pig}}(b') = \mathtt{pa}_{\mathcal{G}_{new}}(b) \setminus \{p\}$, adding a common latent parent between a and a' if it is not already present and adding a common latent parent between b and b' if it is not already present. If $\mathcal{G}_{pig}$ has a QC gap that includes distributions where $X_{a'}$ and $X_{b'}$ are perfectly correlated, then $\mathcal{G}_{new}$ also has a QC gap.*

Proof. Construct a pDAG $\mathcal{G}'_{\text{new}}$ by starting from $\mathcal{G}_{\text{new}}$ and adding visible nodes a' and b' such that $\texttt{ch}_{\mathcal{G}_{\text{pig}}}(a') = \texttt{ch}_{\mathcal{G}_{\text{new}}}(a)$, $\texttt{pa}_{\mathcal{G}_{\text{pig}}}(a') = \texttt{pa}_{\mathcal{G}_{\text{new}}}(a)$, $\texttt{ch}_{\mathcal{G}_{\text{pig}}}(b') = \texttt{ch}_{\mathcal{G}_{\text{new}}}(b)$ and $\texttt{pa}_{\mathcal{G}_{\text{pig}}}(b') = \texttt{pa}_{\mathcal{G}_{\text{new}}}(b)$, and there is at least one latent common cause between a and a' and between b and b'. Since $\mathcal{G}_{\text{new}}$ has that $a \perp p | \texttt{pa}_{\mathcal{G}_{\text{new}}}(b) \setminus \{p\}$, $\mathcal{G}'_{\text{new}}$ has that $a' \perp p | \texttt{pa}_{\mathcal{G}'_{\text{new}}}(b') \setminus \{p\}$. The only difference between $\mathcal{G}'_{\text{new}}$ and $\mathcal{G}_{\text{pig}}$ is the arrow $p \to b'$.

It is easy to see that, if $\mathcal{G}'_{\text{new}}$ has a QC gap, then $\mathcal{G}_{\text{new}}$ also has a QC gap. This is because, if we start from a distribution over $X_{\texttt{vis}(\mathcal{G}'_{\text{new}})}$ that is quantumly but not classically realizable by $\mathcal{G}'_{\text{new}}$, we can obtain a distribution over $X_{\texttt{vis}(\mathcal{G}_{\text{new}})}$ that is quantumly but not classically realizable by $\mathcal{G}_{\text{new}}$ by associating the node a to the variable $X_{a'} \times X_a$ and associating the node b to the variable $X_{b'} \times X_b$. Now, we show that if $\mathcal{G}_{\text{pig}}$ has a QC gap, then $\mathcal{G}'_{\text{new}}$ has a QC gap.

Let $\tilde{P}(X_{\texttt{vis}(\mathcal{G}_{\text{pig}})})$ be a probability distribution that is realizable by $\mathcal{G}_{\text{pig}}$ when its latent nodes are quantum but not when its latent nodes are classical, and that is such that $X_{a'} = X_{b'}$. Since $\mathcal{G}'_{\text{new}}$ is obtained from $\mathcal{G}_{\text{pig}}$ by adding the edge $p \to b'$, it clearly can quantumly realize $\tilde{P}$. Now, we proceed to show that $\mathcal{G}'_{\text{new}}$ cannot classically realize $\tilde{P}$.

Let $P(X_{\texttt{vis}(\mathcal{G}'_{\text{new}})})$ be a distribution that is classically realizable by $\mathcal{G}'_{\text{new}}$, and that is such that $X_{a'} = X_{b'}$. This implies

$$P(X_{a'}, X_{b'}, X_{\texttt{pa}_{\mathcal{G}'_{\text{new}}}(b')\setminus\{p\}}, X_p) = P(X_{a'}, X_{\texttt{pa}_{\mathcal{G}'_{\text{new}}}(b')\setminus\{p\}}, X_p)\delta_{X_{b'}, X_{a'}} \tag{D.3}$$

$$P(X_{a'}, X_{b'}, X_{\texttt{pa}_{\mathcal{G}'_{\text{new}}}(b')\setminus\{p\}}, X_p) = P(X_{b'}, X_{\texttt{pa}_{\mathcal{G}'_{\text{new}}}(b')\setminus\{p\}}, X_p)\delta_{X_{a'}, X_{b'}} \tag{D.4}$$

$$\Downarrow$$

$$P(X_{a'}, X_{\texttt{pa}_{\mathcal{G}'_{\text{new}}}(b')\setminus\{p\}}, X_p) = P(X_{b'}, X_{\texttt{pa}_{\mathcal{G}'_{\text{new}}}(b')\setminus\{p\}}, X_p) \tag{D.5}$$

$$\Downarrow$$

$$P(X_{a'} | X_{\texttt{pa}_{\mathcal{G}'_{\text{new}}}(b')\setminus\{p\}}, X_p) = P(X_{b'} | X_{\texttt{pa}_{\mathcal{G}'_{\text{new}}}(b')\setminus\{p\}}, X_p). \tag{D.6}$$

By a similar logic, we have that $P(X_{a'} | X_{\texttt{pa}_{\mathcal{G}'_{\text{new}}}(b')\setminus\{p\}}) = P(X_{b'} | X_{\texttt{pa}_{\mathcal{G}'_{\text{new}}}(b')\setminus\{p\}})$. Since $a' \perp p | \texttt{pa}_{\mathcal{G}'_{\text{new}}}(b') \setminus \{p\}$ in $\mathcal{G}'_{\text{new}}$, we know that $P(X_{a'} | X_{\texttt{pa}_{\mathcal{G}'_{\text{new}}}(b')\setminus\{p\}}, X_p) = P(X_{a'} | X_{\texttt{pa}_{\mathcal{G}'_{\text{new}}}(b')\setminus\{p\}})$. Therefore, we learn that $P(X_{b'} | X_{\texttt{pa}_{\mathcal{G}'_{\text{new}}}(b')\setminus\{p\}}, X_p) = P(X_{b'} | X_{\texttt{pa}_{\mathcal{G}'_{\text{new}}}(b')\setminus\{p\}})$. That is, this distribution can also be classically realized by the pDAG obtained by removing the arrow $p \to b'$ from $\mathcal{G}'_{\text{new}}$, which is $\mathcal{G}_{\text{pig}}$. By contradiction, then, $\tilde{P}$ is not classically realizable by $\mathcal{G}'_{\text{new}}$.

□

Proposition 28 (Interruption Piggyback)**.** *Let $\mathcal{G}_{new}$ be a pDAG, let $c \in \texttt{vis}(\mathcal{G}_{new})$ be a visible node of $\mathcal{G}_{new}$ that has no parents and $p \in \texttt{vis}(\mathcal{G}_{new})$ be a visible node of $\mathcal{G}_{new}$ that has no children, that are such that p is not an ancestor of c in $\mathcal{G}_{new}$. Let $\mathcal{G}_{pig}$ be the pDAG obtained by deleting c and p from $\mathcal{G}_{new}$, and adding a visible node $b \in \texttt{vis}(\mathcal{G}_{pig})$ such that $\texttt{pa}_{\mathcal{G}_{pig}}(b) = \texttt{pa}_{\mathcal{G}_{new}}(c)$ and $\texttt{ch}_{\mathcal{G}_{pig}}(b) = \texttt{ch}_{\mathcal{G}_{new}}(p)$. If $\mathcal{G}_{pig}$ has a QC gap, then $\mathcal{G}_{new}$ also has a QC gap.*

Proof. This idea is based on the interruption lemma of Ref. [81], which is in turned based on the single-world intervention graphs introduced in Ref. [64].

Let $P\big(X_{\texttt{vis}(\mathcal{G}_{\text{pig}})}\big)$ be a probability distribution that is realizable by $\mathcal{G}_{\text{pig}}$ when its latent nodes are quantum but not when its latent nodes are classical. Let $\big\{\mathcal{P}_{y|\texttt{pa}_{\mathcal{G}_{\text{pig}}}(y)}\big\}_{y\in\texttt{nodes}(\mathcal{G}_{\text{pig}})}$ be the set of parameters that are used to quantumly realize P in $\mathcal{G}_{\text{pig}}$. Construct a quantum model in $\mathcal{G}_{\text{new}}$ as follows: for nodes $y \in \texttt{nodes}(\mathcal{G}_{\text{new}}) \setminus \{c, d\}$, use the same parameters $\mathcal{P}_{y|\texttt{pa}_{\mathcal{G}_{\text{pig}}}(y)}$ that were used to quantumly realize P in $\mathcal{G}_{\text{pig}}$, substituting b for p whenever $b \in \texttt{pa}_{\mathcal{G}_{\text{pig}}}(y)$ (this is possible because $\texttt{ch}_{\mathcal{G}_{\text{pig}}}(b) = \texttt{ch}_{\mathcal{G}_{\text{new}}}(p)$). For the node c, use the parameter $\mathcal{P}_{b|\texttt{pa}_{\mathcal{G}_{\text{pig}}}(b)}$ that was associated with b in the quantum realization of P in $\mathcal{G}_{\text{pig}}$ (this is possible because $\texttt{pa}_{\mathcal{G}_{\text{pig}}}(b) = \texttt{pa}_{\mathcal{G}_{\text{new}}}(c)$). For the node p, associate the variable X_p to an uniform distribution over its possible outcomes. This model generates a distribution

$$P'\big(X_{\texttt{vis}(\mathcal{G}_{\text{new}})\setminus\{c,p\}}, X_c, X_p\big) = \prod_{y\in\texttt{nodes}(\mathcal{G}_{\text{new}})\setminus\{c,p\}} \mathcal{P}_{y|\texttt{pa}_{\mathcal{G}_{\text{pig}}}(y)\setminus\{b\},X_p}\mathcal{P}_{X_c|\texttt{pa}_{\mathcal{G}_{\text{pig}}}(b)}\frac{1}{|X_p|}, \tag{D.7}$$

where $\mathcal{P}_{y|\texttt{pa}_{\mathcal{G}_{\text{pig}}}(y)\setminus\{b\},X_p}$ indicates that we substitute X_b for X_p every time that $b \in \texttt{pa}_{\mathcal{G}_{\text{pig}}}(y)$, and $P(X_p) = \frac{1}{|X_p|}$ because X_p is uniformly distributed. If we look at the specific case where $X_c = X_p = \bar{x}$, we get:

$$P'\big(X_{\texttt{vis}(\mathcal{G}_{\text{new}})\setminus\{c,p\}}, X_c = \bar{x}, X_p = \bar{x}\big) = \prod_{y\in\texttt{nodes}(\mathcal{G}_{\text{new}})\setminus\{c,p\}} \mathcal{P}_{y|\texttt{pa}_{\mathcal{G}_{\text{pig}}}(y)\setminus\{b\},X_p=\bar{x}}\mathcal{P}_{X_c=\bar{x}|\texttt{pa}_{\mathcal{G}_{\text{pig}}}(b)}\frac{1}{|X_p|} \tag{D.8}$$

$$= P\big(X_{\texttt{vis}(\mathcal{G}_{\text{pig}})\setminus\{b\}}, X_b = \bar{x}\big)\frac{1}{|X_p|}. \tag{D.9}$$

In other words, we have that

$$P'\big(X_{\texttt{vis}(\mathcal{G}_{\text{new}})\setminus\{c,p\}}, X_c = \bar{x}\big|X_p = \bar{x}\big) = P\big(X_{\texttt{vis}(\mathcal{G}_{\text{pig}})\setminus\{b\}}, X_b = \bar{x}\big). \tag{D.10}$$

If any distribution $P'(X_{\mathtt{vis}(\mathcal{G}_{\text{new}})})$ that obeys Eq. (D.10) were classically realizable by $\mathcal{G}_{\text{new}}$, then $P(X_{\mathtt{vis}(\mathcal{G}_{\text{pig}})})$ would be classically realizable by $\mathcal{G}_{\text{pig}}$. This is because we could use in $\mathcal{G}_{\text{pig}}$ exactly the same parameters that were used to classically realize P' in $\mathcal{G}_{\text{new}}$, with the only difference that the values of the variables X_c and X_p are always the same (corresponding to the variable X_b).

Therefore, $P'(X_{\mathtt{vis}(\mathcal{G}_{\text{new}})})$ is not classically realizable by $\mathcal{G}_{\text{new}}$. However, by construction it is quantumly realizable by $\mathcal{G}_{\text{new}}$, so $\mathcal{G}_{\text{new}}$ has a QC gap. □

Proposition 29 (Postselection Piggyback)**.** *Let $\mathcal{G}_{new}$ be a pDAG, and let $\mathcal{G}_{pig}$ be the S-pDAG obtained by transforming a visible node v from $\mathcal{G}_{new}$ into a selected node. If $\mathcal{G}_{pig}$ has a QC gap, then $\mathcal{G}_{new}$ also has a QC gap.*

Proof. Let $P(X_{\mathtt{vis}(\mathcal{G}_{\text{pig}})}|X_v = 0)$ be a probability distribution that is realizable by $\mathcal{G}_{\text{pig}}$ when its latent nodes are quantum but not when its latent nodes are classical. By the definition of realizability of a distribution by an S-pDAG, this conditional distribution comes from a joint distribution $P(X_{\mathtt{vis}(\mathcal{G}_{\text{pig}})}, X_v = 0)$ that is realizable by $\mathcal{G}_{\text{new}}$ when its latent nodes are quantum but not when its latent nodes are classical. So, $\mathcal{G}_{\text{new}}$ has a QC gap. □

D.2 Proofs of observational dominance rules in the quantum case

Lemma 17 (Observational equivalence via teleportation)**.** *Let $\mathcal{G}$ be a pDAG, and let $\lambda \in \mathtt{lat}(\mathcal{G})$ be one of its latent nodes. Associate to $\mathcal{G}$ a semantics $\mathscr{T}$ where λ is classical. Let $\mathtt{Qpa}_{\mathcal{G}}(\lambda)$ denote the set of quantum parents of λ in $\mathcal{G}$ under $\mathscr{T}$, and $\mathtt{Cpa}_{\mathcal{G}}(\lambda)$ denote the set of classical parents of λ in $\mathcal{G}$ under $\mathscr{T}$. Let $T = \mathtt{ch}_{\mathcal{G}}(\lambda) \cap \mathtt{ch}_{\mathcal{G}}(\mathtt{Qpa}_{\mathcal{G}}(\lambda))$ denote the set of children of λ that share a quantum parent with λ.*

Let $\mathcal{G}'$ be the pDAG obtained by

1. *Adding one arrow from every node in $\mathtt{Qpa}_{\mathcal{G}}(\lambda)$ to every node in T, and*
2. *Adding one arrow from every node in $\mathtt{Cpa}_{\mathcal{G}}(\lambda)$ to every node in $\mathtt{ch}_{\mathcal{G}}(\lambda)$.*

Then, under $\mathscr{T}$, $\mathcal{G}$ is observationally equivalent to $\mathcal{G}'$.

Proof. The fact that the arrows described in 2 can be added is a simple consequence of the classical exogenization rule. Now, we prove that the arrows described in 1 can be added.

Let $c \in T$ be a child of λ that shares a quantum parent $\alpha_q \in \mathtt{Qpa}_{\mathcal{G}}(\lambda) \cap \mathtt{pa}_{\mathcal{G}}(c)$ with it. Furthermore, let $\beta_q \in \mathtt{Qpa}_{\mathcal{G}}(\lambda) \setminus \mathtt{pa}_{\mathcal{G}}(c)$ be another quantum parent of λ, which is *not* a parent of c in $\mathcal{G}$.

Say that a certain observational distribution is realizable in $\mathcal{G}'$ by associating the node α_q to a Hilbert space $\mathcal{H}_\alpha$ and to a specific state $\rho_\alpha \in \mathcal{B}(\mathcal{H}_\alpha)$, and associating the node β_q to a Hilbert space $\mathcal{H}_\beta$ and to a specific state $\rho_\beta \in \mathcal{B}(\mathcal{H}_\beta)$. In $\mathcal{G}'$, the node β_q is a parent of c, so we have that $\mathcal{H}_\beta = \mathcal{H}_\beta^c \otimes \mathcal{H}_\beta^{\text{rest}}$, where the system $\mathcal{H}_\beta^c$ is sent to c.

In $\mathcal{G}$, we associate β_q to the same system and state $\rho_\beta \in \mathcal{B}(\mathcal{H}_\beta)$, but now send the subsystem $\mathcal{H}_\beta^c$ to λ. Furthermore, we associate α_q to the Hilbert space $\mathcal{H}_\alpha \otimes \mathcal{H}_\alpha^c \otimes \mathcal{H}_\alpha^\lambda$, where $\mathcal{H}_\alpha$ is in the same state ρ_α as in $\mathcal{G}'$, and $\mathcal{H}_\alpha^c \otimes \mathcal{H}_\alpha^\lambda$ is a bipartite system such that $\dim(\mathcal{H}_\alpha^c) = \dim(\mathcal{H}_\alpha^\lambda) = \dim(\mathcal{H}_\beta^c)$ that is in a maximally entangled state such as in Equation (5.5). The system $\mathcal{H}_\alpha$ is subject to the same processes as in $\mathcal{G}'$, the system $\mathcal{H}_\alpha^c$ is sent to c and the system $\mathcal{H}_\alpha^\lambda$ is sent to λ.

Now, λ and c can use the maximally entangled state of $\mathcal{H}_\alpha^c \otimes \mathcal{H}_\alpha^\lambda$ to perform the teleportation protocol: λ performs a generalized Bell basis measurement [10] on the system $\mathcal{H}_\beta^c \otimes \mathcal{H}_\alpha^\lambda$ that it received, and sends the classical outcome to c. Then, c uses it to correct the state of its system $\mathcal{H}_\alpha^\lambda$, and ends up with exactly the state that the system $\mathcal{H}_\beta^c$ was in initially. In this way, it is possible to reproduce the situation of $\mathcal{G}'$, where $\mathcal{H}_\beta^c$ was being directly sent to c. □

Lemma 18 (Observational dominance via entanglement swapping)**.** *Let $\mathcal{G}$ be an S-pDAG, and let $s \in \mathtt{sel}(\mathcal{G})$ be one of its selected nodes. Associate to $\mathcal{G}$ a semantics $\mathscr{T}$, and let $\mathtt{Qpa}_{\mathcal{G}}(s)$ denote the set of quantum parents of s in $\mathscr{T}$.*

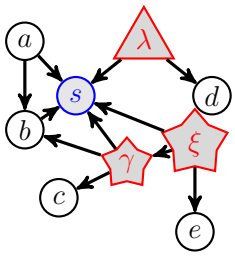
Let $\mathcal{G}'$ be the S-pDAG obtained by

1. *Adding one latent node α that has no parents and is a parent of all of the nodes in $\mathtt{Qpa}_{\mathcal{G}}(s)$,*
2. *Adding one latent node β that has no parents and is a parent of all of the nodes in $\mathtt{pa}_{\mathcal{G}}(s)$, and*
3. *Deleting s.*

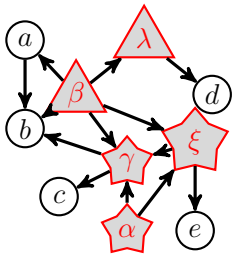
Suppose that the node s has no ancestors that are not its parents in $\mathcal{G}$. Then, $\mathcal{G}$ under $\mathscr{T}$ observationally dominates $\mathcal{G}'$ when α is quantum, β is classical and the remaining latent nodes follow $\mathscr{T}$.

Proof. We will start by proving observational dominance in one concrete example (Fig. D.2.1). Before going through this example, we recommend the reader to first go through the simpler example presented in Section 5.1.5, when we showed that $\mathcal{G}_{\text{flag}}$ under selection on d quantumly observationally dominates the Instrumental scenario.

Consider the pDAG $\mathcal{G}$ of Fig. D.2.1(a). Here, we indicated the semantics $\mathscr{T}$ in the picture: while the latent node λ is classical, the latent nodes γ and ξ are quantum. This pDAG is such that all of the ancestors of the selected node s are also its parents. We will show that the two pDAGs of Fig. D.2.1 are observationally equivalent under the semanticss indicated in the figures.



(a) A pDAG $\mathcal{G}$ with a choice of semantics: λ is classical, while ξ and γ are quantum.



(b) A pDAG $\mathcal{G}'$ with a choice of semantics: β and λ are classical, while α, ξ and γ are quantum.

Figure D.2.1: DAGs used in to exemplify the proof of Lemma 18.

Let $P(X_aX_bX_cX_dX_e)$ be a probability distribution realizable by $\mathcal{G}'$ of Fig. D.2.1(b) under the semantics indicated in the figure. This distribution can be written as:

$$P(X_aX_bX_cX_dX_e) = \sum_{X_\beta} \sum_{E_aE_dE_\lambda} P(X_\beta)\delta_{X_\lambda, f_\lambda(X_\beta,E_\lambda)}P(E_\lambda)\delta_{X_a,f_a(X_\beta,E_a)}P(E_a)\delta_{X_d,f_d(X_\lambda,E_d)}P(E_d)$$
$$\text{Tr}\left[\Pi^{X_b|X_\beta X_a}_{\mathcal{H}^{\text{in}}_b} \otimes \Pi^{X_c}_{\mathcal{H}^{\text{in}}_c} \otimes \Pi^{X_e}_{\mathcal{H}^{\text{in}}_e} \mathcal{E}^{X_\beta}_{\mathcal{H}^{\text{out}}_\xi|\mathcal{H}^{\text{in}}_\xi} \otimes \mathcal{E}^{X_\beta}_{\mathcal{H}^{\text{out}}_\gamma|\mathcal{H}^{\text{in}}_\gamma} \sigma_{\mathcal{H}^{\text{out}}_\alpha}\right]. \tag{D.11}$$

Where $\mathcal{H}^{\text{out}}_\alpha = \mathcal{H}^\gamma_\alpha \otimes \mathcal{H}^\xi_\alpha$ is the quantum system associated with α, $\mathcal{H}^{\text{in}}_\xi = \mathcal{H}^\xi_\alpha$ is the quantum system that ξ receives from α, $\mathcal{H}^{\text{out}}_\xi = \mathcal{H}^\gamma_\xi \otimes \mathcal{H}^{\text{in}}_e$ is the composite of the systems that ξ sends to γ and e, $\mathcal{H}^{\text{in}}_\gamma = \mathcal{H}^\gamma_\alpha \otimes \mathcal{H}^\gamma_\xi$ is the composite of the systems that γ receives

from α and ξ, and $\mathcal{H}_\gamma^{\text{out}} = \mathcal{H}_b^{\text{in}} \otimes \mathcal{H}_c^{\text{in}}$ is the composite of the systems that γ sends to b and c.

Now, we will make a choice of parameters in $\mathcal{G}$ that reproduces this distribution. Our choice of parameters is as follows:

1. The nodes c, d and e react to their parents in the same way they did in $\mathcal{G}'$. The error variable E_d is distributed in the same way as it was in $\mathcal{G}'$.

2. The variables X_a and X_b are uniformly distributed over their supports. The variable X_λ is uniformly distributed over the support that it had in Eq. (D.11).

3. The node ξ is associated with the quantum system $\mathcal{H}_\xi = \mathcal{H}_e^{\text{in}} \otimes \mathcal{H}_\xi^\gamma \otimes \tilde{\mathcal{H}}_\alpha^\xi \otimes \mathcal{H}_{\xi,\beta}$, where $\dim(\tilde{\mathcal{H}}_\alpha^\xi) = \dim(\mathcal{H}_\alpha^\xi)$ and $\dim(\mathcal{H}_{\xi,\beta}) = |X_\beta|$ is the cardinality of the variable X_β in Eq. (D.11). This system is prepared in the state

$$\frac{1}{|X_\beta|} \sum_{x_\beta} \mathcal{E}^{X_\beta = x_\beta}_{\mathcal{H}_e^{\text{in}} \otimes \mathcal{H}_\xi^\gamma | \mathcal{H}_\alpha^\xi} |\Phi^+\rangle_{\mathcal{H}_\alpha^\xi \otimes \tilde{\mathcal{H}}_\alpha^\xi} \langle \Phi^+| \otimes |x_\beta\rangle \langle x_\beta|_{\mathcal{H}_{\xi,\beta}}, \tag{D.12}$$

where

$$|\Phi^+\rangle_{\mathcal{H}_\alpha^\xi \otimes \tilde{\mathcal{H}}_\alpha^\xi} = \frac{1}{\sqrt{\dim(\mathcal{H}_\xi^\alpha)}} \sum_{i=1}^{\dim(\mathcal{H}_\xi^\alpha)-1} |i\rangle_{\mathcal{H}_\xi^\alpha} \otimes |i\rangle_{\tilde{\mathcal{H}}_\xi^\alpha}, \tag{D.13}$$

and $\mathcal{E}^{X_\beta}_{\mathcal{H}_e^{\text{in}} \otimes \mathcal{H}_\xi^\gamma | \mathcal{H}_\alpha^\xi}$ is the same as the channel $\mathcal{E}^{X_\beta}_{\mathcal{H}_\xi^{\text{out}} | \mathcal{H}_\xi^{\text{in}}}$ from Eq. (D.11).

ξ then sends the system $\mathcal{H}_e^{\text{in}}$ to e, the system $\mathcal{H}_\xi^\gamma$ to γ and the system $\tilde{\mathcal{H}}_\alpha^\xi \otimes \mathcal{H}_{\xi,\beta}$ to s.

4. The node γ receives the system $\mathcal{H}_\xi^\gamma$ from ξ. It applies a quantum channel on $\mathcal{H}_\xi^\gamma$ that has output in $\mathcal{H}_b^{\text{in}} \otimes \mathcal{H}_c^{\text{in}} \otimes \tilde{\mathcal{H}}_\alpha^\gamma \otimes \mathcal{H}_{\gamma,\beta}$, where $\dim(\tilde{\mathcal{H}}_\alpha^\gamma) = \dim(\mathcal{H}_\alpha^\gamma)$ and $\dim(\mathcal{H}_{\gamma,\beta}) = |X_\beta|$ is the cardinality of the variable X_β in Eq. (D.11). The quantum channel is:

$$\frac{1}{|X_\beta|} \sum_{\tilde{x}_\beta} \mathcal{E}^{X_\beta = \tilde{x}_\beta}_{\mathcal{H}_b^{\text{in}} \otimes \mathcal{H}_c^{\text{in}} | \mathcal{H}_\alpha^\gamma \otimes \mathcal{H}_\xi^\gamma} |\Phi^+\rangle_{\mathcal{H}_\alpha^\gamma \otimes \tilde{\mathcal{H}}_\alpha^\gamma} \langle \Phi^+| \otimes |\tilde{x}_\beta\rangle \langle \tilde{x}_\beta|_{\mathcal{H}_{\gamma,\beta}}, \tag{D.14}$$

where $|\Phi^+\rangle_{\mathcal{H}_\alpha^\gamma}$ is defined analogously to Eq. (D.13) and $\mathcal{E}^{X_\beta}_{\mathcal{H}_b^{\text{in}} \otimes \mathcal{H}_c^{\text{in}} | \mathcal{H}_\alpha^\gamma \otimes \mathcal{H}_\xi^\gamma}$ is the same as the channel $\mathcal{E}^{X_\beta}_{\mathcal{H}_\gamma^{\text{out}} | \mathcal{H}_\gamma^{\text{in}}}$ from Eq. (D.11).

γ then sends the system $\mathcal{H}_c^{\text{in}}$ to c and the systems $\mathcal{H}_b^{\text{in}}$ and $\tilde{\mathcal{H}}_\alpha^\gamma \otimes \mathcal{H}_{\gamma,\beta}$ to s.

5. The node s receives quantum systems $\tilde{\mathcal{H}}_\alpha^\xi \otimes \mathcal{H}_{\xi,\beta} \otimes \tilde{\mathcal{H}}_\alpha^\gamma \otimes \mathcal{H}_{\gamma,\beta} \otimes \mathcal{H}_b^{\text{in}}$. It is such that $X_s = 0$ if and only if all of the conditions described below are satisfied.

   - The node s makes a measurement on $\tilde{\mathcal{H}}_\alpha^\xi \otimes \tilde{\mathcal{H}}_\alpha^\gamma$, and a necessary condition for $X_s = 0$ is that it obtains the measurement outcome associated with the POVM element $\sigma^\intercal$, where σ is the same as the state of the Hilbert space $\mathcal{H}_\alpha^{\text{in}}$ in Eq. (D.11) and the transpose is defined relative to the bases $\{|i\rangle_{\tilde{\mathcal{H}}_\alpha^\xi}\}_{\dim(\tilde{\mathcal{H}}_\alpha^\xi)}$ and $\{|i\rangle_{\tilde{\mathcal{H}}_\alpha^\gamma}\}_{\dim(\tilde{\mathcal{H}}_\alpha^\gamma)}$.

     As we will see, the selection on $X_s = 0$ steers systems $\mathcal{H}_\alpha^\xi \otimes \mathcal{H}_\alpha^\gamma$ to the state σ.

   - The node s has an internal random variable X_β distributed as $P(X_\beta)$. Say that in a certain round, this variable takes the value $X_\beta = x_\beta^*$. Then, s makes a measurement on $\mathcal{H}_{\xi,\beta} \otimes \mathcal{H}_{\gamma,\beta}$, and a necessary condition for $X_s = 0$ is that it obtains measurement outcome associated with the POVM element

   $$|x_\beta^*, x_\beta^*\rangle_{\mathcal{H}_{\xi,\beta}\otimes\mathcal{H}_{\gamma,\beta}} \langle x_\beta^*, x_\beta^*| . \tag{D.15}$$

   - The node s also has internal random variables E_a and E_λ, distributed as $P(E_a)$ and $P(E_\lambda)$. Say that in a certain round, we have $X_\beta = x_\beta^*$, $E_a = e_a^*$ and $E_\lambda = e_\lambda^*$. Another necessary condition for $X_s = 0$ is that $X_a = f_a(X_\beta = x_\beta^*, E_a = e_a^*)$ and $X_\lambda = f_\lambda(X_\beta = x_\beta^*, E_\lambda = e_\lambda^*)$.

   - The node s measures the system $\mathcal{H}_b^{\text{in}}$ in the basis

   $$\left\{\Pi_{\mathcal{H}_b^{\text{in}}}^{X_b = x_b | X_\beta = x_\beta^*, X_a = f_a(x_\beta^*, E_a)}\right\}_{x_b = 1,\ldots,|X_b|}$$

     from Eq. (D.11), and obtains a certain outcome x_b^*. Then, another necessary condition for $X_s = 0$ is that $X_b = x_b^*$. Note that this was only possible because s has access to all of the parents of b.

A probability distribution $\tilde{P}(X_a X_b X_c X_d X_e | X_s = 0)$ is realizable by $\mathcal{G}$ of Fig. D.2.1(a) under the semantics indicated in the figure if it can be written as:

$$\begin{aligned}\tilde{P}(X_a X_b X_c X_d X_e | X_s = 0) = &\frac{1}{\tilde{P}(X_s = 0)} \sum_{X_\lambda} \sum_{E_d} \tilde{P}(X_\lambda) \tilde{P}(X_a) \delta_{X_d, \tilde{f}_d(X_\lambda, E_d)} \tilde{P}(E_d) \\ &\text{Tr}\left[\tilde{\Pi}_{\tilde{\mathcal{H}}_b^{\text{in}}}^{X_b|X_a} \otimes \tilde{\Pi}_{\tilde{\mathcal{H}}_c^{\text{in}}}^{X_c} \otimes \tilde{\Pi}_{\tilde{\mathcal{H}}_e^{\text{in}}}^{X_e} \otimes \tilde{\Pi}_{\tilde{\mathcal{H}}_s^{\text{in}}}^{X_s=0|X_a X_b X_\lambda} \tilde{\mathcal{E}}_{\tilde{\mathcal{H}}_\gamma^{\text{out}}|\tilde{\mathcal{H}}_\gamma^{\text{in}}} \rho_{\tilde{\mathcal{H}}_\xi^{\text{out}}}\right],\end{aligned} \tag{D.16}$$

where

$$\begin{aligned}\tilde{P}(X_s = 0) = &\sum_{X_a X_b X_c X_d X_e} \sum_{X_\lambda} \sum_{E_d} \tilde{P}(X_\lambda)\tilde{P}(X_a)\delta_{X_d,\tilde{f}_d(X_\lambda,E_d)}\tilde{P}(E_d) \\ &\operatorname{Tr}\left[\tilde{\Pi}^{X_b|X_a}_{\tilde{\mathcal{H}}^{\text{in}}_b} \otimes \tilde{\Pi}^{X_c}_{\mathcal{H}^{\text{in}}_c} \otimes \tilde{\Pi}^{X_e}_{\mathcal{H}^{\text{in}}_e} \otimes \tilde{\Pi}^{X_s=0|X_a X_b X_\lambda}_{\mathcal{H}^{\text{in}}_s} \tilde{\mathcal{E}}_{\tilde{\mathcal{H}}^{\text{out}}_\gamma|\tilde{\mathcal{H}}^{\text{in}}_\gamma \rho \mathcal{H}^{\text{out}}_\xi}\right] \\ = &\sum_{X_a X_b} \sum_{X_\lambda} \tilde{P}(X_\lambda)\tilde{P}(X_a) \operatorname{Tr}\left[\tilde{\Pi}^{X_b|X_a}_{\tilde{\mathcal{H}}^{\text{in}}_b} \otimes \tilde{\Pi}^{X_s=0|X_a X_b X_\lambda}_{\tilde{\mathcal{H}}^{\text{in}}_s} \tilde{\mathcal{E}}_{\tilde{\mathcal{H}}^{\text{out}}_\gamma|\tilde{\mathcal{H}}^{\text{in}}_\gamma \rho \tilde{\mathcal{H}}^{\text{out}}_\xi}\right] \end{aligned} \tag{D.17}$$

Now, we apply the choice of parameters above to these equations. Equation (D.16) becomes

$$\begin{aligned}&\tilde{P}(X_a X_b X_c X_d X_e | X_s = 0) = \\ &\frac{1}{\tilde{P}(X_s = 0)} \sum_{X_\lambda} \sum_{E_d} \frac{1}{|X_\lambda|} \frac{1}{|X_a|} \delta_{X_d, f_d(X_\lambda, E_d)} P(E_d) \frac{1}{|X_b|} \operatorname{Tr}\Big[\Pi^{X_c}_{\mathcal{H}^{\text{in}}_c} \otimes \Pi^{X_e}_{\mathcal{H}^{\text{in}}_e} \otimes \tilde{\Pi}^{X_s=0|X_a X_b X_\lambda}_{\tilde{\mathcal{H}}^{\text{in}}_s} \\ &\left(\frac{1}{|X_\beta|} \sum_{\tilde{x}_\beta} \mathcal{E}^{X_\beta=\tilde{x}_\beta}_{\mathcal{H}^{\text{in}}_b \otimes \mathcal{H}^{\text{in}}_c | \mathcal{H}^\gamma_\alpha \otimes \mathcal{H}^\gamma_\xi} |\Phi^+\rangle_{\mathcal{H}^\gamma_\alpha \otimes \tilde{\mathcal{H}}^\gamma_\alpha} \langle \Phi^+| \otimes |\tilde{x}_\beta\rangle \langle \tilde{x}_\beta|_{\mathcal{H}_{\gamma,\beta}}\right) \\ &\left(\frac{1}{|X_\beta|} \sum_{x_\beta} \mathcal{E}^{X_\beta=x_\beta}_{\mathcal{H}^{\text{in}}_e \otimes \mathcal{H}^\gamma_\xi | \mathcal{H}^\xi_\alpha} |\Phi^+\rangle_{\mathcal{H}^\xi_\alpha \otimes \tilde{\mathcal{H}}^\xi_\alpha} \langle \Phi^+| \otimes |x_\beta\rangle \langle x_\beta|_{\mathcal{H}_{\xi,\beta}}\right)\Big],\end{aligned} \tag{D.18}$$

where

$$\begin{aligned}\tilde{\Pi}^{X_s=0|X_a X_b X_\lambda}_{\tilde{\mathcal{H}}^{\text{in}}_s} = &\sum_{E_a E_\lambda} P(E_a)P(E_\lambda)\sigma^{\intercal}_{\tilde{\mathcal{H}}^\gamma_\alpha \otimes \tilde{\mathcal{H}}^\xi_\alpha} \otimes \sum_{x^*_\beta} \Big(P(X_\beta = x^*_\beta) |x^*_\beta, x^*_\beta\rangle_{\mathcal{H}_{\xi,\beta} \otimes \mathcal{H}_{\gamma,\beta}} \langle x^*_\beta, x^*_\beta| \\ &\otimes \delta_{X_a, f_a(x^*_\beta, E_a)} \delta_{X_\lambda, f_\lambda(x^*_\beta, E_\lambda)} \sum_{x^*_b} \delta_{X_b, x^*_b} \Pi^{X_b = x^*_b | X_\beta = x^*_\beta, X_a = f_a(x^*_\beta, E_a)}_{\mathcal{H}^{\text{in}}_b}\Big).\end{aligned} \tag{D.19}$$

Substituting Equation (D.19) into Equation (D.18) and tracing out $\mathcal{H}_{\xi,\beta} \otimes \mathcal{H}_{\gamma,\beta}$, we get

$$\tilde{P}(X_a X_b X_c X_d X_e | X_s = 0) =$$
$$\frac{1}{\tilde{P}(X_s = 0)} \sum_{X_\lambda} \sum_{E_a E_\lambda} P(E_a) P(E_\lambda) \frac{1}{|X_\lambda|} \frac{1}{|X_a|} \delta_{X_d, f_d(X_\lambda, E_d)} P(E_d) \frac{1}{|X_b|} \operatorname{Tr}\Big[\Pi^{X_c}_{\mathcal{H}^{\text{in}}_c} \otimes \Pi^{X_e}_{\mathcal{H}^{\text{in}}_e} \otimes$$
$$\sigma^{\intercal}_{\tilde{\mathcal{H}}^\gamma_\alpha \otimes \tilde{\mathcal{H}}^\xi_\alpha} \otimes \sum_{x^*_\beta} P(X_\beta = x^*_\beta) \delta_{X_a, f_a(x^*_\beta, E_a)} \delta_{X_\lambda, f_\lambda(x^*_\beta, E_\lambda)} \Pi^{X_b | X_\beta = x^*_\beta, X_a = f_a(x^*_\beta, E_a)}_{\mathcal{H}^{\text{in}}_b}$$
$$\left(\frac{1}{|X_\beta|} \mathcal{E}^{X_\beta = x^*_\beta}_{\mathcal{H}^{\text{in}}_b \otimes \mathcal{H}^{\text{in}}_c | \mathcal{H}^\gamma_\alpha \otimes \mathcal{H}^\gamma_\xi} |\Phi^+\rangle_{\mathcal{H}^\gamma_\alpha \otimes \tilde{\mathcal{H}}^\gamma_\alpha} \langle \Phi^+| \right) \left(\frac{1}{|X_\beta|} \mathcal{E}^{X_\beta = x^*_\beta}_{\mathcal{H}^{\text{in}}_e \otimes \mathcal{H}^\gamma_\xi | \mathcal{H}^\xi_\alpha} |\Phi^+\rangle_{\mathcal{H}^\xi_\alpha \otimes \tilde{\mathcal{H}}^\xi_\alpha} \langle \Phi^+| \right) \Big], \tag{D.20}$$

We will simplify this by using the identity

$$\operatorname{Tr}_{\mathcal{H}_B}\left[(A \otimes B)(|\Phi^+\rangle_{\mathcal{H}_{AB}} \langle \Phi^+|) \right] = \frac{1}{d} A B^{\intercal}, \tag{D.21}$$

where $d = \dim(\mathcal{H}_A) = \dim(\mathcal{H}_B)$. Using this identity two times on Equation (D.20), we get

$$\tilde{P}(X_a X_b X_c X_d X_e | X_s = 0) =$$
$$\frac{1}{\tilde{P}(X_s = 0)} \sum_{X_\lambda} \sum_{E_a E_\lambda} P(E_a) P(E_\lambda) \frac{1}{|X_\lambda|} \frac{1}{|X_a|} \delta_{X_d, f_d(X_\lambda, E_d)} P(E_d) \frac{1}{|X_b|} \operatorname{Tr}\Big[\Pi^{X_c}_{\mathcal{H}^{\text{in}}_c} \otimes \Pi^{X_e}_{\mathcal{H}^{\text{in}}_e} \otimes$$
$$\sum_{x^*_\beta} P(X_\beta = x^*_\beta) \delta_{X_a, f_a(x^*_\beta, E_a)} \delta_{X_\lambda, f_\lambda(x^*_\beta, E_\lambda)} \Pi^{X_b | X_\beta = x^*_\beta, X_a = f_a(x^*_\beta, E_a)}_{\mathcal{H}^{\text{in}}_b}$$
$$\frac{1}{\dim(\mathcal{H}^\xi_\alpha) \dim(\mathcal{H}^\gamma_\alpha)} \frac{1}{|X_\beta|^2} \mathcal{E}^{X_\beta = x^*_\beta}_{\mathcal{H}^{\text{in}}_b \otimes \mathcal{H}^{\text{in}}_c | \mathcal{H}^\gamma_\alpha \otimes \mathcal{H}^\gamma_\xi} \mathcal{E}^{X_\beta = x^*_\beta}_{\mathcal{H}^{\text{in}}_e \otimes \mathcal{H}^\gamma_\xi | \mathcal{H}^\xi_\alpha} \sigma_{\mathcal{H}^\gamma_\alpha \otimes \mathcal{H}^\xi_\alpha} \Big]. \tag{D.22}$$

Doing all of the same steps as above to Equation (D.17), we get:

$$\tilde{P}(X_s = 0) = \frac{1}{|X_\lambda|} \frac{1}{|X_a|} \frac{1}{|X_b|} \frac{1}{\dim(\mathcal{H}^\xi_\alpha) \dim(\mathcal{H}^\gamma_\alpha)}. \tag{D.23}$$

Substituting Equation (D.23) into Equation (D.22), we get exactly the same form as Equation (D.11). Therefore, this choice of parameters shows us that any distribution that is realizable by $\mathcal{G}'$ of Fig. D.2.1(b) under $\mathscr{T}$ with quantum α and classical β can also be realized by $\mathcal{G}$ of Fig. D.2.1(a) under $\mathscr{T}$.

The proof of the lemma in the general case is obtained by a tedious but straightforward generalization of the steps presented here.

□