Exploring Non-Classical correlations in Causal Scenarios

Shashaank Khanna

PhD

University of York
Mathematics

September 2024

Abstract

Quantum Mechanics has been one of our most successful theory. However its foundations are shrouded with various conceptual and interpretational challenges. One of them has been Quantum Non-Locality. Bell's eponymous theorem implies that Quantum Mechanics is incompatible with Local Causality. Here we study Causality in the context of Quantum Mechanics by resorting to Classical Causal Models. The classical causal relations between a set of variables, some observed and some latent, can be studied using the formalism of Directed Acyclic Graphs (DAGs) and Bayesian Networks. In quantum foundations the challenging task of ascertaining when correlations are Non-Classical within a directed acyclic graph (or causal scenario) remains a significant area of research. In this work we study how to distinguish between correlations that are Classically explainable from those that are Non-Classical and potentially realizable by Quantum Mechanics in various causal scenarios. We provide various sufficient conditions to check which causal scenarios can support Non-Classical correlations, meanwhile providing evidence towards a conjecture. We characterize all but three causal scenarios of up to four visible nodes into those which can sustain Non-Classical correlations and those which cannot. For specific scenarios, we also focus on understanding when a set of probability distributions can be explained by quantum mechanics but not by any classical theory. It is also shown here that there exist a range of causal scenarios beyond just the Bell scenario which can support Non-Classical correlations but their causal explanation requires Fine-Tuning. The multi-partite Bell scenario is conjectured to be an example of such a scenario. On the other hand, for a number of causal scenarios, which exhibit Non-Classically explainable Quantum-Correlations, we show that we can always explain such Non-Classical Quantum-Correlations by resorting instead to another causal scenario perfectly Classically. Digressing into the subject of elimination of variables from system of linear inequalities, we also provide some conditions to accelerate the algorithm to do so, the Fourier-Motzkin elimination algorithm.

Dedication

To my parents.

CONTENTS

Abstract **2**

Contents **4**

List of Figures **7**

List of Tables **11**

Acknowledgments **14**

Author's declaration **16**

1 Introduction **19**

1.1 Preface 19

1.2 Structure of the Thesis 22

1.3 Notational Conventions 23

2 Mathematical Introduction to Causality **25**

2.1 Introduction 25

2.2 Causal Explanations of Observational Data 26

2.3 Directed Acyclic Graphs 27

2.4 DAGs as Causal Structures 28

2.5 d-separation: A graphical criterion for conditional independence 29

2.6 Latent-Permitting DAGs 33

2.7 Faithfulness and Minimality 34

2.8 Classically Causal Explanation of correlations 36

2.9 Non-Classical Causal Models 38
2.10 Causal Discovery Algorithms 41

3 Classifying INTERESTING Causal Scenarios 47
3.1 Introduction 47
3.2 What are INTERESTING Causal Scenarios 51
3.3 Simplifying the problem by using mDAG formalism 54
3.4 The HLP criterion 56

4 Methods to determine INTERESTINGNESS 59
4.1 Using nonmaximality to prove INTERESTINGNESS 59
4.2 Using setwise nonmaximality to prove INTERESTINGNESS 62
4.3 Using d-separation to prove INTERESTINGNESS 65
4.4 Using e-separation to prove INTERESTINGNESS 67
4.5 Using incompatible supports to prove INTERESTINGNESS 69
4.5.1 Rapidly testing supports (without comparing to any latent-free graph) 71
4.5.2 Incompatible supports subsumes all other methods 76
4.6 Computational Results 78
4.7 Entropy Vector method to classify INTERESTING scenarios 80
4.7.1 Shannon-type inequalities 80
4.8 On the potential INTERESTINGNESS of the remaining 3 mDAGs 83
4.9 Conclusions 84

5 Causal Scenarios which cannot support Quantum Correlations without Fine-Tuning 87
5.1 Introduction 87
5.2 Non-Classical Correlations 90
5.3 Results 93
5.4 FINE-TUNING for multi-partite Bell scenarios 101
5.5 Summary of the results 114
5.6 Discussion on Modifying the IC^* algorithm for Causal Discovery 115
5.7 Conclusion 115

6 Classical-Quantum gap in Causal Scenarios 117
6.1 Introduction 117
6.2 Results 117
6.3 Comments and Discussion 126

6.4 Conclusion . . . 128

7 Digression into Accelerating Fourier-Motzkin Elimination 131

7.1 Introduction . . . 131

7.2 Fourier-Motzkin Elimination . . . 133

7.2.1 Constraints and Complexity . . . 138

7.2.2 Imbert's acceleration theorems . . . 139

7.2.3 Incompatibility of Imbert's acceleration theorems with Linear Programming to remove redundancies . . . 140

7.3 Eliminating Variables in an Efficient Order . . . 143

7.3.1 Eliminating the variable that produces the least number of redundant inequalities in each iteration of the algorithm . . . 144

7.3.2 Eliminating the variable that produces the least number of non-redundant inequalities in each iteration of the algorithm . . . 146

7.4 Shannon-type inequalities for Causal Scenarios (description of the computational problem) . . . 146

7.5 Computational Results . . . 148

7.6 Conclusion . . . 151

8 Discussion and Conclusion 152

8.0.1 A Future Project Proposal . . . 156

A Further important results in the classification of INTERESTING DAGs 157

A.1 HLP's reduction techniques . . . 157

A.2 Pienaar's e-separation theorem is incorrect . . . 161

A.3 Revisiting the Skeleton Condition . . . 163

A.4 Rapid d-separation test . . . 165

A.5 Supports subsumes e-separation as a test of observational inequivalence . . . 167

A.6 Results for mDAGs of 5 observed nodes . . . 169

A.7 Comparison of classification of INTERESTING causal scenarios based on the total number of all the nodes vs total number of just observed nodes . . . 170

A.7.1 Computational Results for mDAGs without using HLP's reduction techniques 173

A.7.2 Computational Results for mDAGs and GDAGs using HLP's reduction techniques 174

A.8 Finding a difference between the Shannon polytopes corresponding to $\mathcal{C}$ and $\mathcal{I}$ for a DAG 176

References 182

LIST OF FIGURES

2.1 Example of a DAG. 25

2.2 Two hypotheses for the causal relationships between observing clouds, rain and wet floor. By intervening on the experiment we can attest that (a) is the correct explanation, but we cannot attest that if we only passively look at the correlations between those events. . . . 26

2.3 Example of a directed acyclic graph (DAG). The probability distributions that are classically compatible with this DAG are those that can be decomposed as in Equation (2.2). 28

2.4 Another example of a directed acyclic graph (DAG). The probability distributions that are classically compatible with this DAG are those that can be decomposed as in Equation (2.3). 28

2.5 According to the d-separation criterion, this DAG imposes among other conditional independences, $A \perp\!\!\!\perp_{\text{CI}} C|B$ on the probability distributions classically compatible with it. 30

2.6 The Bell DAG. It encompasses the assumptions of Bell's theorem for a Bell Scenario where X and Y are the measurement settings of Alice and Bob, A and B are their outcomes and Λ is a classical hidden variable. The probability distributions that are classically compatible with this DAG are those that decompose as in Equation (2.9). 33

2.7 Direct causal influence from A to C can lead to positive correlation between the two nodes (shown with + over the causal arrow). Whereas indirect casual influence from A to C via B can lead to equivalent negative correlation between the two nodes (shown with $-$ over the causal arrows). The net result could be that A and C are independent of each other even though $A \perp_{\text{d}} C$ does not hold in the DAG. This is an example of FINE-TUNING. . . . 36

2.8 The Bell GDAG. Here the latent common cause, Λ denotes a quantum state 40

2.9 Interpretation of a bi-directed edge in the pattern as a DAG. 42

2.10 Interpretation of a directed edge in the pattern as a DAG. 43

2.11 Interpretation of an undirected edge in the pattern as a DAG. 44

2.12 Pattern returned for conditional independences $(AC \perp\!\!\!\perp_{\text{CI}} B)$. 44

2.13 Causal scenarios compatible with conditional independences $(AC \perp\!\!\!\perp_{\text{CI}} B)$. 45

2.14 Pattern returned by the IC^* algorithm for conditional independences $X \perp\!\!\!\perp_{\text{CI}} BY$ and $Y \perp\!\!\!\perp_{\text{CI}} XA$. 46

3.1 The Bell DAG. It encompasses the assumptions of Bell's theorem for a Bell Scenario where X and Y are the measurement settings of Alice and Bob, A and B are their outcomes and Λ is a classical hidden variable. The probability distributions that are classically compatible with this DAG are those that decompose as in Equation (2.9). 47

3.2 The DAGs (a) and (b) are classically observationally equivalent to the mDAG of (c). The step (a)→(b) can be shown by Lemma 20, and the step (b)→(c) can be shown by Lemma 21. 55

3.3 An example of the application of the HLP criterion. All of these three mDAGs have the same set of d-separation relations: $A \perp_{\text{d}} E$, $A \perp_{\text{d}} E|C$,$A \perp_{\text{d}} C$ and $A \perp_{\text{d}} C|E$. Since (c) is latent-free, we can conclude that (a) is NON-INTERESTING. 57

4.1 (a) An example of maximal DAG, which so happens to be INTERESTING. (b) An example of an nonmaximal DAG, namely, The Unrelated Confounders scenario introduced in Ref. [26]. It is not maximal because nodes D and E are not d-separable, but are nevertheless not adjacent. Like *all* nonmaximal DAGs, the Unrelated Confounders scenario is INTERESTING. 61

4.2 A DAG with 4 observed nodes that is shown to be INTERESTING per Theorem 33. This can be seen because G and F are not d-separable, i.e. none of the d-separation relations $(G \perp_{\text{d}} F|E)$, $(G \perp_{\text{d}} F|D)$ or $(G \perp_{\text{d}} F|D, E)$ hold, but they are not adjacent. 61

4.3 Examples of maximal DAGs which are setwise-nonmaximal. (a) is the Triangle scenario, where the set $\{A, B, C\}$ (all visible nodes) is setwise d-unrestricted but not setwise adjacent. In (b), the set $\{A, B, C\}$ is setwise d-unrestricted (in fact, there are no d-separation relations between observed nodes) but not setwise adjacent; by contrast, the larger set $\{A, B, C, D\}$ is *both* setwise d-unrestricted *and* setwise adjacent, despite (b) exhibiting the d-separation relation $(A \perp_{\text{d}} D|B)$, as that d-separation relation involves conditioning on a node *within the set*. In (c), both of the sets $\{A, B, C\}$ and $\{A, B, C, D\}$ are d-unrestricted but not setwise adjacent. Note that $\{A, B, C, D\}$ is d-unrestricted in the DAG (c) despite that DAG exhibiting the d-separation relation $(A \perp_{\text{d}} D|B, C)$, as that d-separation relation involves conditioning on nodes *within the set*. 65

4.4 A DAG with 4 observed nodes and 7 total nodes whose INTERESTINGNESS can be shown by Fraser's algorithm for compatible supports. 75

5.1 Bell DAG embedded in space-time where a NON-CLASSICAL $P(A, B, X, Y)$ is NON-LOCAL and FINE-TUNED. $P(A, B, X, Y)$ is NON-CLASSICAL because it can violate Bell's inequality. It is NON-LOCAL because two space-like separated nodes are correlated and the correlation is not coming completely from their common past light cone but from somewhere outside their past light cones. It is FINE-TUNED because any causal model that can explain it will require some kind of FINE-TUNING. 93

5.2 Bell DAG embedded in space-time where a NON-CLASSICAL $P(A, B, X, Y)$ is LOCAL but FINE-TUNED. $P(A, B, X, Y)$ is NON-CLASSICAL because it can again violate a Bell-inequality. It is LOCAL because both the outcomes and settings of the opposite parties are time-like separated and so all causal influences can propagate subluminally. It is FINE-TUNED because any causal model that can explain it will require some kind of FINE-TUNING. 94

5.3 A FINE-TUNED and LOCAL explanation of NON-CLASSICAL $P(A, B, X, Y)$ Bell-inequality violating QUANTUM-CORRELATIONS. $P(A, B, X, Y)$ can again be NON-CLASSICAL (in the Bell scenario) because it can violate Bell's inequalities. Though it is local because all the nodes are time-like separated and thus support subluminal signals between them. It is FINE-TUNED since corresponding to the observed conditional independence $B \perp\!\!\!\perp_{\mathrm{CI}} X|Y$ there is no corresponding d-separation relation $B \perp_{\mathrm{d}} X|Y$. Via this unfaithful scenario, de-Broglie-Bohmian mechanics can explain Bell inequality violating LOCAL or NON-LOCAL QUANTUM-CORRELATIONS. Note that $P(A, B, X, Y)$ is not NON-CLASSICAL in the DAG of Figure 5.3, it is just FINE-TUNED. 94

5.4 Latent free casual scenario reproducing all the conditional independendces corresponding to scenarios (1), (3) and (6) in Tables 5.1, 5.2.¹ 97

5.5 Latent free casual scenario reproducing all the conditional independendces corresponding to the scenario (5) in Table 5.2.² 100

5.6 Multi-partite Bell scenarios with independent settings 107

5.7 Edges that can be added from Hidden-Variable-to-Output 112

5.8 Edges that can be added from Hidden-Variables-to-Hidden-Variables 113

5.9 Tri-partite and its n-party generalized multi-partite Bell scenarios. 114

6.1 Causal scenario exhibiting a Quantum-Classical gap 118

6.2 Causal scenario exhibiting a Quantum-Classical gap 124

6.3 Another DAG of total six nodes for which $\mathcal{C}_H \subset \mathcal{Q}_H$ by our technique. 126

6.4 The triangle scenario. 128

7.1 Polyhedron corresponding to the constraints in Equation 7.10. 137

7.2 Polyhedron corresponding to the constraints in Equation 7.10 as seen from the perspective of the $x - y$ plane. 137

7.3 Elimination of z in Equations 7.10 results in the projection of the polyhedron in Figure 7.1 on the $x - y$ plane. 138

7.4 Computational benefit of our strategy in causal scenario 1 149

7.5 Computational benefit of our strategy in causal scenario 2 149

7.6 Computational benefit of our strategy in causal scenario 3 150

7.7 Computational benefit of our strategy in causal scenario 4 150

8.1 An interesting causal scenario which can support NON-CLASSICAL correlations. The problem is to check whether it can support NON-CLASSICAL QUANTUM-CORRELATIONS. As for the other figures in this thesis, triangular nodes represent observed nodes and circular nodes represent latent variables which can be entangled quantum states. 155

A.1 (a) depicts a scenario characterised correctly by Pienaar's theorem as INTERESTING, whereas (b) depicts an NON-INTERESTING scenario which Pienaar's theorem incorrectly deems INTERESTING. 162

A.2 A maximal DAG has a set of d-separation relations unmatched by any latent-free DAG (and is thus INTERESTING by Theorem 44) if and only if it has one of these eighteen patterns as a subgraph. 179

A.3 nonmaximal DAG. It does not have any of the DAGs of Figure A.2 as a subgraph, but it nevertheless has a d-separation pattern that does not correspond to any latent-free DAG. . 180

A.4 A DAG with 4 visible nodes and 7 total nodes that is shown to be INTERESTING by our corrected e-separation theorem. This can be seen because it has $(G \perp_{\mathrm{d}} F|E)_{del_D}$ while E is not a descendant of D and none of the d-separation relations $(G \perp_{\mathrm{d}} F|E)$, $(G \perp_{\mathrm{d}} F|D)$ or $(G \perp_{\mathrm{d}} F|ED)$ hold. 180

A.5 Another DAG with 4 visible nodes and 7 total nodes that is shown to be INTERESTING by our corrected e-separation theorem. This can be seen because it has $(G \perp_{\mathrm{d}} F)_{del_{DE}}$ while none of the d-separation relations $(G \perp_{\mathrm{d}} F|E)$, $(G \perp_{\mathrm{d}} F|D)$ or $(G \perp_{\mathrm{d}} F|ED)$ hold. . . 181

LIST OF TABLES

4.1 The only 4 mDAGs with 4 observed nodes for which every support over binary observed variables is classically compatible but which are nevertheless provably INTERESTING by virtue of the higher-cardinality support given in Eq. (4.10) being incompatible with all 4 DAGs here. 77

4.2 The mDAGs of 4 observed nodes whose INTERESTINGNESS remains unresolved. 79

4.3 A summary of our findings: apart from the HLP criterion, which shows NON-INTERESTINGNESS, all of the other conditions listed show INTERESTINGNESS. Note that here we are counting by *unlabelled* DAGs. That is, two labelled DAGs which are equivalent under a relabelling of the observed nodes and/or a relabelling of the hidden nodes are represented by a single unlabelled DAG in these enumerations. 85

5.1 Causal scenarios and their corresponding patterns returned by the IC^* algorithm. 98

5.2 Causal scenarios and their corresponding patterns returned by the IC^* algorithm. 99

5.3 Causal scenarios and their corresponding patterns returned by the IC^* algorithm. 100

6.1 The non-zero $P\Big(F, E, D, C\Big)$ obtained quantum mechanically according to the given description. 129

6.2 The required $P\Big(F, E, D, C\Big)$ obtained quantum mechanically according the description given in the text. 130

7.1 Results for Random Polytopes 151

A.1 Results for mDAGs of 5 observed nodes. 169

A.2 The mDAGs of 4 observed nodes whose INTERESTINGNESS remains unresolved. 175

A.3 We enumerate the 2809 possible mDAGs of four visible nodes. Applying HLP's criterion for $\mathcal{C} = \mathcal{I}$ reduces them to 996. Then we check if any of the left 996 mDAGs have conditional independences that cannot be reproduced by any latent free DAG. These mDAGs as explained are definitely Interesting. After accounting for these mDAGs we are left with 978 mDAGs which have conditional independences that can all be reproduced by some latent free DAG. So we check how many mDAGs out of 978 are such that they have at least one e-separation relation that cannot be achieved in their corresponding latent free mDAG which had all the same d-separation relations as the mDAG itself. Again such mDAGs are definitely Interesting. Accounting for this leaves us with a figure of 168 mDAGs that could be potentially Interesting. To affirm their Interestingness we refer to Fraser's support algorithm and find supports that are not infeasible in the mDAG but are feasible in its corresponding latent free mDAG with all the same d-separation and e-separation relations as the mDAG, for all but just three mDAGs. 176

A.4 We enumerate both mDAGs of four visible nodes and GDAGs of total nodes seven. Applying HLP's criterion for $\mathcal{C} = \mathcal{I}$ reduces them to 996 and $1,618,679$ respectively. We further apply HLP's reduction criteria on them which returns that 909 $(996 - 87)$ left mDAGs are reducible to 87 mDAGs and $1,618,600$ $(1,618,679 - 79)$ left GDAGs are reducible to 79 DAGs. This just means that if the corresponding 87 mDAGs and 79 GDAGs are Interesting then all the 996 mDAGs and all the $1,618,679$ GDAGs are also Interesting. Hence we are finally left with 87 mDAGs and 79 GDAGs to check for Interestingness. Application of our theorem based on e-separation leaves 54 mDAGs and 20 GDAGs to be checked out of which Fraser's algorithm provides classically infeasible supports for 51 mDAGs and 19 GDAGs. Note that out of these 20 DAGs only 16 are mDAGs and the rest of the 4 GDAGs are equivalent to either of these 16 mDAGs but the one for which we couldn't find any classically infeasible supports. 177

LIST OF ALGORITHMS

1 Using Fourier-Motzkin elimination along with Imbert's theorems and linear programming to detect redundancies. 144

2 Accelerating Fourier-Motzkin elimination by eliminating that variable which produces the least number of non-redundant constraints 147

Acknowledgments

First and the foremost I would like to express my deepest gratitude to my parents. Their unwavering support, endless encouragement and boundless love have been my foundation. They have always believed in me and my dreams that they have supported throughout. Whether through words of wisdom or affirming and uplifting advice, they have provided me the strength and motivation to reach this point. Their belief in my dreams has driven me and guided me during challenging times. Without their continuous encouragement and backing, which have been invaluable, this achievement would not have been possible at all. They have always provided me with resources and opportunities necessary to pursue my ambitions for which I am eternally grateful to them. Thanks in believing in me a lot.

I am profoundly indebted to my supervisors, Matthew Pusey and Roger Colbeck. Their expertise, guidance and mentorship have been crucial to my development as a researcher. Matt, your insightful feedback and innovative thinking have profoundly shaped my work. Meticulous discussions with you on various foundational issues such as interpretations of Bell's theorem and Quantum Causality have inspired my thinking a lot. Your ability to challenge my ideas and push me to think critically has been instrumental in refining my research. Roger, your patience, encouragement and support have been unwavering. Your guidance, your way of approaching a problem and constructive criticism have significantly improved the quality of my work. I am deeply appreciative of the time and effort that you have both invested in me. It has been a pleasure working with both of you. I hope to continue working with you both.

I also deeply thank my collaborators, Marina Maceil Ansanelli and Elie Wolfe. Without their consistent inputs, this work would not have been possible. My deepest thanks goes to Chris Fewster for teaching me General Relativity and Quantum Field Theory in a way I had never learnt before. I am deeply indebted to Stefan Weigert too. Without his guidance and support, I would not have been doing this PhD. Thanks to all the faculty, whose courses I took as a postgraduate, for answering all my out of the syllabus questions. My deepest gratitude goes to the Department of Mathematics as well, for not

only just awarding me with the funding to carry on this PhD but also all the funds I got for attending the conferences I did. Thank you all the people behind it.

I would also like to extend my heartfelt thanks to my friends. Your continuous support and camaraderie have been a source of great comfort and joy. Special thanks to Zosia and Rutvij, for being by my side when I needed them, to David, Jack, Jintao and Hamid for all for all the fun moments we had, to Rutvij (again), Vasilios, Laurens, Max and Vilasini for all the fruitful discussions we had ranging from implications of Bell's theorem to the nature of reality. With Max I had very enriching discussions on the Theory of Relativity, tachyonic signals and symmetries of space time for which I am very thankful to him. I am thankful to Alice for helping me out at various places in the Mathematics when I needed her help. Lastly, I would also like to thank Ding, Berend, Sam, Philip and Simon for being good friends throughout my PhD and helping me at various places. Whether through offering a listening year, providing words of reassurance, genuine advice, enriching discussion or simply sharing a laugh, you have made this journey bearable but also enjoyable. The moments of shared joy and support during tough times have made this journey memorable.

To all of you, I extend my sincerest thanks. Your support and encouragement have made this achievement possible.

Author's declaration

I declare that the work presented in this thesis, except where otherwise stated, is based on my own research carried out at the University of York and has not been submitted previously for any degree at this or any other university. Sources are acknowledged by explicit references.

Research work included in this Thesis

Chapters 3 and 4 constitute the content of the published work: Shashaank Khanna, Marina Maciel Ansanelli, Matthew F. Pusey, and Elie Wolfe, **"Classifying causal structures: Ascertaining when classical correlations are constrained by inequalities" [53]**. It has been published in Physical Review Research **6**, 023038 (2024), https://journals.aps.org/prresearch/pdf/10.1103/PhysRevResearch.6.023038.

Chapter 5 comprises the work: Shashaank Khanna, Matthew F. Pusey, Roger Colbeck, **"Finding causal scenarios which cannot support QUANTUM-CORRELATIONS without FINE-TUNING "**. It is intended to be submitted for publication this year.

Chapters 6 and 7 have been written with the intention of being published as well. Both the chapters will constitute a Regular Article each. Consequently, the corresponding chapters have been written as articles themselves.

Accompanying Code

All the chapters required significant computational work. The necessary code [47, 46, 48] used can be found here: https://github.com/shashaank38/gdag_code and here https:

//github.com/eliewolfe/mDAG-analysis.

Research work not included in this Thesis

The following articles, not included in this thesis, were published during the period in which this thesis was written.

1. [103] L. Walleghem, S. Khanna, and R. Bhavsar. "Comment on a no-go theorem for ψ-ontic models". In: arXiv preprint arXiv:2402.13140 (2024). https://arxiv.org/pdf/2402.13140
2. [54] S. Khanna, S. Halder, and U. Sen. "Quantum entanglement percolation under a realistic restriction". In: Physical Review A 109.1 (2024), p. 012419. https://journals.aps.org/pra/pdf/10.1103/PhysRevA.109.012419

Portions of this thesis were proof read by online tools such as Grammarly and ChatGPT. All the assistance taken was restricted to proofreading, identification of grammatical errors, usage of correct sentence structure and synonyms.

— 1 —

Introduction

1.1 Preface

Quantum Mechanics has been one of the two leading theories of the last century. However, since its inception it has been surrounded by counter-intuitive ideas ranging from particles exhibiting interference and superposition properties [67] to particles being in an "entangled" state [3] where the measurement on one particle seems to have an instantaneous effect on the other particle as well, no matter how large the distance between the two particles is. Along with this the theory has been masked with philosophical difficulties ranging from the nature of quantum state of a particle, whether it is real or epistemic [86, 43], to the collapse of the wave-function. These problems have not only puzzled the founders of the theory but a generation of researchers. Solutions to some of these problems like the nature of the wave-function and its collapse have resulted in various interpretations of Quantum Mechanics.

However, one the most non-classical features that Quantum Theory exhibits over any of the classical theories was only discovered in 1964 when John Bell [10, 12, 41] published his eponymous theorem. Einstein, Podolski and Rosen (EPR from here-on) [25, 71] in 1935 had shown that to respect a particular sensible notion of "Locality", Quantum Mechanics must be an incomplete theory. That is, if Quantum Theory has to be local in the sense of EPR then it must possess certain hidden variables, which would then restore completeness to the theory. Bell in 1964 showed that no matter what hidden variables are assumed to make quantum theory complete, they cannot render the theory local. In other words, Bell showed that any theory that reproduces all the results of Quantum Mechanics cannot be local in the sense of EPR. Bell called this condition Local Causality.[1] Thus according to Bell's theorem, Quantum Mechanics violates Local Causality.

[1] Actually another assumption, called measurement independence is also needed to arrive at Bell's result.

Having said all about the various perplexing concepts in quantum theory, it must be noted that it has been validated again and again via various rigorous experiments. Be it the study of spectra of various atoms, the prediction of different fundamental particles, technological uses like in semiconductors, or nuclear physics, quantum theory has always succeeded with utmost precision wherever it has been put to test. Crucial questions asked by a generation of researchers to answer their curiosity have led Quantum Mechanics to many new fields and revolutionary ideas such as: Information processing using laws of quantum mechanics, quantum computing, quantum optics, quantum thermodynamics, positron emission tomography and lasers. Nevertheless, several foundational questions remain to be answered. What physical principles define Quantum Theory? Does Quantum Theory violate Relativity [11, 75, 105, 22, 24, 68]? What is an axiomatic development of Quantum Theory? All these fundamental questions remain a topic of interest for a new generation of researchers.

Coming to Bell's theorem again, loosely it means that in certain cases in Quantum Mechanics, causes seem to have effects even outside their light cones. Thus, Bell's theorem has sparked various ideas at the interface of Quantum Mechanics and Causality, and the importance of studying causality in the context of Quantum Mechanics has been highlighted by several researchers. Wood and Spekkens [108] showed that any classical causal model, based on the framework of classical causality, cannot truly replicate quantum correlations. They show that correlations obtained in a certain scenario or an experiment using entangled quantum sources have no classical causal explanation unless one resorts to some fine-tuning of the parameters in the model. In other words, the field of classical causality is insufficient to explain Quantum Mechanics. This quantum scenario they studied is exactly what John Bell considered in 1964 and their result is Bell's theorem, recast using the language of causality and directed acyclic graphs.

Further in their seminal paper, Henson, Lal, and Pusey [44] have also shown that besides the scenario Bell considered there exist many more quantum scenarios exhibiting peculiar correlations that offer no classical causal explanation. What this means is that besides what Bell considered there exist numerous more examples of correlations where an observed event does not have a cause that can classically explain it. Such scenarios are called INTERESTING in the literature and lay largely unexplored for quantum advantages.

The non-classical correlations obtained in such scenarios form a semi-algebraic set of correlations in the probability space. Therefore, the rigorous mathematical name for these scenarios is also NON-ALGEBRAIC scenarios. A large portion of this theses deals with methods to characterise these INTERESTING scenarios. Therefore, it is important to point out why their characterisation is an important subject and why it should be highly investigated by researchers in this field.

Why is studying INTERESTING scenarios Important?

Classical causality cannot always explain cause-and-effect relationships at the atomic level [108]. Therefore, the first step in studying causality-based problems related to quantum physics is to characterise scenarios where classical causality fails to causally explain the observed effects. This is because discrepancies between classical and quantum physics can possibly show up in these scenarios only. For this precise reason such scenarios are termed INTERESTING in the literature. Once we know the scenarios where classical causality fails to deliver any explanation, we can proceed to investigate such INTERESTING scenarios. Characterising such scenarios is what we study in [53]. Once such INTERESTING scenarios have been characterised, we can try to modify the theory of classical causality in such a way that it now renders an explanation for even these INTERESTING scenarios. This constitutes the subject of Quantum Causal Modelling [4] and is of significant importance from foundational perspectives.

Understanding the issue from a different angle, suppose that there is a theory more fundamental than quantum mechanics that remains unknown. How do we find out if there exists such a theory [8, 83]? Where do we look to find the traces of the existence of such a theory? It is precisely these INTERESTING scenarios that can manifest such confirmations. Only such INTERESTING scenarios can support any features that such a deeper theory might exhibit. Therefore, observing correlations corresponding to such INTERESTING scenarios that cannot be explained classically is a direct proof of the subsistence of such a deeper theory. This is similar to how evidences of quantum theory were found in the black-body spectrum and atomic transitions.

Moving on from the foundational advantages of this study, there are several practical implications of this work as well. A branch stemming from quantum theory is that of Quantum Information. Quantum information refers to the study and application of information processing using principles from quantum mechanics. In classical computing, information is stored and manipulated using bits, which can be either 0 or 1. However, in quantum computing, quantum bits or qubits are used. Qubits can exist in a quantum superposition or importantly be entangled which means they can support correlations stronger than any classical theory. These stronger correlations find applications in many information processing tasks such as various device-independent protocols [2, 99, 19, 1, 82]. Device independence in quantum information refers to a concept in quantum cryptography and quantum information processing in which the security or functionality of a quantum protocol is guaranteed even if the devices used to implement the protocol are not trusted. In other words, the security of quantum communication or computation is not dependent on the properties or honesty of the quantum devices themselves.

Examples of device-independent protocols are quantum key distribution and quantum random number generators. Device-independent quantum key distribution protocols aim to ensure the security of quantum key distribution, a method for securely sharing cryptographic keys over a quantum channel.

Even if the quantum devices used to generate and measure quantum states are potentially flawed or manipulated, the security of the key exchange remains intact. On the other hand, device-independent methods are also used to generate random numbers in a quantum setting. Even if the devices used for quantum random number generation are untrusted or subject to adversarial behaviour, the randomness properties of the generated numbers can still be guaranteed.
But the question is in what settings should we expect the advantages of these device-independent protocols over already existing classical strategies? These INTERESTING scenarios and none apart from them are the precise candidates that need to be studied to map the advantages of these device-independent protocols over the existing classical strategies. Hence, these INTERESTING scenarios which so far have not been investigated are potential candidates for various device-independent advantages. Having understood the importance of characterizing INTERESTING scenarios, we can now delve into a summary of the main contributions of this thesis.

1.2 Structure of the Thesis

This thesis consists of eight chapters in total. Chapters 2-6 focus on studying INTERESTING causal structures whereas Chapter 7 is a digression into Fourier-Motzkin Elimination algorithm. Chapters 4,5,6,7 constitute the original contributions of this thesis and they are based on the published and upcoming work [53, 55, 56]. We now briefly summarize the contents of all the Chapters except this one.

Chapter 2 introduces the mathematical background and the preliminaries needed to understand the rest of this thesis. The concept of Directed Acyclic Graphs (DAGs) is introduced along with the causal explanation of observed data using the formalism of DAGs. Special emphasis is placed on graphical conditions like d-separation via examples. DAGs permitting latent variables and non classical causal models are introduced. We briefly venture into Causal Discovery Algorithms focussing on one of them, in particular the IC^* algorithm.

Chapter 3 introduces the concept of an INTERESTING DAG or an INTERESTING causal scenario. The mDAG formalism and the HLP criterion are explained with references to specific examples. Chapter 4 consists of the majority of the results in this thesis. Specifically we talk about all the methods to determine INTERESTINGNESS of causal scenarios (DAGs). Non-maximality, d-separation, e-separation and incompatible supports are the methods explained to determine INTERESTINGNESS of a DAG. The entropy vector method and Shannon-type inequalities to classify INTERESTING scenarios is also explained. Lastly the three mDAGs which we could not classify as INTERESTING or not are shown.

Chapter 5 deals with applying the IC^* algorithm to certain INTERESTING causal scenarios and finding those INTERESTING causal scenarios which are the only possible candidates to explain the observed conditional independences supported by them respectively. Explaining certain QUANTUM-CORRELATIONS in them would require FINE-TUNING. It is then proven that any multi-partite Bell scenario will require FINE-TUNING to explain certain Bell inequality violating QUANTUM-CORRELATIONS.

Chapter 6 deals with finding a quantum-classical gap in two specific scenarios. By introducing a method based on the observation of certain extra observed conditional indpendences in the scenarios, we are able to embed the Bell scenario in them and use the known result $\mathcal{C} \neq \mathcal{Q}$[2] for the Bell scenario to show that $\mathcal{C} \neq \mathcal{Q}$ also holds for them.

Finally, in chapter 7, we digress into Fourier-Motzkin elimination. Discussing Imbert's theorem we show why combining Imbert's theorem along with linear programming for redundancy removal is not so trivial. However, we provide a way to combine them both. The main result of the work is then presented; a technique based on the order of elimination of variables to accelerate the whole Fourier-Motzkin elimination process. We provide an algorithm to do so.

Chapter 8 in the end provides a combined conclusion where certain ideas for related problems and approaches towards their solution are proposed.

1.3 Notational Conventions

Unless otherwise specified, throughout this thesis we will employ the following notational conventions.

- **Random Variables, Probability Distributions and Entropy:** We will use capital letters to denote random variables and use their corresponding small letters to denote the value the random variable takes. For example: $X : S \to \mathbb{R}$ denotes the random variable "outcome of a coin toss", where $S = \{H, T\}$ is the sample space of the random variable X. The possible realisations of X are thus denoted by $x \in \mathbb{R}$, with, say, $x = -1$ denoting "heads" and $x = 1$ denoting "tails". Probability distributions over a set of random variables $\{X_1, X_2, \dots X_n\}$ will be denoted by $P_{X_1,X_2,\dots X_n} \in \mathcal{P}_{X_1,X_2,\dots X_n}$, where $\mathcal{P}_{X_1,X_2,\dots X_n}$ denotes the space of all probability distributions over the set of variables $\{X_1, X_2, \dots X_n\}$. Sometimes we will denote $P_{X_1,X_2,\dots X_n}$ as simply $P(X_1, X_2, \dots X_n)$. When we have to mention the specific values the random variables take, we will abbreviate $P_{X_1,X_2\dots X_n}(X_1 = x_1, X_2 = x_2, \dots X_n = x_n)$ using $P(X_1 = x_1, X_2 = x_2, \dots X_n = x_n)$ or more simply by just $P(x_1, x_2, \dots x_n)$. Any equation concerning probabilities written only in terms of the random variables must be understood to hold for all the value of the random variables. For example $P_{XY} = P_X P_Y$, denoting the

[2] Sets $\mathcal{C}$ and $\mathcal{Q}$ are defined in Chapter 3.

conditional indepedence of X and Y, is the same as $P(X,Y) = P(X)P(Y)$ and it must be interpreted as $P(X = x, Y = y) = P(X = x)P(Y = y)\ \forall\ x, y$. And thus, the conditional independence of the random variable A of the variable C given the variable B in this notation means $P(A,C|B) = P(A|B)P(C|B)$. Only in Chapter 5, section 5.4 even small letters are used to denote just the random variable itslef rather than its value. The union of two random variables X, Y will be denoted as XY. The entropy of a random variable X will be denoted as $H(X)$. Joint entropies of random variables $XY\dots$ will be denoted as $H(XY\dots)$.

- **Quantum Formalism:** The Hilbert Space associated with any quantum system An will be denoted as $\mathcal{H}_A$. $S(\mathcal{H}_A)$ will denote the set of all positive semi-definite trace one operators on $\mathcal{H}_A$. A quantum state i.e., a density matrix corresponding to a quantum system A will be denoted as ρ_A. Hence $\rho_A \in S(\mathcal{H}_A)$. Hermitian conjugate will be denoted by a dagger, for example $\rho_A^\dagger$. Transpose of a matrix will be denoted by T, for example ρ_A^T. Rank one density operators, that is pure states, will be denoted via the usual bra-ket notation. So $|\psi\rangle_A \in \mathcal{H}_A$ and $|\psi\rangle_A^\dagger = {}_A\langle\psi| \in \mathcal{H}_A^*$, where $\mathcal{H}_A^*$ denotes the dual space of $\mathcal{H}_A$. Quantum measurements will be described by positive operator-valued measures (POVMs). A POVM acting on a system A will be denoted as $\{E_X^A\}$. It is a set of all $E_X^A \in P(\mathcal{H}_A)$, indexed by the values of the random variable X and where $P(\mathcal{H}_A)$ is the set of all positive, semi definite operators on $\mathcal{H}_A$, such that $\sum_X E_X^A = \mathbb{1}$. Here the summation is understood to be over the values of the random variable X. We will this convention throughout this thesis, especially in Chapter 5. Further, $\otimes$ will be used to denote the tensor product of Hilbert Spaces, for example $\mathcal{H}_{AB} = \mathcal{H}_A \otimes \mathcal{H}_B$. Similarly for simplicity we will use $|\psi\rangle_A \otimes |\phi\rangle_B = |\psi\phi\rangle_{AB} \in \mathcal{H}_{AB}$ to denote the tensor product of two states.
- **Causal and space time diagrams:** Throughout the thesis regular arrows i.e., → between two random variables will denote a possible causal influence from one variable to the other. For example $A \to B$ denotes that A is the cause of B. In Section 2.10 and Chapter 5, we will also use bi-directed (↔) and undirected (−) arrows. Their meanings will be made clearer in Section 2.10. Also, space time diagrams in this thesis will have space running along the X axis and time increasing along the Y axis.
- **Logical operators and others:** We will use $\neg$ to denote the standard "not" operation, $\wedge$ and $\vee$ to denote "and" and "or" operations respectively. $\implies$ would be used to denote if-then propositions and $\iff$ would be used to denote equivalency of two mathematical statements. $\neq$ means not equal to and $\oplus$ means addition modulo 2 of two classical bits (or equivalently their XOR).

— 2 —

Mathematical Introduction to Causality

2.1 Introduction

Since ages people have wondered about causes of observed effects and there have been various philosophical approaches to understanding causality including Aristotelian and Kantian causality. However, in this thesis we will not discuss about the various philosophical implications of causality, but rather begin with Reichenbach's [87] notion that observed correlations must have a causal explanation.

Many times in science we are not only interested in the observed correlations between events, but also in the underlying causal explanations that govern such observed phenomena. In the mathematical framework for causal inference [76, 101, 77, 95, 59], the candidates for these causal explanations are represented by directed acyclic graphs (DAGs), where each node is associated with a variable and each edge represents a direct causal influence.

A DAG imposes constraints on the probability distributions that can be causally explained by it. For example, a probability distribution over variables $\{A, B, C\}$ where A and C remain correlated even after a value of B is conditioned upon *cannot* be causally explained by the DAG of Figure 2.1. All of the distributions over $\{A, B, C\}$ that can be explained by Figure 2.1 need to satisfy $P(A, C|B) = P(A|B)P(C|B)$.[1]

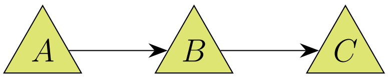


Figure 2.1: Example of a DAG.

The causal compatibility constraint described above, denoted by $A \perp\!\!\!\perp_{\text{CI}} C|B$, is a *conditional*

[1] We will denote probabilities by the concise notation $P(XY) \equiv P_{X,Y}(X = x, Y = y)$. Only in Chapter 5, section 5.4 even small letters will denote the random variable rather than its value.

independence relation. That is, it says that one set of variables (A) becomes independent of a second set of variables (C) when conditioning on a third (B). In general, however, a DAG can impose more complicated types of constraints on the compatible probability distributions. This only happens for DAGs that have latent nodes, i.e. nodes associated with variables that do not appear in the probability distributions of interest.

From now on, the variables that do appear in the probability distributions of interest will be called *observed variables*, while the ones that do not will be called *latent variables*. Nodes associated with observed variables will be called *observed nodes* and will be represented by triangles, while nodes associated with latent variables will be called *latent nodes* and will be represented by circles.

To understand how cause and effect relations between observed and latent variables can be understood, first we need to brush up some basic concepts in the field of Causality.

2.2 Causal Explanations of Observational Data

The area of causal inference is concerned with finding potential causal explanations for observed events. For example, imagine that we want to find out what is the causal relationship between three events: a cloudy sky, rain and the floor being wet. Figure 2.2 depicts two possible causal structures between these three events; in 2.2(a) we hypothesize that the clouds cause the rain and the rain makes the floor wet. In 2.2(b), on the other hand, we hypothesize that the wet floor causes the rain, and the rain makes the sky cloudy.

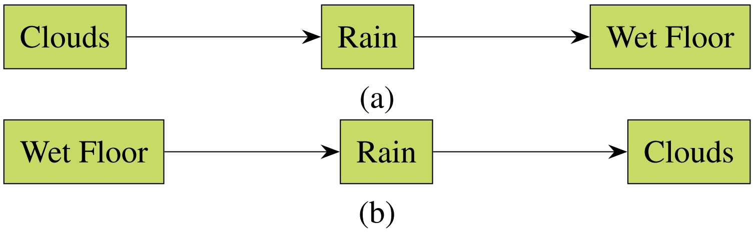


Figure 2.2: Two hypotheses for the causal relationships between observing clouds, rain and wet floor. By intervening on the experiment we can attest that (a) is the correct explanation, but we cannot attest that if we only passively look at the correlations between those events.

There is an easy way we can check that Figure 2.2(b) is not the correct causal hypothesis: if we pour water on the floor in a sunny day, it will not start raining.

Note that this method presupposes that we can *intervene* on our experiment, meaning that we can force one of the variables of the experiment (wet floor) to assume the value we want. However, it is not always possible to do that; sometimes it is unethical or there are technical or fundamental limitations to do so.

If the experimenter cannot make interventions on the variables of interest, it is still possible for them to draw some conclusions from a passive observation of the correlations between the events of interest. As we will see, each causal structure imposes constraints on which probability distributions obtained from passive observations can be classically explained by it. These constraints can be tested against the observed probability distribution to see if the given causal structure is a valid causal hypothesis for the observed phenomena. In Figure 2.2 it happens that both causal structures impose the same constraints on probability distributions, so they are not distinguishable by passive observations. Precisely, any probability distribution $P(A, B, C)$ satisfying $P(A, C|B) = P(A|B)P(C|B)$, with the variables A interpreted as "Clouds", B as "Rain" and C as "Wet Floor", is compatible with both the causal scenarios of Figure 2.2.

2.3 Directed Acyclic Graphs

In the framework we use here, the mathematical object that describes a causal structure is a directed acyclic graph (DAG).

Definition 1 (DAGs)**.** *A directed graph G is a pair $(\boldsymbol{A}, \boldsymbol{E})$, where $\boldsymbol{A}$ is a set of nodes and $\boldsymbol{E} \subseteq \boldsymbol{A} \times \boldsymbol{A}$ is a set of directed edges. A directed acyclic graph (DAG) is a directed graph that has no directed cycles. Below, we introduce some definitions regarding DAGs that will be useful later.*

Definition 2 (Children, Parents, Descendants, Ancestors). *Let X be a node of a DAG G. If Y is another node of G such that there is a directed edge $X \rightarrow Y$, then Y is called a child of X, and X is called a parent of Y. The set of all children of X is denoted as CH$_G(X)$, and the set of all parents of X is denoted as PA$_G(X)$.*

A directed path is a sequence of nodes $X^1, X^2, X^3......X^n$ such that $X^i \rightarrow X^{i+1}$ for $i = 1, ..., n$. The descendants of X in G are all the nodes in G that can be reached from X by a directed path. Conversely, all the nodes that have X as a descendant are called ancestors of X. The set of all ancestors of X is denoted as ANC$_G(X)$, meanwhile the set of all descendants of X can be denoted as DES$_G(X)$. [2]

[2] In this thesis we follow the convention of Refs. [90, 97, 112] and others, namely, the convention in which $X \in \text{ANC}_G(X)$ and $X \in \text{DEC}_G(X)$ but $X \notin \text{PA}_G(X)$ and $X \notin \text{CH}_G(X)$. That is, a node is considered it own ancestor and its own descendant, but not its own parent or its own child.

For example, in the DAG of Figure 2.3 we have that $\mathrm{PA}_G(B) = \{A, D\}$, and that E is a descendant of D, even if it is not its child.

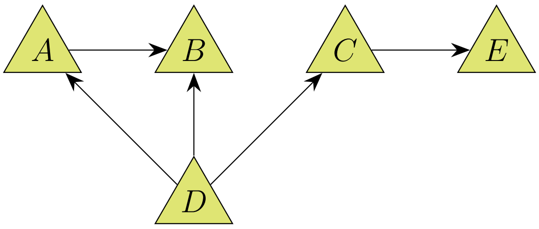


Figure 2.3: Example of a directed acyclic graph (DAG). The probability distributions that are classically compatible with this DAG are those that can be decomposed as in Equation (2.2).

As another example consider the DAG of Figure 2.4. In it we have that $\mathrm{PA}_G(B) = \{A, D, C, F\}$, and that E is a descendant of D, even if it is not its child. F is a non-descendant of A even though F is not an ancestor of A. This is a valid DAG since it does not contain any directed path both beginning and ending at the same node (i.e a directed cycle). Also, note that we draw the nodes as triangles.

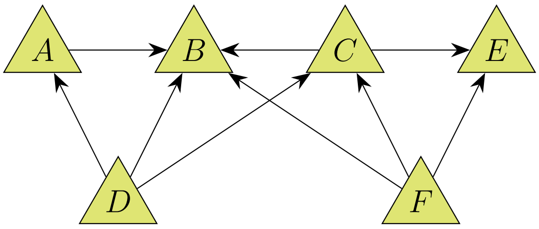


Figure 2.4: Another example of a directed acyclic graph (DAG). The probability distributions that are classically compatible with this DAG are those that can be decomposed as in Equation (2.3).

2.4 DAGs as Causal Structures

When we associate each node of a DAG with an event of interest, the DAG is a representation of a causal structure: an edge $X \rightarrow Y$ shows a possibility of a direct causal influence of X on Y. Here, we will indicate the variable associated with a node by the same capital letter as the node itself. If all the events of interest are described by classical random variables, the idea that a probability distribution "can be causally explained" by a certain DAG is formalized through the *Markov condition*:

Definition 3 (Markov Condition). *Let G be a DAG with nodes $\boldsymbol{A}$. A probability distribution P over the variables $\boldsymbol{A}$ is said to be* Markov with respect to G *if it can be factorized as:*

$$P(\boldsymbol{A}) = \prod_i P\left(X^i | PA_G\left(X^i\right)\right) \tag{2.1}$$

Where $\boldsymbol{A} = \{X^1, ..., X^n\}$ *and* $PA_G\left(X^i\right)$ *is the set of parents of the node* X^i *in* G*.*
If P *is Markov with respect to* G*, we also say that it is " classically compatible" with* G*.*

As an example of this definition, a joint probability distribution P_{ABCDE} over the random variables A, B, C, D and E is Markov with respect to the DAG of Figure 2.3 if it can be decomposed as:

$$P(ABCDE) = P(A|D)P(B|A, D)P(C|D)P(D)P(E|C) \tag{2.2}$$

As a second example of this definition, a joint probability distribution P_{ABCDEF} over the random variables A, B, C, D, E and F is Markov with respect to the DAG of Figure 2.4 if it can be decomposed as:

$$P(ABCDEF) = P(A|D)P(B|A, C, D, F)P(C|D, F)P(D)P(E|C, F)P(F) \tag{2.3}$$

Any probability distribution on A, B, C, D, E, F that can be factorized as in Equation 2.3 is compatible with the DAG of Figure 2.4. Equivalently, any probability distribution over A, B, C, D, E, F which cannot be factorised as in Equation 2.3 is incompatible with this DAG.

Therefore, through the Markov condition, each DAG imposes constraints on the probability distributions that are classically compatible with it. A DAG which is classically compatible with *every* probability distribution, that is, a DAG that does not impose any constraint on the classically compatible distributions, is said to be a *saturated* DAG.

2.5 d-separation: A graphical criterion for conditional independence

If P is a probability distribution on a certain set of random variables $\boldsymbol{A}$, we say that the variables of the subset $\boldsymbol{X} \subseteq \boldsymbol{A}$ are *conditionally independent* of the variables of the subset $\boldsymbol{Y} \subseteq \boldsymbol{A}$ given the subset $\boldsymbol{Z} \subseteq \boldsymbol{A}$ if P can be factorized as $P(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{Z}) = P(\boldsymbol{X}|\boldsymbol{Z})P(\boldsymbol{Y}|\boldsymbol{Z})$. This is denoted by $\boldsymbol{X} \perp\!\!\!\perp_{\text{CI}} \boldsymbol{Y}|\boldsymbol{Z}$.

Some of the constraints that a DAG imposes on the probability distributions that are classically compatible with it are of the form of conditional independence: if a probability distribution P *cannot* be factorized according to a certain conditional independence that is imposed by the DAG, then it is not classically compatible with the DAG. As it turns out, there exists a graphical algorithm to obtain all the conditional independence constraints that are imposed by a DAG. This algorithm is called "d-separation", and we will now describe it.

Definition 4 (d-separation)**.** *If* G *is a DAG with nodes* $\boldsymbol{A}$ *and* $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z} \subseteq \boldsymbol{A}$ *are sets of nodes in* G*, the* d*-separation algorithm says whether* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *are "*d*-separated" by* $\boldsymbol{Z}$*.* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *are "*d*-separated"* *by* $\boldsymbol{Z}$ *if all the undirected paths (paths that ignore the direction of the arrows) from* $\boldsymbol{X}$ *to* $\boldsymbol{Y}$ *are* blocked *by* $\boldsymbol{Z}$. *A path is blocked if one or more of the following is true:*

1. *There is a chain of nodes along the path:* $i \rightarrow m \rightarrow j$ *such that,* $i \in X, j \in Y$ *and* $m \in Z$.
2. *There is a fork along the path:* $i \leftarrow m \rightarrow j$ *such that,* $i \in X, j \in Y$ *and* $m \in Z$.
3. *There is a collider along the path:* $i \rightarrow m \leftarrow j$ *such that,* $i \in X, j \in Y$ *and* $m \notin Z$ *and* $d \notin Z$ *for all the descendants* d *of* m.

If X *is* d*-separated of* Y *given* Z *in the DAG under consideration, we denote it as* $X \perp_d Y|Z$. *If* Z*, is the empty set we simply denote that* X *is* d*-separated of* Y *by* $X \perp_d Y$.

For example, for the DAG of Figure 2.1 the d-separation criterion says that $A \perp_{\mathrm{d}} C|B$. This means that the event A should become independent of the event C upon knowledge of B; if your distribution P *does not* satisfy the constraint $P(AC|B) = P(A|B)P(C|B)$, Figure 2.1 is *not* a valid causal explanation for it. Interpreting the variables $\{A, B, C\}$ as, respectively, clouds, rain, and wet floor (such as in Figure 2.2(a)), this d-separation relation says that the occurrence of clouds becomes independent of the floor being wet if we already know whether it is raining.

As a second example, consider the DAG in Fig. 2.5. For it the d-separation criterion says that $A \perp_{\mathrm{d}} C|B$ (due to the colliders at D and E and the fork at B) and $B \perp_{\mathrm{d}} E|AC$ (due to the chains at A and C and collider at D). This means that the event A (B) should become independent of the event C (E) upon knowledge of B (AC). So if for example a distribution P *does not* satisfy the constraint $P(AC|B) = P(A|B)P(C|B)$, then it is not compatible with the DAG of Fig. 2.5 and thus Fig. 2.5 is *not* a valid causal explanation for it.

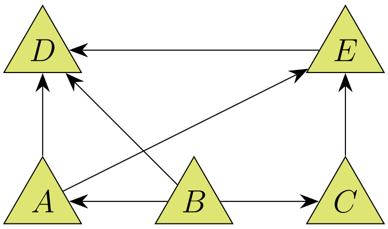


Figure 2.5: According to the d-separation criterion, this DAG imposes among other conditional independences, $A \perp\!\!\!\perp_{\mathrm{CI}} C|B$ on the probability distributions classically compatible with it.

The following theorem, proven in Ref. [101], makes explicit the connection between d- separation relations and conditional independence relations:

Theorem 5 (d-separation and conditional independence [101])**.** *Let* G *be a DAG with nodes* $\boldsymbol{A}$*, and let* $\boldsymbol{X} \subseteq \boldsymbol{A}$*,* $\boldsymbol{Y} \subseteq \boldsymbol{A}$ *and* $\boldsymbol{Z} \subseteq \boldsymbol{A}$ *be three disjoint subsets of* $\boldsymbol{A}$*. Then:*

1. *If G has the d-separation relation $\boldsymbol{X} \perp_d \boldsymbol{Y} | \boldsymbol{Z}$, then all of the probability distributions over the variables $\boldsymbol{A}$ which are Markov with respect to G need to satisfy $\boldsymbol{X} \perp\!\!\!\perp_{CI} \boldsymbol{Y} | \boldsymbol{Z}$.*
2. *If G* does not have *the d-separation relation $\boldsymbol{X} \perp_d \boldsymbol{Y} | \boldsymbol{Z}$, then there exists some probability distribution over the variables $\boldsymbol{A}$ which is Markov with respect to G and* does not *satisfy $\boldsymbol{X} \perp\!\!\!\perp_{CI} \boldsymbol{Y} | \boldsymbol{Z}$.*

When a DAG does not have latent nodes, the *only* constraints that it imposes on the compatible distributions are the conditional independence relations associated with the d-separation relations of the DAG [36, Theorem 3]:

Theorem 6 (d-separation is complete for DAGs without latent nodes [36])**.** *Let G be a latent free DAG with nodes $\boldsymbol{A}$. A probability distribution over the variables $\boldsymbol{A}$ is Markov with respect to G if and only if it satisfies all the conditional independence relations associated with the d-separation relations of G.*

Importantly, in general this is *not* true for DAGs that include latent nodes, as we will see in the next subsection. In a DAG without any latent nodes, it is possible to verify the Markov condition without relying on the d-separation conditions. The reason being that the probability distribution over all nodes in the graph is known. Thus, by marginalising the joint probability distribution, the probabilities for each node can be obtained and the Markov condition can be confirmed. In other words, in the absence of latent nodes, the Markov condition implies the d-separation relations and vice versa. However, when latent variables are present, satisfying the observed conditional independences may not be sufficient to satisfy the Markov condition. This is because the d-separation relations do not necessarily imply the Markov condition when latent nodes are present.

Remark 7. *It is mentioned in Section 2.1 that all of the distributions over $\{A, B, C\}$ that can be explained by Figure 2.1 need to satisfy $P(A, C|B) = P(A|B)P(C|B)$. However, as this condition stands it does not seem to satisfy the Markov condition for the causal scenario in Figure 2.1. That it nevertheless does satisfy the Markov condition for this scenario can be seen as follows:*

$$P(A, B, C) = P(A, C|B)P(B) = P(A|B)P(C|B)P(B) = P(A)P(B|A)P(C|B), \quad (2.4)$$

where the second equality follows from $P(A, C|B) = P(A|B)P(C|B)$ and the third inequality follows from Bayes' theorem which states that $P(A|B)P(B) = P(B|A)P(A)$.

Similarly for the same causal scenario of Figure 2.1, it can be shown that the Markov condition implies $P(A, C|B) = P(A|B)P(C|B)$, because,

$$P(A, C|B) = \frac{P(A, B, C)}{P(B)} = \frac{P(A)P(B|A)P(C|B)}{P(B)} = P(A|B)P(C|B), \quad (2.5)$$

where for the second equality we have utilised the Markov condition for the causal scenario of Figure 2.1 and for the third equality we have again used Bayes' theorem. Hence we have shown that for a simple chain causal scenario and when there are no latent variables then Markov condition is equivalent to the observed d*-separation relations.*

Now consider the causal scenarios of Figure 2.2 and interpret "Clouds" as A*, "Rain" as* B *and "Wet Floor" as* C*. The Markov condition for the causal scenario in Figure 2.2 (a) results in*

$$P(A,C|B) = \frac{P(A,B,C)}{P(B)} = \frac{P(A)P(B|A)P(C|B)}{P(B)} = \frac{P(A,B)P(C|B)}{P(B)} = \\ \frac{P(B|C)P(C)P(A,B)}{P(B)P(B)} = \frac{P(B|C)P(C)P(A|B)P(B)}{P(B)P(B)} = \frac{P(B|C)P(C)P(A|B)}{P(B)} \tag{2.6}$$

where we have resorted to Bayes' theorem in the fourth equality. But Equation 2.6 results in

$$P(A,C|B)P(B) = P(A,B,C) = P(C)P(B|C)P(A|B). \tag{2.7}$$

That is the Markov condition for the causal scenario of Figure 2.2 (a) implies the Markov condition for the causal scenario of Figure 2.2 (b). And similarly, it can be shown that the Markov condition for the causal scenario of Figure 2.2 (b) implies the Markov condition for the causal scenario of Figure 2.2 (a). Further, consider:

$$P(A,B,C) = P(A,C|B)P(B) = P(A|B)P(C|B)P(B) = P(A|B)P(B|C)P(C), \tag{2.8}$$

where we have used the conditional independence resulting from the d*-separation relation* $A \perp_d C|B$ *in the second equality and Bayes' theorem in the fourth equality. We have thus shown that the* d*-separation relation* $A \perp_d C|B$ *implies the Markov condition for the causal scenario of Figure 2.2 (b). In Equation 2.4, on interpreting the variables correctly, we had already showed that the* d*-separation relation* $A \perp_d C|B$ *implies the Markov condition for the causal scenario of Figure 2.2 (a). This is understandable because in the absence of any latent variables, the observed* d*-separation relations imply and are also implied by the Markov condition for any DAG. Hence, the Markov condition for the causal scenario of Figure 2.2 (a) needs to be necessarily equivalent to the the Markov condition for the causal scenario of Figure 2.2 (b) because both imply the same set of* d*-separation relations i.e.,* $A \perp_d C|B$*. This is the reason that the two causal scenarios of Figure 2.2 are indistinguishable under passive observations since they both imply the same constraints on the observed probability distributions. Our proof of Theorem 6 here was only for a simple chain causal scenario. For a full proof of it refer to Theorem 3 in [36].*

2.6 Latent-Permitting DAGs

As discussed in the introduction, sometimes we want to allow for causal explanations that include *latent variables*, i.e. variables that do not appear in the final probability distribution we are trying to explain. When our DAG of interest can have latent nodes, we call it a *latent-permitting DAG*, as opposed to the *latent-free DAGs* that only have observed nodes.

Let G be a DAG that has the set of nodes $\boldsymbol{A} = \boldsymbol{V} \cup \boldsymbol{L}$, $\boldsymbol{V}$ and $\boldsymbol{L}$ disjoint, where $\boldsymbol{V}$ are the observed nodes and $\boldsymbol{L}$ are the latent nodes. Mimicking the terminology used for the latent-free case, we will say that a probability distribution $P(\boldsymbol{V})$ over the observed variables is *classically compatible with* G if there exists some probability distribution $P(\boldsymbol{V}, \boldsymbol{L})$ over $\boldsymbol{V} \cup \boldsymbol{L}$ such that:

- $P(\boldsymbol{V}, \boldsymbol{L})$ is Markov with respect to G.
- The marginal of $P(\boldsymbol{V}, \boldsymbol{L})$ over $\boldsymbol{V}$ is the original probability distribution that we are interested in, i.e. $\sum_{\boldsymbol{L}} P(\boldsymbol{V}, \boldsymbol{L}) = \sum_{\boldsymbol{L}} P(\boldsymbol{V}|\boldsymbol{L})P(\boldsymbol{L}) = P(\boldsymbol{V})$.

As an example, a probability distribution P_{ABXY} over the random variables A, B, X and Y is classically compatible with the Bell DAG[3] (Figure 3.1) if it can be decomposed as:

$$P(ABXY) = \sum_{\Lambda} P(A|X,\Lambda)P(B|Y,\Lambda)P(X)P(Y)P(\Lambda) \tag{2.9}$$

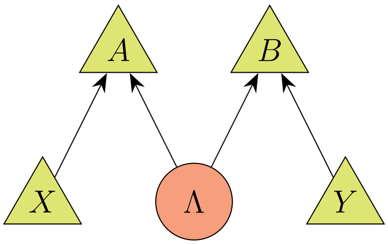


Figure 2.6: The Bell DAG. It encompasses the assumptions of Bell's theorem for a Bell Scenario where X and Y are the measurement settings of Alice and Bob, A and B are their outcomes and Λ is a classical hidden variable. The probability distributions that are classically compatible with this DAG are those that decompose as in Equation (2.9).

When dealing with latent-permitting DAGs, we will call the conditional independence relations that only involve observed variables the *observed conditional independence relations*, and similarly the d-separation relations that only involve observed nodes will be called *observed d-separation relations*. From the definition of classical compatibility with a latent-permitting DAG, we can see that the conclusions of Theorem 5 are also valid for latent-permitting DAGs:

[3] We will study the Bell DAG and other important DAGs in subsequent chapters in this theses.

Theorem 8 (Observed d-separation in latent-permitting DAGs [101])**.** *Let G be a DAG with nodes $\boldsymbol{A} = \boldsymbol{V} \cup \boldsymbol{L}$, where $\boldsymbol{V}$ are observed nodes and $\boldsymbol{L}$ are latent nodes. Let $\boldsymbol{X} \subseteq \boldsymbol{V}$, $\boldsymbol{Y} \subseteq \boldsymbol{V}$ and $\boldsymbol{Z} \subseteq \boldsymbol{V}$ be three disjoint sets of observed nodes of G. Then:*

1. *If G has the d-separation relation $\boldsymbol{X} \perp_d \boldsymbol{Y} | \boldsymbol{Z}$, then all of the probability distributions over the variables $\boldsymbol{V}$ which are classically compatible with G need to satisfy $\boldsymbol{X} \perp\!\!\!\perp_{CI} \boldsymbol{Y} | \boldsymbol{Z}$.*
2. *If G* does not have *the d-separation relation $\boldsymbol{X} \perp_d \boldsymbol{Y} | \boldsymbol{Z}$, then there exists some probability distribution over the variables $\boldsymbol{V}$ which is classically compatible with G and* does not *satisfy $\boldsymbol{X} \perp\!\!\!\perp_{CI} \boldsymbol{Y} | \boldsymbol{Z}$.*

As seen in Theorem 6, the only constraints that latent-free DAGs impose on the compatible probability distributions are the conditional independence relations, that can be obtained from d-separation. Consequently, if a DAG G has nodes $\boldsymbol{A} = \boldsymbol{V} \cup \boldsymbol{L}$, all the constraints that it imposes on the compatible joint probability distributions $P(\boldsymbol{A}) = P(\boldsymbol{V}, \boldsymbol{L})$ are the conditional independence relations. However, if we are interested only in distributions over the observed variables $\boldsymbol{V}$, sometimes the conditional independence constraints that involve the latent variables $\boldsymbol{L}$ might induce more complicated extra constraints on the distributions over the observed variables $\boldsymbol{V}$. If a probability distribution $P(\boldsymbol{V})$ over the observed variables satisfies both the observed conditional independence relations (obtained from d-separation) and the extra constraints derived from the conditional independence relations that involve the latent variables $\boldsymbol{L}$, then it is classically compatible with G.

Therefore, in principle one could find all the d-separation relations of a DAG (involving both observed and latent nodes), thus getting conditional independence constraints on $P(\boldsymbol{V}, \boldsymbol{L})$, and from there infer the constraints on the probability distribution over the observed variables, $P(\boldsymbol{V})$. This process will be exemplified with the Bell DAG in subsequent chapters. However, inferring constraints on $P(\boldsymbol{V})$ from the conditional independence relations of $P(\boldsymbol{V}, \boldsymbol{L})$, is generally very complicated.

2.7 Faithfulness and Minimality

Note that from Theorem 5 every d-separation relation implies a conditional independence on the classically compatible probability distributions, but *not all* conditional independence constraints presented by a specific classically compatible probability distribution P have to be associated with corresponding d-separation relations of the DAG. There might not be any d-separation relation corresponding to a conditional independence. This leads us to the following definition.

Definition 9 (Faithfulness or NO-FINE-TUNING)**.** *If the only conditional independences present in a probability distribution P, which is compatible with a DAG G, are those that follow from the*

d-separation relations in the DAG G, then we say that the DAG G explains it faithfully or with No-Fine-Tuning. On the other hand if the probability distribution has some conditional independence corresponding to which there is no d-separation relation in the DAG G,[4] *then we say that the DAG G is a Fine-Tuned explanation of P.*

In other words all the conditional independences present in the probability distribution should be a consequence of the causal scenario alone. There should not be any extra conditional independences that are not explained by the d-separation relations in the DAG for it to be a faithful explanation of the probability distribution. For example, suppose that there is some conditional independence $X \perp\!\!\!\perp_{\mathrm{CI}} Y|Z$ that is observed in a distribution P compatible with a DAG G that *does not* correspond to a d-separation relation $X \perp_{\mathrm{d}} Y|Z$ in G. If we adopt G as the causal explanation for P, this implies that there would be some correlations that for some conspiratorial reason do not appear in nature (because $X \perp\!\!\!\perp_{\mathrm{CI}} Y|Z)$, even if they are allowed by the underlying causal structure, G (as $X \perp_{\mathrm{d}} Y|Z$ does not hold in G). Then G could be considered as a bad causal explanation [108] of this probability distribution, and there could be other DAGs that explain all the conditional independences in this probability distribution through their corresponding d-separation relations.

As an explicit example, taken from [102, 93], consider the DAG in Figure 2.7. In the DAG, there is a direct arrow from A to C, hence $A \perp_{\mathrm{d}} C$ does not hold. Any probability distribution P is compatible with the DAG because by the definition of conditional probabilities any P can be written as $P(ABC) = P(A)P(B|A)P(C|AB)$. Therefore, the Markov condition is satisfied for any P and there are no d-separation relations in the DAG to impose any additional constraints on P. However, consider the distribution over all the nodes where A and C are independent, i.e $\sum_B P(ABC) = P(A)P(C)$. Such a distribution is compatible with the DAG because it satisfies the Markov constraint, since we have noted that any probability distribution over A, B and C satisfies the Markov condition with respect to the given DAG. Further such a distribution could arise because the indirect causal influence from A to C via B could correlate A and C negatively whereas the direct influence from A to C could correlate them equally positively thereby rendering C independent of A. Here the DAG is a Fine-Tuned explanation of such a probability distribution.

Further conditional independence-based algorithms follow one more assumption, which is minimality, and we explain it now. To understand the assumption of minimality, it is important to understand when a causal scenario A *can simulate* another causal scenario B.

Definition 10 (Simulate). *In the context of causal modelling, a causal scenario A is said to simulate another causal scenario B, both on the same set of variables V, if for any possible causal-statistical*

[4] Note that such a probability distribution will still satisfy the Markov condition because we are only considering distributions that are compatible with the DAG G.

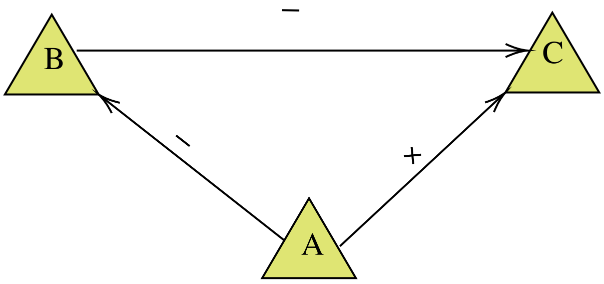


Figure 2.7: Direct causal influence from A to C can lead to positive correlation between the two nodes (shown with + over the causal arrow). Whereas indirect casual influence from A to C via B can lead to equivalent negative correlation between the two nodes (shown with − over the causal arrows). The net result could be that A and C are independent of each other even though $A \perp_{\mathrm{d}} C$ does not hold in the DAG. This is an example of FINE-TUNING.

parameters chosen for the latter scenario, there exists a corresponding set of causal-statistical parameters for the former scenario that produces the same probability distribution on the set of variables V.

Definition 11 (Minimality)**.** *If two causal scenarios A and B produce the same probability distribution on a set of observed variables V, and B can simulate A on V but A cannot simulate B on V, then A is considered to be a better causal explanation for the probability distribution over V than B.*

Choosing A over B even when B explains more might seem a bit counter-intuitive but the fact that A explains less than B makes A more falsifiable than B. So when choosing a causal scenario to explain the observed distribution we will choose the scenario which can be proven to be a wrong explanation more easily. This is in-fact what causal discovery algorithms do.

2.8 Classically Causal Explanation of correlations

Now we have all the mathematical definitions needed to understand what a classical causal explanation of any observed correlations is.

Definition 12 (Classically Causal Explanation of any correlations)**.** *Let P be any observed probability distribution over a set of visible nodes, V. Let there exist a DAG, G, over V visible nodes and L latent nodes. If P is both Markov and NOT-FINE-TUNED with respect to G, then we say that G is a* classical causal faithful explanation *of P or the observed correlations. If P is just Markov and not faithful with respect to G, then we say that G is a* classically causal unfaithful explanation *of P.*

Notice that given a DAG G the only difference between a “classically compatible” probability distribution P with respect to G and a probability distribution P' whose “classically causal faithful explanation” is the DAG G, is just that P could be FINE-TUNED with respect to G, whereas P' has to be NON-FINE-TUNED with respect to G. This is because of the “Faithfulness” assumption introduced in Section 2.7.

For example, consider the Bell DAG of Figure 3.1 and consider any probability distribution P' which obeys Equation 2.9 and thus all the Bell inequalities. Further, consider that all the observed conditional independences in P' are the ones that are implied by the observed d-separation relations in the Bell DAG. Then we say that the Bell DAG (of Figure 3.1) is a “classically causal faithful explanation” of such a probability distribution P'. On the contrary, consider a probability distribution P which obeys Equation 2.9 and (thus) the Bell inequalities but has more conditional independences amongst the observed nodes than those just implied by the observed d-separation relations in the Bell DAG. Then we say that the Bell DAG is a classically causal unfaithful explanation of P.

An example of such a probability distribution P which respects Equation 2.9 and the Bell inequalities but has more observed conditional independences than just those implied by the observed d-separation relations in the Bell DAG is just a probability distribution in which one or each party’s output is also independent of its respective setting. That is, a probability distribution $P(A, B, X, Y)$ that not only satisfies Equation 2.9 but also satisfies $\sum_{B,Y} P(A, B, X, Y) = P(A)P(X)$ is an example of a probability distribution that is CLASSICAL but FINE-TUNED with respect to the Bell DAG of Figure 3.1 and consequently the Bell DAG (of Figure 3.1) is a classically causal unfaithful explanation of such a $P(A, B, X, Y)$. In other words, for any distribution $P(A, B, X, Y)$ that is CLASSICAL with respect to the Bell DAG we have, $P(ABXY) = \sum_{\Lambda} P(A|X, \Lambda)P(B|Y, \Lambda)P(X)P(Y)P(\Lambda)$. If we now define $P(A|X, \Lambda) := P(A|\Lambda)$ (i.e., take $(A \perp\!\!\!\perp_{\text{CI}} X|\Lambda)$), we still maintain the Markov condition for the Bell DAG but together with it we also now satisfy $\sum_{B,Y} P(A, B, X, Y) = P(A)P(X)$, because $(A \perp\!\!\!\perp_{\text{CI}} X|\Lambda)$ and $(X \perp\!\!\!\perp_{\text{CI}} \Lambda)$ (from the Markov condition for the Bell DAG) together imply $(A \perp\!\!\!\perp_{\text{CI}} X)$ via the contraction semi-graphoid axiom [79] which can be easily proven. Hence, a probability distribution satisfying the Markov condition for the Bell DAG and also having $P(A|X, \Lambda) = P(A|\Lambda)$ also satisfies the extra conditional independence $\sum_{B,Y} P(A, B, X, Y) = P(A)P(X)$ which does not follow from the observed d-separation relations in the Bell DAG.

Another example of such a CLASSICAL and FINE-TUNED distribution with respect to the Bell DAG is a probability distribution (PA, B, X, Y), which respects Equation 2.9 but also exhibits the extra conditional independence that the outputs A and B are independent of each other, i.e., $\sum_{X,Y} P(A, B, X, Y) = P(A)P(B)$. Again the Bell DAG is a FINE-TUNED CLASSICAL explanation of such a probability distribution.[5]

We can also have a distribution that does not respect the Markov condition but nevertheless respects all the observed d-separation relations (later termed NON-CLASSICAL distributions with respect to the considered DAG) in the causal scenario, but yet is FINE-TUNED with respect to the causal scenario, in that it exhibits more observed conditional independences than just those that follow from observed d-separation relations in the causal scenario. As an example consider the Bell DAG and consider the probability distribution that violates the CHSH [18] inequality maximally. It is well known that beyond just respecting the no-signaling conditions and the independence of the settings of each other which are the only observed conditional independences that follow from the observed d-separation relations in the Bell DAG, this distribution also satisfies the conditions that the outputs of each party are independent of their respective settings. It will be seen that such a distribution is NON-CLASSICAL and FINE-TUNED with respect to the Bell DAG.

Also note that there might exist a probability distribution which cannot be explained by any causal scenario unless we introduce FINE-TUNING. For example consider a distribution over three variables A, B and C where A and C are independent and $B = A \oplus C$. Then B is independent of A and also of C but B is not independent of AC (i.e., B is not independent of A and C together or that B is dependent on the variable $X = (A, C)$). It is easy to see that no causal scenario can explain this probability distribution without the introduction of FINE-TUNING because $(B \not\perp\!\!\!\perp_{\mathrm{CI}} AC) \implies (B \not\perp_{\mathrm{d}} AC)$. Since $(B \perp_{\mathrm{d}} AC) \implies (B \perp_{\mathrm{d}} A)$ and $(B \perp_{\mathrm{d}} C)$, we have that $(B \not\perp_{\mathrm{d}} AC)$ implies $(B \not\perp_{\mathrm{d}} A)$ or $(B \not\perp_{\mathrm{d}} C)$ or both. In either case $(B \perp\!\!\!\perp_{\mathrm{CI}} A)$ or $(B \perp\!\!\!\perp_{\mathrm{CI}} C)$ or both can be satisfied only by the introduction of FINE-TUNING.

2.9 Non-Classical Causal Models

Now we can extend the framework of DAGs we have built till now to go beyond classical theories. This will serve to study more deeper theories of reality, like Quantum Mechanics and other Generalized Probabilistic theories [8].

Definition 13 (Generalized DAGs or GDAGs [44])**.** *Let G be a DAG with nodes $\boldsymbol{A} = \boldsymbol{V} \cup \boldsymbol{L}$, where $\boldsymbol{V}$ are observed nodes and $\boldsymbol{L}$ are latent nodes. All the observed nodes are associated with classical random variables. If the latent nodes are permitted to be more generalized sources than being just classical random variables, then we call the DAG G, a Generalized DAG (GDAG). More generalized*

[5] It is obvious that such a probability distribution is CLASSICAL and will never violate the Bell inequalities because to violate the Bell inequalities the outputs A and B need to be correlated.

sources could be Quantum sources or sources corresponding to other Generalized Probabilistic Theories [8].

Classical causal structures (DAGs) are an example of GDAGs where the latent nodes correspond to random variables as well. Another example of GDAGs are Quantum Causal Structures [4] where the latent nodes correspond to quantum states.

Quantum Causal Structures: There are several equivalent frameworks to study Quantum Causal structures [4, 9, 20, 44] but while providing a short description of them, we stick to the GDAG framework of [44]. In our presentation we follow [102]. The reader who is interested in specific details about Quantum Causal Structures is referred to [44].

Any unobserved (latent) node A without any incoming edges in a GDAG corresponds to a density matrix $\rho_A \in S(\mathcal{H}_A)$. Each directed edge from an unobserved (latent) node in the GDAG corresponds to a quantum system and thus to a Hilbert Space which can be labelled by the starting and the ending nodes of the edge [102]. Consider, for example, a latent node Λ that has two children A and B. Then the Hilbert space associated with Λ can be factorised as $\mathcal{H}_\Lambda = \mathcal{H}_{\Lambda_A} \otimes \mathcal{H}_{\Lambda_B}$, where subsystem $\mathcal{H}_{\Lambda_A}$ denotes the edge from Λ to A and subsystem $\mathcal{H}_{\Lambda_B}$ denotes the edge from Λ to B. If Λ is a classical node then the the two subsystems Λ_A and Λ_B can be taken to be just copies of Λ since classical information can be copied [102]. To any observed node A, there is associated a POVM, $\{E^A\}$ on the incoming quantum states. The measurement performed can depend on the classical output values of any observed parents of A. If A has only observed parents and no unobserved ones then its output can be seen as generated by the probability distribution on the incoming observed random variables. If A is parent-less then its output is just a classical probability distribution since it can be understood as a POVM on a trivial quantum state (i.e., a quantum state in a Hilbert space of dimension one). Consequently, if an unobserved node has an observed parent, then the density matrix associated with the unobserved node can be understood to depend on the classical output incoming from the observed parent. If however an unobserved node A has an unobserved parent B then we can understand that just as a quantum channel, a CPTP map Φ from B to A, i.e., $\Phi(\rho_B) \mapsto \rho_A$.

We have seen that the Markov condition provides a condition to the joint probability distribution over all nodes for compatibility with a DAG. The probability distribution over the observed nodes can be obtained by appropriate marginalization of the joint distribution over all the nodes that one gets by applying the Markov condition. However, in the case of GDAGs, where one can have more general and non-classical sources as well, a joint distribution over all the nodes (including the latent nodes) might not exist. For example if the latent variables correspond to quantum sources, then the quantum source does not co-exist with the measurement outcomes on it. Hence assigning a joint distribution over all the nodes which includes the quantum source as well as the measurement outcomes on it maybe impossible [102]. Hence, a more general compatibility condition is required when the latent variables are allowed to be more generalized sources. For more details about this compatibility condition, the reader is referred to [44].

As an example of a quantum causal structure, consider the Bell GDAG of Figure 2.8. One difference it has from the Bell DAG of Figure 3.1, is that the latent common cause, Λ corresponds to a quantum state ρ_Λ, such that $\rho_\Lambda \in \mathcal{S}(\mathcal{H}_\Lambda) = \mathcal{S}(\mathcal{H}_{\Lambda_A} \otimes \mathcal{H}_{\Lambda_B})$ and not a classical random variable. For the Bell DAG, any probability distribution, P, classically compatible with it needed to satisfy Equation 2.9. For the Bell GDAG, the compatibility condition is a bit different. The observed nodes A, B correspond to POVMs, $\{E^A_X\}$ and $\{E^B_Y\}$ that act on subsystems $\mathcal{H}_{\Lambda_A}$ and $\mathcal{H}_{\Lambda_B}$ and depend on the inputs X and Y respectively. Any probability distribution over the observed nodes, $P(ABXY)$, compatible with the Bell GDAG must satisfy the following factorization:

$$P(ABXY) = \mathrm{Tr}((E^A_X \otimes E^B_Y)\rho_\Lambda)P(X)P(Y) \tag{2.10}$$

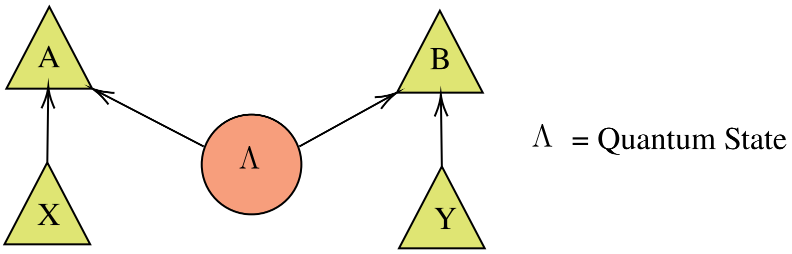


Figure 2.8: The Bell GDAG. Here the latent common cause, Λ denotes a quantum state

Any probability distribution $P(ABXY)$ that satisfies this factorization as in Equation 2.10 is termed as a *Quantum-Correlation*[6] compatible with the respective GDAG.

A few definitions and theorems due to HLP [44] are now in order.

Definition 14 (Generalized Markov condition [44, 102])**.** *Let G be a GDAG with nodes $\boldsymbol{A} = \boldsymbol{V} \cup \boldsymbol{L}$, where $\boldsymbol{V}$ are observed nodes and $\boldsymbol{L}$ are latent nodes. A probability distribution P is said to be Generalized Markov with respect to the GDAG, G, if there exists a casual-operational-probabilistic theory, a test corresponding to each node of the GDAG and a system corresponding to each edge of the GDAG which can generate the given probability distribution.*

[6] Throughout this thesis we assume that the only conditional independences exhibited by the observed QUANTUM-CORRELATIONS are the ones that follow from the observed d-separation relations of the corresponding GDAG G.

Theorem 15 (Generalized d-separation [44][7])**.** *Let X, Y and Z be disjoint sets of observed nodes in a GDAG G. Further let $W = G \setminus J^{-}(X \cup Y \cup Z)$. X and Y are d-separated by Z if and only if there exist sets of nodes U and V such that $\{U, V, Z, W\}$ is a partition of G, and*

$$\begin{gathered} X \subseteq U,\ Y \subseteq V \\ m(U) \cap m(V) \subseteq W \end{gathered} \tag{2.11}$$

where $m(U)$ denotes the union of the set of nodes U with all the children of each of the nodes in U (and similarly for $m(V)$) and $J^{-}(U)$ to be the union of U with the set of all ancestors of nodes in U (the entire 'past' of U), then $X \perp_d Y|Z$ in GDAG G.

This leads to the following d-separation theorem:

Theorem 16 (d-separation: general case [102, 44])**.** *Let G be a GDAG, with disjoint observed nodes X, Y, Z. For any probability distribution P that is a Quantum Correlation with respect to G, then $X \perp_d Y|Z \implies X \perp\!\!\!\perp_{CI} Y|Z$. Secondly if $X \perp\!\!\!\perp_{CI} Y|Z$ for all probability distributions that are generalized Markov with respect to G, then $X \perp_d Y|Z$ in G.*

2.10 Causal Discovery Algorithms

Causal discovery is the study responsible for generating possible causal models representing the properties corresponding to some observed data, usually conditional independences. Causal discovery algorithms are responsible to answer this problem. They return possible causal models corresponding to the particular observed conditional independences in the data. Causal discovery algorithms constitute a major active area of research in the fields of Machine Learning and Artificial Intelligence. There are several causal discovery algorithms like IC^*, (Inductive Causation in the presence of latent variables), IC (Inductive Causation in the absence of latent variables), PC (Peter-Clarke), CI (Causal Inference) [76, 95, 39, 80, 100]. However, in this thesis we use only the IC^* algorithm to study several conditional independences exhibited by certain causal scenarios.

In Section 2.5 we saw how a given causal scenario implies conditional independences. Causal discovery algorithms try to answer the inverse problem. Given some observed conditional independences) between observed variables, how can we find a causal scenario that explains them faithfully? Causal discovery in the presence of possible latent variables is significantly more difficult than in the absence of latent variables. IC^* [76, 39] algorithm (inductive causation in the presence of latent variables) which is equivalent to the Causal Inference algorithm (CI)[95, 59] takes as input the observed

[7] Taken from [44].

conditional independences and returns a (graph) pattern in which the random variables obey those conditional independences. It is one of the well-known algorithms to study causal discovery in the presence of latent variables. If there are causal structures which have pairwise common causes and which are faithful to the observed conditional independences then this algorithm returns a minimal set of these causal structures as a pattern.[8]

A pattern is a graph that encodes all the possible causal scenarios consistent with it. It consists of observed nodes and undirected, directed or bi-directed edges which represent the possible causal scenarios (DAGs) consistent with it. We do not review the algorithm here, but just explain how to interpret the pattern it returns subject to some observed conditional independences. We follow this up with a simple example where we show how the algorithm works. For a review of the IC^* algorithm we point the reader towards [76, 100].

The interpretation of the possible edges in the pattern returned by the algorithm in terms of a DAG is shown in Figures 2.9, 2.10 and 2.11. In all of them we show the pattern returned on the left and its interpretation in terms of a DAG on the right.

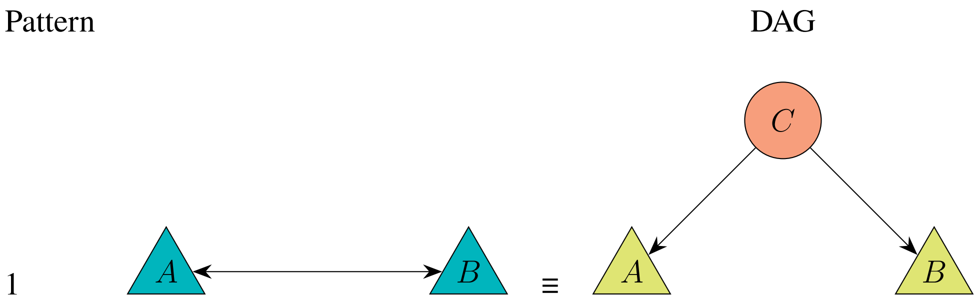


Figure 2.9: Interpretation of a bi-directed edge in the pattern as a DAG.

The process of selecting the causal scenarios which can be the possible explanations from the returned pattern is easily done by following the steps below:

1. Assume a particular causal order (that is choosing the order of the occurrence of the variables in time) for the variables.

2. Draw all the possible causal scenarios from the returned pattern which obey the observed conditional independences fed into the algorithm. These are the possible faithful explanations of the observed conditional independences.

[8] The algorithm only searches for pairwise common causes. So if in a scenario a common cause has three children, the algorithm would return that the three pairs of two children each have independent common causes.

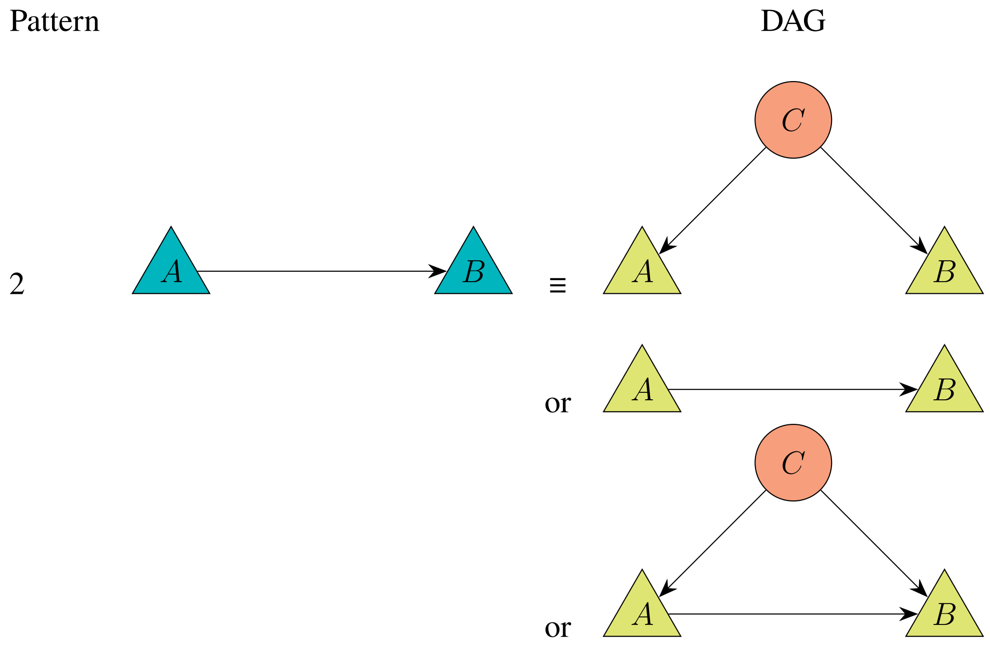


Figure 2.10: Interpretation of a directed edge in the pattern as a DAG.

The importance of point 1 can be seen through the examples of Figure 2.2(a,b). The conditional independences for both the scenarios in Fig 2.2 (a) and Fig 2.2 (b) are the same, precisely that there are no conditional independences for either of them. So if no conditional independences are fed into the IC^* algorithm then it will return both the causal scenarios of Figure 2.2 as possible explanations. It is evident that if we need to select one out of the two cases returned we need to enforce a causal order on the variables considered: Rain occurs after clouds.

Point 2 above says that we need to manually select only those causal scenarios from the returned pattern which respect the observed conditional independences. That means that the pattern returned could lead to scenarios which violate some observed conditional independences. This sort of defeats the purpose of the algorithm as we would expect the algorithm to itself only return those causal scenarios which are consistent with the observed conditional independences. Nevertheless, the algorithm and its python implementation [48] that we have used here returns a pattern from which we manually need to select the possible causal scenarios.

Now we give a simple example to show the working of the algorithm. Suppose we observe some data over three random variables A, B and C. Suppose that the only conditional independences that we observe are: $(AC \perp\!\!\!\perp_{\text{CI}} B)$. We feed them as inputs to the algorithm and the algorithm works out all the possible scenarios that can support these two conditional independences between the three

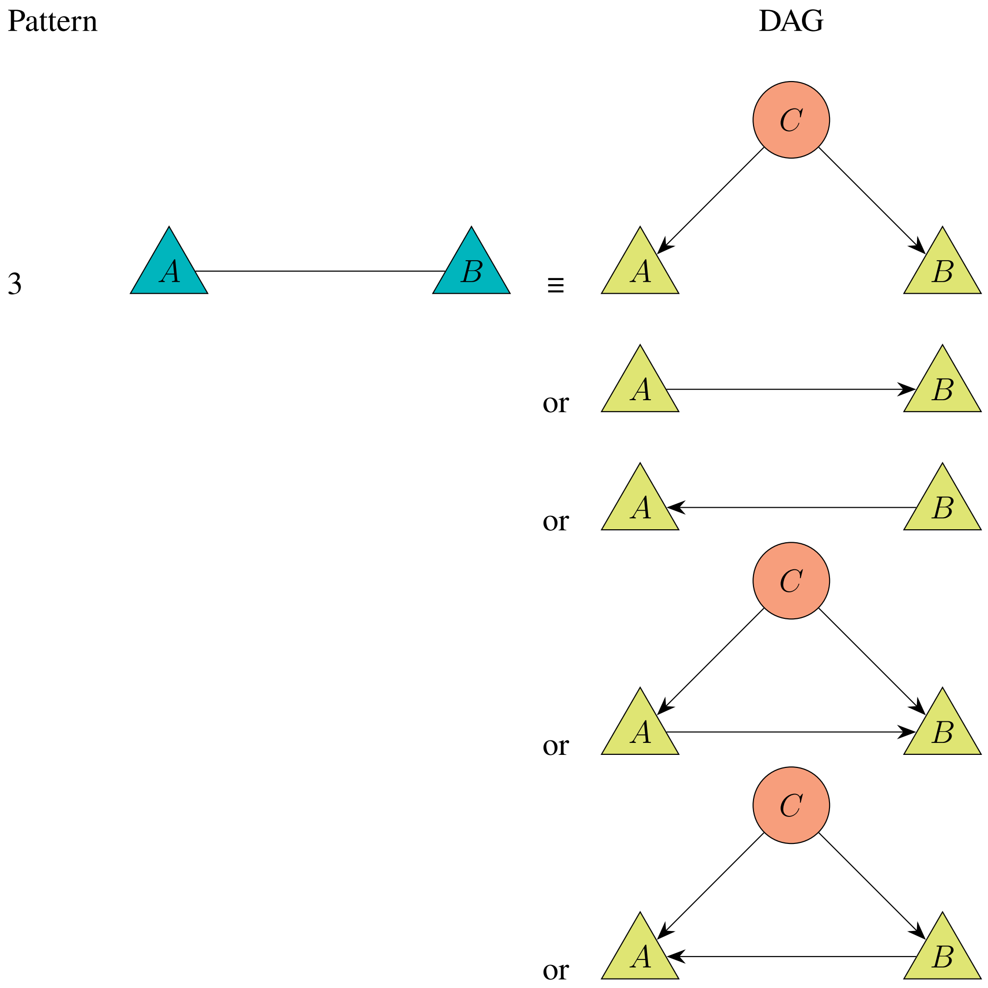


Figure 2.11: Interpretation of an undirected edge in the pattern as a DAG.

given random variables A, B and C. The pattern (or graph) that the algorithm returns is shown in Figure 2.12.

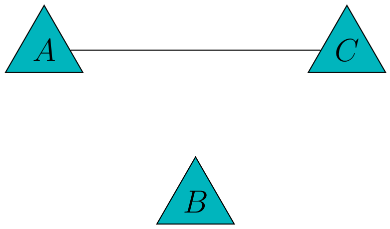


Figure 2.12: Pattern returned for conditional independences $(AC \perp\!\!\!\perp_{\text{CI}} B)$.

The undirected edge between nodes A and C needs to be interpreted according to the explanation in Figure 2.11. Thus the causal scenarios shown in Figure 2.13 are compatible with the returned pattern and with the input observed conditional independences: $(AC \perp\!\!\!\perp_{\text{CI}} B)$.

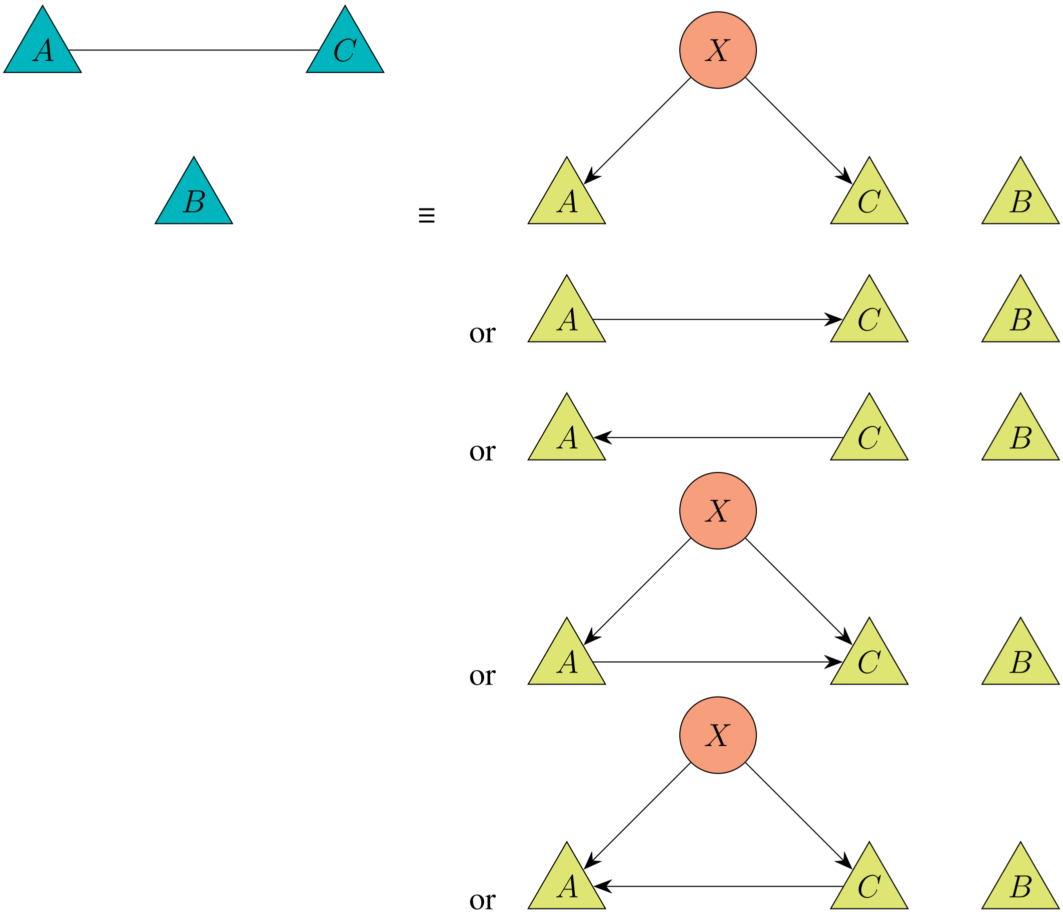


Figure 2.13: Causal scenarios compatible with conditional independences $(AC \perp\!\!\!\perp_{\text{CI}} B)$.

Unless we choose a specific causal order for the variables, we cannot narrow down on the choices of the causal scenarios compatible with the conditional independences $(AC \perp\!\!\!\perp_{\text{CI}} B)$. We may assume that A occurs before C. In that case the only compatible scenarios would be the the first, second and fourth ones in Figure 2.13. If on the other hand we assume that C occurs before A, then the compatible scenarios would be the the first, third and fifth ones in Figure 2.13. Note that in this simple example the observed conditional independences $(AC \perp\!\!\!\perp_{\text{CI}} B)$ are satisfied by all the scenarios resulting from the pattern. In a more complicated case this might not necessarily be true and we would need to consider only those scenarios resulting from the returned pattern which obey the given observed conditional independences. Wood and Spekkens [108] have explained this manual selection through nice pedagogical examples which we do not repeat here but just use the above procedure of manual selection on the causal scenarios we are interested in.

It is interesting to see what pattern the algorithm returns if we input the conditional independences corresponding to the Bell DAG of Figure 2.8. The generating set of conditional independences for the Bell DAG of Figure 2.8 are $(X \perp\!\!\!\perp_{\text{CI}} BY)$ and $(Y \perp\!\!\!\perp_{\text{CI}} XA)$. The pattern returned by the algorithm corresponding to these conditional independences is shown in Figure 2.14.

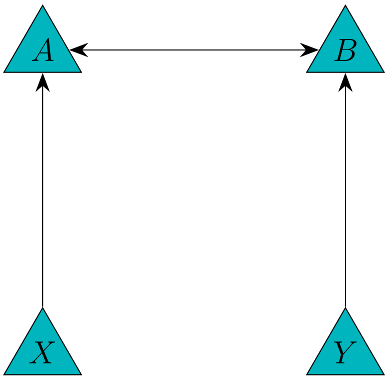


Figure 2.14: Pattern returned by the IC^* algorithm for conditional independences $X \perp\!\!\!\perp_{\text{CI}} BY$ and $Y \perp\!\!\!\perp_{\text{CI}} XA$.

Observation shows that the Bell DAG of Figure 2.8 is the only possible causal scenario compatible with the pattern returned in Figure 2.14 unless one resorts to fine-tuning. This was first pointed out by Wood and Spekkens [108]. But the Bell DAG cannot explain certain quantum correlations and hence is not an explanation of the set of all possible quantum correlations in this scenario.

It must not be forgotten that the causal discovery algorithms can go wrong because at best they are a heuristic explanation of the observed distribution. As Wood and Spekkens have shown this algorithm cannot distinguish between Bell inequality violating and non-violating correlations and hence cannot return a causal scenario which explains certain (quantum) correlations obtained in a Bell test faithfully. All explanations of Bell inequality violating correlations require fine tuning. This shortcoming of the algorithm is not a surprise because the algorithm only takes observed conditional independences as input and consequently does not know anything about the strength of possible correlations the respective conditional dependences can give rise to.

This brings us to an end of the mathematical preliminaries needed to understand the causality-based results in this theses. Now we are ready to dive into the specific results this thesis is concerned with.

— 3 —

Classifying Interesting Causal Scenarios

3.1 Introduction

As discussed in the last chapter, a DAG imposes causal compatibility constraints on the probability distributions that can be causally explained by it. An example of a DAG that imposes more complicated types of constraints on the compatible distributions over observed variables is the Bell DAG, presented in Figure 3.1. This DAG is of interest to physicists, as it encompasses the causal assumptions of Bell's theorem.

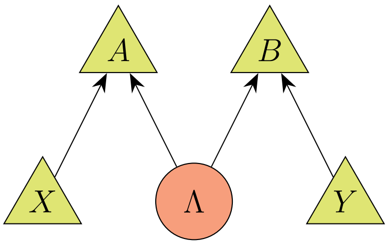


Figure 3.1: The Bell DAG. It encompasses the assumptions of Bell's theorem for a Bell Scenario where X and Y are the measurement settings of Alice and Bob, A and B are their outcomes and Λ is a classical hidden variable. The probability distributions that are classically compatible with this DAG are those that decompose as in Equation (2.9).

Bell's theorem [12, 10] is central to the foundations of quantum mechanics [72, 14, 23, 25], as it says that no locally causal theory in which the observers can choose their measurements independently of the source can ever be capable of reproducing all the operational predictions of quantum theory [73, 74, 40, 69, 70]. These assumptions are encoded in the Bell DAG, as there the settings X as Y are not causally connected to the source Λ, as well as the setting in one wing does not causally influence the outcome in the other wing.

As it turns out, the Bell DAG imposes causal compatibility constraints on the compatible distributions over $\{A, B, X, Y\}$ that take the form of inequalities, which are precisely the Bell inequalities. This means that all the distributions which violate Bell's inequalities, including some of the quantum predictions for this scenario, *cannot* be causally explained by the Bell DAG.

With the goal of causally explaining the violation of Bell's inequalities without changing the causal assumptions embedded in the structure of the Bell DAG, Henson, Lal and Pusey (HLP) [44] developed a generalization of Pearl's causal inference [76]. As explained in the last chapter, in this generalized framework, the latent nodes can be associated with quantum or other generalized probabilistic theory (GPT) systems. HLP also proved that the conditional independence constraints remain the same independently of the theory that describes the latent nodes; other types of constraints, like Bell's inequalities, can be violated.

The Bell DAG is just one of the many causal structures for which inequality constraints can be violated when the latent nodes are associated with nonclassical systems [44, 33, 32, 16, 88]. In the Bell scenario, correlations that violate Bell inequalities have cryptographic applications precisely because of their non-classicality, so it seems reasonable to hope that other scenarios allowing non-classical correlations will have similar applications. Finding which causal structures imply inequalities which potentially admit quantum violation is a critical step towards such potential applications.

In HLP, the concept of INTERESTINGNESS of a causal structure was defined. A causal structure is said to be NON-INTERESTING when all of its causal compatibility constraints are of the form of conditional independence relations, even in the classical case. Conversely, if the causal structure imposes more complicated constraints on the compatible probability distributions, it is said to be INTERESTING. As proven in HLP, only the INTERESTING scenarios are capable of witnessing a difference between the sets of classically and quantumly achievable probability distributions.

In HLP, a sufficient condition for NON-INTERESTINGNESS of a causal structure was developed; it is referred to here as the *HLP criterion*. A central motivation behind this work is that, at present, it is not known whether the HLP criterion is also necessary for NON-INTERESTINGNESS. That is, if a DAG cannot be proven NON-INTERESTING by virtue of the HLP criterion, is the DAG necessarily INTERESTING? The conjecture that the HLP criterion is indeed necessary for NON-INTERESTINGNESS will be referred to as the *HLP conjecture*.

How might we evaluate the HLP conjecture? To disprove it, we need to find only one DAG for which the HLP criterion does not apply, but which can be proven NON-INTERESTING by some other method. We did not pursue a search for such a counterexample, simply because we are unaware of any means to prove NON-INTERESTINGNESS when the HLP criterion does not apply. We therefore concentrate on providing evidence in support of the conjecture being true. Namely, we show that as one considers "larger and larger" DAGs, we can still certify the INTERESTINGNESS of (almost) all DAGs for which the HLP criterion does not apply.

To accomplish these goals we must clarify two preliminary questions. Firstly, how should we *enumerate* DAGs? One enumeration style — the original enumeration choice employed by HLP— is to count DAGs by their total number of nodes. We can thus consider DAGs with five total nodes, with 6 total nodes, with seven total nodes, etc. While this enumeration style has the advantage of simplicity, the arguably more natural enumeration which will be used here is to count by the total number of observed nodes. We can thus consider DAGs with two observed nodes, three observed nodes, four observed nodes, etc. From naive structural considerations alone, however, one might imagine that there are infinitely many DAGs with any fixed number of observed nodes. Motivated in-part by avoiding such infinities, when enumerating by the number of observed nodes we elect to work within Evans' framework of marginalized DAGs (mDAGs) [27], which will be explained in Section 3.3.

The second preliminary question is how to prove that a DAG for which the HLP criterion does not apply is indeed Interesting? Since we here seek to consider hundreds if not thousands of such DAGs, we are heavily invested in identifying *algorithmic* techniques for proving Interestingness, within which we deprioritize computationally expensive methods. In stark contrast to the approach of HLP, we extensively leverage a new result due to Evans [28], who has shown that a DAG G is Non-Interesting if and only if it is observationally equivalent to *some* DAG that does not have latent variables — where two DAGs are said to be (classically) observationally equivalent when their sets of (classically) compatible probability distributions are the same. As such, we herein primarily exploit necessary conditions for a graph to be observationally *in*equivalent to every latent-free graph.

One such condition is that when the DAG is *nonmaximal*, i.e. when it has a pair of nodes that are not adjacent (not connected by an arrow nor by a shared latent common cause) but are nevertheless not d-separated by any set of observed nodes, then the DAG is *not* observationally equivalent to any latent-free DAG. Another condition says that for two DAGs to be observationally equivalent, they must admit the same e-separation (introduced in Chapterr 4) relations over their observed nodes. A third condition says that for two DAGs to be observationally equivalent, they must admit the same set of compatible *supports*. Although the latter condition subsumes the two former ones (as discussed in Section 4.5), the former conditions can be evaluated more efficiently. The latter condition involving supports (remarkably!) can be assessed using Fraser's algorithm [31], which generally requires higher computational overhead. All of these conditions can be utilized to prove the Interestingness of a given DAG, via proving the classical observational inequivalence of the said DAG with *every* latent-free graph which has the same number of observed nodes.

It is worth contrasting the tools we employ here to certify Interestingness with those employed in prior literature by HLP and Pienaar [81].[1] HLP themselves attempted to explore all DAGs with 6 total nodes for which their sufficient condition for NON-INTERESTINGNESS did not apply. For all but five of these DAGs, HLP proved INTERESTINGNESS by means of deriving DAG-specific entropic inequalities and showing that those entropic inequalities could be violated by a DAG-specific distribution which nevertheless satisfied the conditional independence constraints of the DAG. One of the remaining five cases is the Bell DAG, which is long-since established as INTERESTING. Another one of the five is the so-called triangle scenario, whose INTERESTINGNESS was proven using a one-off proof technique which HLP did not generalize. The remaining three DAGs with 6 total nodes were eventually established as INTERESTING in a separate work by Pienaar [81], who employed a proof technique using fine-grained entropic inequalities.

At first glance, the algorithmic proofs of INTERESTINGNESS we employ here may seem unrelated to those utilized by HLP or Pienaar. However, our theorem relating e-separation to INTERESTINGNESS turns out to recover all but one of the INTERESTINGNESS results that HLP achieved by appealing to entropic inequalities. Additionally, our application of Fraser's algorithm further witnesses the INTERESTINGNESS of every other DAG conjectured by HLP to be INTERESTING, including the three of which were only *proven* INTERESTING in the later work of Pienaar [81].

We are therefore confident that our plethora of techniques likely supersede those earlier employed by HLP and Pienaar [81], despite not having a formal proof yet. We have made the Python code implementing the filters for INTERESTINGNESS proposed throughout this manuscript publicly available [46], though without accompanying documentation at this time.

As discussed, this work is of interest to quantum physicists because the INTERESTING DAGs are the possible candidates to explore quantum advantages in device independent information processing protocols [2, 99]. These DAGs are also the ones that should be looked into to compare quantum theory to more general probabilistic theories (GPTs) [8, 83].

On the other hand, our problem is also of central interest for purely classical causal inference. The set of probability distributions which are classically compatible with a NON-INTERESTING DAG is constrained only by conditional independence relations, which can be obtained from the d-separation relations of the DAG. Isolating a sufficient set of d-separation relations in a graph[2] is a well-studied problem. It is of paramount value to a classical data scientist, therefore, to know if the causal hypothesis encoded in a DAG may or may not be falsified by accounting for nontrivial inequality constraints.

[1] In doing so, we expose a critical error in one of the theorem's in Ref. [81] which we then rectify here (Happily, the handful of explicit DAGs which were declared as INTERESTING pursuant to the fallacious theorem in Ref. [81] are ultimately indeed INTERESTING nevertheless, and moreover their INTERESTINGNESS follows from our replacement nonmaximality theorem here). This discussion is made in Appendix A.2.

[2] A set of d-separation relations is said to be sufficient for a DAG if all other d-separation relations in the DAG can be inferred from the sufficient subset by application of semigraphoid axioms [79]

Such inequality constraints, when present, are often difficult to characterize explicitly .

3.2 What are Interesting Causal Scenarios

One of the goals of this work is to classify DAGs as Non-Interesting or not, following the concept introduced in HLP. Before establishing this concept in full generality, we will explore the example of the Bell DAG (Figure 3.1) and explain its Interestingness.

As discussed in Section 2.6, the d-separation criterion gives us all the classical compatibility constraints imposed on the probability distribution over *all the nodes*, $P(\boldsymbol{V},\boldsymbol{L})$. In the case of the Bell DAG, we have $P(\boldsymbol{V},\boldsymbol{L}) = P(ABXY\Lambda)$, while $P(\boldsymbol{V}) = P(ABXY)$.

The conditional independence relations of $P(ABXY\Lambda)$, that come from the d-separation relations of the Bell DAG, are:

$$P(\Lambda|XY) = P(\Lambda) \tag{3.1}$$

$$P(AB|XY\Lambda) = P(A|X\Lambda)P(B|Y\Lambda) \tag{3.2}$$

$$\begin{cases} P(A|XY) & = P(A|X) \\ P(B|XY) & = P(B|Y) \\ P(XY) & = P(X)P(Y) \end{cases} \tag{3.3}$$

These equations will give rise to the constraints imposed by the Bell DAG on $P(ABXY)$ through the marginalization of Λ. Equations (3.3), that do not involve Λ, are automatically transported as observed conditional independence constraints imposed by the Bell DAG on $P(ABXY)$. Equations (3.1) and (3.2), that do involve the latent variable Λ, will give rise to another type of constraint on $P(ABXY)$: the Bell's inequality. In fact, Equation (3.1) encodes the so-called no-superdeterminism assumption and Equation (3.2) encodes the local causality assumption which together are used to derive Bell's inequality.

Therefore, to study the Bell DAG, it is *not enough* to just look at the conditional independence constraints on the compatible distributions $P(ABXY)$ (Equations (3.3)). If one does that, they would miss important information that is encoded in Bell's inequality. This is so because the set of probability distributions $P(ABXY)$ that satisfy only the conditional independence relations of Equations (3.3) is *strictly larger* than the set of probability distributions that satisfies these conditional independence relations *and* Bell's inequality. In other words, Bell's inequality is not implied by Equations (3.3).

This is the core of the concept of Interestingness: an Interesting DAG imposes nontrivial inequality constraints on the classically compatible distributions. A "nontrivial" inequality constraint is an inequality that is not implied by the observed conditional independence relations of the DAG, along with nonnegativity of all probabilities and normalization.[3] Note that a NON-INTERESTING DAG can impose *trivial* inequality constraints on the compatible distributions: for example, if the node Λ in the Bell DAG was treated as an observed node, then the Bell inequalities would still be satisfied by the compatible distributions $P(ABXY\Lambda)$. However, in this case the Bell inequalities would be trivial, because they would be implied by *observed* conditional independence relations (Eqs. (3.1) and (3.2)).

To formalize the idea of INTERESTINGNESS, we will introduce a few definitions:

Definition 17. *Let G be a DAG. The sets $\mathcal{C}_G$, $\mathcal{Q}_G$, $\mathcal{G}_G$ and $\mathcal{I}_G$ of probability distributions over observed variables are defined as follows:*

1. *$\mathcal{C}_G$: Set of probability distributions that are classically compatible with G.*
2. *$\mathcal{Q}_G$: Set of all probability distributions that are "generalized Markov" for Quantum Theory with respect to G, as defined in HLP [44]. In other words this the set of all QUANTUM-CORRELATIONS compatible with G. For further details refer to HLP [44].*
3. *$\mathcal{G}_G$: Set of probability distributions that are "generalized Markov" for any operational theory with respect to G, as defined in HLP [44]. For further details refer to HLP [44].*
4. *$\mathcal{I}_G$: Set of probability distributions that satisfy all the conditional independence constraints that follow from the observed d-separation relations of G.*

For the case of the Bell DAG, $\mathcal{I}_{\text{Bell}}$ represents the set of distributions that obey the Equations (3.3). By contrast, $\mathcal{C}_{\text{Bell}}$ consists of a strict subset of $\mathcal{I}_{\text{Bell}}$, where we additionally restrict the conditional probabilities $P(AB|XY)$ to satisfy Bell's inequalities and thus lie in the local polytope. Set $\mathcal{Q}_{Bell}$ represents the set of all probability distributions that obey Equation 2.10.

Theorem 8 shows that $\mathcal{C}_G \subseteq \mathcal{I}_G$ for every DAG G. This is so because all the probability distributions that are classically compatible with G have to satisfy the conditional independence constraints that come from its observed d-separation relations. When the observed conditional independence relations are the only constraints imposed by a DAG on the compatible probability distributions over observed variables, the DAG is said to be NON-INTERESTING:

Definition 18 (INTERESTINGNESS). *Let G be a DAG. If $\mathcal{C}_G = \mathcal{I}_G$, then G is said to be NON-INTERESTING. Conversely, if $\mathcal{C}_G \subsetneq \mathcal{I}_G$, then G is said to be INTERESTING.*

[3] There is another type of equality constraint that a DAG can impose on the compatible distributions, apart from conditional independence relations: the "nested Markov constraints". In Ref. [28] it was proven that every DAG that presents nested Markov equality constraints also presents nontrivial inequalities. Consequently, INTERESTINGNESS may be equivalently defined relative to *all implied equality constraints* or relative to *all implied conditional independence relations*.

An Non-Interesting DAG corresponds to an algebraic set of classically compatible probability distributions: An algebraic set is defined by polynomial equalities (or more generally, by some finite union of sets each of which is defined by polynomial equalities). Semialgebraic sets, by contrast, are characterised by both polynomial equalities and polynomial inequalities and they correspond to an Interesting DAG's set of classicaly compatible probability distributions. To emphasise that a DAG's set of (classically) compatible distributions is defined by *more* than just the conditional independence (notably, equality) constraints, we therefore elect to speak of such a DAG as Interesting.

As proven by HLP, the observed conditional independence constraints imposed by a DAG on the compatible probability distributions do not change when the latent variables of the DAG are associated with quantum systems or other GPT systems. Therefore, if one is interested in studying causal structures that provide any quantum or GPT observational advantage, then there is only hope among the Interesting scenarios. If a DAG is Non-Interesting, then *all* the probability distributions that exhibit the conditional independence relations associated with its observed d-separation relations can be explained by this DAG *classically*.

Theorem 6 implies that every latent-free DAG is Non-Interesting. In HLP, a stronger sufficient condition for Non-interestingness is provided, together with an algorithmic strategy to check it. This condition, called the *HLP criterion*, will be discussed in Section 3.4. It is still not known whether the HLP criterion is also necessary for Non-interestingness; its possible outcomes for a given DAG are either that it is Non-Interesting or that it is "unresolved". The unresolved DAGs thus need to be assessed by some other method.

Based on the HLP criterion and certain types of entropic inequalities, HLP and Pienaar [81] classified the Non-interestingness or Interestingness of all DAGs of up to 6 *total* nodes (observed and latent), thus leaving no DAGs of 6 total nodes with inconclusive status. In this work, however, we elect to count DAGs not by their *total* node count but rather by their number of *observed* nodes. It turns out that HLP's complete classification of DAGs with *total* node count up to 6 meant that they resolved *all* DAGs with 3 observed nodes, a *few* DAGs with 4 observed nodes, and *no* DAGs with 5 or more observed nodes. Here we attempt to tackle the Non-interestingness classification of all causal structures with up to 4 observed nodes.[4] To do so, we will utilise the *mDAG formalism* introduced by Evans in Ref. [27].

[4] Note that the set of all DAGs with 4 observed nodes *includes* the set of all DAGs with 7 total nodes which persist under HLP's reduction techniques. Thus, this work can *also* be considered an extension of HLP's classification from up-to-6 to up-to-7 total nodes. As noted in Section 4.6, we ultimately resolve the Non-interestingness or Interestingness of *all but three* DAGs of 4 observed nodes. Only *one* of those 3 has 7 total nodes. Thus, we ultimately resolve *all but one* DAG with 7 total nodes.

3.3 Simplifying the problem by using mDAG formalism

Two DAGs G and H are said to be *classically observationally equivalent* when $\mathcal{C}_G = \mathcal{C}_H$. Note that $\mathcal{C}_G = \mathcal{C}_H$ implies that $\mathcal{I}_G = \mathcal{I}_H$: by Theorem 8, if a certain d-separation relation is *not* present in a DAG, it is always possible to find a probability distribution that violates the conditional independence corresponding to this d-separation relation and is classically compatible with the DAG. In other words, if two DAGs are classically compatible with the same sets of distributions, they have to present the same d-separation relations.[5]

In short,

$$\mathcal{C}_G = \mathcal{C}_H \implies \mathcal{I}_G = \mathcal{I}_H, \tag{3.4}$$

with contrapositive

$$\mathcal{I}_G \neq \mathcal{I}_H \implies \mathcal{C}_G \neq \mathcal{C}_H. \tag{3.5}$$

In particular, this means that if a DAG G is NON-INTERESTING (INTERESTING), then all of the DAGs H that are compatible with it are also NON-INTERESTING (INTERESTING).

In this section, we will present two results of [27] that prove classical observational equivalence, thus reducing the number of DAGs that have to be examined for NON-INTERESTINGNESS. After presenting the two results, we will show that they allow for a definition of a new object, called mDAG, that encompasses this simplification.

To do so, we start with the definition of *exogenization*. It might be easier to understand this definition by following Figure 6.4, where DAG 6.4(b) is obtained from DAG 6.4(a) by exogenizing node B.

Definition 19 (Exogenized DAG). *Let G be a DAG and let λ be a latent variable of G. We define the exogenized DAG $\mathcal{E}(G, \lambda)$ as follows: take the vertices and edges of G and (1) add an edge $m \to n$ from every $m \in PA_G(\lambda)$ to every $n \in CH_G(\lambda)$ and (2) delete edges $m \to \lambda$ for every $l \in PA_G(\lambda)$. All other edges remain the same.*

With this definition at hand, we state the Lemma 3.7 of [27]:

Lemma 20 (Exogenization)**.** *Let G be DAG with observed nodes V and latent nodes L, and let $\lambda \in L$ be a latent node of G. Furthermore, let $\tilde{G} = \mathcal{E}(G, \lambda)$. Then, $\mathcal{C}_G = \mathcal{C}_{\tilde{G}}$.*

Now, we state the Lemma 3.8 of [27]. This Lemma is also illustrated in Figure 6.4, where it is applied to go from 6.4(b) to 6.4(c).

[5] On the other hand, if DAGs G and H are such that $\mathcal{I}_G = \mathcal{I}_H$, this does *not* imply that $\mathcal{C}_G = \mathcal{C}_H$.

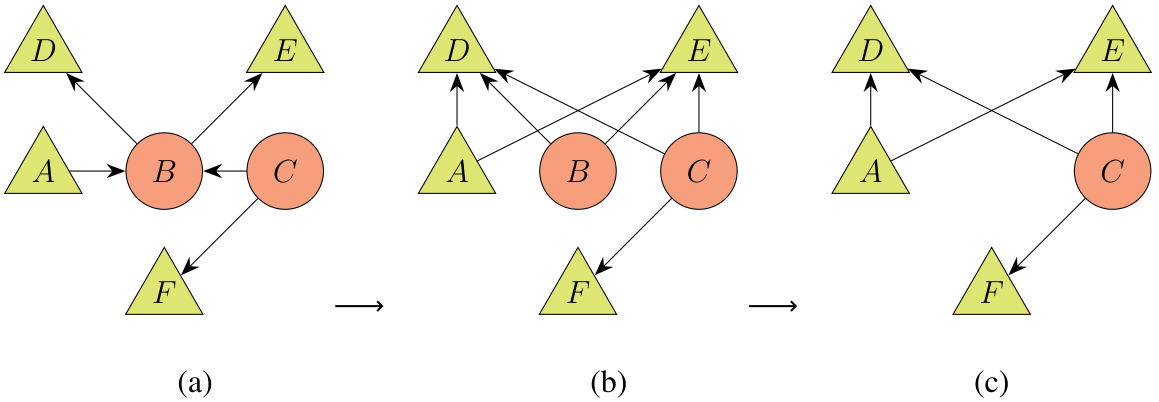


Figure 3.2: The DAGs (a) and (b) are classically observationally equivalent to the mDAG of (c). The step (a)→(b) can be shown by Lemma 20, and the step (b)→(c) can be shown by Lemma 21.

Lemma 21 (No redundant latents)**.** *Let G be a DAG with observed nodes V and latent nodes L. Let $\lambda \in L$ and $\mu \in L$ be latent nodes of G such that $\lambda \neq \mu$, $PA_G(\lambda) = PA_G(\mu) = \varnothing$ and $CH_G(\lambda) \subseteq CH_G(\mu)$. In this case, we say that the node λ is "redundant". Let G' be the DAG obtained after deleting the node λ. Then, $\mathcal{C}_G = \mathcal{C}_{G'}$.*

Like Lemma 20, Lemma 21 also states conditions under which proving the INTERESTINGNESS of one DAG automatically gives you the INTERESTINGNESS of another.

For example, Lemmas 20 and 21 show that all the three DAGs of Figure 6.4 have the same sets $\mathcal{C}$ and $\mathcal{I}$. Thus, we need only to examine one DAG out of the these. It makes sense to pick the DAG of Figure 6.4(c), as it is the simplest.

Following this same idea, we can work with the concept of an mDAG, first defined in Ref. [27]. The definition below is different than, but equivalent to the one presented in Ref. [27].

Definition 22 (mDAG). *An mDAG is a DAG where none of the latent nodes is redundant (as defined in Lemma 21) nor has any parents.*

For example, the DAG of Figure 6.4(c) is an mDAG. For a fixed number of observed nodes, there is a finite number of mDAGs. In particular, from our computations [47, 46] we know that for 3 observed nodes there are 46 mDAGs, while for 4 observed nodes there are 2809 mDAGs. Lemmas 20 and 21 show that the mDAG encodes all the necessary information of the DAG if you only want to talk about the sets $\mathcal{C}$ and $\mathcal{I}$.

The Lemmas here presented also provide an argument in favour of counting the causal structures in terms of the number of observed nodes instead of in terms of the total number of nodes; Lemma 21 shows that DAGs 6.4(a) and 6.4(b), that have 6 total nodes, actually do not have to be analyzed; their INTERESTINGNESS can be examined by looking at 6.4(c), that has 5 total nodes.

3.4 The HLP criterion

In this section, we describe the sufficient criterion for NON-INTERESTINGNESS that was developed in HLP. The criterion describes some transformations that take a DAG G to another DAG H such that $\mathcal{C}_H \subseteq \mathcal{C}_G$ while $\mathcal{I}_H = \mathcal{I}_G$. If the final DAG H is known to be NON-INTERESTING (for example, by being latent-free), then the original DAG G is also NON-INTERESTING.

These transformations, adjusted to the language of mDAGs, are presented in Theorem 23.

Theorem 23. *Let G and H be two mDAGs. Suppose that H can be obtained by starting from G and applying one or more of the following transformations:*

1. *Removal of an edge.*
2. *Addition of an edge $X \to Y$, where previously $PA(X) \subseteq PA(Y)$ and $PA(X)$ contained at least one latent node.*

Then, $\mathcal{C}_H \subseteq \mathcal{C}_G$.

Now we state the *HLP criterion* as a corollary:

Corollary 24 (HLP Criterion)**.** *Let G be an mDAG. Suppose that by a sequence of the transformations defined in Theorem 23 it is possible to start from G and reach another mDAG H such that:*

1. *H does not have latent nodes.*
2. *The set of observed d-separation relations of H and G is the same, i.e. $\mathcal{I}_H = \mathcal{I}_G$.*

Then, the original mDAG G is NON-INTERESTING.

Proof. From Theorem 23, $\mathcal{C}_H \subseteq \mathcal{C}_G$. Since H is latent-free, by Theorem 6 it is NON-INTERESTING, i.e. $\mathcal{C}_H = \mathcal{I}_H$. Therefore, we have $\mathcal{I}_G = \mathcal{I}_H = \mathcal{C}_H \subseteq \mathcal{C}_G$. As noted before, $\mathcal{C} \subseteq \mathcal{I}$ is valid for any DAG, due to Theorem 8. Therefore, $\mathcal{C}_G = \mathcal{I}_G$, meaning that G is NON-INTERESTING. □

Figure 3.3(a) exemplifies an mDAG that can be shown NON-INTERESTING by the HLP Criterion: by a sequence of transformations defined in Theorem 23 we can obtain the mDAG shown in Figure 3.3(c), that obeys the conditions presented in Corollary 24.

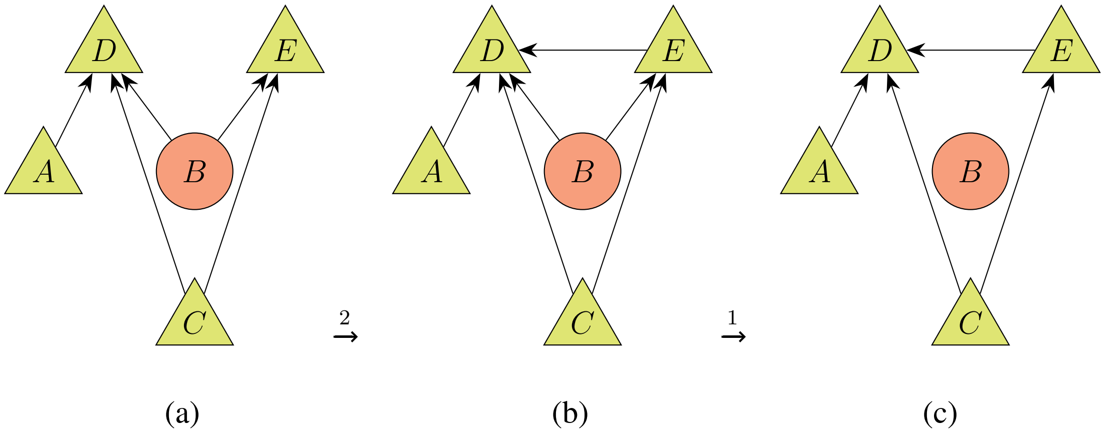


Figure 3.3: An example of the application of the HLP criterion. All of these three mDAGs have the same set of d-separation relations: $A \perp_{\mathrm{d}} E$, $A \perp_{\mathrm{d}} E|C$,$A \perp_{\mathrm{d}} C$ and $A \perp_{\mathrm{d}} C|E$. Since (c) is latent-free, we can conclude that (a) is NON-INTERESTING.

The natural question to ask here is whether the HLP criterion is also necessary for an mDAG to be NON-INTERESTING. The conjecture that the HLP criterion is also necessary will be called the *HLP conjecture*:

Conjecture 25 (HLP Conjecture)**.** *Let G be an NON-INTERESTING mDAG. Then, by a sequence of transformations defined in Theorem 23, it is possible to start from G and reach another mDAG H such that:*

1. *H does not have latent nodes.*
2. *The set of observed d-separation relations of H and G is the same, i.e. $\mathcal{I}_H = \mathcal{I}_G$.*

As we will see, our current results do not prove this conjecture, but give hints towards its validity. In Ref. [28], Evans has shown that:

Theorem 26. *A latent-permitting DAG G is NON-INTERESTING if and only if it is classically observationally equivalent to a latent-free DAG H.*

This result allows us to restate the HLP conjecture in a different manner:

Conjecture 27 (Reformulation of the HLP conjecture)**.** *Let G be an mDAG. G is classically observationally equivalent to a latent-free mDAG H if and only if:*

1. *The set of observed d-separation relations of H and G is the same, i.e. $\mathcal{I}_H = \mathcal{I}_G$.*
2. *Through the transformations defined in Theorem 23, it is possible to go from G to some latent-free mDAG H' which has the same set of observed d-separation relations as G and H, i.e. $\mathcal{I}_{H'} = \mathcal{I}_H = \mathcal{I}_G$.*

Therefore, proving the HLP conjecture would also be of relevance to the problem of classifying causal structures into classical observational equivalence classes.

— 4 —

Methods to determine Interestingness

In this chapter we discuss the methods we used to prove the Interestingness of a large number of DAGs that do not respect the HLP criterion.

4.1 Using nonmaximality to prove Interestingness

The first method to show Non-interestingness that we will present relies on the concept of *maximality* [91, 28]. To define this, we will first define what *adjacent* and d*-separable* pairs of nodes are.

Definition 28 (**Adjacency**). *Let G be an mDAG, and let A and B be a pair of observed nodes of G. We say that A and B are* adjacent *in G if A is a parent of B, or B is a parent of A, or A and B share a common latent parent.* [1]

Definition 29 (**d-(un)separable pair of observed nodes**). *Let G be a DAG with nodes $\boldsymbol{A} = \boldsymbol{V} \cup \boldsymbol{L}$, where $\boldsymbol{V}$ are observed nodes and $\boldsymbol{L}$ are latent nodes. A pair of observed nodes $\{A, B\}$ for $A \in \boldsymbol{V}$, $B \in \boldsymbol{V}$ is said to be d*-separable *if there is some subset $\boldsymbol{Z} \subseteq \boldsymbol{V}$ of the remaining observed nodes such that $(A \perp_d B|\boldsymbol{Z})$, otherwise, A and B are said to be d-unseparable.*

These two definitions are related, in that they are criteria which relate to the compatibility of a particular distribution, namely, perfect correlation between one pair of nodes while all other nodes are point distributed. Consider the following distribution:

$$P(A, B, C, D...) = \frac{1}{2}(\delta_{A,0}\delta_{B,0} + \delta_{A,1}\delta_{B,1})\delta_{C,0}\delta_{D,0}... \tag{4.1}$$

[1] For DAGs which are not exogenized (see Definition 19) we additionally say that A and B are adjacent if there is some *latent treck* from on to the other. Which is to say, A and B are said to be *adjacent* in a non-exogenized DAG G iff they are adjacent in the mDAG resulting from exogenizing G per Definition 19.

Eq. (4.1) describes a probability distribution in which A and B are random but perfectly correlated, and every other observed node is point-distributed (i.e., one value of the node occurs with probability one and all the other values of the node have probability zero) at the value 0. Then,

Proposition 30 (Adjacency $\Leftrightarrow P_{(4.1)} \in \mathcal{C}_G$)**.** *The distribution in Eq.* (4.1) *is classically compatible with the graph G if and only if A and B are adjacent nodes in G.*

Proof. See Refs. [27, 26]. Clearly $P_{(4.1)} \in \mathcal{C}_G$ when A and B *are* adjacent. Whenever A and B are *not* adjacent, then evidently A and B are e-separated[2] upon removal of $\boldsymbol{V} \setminus \{A, B\}$, which also means that $P_{(4.1)}$ would violate the entropic inequality in Theorem 5 of Ref. [29]. □

Proposition 31 (d-unseparability $\Leftrightarrow P_{(4.1)} \in \mathcal{I}_G$)**.** *The distribution in Eq.* (4.1) *satisfies all the conditional independence constraints that follow from the observed d-separation relations relations of graph G if and only if A and B are d-unseparable in G.*

Proof. The only conditional independence relations that the distribution in Eq. (4.1) *fails* to satisfy are those of the form $A \perp\!\!\!\perp_{\text{CI}} B|\boldsymbol{Z}$. Such conditional independence relations follow from the observed d-separation relations relations of graph G if and only if A and B are d-separable in G. □

Putting Propositions 30 and 31 together, we find that $P_{(4.1)} \in \mathcal{I}_G$ but $P_{(4.1)} \notin \mathcal{C}_G$ whenever a graph has A and B nonadjacent but also d-unseparable. This leads us to the concept of *maximality* [91, 28]:

Definition 32 (Maximal DAG). *Let G be a DAG. If all of the pairs of nodes of G which are* not *d-separable are also adjacent then G is said to be* maximal*, otherwise G is said to be* nonmaximal.

Figure 4.1 shows an example of a maximal DAG (4.1a) and an example of a nonmaximal DAG (4.1b).

Theorem 33 (Nonmaximal). *Every NON-INTERESTING DAG is maximal, that is, every nonmaximal DAG is INTERESTING.*

Proof. As we have seen from Propositions 30 and 31, if a graph G is nonmaximal, then there is some pair of observed nodes such that the distribution given by "those two nodes are random and perfectly correlated while all other observed nodes are point distributed" lies in some *gap* between $\mathcal{I}_G$ and $\mathcal{C}_G$. Alternatively, note that nonmaximality implies INTERESTINGNESS also follows from the fact that all latent-free graphs are maximal [91, Prop. 3.19]. We then simply note that a nonmaximal graph

[2] The concept of e-separation will be explained upon in Section 4.4 here.

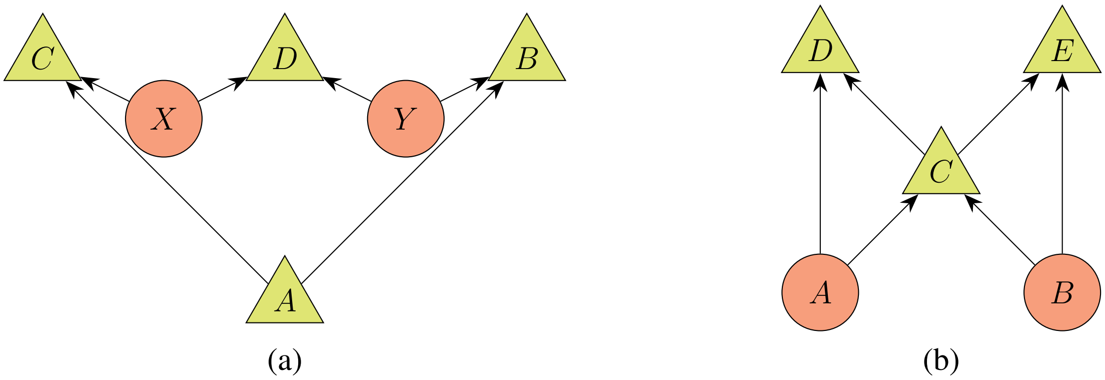


Figure 4.1: (a) An example of maximal DAG, which so happens to be INTERESTING. (b) An example of an nonmaximal DAG, namely, The Unrelated Confounders scenario introduced in Ref. [26]. It is not maximal because nodes D and E are not d-separable, but are nevertheless not adjacent. Like *all* nonmaximal DAGs, the Unrelated Confounders scenario is INTERESTING.

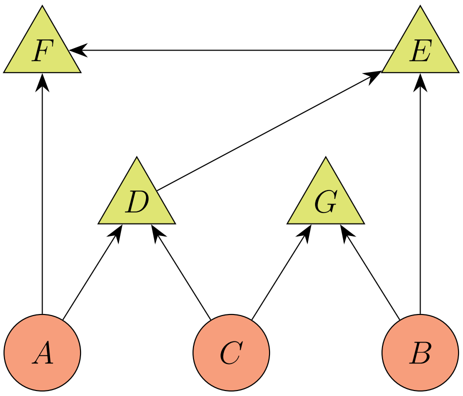


Figure 4.2: A DAG with 4 observed nodes that is shown to be INTERESTING per Theorem 33. This can be seen because G and F are not d-separable, i.e. none of the d-separation relations $(G \perp_{\mathrm{d}} F|E)$, $(G \perp_{\mathrm{d}} F|D)$ or $(G \perp_{\mathrm{d}} F|D, E)$ hold, but they are not adjacent.

cannot be observationally equivalent to any maximal graph: It follows from Eq. 3.5 that every pair of observationally equivalent DAGs must agree on their sets of d-(un)separable observed-node pairs. Furthermore, pursuant to Lemma 76 (discussed in Appendix A.3), agreement with respect to adjacency structure is *also* a prerequisite for observational equivalence [27, 26]. These facts imply that if a DAG G is nonmaximal, then it is *not* going to be observationally equivalent to any latent-free DAG (and will thus be INTERESTING by Theorem 26). □

Figure A.5 shows an example of a 4-observed-nodes mDAG whose INTERESTINGNESS may be shown by Theorem 33. The reasoning is that G and F are not d-separable, i.e. none of the d-separation relations $(G \perp_{\mathrm{d}} F|E)$, $(G \perp_{\mathrm{d}} F|D)$ or $(G \perp_{\mathrm{d}} F|D, E)$ hold, but they are not adjacent.

To find out which sets of pairs of observed nodes are d-separable in G and then check whether it is maximal, is it necessary to first obtain *all* the d-separation relations of G? The following accessory lemma shows that this is not the case, thus simplifying the application of Theorem 33 in practice.

Lemma 34 (Rapid test for d-separability of pairs)**.** *A pair of node subsets $\boldsymbol{A}$ and $\boldsymbol{B}$ of a DAG G are d-separable by* some *set of observed nodes if and only if they are d-separable by a* particular *set of observed nodes, namely, the set of all and only those observed nodes which are ancestors of either $\boldsymbol{A}$ or of $\boldsymbol{B}$. In other words, $\boldsymbol{A} \perp_d \boldsymbol{B} \mid \mathrm{ANC_G}(\boldsymbol{A}) \cup \mathrm{ANC_G}(\boldsymbol{B})$ whenever there exists some set $\boldsymbol{Z}$ such that $\boldsymbol{A} \perp_d \boldsymbol{B}|\boldsymbol{Z}$.*

Proof. This follows from Theorem 6 of Ref. [98]. In particular, it follows from the use of Ref. [98]'s Algorithm 3 to solve Ref. [98]'s Problem 4. □

In their Appendix E, HLP effectively utilized Shannon-type inequalities to certify the INTERESTINGNESS of many DAGs with 5 or 6 total nodes. In particular, they found that the Shannon-type inequalities for those DAGs could by violated by distributions that satisfy all of the conditional independence relations imposed by the DAG: those distributions were such that two particular variables are perfectly correlated and all other observable variables are point distributed, identical in nature to the construction of Eq. (4.1). From this, note that Proposition 30 implies that the nodes corresponding to these perfectly correlated variables are *not* adjacent, and Proposition 31 implies that they are not d-separable. Thus, our Theorem 33 also certifies the INTERESTINGNESS of all those examples. Conversely, *every* pair of nodes which are not d-separable but also not adjacent generate a distribution from Eq. (4.1) that violates a Shannon-type inequality; this is a consequence Theorem 5 of Ref. [29], that says that every e-separation relation (which will be defined in Sec. 4.4) implies on a Shannon-type inequality. This result was already used in the proof of Proposition 30.

4.2 Using setwise nonmaximality to prove INTERESTINGNESS

In the previous subsection we found that that membership in $\mathcal{I}_G$ or $\mathcal{C}_G$ of the distribution in which *pair* of observed nodes is perfectly correlated while all other observed nodes are point distributed is directly related to the concepts of d-unseparability and adjacency, respectively. Here we formulate analogous criteria in order to assess perfect correlation of three-or-more variables when all other variables in the distribution are point-distributed.

Definition 35 (Setwise adjacency). *Let G be a DAG with nodes $\boldsymbol{A} = \boldsymbol{V} \cup \boldsymbol{L}$, where $\boldsymbol{V}$ are observed nodes and $\boldsymbol{L}$ are latent nodes. Then, the subset of observed nodes $\{V_1 ... V_k\}$ is setwise adjacent in G if* *and only if there is some node X (possibly but not necessarily within $\{V_1...V_k\}$) such that X is an ancestor of* every *node in $\{V_1...V_k\}$ not only in the DAG G but also in the* subgraph *of G formed by deleting all nodes $\boldsymbol{V} \setminus \{V_1...V_k\}$ from G.*[3]

Definition 36 (Setwise d-unrestriction). *Let $\boldsymbol{V}$ be the set of all observed nodes in some DAG G, and let $\boldsymbol{S}$ be some subset of $\boldsymbol{V}$. Then, the nodes $\boldsymbol{S}$ are setwise d-unrestricted in G if and only if there does* not *exist any pair of nodes $\{S_i, S_j\} \subset \boldsymbol{S}$ along with some (possibly empty) set of observed nodes $\boldsymbol{Z} \subset \boldsymbol{V} \setminus \boldsymbol{S}$ such that $(S_i \perp_d S_j | \boldsymbol{Z})$.*

We next show that these definitions for setwise adjacency and setwise d-unrestriction have the desired properties. Consider the following distribution:

$$P(\boldsymbol{V}) = \frac{1}{2}(\delta_{V_1,0}\delta_{V_2,0}...\delta_{V_k,0} + \delta_{V_1,1}\delta_{V_2,1}...\delta_{V_k,1}) \\ \delta_{V_{k+1},0}\delta_{V_{k+2},0}... \tag{4.2}$$

Eq. (4.2) describes a probability distribution in which the first k observed variables are random but perfectly correlated while all other observed variables are point-distributed at the value 0. Then,

Proposition 37 (Setwise Adjacency $\Leftrightarrow$ $P_{(4.2)} \in \mathcal{C}_G$)**.** *The distribution in Eq.* (4.2) *is classically compatible with graph G if and only if $\{V_1...V_k\}$ are setwise adjacent in G.*

Proposition 37 follows from the two accessory lemmas below.

Lemma 38 (Partial point distribution $\implies$ subgraph compatibility)**.** *Suppose $P(\boldsymbol{V})$ is some distribution wherein the variables $\boldsymbol{V} \setminus \{V_1...V_k\}$ are point-distributed and moreover all the variables in $\{V_1...V_k\}$ have finite cardinality. Then, $P(\boldsymbol{V})$ is classically compatible with G if and only if $P(\{V_1...V_k\})$ is compatible with the* subgraph *of G formed be deleting all nodes $\boldsymbol{V} \setminus \{V_1...V_k\}$ from G.*

Proof. This lemma is an immediate consequence of the e-separation theorem central in Ref. [26]. □

Lemma 39 (setwise correlation $\implies$ common ancestor)**.** *Let $P_{\textit{perfect correlation}}(\boldsymbol{A})$ be the distribution in which all variables in $\boldsymbol{A}$ are random but perfectly correlated with each other. Then, $P_{\textit{perfect correlation}}(\boldsymbol{A})$ is compatible with a DAG G if and only if all the nodes in $\boldsymbol{A}$ share some common ancestor in G.*

Proof. The "if" direction is trivial; the "only if" direction follows from Ref. [97, Theorem 2]. □

[3] Note that we are using the convention that a node counts as its own ancestor, a la Refs. [90, 97, 112].

Note that Proposition 37 implies Proposition 30 as a special case: A pair of observed nodes share a common ancestor upon removing all other observed nodes from a DAG G if and and only they are adjacent in G.

We also highlight the utility of the definition of a setwise d-unrestricted set:

Proposition 40 (Setwise d-unrestriction $\iff P_{(4.2)} \in \mathcal{I}_G$)**.** *The distribution in Eq. (4.2) satisfies all the conditional independence constraints that follow from the observed d-separation relations relations of graph G if and only if $\{V_1...V_k\}$ are setwise d-unrestricted in G.*

Proof. Suppose that $\{V_1...V_k\}$ are *not* setwise d-unrestricted in G. Then, there exists a pair of nodes $\{S_i, S_j\} \subseteq \{V_1...V_k\}$ such that G exhibits the d-separation relation $(S_i \perp_{\mathrm{d}} S_j | \boldsymbol{Z})$. In this case, the distribution [4]

$$P(\boldsymbol{V}) = \frac{1}{2}\left(\delta_{\{V_1...V_k\},\vec{0}^k} + \delta_{\{V_1...V_k\},\vec{1}^k}\right) \tag{4.3}$$

$$\delta_{\boldsymbol{V}\setminus\{V_1...V_k\},\vec{0}^{(|\boldsymbol{V}|-k)}} \tag{4.4}$$

violates the conditional independence relation $S_i \perp\!\!\!\perp_{\mathrm{CI}} S_j | \boldsymbol{Z}$, and therefore $P_{(4.2)} \notin \mathcal{I}_G$. For the other direction, note that if $(S_i \not\perp\!\!\!\perp_{\mathrm{d}} S_j | \boldsymbol{Z})$ then so too $(\boldsymbol{S}_i \not\perp\!\!\!\perp_{\mathrm{d}} \boldsymbol{S}_j | \boldsymbol{Z})$ for any disjoint sets $\boldsymbol{S}_i$ and $\boldsymbol{S}_j$ wherein $S_i \in \boldsymbol{S}_i$ and $S_j \in \boldsymbol{S}_j$. Consequently, when $\{V_1...V_k\}$ are setwise d-unrestricted in G, there is no way to d-separate *any* subset of $\{V_1...V_k\}$ from any other subset of $\{V_1...V_k\}$ by any subset of the observed nodes outside of $\{V_1...V_k\}$, which are the *only* d-separation relations whose corresponding conditional independence relations would exclude the distribution of Eq. (4.3). This means that, in this case, $P_{(4.2)} \in \mathcal{I}_G$. □

Putting Propositions 37 and 40 together lead us to a natural generalization of maximality, which we now define and employ in a theorem.

Definition 41 (Setwise Maximal DAG). *Let G be DAG. If every subset of the observed nodes of G which is setwise d-unrestricted is also setwise adjacent then G is said to be* setwise-maximal*, otherwise G is said to be* setwise nonmaximal.

Theorem 42 (Setwise Nonmaximal)**.** *Every NON-INTERESTING DAG is setwise-maximal, that is, every setwise-nonmaximal DAG is INTERESTING.*

Proof. Theorem 42 follows immediately from Propositions 37 and 40. □

[4] Here $|\mathbf{X}|$ denotes the number of elements of the set $\mathbf{X}$, and $\vec{0}^n$ represents the vector with n entries equalling 0.

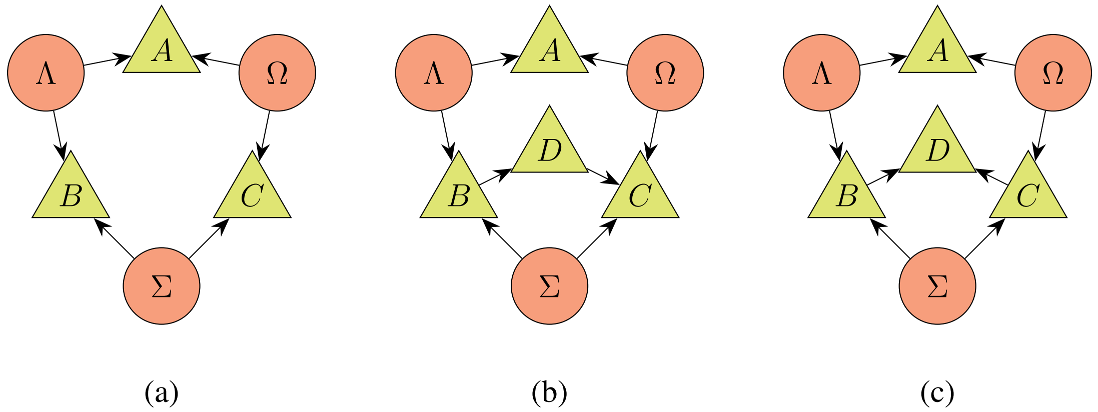


Figure 4.3: Examples of maximal DAGs which are setwise-nonmaximal. (a) is the Triangle scenario, where the set $\{A, B, C\}$ (all visible nodes) is setwise d-unrestricted but not setwise adjacent. In (b), the set $\{A, B, C\}$ is setwise d-unrestricted (in fact, there are no d-separation relations between observed nodes) but not setwise adjacent; by contrast, the larger set $\{A, B, C, D\}$ is *both* setwise d-unrestricted *and* setwise adjacent, despite (b) exhibiting the d-separation relation $(A \perp_{\mathrm{d}} D|B)$, as that d-separation relation involves conditioning on a node *within the set*. In (c), both of the sets $\{A, B, C\}$ and $\{A, B, C, D\}$ are d-unrestricted but not setwise adjacent. Note that $\{A, B, C, D\}$ is d-unrestricted in the DAG (c) despite that DAG exhibiting the d-separation relation $(A \perp_{\mathrm{d}} D|B, C)$, as that d-separation relation involves conditioning on nodes *within the set*.

The DAGs of Fig. 4.3 are setwise-nonmaximal, so they are shown INTERESTING by Theorem 42 (even if they are *not* shown INTERESTING by Theorem 33). On the other hand, the mDAG of Fig. 4.1(a) is both maximal and setwise maximal, as all maximal DAGs are also setwise maximal.

Note that there are precisely five mDAGs with 3 observed nodes which are INTERESTING (the Triangle scenario, the Unrelated Confounders scenario and three observationally equivalent versions of the Instrumental scenario). Theorem 42 certifies the INTERESTINGNESS of *all five*. Indeed, Theorem 42 turns out to be an extremely powerful filter for recognizing INTERESTING mDAGs with 4 or 5 observed nodes as well, as discussed in Section 4.6.

4.3 Using d-separation to prove INTERESTINGNESS

With Eq. (3.5), it was noted that any two DAGs that have different sets of observed d-separation relations are *not* classically observationally equivalent. In other words, imposing the same observed conditional independence constraints on the compatible distributions is a necessary condition for two DAGs to be classically observationally equivalent.

This fact can be used to establish the INTERESTINGNESS of some DAGs: if we can prove that the set of observed d-separation relations of a latent-permitting DAG *does not* match the set of d-separation relations of *any* latent-free DAG, then our latent-permitting DAG is classically observationally *inequivalent* to all latent-free DAGs. Via Evans' [28] Theorem 26, then, we can conclude that our latent-permitting DAG is INTERESTING.

This type of reasoning, that says that when a certain property of a DAG G is unmatched by all latent-free DAGs then G is INTERESTING, will be used a few times in this subsection and in the two next ones. As such, it will be useful to define the auxiliary term *Not Achievable in Latent-Free (NALF)*:

Definition 43 (NALF property of a DAG)**.** *Let G be a latent-permitting DAG. If G has a certain property that* does not match *that same property of any latent-free DAG, we say that this property of G is* NALF (not achievable in latent-free).

For example, we can have a DAG G whose set of observed d-separation relations is NALF. If there is a proof that a DAG is observationally inequivalent to all latent-free DAGs whenever certain property of the DAG is NALF, then this can be used to show INTERESTINGNESS. As discussed, this is the case for d-separation.

Theorem 44 (NALF d-sep)**.** *Let G be a DAG. Suppose that the set of observed d-separation relations of G is NALF (as per Definition 43). Then, G is INTERESTING.*

Proof. From Eq. (3.5), G is not classically observationally equivalent to any latent-free DAG. Thus, via Theorem 26, G is INTERESTING. □

The mDAG presented in Figure 4.1(a), which was not shown INTERESTING by Theorem 42 due to being setwise maximal, *can* be shown INTERESTING by Theorem 44: its set of observed d-separation relations, $A \perp_{\mathrm{d}} D|\varnothing$ and $B \perp_{\mathrm{d}} C|A$, is not matched by any latent-free DAG. This mDAG can be considered a special case of the bilocality scenario [13] restricted to have the same setting employed at both the extreme wings. Remarkably, it can support non-classical correlations even though the "setting" for the extreme wings is the same (in stark contradiction with the Bell scenario).

On the other hand, note that the Evans scenario (Figure 4.1(b)), which was shown INTERESTING by Theorem 33, cannot be shown INTERESTING via Theorem 44: it does not have any d-separation relation, just like the saturated latent-free DAGs. Therefore, Theorems 33 and 44 are not redundant to each other. Similarly, Theorems 42 and 44 are not redundant to each other.

There are a variety of different practical methods to ascertain whether or not Theorem 44 is satisfied for a given latent-permitting DAG G. Naively, one could construct *all* the latent-free DAGs with the same number of observed nodes as G, and one-by-one check whether any of them have observed d-separation relations matching those of G. Alternatively, one could envision employing a constraint-based causal discovery algorithm where the input to the algorithm is precisely the observed

d-separation relations of G: If the output of the causal discovery algorithm fails to include *any* latent-free DAG as a viable explanation given the input constraints, then evidently G is INTERESTING per Theorem 44. While such "brute force" approaches are viable for a small number of observed nodes, as the number of observed nodes increases one would need to contend with potential combinatorial explosion. As it turns out, however, when the DAG is maximal (and thus not shown INTERESTING by Theorem 33) there is an efficient way to check whether or not its set of observed d-separation relations is NALF; the efficient algorithm is presented in Appendix A.4.

As well as our previous methods,Theorem 44 is only a *sufficient* condition for INTERESTINGNESS, and not necessary. If there *exists* a latent-free DAG H which has the same observed d-separation relations as G (i.e. $\mathcal{I}_H = \mathcal{I}_G$), this *does not* imply that $\mathcal{C}_G = \mathcal{C}_H$, and as such we cannot conclude anything about the relation between $\mathcal{C}_G$ and $\mathcal{I}_G$. We will now discuss other sufficient conditions for INTERESTINGNESS that can be used when Theorems 33 and 44 fail.

4.4 Using *e*-separation to prove INTERESTINGNESS

Just as the mismatch of d-separation relations is a witness of classical observational inequivalence, we can show that the mismatch of e-separation relations, a concept that will be defined below, also witnesses classical observational inequivalence. As such, we can mimic the same logic used in the last section, thus showing that e-separation relations can be used to attest INTERESTINGNESS. An e-separation relation is defined as:

Definition 45 (***e*-separation**)**.** *Let G be a DAG, and let $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$ and $\boldsymbol{W}$ be four disjoint sets of nodes of G. Let $G_{del_{\boldsymbol{W}}}$ be the DAG obtained by starting from G and deleting the nodes of $\boldsymbol{W}$. The sets $\boldsymbol{X}$ and $\boldsymbol{Y}$ are said to be* e-separated by $\boldsymbol{Z}$ upon deletion of $\boldsymbol{W}$ *in G, denoted $(\boldsymbol{X} \perp_e \boldsymbol{Y} | \boldsymbol{Z})_{del_{\boldsymbol{W}}}$, if $\boldsymbol{X} \perp_d \boldsymbol{Y} | \boldsymbol{Z}$ holds in $G_{del_{\boldsymbol{W}}}$.*

Note that if $\boldsymbol{W} = \varnothing$, the concept of e-separation reduces to d-separation. As well as for the case of d-separation, we will say that an e-separation relation is an *observed e-separation relation* when the sets $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$ and $\boldsymbol{W}$ only involve observed nodes. Matching observed e-separation relations is a prerequisite for observational equivalence:

Lemma 46 (e-separation condition for observational equivalence)**.** *Let G and H be two DAGs. If they are classically observationally equivalent (i.e. $\mathcal{C}_G = \mathcal{C}_H$), then their sets of observed e-separation relations must be identical.*

Proof. In Ref. [29] it is shown that e-separation relations imply inequalities that must be satisfied by the compatible probability distributions. In Appendix E of that same reference, it is further shown that if a DAG *does not* exhibit an e-separation relation, then there must exist a compatible probability distribution which violates the inequality associated with that e-separation relation. This implies that, if a DAG G exhibits an e-separation relation which is not exhibited by another DAG H, then it is possible to find a probability distribution that is compatible with H but not with G. □

Using this lemma, we can derive the analog of Theorem 44 for the case of e-separation:

Theorem 47 (NALF e-sep)**.** *Let G be a DAG. Suppose that the set of observed e-separation relations of G is NALF (as per Definition 43). Then, G is* INTERESTING.

Proof. Follows directly from Lemma 46 and Theorem 26. □

To use Theorem 47 in practice, one might enumerate the e-separation relations exhibited by every latent-free DAG with the same number of observed nodes of G and compare them to the observed e-separation relations of G. Far more efficiently, one need only check against any *one* latent-free graph which matches the observed d-separation relations of G. (Such a latent-free DAG must exist if G is not already certified as INTERESTING by Theorem 44[5]). After all, if two latent-free graphs share the same d-separation relations, they will also share the same e-separation relations, per Theorem 6 and Lemma 46. We thus advise verifying that a DAG G in question is not already certified as INTERESTING by Theorem 44 before invoking Theorem 47.

It is clear that every DAG that can be shown INTERESTING by Theorem 44 can also be shown INTERESTING by Theorem 47, since d-separation is a special case of e-separation. A little less trivial, every DAG that can be shown INTERESTING by Theorem 33 can also be shown INTERESTING by Theorem 47: all nonmaximal DAGs have a set of e-separation relations which is NALF.

Proposition 48 (Theorem 47 subsumes Theorem 33)**.** *Let G be a nonmaximal DAG. Then, the set of e-separation relations of G is NALF.*

Proof. If G is nonmaximal, then there are at least two nodes A and B which are not d-separable, but are also not adjacent in G. If they are not adjacent in G, it is clear that A is e-separated from B by deletion of *every other* observed node of G.

Let us make a proof by contradiction: suppose that the set of e-separation relations of G *is not* NALF. Then, there is a latent-free DAG H which has exactly the same set of e-separation relations as G. This means that A should be e-separated from B by deletion of every other node of H, which implies that A and B are *not* adjacent in H. In a latent-free graph, two nodes are *nonadjacent* if and

[5] We do not discuss *how* to find a latent-free DAG with the same d-separation relations as G, as ultimately we conclude that there is apparently no advantage in doing so, as discussed in Appendix A.4.

only if said pair of nodes are d-separable [91, Prop. 3.19]. However, if A and B are d-separable in H but not in G, then their sets of d-separation relations do not coincide, which is a contradiction. □

It is still an open question whether the conjunction of Theorems 33 and 44 is as good as Theorem 47 to show Interestingness. We did not find any DAG that was shown Interesting by Theorem 47 but not by any of these two previous methods. (See Appendix A.6 for further discussion of this question.)

Note, however, that Theorem 47 *does not* subsume Theorem 42. For example, the Triangle scenario (Fig. 4.3(a)), that is shown Interesting by Theorem 42, does *not* have any e-separation relation (just like the saturated latent-free DAGs).

Appendices A.2 and A.3 relate Theorems 33 and 47 to results of prior literature. In particular, in Appendix A.2 we show that a version of Theorem 33 in terms of e-separation that was presented in Ref. [81] is incorrect.

4.5 Using incompatible supports to prove Interestingness

The final method we used to prove Interestingness is based on the classical feasibility of supports.

Given a set of random variables $\{X_1, ..., X_n\}$, a specific set $\{X_1 = x_1, ..., X_n = x_n\}$ of values that these random variables can take is called an *event*. The *support* $\mathcal{S}(P_{X_1,...,X_n})$ of a probability distribution $P_{X_1,...,X_n}$ over the variables $\{X_1, ..., X_n\}$ is the set of events that have non-zero probability:

$$\mathcal{S}(P_{X_1,...,X_n}) = \{\{x_1, ..., x_n\} \mid P_{X_1,...,X_n}(x_1, ..., x_n) > 0\} \tag{4.5}$$

We previously defined what it means for a probability distribution to be classically compatible with a DAG. Here, we define what it means for a support to be classically compatible with a DAG:

Definition 49 (Compatibility of a support with a DAG). *Let G be a DAG with observed nodes $\boldsymbol{A} = \boldsymbol{V} \cup \boldsymbol{L}$, where $\boldsymbol{V}$ are observed nodes and $\boldsymbol{L}$ are latent nodes. Let $\mathcal{S}$ be a set of events over the variables $\boldsymbol{V}$. We say that $\mathcal{S}$ is a support* classically compatible with *G if there exists a probability distribution $P_{\boldsymbol{V}}$ over $\boldsymbol{V}$ that is classically compatible with G (i.e., $P_{\boldsymbol{V}} \in \mathcal{C}_G$) and whose support is $\mathcal{S}(P_{\boldsymbol{V}}) = \mathcal{S}$. We say that $\mathcal{S}$ is a support* compatible-up-to-CI *with G if there exists a probability distribution $P_{\boldsymbol{V}}$ over $\boldsymbol{V}$ such that $P_{\boldsymbol{V}} \in \mathcal{I}_G$) and whose support is $\mathcal{S}(P_{\boldsymbol{V}}) = \mathcal{S}$.*

As an example, the following support is *not* compatible with the Bell DAG (Figure 3.1), as it corresponds to the Popescu-Rohrlich box [52, 84]:

$$\mathcal{S}_{\text{Bell}} = \left\{\begin{bmatrix} \{X=0,Y=0,A=0,B=0\} \\ \{X=0,Y=0,A=1,B=1\} \\ \{X=0,Y=1,A=0,B=0\} \\ \{X=0,Y=1,A=1,B=1\} \\ \{X=1,Y=0,A=0,B=0\} \\ \{X=1,Y=0,A=1,B=1\} \\ \{X=1,Y=1,A=1,B=0\} \\ \{X=1,Y=1,A=0,B=1\} \end{bmatrix}\right\} \tag{4.6}$$

Note that if a support *is* compatible with DAG G, that does *not* mean that every distribution with that support will be compatible with G. There are countless counterexamples, but let us simply note that the full support (the one where all events have positive probability) is compatible with any DAG, but at the same time we know of many incompatible distributions which nevertheless have full support.

Naturally, admitting the same set of compatible supports is a prerequisite for two DAGs to admit the same set of compatible distributions:

Lemma 50 (Supports condition for observational equivalence)**.** *Let G and H be two DAGs. If they are classically observationally equivalent (i.e. $\mathcal{C}_G = \mathcal{C}_H$), then their sets of classically compatible supports must be identical.*

It remains an open question whether the condition of Lemma 50 is also necessary for observational equivalence. In particular, it is not known whether or not there exists a DAG for which some distributions are incompatible (due to inequalities) but for which all *supports* are compatible.

As before, this necessary condition for observational equivalence immediately translates into a method for proving INTERESTINGNESS:

Theorem 51 (NALF Supports)**.** *Let G be a DAG. Suppose the set of classically compatible supports of G is NALF (as per Definition 43). Then, G is INTERESTING.*

To exploit Theorem 51 in practice, we need an an algorithm capable of assessing whether or not a given support is compatible with a given DAG. Such algorithm was developed in Ref. [31], and is referred to here as *Fraser's algorithm*. We have implemented Fraser's algorithm in Python and scripted it to yield *all* the supports that are classically incompatible with a given DAG (for a certain assignment of the cardinalities of the observed variables).

In general, Fraser's algorithm is much more computationally expensive than simply assessing whether or not a graph exhibits some d-separation or e-separation relation. Consequently, we consider Theorem 51 a method of last resort to show INTERESTINGNESS.

4.5.1 RAPIDLY TESTING SUPPORTS (WITHOUT COMPARING TO ANY LATENT-FREE GRAPH)

As well as for the case of e-separation, Theorem 51 has the downside that it requires one to find the supports compatible with the DAG G and then check the compatible supports of all the latent-free DAGs with the same number of observed nodes (or alternatively to find the latent-free H that has the same set of d-separation relations as G, and then check which supports are compatible with H). Since Fraser's algorithm is computationally expensive, doing this in practice can be cumbersome.

Luckily, it is possible to develop a rapid supports test where it is not even necessary to find *all* of the supports compatible with G. The idea of the rapid supports test comes from noting that sometimes we can prove the *incompatibility* of a given support with a DAG G by recognizing that the given support conflicts with a d-separation relation exhibited by G.

Suppose, for instance, that a DAG G has the (unconditional) d-separation relation $A \perp_{\mathrm{d}} B$. Then, the support in Eq. (4.7) is clearly incompatible with G, since any probability distribution with that support must have $P_{AB}(1,1) = 0 \neq P_A(1)P_B(1) > 0$, contradicting $A \perp\!\!\!\perp_{\mathrm{CI}} B$.

$$\mathcal{S}_{4.4} = \begin{Bmatrix} \{A=0, B=0\} \\ \{A=0, B=1\} \\ \{A=1, B=0\} \end{Bmatrix} \tag{4.7}$$

Indeed, we can formally categorize all such "trivial" proofs of support incompatibility through the following two definitions:

Definition 52 (Support conflicting with a conditional independence relation). *Let $\mathcal{S}$ be a support over a set $\boldsymbol{V}$ of variables, and let $\boldsymbol{A} \subseteq \boldsymbol{V}$, $\boldsymbol{B} \subseteq \boldsymbol{V}$ and $\boldsymbol{C} \subseteq \boldsymbol{V}$ be three disjoint subsets of $\boldsymbol{V}$. We say that $\mathcal{S}$* conflicts with the conditional independence relation $\boldsymbol{A} \perp\!\!\!\perp_{CI} \boldsymbol{B}|\boldsymbol{C}$ *if there exists a set $\{\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}\}$ of values of the variables in $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ such that the events $\{\boldsymbol{A} = \boldsymbol{a}, \boldsymbol{C} = \boldsymbol{c}\}$ and $\{\boldsymbol{B} = \boldsymbol{b}, \boldsymbol{C} = \boldsymbol{c}\}$ occur in $\mathcal{S}$, but the event $\{\boldsymbol{A} = \boldsymbol{a}, \boldsymbol{B} = \boldsymbol{b}, \boldsymbol{C} = \boldsymbol{c}\}$ does not occur in $\mathcal{S}$.*

For example, the support of Eq. (4.7) conflicts with the conditional independence relation $A \perp\!\!\!\perp_{\mathrm{CI}} B$: both the events $A = 1$ and $B = 1$ occur in the support, but the event $\{A = 1, B = 1\}$ does not.

If a support $\mathcal{S}$ conflicts with a conditional independence relation $\boldsymbol{A} \perp\!\!\!\perp_{\mathrm{CI}} \boldsymbol{B}|\boldsymbol{C}$, then there is no probability distribution with support $\mathcal{S}$ that obeys $\boldsymbol{A} \perp\!\!\!\perp_{\mathrm{CI}} \boldsymbol{B}|\boldsymbol{C}$. This will be seen explicitly in the proof of Lemma 54.

Definition 53 (**Triviality of support incompatibility**)**.** *A support $\mathcal{S}$ is said to be* trivially incompatible *with a given DAG whenever the DAG exhibits some d-separation relation whose associated conditional independence relation conflicts with $\mathcal{S}$ (as in Definition 52).*

By generalizing the discussion made around Eq. (4.7), we see that this definition indeed implies in classical incompatibility of the support with the DAG:

Lemma 54 (Trivial incompatibility implies incompatibility)**.** *If a support $\mathcal{S}$ is trivially incompatible with a DAG G, then it is classically incompatible with G.*

Proof. Let $\boldsymbol{A} \perp_{\mathrm{d}} \boldsymbol{B}|\boldsymbol{C}$ be a d-separation relation of G which is in conflict with $\mathcal{S}$. Furthermore, let $\{\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}\}$ be a set of values of the variables in $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ that witnesses this conflict. This means that all of the probability distributions which have the support $\mathcal{S}$ must have $P_{\boldsymbol{AB}|\boldsymbol{C}}(\boldsymbol{ab}|\boldsymbol{c}) = 0 \neq P_{\boldsymbol{A}|\boldsymbol{C}}(\boldsymbol{a}|\boldsymbol{c})P_{\boldsymbol{B}|\boldsymbol{C}}(\boldsymbol{b}|\boldsymbol{c}) > 0$. Therefore, $\mathcal{S}$ is not classically compatible with G. □

The key insight which allows us to accelerate the application of Theorem 51 is that for a latent-free DAG, the *only* supports incompatible with it are those which are *trivially* incompatible with it.

Lemma 55 (Latent-free support compatibility)**.** *Let H be a latent-free DAG. If $\mathcal{S}$ is a support which is not trivially incompatible with H as per Definition 53, then $\mathcal{S}$ is classically compatible with H as per Definition 49.*[6]

Proof. Let $X_1, \ldots, X_n$ be the nodes of H in some topological order, i.e. an order where X_i is a non-descendant of X_{i+1}. We will make this proof by explicitly constructing a probability distribution $P(X_1, ..., X_n)$ which is classically compatible with H and has the support $\mathcal{S}$.

The explicit construction is given by $P(X_1, ..., X_n) = \prod_i P(X_i|\mathrm{PA}_H(X_i))$, where each $P(X_i|\mathrm{PA}_H(X_i))$ is uniformly distributed over the values of X_i which occur along with the given values of $\mathrm{PA}_H(X_i)$ (and any value of the remaining variables) in $\mathcal{S}$. It is clear that this distribution is classically compatible with H, since it takes the form of the Markov condition. Now, we will show that its support is in fact $\mathcal{S}$.

We will make a proof by induction on m that this distribution has the correct support on $X_1, ..., X_m$. For $m = 1$ we simply have $P(X_1)$ uniformly distributed on values of X_1 that are possible under $\mathcal{S}$, so the basis case is immediately satisfied. Now, we assume that $X_1, ..., X_k$ has the correct support, and we will prove that $X_1, ..., X_{k+1}$ also does.

[6] More generally, we conjecture that for *any* DAG G, if $\mathcal{S}$ is a support which is not trivially incompatible with G as per Definition 53, then $\mathcal{S}$ is compatible *up-to-CI* with G. We did not prove this, however, except for the special case when G is latent-free.

First, consider some $(x_1, \dots, x_k, x_{k+1})$ that occurs in $\mathcal{S}$. Since we assumed that $X_1, ..., X_k$ has the correct support, we know that the corresponding $P(X_1, ..., X_k)$ is non-zero. All of the parents of X_{k+1} are elements of $\{X_1, ..., X_k\}$; therefore, by definition, the corresponding $P(X_{k+1}|\text{PA}_H(X_{k+1}))$ is also non-zero. Therefore, $P(X_1, ..., X_{k+1}) = P(X_1, ..., X_k)P(X_{k+1}|\text{PA}_H(X_{k+1}))$ is non-zero as required.

Now, consider some $(x_1, \dots, x_k, x_{k+1})$ that *does not* occur in $\mathcal{S}$. There are two possibilities: the set of values $(x_1, \dots, x_k)$ could occur or not occur in $\mathcal{S}$. If $(x_1, \dots, x_k)$ also does not occur in $\mathcal{S}$, the proof is simple: by the inductive hypothesis, the corresponding $P(X_1, ..., X_k)$ is zero, so $P(X_1, ..., X_{k+1})$ has to be zero as required.

Suppose now that $(x_1, \dots, x_k, x_{k+1})$ does not occur in $\mathcal{S}$ but $(x_1, \dots, x_k)$ *occurs*. Let $C_1, \dots, C_p$ denote the parents of X_{k+1} and $B_1, \dots, B_{k-p}$ the remaining variables among $X_1, \dots, X_k$, which we know are non-descendants of X_{k+1} in H. The DAG H must have the d-separation relation $X_{k+1} \perp_{\text{d}} \{B_1, \dots, B_{k-p}\}|\{C_1, \dots, C_p\}$; since $\mathcal{S}$ is *not* trivially incompatible with H, then it must *not* be in conflict with the associated conditional independence relation. Since $(x_1, ..., x_k) = (b_1, ..., b_{k-p}, c_1, ..., c_p)$ occurs in $\mathcal{S}$ but $(x_1, ..., x_k, x_{k+1}) = (b_1, ..., b_{k-p}, c_1, ..., c_p, x_k + 1)$ does not occur, we can then conclude that $(x_{k+1}, c_1, \dots, c_p)$ *must not occur* in $\mathcal{S}$. Therefore, by definition the associated $P(X_{k+1}|\text{PA}_H(X_{k+1})) = P(X_{k+1}|C_1, ..., C_p)$ is zero and hence so is $P(X_1, \dots, X_k, X_{k+1})$, as required. □

Accordingly, we have the following upgrade to Theorem 51:

Theorem 56 (Rapid Supports). *Let G be a DAG. If there is a support $\mathcal{S}$ which is incompatible with G despite* not *being trivially incompatible with G, then G is* INTERESTING.

Proof. We start by noting that if G was NON-INTERESTING, then it would be classically observationally equivalent to some latent-free DAG H, per Theorem 26. If so, then H and G would (at least!) share the same d-separation relations, and hence a support would only *not* be trivially incompatible with G if was not trivially incompatible with H. But by Lemma 55, if a support is not trivially incompatible with H then it must be compatible with H. Since our starting premise is that this support is not compatible with G, then by Lemma 50 it follows that $\mathcal{C}_G \neq \mathcal{C}_H$, and hence G is INTERESTING. □

Note that Theorem 56 allows us to leverage tools distinct from Fraser's algorithm for assessing support incompatibility. Fraser's algorithm is a *necessary and sufficient* test for support compatibility. For the purposes of Theorem 56, however, we can instead consider variant algorithms related to Inflation [107] which cannot certify support compatibility but which can often efficiently detect support incompatibility. Some such algorithms are discussed in Ref [89], for example.

It is clear that the application of Theorem 56 is much more efficient than the application of Theorem 51. We can also show that both theorems are equally powerful:

Proposition 57. *Let G be a* Interesting *DAG. If the* Interestingness *of G can be shown via Theorem 51, then it can also be shown via Theorem 56.*

Proof. First, suppose that there exists a latent-free DAG H such that $\mathcal{I}_H = \mathcal{I}_G$. If G is Interesting, then $C_G \subsetneq C_H$, and hence every support which is incompatible with G must also be incompatible with H. If Theorem 51 shows the Interestingness of G, then there must be a support which is *compatible* with H but not G. Since the set of supports incompatible with H is exactly the set of trivially incompatible supports as per Lemmas 54 and 55, it follows that the support which is incompatible with G but not H must *not* be trivially incompatible with G. Therefore, the Interestingness of G can be shown by Theorem 56.

Now, suppose that there is *no* latent-free DAG that has the same set of d-separation relations as G, i.e. that the set of d-separation relations of G is NALF. If G is not maximal, in the proof of Theorem 33 we showed a distribution which is not compatible with G (Eq. 4.1). In reality, *all* the distributions which have the same support as this one will be incompatible with G. Furthermore, since Eq. 4.1 is an element of $\mathcal{I}_G$, this support is not trivially incompatible with G. Therefore, the support of the distribution of Eq. 4.1 is an incompatible support which is not trivially incompatible.

If the set of d-separation relations of G is NALF and G *is* maximal, per Theorem 79 we know that G has one of the eighteen mDAGs of Figure A.2 as a subgraph. One can explicitly check that these eighteen mDAGs *have* incompatible supports that are not trivially incompatible; therefore, G itself also must have incompatible supports that are not trivially incompatible, namely, by taking all its observed variables outside of the pertinent subgraph to be point distributed. Therefore, this case also falls under the scope of Theorem 56. □

It is also worth noting that, when there *is* a latent-free DAG H with the same set of d-separation relations as G, Theorem 56 is *constructive*: the distribution P constructed for H in the proof of Lemma 55 is such that $P \in \mathcal{I}_G$ yet $P \notin \mathcal{C}_G$.

An example of a support which is incompatible — but not *trivially* incompatible — with the Evans scenario (Figure 4.1(b)) is the following:

$$\mathcal{S}_{\text{Evans}} = \begin{Bmatrix} \{C=0, D=0, E=0\} \\ \{C=0, D=1, E=1\} \end{Bmatrix} \tag{4.8}$$

The Evans scenario does not have any d-separation relation. Nevertheless, the support $\mathcal{S}_{\text{Evans}}$ of Equation (4.8) is not compatible with it. This can be seen by noting that the variable C in $\mathcal{S}$ is associated with a point distribution: it always takes the value 0. Since in the Evans scenario all the correlation between D and E is established through C, it is impossible to have perfect correlation between D and E while C takes a point distribution.

An example of DAG whose INTERESTINGNESS was first certified in Ref. [31] via the discovery of an incompatible support is presented in Figure 4.4. By means of his eponymous algorithm for compatible supports, Fraser showed that the following support is classically incompatible with the DAG of Figure 4.4[7] :

$$\mathcal{S}_{4.4} = \begin{Bmatrix} \{A=0, B=0, C=0, D=0\} \\ \{A=0, B=0, C=0, D=1\} \\ \{A=0, B=1, C=0, D=0\} \\ \{A=1, B=0, C=1, D=0\} \end{Bmatrix} \tag{4.9}$$

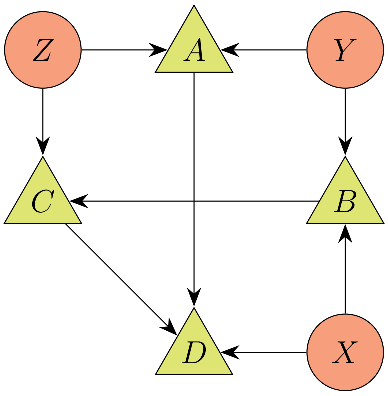


Figure 4.4: A DAG with 4 observed nodes and 7 total nodes whose INTERESTINGNESS can be shown by Fraser's algorithm for compatible supports.

When using Theorem 56 to attest INTERESTINGNESS, we start by checking supports at binary cardinalities of the observed variables. If such an incompatible but not trivially incompatible support is found, we can try to search for such a support at higher cardinalities of the observed variables.

Remark 58. *We anticipated that if a DAG has* any *incompatible support (for* any *cardinality), then it seemed likely that we should expect to find* some *incompatible support where all variables have binary cardinality. Indeed, prior to this work, we are not aware of any counterexample. Even the three challenging DAGs identified in Figure 14 of Ref. [27] were eventually found to have incompatible supports with merely binary variables. However, this intuition turns out to be misplaced: We identified*

[7] Note that the support is not *trivially* incompatible, since the DAG in question does not exhibit any d-separation relations involving observed nodes

4 mDAGs for which the high-cardinality support of Eq. 4.10 is identified as incompatible, but where nevertheless every *support over binary variables is provably compatible. These mDAGs are depicted in Table 4.1.*

$$\mathcal{S}_{\text{for Table 4.1}} = \begin{Bmatrix} \{A = 0, B = 0, C = 0, D = 0\} \\ \{A = 0, B = 0, C = 1, D = 0\} \\ \{A = 0, B = 1, C = 0, D = 0\} \\ \{A = 1, B = 0, C = 0, D = 0\} \\ \{A = 1, B = 1, C = 0, D = 0\} \\ \{A = 2, B = 0, C = 0, D = 1\} \\ \{A = 2, B = 1, C = 1, D = 0\} \end{Bmatrix} \tag{4.10}$$

4.5.2 INCOMPATIBLE SUPPORTS SUBSUMES ALL OTHER METHODS

The methods to show INTERESTINGNESS that we presented here are not independent of each other. For example, it is clear that Theorem 47 subsumes Theorem 44, and Proposition 48 shows that Theorem 47 also subsumes Theorem 33. However, it is more efficient to start by checking maximality of the DAG of interest or the d-separation relations of latent-free DAGs instead of directly checking all of their e-separation relations; this is the reason why we presented these results separately.

As it turns out, we can also show that Theorem 51 subsumes Theorem 47: every DAG that can be shown INTERESTING via its e-separation relations (through Theorem 47) can also be shown INTERESTING via its incompatible supports (through Theorem 51). This result will be presented here as Corollary 60. All of the discussion about e-separation was made here because checking e-separation relations is in general much faster than checking incompatible supports.

Theorem 59 (Supports subsumes e-separation for observational inequivalence)**.** *Let G and H be two DAGs. If their sets of observed e-separation relations are different, then it is possible to find a support* over binary variables *which is compatible with one of them but incompatible with the other.*

Proof. Presented in Appendix A.5. □

Note that Theorem 59 is a generic result about observational equivalence between two DAGs, not restricted to the comparison with latent-free DAGs. A direct corollary is:

Corollary 60 (Supports subsumes e-separation for INTERESTINGNESS)**.** *Let G be a DAG. If there is no latent-free DAG which presents the same set of e-separation relations as G, then there is also no*

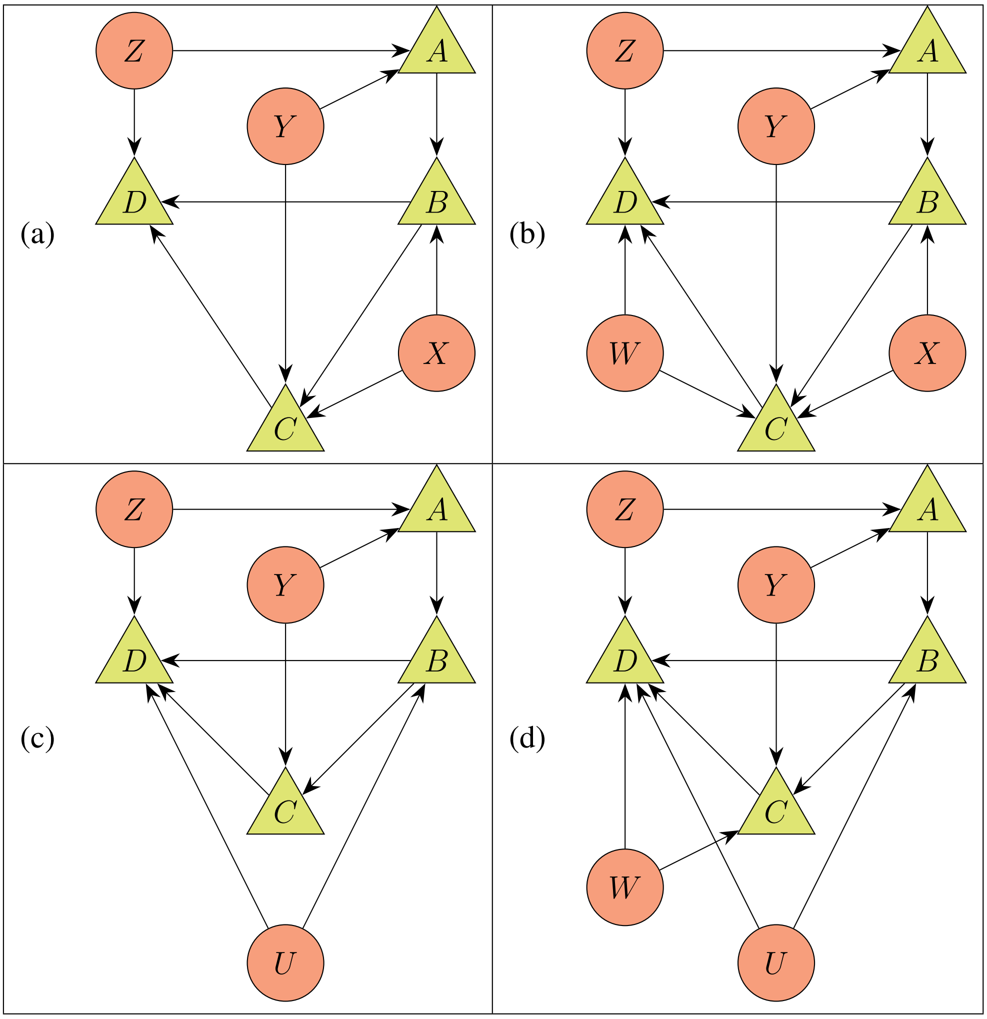


Table 4.1: The only 4 mDAGs with 4 observed nodes for which every support over binary observed variables is classically compatible but which are nevertheless provably INTERESTING by virtue of the higher-cardinality support given in Eq. (4.10) being incompatible with all 4 DAGs here.

latent-free DAG which presents the same set of classically compatible supports as G. In other words, if G can be shown INTERESTING *by Theorem 47, then it can also be shown* INTERESTING *by Theorem 51.*

Since Proposition 57 shows that the rapid supports test (Theorem 56) can show INTERESTINGNESS in all the cases covered by Theorem 51, from Corollary 60 we can conclude that Theorem 56 is the most powerful tool we have so far to show INTERESTINGNESS. It is still an open question whether detection by Theorem 56 is a necessary condition for INTERESTINGNESS.

Note that Theorem 56 also subsumes Theorem 42, as all distributions *with same support* of that in Eq. (4.2) are in $\mathcal{I}$ but not $\mathcal{C}$, and hence every DAG which is provably INTERESTING via Theorem 42 will also be provably INTERESTING via Theorem 56. In particular, merely by considering compatible versus incompatible supports *over two events*, Theorem 56 already subsumes Theorem 42.

4.6 Computational Results

We consider the problem of certifying the INTERESTINGNESS of those mDAGs of 4 observed nodes which are not shown NON-INTERESTING by the HLP criterion (Corollary 24). We start by the method described in Subsection 4.1, that says that nonmaximal DAGs are INTERESTING. Among the 996 mDAGs which are left as *potentially* interesting after applying the HLP criterion, we find that 810 are nonmaximal. Therefore, nonmaximality seems to be a powerful tool to show INTERESTINGNESS; at this stage, we are left with only 186 unresolved mDAGs after this preliminary filtering.

We then exploit the d-separation test for INTERESTINGNESS per Subsection 4.3. That is, among those 186 remaining mDAGs we filter out any mDAGs which possess observed d-separation relations *not* matching those of some latent-free DAG (NALF d-separation relations). We find that 168 mDAGs remain as-yet unresolved — the 18 mDAGs that are shown INTERESTING at this stage are the ones presented in Figure A.2.

The e-separation test for INTERESTINGNESS presented in Subsection 4.4 does not resolve *any* of these 168 unresolved cases. As mentioned, it is still an open problem whether the nonmaximality test and the d-separation test together will always subsume the e-separation test. We then turn to the method of supports analysis per Subsection 4.5, which ultimately leaves us with only 3 remaining unresolved mDAGs.

More specifically, by considering supports with *binary* cardinalities of the observed variables we were able to certify the INTERESTINGNESS of 161 out of the 168 remaining mDAGs. We could not, however, find *any* classically incompatible supports for the remaining 7 mDAGs when only considering binary cardinality variables. But by increasing the cardinality of variable A to be three we were able to identify a support that is incompatible — but not *trivially* incompatible — with the four mDAGs in Table 4.1, hence certifying their INTERESTINGNESS. Said support is explicitly reproduced in Eq. (A.2); it is obviously not trivially incompatible with any of the mDAGs in Table 4.1, since none of those four mDAGs exhibits any d-separation relations over its observed nodes. The remaining 3 mDAGs which we were *unable* to resolve as INTERESTING via supports considerations — up to computational tractability limits — are depicted in Table 4.2.

The exact number of mDAGs that we are able to characterize as INTERESTING or NON-INTERESTING at each stage is summarized in Table 4.3.

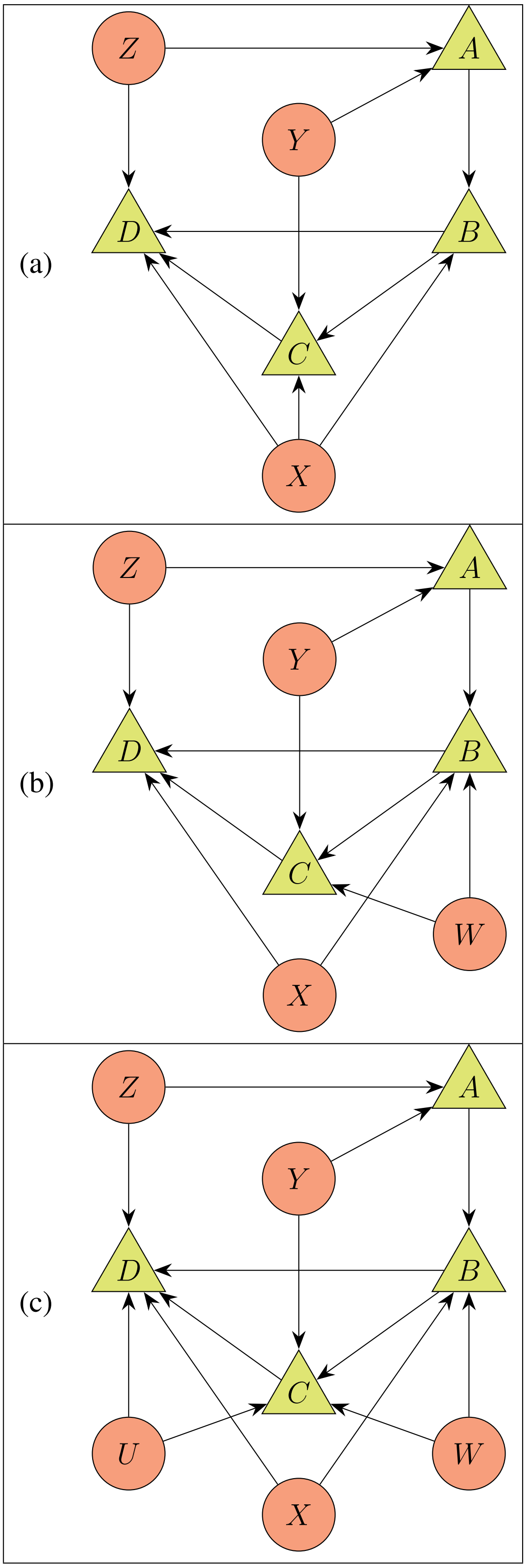


Table 4.2: The mDAGs of 4 observed nodes whose Interestingness remains unresolved.

4.7 Entropy Vector method to classify INTERESTING scenarios

In an arbitrary scenario with multiple unobserved common causes, the set of classically compatible distributions need not form a convex set. This motivates, the Entropy Vector Method for classifying INTERESTING scenarios. The idea is to check the membership problem in the entropy space where the relevant sets become convex. Here we briefly overview the entropy vector method to classify INTERESTING causal scenarios. For details we point the reader towards Refs [104, 15, 63]

Definition 61 (Entropy Vector)**.** . *Let $P(X_1, \dots X_N)$ be a probability distribution over N random variables $X_1, \dots X_N$. The entropy vector of $P(X_1, \dots X_N)$ is defined as the vector with 2^N components as follows:*

$$(H(\phi), H(X_1), H(X_2), \dots H(X_N), H(X_1, X_2), \dots H(X_1, \dots X_N)) \tag{4.11}$$

That is the entropy vector has a component corresponding to every element of the power set of $\{X_1, \dots X_N\}$ where the component itself is the entropy of the corresponding element of the power set. The entropy vector of a distribution $P(X_1, \dots X_N)$ is denoted by $H(P(X_1, \dots X_N)) \in \mathbb{R}^{2^N}$.

Definition 62 (Entropy Cone [102])**.** *Let Γ_N^* be the set of all entropy vectors, that is $\Gamma_N^* = \{v \in \mathbb{R}^{2^N} : \exists P(X_1, \dots X_N) \;\; s.t \;\; v = H(P(X_1, \dots X_N))\}$. The closure of Γ_N^*, i.e. $\bar{\Gamma}_N^*$ is known as the entropy cone.*

4.7.1 SHANNON-TYPE INEQUALITIES

Shannon, first introduced the concept of entropy in the context of information theory to characterize the amount of information in a certain data [94]. Many different measures of entropy have been introduced since then. In this work we stick to Shannon's definition of Information Entropy. Following the conventions in [104, 111], the Shannon entropy for a discrete random variable Y taking values in $\mathbb{R}$ with a probability distribution $P(Y)$, is given by

$$H(Y) := -\sum_{y \in \mathbb{R}} P(Y = y) \log_2(P(Y = y)) \tag{4.12}$$

where $0 \log_2 0$ is taken to be 0 because $\lim_{y \to 0} \;\; y \log_2 y \;\; = \;\; 0$

It is trivial to see that a basic necessity of Shannon's entropy measure is $H(Y) \geq 0$.[8]

[8] This is because $0 \leq P(Y = y) \leq 1 \; \forall y$.

Further, the *Conditional Entropy* of Y conditioned on Z is defined as

$$H(Y|Z) := H(YZ) - H(Z). \tag{4.13}$$

The *Mutual Information* of Y and Z is defined as

$$I(Y:Z) := H(Y) + H(Z) - H(YZ) \tag{4.14}$$

and the *Conditional Mutual Information* of Y and Z conditioned on W is defined as

$$I(Y:Z|W) := H(YW) + H(ZW) - H(W) - H(YZW). \tag{4.15}$$

For an extensive review of Shannon's information measures we point the reader towards [111, 110, 114, 113]. The upshot is that the non-negativity of all Shannon's information measures form a set of inequalities called the basic inequalities. This set of basic inequalities is not however minimal. For example:

$$H(X|Y) \geq 0 \tag{4.16}$$

and

$$I(X:Y) \geq 0 \tag{4.17}$$

which are both basic inequalities involving random variables X and Y imply,

$$H(X) = H(X|Y) + I(X:Y) \geq 0 \tag{4.18}$$

which again is a basic inequality involving random variables X and Y.

The important point is that any Shannon's information measure can be expressed as the sum of Shannon's information measures of the following two elemental forms:

$$H(X_i|X_{\mathcal{N}_n-\{i\}}) \tag{4.19}$$

where $i \in \mathcal{N}_n$ (the set of all Natural numbers) and

$$I(X_i : X_j|X_k) \tag{4.20}$$

and where $i \neq j$ and $k \subset \mathcal{N}_n - \{i, j\}$

Thus, if given a certain number, N, of random variables, we can construct the minimal set of basic inequalities formed by Shannon's information measures by using equations (4.19) and (4.20).

This minimal set is called the set of *Shannon-type Inequalities* for the given random variables. It is easy to check [111] that the total number of Shannon type inequalities for N random variables are just given by:

$$m = n +^{n} C_2 2^{n-2} \tag{4.21}$$

A linear programming approach is used to list out all the Shannon type inequalities for a given number of random of variables.

Definition 63 (Shannon-type inequalities)**.** *The minimal set of basic linear inequalities formed by Shannon's information measures by using equations* (4.19) *and* (4.20) *constitutes the set of Shannon-type inequalities or Shannon constraints. The set of all vectors $u \in \mathbb{R}^{2^N}$ that obey these Shannon-type inequalities make up a convex cone in the entropy space, called the Shannon Cone Γ_n.*

The Shannon cone is an outer approximation to $\overline{\Gamma}^*_N$, i.e. $\overline{\Gamma}^*_N \subseteq \Gamma_N$. Hence, there are are vectors $u \in \mathbb{R}^{2^N}$ obeying the Shannon type inequalities that are not entropy vectors. On the other hand entropy vector corresponding to any probability distribution $P(X_1 \ldots X_N)$ satisfies the Shannon type inequalities.

So far, all the constraints on the entropy vector are independent of the particular causal structure under consideration. However, if we consider the causal structure we know that the d-separation relations in it imply conditional independences. These conditional independences therefore enforce more constraints on the entropy vectors corresponding to the sets of distributions compatible with a causal scenario. For eg: the constraint $X \perp\!\!\!\perp_{\text{CI}} Y|Z \implies I(X:Y|Z) = 0$. Such constraints will be called entropic causal constraints from hereon.

The set of all entropy vectors, $v \in \Gamma^*_N$, corresponding to the classically compatible distributions of a causal scenario G and which additionally satisfy the entropic causal constraints (arising from the Markov condition in the scenario) is denoted by $\Gamma^*_N(C_G)$. The set of all entropy vectors corresponding to the classically compatible distributions of a causal scenario, G, in the outer approximation of Γ^*_N which also satisfy the entropic causal constraints (from the Markov condition in the scenario) is denoted as $\Gamma_N(C_G)$. By construction we have $\Gamma^*_N(C_G) \subseteq \Gamma_N(C_G)$. Further it can be shown [104] that $\overline{\Gamma}^*_N(C_G)$ is convex, even though its characterisation is difficult. For this purpose, $\Gamma_N(C_G)$ is used to get an outer approximation of $\overline{\Gamma}^*_N(C_G)$. We call $\Gamma_N(C_G)$ the Shannon-cone of the classically compatible set of distributions (i.e. C_G) corresponding to a causal scenario, G . Since a causal scenario might consist of unobserved nodes as well, the characterisation of $\Gamma_N(C_G)$ over the visible nodes requires a projection of the full Shannon-cone, $\Gamma_N(C_G)$, over all the nodes onto the visible nodes. This requires quantifier elimination, normally the Fourier-Motzkin Elimination algorithm which can be computationally quite costly. Similarly, the Shannon-cone of the set of distributions I_G of a causal scenario G is denoted as $\Gamma_N(I_G)$ and it is an outer approximation to the entropy cone, $\Gamma^*_N(I_G)$ corresponding to the set of distributions I_G of a the causal scenario G.

The entropy vector method to classify a distinction between the sets C_G and I_G for a causal scenario G is based on comparing the Shannon-Cones corresponding to the sets C_G and I_G, i.e, on comparing $\Gamma_N(C_G)$ and $\Gamma_N(I_G)$. If the Shannon-cone $\Gamma_N(C_G)$ is contained strictly inside the Shannon cone $\Gamma_N(I_G)$, the set $\Gamma^*_N(C_G)$ could be a strict subset of $\Gamma^*_N(I_G)$ but not necessarily. This is because $\Gamma_N(C_G)$ and $\Gamma_N(I_G)$ are only outer approximations of $\Gamma^*_N(C_G)$ and $\Gamma^*_N(I_G)$ respectively. If $\Gamma_N(C_G)$ is strictly contained inside $\Gamma_N(I_G)$, there would be some Shannon-type inequalities constraining $\Gamma_N(C_G)$ but not constraining $\Gamma_N(I_G)$. Finding such inequalities again does not necessarily imply that the entropy cone $\Gamma^*_N(C_G)$ is contained inside the entropy cone $\Gamma^*_N(I_G)$. This is because the found inequality could be a Non-Shannon type inequality for $\Gamma^*_N(I_G)$. However, finding a probability distribution in I_G that violates the inequality that is Shannon-type for $\Gamma_N(C_G)$ but not Shannon-type for $\Gamma_N(I_G)$ would imply that the concerned inequality is also not Non-Shannon-type for $\Gamma_N(I_G)$. This would lead to the conclusion that $\Gamma^*_N(C_G)$ is strictly contained in $\Gamma^*_N(I_G)$ and thus $C_G \subset I_G$. In this way we can attest the INTERESTINGNESS of the causal scenario using the entropy vector approach. However, for a causal scenario, it might turn out that $\Gamma_N(C_G) = \Gamma_N(I_G)$, in which case there will be no Shannon-type inequalities that constrain $\Gamma_N(C_G)$ but do not constrain $\Gamma_N(I_G)$. In such a scenario we would not be able to conclude anything about the entropy cones corresponding to the two sets of distributions C_G and I_G unless we resort to Non-Shannon type inequalities.

4.8 On the potential INTERESTINGNESS of the remaining 3 mDAGs

The 3 as-yet unresolved mDAGs with 4 observed nodes are depicted in Table 4.2. For these 3 mDAGs we could not find *any* incompatible supports [46], at least up to the small cardinalities of the observed nodes that we checked. Searching for incompatible supports at higher cardinalities of the observed nodes using Fraser's algorithm is computationally expensive, as the algorithm's complexity increases significantly on increasing the cardinalities. Perhaps future acceleration of Fraser's algorithm may allow us to probe supports for higher cardinalities. For the present work, however, we considered one final attempt to prove the INTERESTINGNESS of these 3 mDAGs, namely, by exploring entropic inequalities. For a more comprehensive introduction to entropic inequalities and Shannon cones see Refs. [104, 63].

Also in particular the complete characterization of $\Gamma_N(C_G)$ over only the visible nodes was computationally not feasible due to the Fourier-Motzkin algorithm generating redundant inequalities in the process at a doubly exponential rate [38]. Hence isolating some Shannon-type inequalities that constrain $\Gamma_N(C_G)$ and thus implicitly $\mathcal{C}_G$, but are not Shannon-type inequalities for $\Gamma_N(I_G)$ and thus not implicitly constraining $\mathcal{I}_G$ was computationally intractable.

We can nevertheless certify that the Shannon cones corresponding to $\mathcal{C}_G$ and $\mathcal{I}_G$ are indistinguishable (i.e. $\Gamma_N(C_G)$= $\Gamma_N(I_G)$) for these 3 remaining mDAGs *without* explicitly constructing the Shannon cone for $\mathcal{C}_G$. We do so as follows:

1. From all the Shannon type inequalities corresponding to $\Gamma_N(I_G)$ (and implicitly to $\mathcal{I}_G$), generate the extremal rays of this cone. Since it is a cone it will have one vertex and other extremal rays.
2. Check whether each of the extremal rays and the vertex are such that they are are implied by all the Shannon-type inequalities over all the variables (thus not using Fourier-Motzkin elimination) corresponding to $\mathcal{C}$ (i.e. Shannon type inequalities constraining $\Gamma_N(C_G)$). If so is the case then the Shannon cones corresponding to $\mathcal{C}$ and $\mathcal{I}$ are the same (i.e $\Gamma_N(C_G)$= $\Gamma_N(I_G)$). If one extremal ray is found which is not implied by the Shannon type inequalities corresponding $\mathcal{C}$ then the Shannon cone corresponding to $\mathcal{C}$ is smaller than the Shannon cone corresponding to $\mathcal{I}$ (i.e. $\Gamma_N(C_G) \subset \Gamma_N(I_G)$).

For the 3 mDAGs in Table 4.2 we find that the Shannon cones corresponding to $\mathcal{C}_G$ and $\mathcal{I}_G$ are the same. That is, we were unable to find any valid Shannon type inequality for $\mathcal{C}_G$ that is not also a Shannon type inequality for $\mathcal{I}_G$ for these 3 mDAGs. Thus, entropic methods are incapable of proving the INTERESTINGNESS of these 3 mDAGs, unless perhaps we explore non-Shannon-type inequalities or entropic inequalities involving non-Shannon entropies. It is interesting to note that the only scenario of up-to seven total nodes apart from these 3 left mDAGs for which there is no difference between the Shannon cones corresponding to $\mathcal{C}$ and $\mathcal{I}$ is the Bell DAG of Figure 3.1.

4.9 Conclusions

In this work, we contributed to causal investigation by categorizing which causal structures of 4 observed nodes present inequality constraints or not. To do so, we developed a plethora of techniques to prove that a causal structure is INTERESTING (has inequality constraints), while we used one single technique to prove that a causal structure is NON-INTERESTING: the HLP criterion.

As can be seen from Table 4.3, out of the 2809 mDAGs with 4 observed nodes, the HLP criterion shows that that 1813 are NON-INTERESTING. Out of the remaining 996 mDAGs, our techniques showed 993 of them to be INTERESTING, while we are still uncertain about the status of 3 mDAGs (presented in Table 4.2). While these 3 remaining mDAGs are still potential counter-examples to the HLP conjecture (which says that the HLP criterion is necessary and sufficient for NON-INTERESTINGNESS), we believe that our numerical results are a hint towards the validity of this conjecture. A truly thorough analysis of all mDAGs with 5 observed nodes proved to be quite computationally demanding. Nevertheless, in Appendix A.6 we show that — among those 5-node mDAGs which the HLP criterion fails to certify as INTERESTING— *at least* 99% are INTERESTING, which we again elect to interpret as at least *consistent* with the HLP conjecture.

Category	mDAGs with **3** observed nodes	mDAGs with **4** observed nodes
Total Count	46	2809
remaining # for which the HLP criterion does not apply	5	996
remaining # for which our nonmaximality condition does not apply	1	186
remaining # for which our setwise nonmaximality condition does not apply	0	78
remaining # for which our d-separation condition does not apply	0	60
remaining # not resolved by our use of Fraser's algorithm	0	3

Table 4.3: A summary of our findings: apart from the HLP criterion, which shows NON-INTERESTINGNESS, all of the other conditions listed show INTERESTINGNESS. Note that here we are counting by *unlabelled* DAGs. That is, two labelled DAGs which are equivalent under a relabelling of the observed nodes and/or a relabelling of the hidden nodes are represented by a single unlabelled DAG in these enumerations.

It is also interesting to note that all of our techniques to show INTERESTINGNESS give explicit constructions of distributions which are in $\mathcal{I}_G$ but not in $\mathcal{C}_G$, i.e., respect the conditional independence constraints of DAG G but not its inequality constraints. Theorem 33 is related to the construction of Eq. 4.1, as well as Theorem 42 is related to the construction of Eq. 4.2. Theorem 56 is constructive whenever there *is* a latent-free DAG H with the same set of d-separation relations as G (such construction can then be found in the proof of Lemma 55). If G is maximal but there is *no* latent-free DAG H with the same set of d-separation relations as G, then Theorem 79 says that G has one of the eighteen DAGs of Figure A.2 as a subgraph. In the end of Appendix A.4, an explicit distribution which is in $\mathcal{I}_G$ but not in $\mathcal{C}_G$ for these eighteen DAGs is presented: it is the uniform distribution over the events in the Popescu-Rohrlich support presented in Eq. (4.6).

In particular, our Theorem 33 which showed itself to be very powerful in proving INTERESTINGNESS, is a corrected version of the e-separation theorem of [81] (as discussed in Appendix A.2).

By showing practical tools to attest that a causal structure presents inequality constraints, this work simultaneously contributes to purely classical causal inference and advances the question of which causal scenarios might exhibit quantum or post-quantum advantage.

— 5 —

Causal Scenarios which cannot support Quantum Correlations without Fine-Tuning

5.1 Introduction

Bell's eponymous theorem implies that Quantum Mechanics is incompatible with local causality. This leads to tension between Relativity and Quantum Theory because certain Quantum-Correlations cannot be explained Locally. Non-Classicality of such Quantum-Correlations is often realized to arise because of their Non-Local nature [66]. Making use of the framework of classical causality theory and causal discovery algorithms, Wood and Spekkens have shown that for the Bell scenario such Non-Classical correlations cannot be explained causally without resorting to Fine-Tuning. Here we follow Wood and Spekkens and use the framework of classical causality to study the Non-Classicality of such Quantum-Correlations for other arbitrary scenarios. To get to our results we apply the IC^* algorithm to observed conditional independences corresponding to certain causal scenarios and show that there are several other causal scenarios apart from the scenario that Bell analysed which possess Non-Classical correlations that need not necessarily be Non-Local, but their causal explanation necessarily requires Fine-Tuning. We show that being Fine-Tuned is another interesting feature of the Non-Classicality of such Quantum-Correlations apart from just the fact that they could be Non-Local. Moreover, we show that the Bi-locality scenario, the GHZ scenario are also examples of such scenarios. Further, we put forward the conjecture that the general multi-partite Bell scenarios with n independent parties are also an example of such scenarios. On the other hand for a number of scenarios, which exhibit Non-Classically explainable Quantum-Correlations, we show that we can always explain such Non-Classical Quantum-Correlations by resorting instead to another causal scenario perfectly Classically. This work also validates their observation that causal discovery algorithms that try to reproduce a causal hypothesis for a given set of data must also look into properties of the data other than just the observed conditional independences present in it. A candidate for such a property can be the strength of correlations that are allowed to be possible. Since Causal Discovery finds applications in Machine Learning and Artificial Intelligence as well, our work could be potentially useful in these fields as well.

Studying causal relations- i.e., what causes what and how extensively, has been of interest since centuries. Nevertheless this study of causal relations- causality, has been shrouded with confusion since then [78]. Only recently has the concept of causality been given a rigorous mathematical framework [76, 95, 59]. Causal relations unlike correlations are asymmetric dependencies of some variables on others. For example certain negative behavioural changes in an individual can be correlated with depressing environmental factors but it is the environmental factors that *cause* the behavioural changes and not the other way round. Causal relations between variables can be understood by intervening on them unlike correlations which do not require any external interventions.

Causal discovery is the study responsible for generating possible causal models representing the properties corresponding to some observed data, usually conditional independences. Causal discovery algorithms are responsible to answer this problem. They return possible causal models corresponding to the particular observed conditional independences in the data. Causal discovery algorithms constitute a major active area of research in the fields of Machine Learning and Artificial Intelligence. There are several causal discovery algorithms like IC^*, IC (Inductive Causation) [76, 80, 100], PC (Peter-Clarke) [95], CI (Causal Inference) [39]. However, in this chapter we use only the IC^* algorithm to study several conditional independences exhibited by certain causal scenarios.

Any causal hypothesis about some cause and its effects needs a strong experimental backing. This can be done by studying the probability distributions obtained in an actual realization of such an experiment. These probability distributions might possess certain conditional independences and correlations which we would like to be explained by our hypothesis causally. If our hypothesis is able to explain such conditional independences and correlations consistently then indeed our understanding of what is causing the observed effect is along the right direction. Often though, we want to test our belief in a certain causal scenario by asking questions such as - "What are the possible causal scenarios that can possess the particular set of conditional independences observed in the experiment ?" . In answering such questions we fall back on the causal discovery algorithms such as the IC^*.

Asking these questions for certain scenarios is the subject of this chapter. We consider certain causal scenarios for which we know that d-separation [37] relations alone are certainly not the only constraints restricting the set of probability distributions that are Markov (compatible) with respect to them. Such scenarios have been termed as INTERESTING previously in the literature [44]. Analogously, causal scenarios for which d-separation relations alone are the only constraints restricting the set of probability distributions that are Markov (compatible) with respect to them have been termed as NON-INTERESTING [44]. The INTERESTING scenarios are the potential candidates for exhibiting NON-CLASSICAL correlations. The motivation here is to check whether all the INTERESTING scenarios are the only possible scenarios that can model the corresponding observed conditional independences without any FINE-TUNING or are there some NON-INTERESTING causal scenarios that can also potentially explain the respective conditional independences without resorting to FINE-TUNING. If so, can the QUANTUM-CORRELATIONS[1] which are not CLASSICALLY explainable in an INTERESTING scenario be actually explainable CLASSICALLY using another NON-INTERESTING scenario which gives rise to the same observed conditional independences as the INTERESTING one?

We input the observed conditional independences corresponding to these INTERESTING causal scenarios into the IC^* algorithm. Since conditional independences are the only required input that the algorithm needs, we cannot test those INTERESTING causal scenarios which do not have any conditional independences because for all such scenarios (INTERESTING or NON-INTERESTING), the IC^* algorithm will just return a completely connected graph. For many of the scenarios that we test the algorithm returns that the observed conditional independences can be modelled by both an INTERESTING and a NON-INTERESTING scenario. These INTERESTING scenarios are shown by the researchers at the Perimeter Institute and us [56] to support QUANTUM-CORRELATIONS which cannot be modelled in any CLASSICAL if we restrict to explaining them using the INTERESTING scenario itself. But as we show, we can model such QUANTUM-CORRELATIONS in a perfectly CLASSICAL way by changing our causal explanation and resorting to the NON-INTERESTING scenario being the correct causal model. This is because the NON-INTERESTING scenario can also model the observed conditional independneces without FINE-TUNING. In Section A.7.2 we provide explicit examples of such QUANTUM-CORRELATIONS.

Nevertheless, for some scenarios, the algorithm returns that the observed conditional independences can be explained by only an INTERESTING causal scenario. Any QUANTUM-CORRELATIONS that are not CLASSICALLY explainable using these INTERESTING scenarios will also not be CLASSICALLY explainable by *any other causal scenario*. Amongst examples of such scenarios we find the Bi-locality scenario of entanglement swapping and the GHZ scenario. Building on from the GHZ scenario, we also put forward a conjecture, conjecture 68, stating that scenarios with an arbitrary number of independent parties, i.e., *multi-partite Bell scenarios*, are also examples of such scenarios.

Thus, certain correlations corresponding to these causal scenarios cannot be explained using classical causality or equivalently using any CLASSICAL theory without resorting to FINE-TUNING. The important thing to note is that such NON-CLASSICAL correlations need not be NON-LOCAL, their explanation just requires a particular FINE-TUNING. And thus being FINE-TUNED is another important characteristic of these NON-CLASSICAL correlations apart from them being potentially NON-LOCAL.

[1] Throughout this chapter we consider only those QUANTUM-CORRELATIONS which have the only observed conditional independences that are generated from the observed d-separation relations in the causal scenario.

Previously the same has been studied for the INTERESTING Bell scenario [108] by Wood and Spekkens. They showed that all causal explanations of Bell inequality violating correlations require FINE-TUNING. Causal discovery algorithms cannot differentiate between Bell inequality violating and Bell inequality respecting correlations. The reason being the causal discovery algorithms take only the conditional independences as inputs so necessarily they cannot distinguish between probability distributions possessing different possible strengths of correlations. That the strength of correlations observed in an experiment also plays an important role in our causal explanation of it is another of their conclusions.

Hence, there are scenarios beyond the Bell scenario which not only can just support NON-CLASSICAL correlations, but they are the only possible scenarios which can explain the corresponding experimentally observed conditional independences. That is, modelling these particular conditional independences will necessarily lead to NON-CLASSICAL correlations. On the other hand, there also exist scenarios which support NON-CLASSICAL correlations, but these NON-CLASSICAL correlations can be modelled CLASSICALLY by explaining them using a different (NON-INTERESTING) scenario.

5.2 Non-Classical Correlations

Correlation does not imply causation but under the assumption of NO-FINE-TUNING causation does necessarily imply correlation. As shown in Figure 2.7, if A causes C, then it is not necessary that A and C will be correlated. Only if we assume NO-FINE-TUNING then a correlation between A and C is guaranteed if A causes C. On the other hand, any correlations we observe in a set of data[2] between two random variables A and B implies:

1. Either A is the cause of B
2. Or B is the cause of A
3. Or A and B both share a common cause in the past.

The implications are not all mutually exhaustive. A can be a cause of B and B can be a cause of A; cyclic causation. Implication three is Reichenbach's common cause principle [87] which says that if A and B are correlated and none is the cause of the other, then the correlation between the two is due to some unobserved common cause which affects both A and B. Any correlation observed between two variables must be explainable via the above three possibilities. This is the basic assumption of classical causality theory.

[2] Obtaining how the two random variables A and B are correlated is a different question and here we assume that if correlation is observed between A and B then it is not coincidental because coincidental correlations in a data set can be removed by performing the experiment a large number of times.

Explaining QUANTUM-CORRELATIONS causally has been a program since Einstein, Podolski and Rosen [25] who showed that quantum mechanics on its own does not provide a LOCAL explanation of certain QUANTUM-CORRELATIONS (EPR correlations). Further, it has been shown that these EPR correlations can be explained locally by adding hidden variables to quantum mechanics (for example refer to Chapter 8 in [72]) . The notion of LOCALITY here is that any cause of any effect should be in the past light cone of the effect i.e., causal influence must propagate at a speed less than or equal to the speed of light. However, soon Bell [12] was able to show that any theory reproducing the results of quantum theory would be necessarily NON-LOCAL. He showed this by showing that certain quantum mechanical correlations (Bell correlations) cannot be explained locally, or that the causes of certain two observed correlated events, space-like separated, do not all necessarily lie in their past light cones. A pedagogical review of these results can be found in [70, 69, 74, 73, 72, 40]. The NON-LOCAL [45] nature of these QUANTUM-CORRELATIONS (Bell correlations) is often interpreted to be the reason behind them being NON-CLASSICAL. Using the language of DAGs and Bayesian networks introduced in the previous chapters we now explain what are non-classical correlations.

The causal scenario Bell considered is shown in Figure 5.1. Λ is the complete description (including any hidden variables) of all the real things that the theory under consideration posits to be "beables" [58, 11][3] and which could influence the experiment under consideration. Λ may or may not include the quantum state depending on whether our theory posits the quantum state to be real or not. The important thing is that Λ is a complete description of all the things in the past light cone of the experiment.[4] A maximally entangled state is understood to be distributed to two labs which are space like separated. The measurement settings for the labs are denoted by X and Y, respectively, and the corresponding outputs of the labs are A and B. Since the labs are space-like separated, Bell formulated the condition of LOCALITY we have described above in a mathematically rigorous way by the following factorization of the joint probability distribution over the outcomes:

$$P(A,B|X,Y,\Lambda) = P(A|X,\Lambda)P(B|Y,\Lambda) \tag{5.1}$$

Equation 5.1 means that the probability of the outcome A is independent of the outcome B and the setting of B, i.e Y, once the setting of A, i.e X, and the complete description of anything in the past light cone of A, i.e Λ is known. A similar condition holds for the outcome B. Under the assumption of measurement independence

$$P(X,Y,\Lambda) = P(X,Y|\Lambda)P(\Lambda) = P(X,Y)P(\Lambda) \tag{5.2}$$

[3] For the definition of beables refer to [58, 11]

[4] Except the measurement settings which are assumed to be chosen randomly.

and Equation 5.1, Bell derived an inequality which he showed was violated by certain correlations produced by using a maximally entangled bi-partite quantum state; thus showing that certain QUANTUM-CORRELATIONS do not obey the above (Equation 5.1) notion of locality and are hence NON-LOCAL.

But note that Bell's inequality would still be violated by these QUANTUM-CORRELATIONS even if the labs were not space-like separated. This is because even if the labs are not space-like separated but we carry out the same experiment using the same apparatus as compared to when the labs were space-like separated, we can obtain the same probability distribution $P(A, B, X, Y)$ that we obtained when the separation between the labs and the settings was space-like. We can obtain the same probability distribution $P(A, B, X, Y)$ as before for two simple reasons:

1. The outcomes A and B do not change with the distance between the labs.

2. If the labs were not communicating when they were space-like separated, we expect them to not communicate even when they are time-like separated. Just because the labs are now time-like separated does not imply that they should necessarily communicate. Especially, if we use the same labs and the same experimental setups in both the experiments, we expect the same no-communication condition between the labs.

Hence Equation 5.1 which along with measurement independence leads to Bell's inequalities would still hold true and thus $P(A, B, X, Y)$ would still violate the Bell inequalities even even the labs are time-like separated. Such a causal set up is shown in Figure 5.2[5] Nothing NON-LOCAL is going on here and yet the Bell inequalities are violated. If in this setup we call the Bell inequality violating correlations as NON-CLASSICAL, it is evident that these NON-CLASSICAL correlations have characteristic properties beyond just NON-LOCALITY.

On the other hand if a priory we see such correlations when the labs are time-like separated we can explain them LOCALLY by positing for example that the labs signalled to each other because now they could. But the important thing is that they would still violate the Bell inequality. Wood and Spekkens [108] have showed that all the explanations of Bell inequality violating NON-LOCAL QUANTUM-CORRELATIONS are FINE-TUNED. But the same is true for these LOCALLY explainable Bell inequality violating QUANTUM-CORRELATIONS; they are FINE-TUNED as well (that is why they will violate the Bell inequality). Figure 5.3 shows a FINE-TUNED explanation of these LOCAL Bell inequality violating QUANTUM-CORRELATIONS (which are non-classical nonetheless). Together these LOCAL and NON-LOCAL Bell inequality violating QUANTUM-CORRELATIONS form a set of Bell inequality violating correlations that we call NON-CLASSICAL because they cannot be explained using the framework

[5] In the Figures 5.1 and 5.2, it is assumed that Λ does not influence X and Y and that the measurement settings are chosen randomly.

of Classical Causality theory (that is both will require some kind of fine tuning for their causal explanation).

The important takeaway is that NON-LOCALITY is not the only characteristic property of these NON-CLASSICAL correlations. That they are FINE-TUNED is another important property that these NON-CLASSICAL correlation posses even if they are not NON-LOCAL. We show that the same is the case for other scenarios apart from the Bell scenario in Section A.7.2. They possess NON-CLASSICAL correlations which cannot be explained classically without resorting to FINE-TUNING, irrespective of whether measurements are done over space-like separated or time-like separated labs (and thus irrespective of whether the correlations are LOCAL or NON-LOCAL in nature).

Just like the NON-LOCALITY of these NON-CLASSICAL correlations creates a conflict between Relativity and Quantum Mechanics, similarly the fact that their causal explanation requires FINE-TUNING creates a tension between Quantum Mechanics and Classical Causality theory (even if they are local and hence do not create a conflict between Relativity and Quantum Mechanics).

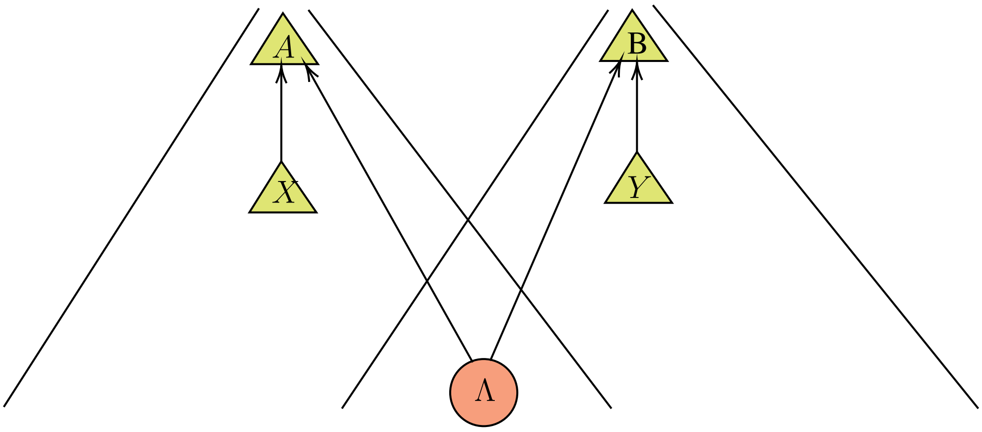


Figure 5.1: Bell DAG embedded in space-time where a NON-CLASSICAL $P(A, B, X, Y)$ is NON-LOCAL and FINE-TUNED. $P(A, B, X, Y)$ is NON-CLASSICAL because it can violate Bell's inequality. It is NON-LOCAL because two space-like separated nodes are correlated and the correlation is not coming completely from their common past light cone but from somewhere outside their past light cones. It is FINE-TUNED because any causal model that can explain it will require some kind of FINE-TUNING.

5.3 Results

We consider some causal scenarios that have already been shown to be capable of possessing NON-CLASSICAL correlations (and are hence called INTERESTING) They have also been shown to definitely support certain QUANTUM-CORRELATIONS that cannot be explained in any CLASSICAL way without

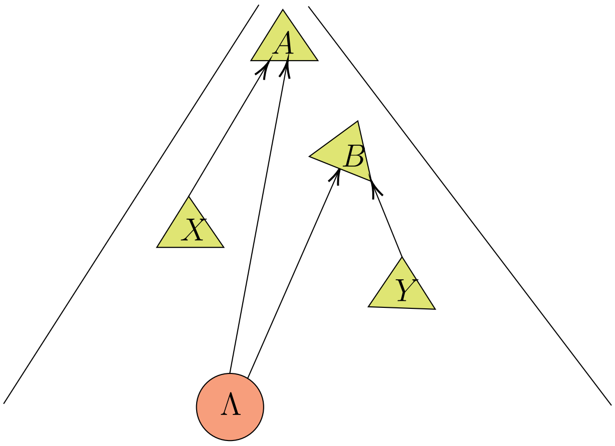


Figure 5.2: Bell DAG embedded in space-time where a NON-CLASSICAL $P(A, B, X, Y)$ is LOCAL but FINE-TUNED. $P(A, B, X, Y)$ is NON-CLASSICAL because it can again violate a Bell-inequality. It is LOCAL because both the outcomes and settings of the opposite parties are time-like separated and so all causal influences can propagate subluminally. It is FINE-TUNED because any causal model that can explain it will require some kind of FINE-TUNING.

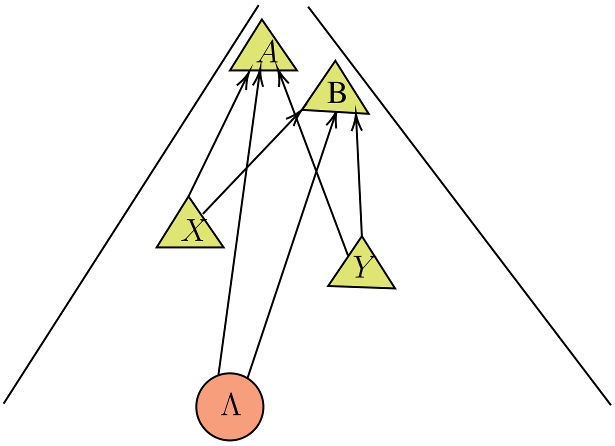


Figure 5.3: A FINE-TUNED and LOCAL explanation of NON-CLASSICAL $P(A, B, X, Y)$ Bell-inequality violating QUANTUM-CORRELATIONS. $P(A, B, X, Y)$ can again be NON-CLASSICAL (in the Bell scenario) because it can violate Bell's inequalities. Though it is local because all the nodes are time-like separated and thus support subluminal signals between them. It is FINE-TUNED since corresponding to the observed conditional independence $B \perp\!\!\!\perp_{\text{CI}} X|Y$ there is no corresponding d-separation relation $B \perp_{\text{d}} X|Y$. Via this unfaithful scenario, de-Broglie-Bohmian mechanics can explain Bell inequality violating LOCAL or NON-LOCAL QUANTUM-CORRELATIONS. Note that $P(A, B, X, Y)$ is not NON-CLASSICAL in the DAG of Figure 5.3, it is just FINE-TUNED.

resorting to Fine-Tuning in an upcoming work by the researchers at the Perimeter Institute. Now in an actual experiment we will observe some conditional independences and correlations. So, given just these observed conditional independences and correlations that we find in the experiment, we would like to identify all the possible causal scenarios can explain them. In particular, we would like to check if along with an Interesting scenario there also exists a Non-Interesting scenario that can explain those observed conditional independences or not. If there exists such a Non-Interesting scenario as well, then we can very well explain those *experimentally observed* conditional independences Classically and the existence of such conditional independences does not imply the existence of Non-Classical correlations. This is because in such a case, we can always explain any Non-Classical correlations in the Interesting scenario Classically in the Non-Interesting scenario supporting the same observed conditional independences. This is proved in Theorem 64.

On the other hand if we find that an Interesting scenario is the only possible scenario that can explain the corresponding *experimentally observed* conditional independences then the existence of Non-Classical correlations is *directly implied just by the observation of some conditional independences*. Further, if the scenario supports some Quantum-Correlations that are Non-Classical then there would exist no other scenario that can support those particular Quantum-Correlations in any Classical way without resorting to Fine-Tuning. Hence, no Classical model would be able to explain such experimentally realized Quantum-Correlations causally. These scenarios are thus *"genuinely interesting"*. Note that as explained in Section 5.2, these observed Non-Classical correlations (including the Quantum-Correlations) need not be necessarily Non-Local. Irrespective of whether these observed correlations can or cannot be explained Locally, they are Non-Classical in the sense that classical causal modelling cannot explain them without Fine-Tuning, an assumption very central to the field.

We find several examples of such scenarios. For example, the Bi-locality scenario that represents the correlations obtained in an entanglement swapping scenario is the only scenario that classical causal modelling says can explain all the corresponding observed conditional independences faithfully. The same is true for the GHZ scenario. Just like the Bell scenario, they support both Non-Local and Local Non-Classical correlations which require Fine-Tuning to be explained Classically. Thus, there are scenarios beyond just the Bell scenario that classical causal theory cannot explain faithfully. In-fact these scenarios highlight the incompleteness of causal modelling.

In Tables 5.1, 5.2 and 5.3 we show the causal scenarios whose corresponding observed conditional independences we input in the IC^* algorithm and the corresponding pattern returned by it. In all the causal scenarios triangular nodes represent observed variables and the circular nodes represent the latent variables. First we prove an important theorem.

Theorem 64. *Let G be an INTERESTING causal scenario that exhibits some QUANTUM-CORRELATIONS that are NON-CLASSICAL. Let G also exhibit certain observed conditional independences. Let H be another NON-INTERESTING scenario free of any latent variables that exhibits exactly the same observed conditional independences as G. Then H can support* ***all*** *the QUANTUM-CORRELATIONS that were NON-CLASSICAL in G in a perfectly CLASSICAL way without resorting to any FINE-TUNING.*

Proof. Let $\mathcal{C}_{\mathcal{G}}$, $\mathcal{C}_{\mathcal{H}}$ be the set of all correlations that are CLASSICALLY explainable in G and H respectively. Similarly, let $\mathcal{Q}_{\mathcal{G}}$, $\mathcal{Q}_{\mathcal{H}}$ be the set of all correlations that are explainable using Quantum theory in G and H, respectively. Finally, let $\mathcal{I}_{\mathcal{G}}$, $\mathcal{I}_{\mathcal{H}}$ be the set of all correlations that respect all the observed conditional independences in G and H, respectively. Since G and H have exactly the same observed conditional independences, we have that $\mathcal{I}_{\mathcal{G}} = \mathcal{I}_{\mathcal{H}}$. Also from [44] we know the following relations between the above sets of correlations:

$$\mathcal{C}_{\mathcal{G}} \subseteq \mathcal{Q}_{\mathcal{G}} \subseteq \mathcal{I}_{\mathcal{G}} \tag{5.3}$$

and

$$\mathcal{C}_{\mathcal{H}} = \mathcal{Q}_{\mathcal{H}} = \mathcal{I}_{\mathcal{H}} = \mathcal{I}_{\mathcal{G}} \tag{5.4}$$

Using Equations 5.3 and 5.4 together we have that

$$\mathcal{Q}_{\mathcal{G}} \subseteq \mathcal{I}_{\mathcal{G}} = \mathcal{I}_{\mathcal{H}} = \mathcal{C}_{\mathcal{H}} \tag{5.5}$$

thus showing that any QUANTUM-CORRELATIONS in G are CLASSICAL in H and that all NON-CLASSICAL correlations in G can be explained CLASSICALLY in H without resorting to FINE-TUNING. □

Causal scenarios (1), (3) and (6) in Tables 5.1, 5.2 can be modelled by the latent free scenario of Figure 5.4. Scenario (5) in Table 5.2 can also be modelled by this same latent free scenario with the simple re-lablelling $D \mapsto F, F \mapsto E, E \mapsto D$. Similarly, scenario (2) in Table 5.1 can be modelled by the latent free scenario of Figure 5.5. Thus, by Theorem 64 any NON-CLASSICAL correlations in causal scenarios (1), (2), (3), (5), (6) can be explained CLASSICALLY by using the latent free scenarios of Figures 5.4 and 5.5 respectively. In an upcoming work by the researchers at the Perimeter Institute and us [56], it is shown that all these scenarios also support QUANTUM-CORRELATIONS that cannot be explained CLASSICALLY using these scenarios. But again by resorting to Theorem 64, we know that such QUANTUM-CORRELATIONS can always be modelled CLASSICALLY by the respective latent free scenarios of Figures 5.4 and 5.5.

Next consider the 4th scenario in Table 5.2. The IC^* algorithm says that the conditional independences corresponding to this scenario can be produced by only one graph which necessarily needs to have two latent variables between $E \longleftrightarrow D$, and $F \longleftrightarrow D$ respectively. This is because any

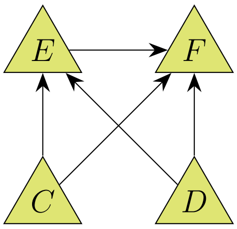


Figure 5.4: Latent free casual scenario reproducing all the conditional independendces corresponding to scenarios (1), (3) and (6) in Tables 5.1, 5.2.[6]

latent common causes between nodes A, E and nodes A, F can be absorbed into node A without any loss of generality. Hence, the 4th scenario in Table 5.1 is the only possible scenario that produces the corresponding observed conditional independences. A close look at the scenario shows that it is similar to the Bi-locality scenario with the exception that the extreme wings share the same settings.

For the 7th scenario in Table 5.2 notice that from all the many scenarios resulting from the pattern corresponding to it, only those preserve the respective conditional independences which respect the following:

1. Given the pattern returned, for $C \perp\!\!\!\perp_{\text{CI}} F$ to be true there cannot be a directed edge $E \rightarrow F$. So there can be an edge $F \rightarrow E$, or $E - F$ can share a common cause or there can be both an edge $F \rightarrow E$ and a common cause between $E - F$.
2. But then for $C \perp\!\!\!\perp_{\text{CI}} E|D$ to hold true there can be no collider at D on an undirected path $C - D - E$, unless there is another collider somewhere along the undirected path $C - D - E$ that makes $C \perp_{\text{d}} E|D$ true. Thus for $C \perp\!\!\!\perp_{\text{CI}} E|D$ to hold true there can be no directed edges $F \rightarrow D$ and $F \rightarrow E$, but the only possibility is that $D - F$ share a common cause and $E - F$ share another common cause only.
3. For the same reason as the last point, there can be no common cause of $D - E$, else $C \perp\!\!\!\perp_{\text{CI}} E|D$ will not hold true.
4. $C - D$ can share a common cause or a directed edge $C \rightarrow D$, or both the edge and the common cause. Without loss of generality we can always absorb the common cause of $C - D$ into C.

A close inspection reveals that there is only one causal scenario resulting from the pattern returned which respects the above conditions and the corresponding observed conditional independences and it is indeed the 7th one in Table 5.2.

Now consider the Bi-locality scenario which is 8th in Table 5.2. Inputting its conditional independences in the algorithm returns the corresponding pattern for which edges $G \longleftrightarrow D$ and

[6] Inverting the direction of the edge $E \rightarrow F$ would also result in a valid scenario.

[7] Similarly, inverting the direction of the edge $D \rightarrow F$ would also result in a valid scenario.

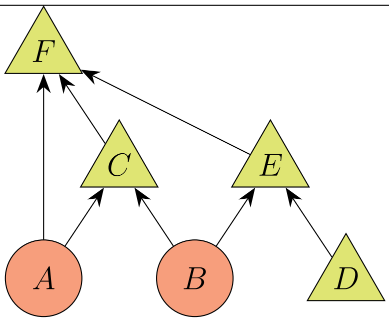
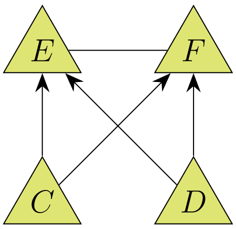
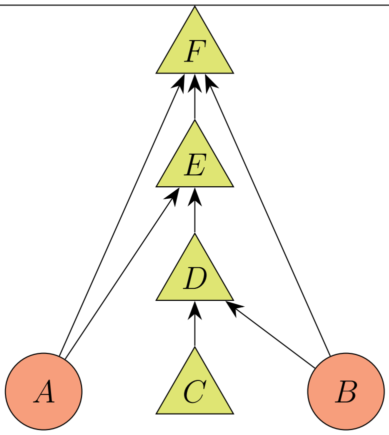
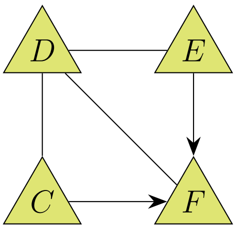
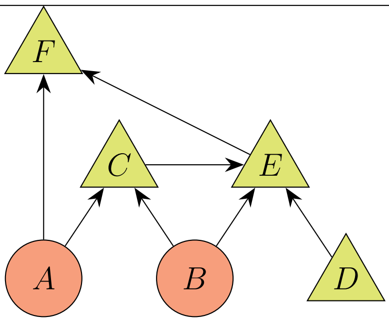
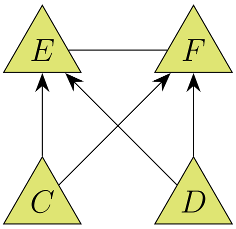
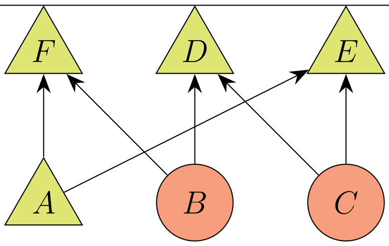
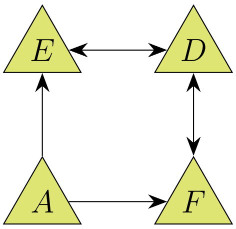
	Causal Scenarios possessing certain observed conditional independences	The pattern obtained by inputting the corresponding observed conditional independences in the IC^* algorithm
(1)	 Conditional Independences = $\{C \perp\!\!\!\perp_{\text{CI}} D\}$	
(2)	 Conditional Independences = $\{C \perp\!\!\!\perp_{\text{CI}} E \mid D\}$	
(3)	 Conditional Independences = $\{C \perp\!\!\!\perp_{\text{CI}} D\}$	
(4)	 Conditional Independences = $\{A \perp\!\!\!\perp_{\text{CI}} D, E \perp\!\!\!\perp_{\text{CI}} F \mid A\}$	

Table 5.1: Causal scenarios and their corresponding patterns returned by the IC^* algorithm.

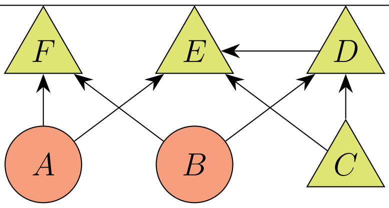
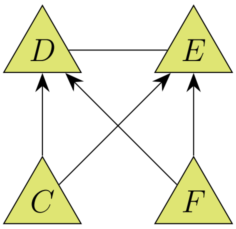
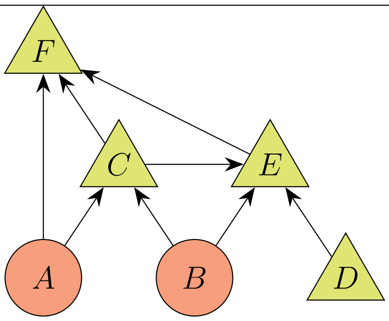
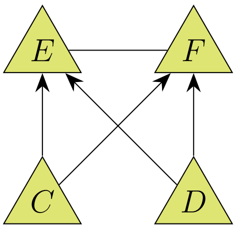
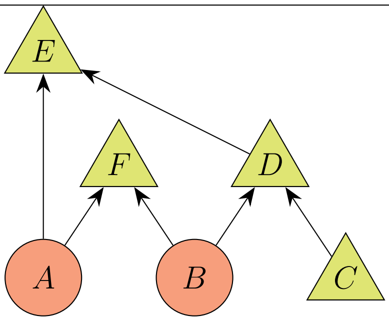
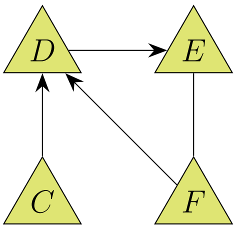
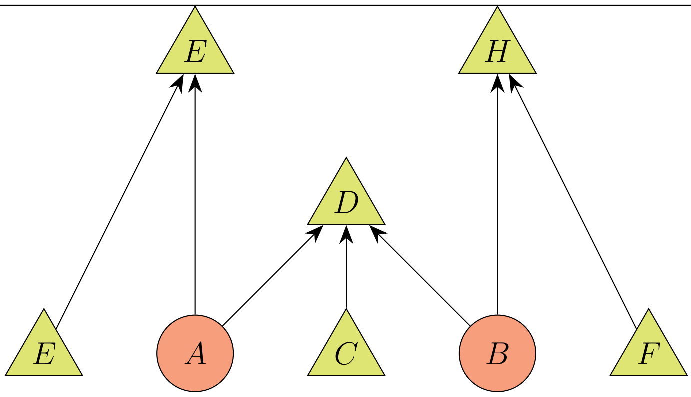
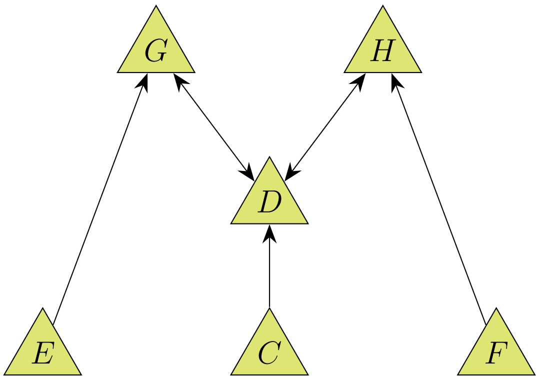
Causal Scenarios possessing certain observed conditional independences	The pattern obtained by inputting the corresponding observed conditional independences in the IC^* algorithm
(5)  Conditional Independences = $\{C \perp\!\!\!\perp_{\mathrm{CI}} F\}$	
(6)  Conditional Independences = $\{C \perp\!\!\!\perp_{\mathrm{CI}} D\}$	
(7)  Conditional Independences = $\{C \perp\!\!\!\perp_{\mathrm{CI}} F,$ $C \perp\!\!\!\perp_{\mathrm{CI}} E \mid D\}$	
(8)  Conditional Independences = $\{E \perp\!\!\!\perp_{\mathrm{CI}} C,$ $F \perp\!\!\!\perp_{\mathrm{CI}} C, G \perp\!\!\!\perp_{\mathrm{CI}} C, E \perp\!\!\!\perp_{\mathrm{CI}} D, F \perp\!\!\!\perp_{\mathrm{CI}} D,$ $E \perp\!\!\!\perp_{\mathrm{CI}} F, G \perp\!\!\!\perp_{\mathrm{CI}} F\ \}$	

Table 5.2: Causal scenarios and their corresponding patterns returned by the IC^* algorithm.

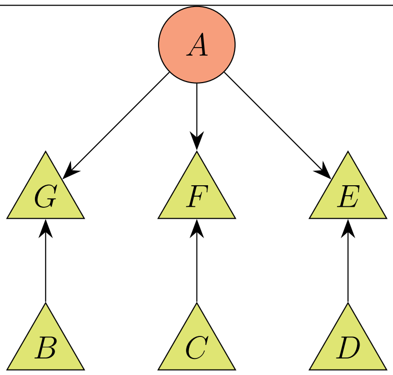
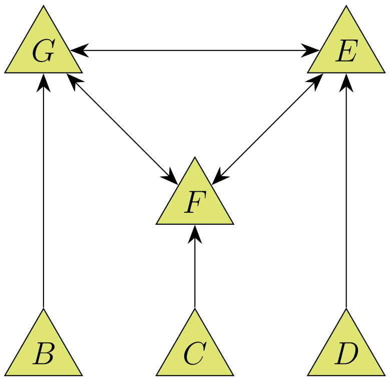
Causal Scenarios possessing certain observed conditional independences	The pattern obtained by inputting the corresponding observed conditional independences in the IC^* algorithm
(9)  Conditional Independences = {$C \perp\!\!\!\perp_{\text{CI}} B$, $D \perp\!\!\!\perp_{\text{CI}} B$, $E \perp\!\!\!\perp_{\text{CI}} B, F \perp\!\!\!\perp_{\text{CI}} B, D \perp\!\!\!\perp_{\text{CI}} C, E \perp\!\!\!\perp_{\text{CI}} C, G \perp\!\!\!\perp_{\text{CI}} C$, $F \perp\!\!\!\perp_{\text{CI}} D, G \perp\!\!\!\perp_{\text{CI}} D$	

Table 5.3: Causal scenarios and their corresponding patterns returned by the IC^* algorithm.

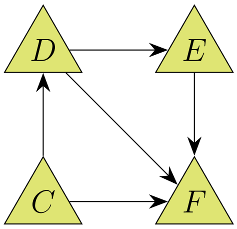


Figure 5.5: Latent free casual scenario reproducing all the conditional independendces corresponding to the scenario (5) in Table 5.2.[7]

$D \longleftrightarrow H$ are bi-directed. Remembering the meaning of a bi-directed edge in a pattern from Section 2.10, we see that the only possible scenario is the Bi-locality one. This is because nodes G and D can share a common cause only without any directed edge between them and the same would be the case for nodes D and H which will share only a common cause as well.[8] So just like the Bell scenario, the Bi-locality scenario is the only faithful scenario which explains the corresponding observed conditional independences.

The conclusion is the same for the GHZ scenario which is 9th in Table 5.3 [9]. It is the only possible scenario that can model the corresponding conditional independences. This is because the bi-directed

[8] For the directed edges between two nodes we can absorb the latent variable into the parent node without any loss of generality.

[9] Note that $C \perp\!\!\!\perp_{\text{CI}} B \Leftrightarrow B \perp\!\!\!\perp_{\text{CI}} C$.

edges at $G \longleftrightarrow F$, $F \longleftrightarrow E$, and $G \longleftrightarrow E$ imply that the corresponding nodes can share only a latent common cause. We conjecture that this result can be generalized to multi-partite Bell scenarios with arbitrary number of parties with respectively independent settings. We state this conjecture as conjecture 68 in the next section and try to prove it.

5.4 FINE-TUNING for multi-partite Bell scenarios

We first introduce the formal definition of a multi-partite Bell scenario.

Definition 65 (Multi-partite Bell scenarios)**.** *Consider n parties with m settings and k outputs, all space-like separated to each other, such that $n > 2, m > 1, k > 1$. We call such a scenario the multi-partite Bell scenario.*

For all the scenarios we have dealt with until now, we have depended on the IC^* algorithm, which worked because the number of conditional independences were not arbitrarily many. But now we consider multi-partite Bell scenarios with an arbitrary number of parties, each one independent of the other. Here we will have arbitrarily many conditional independences and hence an analytical search is needed to find the possible NON-FINE-TUNED scenarios which can support such conditional independences.

Consider a multi-partite Bell scenario with n independent parties. We denote the set of parties as $\mathbb{P} = \{1 \ldots\ldots n \mid n \in \mathbb{N}\}$. The settings and outputs of the respective parties are denoted as $x_1, x_2, x_3, x_4, \ldots\ldots x_n$ and $a_1, a_2, a_3, a_4, \ldots\ldots a_n$ respectively.[10] Thus, $\mathbb{O} = \{a_i \mid i = 1, \ldots\ldots n\}$ is the set of random variables representing all the outputs and $\mathbb{S} = \{x_i \mid i = 1, \ldots\ldots n\}$ is the set of random variables representing all the settings. Here, a_i and x_i correspond to the outcome and the setting of the ith party. The outputs and settings corresponding to any subset $\mathcal{P} \subseteq \mathbb{P}$ of parties is denoted as $\mathcal{O}_{\mathcal{P}} = \{a_i \mid i \in \mathcal{P}\}$ and $\mathcal{S}_{\mathcal{P}} = \{x_i \mid i \in \mathcal{P}\}$ respectively.

Now, we begin with the following assumptions for our conjecture:

1. **Quantum Correlations and Observed Conditionl Independences:** On grounds of relativity, we expect that any subset of the parties cannot signal to any other subset of the parties since they are all space-like separated to each other . That is, once the settings corresponding to any subset of the outputs is given that subset of outputs is independent of all the other settings. We also

[10] Here the small letters x_i, a_i function as the random variable itself. This is different from our usage until now where the capital letters depicted the random variable and small letters depicted the values of the random variables. In Section 5.4, both capital and small letters will denote a random variable.

assume that any subset of the settings are independent of any other disjoint set of settings. Hence, we assume that all the observed quantum correlations and observed probability distributions respect the following conditional independences:

$$\begin{aligned} &(\mathcal{O}_P \perp\!\!\!\perp_{\text{CI}} \mathbb{S} \setminus \mathcal{S}_P \mid \mathcal{S}_P) \quad \forall \ \mathcal{P} \subseteq \mathbb{P} \\ &(No - Signalling\ between\ any\ subsets\ of\ parties) \\ &(\mathcal{S}_P \perp\!\!\!\perp_{\text{CI}} \mathbb{S} \setminus \mathcal{S}_P) \quad \forall \ \mathcal{P} \subseteq \mathbb{P} \\ &(Measurement\ Independence\ between\ any\ subsets\ of\ parties) \end{aligned} \tag{5.6}$$

By the semi-graphoid axioms (in particular the decomposition axiom) we know that the conditional independences in Equation 5.6 also imply the following conditional independences:

$$\begin{aligned} &(a_i \perp\!\!\!\perp_{\text{CI}} x_j \mid x_i) \quad \forall \ i,\ j = 1, \ldots\ldots .n, \ \ j \neq i \\ &(Pairwise\ No - Signalling) \\ &(x_i \perp\!\!\!\perp_{\text{CI}} x_j) \quad \forall \ i,\ j = 1, \ldots\ldots n \ \ j \neq i \\ &(Pairwise\ Measurement\ Independence\ between\ two\ parties) \end{aligned} \tag{5.7}$$

2. **Causal and NON-FINE-TUNED Explanation:** This assumption states that all probability distributions on observed variables must exhibit a NON-FINE-TUNED explanation based on the framework of standard causal modelling. That is there should exist a DAG that is a classically causal faithful explanation of all the observed probability distributions.

Our conjecture 68 is based on the conjunction of the above two assumptions and says that all explanations of Bell inequality violating correlations in any multi-partite Bell scenario can be explained only via some FINE-TUNING. If we obtain some correlations that follow our conjecture 68, then either of the above two assumptions or both are wrong. The way to the proof is to show that between any two parties there will be a Bell scenario and any extra conditional independences either do not change the NON-FINE-TUNED Bell scenario between each pair of parties or change them in such a way that Bell's inequalities are still implied. We begin by introducing two definitions that will be used in the sketch proof of the following conjecture.

Definition 66 (Bell sub-scenarios). *Consider n parties. If any two parties $i, j \in \mathbb{P}$ out of the n parties have their marginal probability distribution factorised as:*

$$P(a_i, a_j, x_i, x_j) = \sum_{\Lambda_{ij}} P(a_i \mid x_i, \Lambda_{ij}) P(a_j \mid x_j, \Lambda_{ij}) P(x_i) P(x_j) P(\Lambda_{ij})^{11} \tag{5.8}$$

with Λ_{ij} being the common cause of the outcomes a_i, a_j, then we say that the two particular parties $i, j \in \mathbb{P}$ out of the n parties under consideration share a Bell sub-scenario.

[11] Note that here the small letters a_i and x_i are the random variables themselves and not their values.

The DAG on all the observed $2n$ nodes (i.e on $(a_1, \ldots\ldots a_n, x_i \ldots\ldots x_n)$) could be any possible DAG such that the marginal probability distribution over the four nodes (a_i, a_j, x_i, x_j) is as in Equation 5.8.

Definition 67 (Cycle DAG over n parties)**.** *Consider the n parties with outcomes $(a_1, \ldots\ldots a_n)$ and respective settings $(x_1, \ldots\ldots x_n)$. Let each pair of party $i, j \in \mathbb{P}$ share a latent common cause Λ_{ij} Now consider the DAG over these $2n$ observed random variables and n latent random variables in which the only edges are $x_i \to a_i$ and $a_i \leftarrow \Lambda_{ij} \to a_j$ $\forall$ $i, j = 1, \ldots\ldots n$ and $i \neq j$. We call such a DAG, the Cycle DAG over the n parties. For $n = 3$, the resulting DAG is shown in Figure 5.6.*

Now, we formally state our conjecture and try to prove it.

Conjecture 68 (FINE-TUNING for multi-partite Bell scenarios)**.** *There are no classically causal faithful explanations of Bell inequality violating correlations in any multi-partite Bell scenario with n independent settings. Or alternatively, any classical causal explanation of such correlations requires FINE-TUNING.*

Way to the proof: First, consider any two particular parties $i, j \in \mathbb{P}$. A few conditional independences involving them are: $(a_i \perp\!\!\!\perp_{\text{CI}} x_j \mid x_i)$, $(a_j \perp\!\!\!\perp_{\text{CI}} x_i \mid x_j)$ and $(x_j \perp\!\!\!\perp_{\text{CI}} x_i)$. If these were the only three conditional independences involving the variables (a_i, a_j, x_i, x_j), then by the result of Wood and Spekkens [108], the only NON-FINE-TUNED causal scenario modelling the marginal probability distributions compatible with the conditional independences $(a_i \perp\!\!\!\perp_{\text{CI}} x_j \mid x_i)$, $(a_j \perp\!\!\!\perp_{\text{CI}} x_i \mid x_j)$ and $(x_j \perp\!\!\!\perp_{\text{CI}} x_i)$ between the parties $i, j \in \mathbb{P}$, would be the Bell scenario, with Λ_{ij} being the common cause of the outcomes of the parties $i, j \in \mathbb{P}$. That is the probability distribution $P(a_i, a_j, x_i, x_j)$ will be modelled as:

$$P(a_i, a_j, x_i, x_j) = \sum_{\Lambda_{ij}} P(a_i \mid x_i, \Lambda_{ij}) P(a_j \mid x_j, \Lambda_{ij}) P(x_i) P(x_j) P(\Lambda_{ij}). \tag{5.9}$$

And thus any two parties $i, j \in \mathbb{P}$ would share a Bell sub-scenario between them. But these are not the only three conditional independences amongst the variables (a_i, a_j, x_i, x_j). There are other conditional independences amongst these same variables that follow from Equation 5.6 which forms the generating set of all the conditional independences that is strictly greater than the set of conditional independences formed by Equation 5.7. The first part of the proof is to recognize that under the assumption of NO-FINE-TUNING the set of conditional independences formed by Equations 5.7 and 5.6 is actually the same. This is because under the assumption of NO-FINE-TUNING the set of all the conditional independences implies the set of all the d-separation relations. Equations 5.7 and 5.6 imply the same set of d-separation relations and thus under the assumption of NO-FINE-TUNING, Equations

5.7 and 5.6 generate the same set of conditional independences.[12] In other words under the assumption of NO-FINE-TUNING we have that:

$$Equation\ \ 5.7 \iff Equation\ \ 5.6 \tag{5.10}$$

and thus the generating set of all the conditional independences is formed by the conditional independences in Equation 5.7 alone.

Therefore, using this fact, we conclude that any two parties $i, j \in \mathbb{P}$ share a Bell sub-scenario between them because the only independent conditional independences between the parties $i, j \in \mathbb{P}$ are in fact just $(a_i \perp\!\!\!\perp_{\text{CI}} x_j \mid x_i)$, $(a_j \perp\!\!\!\perp_{\text{CI}} x_i \mid x_j)$ and $(x_j \perp\!\!\!\perp_{\text{CI}} x_i)$.

Now on considering all the pairs of parties together, one possible scenario between the n parties, or over the $2n$ observed variables $(a_1, \ldots\ldots a_n, x_i \ldots\ldots x_n)$ is just the Cycle DAG over these n parties. This is because if we consider n parties, then the probability distribution over any pair of parties $i, j \in \mathbb{P}$ would just be given by:

$$P(a_i, a_j, x_i, x_j) = \sum_{\mathcal{O}\backslash\{a_i,a_j\}\mathcal{S}\backslash\{x_i,x_j\}} P(a_i, a_j, \ldots\ldots a_n, x_i, x_j, \ldots\ldots x_n) \tag{5.11}$$

$$= \sum_{\mathcal{O}\backslash\{a_i,a_j\}\mathcal{S}\backslash\{x_i,x_j\}} \sum_{\Lambda_{ij},\ldots\ldots\Lambda_{in},\ldots\ldots\Lambda_{jn},\ldots\ldots\Lambda_{n(n-1)}\ldots\ldots} \Big[P(a_i \mid x_i, \Lambda_{ij}, \Lambda_{ik}, \ldots\ldots \Lambda_{in}) P(a_j \mid x_j, \Lambda_{ij}, \Lambda_{jk} \ldots\ldots \Lambda_{jn}) \\ P(a_k \mid x_k, \Lambda_{ik}, \Lambda_{jk}, \ldots\ldots \Lambda_{kn}) \ldots\ldots P(a_n \mid x_n, \Lambda_{in}, \Lambda_{jn}, \Lambda_{kn} \ldots\ldots \Lambda_{n(n-1)}) \\ P(x_i)P(x_j)P(x_k) \ldots\ldots P(x_n) P(\Lambda_{ij}) P(\Lambda_{ik}) P(\Lambda_{jk}) \ldots\ldots P(\Lambda_{in}) P(\Lambda_{jn}) P(\Lambda_{kn}) P(\Lambda_{n(n-1)}) \Big] \tag{5.12}$$

where $P(a_i, a_j, \ldots\ldots a_n, x_i, x_j, \ldots\ldots x_n)$ has been expanded out according the Markov condition in the Cycle DAG over the n parties. Summing over the outcomes and settings of all the rest of the $n-2$ parties except parties $i, j \in \mathbb{P}$ we get:

$$P(a_i, a_j, x_i, x_j) = \sum_{\Lambda_{ij},\ldots\ldots\Lambda_{in},\ldots\ldots\Lambda_{jn},\ldots\ldots\Lambda_{n(n-1)}\ldots\ldots} \Big[P(a_i \mid x_i, \Lambda_{ij}, \Lambda_{ik}, \ldots\ldots \Lambda_{in}) P(a_j \mid x_j, \Lambda_{ij}, \Lambda_{jk} \ldots\ldots \Lambda_{jn}) \\ P(x_i)P(x_j) P(\Lambda_{ij}) P(\Lambda_{ik}) P(\Lambda_{jk}) \ldots\ldots P(\Lambda_{in}) P(\Lambda_{jn}) P(\Lambda_{kn}) P(\Lambda_{n(n-1)}) \Big] \tag{5.13}$$

[12] This is the reason the IC^* algorithm looks only for conditional independences between two nodes given a set of nodes, rather than looking for all the possible conditional independences between sets of nodes.

$$\implies P(a_i,a_j,x_i,x_j) = \sum_{\Lambda_{ij},\ldots\ldots\Lambda_{in},\ldots\ldots\Lambda_{jn},\ldots\ldots\Lambda_{n(n-1)}\ldots\ldots} \Bigg[\frac{P(a_i,\Lambda_{ik},\ldots\ldots\Lambda_{in} \mid x_i,\Lambda_{ij})}{P(\Lambda_{ik},\ldots\ldots\Lambda_{in} \mid x_i,\Lambda_{ij})} \frac{P(a_j,\Lambda_{jk},\ldots\ldots\Lambda_{jn} \mid x_j,\Lambda_{ij})}{P(\Lambda_{jk},\ldots\ldots\Lambda_{jn} \mid x_j,\Lambda_{ij})} P(\Lambda_{ij}) \\ P(\Lambda_{ik},\ldots\ldots\Lambda_{in} \mid x_i,\Lambda_{ij}) P(\Lambda_{jk},\ldots\ldots\Lambda_{jn} \mid x_j,\Lambda_{ij}) \ldots\ldots P(\Lambda_{kn}) P(\Lambda_{n(n-1)}) P(x_i) P(x_j) \Bigg] \tag{5.14}$$

where we have used the fact that from the Cycle DAG over the n parties we have

$$(\Lambda_{ik} \perp_{\mathrm{d}} \Lambda_{il},\ldots\ldots\Lambda_{in}) \ , \ (\Lambda_{il} \perp_{\mathrm{d}} \Lambda_{im},\ldots\ldots\Lambda_{in}) \ , \ \ldots\ldots(\Lambda_{i(n-1)} \perp_{\mathrm{d}} \Lambda_{in}) \ , \\ (\Lambda_{ik},\ldots\ldots\Lambda_{in} \perp_{\mathrm{d}} x_i,\Lambda_{ij}) \\ \implies P(\Lambda_{ik},\ldots\ldots\Lambda_{in} \mid x_i,\Lambda_{ij}) = P(\Lambda_{ik})\ldots\ldots P(\Lambda_{in})$$

and

$$(\Lambda_{jk} \perp_{\mathrm{d}} \Lambda_{jl},\ldots\ldots\Lambda_{jn}) \ , \ (\Lambda_{jl} \perp_{\mathrm{d}} \Lambda_{jm},\ldots\ldots\Lambda_{jn}) \ , \ \ldots\ldots(\Lambda_{j(n-1)} \perp_{\mathrm{d}} \Lambda_{jn}) \ , \\ (\Lambda_{jk},\ldots\ldots\Lambda_{jn} \perp_{\mathrm{d}} x_j,\Lambda_{ij}) \\ \implies P(\Lambda_{jk},\ldots\ldots\Lambda_{jn} \mid x_j,\Lambda_{ij}) = P(\Lambda_{jk})\ldots\ldots P(\Lambda_{jn})$$

thus giving us back the required distribution over the parties $i,j \in \mathbb{P}$ as in Equation 5.9.

$$P(a_i,a_j,x_i,x_j) = \sum_{\Lambda_{ij}} P(a_i \mid x_i,\Lambda_{ij}) P(a_j \mid x_j,\Lambda_{ij}) P(x_i) P(x_j) P(\Lambda_{ij}) \tag{5.15}$$

Hence we have shown that the Cycle DAG over the n parties or over the $2n$ observed random variables $(a_1,\ldots\ldots a_n, x_i \ldots\ldots x_n)$ is a valid possibility. Figure 5.6 shows the case for only three parties for simplicity. Also note that instead of naming the hidden variables as $\Lambda_{12},\Lambda_{13},\Lambda_{23}$ we have simply denoted the hidden variables by $\Lambda_1,\Lambda_2,\Lambda_3$ in Figure 5.6.

Now we need to show two things:

1. If this Cycle DAG over the n parties or over the $2n$ observed random variables $(a_1,\ldots\ldots a_n, x_i \ldots\ldots x_n)$ is the only unique scenario that can be constructed when each pair of parties $i,j \in \mathbb{P}$ share a Bell sub-scenario between them, then it cannot support all quantum correlations unless we resort to Fine-Tuning.

2. If it is not the only possible scenario in such a case, then all the other possible scenarios can also not explain all quantum correlations without FINE-TUNING.

We first show that the Cycle DAG over the n parties cannot support all possible quantum correlations. To see this note that distributions in which *all the output nodes* $(a_1, ..., a_n)$ *are perfectly correlated* respects all the conditional independences in Equations 5.6 and 5.7. But even such a perfectly classically attainable distribution is incompatible with the Cycle DAG over the n parties due to the absence of a n-party common cause of the outputs $(a_1, .., a_n)$ which is the only way to generate perfect correlations between the outputs $(a_1, ..., a_n)$, without introducing FINE-TUNING.

Now, we check whether it is the only possible scenario that can be constructed in such a case or not. To check this, we consider all the possible edges and hidden variables that can be added or deleted in such a circumstance. If this is the only way in which all possible scenarios respecting the observed conditional independences can be generated from the Cycle DAG then our conjecture can be formed into a theorem. But as of now we do not have a proof of this fact and hence we have it only as a conjecture. Also while checking the addition and deletion of all possible edges and hidden variables, it is sufficient to just consider three parties in total because all the possible edges and hidden variables that can be added or deleted between all the nodes of n parties can also be added or deleted between the nodes of three parties. Thus, throughout showing this, we refer to Figure 5.6 for the notation. We begin by first considering the possibility of deletion of any edges.

- It is trivial to see that no edge in any particular Bell sub-scenario can be deleted. This is because the deletion of any edge would imply an extra independent conditional independence that would not be implied by the conditional independences in Equation 5.6. We know this because if the extra conditional independence could be implied by the conditional independences in Equation 5.6, then Wood and Spekkens' result [108] would have already taken it into account and not generated a Bell sub-scenario between the two concerned parties in the first place itself.

Now we consider the possibility of the addition of edges between the same or different Bell sub-scenarios, like between Bell sub-scenarios 1 and 2 in Figure 5.6. We show that either addition of such edges is not possible unless we resort to FINE-TUNING or where it is possible we still recover our result. Nine types of edges or any of their combinations can be added in 3 or n party scenarios. Namely, they are the edges that could be added from a setting-to-a-setting, output-to-output, setting-to-output, output-to-setting, hidden-variable-to-setting, setting-to-hidden-variable, output-to-hidden-variable, hidden-variable-to-output, hidden-variable-to-hidden-variable and any of their combination.

Hence, consider the following nine possible types of edges that could be added one by one:

1. *(Setting-to-Setting):* Addition of edges from one party's setting to another party's setting, for

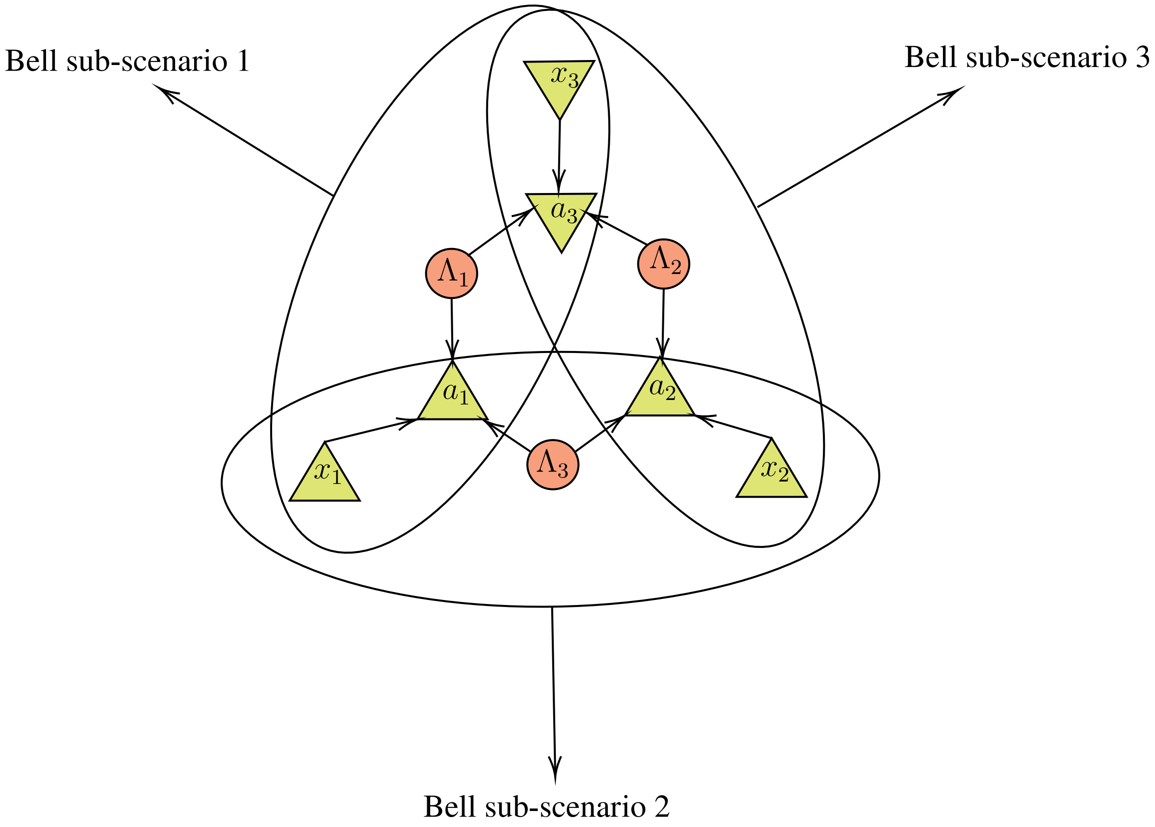


Figure 5.6: Multi-partite Bell scenarios with independent settings

example, the addition of $x_1 \to x_2$ or $x_1 \to x_3$ is not possible as it would lead to $(x_1 \not\perp x_2)$ in Bell sub-scenario 1 or $(x_1 \not\perp x_3)$ in Bell sub-scenario 2 (unless one resorts to FINE-TUNING).

2. *(Output-to-Output):* Addition of edges from an output to another output, like $a_1 \to a_2$ or $a_1 \to a_3$ is not possible as it would lead to violation of $(a_2 \not\perp x_1 \mid x_2)$ in Bell sub-scenario 1 or violation of $(a_3 \not\perp x_1 \mid x_3)$ in Bell sub-scenario 3 (unless again one resorts to FINE-TUNING).

3. *(Setting-to-Output):* Addition of edges from the setting of a party to the output of a party in the same or a different Bell sub-scenario is not possible since they would lead to the violation of the no-signalling conditions. An added edge $x_1 \to a_3$ is prohibited because it leads to $(a_3 \not\perp x_1 \mid x_3)$ in Bell sub-scenario 1, whereas an added edge $x_1 \to a_2$ is prohibited because it leads to $(a_2 \not\perp x_1 \mid x_2)$ in Bell sub-scenario 2.

4. *(Output-to-Setting):* Addition of edges from the output of one party to the setting of another [13] is prohibited because they lead to violation of the measurement independence assumption. For

[13] Addition of an edge from the output of a party to the setting of the same party is trivially not possible as it introduces a directed cycle.

example, addition of the edge $a_1 \rightarrow x_3$ leads to $(x_1 \not\perp\!\!\!\perp x_3)$ in Bell sub-scenario 1 and addition of the edge $a_3 \rightarrow x_2$ leads to $(x_2 \not\perp\!\!\!\perp x_3$ in Bell sub-scenario 3.

5. *(Setting-to-Hidden-Variable):* Addition of an edge from a setting to a hidden variable whether in the same or a different Bell sub-scenario is prohibited because it leads to violation of no-signalling conditions. For example addition of the edge $x_1 \rightarrow \Lambda_3$ leads to $(a_2 \not\perp\!\!\!\perp x_1 \mid x_2)$, while addition of the edge $x_1 \rightarrow \Lambda_2$ again leads to $(a_2 \not\perp\!\!\!\perp x_1 \mid x_2)$ and $(a_3 \not\perp\!\!\!\perp x_1 \mid x_3)$.
6. *(Hidden-Variable-to-Setting):* Addition of an edge from a hidden variable to a setting whether in the same or a different Bell sub scenario is prohibited for the very same reason as the previous one: it leads to violation of no-signalling conditions. For example addition of the edge $\Lambda_3 \rightarrow x_1$ leads to $(a_2 \not\perp\!\!\!\perp x_1 \mid x_2)$, while addition of the edge $\Lambda_2 \rightarrow x_1$ again leads to $(a_2 \not\perp\!\!\!\perp x_1 \mid x_2)$ and $(a_3 \not\perp\!\!\!\perp x_1 \mid x_3)$.
7. *(Output-to-Hidden-Variable):* An edge from any output to any other hidden variable cannot be added because if it is added in the same Bell sub-scenario it creates a directed cycle; whereas if added between different Bell sub scenarios, it leads to violation of no-signalling condition. For example addition of the edge $a_1 \rightarrow \Lambda_2$ leads to $(a_2 \not\perp\!\!\!\perp x_1 \mid x_2)$, and addition of $a_2 \rightarrow \Lambda_1$ leads to $(a_1 \not\perp\!\!\!\perp x_2 \mid x_1)$.
8. *(Hidden-Variable-to-Output):* Addition of an edge from a hidden variable to the output which did not previously have an incoming edge from the same hidden variable *is possible* as its addition does not lead to the violation of any of the existing observed conditional independences. This can be checked by noting that addition of such an edge does not lead to the violation of any of the conditional independences in Equations 5.6 and 5.7 For example an edge $\Lambda_1 \rightarrow a_2$ or an edge $\Lambda_2 \rightarrow a_1$ can be added without introducing FINE-TUNING and simultaneously respecting all the observed conditional independences. This possibility gives rise to a number of possible DAGs, all of which are shown in Figure 5.7, where again due to sufficiency we have shown only three parties and where the hidden variables are again named $\Lambda_1, \Lambda_2, \Lambda_3$.

   All the probability distributions compatible with the DAGs in Figure 5.7 are also compatible with the DAG in Figure 5.9 (a).

   Thus, we can explain any probability distribution compatible with the DAGs in Figure 5.7 by the DAG in Figure 5.9. To show this consider as an example, the first DAG in Figure 5.7. Any distribution compatible with it should be of the form:

$$P(a_1, a_2, a_3, x_1, x_2, x_3) = \sum_{\Lambda_1,\Lambda_2,\Lambda_3} P(a_1 \mid x_1, \Lambda_1, \Lambda_3) P(a_2 \mid x_2, \Lambda_2, \Lambda_3) P(a_3 \mid x_3, \Lambda_1, \Lambda_2, \Lambda_3)$$
$$P(x_1)P(x_2)P(x_3)P(\Lambda_1)P(\Lambda_2)P(\Lambda_3) \tag{5.16}$$

and for the DAG of Figure 5.9 (a) with the tri-partite common cause we have that any distribution compatible with it should have the form:

$$P^*(a_1,a_2,a_3,x_1,x_2,x_3) = \sum_{\Lambda} P^*(a_1 \mid x_1,\Lambda)P^*(a_2 \mid x_2,\Lambda)P^*(a_3 \mid x_3,\Lambda)P^*(x_1)P^*(x_2) \\ P^*(x_3)P^*(\Lambda) \tag{5.17}$$

To explain any $P(a_1,a_2,a_3,x_1,x_2,x_3)$ compatible with the first DAG in Figure 5.7 by the DAG in Figure 5.9 (a) we define

$$\Lambda := (\Lambda_1,\Lambda_2,\Lambda_3) \tag{5.18}$$

$$P(\Lambda) := P(\Lambda_1)P(\Lambda_2)P(\Lambda_3) \tag{5.19}$$

and take

$$\begin{aligned} P^*(a_1 \mid x_1,\Lambda) &:= P(a_1 \mid x_1,\Lambda_1,\Lambda_3) \\ P^*(a_2 \mid x_2,\Lambda) &:= P(a_2 \mid x_2,\Lambda_2,\Lambda_3) \\ P^*(a_3 \mid x_3,\Lambda) &:= P(a_3 \mid x_3,\Lambda_1,\Lambda_2,\Lambda_3) \\ P^*(x_i) &:= P(x_i) \quad \forall \quad i = 1,2,3 \end{aligned} \tag{5.20}$$

Similarly, any distribution compatible with *any* of the DAGs in Figure 5.7 can be reproduced by the DAG in Figure 5.9 (a).

Following the same logic, in the case of n parties, we can add edges from various hidden variables to different outputs that did not previously have an incoming edge from that hidden variable. Thus, instead of a tri-partite common cause, we just use an *n-party common cause* to generate all the compatible distributions. This is shown in Figure 5.9 (b).

9. *(Hidden-Variable-to-Hidden-Variable):* Edges can be added between two hidden variables, like $\Lambda_1 \to \Lambda_2$, $\Lambda_2 \to \Lambda_3$, without violating any of the observed conditional independences and without introducing FINE-TUNING. This is again because the introduction of such edges respects all the observed conditional independences in Equations 5.6 and 5.7. This again leads to a number of possible DAGs all of which are shown in Figure 5.8 all of which consist of only three parties because of the earlier stated sufficiency. Again the hidden variables are written down as $\Lambda_1,\Lambda_2,\Lambda_3$.

All the probability distributions compatible with all the DAGs in Figure 5.8 are also compatible with the DAG in Figure 5.9 (a). The reason for this is just as in the previous point we can absorb the Λ_i's into each other and define a new tri-partite common cause.

For example, consider the first DAG in Figure 5.8. Any distribution compatible with it should be of the form:

$$P(a_1, a_2, a_3, x_1, x_2, x_3) = \sum_{\Lambda_1,\Lambda_2,\Lambda_3} P(a_1 \mid x_1, \Lambda_1, \Lambda_3) P(a_2 \mid x_2, \Lambda_2, \Lambda_3) P(a_3 \mid x_3, \Lambda_1, \Lambda_2) \\ P(x_1)P(x_2)P(x_3)P(\Lambda_1 \mid \Lambda_3)P(\Lambda_2)P(\Lambda_3) \tag{5.21}$$

while any distribution compatible with the DAG in Figure 5.9 (a) should have the form:

$$P^*(a_1, a_2, a_3, x_1, x_2, x_3) = \sum_{\Lambda} P^*(a_1 \mid x_1, \Lambda) P^*(a_2 \mid x_2, \Lambda) P^*(a_3 \mid x_3, \Lambda) P^*(x_1) P^*(x_2) \\ P^*(x_3) P^*(\Lambda) \tag{5.22}$$

And thus to explain any distribution $P(a_1, a_2, a_3, x_1, x_2, x_3)$ compatible with Equation 5.21 (or the first DAG in Figure 5.8) via the DAG in Figure 5.9 (a) we simple define

$$\Lambda := (\Lambda_1, \Lambda_2, \Lambda_3) \tag{5.23}$$

$$P(\Lambda) := P(\Lambda_1|\Lambda_3)P(\Lambda_2)P(\Lambda_3) \tag{5.24}$$

and take

$$\begin{aligned} P^*(a_1 \mid x_1, \Lambda) &:= P(a_1 \mid x_1, \Lambda_1, \Lambda_3) \\ P^*(a_2 \mid x_2, \Lambda) &:= P(a_2 \mid x_2, \Lambda_2, \Lambda_3) \\ P^*(a_3 \mid x_3, \Lambda) &:= P(a_3 \mid x_3, \Lambda_1, \Lambda_2) \\ P^*(x_i) &:= P(x_i) \quad \forall \quad i = 1, 2, 3 \end{aligned} \tag{5.25}$$

Following similar lines we can show that any probability distribution compatible with any DAG in Figure 5.8 can be explained via the DAG in Figure 5.9 (a). And, again the result holds in general for n parties as well. This is because in a n party scenario we can have various hidden variables having edges pointing to different hidden variables. But we can again explain this possibility by using an *n-party common cause*. This possibility is shown in Figure 5.9 (b).

Thus, out of the nine possible edges, we can only add edges from hidden variables to other outputs and other hidden variables. All probability distributions compatible with such resulting DAGs of Figures 5.7 and 5.8 are also compatible with the DAG in Figure 5.9 (a) which has only got a tripartite common cause.

Similarly, types of edges *1-7* and thus any of their combinations cannot be added in an n party case as well. Only edges of the type *8 and 9* and any of their combinations can be added in an n party case. But as shown we can always explain this possibility by using an n-party common cause as shown in Figure 5.9 (b).

Now consider the deletion of any edges after the addition of the above edges. For example, in the tri-partite case, consider the addition of the edge $\Lambda_1 \rightarrow a_2$ and simultaneous deletion of the edges $\Lambda_2 \rightarrow a_2$ and $\Lambda_3 \rightarrow a_2$. This is possible, but, it is clear that in such a situation the tripartite common cause Λ_1 can explain all the compatible probability distributions just as in point (8) above. Similarly, it is clear that in the n party case, an n party common cause can explain all compatible distributions that would result after such deletion of edges.

At last, consider the addition of extra hidden variables. Any new added hidden variable that only points to one node can be absorbed into that node itself, and hence does not change anything. A hidden variable can be introduced that points to the setting and the output of the same party. But such a hidden variable can be absorbed into the setting of the concerned party without any loss of generality. A hidden variable that is the common cause of settings of two different parties cannot be added as it would lead to $(x_i \not\perp x_j)$. Adding a hidden variable which is the common cause of the output of two parties is redundant, as such a hidden variable can be absorbed into the hidden variable that was already the common cause of the two parties. A hidden variable that is the common cause of a hidden variable and its output can be added. But such an added hidden variable can again be absorbed into the already present hidden variable that had an edge pointing to its respective output. A hidden variable that is the common cause of another hidden variable and the output of a different party can also be added. But in such a scenario one can absorb the added hidden variable into the already present hidden variable and explain all the compatible probability distributions by an n party common cause between the concerned parties just as in point (8) above. Finally, a hidden variable can be added that points to multiple other hidden variables. But this possibility can again be explained by an n party common cause between the concerned parties just as in point (9) above. Further, any combination of the above possibilities can be thus dealt with, since each of the individual possibilities can be dealt with.

Thus, in the n party case, all possibilities are subsumed by the DAG in Figure 5.9 (b). But this DAG in Figure 5.9 (b) implies the corresponding Mermin-Ardelahi inequalities [61, 6, 62, 35] which are violated by certain experimentally observed quantum correlations. Thus, even this DAG cannot faithfully explain all the quantum correlations.

Hence we conclude our way to the proof of the conjecture that there is no causal explanation of Bell inequality violating correlations in any multi-partite Bell scenario.

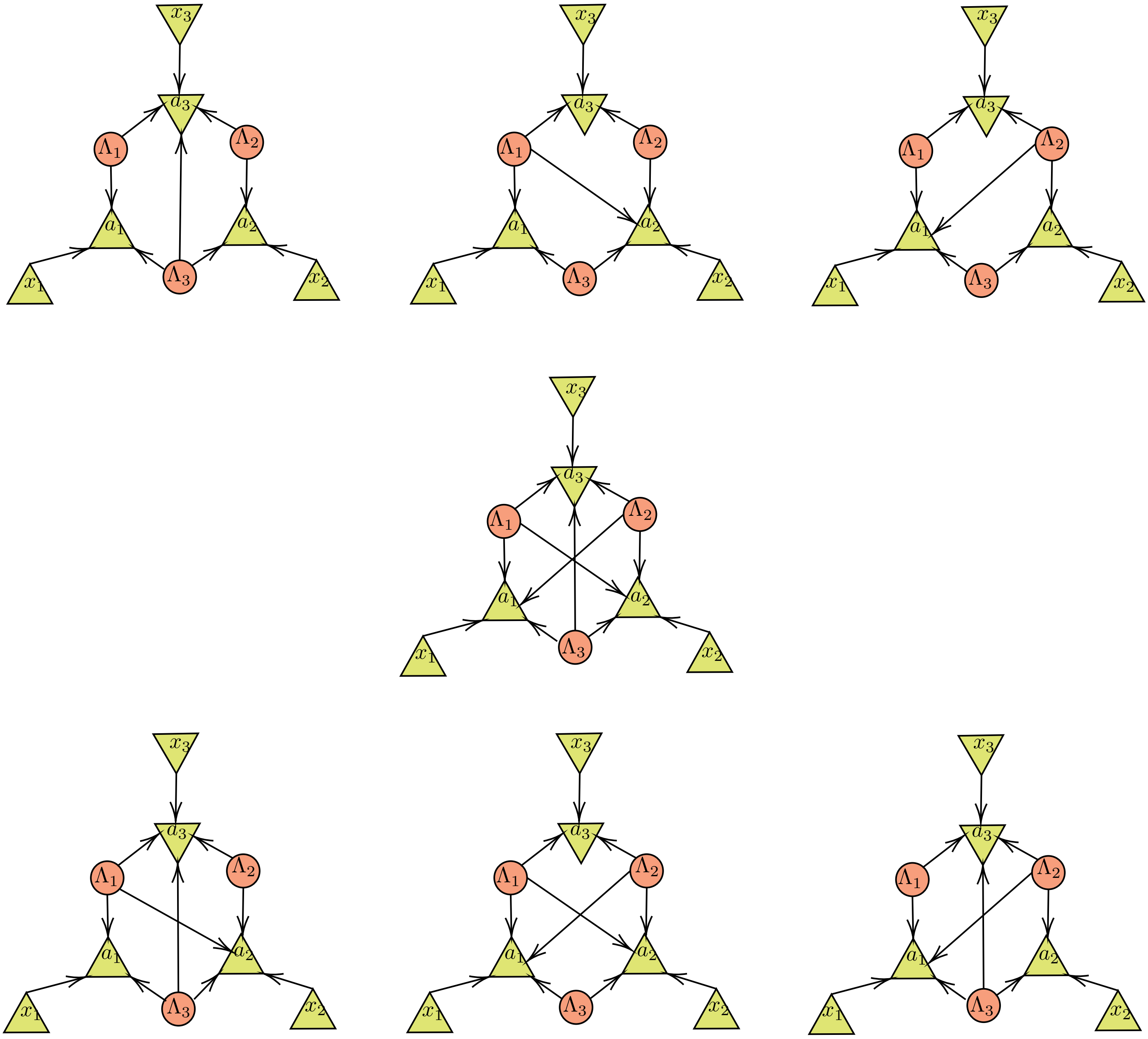


Figure 5.7: Edges that can be added from Hidden-Variable-to-Output

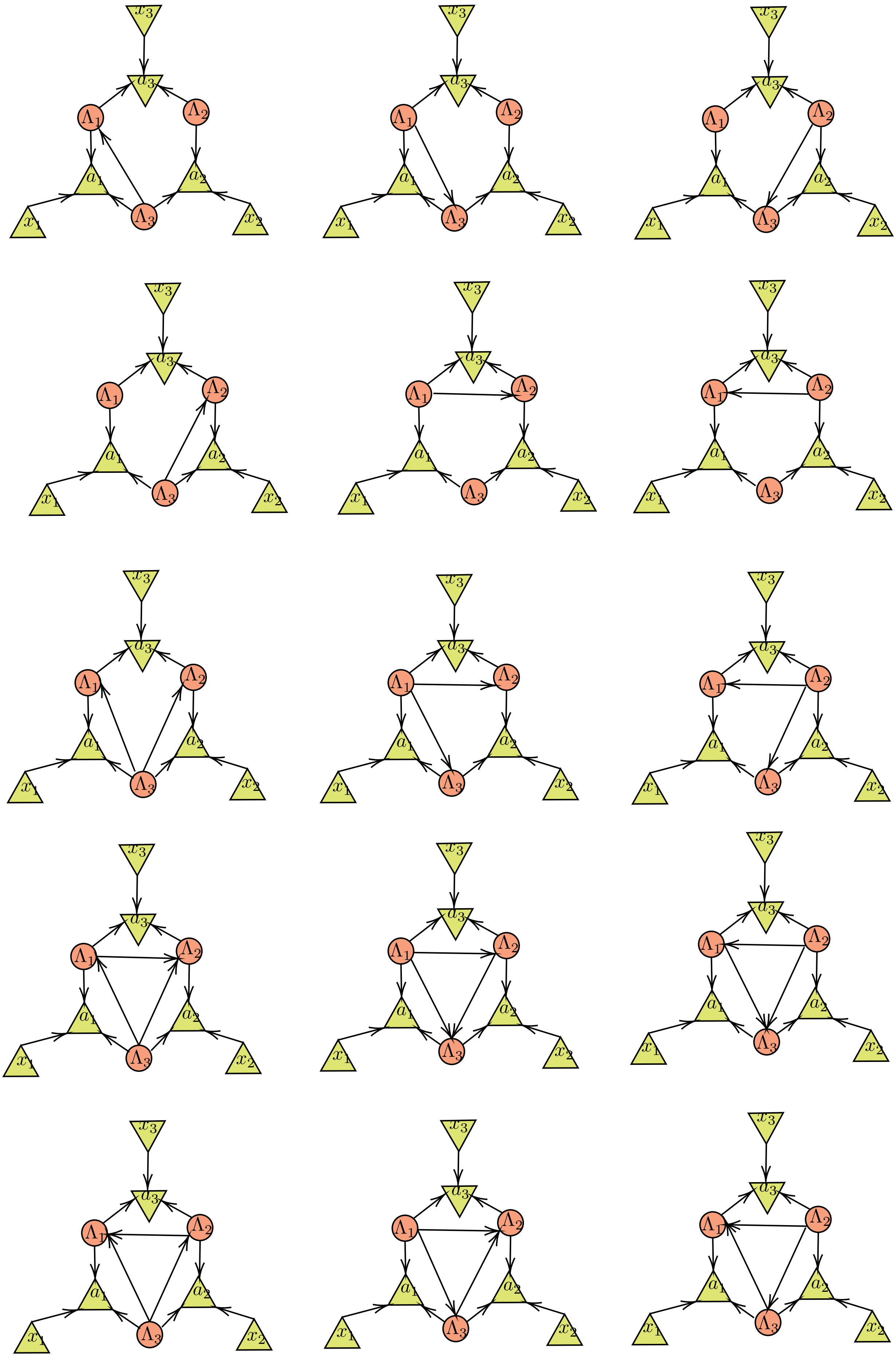


Figure 5.8: Edges that can be added from Hidden-Variables-to-Hidden-Variables

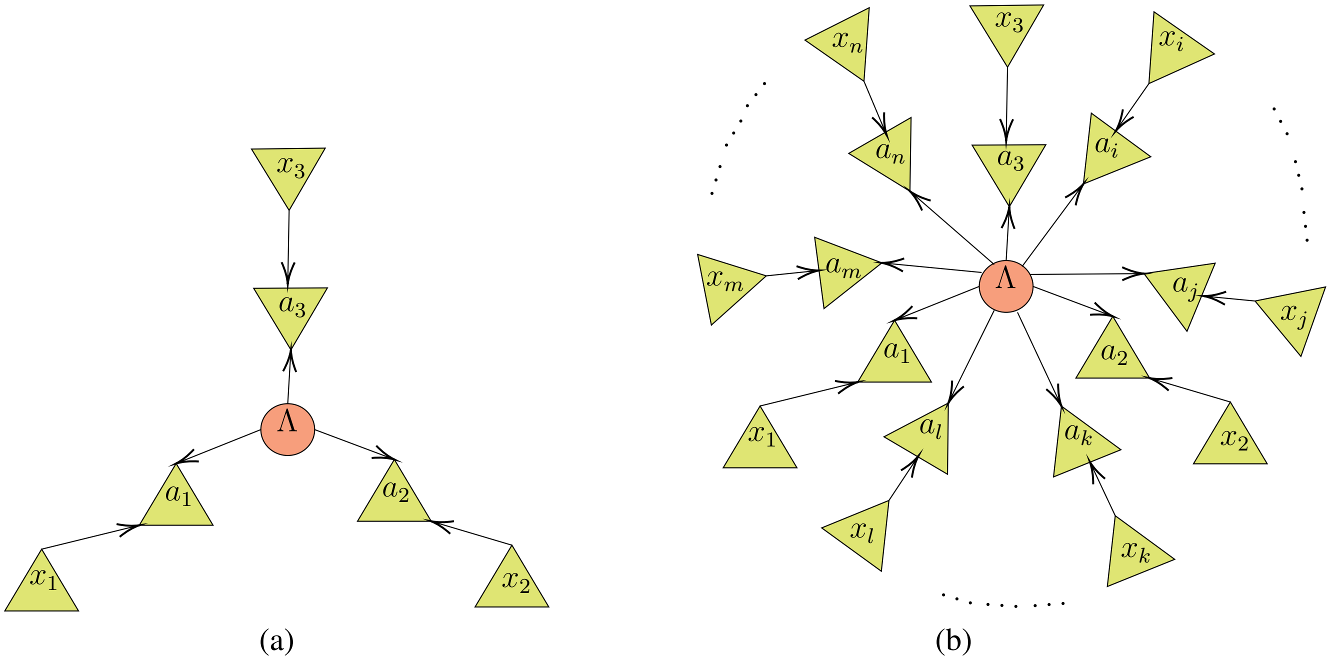


Figure 5.9: Tri-partite and its n-party generalized multi-partite Bell scenarios.

5.5 Summary of the results

In the end we have two results. For concreteness we summarize them now:

1. NON-LOCALITY is not the only defining property of NON-CLASSICAL correlations. Just like the Bell scenario there are several other scenarios that exhibit LOCAL correlations which are NON-CLASSICAL as well. Any causal explanation of these NON-CLASSICAL correlations based on classical causal modelling will require FINE-TUNING just like the Bell scenario. Hence the result of Wood and Spekkens [108] extends to several other scenarios as well. From the scenarios that we modelled, their result extends to scenarios (4th), (7th), Bi-locality, GHZ scenarios in Tables 5.1, 5.2 and 5.3 and also conjectured to extend to the general multi-partite Bell scenario with n independent parties .

2. Suppose we perform some quantum experiment and observe some conditional independences and QUANTUM-CORRELATIONS. Based on our knowledge of the quantum experiment we try to model it using a causal scenario. We find that there is no way to explain these QUANTUM-CORRELATIONS based on our classical causal model without FINE-TUNING. This would suggest that no CLASSICAL theory can reproduce these QUANTUM-CORRELATIONS. But we would still like to search for CLASSICAL explanations of the obtained QUANTUM-CORRELATIONS. Hence we input the observed conditional independences in the IC^* algorithm and find that it returns that there are a number of

scenarios which reproduce these QUANTUM-CORRELATIONS. Amongst them is the scenario that we initially unsuccessfully came up with to model the obtained correlations. But importantly, IC^* outputs that there are a number of other possibly latent free causal scenarios that can also model the observed conditional independences. If we find a latent free causal scenario that can model the observed conditional independences then that scenario will also explain all the obtained QUANTUM-CORRELATIONS classically without FINE-TUNING. This follows from Theorem 64. Thus we will have a valid CLASSICAL explanation of the experimentally observed QUANTUM-CORRELATIONS which we initially thought to have no possible CLASSICAL explanation at all. Amongst the scenarios that we studied, examples of such are scenarios (1), (2), (3), (5) and (6) in Tables 5.1 and 5.2.

5.6 Discussion on Modifying the IC^* algorithm for Causal Discovery

In the last two sections, we have reiterated another result; the IC^* algorithm is not sufficient for causal discovery. This is because the strength of correlations plays an important role determining the causal mechanism at play. Or put differently conditional independences alone do not characterize an observed probability distribution completely. Now we suggest an improvement to the IC^* algorithm.

It is based on the fact that strength of correlations plays an important role in characterizing the classically compatible distributions. Hence we suggest appending the IC^* algorithm with an algorithm that can check the strength of compatible distributions. One such algorithm is provided by Fraser [31]. To get a list of DAGs that obey the observed conditional independences and thus also the observed correlations we can combine the two algorithms in the following way:

1. Input the observed conditional independences in the IC^* algorithm to get extract a list of possibly compatible DAGs.
2. Use Fraser's algorithm to check which of the extracted DAGs support all of the observed correlations.

Only the DAGs which can support all of the observed correlations can be possible explanation of the observed probability distribution.

5.7 Conclusion

Causal discovery algorithms have become an important field of research in Machine Learning and Artificial Intelligence. In this study, we applied the IC^* algorithm to various causal scenarios with up to six total nodes, showing that for most scenarios, the algorithm was able to model the observed conditional independences using both scenarios for which observed d-separation relations are the only constraints on the possible distributions compatible with the scenario and scenarios for which there are constraints beyond observable d-separation that restrict the compatible distributions. However, for three scenarios, the algorithm was only able to model the conditional independences using faithful causal scenarios having constraints beyond observed d-separation relations. This suggests that there may not be a classical causal faithful explanation for these scenarios. We also examined the Bi-Locality, the GHZ and the multi-partite Bell scenarios in quantum mechanics and found that the only faithful scenarios that can explain the corresponding observed conditional independences in the Bi-locality and the GHZ scenarios are in fact just one in each case; namely, the Bi-Locality, GHZ scenarios themselves. Whereas, for the multi-partite Bell scenario we conjecture that it is the only only one reproducing the respectively observed conditional independences. This result also highlights the importance of exploring properties of data beyond just observed conditional independences when attempting to reproduce a causal hypothesis for a given set of data.

— 6 —

Classical-Quantum gap in Causal Scenarios

6.1 Introduction

In the previous chapters we have considered causal scenarios of four observed nodes and shown that all but three scenarios of four observed nodes to which HLP's condition does not apply are actually INTERESTING. And all scenarios of six total nodes to which HLP's condition does not apply are INTERESTING. Put in other words these scenarios (and their corresponding DAG G) have a $\mathcal{C}_G \subset \mathcal{I}_G$. The question that still persists is whether they have $\mathcal{C}_G \subset \mathcal{Q}_G$ or not. One way to answer this question is to check whether for these DAGs the probability distributions that we show are classically infeasible are quantumly realizable or not. If they are quantumly realizable then we have that $\mathcal{C}_G \subset \mathcal{Q}_G$.

Here we develop a method to show that some of these INTERESTING DAGs have $\mathcal{C}_G \subset \mathcal{Q}_G$. We apply the method on two DAGs of six total nodes for which it is unknown whether $\mathcal{C}_G \subset \mathcal{Q}_G$ or not and find that $\mathcal{C}_G \subset \mathcal{Q}_G$ is indeed true for them.

6.2 Results

The causal scenarios of Figure 6.1 and Figure 6.3, which have been taken from the appendix of [44] are the two scenarios of six total nodes that we consider here. Here we show that $\mathcal{C}_G \subset \mathcal{Q}_G$ holds true for both of them.

Theorem 69. *There are correlations that can be achieved quantum mechanically but not classically in the causal scenario of Figure 6.1.*

Proof. We first prove the result numerically and then present the analytical proof.

We begin by constructing a distribution that is compatible with the DAG in Figure 6.1 and can be reproduced quantum mechanically.

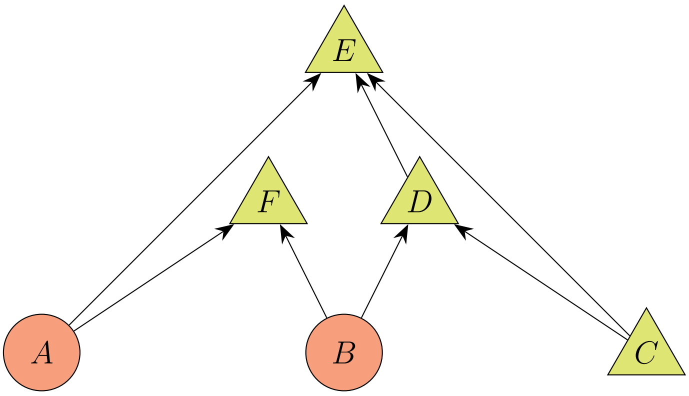


Figure 6.1: Causal scenario exhibiting a Quantum-Classical gap

Consider the FINE-TUNED subset of distributions, where F is two bits, F_O and F_S, and E is a bit perfectly correlated with F_S. Further, let E be independent of the bits C and D. This results in FINE-TUNED distributions because E is independent of C and D even through there are directed edges $C \rightarrow E$ and $D \rightarrow E$.

We show that such a probability distribution can be generated by resorting to the following quantum model:

1. Let B be the maximally entangled state $|\psi\rangle = \left(\frac{|00\rangle+|11\rangle}{\sqrt{2}}\right)$, and let B send the first qubit of $|\psi\rangle$ to F_O and the second qubit to D.

2. Let A be the common cause of E and F which perfectly correlates E and F_S and where A is taken to be the perfectly classically correlated state of two qubits, i.e., $A = \frac{1}{2}\left(|00\rangle\langle 11| + |11\rangle\langle 11|\right)$. Let E receive the first qubit from A and F_S receive the second qubit from A. Further, let both E and F_S measure the qubits received from A in the computational basis. The outputs of the perfectly correlated bits E and F_S are $\{0, 1\}$.

3. When $E = F_S = 0$, a measurement in the $\left\{|0\rangle, |1\rangle\right\}$ basis is carried out at F_O on the qubit received from B. When $E = F_S = 1$, a measurement in the $\left\{\frac{|0\rangle+|1\rangle}{\sqrt{2}}, \frac{|0\rangle-|1\rangle}{\sqrt{2}}\right\}$ basis is carried out at F_O on the qubit received from B. The results of these measurements are exhibited by the bit F_O as $\left\{0, 1\right\}$ respectively.

4. When $C = 0$, a measurement in the $\Big\{ \cos(\theta)|0\rangle + \sin(\theta)|1\rangle, \quad -\sin(\theta)|0\rangle + \cos(\theta)|1\rangle \Big\}$ basis is carried out at D on the qubit received from B. When $C = 1$, a measurement in the $\Big\{ \cos(\theta)|0\rangle - \sin(\theta)|1\rangle, \ \sin(\theta)|0\rangle + \cos(\theta)|1\rangle \Big\}$ basis is carried out at D on the qubit received from B. D contains the results of these measurements as $\Big\{0, 1\Big\}$ respectively.

Take $F = (F_O, F_S)$ such that $F = (0, 0) = 0,\ F = (0, 1) = 1,\ F = (1, 0) = 2,\ F = (1, 1) = 3$. When $E = 0$, $F = (0, 0) = 0$ implies that the post measurement state at F is $\Big(|0\rangle\Big)$ and $F = (1, 0) = 2$ implies that the post measurement state at F is $\Big(|1\rangle\Big)$. When $E = 1$, $F = (0, 1) = 1$ implies that the post measurement state at F is $\left(\frac{|0\rangle+|1\rangle}{\sqrt{2}}\right)$ and $F = (1, 1) = 3$ implies that the post measurement state at F is $\left(\frac{|0\rangle-|1\rangle}{\sqrt{2}}\right)$. Similarly, when $C = 0$, $D = 0$ implies that the post measurement state at D is $\Big(\cos(\theta)|0\rangle + \sin(\theta)|1\rangle \Big)$ and $D = 1$ implies that the post measurement state at D is $\Big(-\sin(\theta)|0\rangle + \cos(\theta)|1\rangle \Big)$. When $C = 1$, $D = 0$ implies that the post measurement state at D is $\Big(\cos(\theta)|0\rangle - \sin(\theta)|1\rangle \Big)$ and $D = 1$ implies that the post measurement state at D is $\Big(\sin(\theta)|0\rangle + \cos(\theta)|1\rangle \Big)$. We further take $P\Big(E\Big)$ and $P\Big(C\Big)$ to be uniform. Quantum mechanically, via the Born rule we have

$$P\Big(F, E, D, C\Big) = Tr\Big[\Big(M_E^F \otimes M_C^D\Big)\rho_{B_{FD}}\Big]P\Big(E\Big)P\Big(C\Big) = P\Big(F, E|D, C\Big)P\Big(E\Big)P\Big(C\Big)^{1}. \tag{6.1}$$

where for each value of C and E, $\{M_E^F\}$ and $\{M_C^D\}$ are the respective POVMs indexed by the values F and D take.

The non-zero probabilities obtained for the distribution $P\Big(F, E, D, C\Big)$ are shown in Table A.4.[2]

Now we only need to show that this distribution cannot be achieved classically. This can be shown by resorting to Fraser's algorithm for infeasible supports [31]. If the support of the probability

[1] Where P(EC)=P(E)P(C) has been used since E is independent of C.

[2] Since E and F_S are perfectly correlated there are only sixteen non-zero probabilities instead of the full thirty two.

distribution in Table A.4 is classically infeasible then so is the probability distribution of Table A.4 as well [53]. However, when we test the support of this distribution, Fraser's algorithm returns that the support is classically feasible. Hence we cannot draw any conclusion about the classical feasibility of the probability distribution in Table A.4. So instead we try to construct a Hardy type bipartite setup [42, 60] in the scenario of Figure 6.1.

For this consider the following quantum model generating the observed FINE-TUNED probability distributions in which E is perfectly correlated with the bit F_S and is also simultaneously independent of the bits D and C:

1. Let B be the entangled state $|\psi_\theta\rangle = \frac{1}{\sqrt{1+\cos^2\theta}}\Bigg(\sin(\theta)|11\rangle + \cos(\theta)|01\rangle + \cos(\theta)|10\rangle\Bigg)$, where $0 < \theta < \pi/2$ [60]. Further let B distribute the first qubit of $|\psi_\theta\rangle$ to F and the second qubit to D.
2. Let A be the common cause of E and F which perfectly correlates E and F_S and where A is taken to be the perfectly classically correlated state of two qubits, i.e., $A = \frac{1}{2}\Bigg(|00\rangle\langle 11| + |11\rangle\langle 11|\Bigg)$. Let E receive the first qubit from A and F_S receive the second qubit from A. Further, let both E and F_S measure the qubits received from A in the computational basis. The outputs of the perfectly correlated bits E and F_S are $\{0, 1\}$.
3. When $E = F_S = 0$, a measurement in the $\Bigg\{|0\rangle, |1\rangle\Bigg\}$ basis is carried out at F on the qubit received from B. When $E = F_S = 1$, a measurement in the $\Bigg\{\Bigg(\sin(\theta)|0\rangle - \cos(\theta)|1\rangle\Bigg), \Bigg(\cos(\theta)|0\rangle + \sin(\theta)|1\rangle\Bigg)\Bigg\}$ basis is carried out at F on the qubit received from B (and where $0 < \theta < \pi/2$) . The results of these measurements are exhibited by the bit F_O as $\Big\{0, 1\Big\}$ respectively.
4. When $C = 0$, a measurement in the $\Bigg\{|0\rangle, |1\rangle\Bigg\}$ basis is carried out at D on the qubit received from B. When $C = 1$, a measurement in the $\Bigg\{\Bigg(\sin(\theta)|0\rangle - \cos(\theta)|1\rangle\Bigg), \Bigg(\cos(\theta)|0\rangle + \sin(\theta)|1\rangle\Bigg)\Bigg\}$ basis is carried out at D on the qubit received from B (and where $0 < \theta < \pi/2$). D contains the results of these measurements as $\Big\{0, 1\Big\}$ respectively.

Take $F = (F_O, F_S)$ such that $F = (0, 0) = 0,\ F = (0, 1) = 1,\ F = (1, 0) = 2,\ F = (1, 1) = 3.$

When $E = 0$, $F = (0,0) = 0$ implies that the post measurement state at F is $\Big(|0\rangle\Big)$ and $F = (1,0) = 2$ implies that the post measurement state at F is $\Big(|1\rangle\Big)$. When $E = 1$, $F = (0,1) = 1$ implies that the post measurement state at F is $\Bigg(\sin(\theta)|0\rangle - \cos(\theta)|1\rangle\Bigg)$ and $F = (1,1) = 3$ implies that the post measurement state at F is $\Bigg(\cos(\theta)|0\rangle + \sin(\theta)|1\rangle\Bigg)$. Similarly, when $C = 0$, $D = 0$ implies that the post measurement state at D is $\Big(|0\rangle\Big)$ and $D = 1$ implies that the post measurement state at D is $\Big(|1\rangle\Big)$. When $C = 1$, $D = 0$ implies that the post measurement state at D is $\Bigg(\sin(\theta)|0\rangle - \cos(\theta)|1\rangle\Bigg)$ and $D = 1$ implies that the post measurement state at D is $\Bigg(\cos(\theta)|0\rangle + \sin(\theta)|1\rangle\Bigg)$. Again we take $P\Big(E\Big)$ and $P\Big(C\Big)$ to be uniform. Quantum mechanically, via the Born rule we have,

$$P\Big(F,E,D,C\Big) = Tr\Bigg[\Big(M_E^F \otimes M_C^D\Big)\rho_{B_{FD}}\Bigg]P\Big(E\Big)P\Big(C\Big) = P\Big(F,E|D,C\Big)P\Big(E\Big)P\Big(C\Big)^{3}. \quad (6.2)$$

where again for each value of C and E, $\{M_E^F\}$ and $\{M_C^D\}$ are the respective POVMs indexed by the values F and D take.

The non-zero probabilities obtained in this distribution $P\Big(F,E,D,C\Big)$ are shown in Table 7.1.[4]

If we test the support of this distribution then, Fraser's algorithm determines it to be a classically infeasible support. Hence, the probability distribution of Table 7.1 is also classically infeasible [53]. Thus we have found a probability distribution that is impossible classically but can be achieved using quantum mechanics. This completes the numerical proof.

Now, we present the analytical proof.

Any distribution over the four visible nodes of the causal scenario of Figure 6.1, by the Markov condition takes the form:

$$\sum_{AB} P(A,B,C,D,E,F) = \sum_{AB} P(A|B,C,D,E,F)P(F|B,C,D,E)P(E|B,C,D)$$
$$P(D|B,C)P(B|C)P(C). \quad (6.3)$$

[3] Where P$\Big($EC$\Big)$=P$\Big($E$\Big)$P$\Big($C$\Big)$ has been used since E is independent of C.

[4] Since E and F_S are perfectly correlated there are only sixteen non-zero probabilities instead of the full thirty two.

Summing over A and using the global Markov conditions, $(E \perp_{\mathrm{d}} B|CD) \implies (E \perp\!\!\!\perp_{\mathrm{CI}} B|CD)$ and $(B \perp_{\mathrm{d}} C) \implies (B \perp\!\!\!\perp_{\mathrm{CI}} C)$ in the causal scenario we get:

$$\sum_{AB} P(A,B,C,D,E,F) = \sum_{B} P(F|B,C,D,E)P(E|C,D)P(D|B,C)P(B)P(C). \tag{6.4}$$

Exploiting the fact that in our observed distribution ($E \perp\!\!\!\perp_{\mathrm{CI}} CD$), we get:

$$\sum_{AB} P(A,B,C,D,E,F) = \sum_{B} P(F|B,C,D,E)P(E)P(D|B,C)P(B)P(C). \tag{6.5}$$

Since $F = (F_O, F_S)$ and $E = F_S$, we get:

$$\sum_{A,B} P(A,B,C,D,E,F_O) = \sum_{A,B,F_S} P(A,B,C,D,E,F) = \sum_{B,F_S} P(F_O,F_S|B,C,D,E)P(E)P(D|B,C)P(B)P(C). \tag{6.6}$$

Hence,

$$P(C,D,E,F_O) = \sum_{A,B,F_S} P(A,B,C,D,E,F) = \sum_{B} P(F_O|B,C,D,E)P(E)P(D|B,C) P(B)P(C). \tag{6.7}$$

$$= \sum_{B} \frac{P(F_O,E|B,C,D)}{P(E|B,C,D)} P(E)P(D|B,C)P(B)P(C). \tag{6.8}$$

Again using $(E \perp_{\mathrm{d}} B|CD) \implies (E \perp\!\!\!\perp_{\mathrm{CI}} B|CD)$ and ($E \perp\!\!\!\perp_{\mathrm{CI}} CD$) (from the property of the observed distribution) we get:

$$= \sum_{B} \frac{P(F_O,E|B,C,D)}{P(E)} P(E)P(D|B,C)P(B)P(C). \tag{6.9}$$

Hence we get (using $E = F_S$ and $F = (F_O, F_S)$) ,

$$P(C,D,E,F_O) = \sum_{A,B,F_S} P(A,B,C,D,E,F) = \sum_{B} P(F|B,C,D)P(D|B,C)P(B)P(C). \tag{6.10}$$

Now we use $(F \perp_{\mathrm{d}} CD|B) \implies (F \perp\!\!\!\perp_{\mathrm{CI}} \; CD|B)$ (from the global Markov condition) to get:

$$P(C,D,E,F_O) = \sum_{A,B,F_S} P(A,B,C,D,E,F) = \sum_{B} P(F|B)P(D|B,C)P(B)P(C). \tag{6.11}$$

$$= \sum_{B} P(F_O,E|B)P(D|B,C)P(B)P(C). \tag{6.12}$$

$$= \sum_{B} P(F_O, |E, B)P(E|B)P(D|B, C)P(B)P(C). \tag{6.13}$$

Now we use the fact that $P(E) = P(E|C, D) = P(E|B, C, D)$[5]. Hence, $P(E) = P(E|B, C, D)$. This in turn implies that:

$$P(E, B, C, D) = P(E)P(B, C, D) \tag{6.14}$$

and thus,

$$\sum_{C,D} P(E, B, C, D) = \sum_{C,D} P(E)P(B, C, D). \tag{6.15}$$

Therefore, we get:

$$\sum_{C,D} P(E, B, C, D) = P(E, B) = P(E|B)P(B) = \sum_{C,D} P(E)P(B, C, D) = P(E)P(B) \tag{6.16}$$

and conclude that:

$$P(E|B) = P(E). \tag{6.17}$$

Using this in Equation 6.13 we get:

$$P(C, D, E, F_O) = \sum_{A,B,F_S} P(A, B, C, D, E, F) = \sum_{B} P(F_O, |E, B)P(E)P(D|B, C)P(B)P(C). \tag{6.18}$$

Thus we have,

$$P(C, D, E, F_O) = \sum_{B} P(F_O, |E, B)P(D|B, C)P(E)P(B)P(C). \tag{6.19}$$

But this is just Bell's Local Causality [10] condition for the random variables C, D, E, F_O. Equation 6.19 implies Bell's inequality constraints [18] for $P(F, E, D, C)$ which are known to be violated by both the probability distributions of Table A.4 and Table 7.1. Hence the probability distributions of Table A.4 and Table 7.1 in which E is perfectly correlated with F_S and $(E \perp\!\!\!\perp_{\text{CI}} CD)$ are in the quantum set for the given causal scenario but not in the classical set for the same. This completes the analytical proof. □

The second scenario we consider is quite similar to the first one and the proof follows the same logic.

Theorem 70. *There are correlations that can be achieved quantum mechanically but not classically in the causal scenario of Figure 6.3.*

[5] Where we have invoked that $(E \perp_{\text{d}} B|C, D) \implies (E \perp\!\!\!\perp_{\text{CI}} B|C, D)$ in the causal scenario.

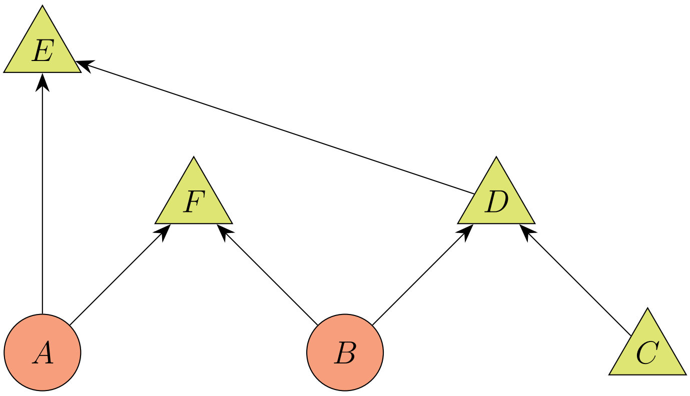


Figure 6.2: Causal scenario exhibiting a Quantum-Classical gap

Proof. To prove the theorem, we construct a quantum mechanically feasible probability distribution and show that it is classically infeasible in this scenario. Consider the FINE-TUNED subset of probability distributions, where F is two bits, F_O and F_S, and E is a bit perfectly correlated with F_S. Further, E is independent of the bits C and D. This is a FINE-TUNED distribution because E is independent of C and D even through there are directed edges $C \to D$ and $D \to E$. From the proof of Theorem 69, we see that in the quantum model generating the consequent probability distributions E did not depend on C. Hence the construction of the quantumly feasible distribution in this scenario of Figure 6.3 can be taken to be exactly the same as in the last scenario we considered (i.e. scenario of Figure 6.1). Therefore, the distributions of Table A.4 and Table 7.1 are quantum mechanically realizable in this scenario as well. The proof that they are classically infeasible follows the same line of thought as the proof of the last scenario did, except that in this scenario there is no directed edge from C to E. Note that in the observed distribution we again have $(E \perp\!\!\!\perp_{\text{CI}} DC)$.

Any distribution over the four visible nodes of the causal scenario of Figure 6.1 takes the form:

$$\sum_{AB} P(A,B,C,D,E,F) = \sum_{AB} P(A|B,C,D,E,F)P(F|B,C,D,E)P(E|B,C,D) P(D|B,C)P(B|C)P(C). \quad (6.20)$$

Summing over A and using $(E \perp_{\text{d}} BC|D) \implies (E \perp\!\!\!\perp_{\text{CI}} BC|D)$ and $(B \perp_{\text{d}} C) \implies (B \perp\!\!\!\perp_{\text{CI}} C)$ (from global Markov conditions), we get,

$$\sum_{AB} P(A,B,C,D,E,F) = \sum_{AB} P(F|B,C,D,E)P(E|D)P(D|B,C)P(B)P(C). \quad (6.21)$$

Now using $(E \perp\!\!\!\perp_{\text{CI}} D)$ from the property of the observed distribution, we get

$$\sum_{AB} P(A,B,C,D,E,F) = \sum_{B} P(F|B,C,D,E)P(E)P(D|B,C)P(B)P(C). \quad (6.22)$$

Now we exploit the fact that $F = (F_O, F_S)$ to get,

$$P(C, D, E, F_O) = \sum_{F_S,B} P(A, B, C, D, E, F) = \sum_{F_S,B} P(F_O, F_S|B, C, D, E)P(E) \\ P(D|B, C)P(B)P(C). \quad (6.23)$$

Summing over F_S, we get.

$$P(C, D, E, F_O) = \sum_{B} P(F_O|B, C, D, E)P(E)P(D|B, C)P(B)P(C). \quad (6.24)$$

Now using $(E \perp_{\text{d}} BC|D) \implies (E \perp\!\!\!\perp_{\text{CI}} BC|D)$ first and then $(E \perp\!\!\!\perp_{\text{CI}} D)$ (from the property of the observed distribution) we get,

$$P(C, D, E, F_O) = \sum_{B} P(F_O, E|B, C, D)P(D|B, C)P(B)P(C). \quad (6.25)$$

Since, $F_S = E$, we have that $F = (F_O, F_S) = (F_O, E)$, thereby giving

$$P(C, D, F) = \sum_{B} P(F|B, C, D)P(D|B, C)P(B)P(C). \quad (6.26)$$

Using $(F \perp_{\text{d}} CD|B) \implies (F \perp\!\!\!\perp_{\text{CI}} CD|B)$ (from the global Markov condition) we get,

$$P(C, D, F) = \sum_{B} P(F|B)P(D|B, C)P(B)P(C). \quad (6.27)$$

Again, using $F = (F_O, E)$ we have that,

$$P(C, D, E, F_O) = \sum_{B} P(F|B)P(E|B)P(D|B, C)P(B)P(C). \quad (6.28)$$

Next we show that $P(E|B) = P(E)$. This follows from the fact that $P(E) = P(E|D)$ (since $(E \perp\!\!\!\perp_{\text{CI}} D)$ in the observed distribution) and $P(E) = P(E|BD)$ (from the global Markov condition $(E \perp_{\text{d}} B|D) \implies (E \perp\!\!\!\perp_{\text{CI}} B|D)$). Thus, $P(E) = P(E|BD)$ giving that $(E \perp\!\!\!\perp_{\text{CI}} BD)$ and following from that via the semi-graphoid axioms [79], we have have $(E \perp\!\!\!\perp_{\text{CI}} B)$ or that $P(E|B) = P(E)$. Thus Equation 6.28 reduces to:

$$P(C, D, E, F_O) = \sum_{B} P(F|B)P(E)P(D|B, C)P(B)P(C). \quad (6.29)$$

This is again Bell's Local Causality condition [10] which leads to Bell's inequalities [18] for variables C, D, E, F_O. If, the probability distributions of Table A.4 and Table 7.1, which have the extra conditional independence $(E \perp\!\!\!\perp_{\text{CI}} DC)$ and for which E is perfectly correlated with F_S, are to be classically feasible then they must obey Equation 6.28 and the Bell's inequalities resulting from it. However, it is known that both these distributions violate Bell's inequalities hence they are quantum mechanically feasible but not classically feasible in the scenario of Figure 6.3. □

6.3 Comments and Discussion

The approach presented here can be taken for some other DAGs in the Appendix E of HLP [44]. For example consider the DAG H in Figure 6.3. This is the same DAG that we considered in Theorem 70. Now we present an easier way to show that it supports NON-CLASSICAL QUANTUM-CORRELATIONS. The only difference it has from the DAG G in Figure 6.1 is that there is no arrow from C to E. Examining the quantum model for the DAG G in Figure 6.1 we see that E does not use C in the quantum model used in Theorem 69. Thus, we can use the same quantum model to get the same observed probabilities as in Table A.4 for this DAG H. Hence this distribution is in $\mathcal{Q}_H$. That it is not in $\mathcal{C}_H$ can be seen as:

$$\begin{aligned} \text{If } P \in \mathcal{C}_H &\implies \text{then } P \in \mathcal{C}_G \\ \text{If } P \notin \mathcal{C}_G &\implies \text{then } P \notin \mathcal{C}_H. \end{aligned} \tag{6.30}$$

Where the first implication follows because $\forall\ P(F,E,D,C) \in \mathcal{C}_H$ we have:

$$\begin{aligned} P(F,E,D,C) = \sum_{A,B} P(F,E,D,C,B,A) = \\ \sum_{A,B} P(F|A,B)\, P(E|A,D)\, P(D|B,C)\, P(C)\, P(B)\, P(A) \end{aligned} \tag{6.31}$$

and thus this same $P(F,E,D,C)$ as in Equation 6.31 is also in $\mathcal{C}_G$ because we can always define:

$$P(E|A,D,C) := P(E|A,D), \tag{6.32}$$

which gives us back the Markov condition: $P(F,E,D,C) = \sum_{A,B} P(F,E,D,C,B,A) = \sum_{A,B} P(F|A,B)\, P(E|A,D,C)\, P(D|B,C)\, P(C)\, P(B)\, P(A)$ for the DAG G in Figure 6.1. Thus, the same distribution that is displayed in Table A.4 is in $\mathcal{Q}_H$ but not in $\mathcal{C}_H$ and therefore we have $\mathcal{C}_H \subset \mathcal{Q}_H$, proving a classical-quantum gap for the DAG in Figure 6.3 as well. A close look at $\mathcal{Q}_G$ in our proof of Theorem 69 $\mathcal{Q}_G$ shows that any probability distribution that is NON-CLASSICAL but

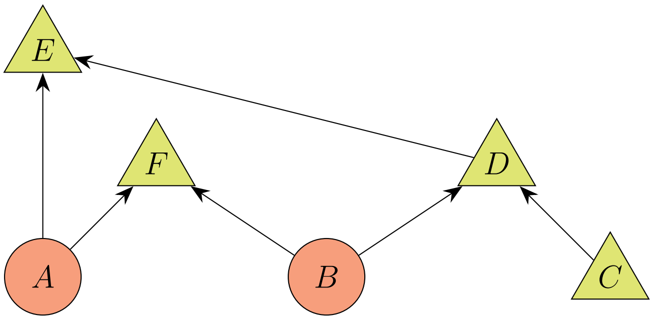


Figure 6.3: Another DAG of total six nodes for which $\mathcal{C}_H \subset \mathcal{Q}_H$ by our technique.

achievable in the Bell DAG is also Non-Classical and achievable in the DAG in Figure 6.1 that we have considered in Theorem 69. For example we can construct a Hardy type quantum distribution for the DAG in Figure 6.1. The quantum state at B and the POVMs at F and D would be different but the proof of it Non-Classicality follows the same as what we have presented here.

Similarly, our technique can be employed for even the triangle scenario [34] in Figure 6.4. Here a distribution that is feasible in the triangle scenario is generated by taking $a := (A, X),\ b := (B, Y),\ c := (X, Y),\ \lambda_2 = \lambda_3 = \frac{1}{2}(|00\rangle\langle 11| + |11\rangle\langle 11|)$ and $\lambda_1 = |\psi\rangle$, a maximally entangled state and where X, Y are the measurement settings for the two parties (a and b with respective outputs A and B) in bases so chosen that the CHSH inequality is violated by $|\psi\rangle$. Now, any distribution that is Classical in the triangle scenario follows:[6]

$$\begin{aligned}\sum_{c,\lambda_1,\lambda_2,\lambda_3} & P(a,b,c|\lambda_1,\lambda_2,\lambda_3) \\ &= \sum_{c,\lambda_1,\lambda_2,\lambda_3} P(c|a,b,\lambda_1,\lambda_2,\lambda_3)P(\lambda_2|a,b,\lambda_1,\lambda_3) \\ &\qquad P(\lambda_3|a,b\lambda_1)P(a,b,\lambda_1). \quad (6.33)\end{aligned}$$

Summing over $c, \lambda_1, \lambda_2, \lambda_3$, having $a := (A, X),\ b := (B, Y),\ c := (X, Y)$ and using $(a \perp\!\!\!\perp_{\text{CI}} b|\lambda_1)$ along with, $(c \perp\!\!\!\perp_{\text{CI}} \lambda_1) \implies (X \perp\!\!\!\perp_{\text{CI}} \lambda_1), (Y \perp\!\!\!\perp_{\text{CI}} \lambda_1)$ we get:

$$\begin{aligned}P(a,b,\lambda_1) &= \sum_{\lambda_1} P(a|\lambda_1)P(b|\lambda_1)P(\lambda_1) \\ &= \sum_{\lambda_1} P(A|X,\lambda_1)P(b|Y,\lambda_1)P(X|\lambda_1)P(Y|\lambda_1)P(\lambda_1) \\ &= \sum_{\lambda_1} P(A|X,\lambda_1)P(b|Y,\lambda_1)P(X)P(Y)P(\lambda_1) \quad (6.34)\end{aligned}$$

Thus, our achievable distribution in the triangle scenario will need to respect the CHSH inequalities [18] which it does not, thereby giving us a Classical-Quantum gap in the triangle scenario. Notice that $(c \perp\!\!\!\perp_{\text{CI}} \lambda_1) \implies (X \perp\!\!\!\perp_{\text{CI}} \lambda_1), (Y \perp\!\!\!\perp_{\text{CI}} \lambda_1)$ is equivalent to Fritz's corollary 2.15 [34] but here we have proven the Classical infeasibility of the given probability distribution in a much simpler way and without resorting to any entropic inequalities unlike Fritz had to [34].

[6] The essential trick which was also implicitly used in the proof of Theorem 69 is that one should factorize the probabilities, here $P(a,b,c|\lambda_1,\lambda_2,\lambda_3)$, in such a way that the sum over all the variable that are not required in the final distribution can be performed at the first instance.

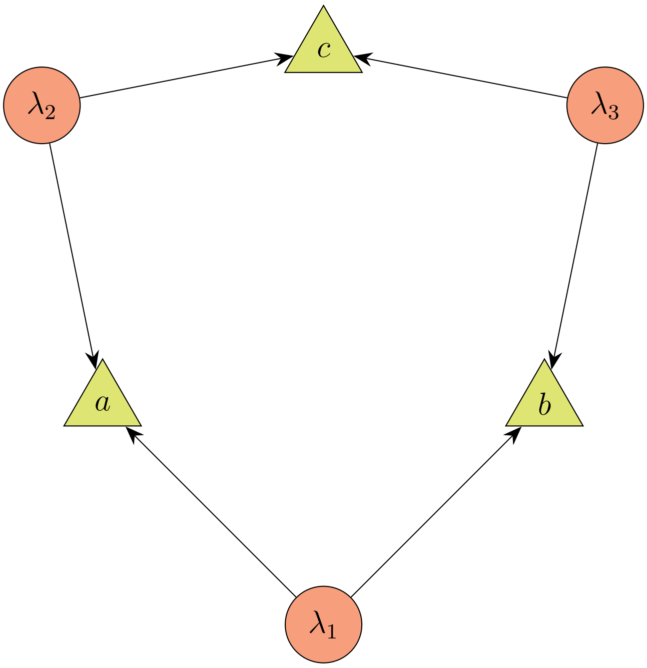


Figure 6.4: The triangle scenario.

6.4 Conclusion

In this chapter we formed a technique to embed the Bell scenario in various DAGs and under the assumption of particular observed conditional independences, we could show that the DAG under consideration in-fact supports a classical-quantum gap, that is for the DAG we have $\mathcal{C}_G \subset \mathcal{Q}_G$. Together with an upcoming work with researchers at the Perimeter Institute, we have collectively shown that all scenarios up to total six nodes which have $\mathcal{C}_G \neq \mathcal{I}_G$ also have $\mathcal{C}_G \neq \mathcal{Q}_G$. It is interesting to note that all the causal scenarios of up to total six nodes which HLP characterize as supporting NON-CLASSICAL correlations can in fact support NON-CLASSICAL QUANTUM-CORRELATIONS. This could be a potential hint towards the conjecture that all INTERESTING scenarios as classified in [53] can in fact support NON-CLASSICAL QUANTUM-CORRELATIONS.

F	**E**	**D**	**C**	**P(F,E,D,C)**
0	0	0	0	$\frac{1}{8}\cos^2(\theta)$
2	0	0	0	$\frac{1}{8}\sin^2(\theta)$
0	0	0	1	$\frac{1}{8}\cos^2(\theta)$
2	0	0	1	$\frac{1}{8}\sin^2(\theta)$
0	0	1	0	$\frac{1}{8}\sin^2(\theta)$
2	0	1	0	$\frac{1}{8}\cos^2(\theta)$
0	0	1	1	$\frac{1}{8}\sin^2(\theta)$
2	0	1	1	$\frac{1}{8}\cos^2(\theta)$
1	1	0	0	$\frac{1}{16}(\cos(\theta) + \sin(\theta))^2$
3	1	0	0	$\frac{1}{16}(\cos(\theta) - \sin(\theta))^2$
1	1	0	1	$\frac{1}{16}(\cos(\theta) - \sin(\theta))^2$
3	1	0	1	$\frac{1}{16}(\cos(\theta) + \sin(\theta))^2$
1	1	1	0	$\frac{1}{16}(\cos(\theta) - \sin(\theta))^2$
3	1	1	0	$\frac{1}{16}(\cos(\theta) + \sin(\theta))^2$
1	1	1	1	$\frac{1}{16}(\cos(\theta) + \sin(\theta))^2$
3	1	1	1	$\frac{1}{16}(\cos(\theta) - \sin(\theta))^2$

Table 6.1: The non-zero $P\Big(F, E, D, C\Big)$ obtained quantum mechanically according to the given description.

F	**E**	**D**	**C**	**P$\Big($F,E,D,C$\Big)$**
0	0	0	0	0
2	0	0	0	$\frac{1}{4}\frac{\cos^2(\theta)}{(1+\cos^2(\theta))}$
0	0	0	1	$\frac{1}{4}\frac{\cos^4(\theta)}{(1+\cos^2(\theta))}$
2	0	0	1	0
0	0	1	0	$\frac{1}{4}\frac{\cos^2(\theta)}{(1+\cos^2(\theta))}$
2	0	1	0	$\frac{1}{4}\frac{\sin^2(\theta)}{(1+\cos^2(\theta))}$
0	0	1	1	$\frac{1}{4}\frac{\cos^2(\theta)\sin^2(\theta)}{(1+\cos^2(\theta))}$
2	0	1	1	$\frac{1}{4}\frac{1}{(1+\cos^2(\theta))}$
1	1	0	0	$\frac{1}{4}\frac{\cos^4(\theta)}{(1+\cos^2(\theta))}$
3	1	0	0	$\frac{1}{4}\frac{\cos^2(\theta)\sin^2(\theta)}{(1+\cos^2(\theta))}$
1	1	0	1	$\frac{1}{4}\frac{\cos^4(\theta)\sin^2(\theta)}{(1+\cos^2(\theta))}$
3	1	0	1	$\frac{1}{4}\frac{\cos^6(\theta)}{(1+\cos^2(\theta))}$
1	1	1	0	0
3	1	1	0	$\frac{1}{4}\frac{1}{(1+\cos^2(\theta))}$
1	1	1	1	$\frac{1}{4}\frac{\cos^6(\theta)}{(1+\cos^2(\theta))}$
3	1	1	1	$\frac{1}{4}\frac{(\sin^3(\theta)+2\cos(\theta)\sin(\theta))^2}{(1+\cos^2(\theta))}$

Table 6.2: The required $P\Big(F, E, D, C\Big)$ obtained quantum mechanically according the description given in the text.

— 7 —

Digression into Accelerating Fourier-Motzkin Elimination

7.1 Introduction

Developed by Fourier and rediscovered by Motzkin, Fourier-Motzkin elimination is an algorithm for eliminating variables from a system of inequalities and consequently determining their solution set. It could be computationally a very expensive algorithm, especially when the system of inequalities is very large or contains a lot of variables. Here, we propose an algorithmic approach to make the computation more feasible. Our algorithmic approach is based on how to select a faster order of elimination of the variables. In the problems we considered and with standard Fourier-Motzkin elimination, our computation was was so slow that it was almost infeasible, but with our algorithmic approach we could significantly accelerate Fourier-Motzkin elimination and make the computation faster. Apart from that, our strategy can be computationally implemented simply by adding it to the normal Fourier-Motzkin elimination algorithm. We also compare our results to show that Fourier-Motzkin elimination used together with our algorithmic strategy is much faster for random polytopes than just using Fourier-Motzkin elimination by itself. Lastly, we provide a python implementation of our accelerated Fourier-Motzkin elimination algorithm [47] including the usage of Imbert's theorems as well.

Fourier-Motzkin (FM from here-on) [30] elimination is a method for solving a system of linear inequalities by successive elimination of the unknown variables in the system. For some given $A \in \mathbb{R}^{m,n}$, $b \in \mathbb{R}^m$, consider a system of linear inequalities in variables $y \in \mathbb{R}^n$, where the following vector inequality is to be considered component wise,

$$Ay \leq b. \tag{7.1}$$

This is a system of m inequalities in n variables. The solution set to a system of linear inequalities (like equation 7.1) is called a polyhedron[1], usually denoted by P, where

$$P = \{y \in \mathbb{R}^n : Ay \leq b\}. \tag{7.2}$$

The FM algorithm is a generalization of Gaussian elimination to a system of linear inequalities. The idea builds upon successively eliminating variables from the system of linear inequalities at each iteration to get a new system of linear inequalities, such that the solution to this new system with a variable eliminated is the same as the solution to the original system over the remaining variables. After eliminating all the variables but one, say y_i, we obtain the solution set for y_i for which the original system of inequalities will also be satisfied. Substituting y_i into the system of inequalities in the last iteration will give the set of solutions for y_{i-1}. We can continue so on to get the full solution set to the initial system of inequalities. More over any solution to the initial system can be found in this way. In most cases though, one is not concerned with the solution set to the system of inequalities but just with the elimination of variables from the system. Elimination of variables just projects the original set of solutions over all the variables onto the set of the remaining variables. If the solution to the system of inequalities is imagined as a polyhedron in $\mathbb{R}^n$, then the elimination of a variable projects the polyhedron onto $\mathbb{R}^{n-1}$, such that the range of the remaining variables remains the same.

There are a couple of important applications of Fourier-Motzkin elimination, some of the most important ones being:

1. Extension of Fourier-Motzkin elimination to Integer Programming Problems [21, 106].
2. Generating Shannon and non-Shannon-type entropic inequalities corresponding to particular sets of probability distributions [17].
3. In adjustable robust optimisation problems [115, 109].
4. Integration over a higher dimensional polyhedron by using Fourier-Motzkin elimination to decompose the domain of integration into simplified different regions [7, 92, 96, 85].
5. Projecting a higher-dimensional polyhedron onto a lower-dimensional polyhedron.
6. For converting between Half-Space (H-representation) and Vertex representation (V-representation) of a polyhedron.

In what follows, we show the standard Fourier-Motzkin algorithm and then give an example to demonstrate it.

[1] We use the definition of the polyhedron according to which it need not necessarily be bounded and can very well extend to infinity

7.2 Fourier-Motzkin Elimination

Consider the system of linear inequalities

$$Ay \leq b \tag{7.3}$$

where $A \in \mathbb{R}^{m,n}$, $b \in \mathbb{R}^m$ and $y \in \mathbb{R}^n$ and we take a set $\mathcal{S} := \{1, 2, 3, 4, ..., m\}$ to index the inequalities.

Writing the inequalities explicitly, we have:

$$\begin{aligned}
a_{11}y_1 + a_{12}y_2 + a_{13}y_3 + a_{14}y_4 + \ldots\ldots + a_{1n}y_n &\leq b_1 \\
a_{21}y_1 + a_{22}y_2 + a_{23}y_3 + a_{24}y_4 + \ldots\ldots + a_{2n}y_n &\leq b_2 \\
a_{31}y_1 + a_{32}y_2 + a_{33}y_3 + a_{34}y_4 + \ldots\ldots + a_{3n}y_n &\leq b_3 \\
a_{41}y_1 + a_{42}y_2 + a_{43}y_3 + a_{44}y_4 + \ldots\ldots + a_{4n}y_n &\leq b_4 \\
\vdots \qquad\qquad\qquad\qquad & \vdots \\
\vdots \qquad\qquad\qquad\qquad & \vdots \\
a_{m1}y_1 + a_{m2}y_2 + a_{m3}y_3 + a_{m4}y_4 + \ldots + a_{mn}y_n &\leq b_m
\end{aligned} \tag{7.4}$$

Suppose we wish to eliminate the variable y_s where $s \in \{1, \ldots n\}$. We adopt the following steps:

1. We begin by partitioning the original system of inequalities $Ay \leq b$ into three sets, I^+, I^-, I^0. I^+ denotes the set of all inequalities in which the coefficient of y_s is strictly positive, I^- denotes the set of all inequalities in which the coefficient of y_s is strictly negative and I^0 denotes the set of all inequalities in which the coefficient of y_s is 0.

2. If I^+ is empty, then the inequalities in I^- are ignored. If I^- is empty, then the inequalities in I^+ are ignored. And in both the cases the inequalities in I^0 are copied down to form the new system of inequalities in which the variable y_s is eliminated.[2] If both I^+ and I^- are non-empty then the following steps are taken to eliminate the variable y_s from the system of inequalities.

    - For every inequality $i \in I^+$ where the coefficient of y_s is strictly positive, i.e., $a_{is} > 0$, we multiply each of the i^{th} inequality by $\frac{1}{|a_{is}|}$ and rewrite every inequality $i \in I^+$ as:

    $$y_s \leq \frac{1}{|a_{is}|}\left(b_i - \sum_{k \neq s} A_{ik}y_k\right) \quad \forall i \in I^+. \tag{7.5}$$

    There are $|I^+|$ such inequalities.

[2] If even I^0 is empty then the new system of inequalities which does not contain the variable y_s is simply the null system.

- For every inequality $j \in I^-$ where the coefficient of y_s is strictly negative, i.e., $a_{js} < 0$, we multiply each of the j^{th} inequality by $\frac{1}{|a_{js}|}$ and rewrite every inequality $j \in I^-$ as:

$$-y_s \leq \frac{1}{|a_{js}|}\left(b_j - \sum_{k \neq s} A_{jk} y_k\right) \quad \forall j \in I^-. \tag{7.6}$$

  There are $|I^-|$ such inequalities.

- Now a new system of linear inequalities is formed by:

  a) copying down all the inequalities $l \in I^0$ in which the coefficient of y_s, i.e., $a_{is} = 0$. There are $|I^0|$ such inequalities.

  b) $|I^+| \times |I^-|$ new inequalities are obtained by adding each inequality $i \in I^+$ in which the coefficient of y_s is strictly positive, i.e., $a_{is} > 0$ to each inequality $j \in I^-$ in which the coefficient of y_s is strictly negative, i.e., $a_{js} < 0$. These $|I^+| \times |I^-|$ inequalities are of the form:

$$0 \leq \frac{1}{|a_{js}|}\left(b_j - \sum_{k \neq s} A_{jk} y_k\right) + \frac{1}{|a_{is}|}\left(b_i - \sum_{k \neq s} A_{ik} y_k\right) \quad \forall i \in I^+, \quad \forall j \in I^-. \tag{7.7}$$

  and thereby do not contain the variable y_s.

3. The new system contains $|I^+| \times |I^-| + |I^0|$ inequalities without the variable y_s.

By eliminating the variable y_s we get a new polyhedron P', which is the projection of $P = \{y \in \mathbb{R}^n : Ay \leq b, A \in \mathbb{R}^{m,n}, b \in \mathbb{R}^m\}$ into the space $\mathbb{R}^{n-1}$ spanned by the variables $y_1, \ldots y_{s-1}, y_{s+1}, \ldots y_n$. Similarly, we can eliminate the other variables $y_1, \ldots y_{s-1}, y_{s+1}, \ldots y_{n-1}$ in the next successive elimination steps. Finally, we get a system of the form $l \leq y_n \leq u$ where l and u are some constant numbers. If $l > u$, then the system $Ay \leq b$ does not have a solution. Otherwise, $y_n \in [l, u]$ and we can then choose y_{n-1} in a interval that depends on y_n etc. Proceeding with this back substitution, one gets a set of solutions $y = (y_1, y_2, y_3, y_4 \ldots \ldots y_n)$ for the system of equations $Ay \leq b$. Moreover, any solution to the system $Ay \leq b$ can be produced in this way.

We provide a simple example to demonstrate the working of this algorithm. It is clear that the Fourier-Motzkin algorithm works equally well for an inequality system $Ay \geq b$ because this system can be equivalently written as $-Ay \leq -b$. Also it is trivial to see that the definition of the sets I^+, I^-, I^0 remains exactly the same when considering the inequality system $Ay \geq b$. This is simply because the Fourier-Motzkin algorithm works equivalently for such an inequality system. Hence, consider the following system of linear inequalities:

$$\begin{cases} 2x - 3y \geq 1 \\ 5x + 3y \geq 1 \\ 8x - 3y \geq 1 \\ x \geq 0 \\ 3y \geq 0 \end{cases} \tag{7.8}$$

We wish to first eliminate the variable y. For the elimination of the variable y, inequalities two and five constitute the set I^+, inequalities one and three constitute the set I^- and the fourth inequality constitutes the set I^0. Following the algorithm, we eliminate the variable y by adding the first and the second inequality, the second and the third inequality, the first and the fifth inequality and the third and the fifth inequality to get:[3]

$$\begin{cases} 7x \geq 2 \\ 2x \geq 1 \\ 13x \geq 1 \\ 8x \geq 0 \\ x \geq 0 \end{cases} \tag{7.9}$$

Hence for x, we get that $x \geq \frac{1}{2}$, or that $x \in [0.5, \infty)$. Since, this new system of inequalities in 7.9 has a solution for x, we know that the original system of inequalities in 7.8 has a solution for (x, y). Now, if we set set $x = 0.5$, for y, we get, $y \leq 0$, $y \geq -0.5$, $y \leq 1$, $y \geq 0$ and thus y satisfies $y = 0$ and *one* solution for the system of inequalities in 7.8 is $(x, y) = (0.5, 0)$. However, if we set $x = 1$, we get $y \geq 0$, $y \leq 0.3333$, $y \geq -1.3333$, $y \leq 2.3333$ and thus for y we have, $y \in [0, 0.3333]$. Therefore, *another* solution to the system of inequalities in 7.8 is $(x, y) = (1, 0.3333)$. What we have done is that we have projected our initial 2-dimensional polyhedron from the $x - y$ plane onto the x axis. The resulting 1-dimensional polyhedron represents constraints on just the variable x. Another simpler example but in in three dimensions i.e., in $\mathbb{R}^3$ will help demonstrate the projection of a higher dimensional polyhedron onto a lower dimensional polyhedron via elimination of variables. Consider

[3] The fourth inequality is retained since the coefficient of y is 0 in it.

the system of inequalities:

$$\begin{cases} x \geq -2 \\ x \leq 2 \\ y \geq -2 \\ y \leq 2 \\ z \geq -2 \\ z \leq 2 \\ 6x - y + 2z \leq 4 \\ x + y - z \leq -1 \end{cases} \tag{7.10}$$

This system of Equations 7.10 defines the polyhedron in Figure 7.1. The region from the perspective of the $x - y$ plane is shown in Figure 7.2. On following the algorithm for eliminating z, we obtain the following system:

$$\begin{cases} x \geq -2 \\ x \leq 2 \\ y \geq -2 \\ y \leq 2 \\ 0 \leq 4 \\ 8x + y \leq 2 \\ x + y \leq 1 \\ 6x - y \leq 8 \end{cases} \tag{7.11}$$

This system represents constraints on just the variables x, y because z has been eliminated. The lower dimensional polyhedron that corresponds to Equations 7.11 is shown in Figure 7.3 and we get it by projecting the polyhedron in Figure 7.1 on the $x - y$ plane. Notice how the polyhedron in Figure 7.3 represents the projection of the polyhedron in Figure 7.2 onto the $x - y$ plane.

If on the other hand the original system of equations has no solution for example as for the following

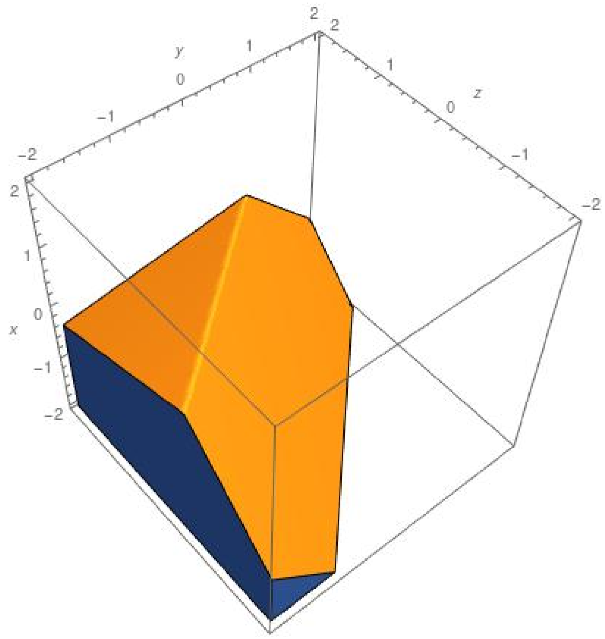


Figure 7.1: Polyhedron corresponding to the constraints in Equation 7.10.

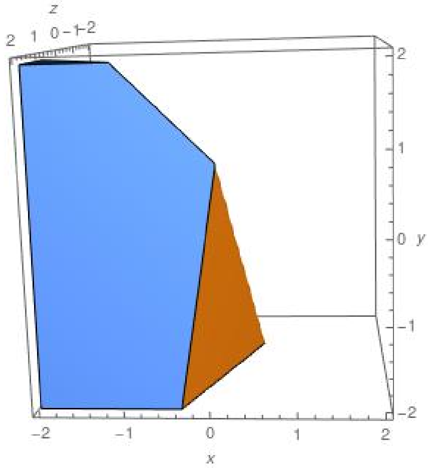


Figure 7.2: Polyhedron corresponding to the constraints in Equation 7.10 as seen from the perspective of the $x - y$ plane.

system of inequalities:

$$\begin{cases} x \geq 1 \\ x \leq -1 \\ y \geq -1 \\ y \leq 1 \end{cases} . \tag{7.12}$$

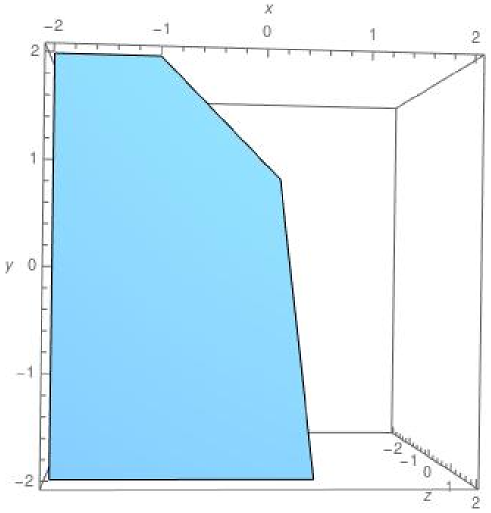


Figure 7.3: Elimination of z in Equations 7.10 results in the projection of the polyhedron in Figure 7.1 on the $x - y$ plane.

Then on eliminating x, we get:

$$\begin{cases} 0 \geq 2 \\ y \geq -1 \\ y \leq 1 \end{cases} . \tag{7.13}$$

That is we get an apparent $0 \geq 2$ contradiction when x is eliminated and thus we get no solution for y as well.

7.2.1 CONSTRAINTS AND COMPLEXITY

As is evident from Equation 7.9, elimination of variables produces redundant constraints (inequalities that are implied by other inequalities[4]). If the system of inequalities to begin with is very big or consists of a lot of variables then number the of redundant constraints produced in the elimination of each variable can grow enormously. In the worst-case scenario, an elimination of n inequalities can result in at the most $\frac{n^2}{4}$ inequalities in each step. Hence, running d successive steps can result in, at most $4(n/4)^{2^d}$ inequalities, a doubly exponential growth. Thus, the algorithm produces many unnecessary and redundant inequalities, which result in this doubly exponential growth. However, it is known that the number of independent constraints grows only as a single exponential [64]. We

[4] For eg: if we have a system of inequalities $x > 2$, $x > 3$, $x > 4$, then the inequalities $x > 2$, $x > 3$ are redundant given the inequality $x > 4$.

can eliminate the redundant inequalities using linear programming or using Imbert's acceleration theorems, or possibly both applied together. But, if we begin with a large number of inequalities in a large number of variables, the problem we may encounter is that of the redundancies growing so large that linear programming (by any interior point algorithm [51]) and Imbert's theorems to eliminate them become computationally infeasible.

7.2.2 Imbert's acceleration theorems

Imbert introduced two acceleration theorems to detect redundancies in the system of linear inequalities generated after Fourier-Motzkin elimination of a variable from an initial system of linear inequalities. We briefly state his theorem here. The interested reader is pointed to [49, 50] for their proof. To state Imbert's theorem, we require a few definitions.

Definition 71. *Let $\mathcal{S}$ be a set of inequalities in n variables. For each inequality $i \in \mathcal{S}$ [49, 50], we define three sets; histories-:$\mathcal{H}_i$, explicit variables-:$\mathcal{E}_i$ and implicit variables-: $\mathcal{I}_i$ as follows-:*

1. *$\mathcal{H}_i$: For each inequality $i \in \mathcal{S}$, $\mathcal{H}_i$ is the set of all the inequalities that were used in its construction.*
2. *$\mathcal{E}_i$: For each inequality $i \in \mathcal{S}$, $\mathcal{E}_i$ is the set of all variables that have been explicitly eliminated in that inequality using Fourier-Motzkin elimination combining its histories to generate it.*
3. *$\mathcal{I}_i$: For each inequality $i \in \mathcal{S}$, $\mathcal{I}_i$ is the set of all variables that have been implicitly eliminated in that inequality using Fourier-Motzkin elimination combining its histories to generate it. A variable is said to be implicitly eliminated for an inequality i, if the following three conditions are satisfied: it occurs in at least one inequality of $\mathcal{H}_i$, it does not occur in the inequality i, it is not explicitly eliminated for the inequality i.*

Further, the set of officially eliminated variables at each iteration is denoted as $\mathcal{O}_k$, where k denotes the index of the current iteration. Now we state Imbert's theorem.

Theorem 72 (Imbert's First Theorem)**.** *If for any inequality $i \in \mathcal{S}$, the condition*

$$Cardinality(\mathcal{H}_i) \leq 1 + Cardinality(\mathcal{E}_i \bigcup (\mathcal{I}_i \bigcap \mathcal{O}_k)) \tag{7.14}$$

does not hold true then that inequality is redundant given the others in the system.

Imbert's first theorem, i.e. Theorem 72 can be used to detect inequalities that are definitely redundant. However the theorem does not detect *all* the possible redundant inequalities. Thus the theorem is only a sufficient condition to detect redundant inequalities. There is a second theorem due to Imbert, but we do not consider it here since it guarantees neither redundancy nor non-redundancy of an inequality.

If we try to use Imbert's first theorem to delete those redundant inequalities that it can detect and then use linear programming to detect further remaining redundancies in this reduced list of inequalities, we expect a total speedup in the process. This is because Imbert's theorem acts only locally on the concerned inequalities. To check the conditions of Theorem 72, we need only the history of the concerned inequality and the list of explicitly and implicitly eliminated variables at that iteration. On the other hand linear programming uses the whole system of inequalities to check whether the concerned inequality is redundant or not. In this sense linear programming acts globally on the whole system of inequalities. The local nature of Imbert's theorems is what makes them acceleration theorems.

However, as we show in the following section Imbert's first Theorem i.e. Theorem 72 is incompatible with linear programming to remove redundancies. Using his second theorem along with linear programming is of no point since it guarantees neither redundancy nor non-redundancy of an inequality, whereas linear programming on the other hand is already both a sufficient and a necessary condition to check non-redundancy of an inequality.

7.2.3 Incompatibility of Imbert's acceleration theorems with Linear Programming to remove redundancies

The first thing we tried to accelerate Fourier-Motzkin elimination was to combine Imbert's theorems along with linear programming to detect redundancies quickly. We did this in the following way:

1. First complete the Fourier-Motzkin algorithm, then apply Imbert's first theorem, i.e. Theorem 72 to check if the generated inequality is redundant. If the conditions of Theorem 72 are not met then the inequality is redundant and we delete it from the new system of inequalities formed after the elimination of a variable. However, if the conditions of Theorem 72 are met then we can not conclude anything about the inequality and in this case we keep the inequality in the newly formed system.

2. The new system formed is evidently smaller in terms of the number of inequalities in it than the new system formed in the case had we not applied Imbert's theorem at all. Finally, we use linear programming to detect redundancies on this smaller system formed after the elimination of a variable.

3. We repeat steps 1 and 2 in that order for all the consecutive variable eliminations.

Since this new system that linear programming checks for redundancies is strictly smaller than the system we would get had we not applied Imbert's theorem at all, we expect a speedup in linear programming and consequently a speedup in the whole Fourier-Motzkin elimination of all the variables. However, in the end after all the eliminations we get a final system of inequalities which is not implied by the final system of inequalities formed had we not used Imbert's theorem anywhere but just used linear programming to detect redundancies everywhere.

The conclusion that Imbert's theorem cannot be used together with linear programming to detect redundancies is best evident through a very simple example. The example we consider shows exactly where the problem occurs.

Consider the following system of inequalities in variables w, x, y, z, with no bounds on the variables.

$$\begin{cases} -x + w \leq 0 \\ x + y \leq 0 \\ -x + y + z - w \leq 0 \\ x - y - z + w \leq 0 \end{cases} . \tag{7.15}$$

We wish to eliminate variables x, z in that order. After eliminating x using the Fourier-Motzkin algorithm we get the following new system of inequalities:

$$\begin{cases} y + w \leq 0 \\ -y - z + 2w \leq 0 \\ 2y + z - w \leq 0 \\ 0y + 0z + 0w \leq 0 \end{cases} . \tag{7.16}$$

Linear programming shows that $y + w \leq 0$ is redundant since $y + w \leq 0 = -y - z + 2w + 2y + z - w \leq 0$ and that $0y + 0z + 0w \leq 0$ is redundant. However, Imbert's first theorem cannot detect either of them as redundant. This can be verified by checking the conditions of Theorem 72. Hence we do not delete them and proceed with the system of inequalities in Equation 7.16 to eliminate the variable z. On eliminating z we get:

$$\begin{cases} y + w \leq 0 \\ y + w \leq 0 \\ 0y + 0z + 0w \leq 0 \end{cases} . \tag{7.17}$$

Imbert's first theorem detects the first occurrence of the inequality $y + w \leq 0$ as redundant. This can again be verified by checking the conditions of Theorem 72. Hence as the final answer we get:

$$\begin{cases} y + w \leq 0 \\ 0y + 0z + 0w \leq 0 \end{cases} . \tag{7.18}$$

This is indeed the correct answer which can be checked by hand. However now consider that we use both linear programming and Imbert's first theorem to detect redundancies. Then since after elimination of x linear programming detects both $y + w \leq 0$ and $0y + 0z + 0w \leq 0$ as redundant we delete them at this step. Thus, after elimination of x, our system is just:

$$\begin{cases} -y - z + 2w \leq 0 \\ 2y + z - w \leq 0 \end{cases} . \tag{7.19}$$

Now eliminating z gives us :

$$\begin{cases} y + w \leq 0 \end{cases} . \tag{7.20}$$

But now Imbert's theorem detects this inequality as redundant. This is because this inequality $y + w \leq 0$ is formed from the inequalities $y - z + 2w \leq 0$ and $2y + z - w \leq 0$. Thus the total history set for the inequality $y + w \leq 0$ is now our whole original system: $\{-x + w \leq 0,\ \ x + y \leq 0,\ \ -x + y + z - w \leq 0,\ \ x - y - z + w \leq 0\}$. The list of explicitly eliminated variables is just $\{x, z\}$. Evidently conditions of Theorem 72 are not met and the inequality is detected as redundant and removed from the system. As the final answer we get the empty system:

$$\begin{cases} \end{cases} \tag{7.21}$$

which is an incorrect answer. Note that when the same inequality $y + w \leq 0$ was generated in Equation 7.16 after elimination of x it was not detected redundant by Imbert' s first theorem because the conditions of Theorem 72 were met as follows:

$$Cardinality\ \ \{x\} + 1 \geq \ \ Cardinality\ \ \{-x + w \leq 0,\ \ x + y \leq 0\} \tag{7.22}$$

But the inequality $y + w \leq 0$ formed in Equation 7.20 after elimination of z by adding the the equations in Equation 7.19 is incorrectly detected as redundant by Imbert's first theorem as shown above in Equation 7.21. This is because linear programming had already detected a version of $y + w \leq 0$ as redundant in Equation 7.16 and deleted it. Imbert's theorems being local in nature could not know this and thus faltered.

Numerous such issues occur when we try to combine the locally acting Imbert's theorem and the globally acting linear programming based redundancy removal. Therefore, we cannot combine Imbert's first theorem along with linear programming based redundancy removal at each step. This can be checked with the code [47] we have provided along with this thesis.

However, two things are still possible:

1. We just use Imbert's theorems to remove redundancies after elimination of each variable. And only in the end after all eliminations have been carried out do we use linear programming to remove the remaining redundancies. However, on trying, this did not turn out to be useful in many cases, especially for the polytopes we consider in Section 7.5.

2. Second thing that is possible is to use Imbert's first theorem to detect redundancies for a fixed number of eliminations and then, in the next elimination use linear programming to remove remaining redundancies. Then reset the whole problem from that particular elimination and begin a new problem of eliminating the remaining variables, repeating the same process again till all the eliminations succeed. In some of the examples that we considered this did not blow up and was computationally better than the case where we used only linear programming to remove redundancies in the same problem and did not use Imbert's theorems at all. Such examples included those polytopes for which Imbert's theorem could detect a lot of redundancies at each iteration. However, in some examples, this technique was actually not better than just using linear programming to remove redundancies at each step. Since this technique worked for some polytopes we have provided the code [47] to implement it along with this thesis. Below we summarise this approach in the form of Algorithm 1.

7.3 Eliminating Variables in an Efficient Order

In the last section we saw how the elimination of a variable leads to a new system of linear inequalities. Until now we did not introduce any particular strategy to select which variable to eliminate during any given iteration of the algorithm. By simple extrapolation of the result of equation 7.9, we can see that if the system of inequalities to begin with is very big then elimination of variables can produce a large number of redundancies and consequently the computation might become infeasible. In section 7.2.2 we saw that we can devise an algorithm to use Imbert's theorems along with linear programming to attack the problem of redundancy removal. Though, as mentioned, this algorithm is faster than compared to using linear programming alone, for just some polytopes and not all.

In what follows, we introduce a strategy in Section 7.3.2 to make the computation more feasible even without resorting to Imbert's acceleration theorems. Our strategy is simpler to implement

Algorithm 1: Using Fourier-Motzkin elimination along with Imbert's theorems and linear programming to detect redundancies.

```
while ($remaining\_variables\_to\_eliminate\ != 0$) do
    if ($counter <= 3$) then
        Run Fourier-Motzkin algorithm with Imbert's theorems to delete redundancies.
        Update the inequality system, histories, explicitly eliminated variables, implicitly
          eliminated variables sets.
        Increment counter by one.
    end
    else
        Run Fourier-Motzkin algorithm along with linear programming to delete redundancies.
        Update the inequality system.
        Initialize the histories, explicitly eliminated variables, implicitly eliminated variables
          sets to default.
        Set counter to zero.
    end
end
```

computationally than Imbert's acceleration theorems. Before presenting it we first show that a strategy 7.3.1 of eliminating variables which looks like advantageous fails drastically.

7.3.1 ELIMINATING THE VARIABLE THAT PRODUCES THE LEAST NUMBER OF REDUNDANT INEQUALITIES IN EACH ITERATION OF THE ALGORITHM

It is trivial to see that the number of (possibly redundant) inequalities in the new system produced after the elimination of a variable depends on which variable is eliminated. Suppose we wish to eliminate the variable y_s from the original system which has n non-redundant inequalities. Say y_s occurs in u inequalities with a positive sign, in v inequalities with a negative sign and in w inequalities with coefficient 0, such that $n = u + v + w$. Then the number of total (possibly redundant) inequalities produced after the elimination of y_s will be $[(u \times v) + w]$. We can choose to eliminate that particular variable in each step which produces the least number ($[(u \times v) + w]$) of total inequalities. If in each step we generate the least number of inequalities we expect a computational speedup which could be possibly seen if the number of inequalities and variables to begin with are both large.In summary the hope is that this technique might even work for complicated examples.

For example consider again the simple system of linear inequalities:

$$\begin{cases} 2x - 3y \geq 1 \\ 5x + 3y \geq 1 \\ 8x - 3y \geq 1 \\ \qquad x \geq 0 \\ \qquad 3y \geq 0 \end{cases} \tag{7.23}$$

Elimination of y gives the new system

$$\begin{cases} 7x \geq 2 \\ 2x \geq 1 \\ 13x \geq 1 \\ 8x \geq 0 \\ x \geq 0 \end{cases} \tag{7.24}$$

which has $[(2 \times 2) + 1] = 5$ inequalities. Now instead of eliminating y first, consider the elimination of x first, which produces the system

$$\begin{cases} 3y \geq 0 \end{cases} \tag{7.25}$$

which has just $[(3 \times 0) + 1] = 1$ inequality. So, we see, that the number of (possibly redundant) inequalities generated after elimination of each variable depends on which variable is eliminated first. This suggests that eliminating the variable that produces the least (possibly redundant) constraints could potentially speed up the algorithm.

But in the case of a large number of inequalities it *counter-intuitively* makes the algorithm even slower than compared to a completely random order of elimination of variables. So, what this shows is that such a greedy algorithm of selecting variables to eliminate is not a good strategy. Choosing variables which generate the least total (possibly redundant) inequalities to eliminate first causes generation of an enormous number of redundant inequalities in the elimination of the other left over variables.

This suggests that we not only need to make each elimination faster (by choosing a specific variable to eliminate first) but also need to make sure that our current choice of variable for elimination is such that it leaves all the further linear programs that we would have to do when we eliminate the other variables feasible. That is, we need an elimination order of variables to globally optimise Fourier-Motzkin elimination rather than just optimising it locally because its local optimisation in almost every case leaves it globally inefficient.

7.3.2 Eliminating the variable that produces the least number of non-redundant inequalities in each iteration of the algorithm

From the approach introduced in the last section we were motivated in eliminating the variable that produces the **least number of non-redundant constraints** instead of eliminating the variable that produces the least total number of (possibly redundant) constraints. We found a significant computational advantage if we eliminated the variable that produced the least number of non-redundant inequalities at each elimination step. In order to do so, at each run of the algorithm we eliminated each remaining variable over different processors and removed redundancies produced after elimination of each variable over the respective processors. We compared the system of inequalities generated by the elimination of each variable over different processors and proceeded ahead in the algorithm with the new system of inequalities generated after the actual elimination of that particular variable from the original system of inequalities which produced the least number of non-redundant inequalities. Algorithm 2 depicts our idea systematically. In Section 7.5, we present and compare our computational results.

Now we digress briefly into Shannon-type inequalities (Sections 4.7.1 and 7.4) for causal scenarios because we consider the polytopes formed by them in Section 7.5.

7.4 Shannon-type inequalities for Causal Scenarios (description of the computational problem)

The entropy vector approach (as discussed in 4.7.1) has been recently applied to causal structures to check for their possible “non-classicality”. We refer the reader to [17] for more details. The result is that, subject to some causal constraints, we can generate sets of Shannon-type inequalities corresponding to certain sets of probability distributions that obey those causal constraints in the causal scenario. These Shannon-type inequalities can possibly help us study whether the causal scenario under consideration can be understood “clasically”. In what follows, we give only a simple example to show the computational problem that might arise by following this approach for characterising causal scenarios.

Consider that we have a causal scenario that has three random variables, A, B and C amongst which A and B depict observed nodes whereas C is a latent variable. We do not need to go into the exact structure of the scenario. There would be certain causal constraints (from the local Markov conditions [76], or as in section 3) in this scenario into which we need not go for the illustration of our

Algorithm 2: Accelerating Fourier-Motzkin elimination by eliminating that variable which produces the least number of non-redundant constraints

```
while ($remaining\_variables\_to\_eliminate$ $!= 0$) do
    function advantage ($inequalities\_system$, $remaining\_variables\_to\_eliminate$)
    Input   : The system of inequalities and the variables to eliminate.
    Output  : Return The smallest non-redundant inequality system and the variable whose
              elimination generates it.
    Begin parallel-processing.
    Run Fourier-Motzkin elimination algorithm with linear programming to remove
      redundancies individually over each core by eliminating different variables over different
      cores.
    Collect the outcomes of each core which is the new non-redundant inequality system
      generated by the elimination of the corresponding variable.
    Compare the lengths of the new non-redundant inequality systems generated by each core.
    End parallel processing.
    Return the smallest new non-redundant inequality system and the variable whose
      elimination generates it.
    Remove this variable from the list of variables left to be eliminated.
    Repeat this process till all the variables that need to be eliminated have been actually
      eliminated.
end
```

simple example. Using equations (4.19) and (4.20) we can generate the set of *Shannon-type inequalities* corresponding to these three random variables, which just contains the following inequalities.

1. $H(A|BC) = H(ABC) - H(BC) \geq 0$
2. $H(B|AC) = H(ABC) - H(AC) \geq 0$
3. $H(C|AB) = H(ABC) - H(AB) \geq 0$
4. $I(A:B|C) = H(AC) + H(BC) - H(C) - H(ABC) \geq 0$
5. $I(A:B|\phi) = H(A) + H(B) - H(\phi) - H(AB) \geq 0$
6. $I(A:C|B) = H(AB) + H(BC) - H(B) - H(ABC) \geq 0$
7. $I(A:C|\phi) = H(A) + H(C) - H(\phi) - H(AC) \geq 0$
8. $I(B:C|A) = H(AC) + H(AB) - H(A) - H(ABC) \geq 0$
9. $I(B:C|\phi) = H(B) + H(C) - H(\phi) - H(BC) \geq 0$

Here ϕ denotes the empty set.

Now, the causal constraints in the actual scenario might convert some of the above inequalities into equalities, but we can neglect that without any loss of generalisation to the actual computational problem that occurs.

We have a system of *nine* inequalities in *seven* variables. Since C is an unobserved variable, we would like to eliminate the variables $H(C), H(AC), H(BC), H(ABC)$ from our system of linear inequalities. This is just the standard Fourier-Motzkin elimination problem. For this system, we can eliminate variables easily. But consider complicated causal scenarios that could have *many* unobserved nodes and thus possibly *hundreds of variables to eliminate* from the initial system of inequalities which itself could be as large *many hundreds of inequalities*. In this case Fourier-Motzkin elimination of variables could generate *thousands of redundant inequalities* and the standard Fourier-Motzkin elimination process could be very slow. In such a situation our technique of eliminating variables is expected to be useful.

Our computational results can be found here [47]. Note that we use Mosek [65] to remove redundant inequalities.

7.5 Computational Results

For Causal Scenarios

To show our results, we first consider, for example, *four* causal scenarios. We show how the standard Fourier-Motzkin elimination encounters a lot of redundancies compared to Fourier-Motzkin elimination used with our technique of eliminating the variables. Encountering a lot of redundancies results in the standard Fourier-Motzkin elimination taking much more time than when implemented along with our Algorithm 2. The exact structure of the causal scenarios considered is not relevant to the present problem. The relevant factor is the number of inequalities in the initial system and the number of variables to eliminate. Each of the four causal scenarios we considered had more than *250* initial inequalities with more than *100* variables to eliminate. Figures 7.4, 7.5, 7.6, 7.7 [5] show in detail how our strategy of eliminating the variable that produces the least number of non-redundant inequalities in each iteration encounters significantly fewer redundancies, which consequently calls for the less time taken by it.

[5] The no strategy techniques corresponds to the standard Fourier-Motzkin elimination.

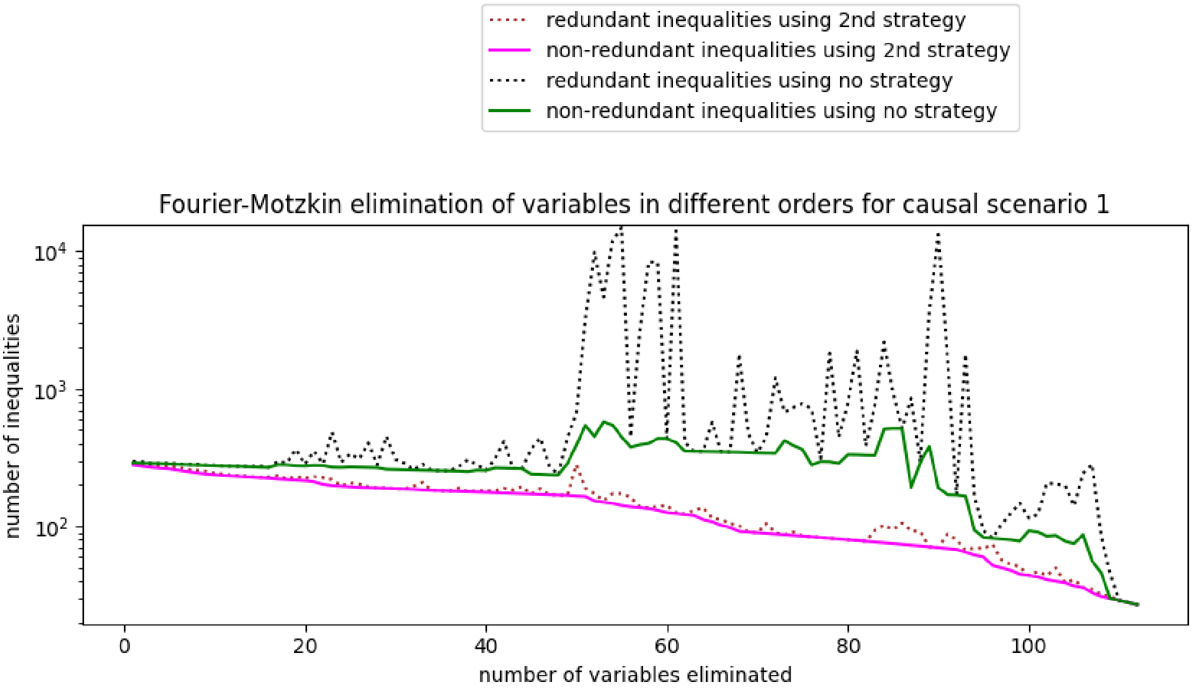


Figure 7.4: Computational benefit of our strategy in causal scenario 1

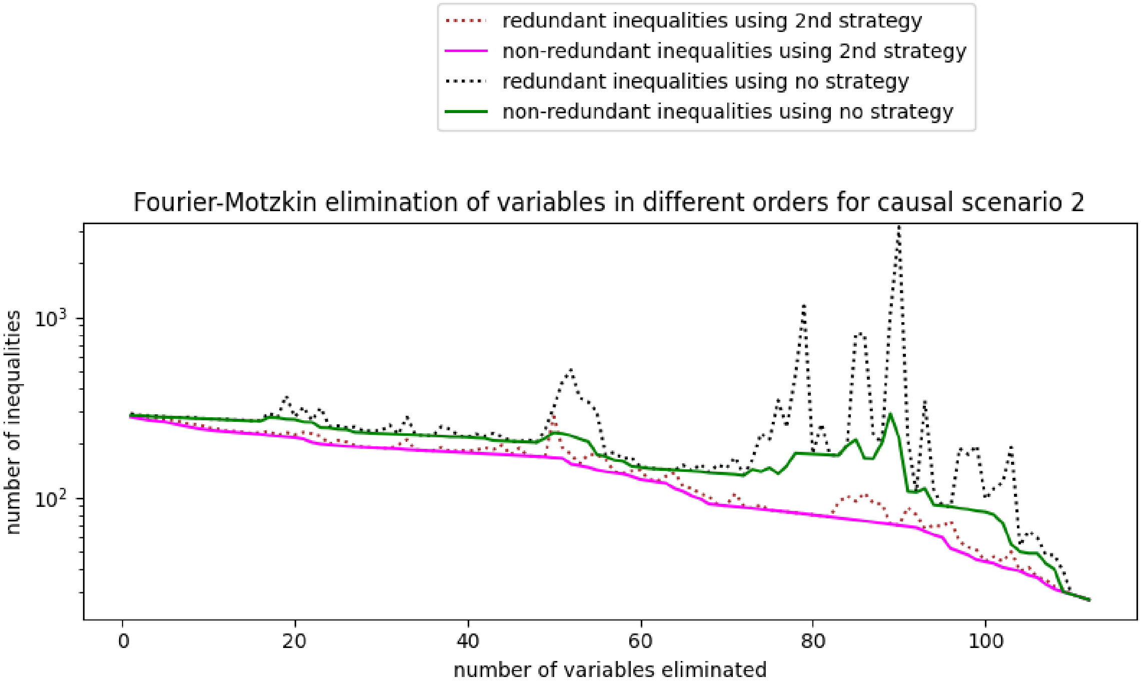


Figure 7.5: Computational benefit of our strategy in causal scenario 2

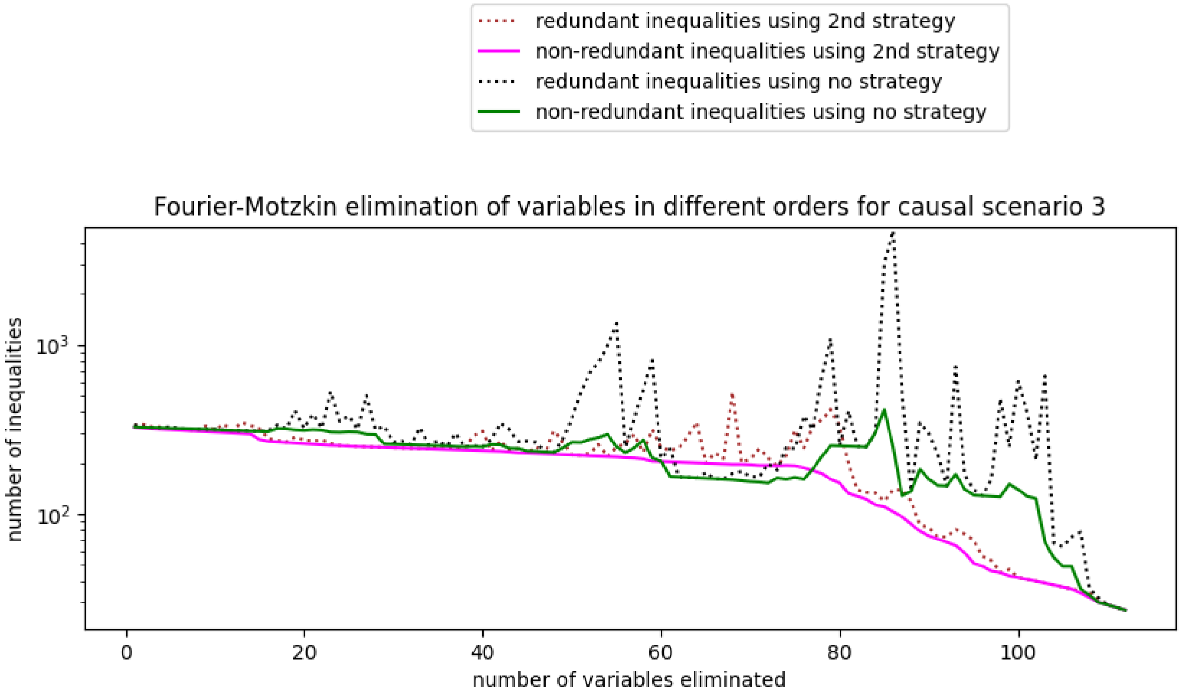


Figure 7.6: Computational benefit of our strategy in causal scenario 3

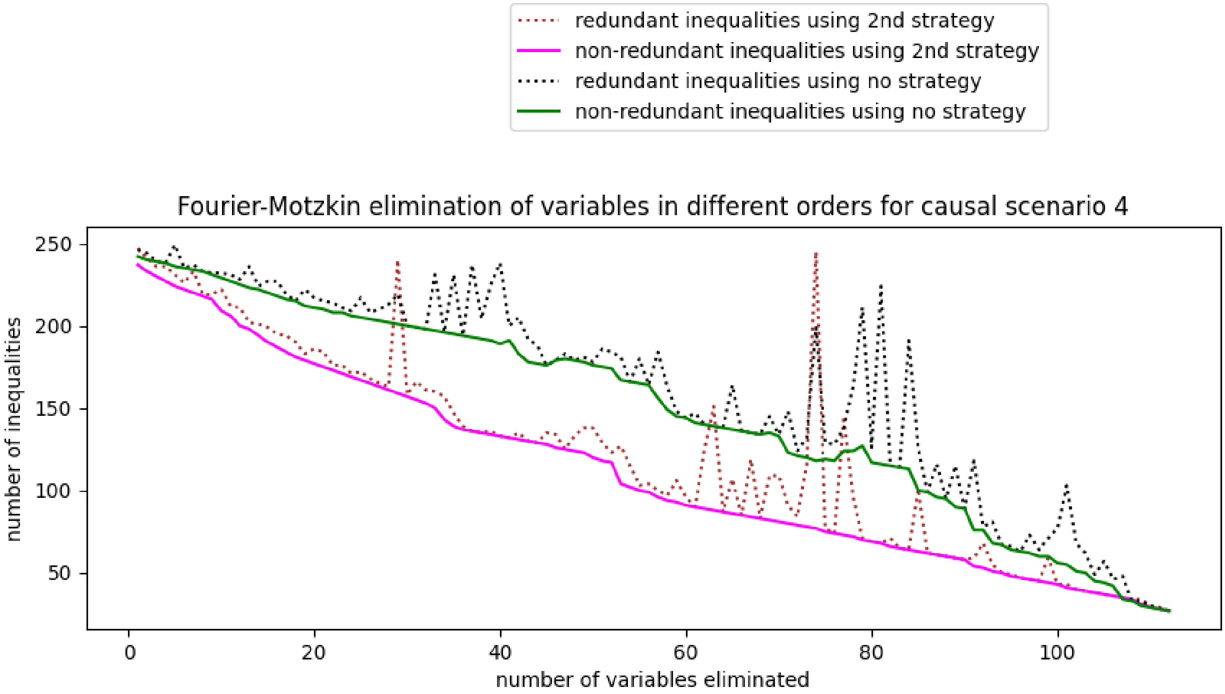


Figure 7.7: Computational benefit of our strategy in causal scenario 4

For Random Polytopes

Now we present our computational results for *six* random polytopes. Table 7.1 shows that our strategy of eliminating variables leads to much faster computation than the standard Fourier-Motzkin elimination.

Random Polytope	Initial Number of Non-Redundant Inequalities	Final Number of Non-Redundant inequalities	Total Number of Variables	Number of Variables to be Eliminated	Time Taken by Standard Fourier Motzkin elimination (sec)	Time Taken by our implementation of Fourier Motzkin elimination (sec)	Reduction Factor
1	52	14	15	12	55496.4560	4126.1571	0.07434
2	44	13	15	12	18152.0970	2536.8207	0.13975
3	46	14	15	12	3620.4586	142.7669	0.03943
4	55	5	15	12	301.6454	49.0690	0.16267
5	56	11	15	12	19652.5220	1164.9615	0.05927
6	53	10	15	12	38530.8608	6513.9060	0.16905

Table 7.1: Results for Random Polytopes

7.6 Conclusion

In this chapter we dealt with accelerating Fourier-Motzkin elimination algorithm. We saw how it is not possible to implement Imbert's theorems along with linear programming trivially. Nevertheless, there is a way around to implement both Imbert's theorems and linear programming together to detect redundancies. We saw that the naive technique of eliminating the variable that produces the least number of possibly redundant inequalities at each iteration is even slower than a random order of elimination of variables. However, our strategy of eliminating that variable which produces the least number of non-redundant inequalities at each iteration brings a significant computational advantage along with it. In chapter 8 we will discuss a few future ideas regarding this work.

— 8 —

Discussion and Conclusion

While many of the independent chapters in this thesis consist of their separate conclusions, we aim to provide here a bigger picture linking together the conclusions of the separate chapters, meanwhile also providing questions that could serve as future research directions.

Chapter 2 introduced the mathematical theory of studying cause and effect relationships among observed variables. It was shown how directed acyclic graphs can be used to model cause-and-effect relations among observed variables, via causal scenarios. Chapter 3 introduced the notion of an Interesting causal scenario. Interesting causal scenarios are those which can support Non-Classical correlations. Specifically it was shown how the HLP criterion helps to filter out all those scenarios which can never support any Non-Classical correlations; hence they are termed Non-Interesting.

We saw how in quantum foundations the challenging task of ascertaining when correlations are Non-Classical within a directed acyclic graph (or causal scenario) remains a significant area of research. Distinguishing between correlations that are Classically explainable from those that are Non-Classical and "quantumly" possible in various causal scenarios has been studied in [53, 44, 108, 55, 5, 56]. Specifically the focus is on understanding when a set of probability distributions can be explained by quantum mechanics but not by any classical theory.

With Chapter 4 began the presentation of the results. Specifically we went through a series of conditions to determine when a causal scenario is Interesting. We completed the classification of causal scenarios of three observed nodes into Interesting and Non-Interesting. For causal scenarios of four observed nodes, we could classify all but three causal scenarios into Interesting or Non-Interesting. It would be interesting to finish this classification for the three left causal scenarios of four observed nodes. This could potentially provide further affirmation of the HLP conjecture. Hence we discuss a few techniques which one can resort to while solving the problem for the three left over causal scenarios of four observed nodes.

Methods suggested to solve the left over three causal scenarios of four observed nodes.

1. *Exploring Non-Shannon-type inequalities:* In Chapter 4 we saw how one cannot solve the problem of $\mathcal{C} \neq \mathcal{I}$ for the three remaining causal scenarios if one resorts to only Shannon type inequalities [15, 17]. The idea was to detect a difference between the two concerned sets in the entropy space. The idea still more or less remains the same but instead of resorting to Shannon-type inequalities, one would now find Non-Shannon type inequalities that would help distinguish between the sets corresponding to $\mathcal{C}$ and $\mathcal{I}$ in the entropy space.
2. *Accelerating Fraser's algorithm:* In Chapter 4 we also used Fraser's algorithm [31] to find classically infeasible supports for causal scenarios. If we could find supports that are classically infeasible corresponding the scenarios, we could attest their NON-CLASSICAL nature; in other words we could show $\mathcal{C} \neq \mathcal{I}$ for them. However to show $\mathcal{C} \neq \mathcal{I}$ and find classically infeasible supports for various scenarios we needed to search at high cardinalities of the involved random variables. The computational implementation of Fraser's algorithm becomes drastically slow when the cardinalities of the involved random variables is high (usually above four). The idea being proposed is to accelerate Fraser's algorithm itself so that one can search for classically infeasible supports at higher cardinalities (higher than four) of the involved random variables.
3. *Machine Learning to solve the problem:* In [57] a neural network-based approach has been developed to check if a probability distribution itself is classically compatible with respect to a causal scenario. The proposed idea is based on generating thousands of random probability distributions and directly checking whether all of them are classically compatible with the respective three causal scenarios. If the algorithm returns a probability distribution whose “classicality" it cannot guarantee, then we can extract its support and use Fraser's algorithm to check if the support is NON-CLASSICAL or not. If the support turns out to be NON-CLASSICAL, then so would be the probability distribution itself [53] and we can certify $\mathcal{C} \neq \mathcal{I}$ for the given causal scenario.

Implementing these ideas could lead to the completion of classification of causal scenarios of four visible nodes. However, note that as mentioned before just concluding $\mathcal{C} \neq \mathcal{I}$ for causal scenarios is not sufficient to also show that $\mathcal{C} \neq \mathcal{Q}$ for them; nevertheless these causal scenarios, for which $\mathcal{C} \neq \mathcal{I}$, are the only causal scenarios for small number of nodes which are candidates for a possible $\mathcal{C} \neq \mathcal{Q}$ relation.

In Chapter 6 we discussed the question of checking whether $\mathcal{C} \neq \mathcal{Q}$ for the causal scenarios for which we concluded $\mathcal{C} \neq \mathcal{I}$. We discussed a particular technique based on the assumption of certain extra observed conditional independences in the observed probability distribution to certify $\mathcal{C} \neq \mathcal{Q}$ for these scenarios for which we have $\mathcal{C} \neq \mathcal{I}$. In particular we discussed two scenarios and showed that based on our technique one can in fact show that $\mathcal{C} \neq \mathcal{Q}$ for them. Coupled with an upcoming work by the researchers at the Perimeter Institute, our results [55] certify that $\mathcal{C} \neq \mathcal{Q}$ for all but one causal scenario of up to six total nodes for which $\mathcal{C} \neq \mathcal{I}$ and that all but one scenario in Appendix E of [44] have in fact got $\mathcal{C} \neq \mathcal{Q}$. This leads us to the following conjecture;

Conjecture: All causal scenarios that support $\mathcal{C} \neq \mathcal{I}$ also support $\mathcal{C} \neq \mathcal{Q}$.

However, all the scenarios for which we [55] and the upcoming work by the researchers at the Perimeter Institute show that $\mathcal{C} \neq \mathcal{Q}$, we do so by embedding the Bell or the Instrumental scenario in the original scenario and by exploiting the fact that each of the Bell scenario and the Instrumental scenario supports $\mathcal{C} \neq \mathcal{Q}$. It would be interesting to actually find non-trivial NON-CLASSICAL QUANTUM-CORRELATIONS in these scenarios. Here by non-trivial NON-CLASSICAL QUANTUM-CORRELATIONS we mean NON-CLASSICAL QUANTUM-CORRELATIONS found not by embedding a causal scenario which has known NON-CLASSICAL QUANTUM-CORRELATIONS, but by finding NON-CLASSICAL QUANTUM-CORRELATIONS intrinsic to the given causal scenario. Put in other words it would be interesting to show that $\mathcal{C} \neq \mathcal{Q}$ for these scenarios by finding some NON-CLASSICAL QUANTUM-CORRELATIONS intrinsic to the scenario itself, rather than showing $\mathcal{C} \neq \mathcal{Q}$ for the scenario by embedding the Bell or the Instrumental scenario in it and exploiting the known $\mathcal{C} \neq \mathcal{Q}$ result for the Bell or the Instrumental scenario respectively.

One can begin checking whether $\mathcal{C} \neq \mathcal{Q}$ or not in one of the simplest scenarios among all, depicted in Figure 8.1. We have shown [44, 53] this scenario exhibits NON-CLASSICAL correlations. Now we want to show that in this scenario we can have quantum mechanically realizable NON-CLASSICAL correlations. The reason to begin with this scenario is because we couldn't embed either the Bell scenario or the Instrumental scenario in it. So, starting from this simple scenario, we would like to build an algorithm to find non-trivial, NON-CLASSICAL QUANTUM-CORRELATIONS in various other scenarios.

Approach to the solution: In the following, we suggest how one may approach solving the problem of finding non-trivial NON-CLASSICAL QUANTUM-CORRELATIONS.

1. **Mathematical Optimization**: Finding QUANTUM-CORRELATIONS that are NON-CLASSICAL in the scenario of Figure 8.1 is a mathematical optimization problem. First we will require to simplify the optimization problem to reduce it to a sequence of semi-definite programs. This would involve implementing the See-Saw algorithm.

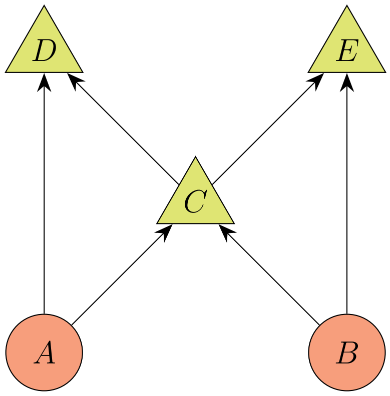


Figure 8.1: An interesting causal scenario which can support NON-CLASSICAL correlations. The problem is to check whether it can support NON-CLASSICAL QUANTUM-CORRELATIONS. As for the other figures in this thesis, triangular nodes represent observed nodes and circular nodes represent latent variables which can be entangled quantum states.

2. **Semi-Definite Programming (SDP)**: The second step would be to solve the sequence of semi-definite programs to find certain QUANTUM-CORRELATIONS that are NON-CLASSICAL for the considered scenario of Figure 8.1.

Beginning with this simple scenario, one can try to build an algorithm to find non-trivial NON-CLASSICAL QUANTUM-CORRELATIONS in various other scenarios for which $\mathcal{C} \neq \mathcal{I}$. Even though for scenarios of total nodes up to six (except the one in Figure 8.1), it is known that $\mathcal{C} \neq \mathcal{Q}$, as previously explained, it would still be interesting to find non-trivial NON-CLASSICAL QUANTUM-CORRELATIONS that are intrinsic to the scenario itself.

Chapter 5 deals with FINE-TUNING in several scenarios required to explain certain QUANTUM-CORRELATIONS. More precisely, we dealt with a list of INTERESTING scenarios and asked the question that which of those were such that they were the only possible scenarios supporting the observed conditional independences seen in them. To explain certain QUANTUM-CORRELATIONS in them would require FINE-TUNING. We conjecture that the multipartite Bell scenario is an example of such a scenario. The above question we have dealt with in this chapter, is for scenarios of six total nodes. The same question can be asked for scenarios of seven total nodes.

Chapter 6 deals in developing a technique to attest if certain INTERESTING causal scenarios are such that they can support NON-CLASSICAL QUANTUM-CORRELATIONS as well. We take a step to complete the problem of finding which causal scenarios of total six nodes that are INTERESTING can also support QUANTUM-CORRELATIONS.

Chapter 7 digresses into Fourier-Motzkin acceleration, and we propose how to combine Imbert's theorems with linear programming to remove redundancies. More importantly, we provide a technique, based on the order of elimination of variables, to accelerate the whole Fourier-Motzkin algorithm. Specifically, we give a method to eliminate the variables in a particular order; at each iteration, eliminate that variable which produces the least number of non-redundant inequalities. To find the particular order in which the variables should be eliminated, parallel programming was required. This order of elimination was seen to accelerate the whole Fourier-Motzkin algorithm than compared to a randomly selected order of elimination of variables. Now an interesting question arises. Is the order of elimination selected by our strategy actually the best and the fastest order of elimination possible or is there any other order of elimination of variables that can further reduce the computational overhead and further accelerate the Fourier-Motzkin elimination algorithm? This would certainly be a good question to look at. If the order of elimination given by us is faster than all possible orders of elimination of the variables, then it would be nice to give a rigorous mathematical proof of the fact.

8.0.1 A Future Project Proposal

In [57], it has been discussed how to train a neural network to try to detect when a probability distribution is classical. The algorithm in [57] tries to check if a given probability distribution can be reproduced classically (using appropriate latent variables). Only if the probability distribution is classical, then the algorithm returns a classical model (using appropriate hidden variables) to reproduce it. In other words if the algorithm is not able to return a classical model for the probability distribution under testing then the probability distribution may be “non-classical" (but not necessarily). It would be interesting to generalize this work and construct a neural network oracle which tries to learn if a given probability distribution can be reproduced using quantum mechanics. Further if the oracle can learn so, it should return the states and measurements required to reproduce the probability distribution quantum mechanically. This would help solve the problem of detecting which causal scenarios can support “non-classical" quantum correlations.

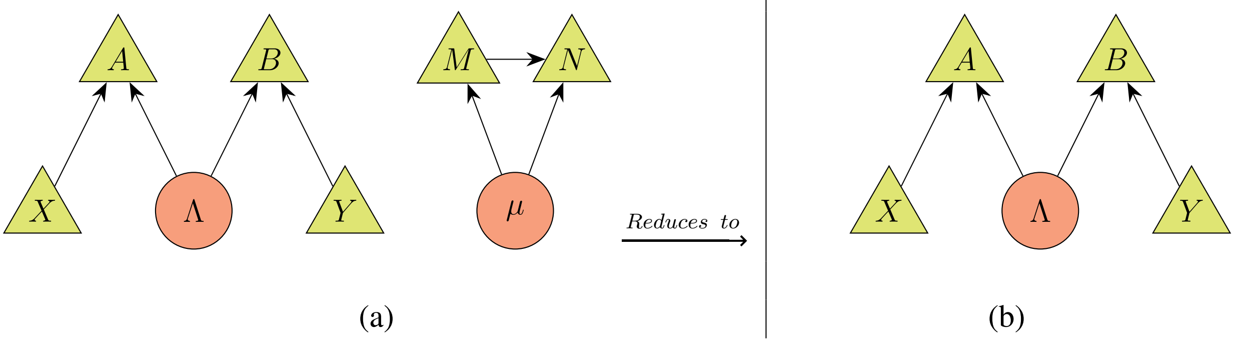
— A —

Further important results in the classification of INTERESTING DAGs

A.1 HLP's reduction techniques

The basic idea of reducibility is to have a notion of when one DAG "reduces" to another DAG such that if the "reduced" smaller DAG is INTERESTING then so is the first one (typically of more number of nodes). We illustrate HLP's reduction techniques using examples. Their proofs can be found in Appendix D of [44]. In all the figures below, if DAG (b), the reduced DAG, is INTERESTING then so is DAG (a), the larger one being reduced.

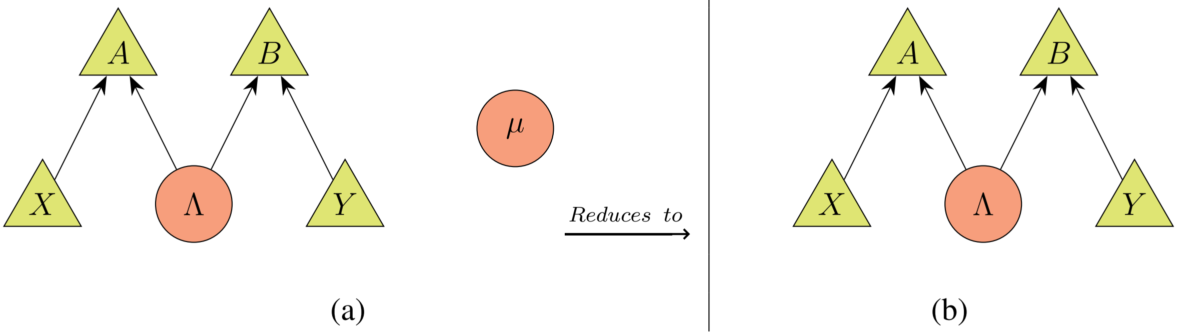
1. Removing a disconnected component:



(a)

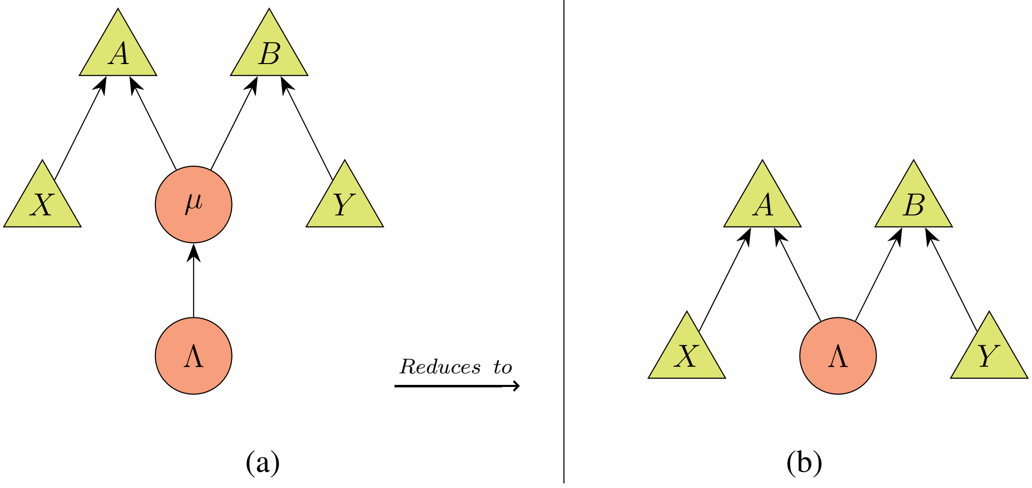
(b)

2. Removing a childless unobserved node:



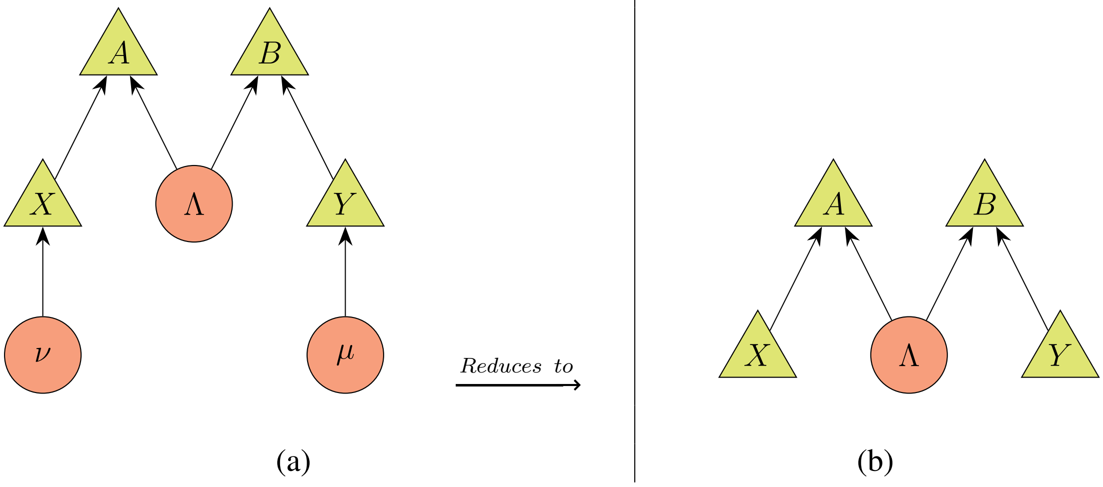
(a) (b)

3. Merging an unobserved node with its sole parent, also unobserved:



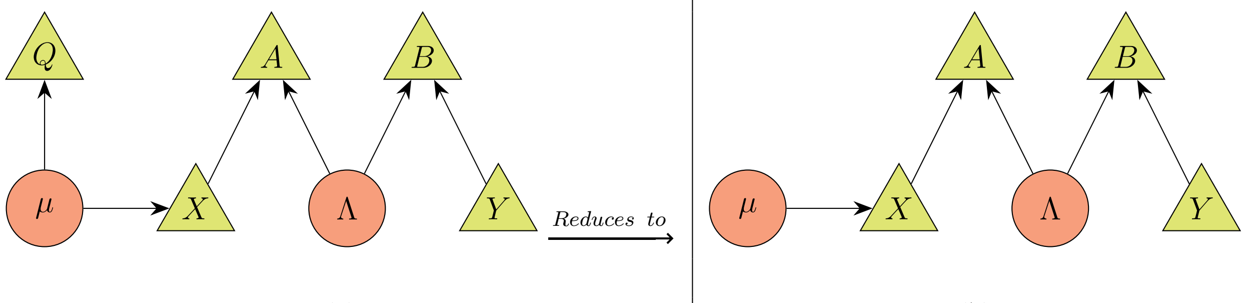
(a) (b)

4. Merging an unobserved node with its sole child:



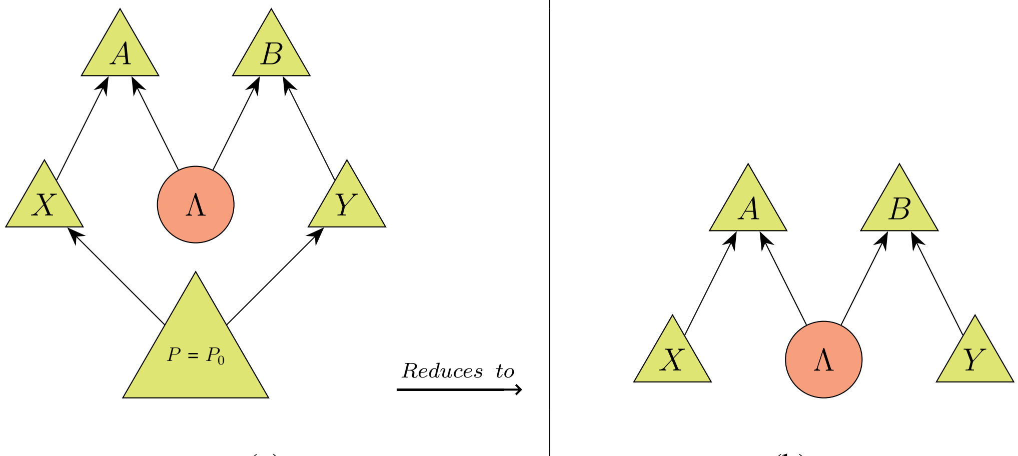
(a) (b)

5. Merging an observed node Q (that has only one sibling, X) with its unobserved parent P (which is itself parentless):



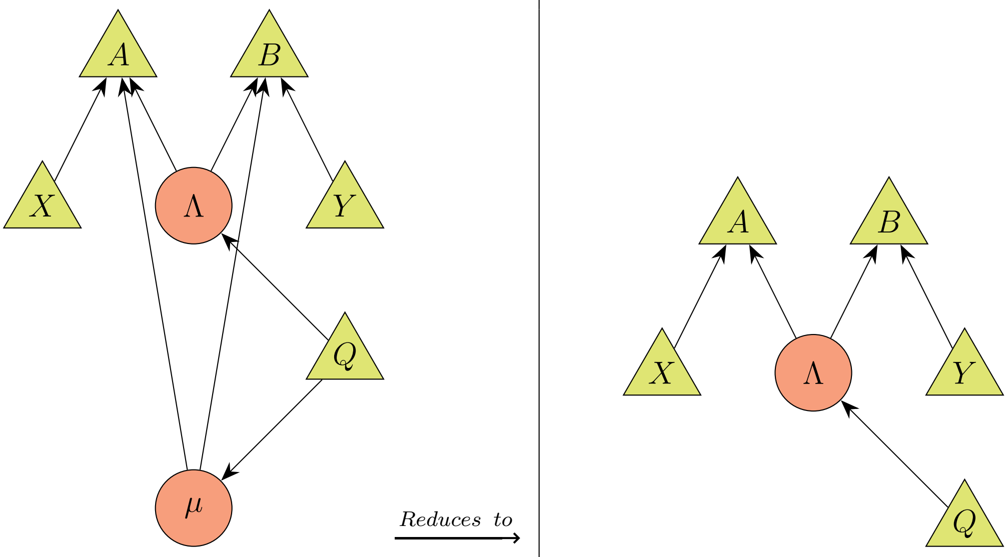
(a) (b)

6. Removing an observed node associated with a 1-outcome variable:



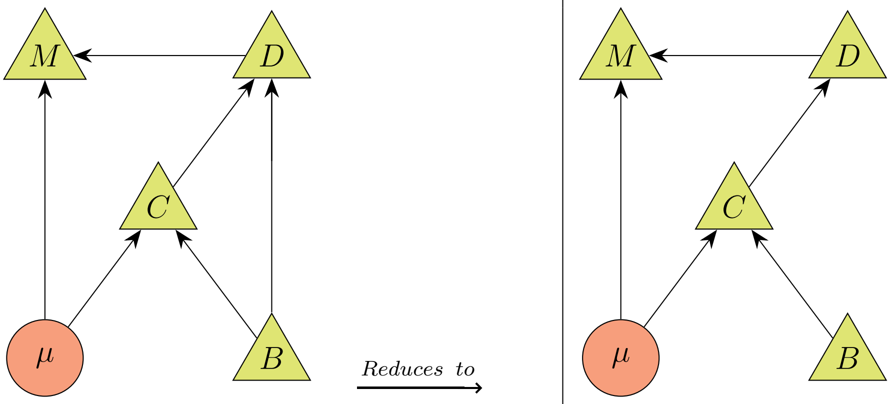
(a) (b)

7. Removing an unobserved node whose parents and children are subsets of the parents and children respectively of another unobserved node:



(a) (b)

8. For an observed node X all of whose parents are observed, removing an edge from a parent Y such that all the observable conditional independences from d-separation after the removal already held beforehand:



(a) (b)

A.2 Pienaar's e-separation theorem is incorrect

In Ref. [81, Lemma 2] Pienaar presented the following claim, which we demonstrate to be incorrect:

Theorem 73 (Pienaar's incorrect e-separation theorem). *Let G be a DAG, and let $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$ and $\boldsymbol{W}$ be disjoint sets of nodes of G such that none of the nodes of $\boldsymbol{Z}$ is a descendant of a node in $\boldsymbol{W}$ and $(\boldsymbol{X} \perp_e \boldsymbol{Y}|\boldsymbol{Z})_{del_{\boldsymbol{W}}}$. Then G is* INTERESTING if and only if *the d-separation relations of G* do not *include any relations of the form $(\boldsymbol{X} \perp_d \boldsymbol{Y}|\boldsymbol{Z}\boldsymbol{S})$, where $\boldsymbol{S}$ is a subset of $\boldsymbol{W}$.*

If this theorem were true, then the problem of classifying INTERESTINGNESS would be completely solved whenever a suitable e-separation is present, as Theorem 73 claims to provides a necessary and sufficient condition for such DAGs that is simple to check. However, the condition turns out to be *neither* necessary *nor* sufficient.

Firstly, note that there are plenty of well-known graphs which are INTERESTING despite not satisfying the conditions of Theorem 73. Examples include the Bell DAG of Figure 3.1, as well as the mDAG in Figure 4.1a, among many others. It is trivial to see that as stated Theorem 73 cannot be an *only if* theorem. That is because there exist DAGs like the Bell DAG which are known to be INTERESTING eventhough they do not have any e-separation condition. Even if take that Theorem 73 could also be an *only if* theorem if there is an e-separation condition true in it, still the theorem cannot be an *only if* one. This can be easily seen by imagining a big DAG which has two smaller disconnected DAGs, one of which is the Bell DAG and the other one is some other DAG for which an e-separation condition holds true but all the conditions of Theoremm 73 do not hold true. Theorem 73 would then conclude the bigger DAG as NON-INTERESTING. But this wrong because the bigger DAG contains the Bell DAG as a disconnected component and hence has to be necessarily INTERESTING. Hence we see that e-separation alone cannot determine INTERESTINGNESS.

Pienaar notably included similar examples in his own work [81], and therefore we believe the inclusion of the "only if" language in Theorem 73 was an oversight, in that Pienaar himself never actually intended to communicate that, but rather only that the e-separation relation $(\boldsymbol{X} \perp_{\mathrm{e}} \boldsymbol{Y}|\boldsymbol{Z})_{del_{\boldsymbol{W}}}$ would automatically follow from the d-separation relation $(\boldsymbol{X} \perp_{\mathrm{d}} \boldsymbol{Y}|\boldsymbol{Z}\boldsymbol{S})$ if that d-separation relation was present in the graph and where $\boldsymbol{S}$ is a subset of $\boldsymbol{W}$.

Regardless, the "if" direction in Theorem 73 *also* turns out to be invalid, though it is a bit more subtle. Consider the DAG in Figure A.1a; Pienaar [81] uses it as an example of a DAG that is deemed (in this case, *correctly* deemed) as INTERESTING pursuant to Theorem 73. In this DAG, $(F \perp_{\mathrm{e}} D|C)_{del_E}$ holds, and node C is not a descendant of node E, and neither $(F \perp_{\mathrm{d}} D|C)$ nor $(F \perp_{\mathrm{d}} D|CE)$ holds true. Therefore, Theorem 73 classifies the DAG of Figure A.1a as INTERESTING. But now consider the DAG of Figure A.1(b) which has the same *structure* as the DAG in Figure A.1(a), with the only difference being that the DAG in Figure A.1(b) has no latent variables. All the conditions of Theorem 73 are again met, however, when applied to Figure A.1. Again, $(F \perp_{\mathrm{e}} D|C)_{del_E}$ holds, and node C is not a descendant of node E, and neither $(F \perp_{\mathrm{d}} D|C)$ nor $(F \perp_{\mathrm{d}} D|CE)$ hold true. So this DAG is again — but this time, wrongly — characterised as INTERESTING by Pienaar's theorem. As previously discussed, all latent-free DAGs are NON-INTERESTING!

In fact, *every* DAG for which the conditions of Theorem 73 hold can be converted into a different latent-free DAG for which the conditions of the theorem would continue to hold by making all the nodes observed. That is, for every DAG which is correctly classified as INTERESTING by the (invalid) condition formulated as Theorem 73 one can construct a latent-free counterexample to Theorem 73.

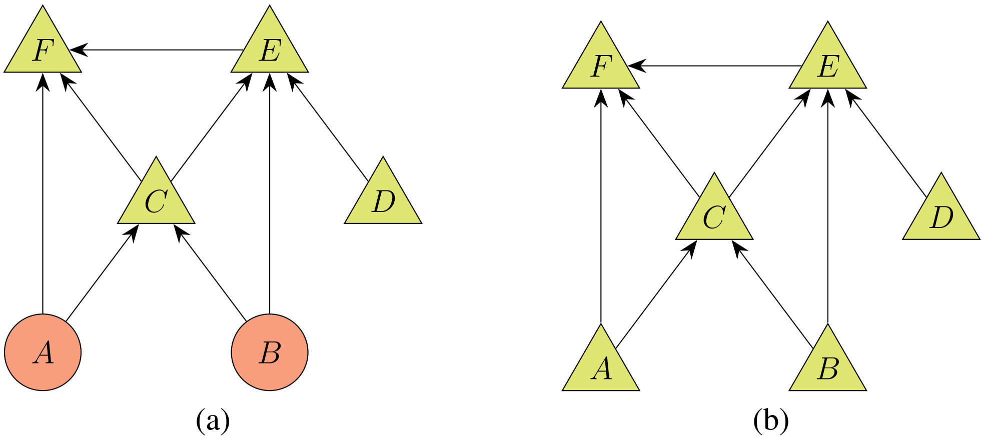


Figure A.1: (a) depicts a scenario characterised correctly by Pienaar's theorem as INTERESTING, whereas (b) depicts an NON-INTERESTING scenario which Pienaar's theorem incorrectly deems INTERESTING.

The flaw with Pienaar's proof of the "if" direction is that it invokes a probability distribution that may not actually be in $\mathcal{I}$, in that the constructed distribution may be inconsistent with some d-separation relation *not mentioned in the statement of Theorem 73*. For example, in Figure A.1, Pienaar's proof would invoke a distribution which posits perfect correlation between D and F while all other variables are point-distributed. While such a distribution is consistent with *all* the observable d-separation relations of Figure A.1a, it is nevertheless *inconsistent* with the d-separation relation $(F \perp_{\mathrm{d}} D|BCE)$ exhibited by Figure A.1b.

This loophole in Pienaar's proof can be closed by *strengthening the conditions of Pienaar's theorem*, namely, to exclude $(F \perp_{\mathrm{d}} D|\boldsymbol{S})$ for *any* subset $\boldsymbol{S}$ of the the remaining observed nodes. In other words,

Theorem 74 (A corrected version of Pienaar's e-separation theorem). *Let G be a DAG, and let $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$ and $\boldsymbol{W}$ be disjoint sets of nodes of G such that none of the nodes of $\boldsymbol{Z}$ is a descendant of a node in $\boldsymbol{W}$ and $(\boldsymbol{X} \perp_e \boldsymbol{Y}|\boldsymbol{Z})_{del_{\boldsymbol{W}}}$. Then G is INTERESTING if the d-separation relations of G* do not *include* *any relations of the form* $(\boldsymbol{X} \perp_d \boldsymbol{Y} | \boldsymbol{S})$*, where* $\boldsymbol{S}$ *may be any subset of the observed nodes of* G *other than* $\boldsymbol{X} \cup \boldsymbol{Y}$*.*

However, a graph can only exclude relations of the form $\boldsymbol{X} \perp_{\mathrm{d}} \boldsymbol{Y} | \boldsymbol{S}$ for any $\boldsymbol{S}$ if for every pair of singleton nodes $\{X, Y\}$ such that $X \in \boldsymbol{X}$ and $Y \in \boldsymbol{Y}$ it holds that $(X \perp_{\mathrm{d}} Y | \boldsymbol{S})$. On the other hand, a pair of nodes $\{X, Y\}$ can only be e-separated by $\boldsymbol{Z}$ upon the removal of $\boldsymbol{W}$ if X and Y are not *adjacent*. Consequently, a DAG can *only* be certified as Interesting by Theorem 74 if there exists a pair of d*-unseparable* but nevertheless non-adjacent nodes. Which is to say, Theorem 74 is ultimately *equivalent*[1] to Theorem 33 of the main text, albeit obfuscated.

A.3 Revisiting the Skeleton Condition

In the main test we noted (Theorem 33) that every nonmaximal DAG is Interesting. Here we revisit that finding, and contrast is with a prior result due to Pienaar [81].

Firstly, we recall that the set of adjacent node pairs in a DAG is conventionally referred to as the DAG's *skeleton*.

Definition 75 (**Skeleton of a DAG.**)**.** *The skeleton of a DAG* G *is the undirected graph wherein* A *and* B *are adjacent in* G*'s skeleton — denoted* A •–• B *— if and only if* A *and* B *are adjacent in* G*.*

Notably, the skeleton of a DAG can be used to witness observational inequivalence. In particular, Evans has shown that if a pair of nodes is adjacent in some mDAG G but not in some other mDAG H, then $\mathcal{C}_G \neq \mathcal{C}_H$ [27, 26]. We formally express this idea here as:

Lemma 76 (Skeleton condition for observational equivalence and dominance)**.** *For any pair of DAGs* G *and* H*,* $\mathcal{C}_G = \mathcal{C}_H$ *only if the skeletons of* G *and* H *are the same, i.e., they agree on all observable node (non)adjacencies. Additionally,* $C_G \subset C_H$ *only if every pair of adjacent nodes in* G *is also adjacent in* H*.*

As with any necessary conditional for observational equivalence, Lemma 76 can then be utilized to formulate a sufficient condition for Interestingness. Pienaar did exactly that when he formulated the following Theorem in Ref. [81, Theorem 1]:

Theorem 77 (Pienaar's skeleton condition for Interestingness)**.** *Let* G *be the DAG we need to check for Interestingness. Suppose that one can find some other DAG* H *with the same observed conditional*

[1] Plainly Theorem 33 can be seen as a special case of Theorem 74, by taking $\boldsymbol{Z}$ to be empty, choosing $\boldsymbol{X}$ and $\boldsymbol{Y}$ to be singleton-sized sets, and letting $\boldsymbol{W}$ be the set of all observed nodes other than $\boldsymbol{X}$ and $\boldsymbol{Y}$.

independences as G and which is certainly known to be NON-INTERESTING, i.e. $\mathcal{I}_G = \mathcal{I}_H$ and $\mathcal{I}_H = \mathcal{C}_H$. Then, if the skeletons of G and H are different, evidently $\mathcal{C}_G \neq \mathcal{C}_H$ per Lemma 76, and hence $\mathcal{C}_G \neq \mathcal{I}_G$, i.e. DAG G is INTERESTING.

Theorem 77 is both correct and useful. Note, however, that it is *superseded* by the Theorem 33 presented in the main text. Suppose that there *does* exists some latent-free DAG H that agrees with the observed d-separation relations of G. Then, among other things, G and H evidently agree on d-unseparable observed node pairs. Since the skeleton of H is defined by H's d-unseparable node pairs [91, Prop. 3.19], the only way for the skeletons of G and H to differ is if G contains some pair of *nonadjacent* d-unseparable observed nodes. Consequently, Theorem 77 can only certify a DAG as INTERESTING when the DAG's INTERESTINGNESS is also certifiable by Theorem 33.

Thus, *both* of Pienaar's conditions for INTERESTINGNESS are subsumed by Theorem 33 here, at least after adjusting Pienaar's condition based on e-separation as per Appendix A.2.

We know that there are DAGs for which one cannot find any DAG H with the required properties to make use of Theorem 77; see Appendix A.4 for examples. In light of Theorem 26 we would want to reformulate Theorem 77 in a manner that is clearly (strictly) more powerful, which removes the caveat about finding such an H.

Theorem 78 (Improved skeleton condition for INTERESTINGNESS)**.** *Let G be the DAG we need to check for INTERESTINGNESS. Consider all latent-free graphs which share the same skeleton as that of G. If no latent-free graph within that set furthermore matches the observed d-separations of G, then G is INTERESTING.*

Proof. Theorem 78 follows directly from combining Lemma 76 and Eq. (3.5) with Theorem 26. □

As it turns out, however, Theorem 78 is *equivalent* to the conjunction of Theorems 33 and 44. First, we can see that Theorem 33 is a special case of Theorem 78. If G is nonmaximal, then it has at least one non-adjacent pair of observed nodes which is d-unseparable. This implies that all the latent-free DAGs which have the same skeleton (same adjacency structure) as G have a different set of d-separable pairs of nodes than that of G, because all latent-free DAGs are maximal [91, Prop. 3.19]. This then implies that they have a different set of d-separation relations than those of G, so Theorem 78 witnesses all nonmaximal DAGs as INTERESTING. Secondly, it is easy to see that Theorem 44 is also a special case of Theorem 78: if there are no latent-free DAGs that match the set of d-separation relations of G, then in particular there are no latent-free DAGs that match the the set of d-separation relations *and* the skeleton of G.

Next, we prove the inverse, that Theorems 33 and 44 *together* subsume Theorem 78. Let G be a DAG which is shown INTERESTING by Theorem 78. If G is nonmaximal, then it is also shown

INTERESTING by Theorem 33. What remains, then, is to show that there are no *maximal* DAGs which *cannot* be proven INTERESTING via Theorem 44 but *can* be seen as INTERESTING via Theorem 78. But if the DAG G is both maximal *and* shares the same d-separation relations as some latent-free DAG H, it *automatically follows* that G and H agree on their skeletons as well. After all, G and G are *both* maximal, which means that their skeletons are dictated by their respective d-unseparable node pairs, which are identical.

Note that Theorem 78 is also subsumed by Theorem 47, since agreeing on e-separation relations implies agreeing on adjacencies (i.e., skeleton) as well as on d-separation relations.

A.4 Rapid d-separation test

In Section 4.3, we described a method to show the INTERESTINGNESS of a DAG when its set of observed d-separation relations is NALF, i.e. it does not match those of any latent-free DAG; this is encoded in Theorem 44. To apply this method in practice, we indicated that one can construct all of the latent-free DAGs that have the same number of observed nodes as the DAG in question, and then compare their sets of d-separation relations. However, when the DAG is *maximal* (as per Definition 32) there is a faster way to apply this method, which will be described now. When the DAG is *not* maximal, its INTERESTINGNESS is automatically attested by Theorem 33.

When the DAG is maximal, to see whether its set of d-separation relations is NALF one just needs to check whether it has one of the DAGs of Figure A.2 as a subgraph. This was recognized in Evans' own proof of Theorem 26 [28], as we argue in the proof of the following theorem:

Theorem 79 (rapid d-separation condition for INTERESTINGNESS)**.** *Let G be an mDAG. If G is not maximal, then it is INTERESTING by Theorem 33. If G is maximal, then it has a set of observed d-separation relations unmatched by any latent-free DAG (NALF) — and is therefore INTERESTING pursuant to Theorem 44 — if and only if it contains one of the eighteen graph patterns (up to relabelling the nodes) presented in Figure A.2 as a subgraph.*

Proof. Evans' [28] has shown that if a latent-permitting DAG G has a NALF set of observed d-separation relations, then either its associated PAG (partial ancestral graph) contains a so-called "locally unshielded collider path" of length 3, and/or it contains a so-called "discriminating path" of length 3.[2] This implies that the PAG associated with G will contain a sub-PAG matching one of the

[2] The phrasing presented in the proof of Proposition 4.2 of Ref. [28] says that either the associated PAG has a locally unshielded collider path of length exactly 3, all options of which are presented in Figure 4 of Ref. [28], or a discriminating path of length *at least* 3. However, Proposition B.1 of [28] further shows that all discriminating paths of length at least

4-node PAGs depicted within Figures 4 and 5(i) of [28] whenever the observed d-separation relations of G are NALF.

A PAG is an abstract graphical representation of a set of d-separation relations. In particular, two nodes are *adjacent* in a PAG if and only the two nodes are d-unseparable in the original DAG. A "path of length 3" in a PAG refers to a set of four nodes $\{X, A, B, Y\}$ such that $\{X, A\}$, $\{A, B\}$, $\{B, Y\}$ all constitute d-unseparable pairs.

The adjacencies of a DAG G and the adjacencies of the PAG associated with G can, in general, differ. For *maximal* DAGs, however, they must coincide: for those, adjacency is equivalent to d-unseparability. The PAG associated with a *maximal* DAG G will exhibit an unshielded collider path or a discriminating path if and only if in the original G we can find four nodes $\{X, A, B, Y\}$ such that $\{X, A\}$, $\{A, B\}$, $\{B, Y\}$ represent adjacent pairs and such that the d-separation relations pertaining exclusively to $\{X, A, B, Y\}$ are of one of "unshielded collider path" type or "discriminating path" type. We do not define these two types for brevity, but we exhibit all such 4-node mDAGs in Figure A.2. □

The set of d-separation relations of each one of the DAGs in Figure A.2 is:

1. $(A \perp_{\mathrm{d}} Y|\varnothing)$ and $(A \perp_{\mathrm{d}} Y|X)$ and $(B \perp_{\mathrm{d}} X|\varnothing)$ and
   $(B \perp_{\mathrm{d}} X|Y)$ and $(X \perp_{\mathrm{d}} Y)$ and $(X \perp_{\mathrm{d}} Y|A)$ and $(X \perp_{\mathrm{d}} Y|B)$ [Figure A.2 (a)-(f)]
2. $(A \perp_{\mathrm{d}} Y|X)$ and $(B \perp_{\mathrm{d}} X|Y)$ [Figure A.2 (g)-(k)]
3. $(A \perp_{\mathrm{d}} Y|\varnothing)$ and $(B \perp_{\mathrm{d}} X|\varnothing)$ [Figure A.2 (l)]
4. $(B \perp_{\mathrm{d}} X|\varnothing)$ and $(A \perp_{\mathrm{d}} Y|X)$ [Figure A.2 (m)-(o)]
5. $(B \perp_{\mathrm{d}} X|\varnothing)$ and $(X \perp_{\mathrm{d}} Y|A)$ [Figure A.2 (p)-(r)]

By checking the d-separation of all the 4-observed-node mDAGs, we know that these are the only mDAGs that present these sets of d-separation relations. While searching for subgraphs is computationally easy, an alternative is to consider all 4-node subsets of a given large mDAG and ask if the d-separation relations *pertaining exclusively to some four nodes* contains all and only one of the patterns listed above, up to relabelling. If yes, and if the large mDAG G is maximal, then G certainly contains one of the patterns of Figure A.2 as a subgraph.

Figure A.3 shows an example of a DAG that does not have any of the DAGs of Figure A.2 as a subgraph, but nevertheless has a NALF set of d-separation relations. However, this is not a problem: this DAG is *not* maximal (A and B are not d-separable but are e-separable by C), so its INTERESTINGNESS follows from Theorem 33.

3 either contain a locally unshielded collider path of length 3 and/or a discriminating path of length exactly 3, which is presented in Figure 5(i) of [28].

Finally, we will show an explicit construction of a distribution which is in $\mathcal{I}_G$ but not in $\mathcal{C}_G$ for the maximal mDAGs which are shown INTERESTING by Theorem 33. First, we note that the Popescu-Rohrlich box support shown in Eq. (4.6) is not compatible with any of the mDAGs of Figure A.2, as can be shown using Fraser's algorithm. This implies that the *uniform* distribution over the events of the Popescu-Rohrlich box support is not classically compatible with any of the mDAGs of Figure A.2, and is thus not an element of $\mathcal{C}_G$ for these mDAGs. On the other hand, it is well-known that said distribution obeys all of the conditional independence relations that come from the d-separation relations of the Bell DAG. As seen above, all the d-separation relations of the mDAGs of Figure A.2 are included in the d-separation relations of the Bell DAG, which implies that such distribution is an element of $\mathcal{I}_G$ for all the mDAGs of Figure A.2.

A.5 Supports subsumes e-separation as a test of observational inequivalence

Here we provide a full proof of Theorem 59, that says that if DAGs G and H can be shown inequivalent by virtue of having different sets of e-separation relations, then they can be shown inequivalent by virtue of having different sets of compatible supports at *binary* variables. Before proving this more general result, we will prove the case for d-separation relations as an auxiliary lemma:

Lemma 80 (d-separation and compatible supports)**.** *Let G be an DAG with observed nodes $\boldsymbol{V}$, with $\boldsymbol{Z} \subseteq V$ being a subset of $\boldsymbol{V}$ and $X \in \boldsymbol{V}$ and $Y \in \boldsymbol{V}$ being two observed nodes. If the DAG G* does not *exhibit the d-separation relation $(X \perp_d Y|\boldsymbol{Z})$, then there is at least one support over binary variables which is compatible with G but at the same time is in conflict with $(X \perp\!\!\!\perp_{CI} Y|\boldsymbol{Z})$, i.e. is trivially incompatible with every other DAG for which $(X \perp_d Y|\boldsymbol{Z})$.*

Proof. We explicitly construct a support such that the marginal events $\{X{=}0, \boldsymbol{Z} = 1\}$ and $\{Y{=}1, \boldsymbol{Z} = 1\}$ both occur, but such that the event $\{X{=}0, Y{=}1, \boldsymbol{Z} = 1\}$ does not occur. Plainly such a support is *trivially incompatible* with any DAG wherein $(X \perp_{\mathrm{d}} Y|\boldsymbol{Z})$ per Lemma 54. Here we show that a distribution with this support *can* arise in every DAG wherein $(X \not\perp_{\mathrm{d}} Y|\boldsymbol{Z})$. We note that this construction is identical to that which appears in Appendix E of [29].

If $(X \not\perp_{\mathrm{d}} Y|\boldsymbol{Z})$ in G, then there exists a path from X to Y in G which is unblocked by $\boldsymbol{Z}$. Every node of the path has either no parents in the path (in which case it is the base of a fork), one parent in the path (if it is the middle node of a chain or an end node of the path) or two parents in the path (if it is a collider). The support is constructed by assigning the following functional dependencies to the nodes of the path:

- A node $F \in \boldsymbol{F}$ (“F" for “Fork") has 0 parents in the path:

$$F = \begin{cases} 0 \text{ with probability } 1/2 \\ 1 \text{ with probability } 1/2 \end{cases}$$

- A node $M \in \boldsymbol{M}$ (“M" for “Mediary") has 1 parent P_M in the path:

$$M = P_M \text{ with unity probability}$$

- Node $C \in \boldsymbol{C}$ (“C" for “Collider") has 2 parents $P_{C,1}$ and $P_{C,2}$ in the path:

$$C = \begin{cases} 0 \text{ if } P_{C,1} \neq P_{C,2} \\ 1 \text{ if } P_{C,1} = P_{C,2} \end{cases}$$

Since the path is unblocked $\boldsymbol{C} \in \boldsymbol{Z}$ while $\boldsymbol{MF} \notin \boldsymbol{Z}$.

This construction leads to a probability distribution compatible with G wherein $P(\boldsymbol{M} = \boldsymbol{F} = 0|\boldsymbol{Z}{=}1) = 1/2$ and $P(\boldsymbol{M} = \boldsymbol{F} = 1|\boldsymbol{Z}{=}1) = 1/2$. Since $\{X, Y\} \subset \boldsymbol{MF}$, we confirm that the marginal event $\{X{=}0, \boldsymbol{Z} = 1\}$ occurs, and such that the marginal event $\{Y{=}2, \boldsymbol{Z} = 1\}$ occurs, yet that the event $\{X{=}0, Y{=}1, \boldsymbol{Z} = 1\}$ does not occur. □

Using Lemma 80, we can now proceed to proof Theorem 59:

Proof. Suppose that G has an e-separation relation $(A \perp_{\mathrm{e}} B|\boldsymbol{C})_{del_{\boldsymbol{D}}}$ that H does not have, where $A, B \in V$ are observed nodes and $\boldsymbol{C}, \boldsymbol{D} \subseteq V$ are sets of observed nodes.

By the definition of e-separation, we know that $G_{\boldsymbol{V}\setminus\boldsymbol{D}}$, the DAG obtained by deleting the nodes $\boldsymbol{D}$ from G, exhibits the d-separation relation $A \perp_{\mathrm{d}} B|\boldsymbol{C}$ while $H_{\boldsymbol{V}\setminus\boldsymbol{D}}$ does not. This implies that *none* of the supports compatible with $G_{\boldsymbol{V}\setminus\boldsymbol{D}}$ are in conflict with $A \perp\!\!\!\perp_{\mathrm{CI}} B|\boldsymbol{C}$, while from Lemma 80 we know that there is at least one support over binary variables which is compatible with $H_{\boldsymbol{V}\setminus\boldsymbol{D}}$ that is in conflict with $A \perp\!\!\!\perp_{\mathrm{CI}} B|\boldsymbol{C}$. This means that it is possible to find a support $\mathcal{S}_{\boldsymbol{V}\setminus\boldsymbol{D}}$ over binary variables which is compatible with $H_{\boldsymbol{V}\setminus\boldsymbol{D}}$ but not compatible with $G_{\boldsymbol{V}\setminus\boldsymbol{D}}$.

Let $\mathcal{S}_{\boldsymbol{V}}$ be the support over $\boldsymbol{V}$ which coincides with $\mathcal{S}_{\boldsymbol{V}\setminus\boldsymbol{D}}$ for the variables in $\boldsymbol{V} \setminus \boldsymbol{D}$ and where the variables in $\boldsymbol{D}$ are set to a point value. This support is incompatible with G, because setting the variables in $\boldsymbol{D}$ to a point value has the same effect on its children as deleting the respective nodes (and thus reaching $G_{\boldsymbol{V}\setminus\boldsymbol{D}}$). Furthermore, $\mathcal{S}_{\boldsymbol{V}}$ is *compatible* with H: since $\mathcal{S}_{\boldsymbol{V}\setminus\boldsymbol{D}}$ is compatible with $H_{\boldsymbol{V}\setminus\boldsymbol{D}}$, we can obtain $\mathcal{S}_{\boldsymbol{V}}$ by functional models where the children of $\boldsymbol{D}$ in G ignore the parents in $\boldsymbol{D}$, as indicated in Lemma 13 of [29].

Therefore, $\mathcal{S}_{\boldsymbol{V}}$ is a support over binary variables that is compatible with H but incompatible with G. □

A.6 Results for mDAGs of 5 observed nodes

Category	mDAGs with **5** observed nodes
Total Count	1,718,596
remaining # for which the HLP criterion does not apply	1,009,961
remaining # for which our nonmaximality condition does not apply	278,964
remaining # for which our setwise nonmaximality condition does not apply	118,278
remaining # which do not contain an Interesting 4-observed-nodes subgraph	12,834
remaining # for which our d-separation condition does not apply	12,834
remaining # for which our e-separation condition does not apply	12,834

Table A.1: Results for mDAGs of 5 observed nodes.

When trying to certify the Interestingness of mDAGs with 5 or more observed nodes, a simple but important consideration should be taken into account. Namely,

Proposition 81. *Consider a DAG G, as well as the subgraph $G_{del\,\boldsymbol{S}}$ of G, where $G_{del\,\boldsymbol{S}}$ is formed by deleting a strict subset of $G^{'}$s observed nodes from G. If $G_{del\,\boldsymbol{S}}$ is Interesting, then G is Interesting as well.*

Proof. Proposition 81 is an immediate consequence of Lemma 38 from the main text. That is, Lemma 38 ensure that $P(\boldsymbol{V} \setminus \boldsymbol{S}) \notin \mathcal{C}_{G_{\text{del}\,\boldsymbol{S}}}$ then $P(\boldsymbol{V}) \notin \mathcal{C}_G$ where $P(\boldsymbol{V}) \coloneqq P(\boldsymbol{V} \setminus \boldsymbol{S})\delta_{\boldsymbol{S},0^{|\boldsymbol{S}|}}$. At the same time, if $P(\boldsymbol{V} \setminus \boldsymbol{S}) \in \mathcal{I}_{G_{\text{del}\,\boldsymbol{S}}}$ then $P(\boldsymbol{V}) \in \mathcal{I}_G$ where again $P(\boldsymbol{V}) \coloneqq P(\boldsymbol{V} \setminus \boldsymbol{S})\delta_{\boldsymbol{S},0^{|\boldsymbol{S}|}}$. That $P(\boldsymbol{V}) \in \mathcal{I}_G$ follows from the fact that no *new* d-separation relations are induced on $\boldsymbol{V} \setminus \boldsymbol{S}$ by embedding the DAG $G_{\text{del}\,\boldsymbol{S}}$ as a subgraph of a larger DAG, namely G. □

The reader might ask why we did not employ Proposition 81 in assessing the Interestingness of mDAGs with *4* observed nodes. After all, there are five mDAGs with *3* observed nodes. However, any 4-observed-nodes mDAGs which would be certifiable as Interesting via Proposition 81 would already be picked up by Theorem 42, since that theorem detects 100% of the Interesting 3-observed-nodes mDAGs by itself.

The application of the conditions for INTERESTINGNESS presented here to mDAGs of 5 observed nodes gives the results shown in Table A.1. Although the application of Fraser's algorithm on the remaining 12,834 mDAGs was computationally infeasible, nevertheless, the application of all the other techniques provides a similar success percentage as compared to mDAGs of 4 and 3 observed nodes. Precisely, for mDAGs of 5 observed nodes all the other techniques (apart from supports) reduce the number of unresolved mDAGs of 5 observed nodes by 99.15%, while this percentage is 99.89% and 97.8% for mDAGs of 4 and 3 observed nodes respectively. This result is consistent with the HLP conjecture, though whether or not it can be considered *evidence* in favor of the conjecture is debatable.

Unsurprisingly, our d-separation condition as articulated in Theorem 44 is now effectively redundant to the conjunction of Proposition 81 and Theorem 33, since a *maximal* mDAG with 5+ observed nodes will have a set of observed d-separation relations inequivalent to any latent-free DAG only if the large mDAG contains one of 18 particular 4-observed-nodes mDAGs, as discussed extensively in Appendix A.4.

The fact that the Theorem 47 does not resolve any further mDAGs as INTERESTING is evidence in favor of a conjecture that Theorem 47 is perhaps subsumed by the conjunction of Theorems 33 and 44.

A.7 Comparison of classification of INTERESTING causal scenarios based on the total number of all the nodes vs total number of just observed nodes

In this chapter, instead of just classifying the causal scenarios by the total observed nodes, we also classify them based on their total number of all the nodes and compare the two procedures. That is, in particular along with using the concept of mDAGs in the classification of INTERESTING causal scenarios, we also use the framework GDAGs (as introduced in Definition 13 in Section 2.9) in the classification of INTERESTING causal scenarios and compare the two approaches. In doing so we again use the concept of e-separation and present a detailed proof of Theorem 74 introduced in Appendix A.2.

To prove Theorem 74 we first present the following Lemma, first proved as Theorem 4.2 of [26]:

Lemma 82. *Let G be a DAG, and let $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$ and $\boldsymbol{W}$ be disjoint sets of nodes of G such that none of the nodes of $\boldsymbol{Z}$ is a descendant of a node of $\boldsymbol{W}$ and $(\boldsymbol{X} \perp_e \boldsymbol{Y}|\boldsymbol{Z})_{del_{\boldsymbol{W}}}$. Let P be a probability distribution compatible with G. Then for any fixed value of $\boldsymbol{W} = w$, there exists another probability distribution P^* for which the conditional independence $\boldsymbol{X} \perp\!\!\!\perp_{CI} \boldsymbol{Y}|\boldsymbol{Z}$ holds and $P(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{W} = w|\boldsymbol{Z}) = P^*(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{W} = w|\boldsymbol{Z})$.*

In fact we will only need a simple case, where the distribution P only supports a single value of $\boldsymbol{W}$ to begin with, in which case we simply have $P^* = P$ and thus:

Corollary 83. *Let G be a DAG, and let $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$ and $\boldsymbol{W}$ be disjoint sets of nodes of G such that none of the nodes of $\boldsymbol{Z}$ is a descendant of a node of $\boldsymbol{W}$ and $(\boldsymbol{X} \perp_d \boldsymbol{Y}|\boldsymbol{Z})_{del_{\boldsymbol{W}}}$. Let P be a probability distribution compatible with G where only one value of $\boldsymbol{W}$ has non-zero probability. Then P satisfies $\boldsymbol{X} \perp\!\!\!\perp_{CI} \boldsymbol{Y}|\boldsymbol{Z}$.*

We can now prove the following equivalent version of our Theorem 74.

Theorem 84. *Let $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$, $\boldsymbol{W}$ be disjoint sets of observable nodes in a DAG G and let $\boldsymbol{V}$ be the set of all observable nodes in the DAG. If $(\boldsymbol{X} \perp_e \boldsymbol{Y}|\boldsymbol{Z})_{del_{\boldsymbol{W}}}$ holds true for the DAG and no member of $\boldsymbol{Z}$ is a descendant from any nodes in $\boldsymbol{W}$ then DAG $\mathcal{G}$ is* Interesting *(i.e has $\mathcal{C} \neq \mathcal{I}$)* if *the observable conditional independence relations in $\mathcal{I}$ exclude all relations of the form $x_i \perp\!\!\!\perp_{CI} y_j|\boldsymbol{S}\ \forall\ x_i \subseteq \boldsymbol{X}, y_j \subseteq \boldsymbol{Y}, \boldsymbol{S} \subseteq \boldsymbol{V} \setminus \{\boldsymbol{X} \cup \boldsymbol{Y}\}$.*

Proof. Note that if $\mathcal{I}$ excludes all relations of the form $x_i \perp_{\rm d} y_j|\boldsymbol{S}\ \forall\ x_i \subseteq \boldsymbol{X}, y_j \subseteq \boldsymbol{Y}, \boldsymbol{S} \subseteq \boldsymbol{V}\setminus\{\boldsymbol{X}\cup\boldsymbol{Y}\}$ then it also excludes all relations of the form $\boldsymbol{A}x_i \perp_{\rm d} \boldsymbol{B}y_j|S\ \forall\ \ x_i \subseteq \boldsymbol{X}, y_j \subseteq \boldsymbol{Y}, S \subseteq \boldsymbol{V} \setminus \{\boldsymbol{X} \cup \boldsymbol{Y}\}$, for arbitrary nodes $\boldsymbol{A} \in \boldsymbol{V} \setminus \{\boldsymbol{X} \cup \boldsymbol{Y}\},\ \ \boldsymbol{B} \in \boldsymbol{V} \setminus \{\boldsymbol{X} \cup \boldsymbol{Y}\}$. Hence, we can construct a probability distribution in $\mathcal{I}$ where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are perfectly correlated and all the other nodes take fixed values, i.e.:

$$P(X,Y,W,Z,A,B.....) = \frac{1}{2}(\delta_{\boldsymbol{X},0}\delta_{\boldsymbol{Y},0} + \delta_{\boldsymbol{X},1}\delta_{\boldsymbol{Y},1})\delta_{\boldsymbol{W},w}\delta_{\boldsymbol{Z},z}\delta_{\boldsymbol{A},a}\delta_{\boldsymbol{B},b}\dots. \tag{A.1}$$

By Corollary 83, this distribution is not in $\mathcal{C}$. □

Consider Figure A.4, that presents a DAG with 4 visible nodes and 7 total nodes that does not satisfy the HLP criterion. This DAG is such that $(G \perp_{\rm e} F|E)_{del_D}$ and E is not a descendant of D. Furthermore, none of the d-separation relations $(G \perp_{\rm d} F|E)$, $(G \perp_{\rm d} F|D)$ or $(G \perp_{\rm d} F|ED)$ holds for this DAG. Therefore, by our corrected e-separation theorem, the DAG of Figure A.4 is Interesting.

Another example is given by Figure A.5. There, the e-separation relation $(G \perp_{\rm e} F)_{del_{DE}}$ holds while none of the d-separation relations $(G \perp_{\rm d} F|E)$, $(G \perp_{\rm d} F|D)$ or $(G \perp_{\rm d} F|ED)$ hold. Therefore, this DAG is also Interesting.

Now we re-discuss a bit about the skeleton of a DAG and show how it can be used to certify when a DAG is Interesting. Skeleton of a DAG can be used to extract information about the classical probability distributions that are Markov compatible with the DAG. Evans has shown that two DAGs that have different skeletons necessarily have different sets $\mathcal{C}$'s [27, 26] . This fact along with Theorem

26 can be used to answer if a DAG is INTERESTING. To understand this first we need to understand what is a skeleton of a DAG.

Definition 85 (Skeleton of a DAG.). *Let G be a mDAG. Its skeleton is defined as the undirected graph in which all the observed nodes which had a directed edge between them or share a latent common cause are connected by the undirected edges.*

Evans [27] and Pienaar [81] have proven the following results:

Lemma 86 (Skeleton condition for observational equivalence). *For any pair of DAGs G and G, $\mathcal{C}_G = \mathcal{C}_H$ only if the skeletons of G and H are the same.*

Corollary 87 (Skeleton method for INTERESTINGNESS.). *Let G be the DAG we need to check for INTERESTINGNESS. Let H be a DAG with the same observed conditional independences as G and which is certainly known to be NON-INTERESTING, i.e. $\mathcal{I}_G = \mathcal{I}_H$ and $\mathcal{I}_H = \mathcal{C}_H$. Then, if the skeletons of G and H are different, evidently $\mathcal{C}_G \neq \mathcal{C}_H$ per Lemma 86, and hence $\mathcal{C}_G \neq \mathcal{I}_G$, i.e. DAG G is INTERESTING.*

The above can indeed be used to check if a DAG is INTERESTING. But it can be better appreciated in light of [28] by unifying Theoroms 44 and 87:

Theorem 88 (Skeleton method for classical observational equivalence.). *Let G be a DAG whose INTERESTINGNESS we need to check. If there is* no *latent free DAG which can reproduce all the observed d-separation relations that DAG G has, then G is is interesting per Theorem 44. If, however, there* is *a latent free DAG H which can reproduce all the observed d-separation relations that DAG G has such that $\mathcal{I}_G = \mathcal{I}_H$ and $\mathcal{I}_H = \mathcal{C}_H$, and if the skeletons of DAGs G and H are different, then $\mathcal{C}_G \neq \mathcal{C}_H$ per Lemma 86, and hence $\mathcal{C}_G \neq \mathcal{I}_G$, i.e. DAG G is INTERESTING.*

Evans scenario of Figure 4.1(b) is an example of a scenario that can be classified as INTERESTING using the above method. Since Evans scenario has no observed conditional independences a latent free graph that can reproduce all the observed conditional independences of Evans scenario is a scenario of 3 visible nodes where each node is connected to every other node. The skeleton of such a scenario is clearly different from the skeleton of Evans scenario which is just $D \leftarrow C \rightarrow E$, thus showing that Evans scenario is INTERESTING.

Having proven Theorem 84, we can begin classifying Interesting causal scenarios. For enumerating the causal scenarios we count the DAGs by both their total number of observed nodes as well as the total number of all the nodes and to do so we depend on the frameworks of mDAGs and GDAGs (as introduced in Definition 13 in Section 2.9) respectively.

A.7.1 Computational Results for mDAGs without using HLP's reduction techniques

First we consider the problem of finding Interesting mDAGs of total visible nodes four without using any reduction techniques of HLP but simply using HLP's sufficient criterion for Non-interestingness and using d-separation and e-separation for finding definitely Interesting scenarios as explained in subsections 4.3 and 4.4. We first enumerate all the mDAGs of total four visible nodes and then filter out the definitely Non-Interesting ones using HLP's criterion for Non-interestingness. On the remaining mDAGs we test how many of those possess observable conditional independences which cannot be reproduced by any latent free mDAG. These scenarios are definitely Interesting. After doing this we have a list of mDAGs which possess observable d-separation relations which can be reproduced by a latent free mDAG. In this list we check which mDAGs have an e-separation condition that their corresponding latent free mDAG having the all same d-separation relations as them doesn't have. Such mDAGs are again definitely Interesting. Now we have a list of mDAGs whose all d-separation and e-separation relations together can be modeled by a latent free mDAG. These are the mDAGs which could be potentially Interesting but we don't know at this step. Using Fraser's support algorithm we compare feasible supports of each mDAG and its corresponding latent free mDAG which has all the same d-separation and e-separation relations as the mDAG itself. At different cardinalities of the nodes we find supports that are infeasible in the mDAG but are feasible in its corresponding latent free mDAG for all the remaining mDAGs but just three. By looking at compatible supports with binary cardinalities of the visible variables, it was possible to show Interestingness for 161 out of the 168 remaining mDAGs with four visible nodes that were unresolved so far. For the rest seven, we couldn't find any classically infeasible supports at binary cardinalities of the variables. However, as mentioned previously, Fraser's algorithm for binary supports can also be applied for higher cardinalities of the visible variables. If we allow the cardinality of the variable A to be three, we can find a support that is incompatible with the mDAGs in Table 4.1(a), 4.1(b), 4.1(c) and 4.1(d) (it does not come from d-separation, since none of the mDAGs in Table 4.1 has any d-separation relation). Therefore, the mDAGs 4.1(a)-(d) are Interesting. Out of the mDAGs with four visible nodes, only three remain undecided. These are shown in Table A.2. These could be potentially INTERESTING.

$$\mathcal{S}_{\text{for Table 4.1}} = \left\{ \begin{matrix} \{a=0,b=0,c=0,d=0\} \\ \{a=0,b=0,c=1,d=0\} \\ \{a=0,b=1,c=0,d=0\} \\ \{a=1,b=0,c=0,d=0\} \\ \{a=1,b=1,c=0,d=0\} \\ \{a=2,b=0,c=0,d=1\} \\ \{a=2,b=1,c=1,d=0\} \end{matrix} \right\} \tag{A.2}$$

In Table A.3 we present the exact number of mDAGs that we are able to characterize as INTERESTING using the techniques introduced in the Chapter 4.

A.7.2 COMPUTATIONAL RESULTS FOR MDAGS AND GDAGS USING HLP'S REDUCTION TECHNIQUES

After enumeration, out of the total of $29,989,052$ GDAGs of total nodes seven, $1,618,679$ remain "unclassified" after the application of HLP's criteria for $\mathcal{C} = \mathcal{I}$. After applying HLP's reduction techniques (in Appendix A.1) on $1,618,679$ left GDAGs, we find that $1,618,600$ GDAGs are reducible to just 79 GDAGs which again implies that if these 79 GDAGs of seven total nodes are INTERESTING then so would be all the $1,618,679$ GDAGs as well. Applying our theorem based on e-separation says that all but twenty GDAGs are definitely INTERESTING. Out of these twenty GDAGs, sixteen are mDAGs and the rest of the four GDAGs are equivalent to either of these sixteen mDAGs. So if these sixteen mDAGs (or interpreted as GDAGs here) are INTERESTING then so would be all the left twenty GDAGs of total nodes seven. We observe that as it should be these sixteen mDAGs (or equivalently GDAGs) which remain unclassified are precisely those sixteen mDAGS of four visible nodes that remained "unclassified" after we had applied HLP's criteria, d-separation and e-separation techniques on the list of 2809 mDAGs of four visible nodes. Fifteen out of these sixteen GDAGs are already shown INTERESTING as mDAGs using Fraser's algorithm for supports. Table A.4 enumerates these results. The one DAG that remains is the mDAG of total nodes seven and visible nodes four. It is the first mDAG in Table A.2.

We provided an example showing the working of HLP's criterion and reduction techniques in Figure 3.3 and Appendix A.1 and showed the working of our theorem based on e-separation in Figures [A.4, A.5].

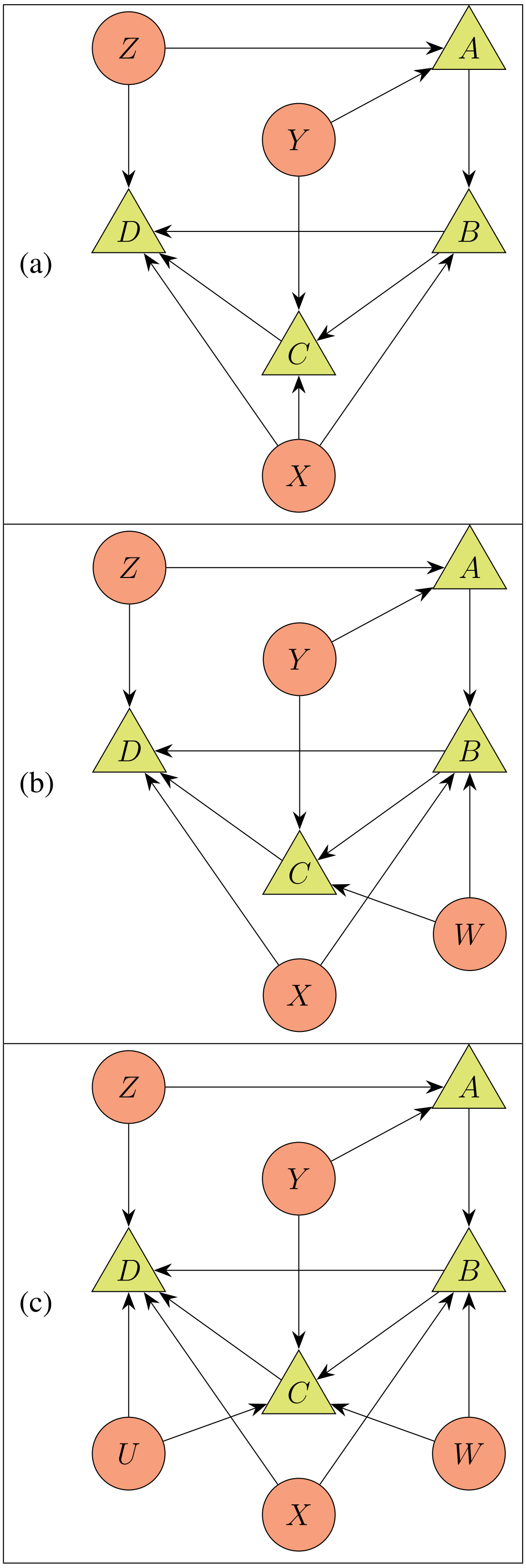


Table A.2: The mDAGs of 4 observed nodes whose Interestingness remains unresolved.

Inference	**Causal scenarios left after**	**Number of mDAGs of 4 visible nodes**	**Number of mDAGs of 5 visible nodes**
1	Enumeration	2809	1718596
2	HLP's $\mathcal{C} = \mathcal{I}$ criterion	996	1009961
3	d-separation to check classical observational equivalence to a latent free DAG	978	1000004
4	e-separation to check classical observational equivalence to a latent free DAG	168	271895
5	Fraser's algorithm to check classical observational equivalence to a latent free DAG	3	...

Table A.3: We enumerate the 2809 possible mDAGs of four visible nodes. Applying HLP's criterion for $\mathcal{C} = \mathcal{I}$ reduces them to 996. Then we check if any of the left 996 mDAGs have conditional independences that cannot be reproduced by any latent free DAG. These mDAGs as explained are definitely INTERESTING. After accounting for these mDAGs we are left with 978 mDAGs which have conditional independences that can all be reproduced by some latent free DAG. So we check how many mDAGs out of 978 are such that they have at least one e-separation relation that cannot be achieved in their corresponding latent free mDAG which had all the same d-separation relations as the mDAG itself. Again such mDAGs are definitely INTERESTING. Accounting for this leaves us with a figure of 168 mDAGs that could be potentially INTERESTING. To affirm their INTERESTINGNESS we refer to Fraser's support algorithm and find supports that are not infeasible in the mDAG but are feasible in its corresponding latent free mDAG with all the same d-separation and e-separation relations as the mDAG, for all but just three mDAGs.

For completeness we also present the mDAGs which we could classify as INTERESTING by finding classically infeasible supports in Figure 4.4 and Table 4.1.

A.8 Finding a difference between the Shannon polytopes corresponding to $\mathcal{C}$ and $\mathcal{I}$ for a DAG

In a work in progress we try to accelerate Fourier-Motzkin elimination and generate such Shannon-type inequalities for $\mathcal{C}$ which are not Shannon-type inequalities for $\mathcal{I}$. Note that even on finding such Shannon- type inequalities for $\mathcal{C}$, it could be that they are non Non-Shannon type inequalities for $\mathcal{I}$ and thus finding them will not necessarily show that the mDAG under consideration is definitely INTERESTING. We would still need to find probability distributions in $\mathcal{I}$ that violate them to confirm

Inference	**Causal scenarios left after**	**Number of mDAGs of 4 visible nodes**	**Number of DAGs of 7 total nodes**
1	Enumeration	2809	29,989,052
2	HLP's $\mathcal{C} = \mathcal{I}$ criterion	996	1,618,679
3	HLP's reduction techniques	87	79
4	Our theorem based on e-separation	54	20
5	Fraser's algorithm for classically infeasible supports	3	1

Table A.4: We enumerate both mDAGs of four visible nodes and GDAGs of total nodes seven. Applying HLP's criterion for $\mathcal{C} = \mathcal{I}$ reduces them to 996 and $1,618,679$ respectively. We further apply HLP's reduction criteria on them which returns that $909 \ \ (996 - 87)$ left mDAGs are reducible to 87 mDAGs and $1,618,600 \ \ (1,618,679 - 79)$ left GDAGs are reducible to 79 DAGs. This just means that if the corresponding 87 mDAGs and 79 GDAGs are Interesting then all the 996 mDAGs and all the $1,618,679$ GDAGs are also Interesting. Hence we are finally left with 87 mDAGs and 79 GDAGs to check for Interestingness. Application of our theorem based on e-separation leaves 54 mDAGs and 20 GDAGs to be checked out of which Fraser's algorithm provides classically infeasible supports for 51 mDAGs and 19 GDAGs. Note that out of these 20 DAGs only 16 are mDAGs and the rest of the 4 GDAGs are equivalent to either of these 16 mDAGs but the one for which we couldn't find any classically infeasible supports.

their Interestingness. As we show in this work, this is useful in generating Shannon type inequalities for $\mathcal{C}$ corresponding to some other causal scenarios for which we either couldn't generate Shannon type inequalities or for which their generation took significant computational time without using the advantages we report in this upcoming work.

Also, the complexity of Fourier-Motzkin elimination to find a difference between the Shannon polytopes corresponding to $\mathcal{C}$ and $\mathcal{I}$ for some DAGs can be avoided using the below techniques.

1. Trying to find valid entropy vectors for $\mathcal{I}$ corresponding to some heuristically chosen probability distributions which might not be valid entropy vectors for $\mathcal{C}$. We couldn't find probability distributions which might give rise to entropy vectors that are not in $\mathcal{C}$ but are in $\mathcal{I}$ for the three mDAGs we were concerned with, but we could use this method to success in some other DAGs of 6 nodes.

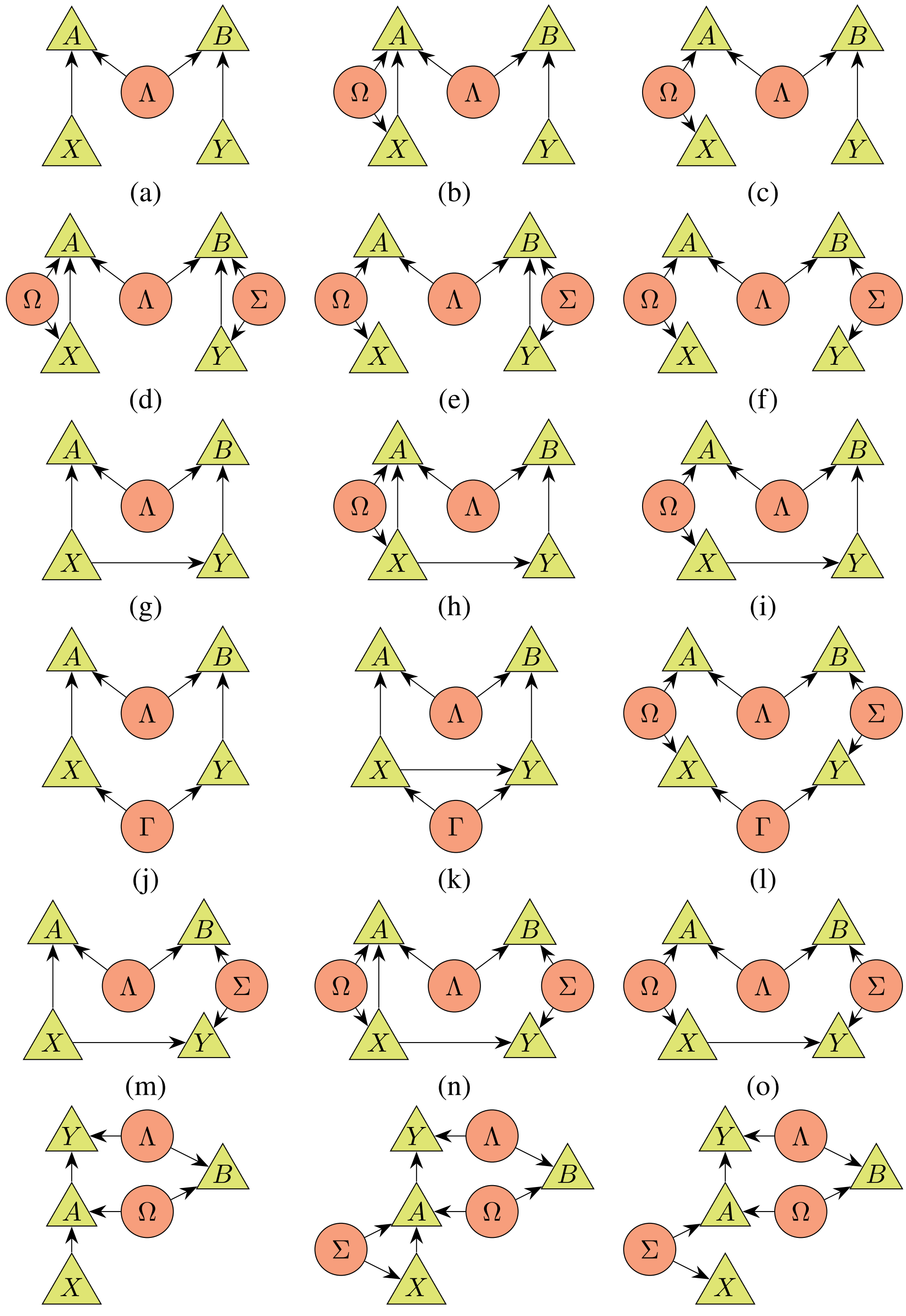
2. Instead of eliminating unobserved variables and trying to project the Shannon cone for $\mathcal{C}$ over the observed variables (using Fourier-Motzkin elimination), it could be better to compare the full Shannon cone over *all* the variables for $\mathcal{C}$ with that of $\mathcal{I}$. To do so we generate random vectors which have non-zero (randomly generated) coefficients corresponding to only the observed

variables. The coefficients corresponding to unobserved variables are zero, because $\mathcal{I}$ is meaningless for the unobserved variables. We then test if this randomly selected vector (a particular direction in the space of all variables) attains a minimum value subject to all the Shannon type inequalities over *all* the variables (hence no usage of Fourier-Motzkin elimination) corresponding to $\mathcal{C}$ that is strictly less than the minimum value the same randomly selected vector takes subject to all the Shannon type inequalities over all variables corresponding to $\mathcal{I}$.[3] If we can find such a random vector then we can attest that Shannon polytope corresponding to $\mathcal{C}$ is strictly smaller than the Shannon polytope corresponding to $\mathcal{I}$. But again for even 2000000 randomly generated vectors and generated many times we couldn't find any such random vector that would help solve the problem for the left three mDAGs. It's interesting to mention that for some smaller DAGs that we tried this method could detect a difference between the Shannon polytopes corresponding to $\mathcal{C}$ and $\mathcal{I}$ very quickly.

3. Applying Fritz trick of perfect predictability [33] to reduce these bigger scenarios to a smaller one and then trying to find a difference between the Shannon cones corresponding to $\mathcal{C}$ and $\mathcal{I}$ for this smaller scenario (for which Fourier-Motzkin elimination is possible) under the condition of certain perfect predictability. The condition of perfect predictability makes the scenarios small and helps in carrying the Fourier-Motzkin elimination to enumerate the Shannon-type inequalities for $\mathcal{C}$. But together with the condition of perfect predictability we could not find any Shannon-type inequality separating the Shannon polytopes corresponding to $\mathcal{C}$ and $\mathcal{I}$ for these three remaining mDAGs. Though this method worked for some smaller scenarios that we tried.

The above strategies do tend to be generally faster in the case of mDAGs where a very long Fourier-Motzkin calculation accompanied with a huge redundancy removal is encountered.

[3] Note that Shannon type inequalities corresponding to $\mathcal{I}$ can be written using all the variables instead of just the observed ones by simply adding all the unobserved variables but simultaneously making their coefficients zero.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

(m) (n) (o)

(p) (q) (r)

Figure A.2: A maximal DAG has a set of d-separation relations unmatched by any latent-free DAG (and is thus INTERESTING by Theorem 44) if and only if it has one of these eighteen patterns as a subgraph.

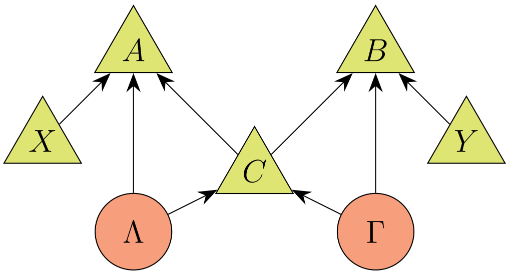


Figure A.3: nonmaximal DAG. It does not have any of the DAGs of Figure A.2 as a subgraph, but it nevertheless has a d-separation pattern that does not correspond to any latent-free DAG.

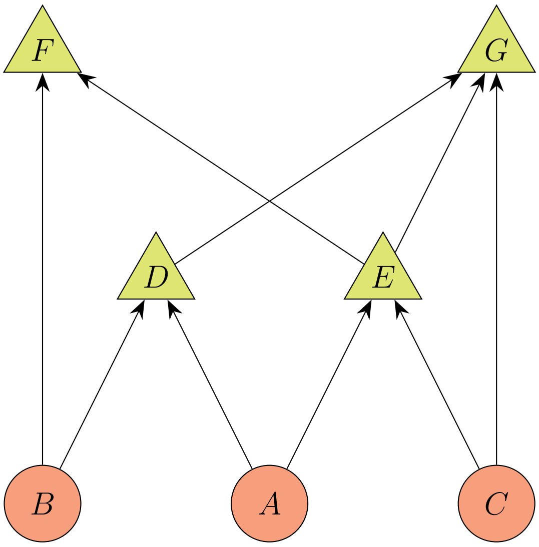


Figure A.4: A DAG with 4 visible nodes and 7 total nodes that is shown to be INTERESTING by our corrected e-separation theorem. This can be seen because it has $(G \perp_{\mathrm{d}} F|E)_{del_D}$ while E is not a descendant of D and none of the d-separation relations $(G \perp_{\mathrm{d}} F|E)$, $(G \perp_{\mathrm{d}} F|D)$ or $(G \perp_{\mathrm{d}} F|ED)$ hold.

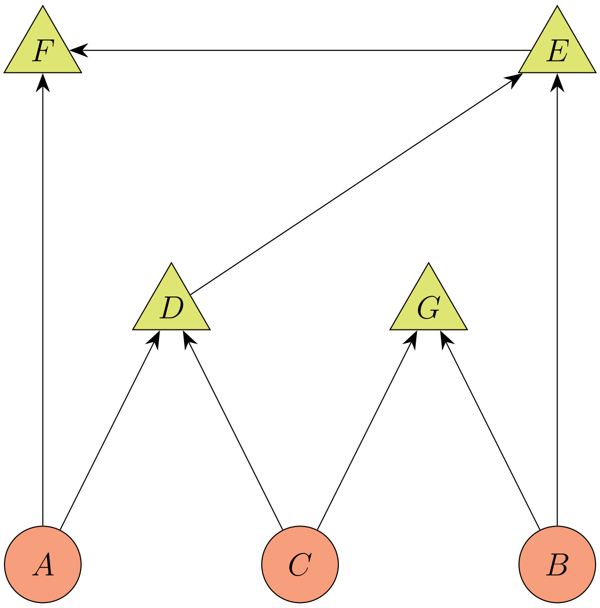


Figure A.5: Another DAG with 4 visible nodes and 7 total nodes that is shown to be INTERESTING by our corrected e-separation theorem. This can be seen because it has $(G \perp_{\mathrm{d}} F)_{del_{DE}}$ while none of the d-separation relations $(G \perp_{\mathrm{d}} F|E)$, $(G \perp_{\mathrm{d}} F|D)$ or $(G \perp_{\mathrm{d}} F|ED)$ hold.

References

[1] A. Acin, N. Gisin, and L. Masanes. “From Bell’s theorem to secure quantum key distribution”. In: *Physical review letters* 97.12 (2006), p. 120405.

[2] A. Acín, N. Brunner, N. Gisin, S. Massar, S. Pironio, and V. Scarani. “Device-independent security of quantum cryptography against collective attacks”. In: *Phys. Rev. Lett.* 98.23 (2007), p. 230501.

[3] A. Acín, S. Massar, and S. Pironio. “Randomness versus Nonlocality and Entanglement”. In: *Phys. Rev. Lett.* 108 (10 Mar. 2012), p. 100402.

[4] J.-M. A. Allen, J. Barrett, D. C. Horsman, C. M. Lee, and R. W. Spekkens. “Quantum common causes and quantum causal models”. In: *Physical Review X* 7.3 (2017), p. 031021.

[5] M. M. Ansanelli, E. Wolfe, and R. Spekkens. “Upcoming Work”. In: (2024).

[6] M. Ardehali. “Bell inequalities with a magnitude of violation that grows exponentially with the number of particles”. In: *Physical Review A* 46.9 (1992), p. 5375.

[7] V. Baldoni, N. Berline, J. A. D. Loera, M. Köppe, and M. Vergne. “How to integrate a polynomial over a simplex”. In: *Mathematics of Computation* 80.273 (July 2010), pp. 297–325.

[8] J. Barrett. “Information processing in generalized probabilistic theories”. In: *Phys. Rev. A* 75 (3 Mar. 2007), p. 032304.

[9] J. Barrett, R. Lorenz, and O. Oreshkov. “Quantum causal models”. In: *arXiv:1906.10726* (2019).

[10] J. S. Bell, M. A. Horne, and A. Zeilinger. “Speakable and Unspeakable in Quantum Mechanics”. In: *Am. J. Phys.* 57.6 (June 1989), pp. 567–567.

[11] J. S. Bell. *The theory of local beables*. Tech. rep. 1975.

[12] J. S. Bell. “On the Einstein-Podolsky-Rosen paradox”. In: *Physics* 1 (1964), pp. 195–200.

[13] C. Branciard, D. Rosset, N. Gisin, and S. Pironio. “Bilocal versus nonbilocal correlations in entanglement-swapping experiments”. In: *Phys. Rev. A* 85 (3 Mar. 2012), p. 032119.

[14] J. Bricmont. *Making Sense of Quantum Mechanics*. Springer International Publishing, 2016.

[15] R. Chaves and T. Fritz. “Entropic approach to local realism and noncontextuality”. In: *Phys. Rev. A* 85 (3 Mar. 2012), p. 032113.

[16] R. Chaves, L. Luft, and D. Gross. “Causal structures from entropic information: geometry and novel scenarios”. In: *New J. Phys.* 16 (2014), p. 043001.

[17] R. Chaves, C. Majenz, and D. Gross. “Information–theoretic implications of quantum causal structures”. In: *Nature Communications* 6.1 (Jan. 2015).

[18] J. F. Clauser, M. A. Horne, A. Shimony, and R. A. Holt. “Proposed Experiment to Test Local Hidden-Variable Theories”. In: *Phys. Rev. Lett.* 23 (15 Oct. 1969), pp. 880–884.

[19] R. Colbeck and A. Kent. “Private randomness expansion with untrusted devices”. In: *Journal of Physics A: Mathematical and Theoretical* 44.9 (2011), p. 095305.

[20] F. Costa and S. Shrapnel. “Quantum causal modelling”. In: *New Journal of Physics* 18.6 (2016), p. 063032.

[21] G. B. Dantzig. *Fourier-Motzkin elimination and its dual*. Tech. rep. STANFORD UNIV CA DEPT OF OPERATIONS RESEARCH, 1972.

[22] D. Dürr, S. Goldstein, T. Norsen, W. Struyve, and N. Zanghì. “Can Bohmian mechanics be made relativistic?” In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 470.2162 (2014), p. 20130699.

[23] D. Dürr and D. Lazarovici. *Understanding Quantum Mechanics*. Springer International Publishing, 2020.

[24] D. Dürr, S. Goldstein, R. Tumulka, and N. Zanghí. “Bohmian mechanics”. In: *Compendium of quantum physics*. Springer, 2009, pp. 47–55.

[25] A. Einstein, B. Podolsky, and N. Rosen. “Can Quantum-Mechanical Description of Physical Reality Be Considered Complete?” In: *Phys. Rev.* 47 (10 May 1935), pp. 777–780.

[26] R. J. Evans. “Graphical methods for inequality constraints in marginalized DAGs”. In: *2012 IEEE International Workshop on Machine Learning for Signal Processing*. IEEE, Sept. 2012, pp. 1–6.

[27] R. J. Evans. “Graphs for Margins of Bayesian Networks”. In: *Scandinavian J. Statistics* 43.3 (Nov. 2015), pp. 625–648.

[28] R. J. Evans. “Latent-free equivalent mDAGs”. In: *arXiv:2209.06534* (2022).

[29] N. Finkelstein, B. Zjawin, E. Wolfe, I. Shpitser, and R. W. Spekkens. “Entropic inequality constraints from e-separation relations in directed acyclic graphs with hidden variables”. In: *Uncertainty in Artificial Intelligence*. PMLR. 2021, pp. 1045–1055.

[30] J. Fourier. “Histoire de l’académie, partie mathématique”. In: *Mémoire de l’Académie des sciences de l’Institut de France* (1824).

[31] T. C. Fraser. “A Combinatorial Solution to Causal Compatibility”. In: *J. Causal Inference* 8.1 (Jan. 2020), pp. 22–53.

[32] T. Fritz. “Beyond Bell’s Theorem II: Scenarios with Arbitrary Causal Structure”. In: *Comm. Math. Phys.* 341.2 (Nov. 2015), pp. 391–434.

[33] T. Fritz. “Beyond Bell’s theorem: correlation scenarios”. In: *New J. Phys.* 14 (2012), p. 103001.

[34] T. Fritz. “Beyond Bell’s theorem: correlation scenarios”. In: *New Journal of Physics* 14.10 (2012), p. 103001.

[35] A. Garg and N. Mermin. “Bell inequalities with a range of violation that does not diminish as the spin becomes arbitrarily large”. In: *Physical Review Letters* 49.13 (1982), p. 901.

[36] D. Geiger and J. Pearl. “On the Logic of Causal Models”. In: *Proc. 4th Conf. UAI* (1988), pp. 136–147.

[37] D. Geiger, T. Verma, and J. Pearl. “Identifying independence in Bayesian Networks”. In: *Networks* 20.5 (1990), pp. 507–534.

[38] T. Gläßle, D. Gross, and R. Chaves. “Computational tools for solving a marginal problem with applications in Bell non-locality and causal modeling”. In: *J. Phys. A* 51.48 (Nov. 2018), p. 484002.

[39] C. Glymour, K. Zhang, and P. Spirtes. “Review of causal discovery methods based on graphical models”. In: *Frontiers in genetics* 10 (2019), p. 524.

[40] S. Goldstein, T. Norsen, D. V. Tausk, and N. Zanghi. “Bell’s theorem”. In: *Scholarpedia* 6.10 (2011). revision #91049, p. 8378.

[41] S. Goldstein, T. Norsen, D. V. Tausk, and N. Zanghì. “Bell’s theorem”. In: *Scholarpedia* 6.10 (2011), p. 8378.

[42] L. Hardy. “Quantum mechanics, local realistic theories, and Lorentz-invariant realistic theories”. In: *Phys. Rev. Lett.* 68 (20 May 1992), pp. 2981–2984.

[43] N. Harrigan and R. W. Spekkens. “Einstein, incompleteness, and the epistemic view of quantum states”. In: *Foundations of Physics* 40 (2010), pp. 125–157.

[44] J. Henson, R. Lal, and M. F. Pusey. “Theory-independent limits on correlations from generalized Bayesian networks”. In: *New J. Phys.* 16.11 (2014), p. 113043. ISSN: 1367-2630.

[45] “How to teach special relativity”. In: *John S Bell On The Foundations Of Quantum Mechanics*. World Scientific, 2001, pp. 61–73.

[46] *https://github.com/eliewolfe/mDAG-analysis*. Note that this repository has not been optimized for public use. We nevertheless include the link in the spirit of transparency.

[47] *https://github.com/shashaank38/gdag _code*. Note that this repository has not been optimized for public use. We nevertheless include the link in the spirit of transparency.

[48] *https://pypi.org/project/causality/*. This package contains tools for causal analysis using observational (rather than experimental) datasets.

[49] J.-L. Imbert. “About redundant inequalities generated by Fourier’s algorithm”. In: *Artificial Intelligence IV*. Elsevier, 1990, pp. 117–127.

[50] J.-L. Imbert. “Fourier’s elimination: Which to choose?” In: *PPCP*. Vol. 1. Citeseer. 1993, pp. 117–129.

[51] N. Karmarkar. “A new polynomial-time algorithm for linear programming”. In: *Proceedings of the sixteenth annual ACM symposium on Theory of computing*. 1984, pp. 302–311.

[52] L. Khalfin and B. Tsirelson. “Steklov Mathematical Institute, Leningrad D-11, UssR”. In: *Symposium on the Foundations of Modern Physics: 50 Years of the Einstein-Podolsky-Rosen Gedankenexperiment, Joensuu, Finland, 16-20 June 1985*. World Scientific Publishing Company Incorporated. 1985, p. 441.

[53] S. Khanna, M. M. Ansanelli, M. F. Pusey, and E. Wolfe. “Classifying causal structures: Ascertaining when classical correlations are constrained by inequalities”. In: *Phys. Rev. Res.* 6 (2 Apr. 2024), p. 023038.

[54] S. Khanna, S. Halder, and U. Sen. “Quantum entanglement percolation under a realistic restriction”. In: *Physical Review A* 109.1 (2024), p. 012419.

[55] S. Khanna, M. Pusey, and R. Colbeck. “Upcoming Work”. In: (2024).

[56] S. Khanna, M. Pusey, and R. Colbeck. “Upcoming Work”. In: (2024).

[57] T. Kriváchy, Y. Cai, D. Cavalcanti, A. Tavakoli, N. Gisin, and N. Brunner. “A neural network oracle for quantum nonlocality problems in networks”. In: *npj Quantum Information* 6.1 (2020), p. 70.

[58] “La nouvelle cuisine”. In: *Quantum Mechanics, High Energy Physics And Accelerators: Selected Papers Of John S Bell (With Commentary)*. World Scientific, 1995, pp. 910–928.

[59] S. L. Lauritzen. *Graphical Models*. Oxford University Press, 1996. ISBN: 0-19-852219-3.

[60] L. Mančinska and S. Wehner. “A unified view on Hardy’s paradox and the Clauser–Horne–Shimony–Holt inequality”. In: *Journal of Physics A: Mathematical and Theoretical* 47.42 (Oct. 2014), p. 424027. ISSN: 1751-8121.

[61] N. D. Mermin. “Extreme quantum entanglement in a superposition of macroscopically distinct states”. In: *Physical Review Letters* 65.15 (1990), p. 1838.

[62] N. D. Mermin. “Quantum mechanics vs local realism near the classical limit: A Bell inequality for spin s”. In: *Physical Review D* 22.2 (1980), p. 356.

[63] N. Miklin, A. A. Abbott, C. Branciard, R. Chaves, and C. Budroni. “The entropic approach to causal correlations”. In: *New J. Phys.* 19.11 (Nov. 2017), p. 113041.

[64] D. Monniaux. “Quantifier elimination by lazy model enumeration”. In: *CAV 2010*. Ed. by B. Cook, P. Jackson, and T. Touili. Vol. 6174. Lecture notes in computer science. Edimburgh, United Kingdom: Springer-Verlag, July 2010, pp. 585–599.

[65] MOSEK-APS. *The MOSEK Optimization Suite 11.0.7 for Python*. 2024.

[66] H. NIKOLIĆ. “Making nonlocal reality compatible with relativity”. In: *Int. J. Quantum Info.* 9.supp01 (2011), pp. 367–377.

[67] H. Nikolić. “Quantum mechanics: Myths and facts”. In: *Foundations of Physics* 37 (2007), pp. 1563–1611.

[68] H. Nikolić. “Relativistic quantum mechanics and the Bohmian interpretation”. In: *Foundations of Physics Letters* 18 (2005), pp. 549–561.

[69] T. Norsen. “Against ‘Realism’”. In: *Found. Phys.* 37.3 (Feb. 2007), pp. 311–340.

[70] T. Norsen. “Einstein’s boxes”. In: *Am. J. Phys.* 73.2 (Feb. 2005), pp. 164–176.

[71] T. Norsen. “EPR and Bell locality”. In: *AIP Conference Proceedings*. Vol. 844. 1. American Institute of Physics. 2006, pp. 281–293.

[72] T. Norsen. *Foundations of Quantum Mechanics*. Springer International Publishing, 2017.

[73] T. Norsen. “John S. Bell’s concept of local causality”. In: *Am. J. Phys.* 79.12 (Dec. 2011), pp. 1261–1275.

[74] T. Norsen. “Local Causality and Completeness: Bell vs. Jarrett”. In: *Found. Phys.* 39.3 (Feb. 2009), pp. 273–294.

[75] T. Norsen. “The theory of (exclusively) local beables”. In: *Foundations of Physics* 40 (2010), pp. 1858–1884.

[76] J. Pearl. *Causality. Models, Reasoning, and Inference*. 2nd ed. Cambridge University Press, 2009. ISBN: 978-0521895606.

[77] J. Pearl. “On the Testability of Causal Models with Latent and Instrumental Variables”. In: *Proceedings of the Eleventh Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-95)*. Morgan Kaufmann, 1995, pp. 435–443.

[78] J. Pearl. “The Art and Science of Cause and Effect”. UCLA 81st Faculty Research Lecture Series. 1996.

[79] J. Pearl and A. Paz. “Graphoids: Graph-Based Logic for Reasoning about Relevance Relations or When would x tell you more about y if you already know z?” In: *Probabilistic and Causal Inference: The Works of Judea Pearl*. 2022, pp. 189–200.

[80] J. Pearl and T. S. Verma. “A theory of inferred causation”. In: *Studies in Logic and the Foundations of Mathematics*. Vol. 134. Elsevier, 1995, pp. 789–811.

[81] J. Pienaar. “Which causal structures might support a quantum–classical gap?” In: *New J. Phys.* 19.4 (Apr. 2017), p. 043021.

[82] S. Pironio, A. Acín, S. Massar, A. B. de La Giroday, D. N. Matsukevich, P. Maunz, S. Olmschenk, D. Hayes, L. Luo, T. A. Manning, et al. “Random numbers certified by Bell’s theorem”. In: *Nature* 464.7291 (2010), pp. 1021–1024.

[83] M. Plávala. “General probabilistic theories: An introduction”. In: *arXiv:2103.07469* (2021).

[84] S. Popescu and D. Rohrlich. “Quantum nonlocality as an axiom”. In: *Found. Phys.* 24.3 (1994), pp. 379–385.

[85] L.-N. Pouchet, C. Bastoul, A. Cohen, and J. Cavazos. “Iterative optimization in the polyhedral model: Part II, multidimensional time”. In: *ACM SIGPLAN Notices* 43.6 (2008), pp. 90–100.

[86] M. F. Pusey, J. Barrett, and T. Rudolph. “On the reality of the quantum state”. In: *Nature Physics* 8.6 (2012), pp. 475–478.

[87] H. Reichenbach. *The direction of time*. Vol. 65. Univ of California Press, 1991.

[88] M.-O. Renou, E. Bäumer, S. Boreiri, N. Brunner, N. Gisin, and S. Beigi. “Genuine Quantum Nonlocality in the Triangle Network”. In: *Phys. Rev. Lett.* 123 (14 Sept. 2019), p. 140401.

[89] A. Restivo, N. Brunner, and D. Rosset. “Possibilistic approach to network nonlocality”. In: *arXiv:2208.13526* (2022).

[90] T. Richardson. “A characterization of Markov equivalence for directed cyclic graphs”. In: *Int. J. Approximate Reasoning* 17.2-3 (Aug. 1997), pp. 107–162.

[91] T. Richardson and P. Spirtes. “Ancestral graph Markov models”. In: *The Annals of Statistics* 30.4 (2002), pp. 962–1030.

[92] M. Schechter. “Integration Over a Polyhedron: An Application of the Fourier-Motzkin Elimination Method”. In: *The American Mathematical Monthly* 105.3 (Mar. 1998), pp. 246–251.

[93] R. Scheines. “An introduction to causal inference”. In: (1997).

[94] C. E. Shannon. “Prediction and entropy of printed English”. In: *Bell system technical journal* 30.1 (1951), pp. 50–64.

[95] P. Spirtes, C. Glymour, R. Scheines, et al. “Causation, Prediction, and Search”. In: *MIT Press Books* 1 (2001).

[96] G. Stehr, H. E. Graeb, and K. J. Antreich. “Analog Performance Space Exploration by Normal-Boundary Intersection and by Fourier–Motzkin Elimination”. In: *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 26.10 (Oct. 2007), pp. 1733–1748.

[97] B. Steudel and N. Ay. “Information-Theoretic Inference of Common Ancestors”. In: *Entropy* 17.4 (Apr. 2015). Note that the journal version and the arXiv version of this citation differ significantly in terms of the enumeration of theorems and examples. See Theorem 2 and Example 2 of the *journal* version., pp. 2304–2327.

[98] J. Tian, A. Paz, and J. Pearl. *Finding minimal d-separators*. Computer Science Department, University of California, 1998.

[99] U. Vazirani and T. Vidick. “Fully device independent quantum key distribution”. In: *Comm. ACM* 62.4 (2019), pp. 133–133.

[100] T. Verma and J. Pearl. “An algorithm for deciding if a set of observed independencies has a causal explanation”. In: *Uncertainty in artificial intelligence*. Elsevier. 1992, pp. 323–330.

[101] T. Verma and J. Pearl. “Causal Networks: Semantics and Expressiveness”. In: *Proceedings of the 4th Workshop on Uncertainty in Artificial Intelligence*. 1988, pp. 352–359.

[102] V. Vilasini. “Approaches to causality and multi-agent paradoxes in non-classical theories”. In: *arXiv preprint arXiv:2102.02393* (2021).

[103] L. Walleghem, S. Khanna, and R. Bhavsar. “Comment on a no-go theorem for ψ-ontic models”. In: *arXiv preprint arXiv:2402.13140* (2024).

[104] M. Weilenmann and R. Colbeck. “Analysing causal structures with entropy”. In: *Proc. Roy. Soc. A* 473.2207 (Nov. 2017), p. 20170483.

[105] K. B. Wharton and N. Argaman. “Colloquium: Bell’s theorem and locally mediated reformulations of quantum mechanics”. In: *Reviews of Modern Physics* 92.2 (2020), p. 021002.

[106] H. Williams. “Fourier-Motzkin elimination extension to integer programming problems”. In: *Journal of Combinatorial Theory, Series A* 21.1 (July 1976), pp. 118–123.

[107] E. Wolfe, R. W. Spekkens, and T. Fritz. “The Inflation Technique for Causal Inference with Latent Variables”. In: *J. Caus. Inf.* 7.2 (July 2019), p. 20170020.

[108] C. J. Wood and R. W. Spekkens. “The lesson of causal discovery algorithms for quantum correlations: Causal explanations of Bell-inequality violations require fine-tuning”. In: *New J. Phys.* 17.3 (2015), p. 033002.

[109] İ. Yanıkoğlu, B. L. Gorissen, and D. den Hertog. “A survey of adjustable robust optimization”. In: *European Journal of Operational Research* 277.3 (Sept. 2019), pp. 799–813.

[110] R. Yeung. “A framework for linear information inequalities”. In: *IEEE Transactions on Information Theory* 43.6 (1997), pp. 1924–1934.

[111] R. W. Yeung. *A first course in information theory*. Springer Science & Business Media, 2002.

[112] B. van der Zander, M. Liśkiewicz, and J. Textor. “Separators and adjustment sets in causal graphs: Complete criteria and an algorithmic framework”. In: *Artificial Intelligence* 270 (May 2019), pp. 1–40.

[113] Z. Zhang and R. Yeung. “A non-Shannon-type conditional inequality of information quantities”. In: *IEEE Transactions on Information Theory* 43.6 (1997), pp. 1982–1986.

[114] Z. Zhang and R. Yeung. “On characterization of entropy function via information inequalities”. In: *IEEE Transactions on Information Theory* 44.4 (1998), pp. 1440–1452.

[115] J. Zhen, D. den Hertog, and M. Sim. “Adjustable Robust Optimization via Fourier–Motzkin Elimination”. In: *Operations Research* 66.4 (Aug. 2018), pp. 1086–1100.