

# Inducing Sets: A New Perspective for Ancestral Graph Markov Models

by

**Bryan Andrews**

BA, Franklin & Marshall College, 2015

MS, University of Pittsburgh, 2017

Submitted to the Graduate Faculty of  
the Dietrich School of Arts and Sciences in partial fulfillment  
of the requirements for the degree of  
**Doctor of Philosophy**

University of Pittsburgh

2022

UNIVERSITY OF PITTSBURGH  
DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Bryan Andrews

It was defended on

August 30 2020

and approved by

Gregory F Cooper, M.D., Ph.D., Department of Biomedical Informatics

Sofia Triantafillou, Ph.D., Department of Biomedical Informatics

Panagiotis Benos, Ph.D., Department of Computational and Systems Biology

Peter Spirtes, Ph.D., Department of Philosophy, Carnegie Mellon University

Thomas S Richardson, Ph.D., Department of Statistics, University of Washington

Dissertation Director: Gregory F Cooper, M.D., Ph.D., Department of Biomedical  
Informatics



Copyright © by Bryan Andrews  
2022

# Inducing Sets: A New Perspective for Ancestral Graph Markov Models

Bryan Andrews, PhD

University of Pittsburgh, 2022

Directed acyclic graphs (DAGs) and their corresponding Markov models have become widely studied and applied in the fields of statistics and causality. The simple directed structure of these models facilitates systematic learning procedures and provides an interpretable representation for causal relationships. However, DAGs are ill-equipped to handle latent variables without explicitly invoking them. This manifests as a lack of stability<sup>1</sup> under marginalization and conditioning and a disparity between statistically and causally valid models. Meanwhile, latent confounding and selection effects occur with some regularity in many domains. The family of maximal ancestral graphs (MAGs) extends the family of DAGs by implicitly taking latent variables into account. In fact, the family of MAGs constitutes the smallest superset of the family of DAGs that is stable under marginalization and conditioning. Accordingly, MAGs and their corresponding Markov models—ancestral graph Markov models—provide a natural choice for statistical and causal modeling in systems with latent confounding and selection effects.

In this work we introduce inducing sets as a new perspective for reasoning about ancestral graph Markov models. In particular, we derive and study  $m$ -connecting sets which are a special case of inducing sets and provide an alternative representation for MAGs. We show that  $m$ -connecting sets admit a characterization of Markov equivalence for MAGs and a factorization criterion equivalent to the global Markov property for directed MAGs. Using the factorization criterion, we formulate a consistent probabilistic score with a closed-form for the Markov models of directed MAGs. Ultimately, we design a local causal discovery algorithm called the ancestral probability (AP) procedure which estimates the posterior probabilities of ancestral relationships. We evaluate the AP procedure on synthetically generated data and a real data set measuring airborne pollutants, cardiovascular health, and respiratory health.

---

<sup>1</sup>A graphical family is stable under marginalization and conditioning if the corresponding set of induced independence models is closed under marginalization and conditioning; see Section 3.4.

## Table of Contents

|  |    |
|--|----|
| <b>1.0 List of Algorithms</b> . . . . .                      | 1  |
| <b>Preface</b> . . . . .                                     | 2  |
| <b>2.0 Introduction</b> . . . . .                            | 3  |
| 2.1 Motivation . . . . .                                     | 3  |
| 2.2 Outline . . . . .  | 11 |
| <b>3.0 Background and Related Work</b> . . . . .             | 13 |
| 3.1 Conditional Independence . . . . .                       | 13 |
| 3.1.1 Probabilistic Conditional Independence . . . . .       | 15 |
| 3.2 Partially Ordered Sets . . . . .                         | 16 |
| 3.2.1 Möbius Inversion . . . . .                             | 18 |
| 3.3 Ancestral Graphs . . . . .                               | 22 |
| 3.3.1 Preliminaries . . . . .                                | 22 |
| 3.3.2 Graphical Conditional Independence . . . . .           | 28 |
| 3.3.3 Markov Properties . . . . .                            | 32 |
| 3.3.4 Maximality . . . . .                                   | 35 |
| 3.3.5 Factorization . . . . .                                | 37 |
| 3.3.6 Markov Equivalence . . . . .                           | 41 |
| 3.4 Stable Mixed Graphs . . . . .                            | 44 |
| 3.4.1 Marginalization and Conditioning . . . . .             | 48 |
| 3.4.2 Latent Projections . . . . .                           | 50 |
| 3.5 Alternative Independence Models . . . . .                | 53 |
| 3.5.1 Elementary and Semi-elementary Imsets . . . . .        | 54 |
| 3.5.2 Multiinformation . . . . .                             | 56 |
| 3.5.3 Structural Imsets as Independence Models . . . . .     | 57 |
| 3.5.4 Characteristic Imsets as Independence Models . . . . . | 58 |
| <b>4.0 Inducing Sets</b> . . . . .                           | 61 |

|            |   |            |
|------------|---|------------|
| 4.1        | Equivalence . . . . .   | 64         |
| 4.1.1      | Characterization of Markov Equivalence . . . . .                          | 65         |
| 4.2        | Relation to Other Work . . . . .  | 67         |
| 4.2.1      | Parametrizing Sets and Characteristic Imsets . . . . .                    | 68         |
| 4.2.2      | The Causal Inference Algorithm . . . . .                                  | 69         |
| 4.3        | Factorization . . . . .   | 70         |
| 4.3.1      | Preliminaries . . . . .   | 72         |
| 4.3.2      | Factorization Implies Markov . . . . .                                    | 91         |
| 4.3.3      | Markov Implies Factorization . . . . .                                    | 118        |
| 4.3.4      | Formalization and Alternatives . . . . .                                  | 120        |
| 4.3.5      | Worked-out Examples . . . . .   | 123        |
| <b>5.0</b> | <b>MAG Curved Exponential Families . . . . .</b>                          | <b>127</b> |
| 5.1        | Conditional Gaussian Probability Measures . . . . .                       | 130        |
| 5.1.1      | Conditional Gaussian Marginalization Condition . . . . .                  | 131        |
| 5.2        | Gaussian Probability Measures . . . . .                                   | 134        |
| 5.2.1      | Gaussian Parameterization . . . . .                                       | 134        |
| 5.3        | Lee and Hastie Probability Measures . . . . .                             | 136        |
| 5.3.1      | Binary Transformation . . . . .   | 136        |
| 5.3.2      | Lee and Hastie MAG condition . . . . .                                    | 139        |
| 5.3.3      | Lee and Hastie Parameterization . . . . .                                 | 140        |
| 5.3.4      | Lee and Hastie as Curved Exponential Families . . . . .                   | 142        |
| <b>6.0</b> | <b>Scoring Criterion and Applications . . . . .</b>                       | <b>147</b> |
| 6.1        | Asymptotic Behavior of Directed MAG Curved Exponential Families . . . . . | 148        |
| 6.1.1      | Theoretical Evaluation . . . . .  | 149        |
| 6.1.2      | Empirical Evaluation . . . . .  | 153        |
| 6.2        | Ancestral Probabilities . . . . .   | 173        |
| 6.2.1      | Synthetic Examples and Background Knowledge . . . . .                     | 175        |
| 6.2.2      | Airborne Pollutants' Effect on Health . . . . .                           | 178        |
| <b>7.0</b> | <b>Discussion and Future Work . . . . .</b>                               | <b>185</b> |
| 7.1        | Discussion . . . . .  | 185        |

|   |   |            |
|---|---|------------|
| 7.2   | Future Work . . . . .                                   | 186        |
| <b>Appendix A. List of Notation . . . . .</b>                             |   | <b>187</b> |
| A.1   | General Terms . . . . .                                 | 187        |
| A.2   | Sets of Numbers . . . . .                               | 187        |
| A.3   | General Sets . . . . .                                  | 187        |
| A.4   | Generic Set Symbols . . . . .                           | 188        |
| A.5   | Probability Measures . . . . .                          | 188        |
| A.6   | Independence Models . . . . .                           | 189        |
| A.7   | Partially Ordered Sets . . . . .                        | 189        |
| A.8   | General Graph Terms . . . . .                           | 189        |
| A.9   | Functions of Vertices . . . . .                         | 190        |
| A.10  | Functions on Graphs . . . . .                           | 190        |
| A.11  | Evans' Partitioning Terms . . . . .                     | 191        |
| A.12  | Interaction Terms . . . . .                             | 191        |
| A.13  | Stable Mixed Graphs . . . . .                           | 191        |
| A.14  | Constrained Subsets . . . . .                           | 191        |
| A.15  | Integer-valued Multisets . . . . .                      | 192        |
| A.16  | Non-m-connecting Sets as Imsets . . . . .               | 192        |
| A.17  | Curved Exponential Families . . . . .                   | 192        |
| A.18  | Parameterization . . . . .                              | 193        |
| <b>Appendix B. Additional Background, Examples, and Results . . . . .</b> |   | <b>194</b> |
| B.1   | Latent Projections . . . . .                            | 194        |
| B.2   | The Causal Inference Algorithm . . . . .                | 197        |
| B.3   | NSI Finds Non-minimal Solutions . . . . .               | 197        |
| B.4   | Necessity of the Adjustment Term . . . . .              | 200        |
| B.5   | Comparison to Bayesian Scoring of Constraints . . . . . | 203        |
| B.6   | Shifted NLL Comparison . . . . .                        | 206        |
| B.7   | Exact Histograms . . . . .                              | 214        |
| B.8   | AP Calibration . . . . .                                | 224        |
| B.9   | Full Airborne Pollutants Tables . . . . .               | 226        |

|  |     |
|--|-----|
| Appendix C. Factorization of Graphs with Five Vertices . . . . . | 228 |
| Bibliography . . . . .   | 274 |

## List of Tables

|     |  |     |
|-----|--|-----|
| 6.1 | Mean run time for graphs (std in parentheses) with 4 vertices in seconds with two decimal places of precision for BIC and $\hat{B}IC$ . Statistical significance at an alpha level of 0.05 is reported as either an overline if $\hat{B}IC$ is better or an underline if the alternative method is better. . . . . | 165 |
| 6.2 | Mean run time for graphs (std in parentheses) with 5 vertices in seconds with one decimal place of precision for BIC and two decimal places of precision for $\hat{B}IC$ (100 reps). . . . .   | 173 |
| 6.3 | NAAQS airborne pollutants and cardiovascular disease results. . . . .  | 182 |
| 6.4 | NAAQS airborne pollutants and respiratory disease results. . . . .   | 182 |
| 6.5 | NESHAP Airborne Pollutants . . . . .   | 183 |
| 6.6 | Exceptions to NESHAP Airborne Pollutants . . . . .   | 184 |
| B1  | Complete airborne pollutants and cardiovascular disease results. . . . .   | 226 |
| B2  | Complete airborne pollutants and respiratory disease results. . . . .  | 227 |

## List of Figures

|     |  |    |
|-----|--|----|
| 2.1 | A causal DAG representing a randomized experiment for an ineffective drug with unpleasant side effects. Colored vertices represent selection effects [65]. . . . .   | 4  |
| 2.2 | DAGs representing a randomized experiment for an ineffective drug with unpleasant side effects: (i) a DAG with a valid casual interpretation, but an invalid statistical interpretation; (ii, iii) DAGs with valid statistical interpretations, but invalid causal interpretations. . . . .                          | 6  |
| 2.3 | Marginalization: (i) a DAG with vertices $\{e, h, r, s, t\}$ ; (ii) a MAG corresponding to the marginalization of $h$ . Grayed vertices represent latent variables to be marginalized. . . . .   | 8  |
| 2.4 | Conditioning: (i) a DAG with vertices $\{e, h, r, s, t\}$ ; (ii) a MAG corresponding to the marginalization of $e$ and conditioning of $s$ . Grayed vertices represent latent variables to be marginalized and colored vertices represent latent variables to be conditioned on. . . . .                             | 8  |
| 2.5 | Marginalization and conditioning: (i) a DAG with vertices $\{e, h, r, s, t\}$ ; (ii) a MAG corresponding to the marginalization of $h$ and $e$ and conditioning of $s$ . Grayed vertices represent latent variables to be marginalized and colored vertices represent latent variables to be conditioned on. . . . . | 9  |
| 2.6 | A structural inset which induced the same independence model as the MAG in Figure 2.4 (ii). . . . .  | 10 |
| 3.1 | The Hasse diagram for a poset $P = \mathcal{P}(\{a, b, c\})$ ordered by inclusion. . . . .   | 17 |
| 3.2 | The zeta function of a poset $P = \mathcal{P}(\{a, b, c\})$ ordered by inclusion—the first and second arguments of the zeta function act as row and column indices respectively. . . . .   | 19 |
| 3.3 | The Möbius function of a poset $P = \mathcal{P}(\{a, b, c\})$ ordered by inclusion—the first and second arguments of the Möbius function act as row and column indices respectively. . . . .   | 20 |
| 3.4 | An application of the zeta and Möbius functions of a poset $P = \mathcal{P}(\{a, b, c\})$ ordered by inclusion. . . . .  | 21 |



|      |   |    |
|------|---|----|
| 3.5  | Mixed graphs with vertices $\{a, b\}$ : (i) a mixed graph with a loop $a - a$ and multiple edges $a \overset{\leftarrow}{\rightrightarrows} b$ ; (ii) a acyclic directed loopless mixed graph with multiple edges $a \overset{\leftrightarrow}{\rightrightarrows} b$ ; (iii) a simple acyclic directed graph. . . . . | 24 |
| 3.6  | A mixed graph with vertices $\{a, b, c, d, e\}$ . . . . .   | 25 |
| 3.7  | Subgraphs of the graph in Figure 3.6: (i) the directed subgraph; (ii) the undirected subgraph. . . . .  | 27 |
| 3.8  | Induced subgraphs of the graph in Figure 3.6: (i) the induced subgraph over $\{a, c, d, e\}$ ; (ii) the induced subgraph over $\{a, b, d, e\}$ . . . . .  | 27 |
| 3.9  | An ancestral graph with vertices $\{a, b, c, d, e\}$ . . . . .  | 30 |
| 3.10 | Ancestral graphs with vertices $\{a, b, c, d\}$ : (i) a non-maximal ancestral graph; (ii) a maximal ancestral graph. . . . .  | 36 |
| 3.11 | ADMGs with vertices $\{a, b, c, d\}$ . . . . .  | 38 |
| 3.12 | The heads and tails for the ADMG illustrated in Figure 3.11 (i) and the Hasse diagram for the corresponding poset over the ADMG's heads. . . . .  | 40 |
| 3.13 | The heads and tails for the ADMG illustrated in Figure 3.11 (ii) and the Hasse diagram for the corresponding poset over the ADMG's heads. . . . .   | 40 |
| 3.14 | A Markov equivalence class of MAGs with vertices $\{a, b, c, d, e\}$ : (i) a maximally informative PAG; (ii) a set of Markov equivalent MAGs. . . . .   | 42 |
| 3.15 | The general form of a discriminating path. . . . .  | 43 |
| 3.16 | Stable mixed graphs: (i) a DAG with latent and selection variables; (ii) the projected ribbonless graph; (iii) the projected summary graph; (iv) the projected ancestral graph. All graphs encode the same independence model over the measured variables using $m$ -separation. . . . .                              | 47 |
| 3.17 | Hasse diagrams for posets of graphical families: (i) families of stable mixed graphs and DAGs ordered by inclusion; (ii) independence models of the families of stable mixed graphs and DAGs ordered by inclusion. . . . .  | 48 |
| 3.18 | An elementary imset: $u_{\langle a, b   c \rangle}$ . . . . .   | 55 |
| 3.19 | The Hasse diagram for an elementary imset: $u_{\langle a, b   c \rangle}$ . . . . .   | 55 |

|      |  |     |
|------|--|-----|
| 3.20 | A DAG with vertices $\{a, b, c\}$ and an application of the zeta and Möbius function of a poset $P = \mathcal{P}(V)$ ordered by inclusion as a transition between the standard and characteristic imsets of the DAG. . . . .         | 59  |
| 4.1  | An illustration of various MAGs $\mathcal{G}$ and their corresponding $m$ -connecting sets $\mathcal{M}(\mathcal{G})$ . . . . .  | 62  |
| 4.2  | A comparison of two Markov equivalent ancestral graphs that are (i) not maximal and (ii) maximal, along with their corresponding $m$ -connecting sets $\mathcal{M}(\mathcal{G})$ ; their $m$ -connecting sets are identical. . . . . | 63  |
| 4.3  | A directed MAG with vertices $\{a, b, c, d, e\}$ and the corresponding $m$ -connecting and non- $m$ -connecting sets for the directed MAG. . . . .   | 84  |
| 4.4  | A visualization of $\text{PAIRS}(\mathcal{G}_{abcde}, c)$ applied to the directed MAG in Figure 4.3 and the corresponding base conditional terms. . . . .  | 85  |
| 4.5  | A visualization of $\text{PAIRS}(\mathcal{G}_{abde}, b)$ applied to the directed MAG in Figure 4.3 and the corresponding base conditional terms. . . . .   | 87  |
| 4.6  | A visualization of $\text{PAIRS}(\mathcal{G}_{ade}, d)$ applied to the directed MAG in Figure 4.3 and the corresponding base conditional terms. . . . .  | 88  |
| 4.7  | A visualization of $\text{PAIRS}(\mathcal{G}_{ae}, a)$ applied to the directed MAG in Figure 4.3 and the corresponding base conditional terms. . . . .   | 89  |
| 4.8  | A visualization of $\text{PAIRS}(\mathcal{G}_e, e)$ applied to the directed MAG in Figure 4.3 and the corresponding base conditional terms. . . . .  | 90  |
| 4.9  | An illustration of the minimal latent sets. . . . .  | 92  |
| 4.10 | The Hasse diagram for the poset over sets ordered by inclusion. . . . .  | 95  |
| 4.11 | An illustration of how various sets interact and partition each other. . . . .   | 96  |
| 4.12 | An illustration of how various sets interact and partition each other. . . . .   | 96  |
| 4.13 | An illustration of the setup of $\text{OLMP}(\mathcal{G}_{\leq}, A_1)$ (step <i>i</i> ). . . . .   | 100 |
| 4.14 | An illustration of $\text{OLMP}(\mathcal{G}_{\leq}, A_1)$ (step <i>ii</i> ). . . . .   | 101 |
| 4.15 | An illustration of $\text{OLMP}(\mathcal{G}_{\leq}, A_1)$ (step <i>iii</i> ). . . . .  | 102 |
| 4.16 | An illustration of $\text{OLMP}(\mathcal{G}_{\leq}, A_1)$ (step <i>iv</i> ). . . . .   | 103 |
| 4.17 | An illustration of $\text{OLMP}(\mathcal{G}_{\leq}, A_1)$ (step <i>v</i> ). . . . .  | 104 |
| 4.18 | An illustration of the setup of $\text{OLMP}(\mathcal{G}_{\leq}, A_2)$ (step <i>vi</i> ). . . . .  | 105 |
| 4.19 | An illustration of $\text{OLMP}(\mathcal{G}_{\leq}, A_2)$ (step <i>vii</i> ). . . . .  | 106 |

|      |   |     |
|------|---|-----|
| 4.20 | An illustration of $\text{OLMP}(\mathcal{G}, \leq, A_2)$ (step <i>viii</i> ). . . . .   | 107 |
| 4.21 | An illustration of $\text{OLMP}(\mathcal{G}, \leq, A_2)$ (step <i>ix</i> ). . . . .   | 108 |
| 4.22 | An illustration of $\text{OLMP}(\mathcal{G}, \leq, A_1)$ (step <i>x</i> ). . . . .  | 109 |
| 4.23 | An illustration of $\text{OLMP}(\mathcal{G}, \leq, A_1)$ (step <i>xi</i> ). . . . .   | 110 |
| 4.24 | An illustration of $\text{OLMP}(\mathcal{G}, \leq, A_1)$ (step <i>xii</i> ). . . . .  | 111 |
| 5.1  | The Hasse diagram for the poset over families of probability measures ordered by inclusion. . . . .   | 129 |
| 5.2  | An illustration of the binary transformation . . . . .  | 137 |
| 5.3  | Lee and Hastie probability measures and violations of the marginalization condition. The contours give three standard deviations and the solid black line gives the first principal component. . . . .  | 140 |
| 6.1  | An evaluation of the approximate log-likelihood and BIC for the specified directed MAG with $n = \{500, 5,000, 50,000\}$ . The approximate and exact shifted negative log-likelihoods are compared for a random parameterization. The approximate BIC ranking of the data generating MEC amongst all MECs is shown using histograms. The rate of recovery for the data generating MEC given by the highest scoring approximate BIC score is compared against several other state-of-the-art algorithms. Statistical significance at an alpha level of 0.05 is reported as either an overline if $\hat{\text{BIC}}$ is better or an underline if the alternative method is better. . . . . | 156 |
| 6.2  | An evaluation of the approximate log-likelihood and BIC for the specified directed MAG with $n = \{500, 5,000, 50,000\}$ . Statistical significance at an alpha level of 0.05 is reported as either an overline if $\hat{\text{BIC}}$ is better or an underline if the alternative method is better. . . . .  | 158 |
| 6.3  | An evaluation of the approximate log-likelihood and BIC for the specified directed MAG with $n = \{500, 5,000, 50,000\}$ . Statistical significance at an alpha level of 0.05 is reported as either an overline if $\hat{\text{BIC}}$ is better or an underline if the alternative method is better. . . . .  | 160 |

|      |  |     |
|------|--|-----|
| 6.4  | An evaluation of the approximate log-likelihood and BIC for the specified directed MAG with $n = \{500, 5,000, 50,000\}$ . Statistical significance at an alpha level of 0.05 is reported as either an overline if $\hat{BIC}$ is better or an underline if the alternative method is better. . . . .  | 162 |
| 6.5  | An evaluation of the approximate BIC for random directed MAGs with specified edge ranges and $n = \{500, 5,000, 50,000\}$ . Statistical significance at an alpha level of 0.05 is reported as either an overline if $\hat{BIC}$ is better or an underline if the alternative method is better. . . . . | 163 |
| 6.6  | An evaluation of the approximate BIC for random directed MAGs with specified edge ranges and $n = \{500, 5,000, 50,000\}$ . Statistical significance at an alpha level of 0.05 is reported as either an overline if $\hat{BIC}$ is better or an underline if the alternative method is better. . . . . | 164 |
| 6.7  | An evaluation of the approximate log-likelihood and BIC for the specified directed MAG with $n = \{500, 5,000, 50,000\}$ . Statistical significance at an alpha level of 0.05 is reported as either an overline if $\hat{BIC}$ is better or an underline if the alternative method is better. . . . .  | 166 |
| 6.8  | An evaluation of the approximate log-likelihood and BIC for the specified directed MAG with $n = \{500, 5,000, 50,000\}$ . Statistical significance at an alpha level of 0.05 is reported as either an overline if $\hat{BIC}$ is better or an underline if the alternative method is better. . . . .  | 168 |
| 6.9  | An evaluation of the approximate log-likelihood and BIC for the specified directed MAG with $n = \{500, 5,000, 50,000\}$ . Statistical significance at an alpha level of 0.05 is reported as either an overline if $\hat{BIC}$ is better or an underline if the alternative method is better. . . . .  | 170 |
| 6.10 | An evaluation of the approximate BIC for random directed MAGs with specified edge ranges and $n = \{500, 5,000, 50,000\}$ . Statistical significance at an alpha level of 0.05 is reported as either an overline if $\hat{BIC}$ is better or an underline if the alternative method is better. . . . . | 171 |

|   |     |
|---|-----|
| 6.11 An evaluation of the approximate BIC for random directed MAGs with specified edge ranges and $n = \{500, 5,000, 50,000\}$ . Statistical significance at an alpha level of 0.05 is reported as either an overline if $\hat{BIC}$ is better or an underline if the alternative method is better. . . . . | 172 |
| 6.12 Precision recall curves for ancestral relationships with and without background knowledge. . . . .   | 176 |
| 6.13 Receiver operator curves for ancestral relationships with and without background knowledge. . . . .  | 176 |
| 6.14 Calibration curves for ancestral relationships with and without background knowledge. . . . .  | 177 |
| 6.15 A comparison of two hypotheses for the underlying causal model: (i) the airborne pollutant is a cause of cardiovascular disease and respiratory disease; (ii) the airborne pollutant is confounded with cardiovascular disease and respiratory disease. . . . .  | 180 |

## 1.0 List of Algorithms

|    |  |     |
|----|--|-----|
| 1  | CAUSAL INFERENCE FROM M-CONNECTING SETS $\text{CIM}(\mathcal{M})$ . . . . .            | 69  |
| 2  | $\text{PAIRS}(\mathcal{G}, b)$ . . . . .   | 76  |
| 3  | NON-M-CONNECTING SETS AS IMSETS $\text{NSI}(\mathcal{G}, \leq)$ . . . . .              | 80  |
| 4  | ORDERED LOCAL MARKOV PROPERTY $\text{OLMP}(\mathcal{G}, \leq, A)$ . . . . .            | 99  |
| 5  | BINARY TRANSFORMATION $z(\mathcal{G})$ . . . . .                                       | 138 |
| 6  | $\text{SIMULATE}(\mathcal{G}, n)$ . . . . .  | 153 |
| 7  | ANCESTRAL PROBABILITIES $\text{AP}(x^1, \dots, x^n, \text{Pr}(\mathcal{G}))$ . . . . . | 175 |
| 8  | $\alpha_{\text{RG}}(\mathcal{G}, L, S)$ . . . . .                                      | 195 |
| 9  | $\alpha_{\text{SG}}(\mathcal{G}, L, S)$ . . . . .                                      | 196 |
| 10 | $\alpha_{\text{AG}}(\mathcal{G}, L, S)$ . . . . .                                      | 196 |
| 11 | CAUSAL INFERENCE $\text{CI}(\mathcal{J})$ . . . . .                                    | 197 |

## Preface

I would like to express my utmost appreciation to my advisors: Greg Cooper for his guidance and patience over the past six years and more recently Peter Spirtes for his insights and encouragement. I would also like to thank the members of my PhD committee for their invaluable suggestions and comments. Lastly, I would like to thank my friends and family for their continual support.

The research reported in this dissertation was supported by: grants R01LM012087 and T15LM007059 from the National Library of Medicine, grant IIS-1636786 from the National Science Foundation, grant U54HG008540 from the National Human Genome Research Institute through funds provided by the trans-NIH Big Data to Knowledge initiative, and grant #4100070287 from the Pennsylvania Department of Health. The content of this dissertation is solely the responsibility of the author and does not necessarily represent the official views of these funding agencies.

## 2.0 Introduction

The formulation and analysis of causal models enables the study of causal relationships, which has provided essential insights in many research areas such as economics, environmental science, and medicine. Randomized experiments where a hypothesized cause is manipulated independently of a hypothesized effect are the gold standard for discovering causal relationships. However, in many domains, these experiments are often infeasible, unethical, or prohibitively expensive. Consequently, there is a growing interest in developing methods for causal inference and discovery without the need for experimentation—methods that work with any available experimental data and the plethora of non-experimental data. One such approach utilizes the dual interpretation of graphical Markov models as statistical and causal models.

### 2.1 Motivation

Graphical Markov models are probabilistic models that leverage conditional independence for modeling and inference. In a graphical Markov model, a graph induces an independence model comprised of conditional independence statements represented in a probability measure—vertices correspond to random variables and absent edges coincide with conditional independence statements. The independence model can be characterized by a graphical separation criterion in conjunction with the global Markov property or a probabilistic factorization criterion—both characterizations may be exploited for modeling and inference. The notions of a conditional independence statement and an independence model are made rigorous in Section 3.1.

In recent years, graphical Markov models have become widely applied in the fields of statistics and causality [35, 45, 46, 50]. At the forefront of these methods are Bayesian networks, whose independence models are induced by directed acyclic graphs (DAGs) [16, 44, 58]. The popularity of Bayesian networks is in part due to their comprehensive theory, which



includes the  $d$ -separation criterion and the recursive factorization criterion. The  $d$ -separation criterion in conjunction with the DAG component of a Bayesian network graphically encodes conditional independence statements represented in the probabilistic component of the Bayesian network. Equivalently, the recursive factorization criterion in conjunction with the DAG component of a Bayesian network algebraically encodes conditional independence statements represented in the probabilistic component of the Bayesian network. Both characterizations of the independence model induced by the DAG component of a Bayesian network facilitate systematic learning procedures. Indeed, there are an abundance of algorithms capable of learning these models from data [13, 14, 64, 86].

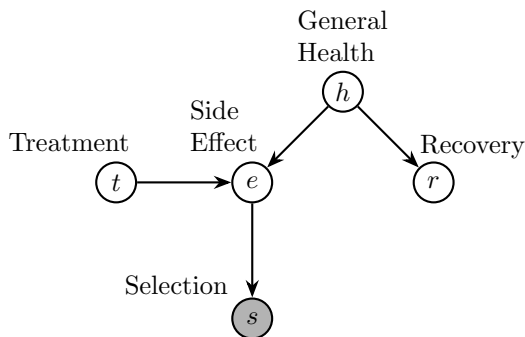


Figure 2.1: A causal DAG representing a randomized experiment for an ineffective drug with unpleasant side effects. Colored vertices represent selection effects [65].

As an example, suppose the DAG depicted in Figure 2.1 induces an independence model comprised of conditional independence statements represented in a probability measure  $P$ . Using the  $d$ -separation criterion and the global Markov property, the following conditional independence statements are represented in  $P$  and graphically encoded by the DAG:

$$r \perp\!\!\!\perp \{e, s, t\} \mid h [P] \quad s \perp\!\!\!\perp \{h, r, t\} \mid e [P] \quad t \perp\!\!\!\perp \{h, r\} [P].$$

This notation is defined in Section 3.1 and attributed to Dawid [17]. Furthermore, the recursive factorization induced by the DAG algebraically encodes the same set of conditional independence statements. The recursive factorization criterion is characterized by the

equivalence of the density admitted by a probability measure with the product of conditional densities defined as a variable conditioned on its parents in the graph. If  $P$  is dominated by a  $\sigma$ -finite product measure  $\nu$  and admits density  $f(x)$ , then the following recursive factorization holds almost everywhere:

$$f(x) = f_{s|e}(x) f_{r|h}(x) f_{e|ht}(x) f_h(x) f_t(x) \quad \text{for } \nu\text{-a.e. } x \in \mathcal{X}.$$

This notation is defined in Section 3.1.

Causal assumptions connect the structural component of a graphical Markov model to causal relationships [79]. These assumptions can be interpreted as an appeal to Occam’s razor—if the true causal model is contained within a family of graphs, then the causal model is a graph that encodes only conditional independence statements represented in the probability measure whose corresponding Markov model has minimal complexity. A causal Bayesian network is a Bayesian network whose independence model is induced by a DAG, whose edges express all the causal relationships and only the causal relationships. These models admit the dual interpretation of graphical Markov models as statistical and causal models. Causal Bayesian networks provide researchers with a means to calculate the effects of intervention without the need for experimentation [51, 59] and have been widely applied in many domains [22, 40, 49, 72, 77].

Unfortunately, the simplicity and theoretical convenience of DAGs comes at the cost of representation power. The set of independence models induced by the family of DAGs is insufficient to represent systems with latent variables without explicitly invoking them. This limitation manifests statistically as a lack of stability under marginalization and conditioning, and causally as a disparity between statistically and causally valid models. Stability under marginalization and conditioning is discussed in Section 3.4. To emphasize this point, consider the following example taken from [65] and attributed to Chris Meek:

The graph [Figure 2.1] represents a randomized [experiment] of an ineffective drug with unpleasant side-effects. Patients are randomly assigned to the treatment or control group [t]. Those in the treatment group suffer unpleasant side-effects [e], the severity of which is influenced by the patient’s general level of health [h], with sicker patients suffering worse side-effects. Those patients who suffer sufficiently severe side-effects are likely to drop out

of the study. The selection variable  $[s]$  records whether or not a patient remains in the study, thus for all those remaining in the study  $[s = \textit{stay in}]$ . Since unhealthy patients who are taking the drug are more likely to drop out, those patients in the treatment group who remain in the study tend to be healthier than those in the control group. Finally health status  $[h]$  influences how rapidly the patient recovers  $[r]$  [65, p.234].

In this example, naïvely comparing the recovery times of the patients remaining in the treatment group against the patients in the control group leads to the incorrect conclusion that the drug is beneficial. The perceived effect is due to the bias towards a good general level of health in the treatment group. Since the remaining patients in the treatment group tend to be healthier, they also tend to recover more quickly. Furthermore, if the patient’s general level of health is allowed to act as a latent confounder, then researchers will be unable to identify this relationship as spurious.

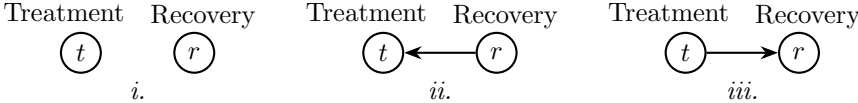


Figure 2.2: DAGs representing a randomized experiment for an ineffective drug with unpleasant side effects: (i) a DAG with a valid casual interpretation, but an invalid statistical interpretation; (ii, iii) DAGs with valid statistical interpretations, but invalid causal interpretations.

Figure 2.2 depicts all possible DAGs over the variables for treatment and recovery—the variables for side effect, general health, and selection are latent. The DAG in (i) is the only valid causal model; it expresses the fact that neither treatment nor recovery cause the other. However, it also implies that treatment and recovery are independent of each other which is false. The DAGs in (ii, iii) correctly imply the dependence between treatment and recovery, but express incorrect causal relationships. Consequently, the family of DAGs is inadequate to represent this example without explicitly invoking the latent variables.

The ubiquity of latent variables necessitates methods capable of dealing with their subtleties. DAGs can model latent variables if the latent variables are explicitly invoked and treated as missing data. However, this approach results in a myriad of problems: there are an infinite number of DAGs with latent variables to consider for each independence model; a DAG with latent variables can encode non-conditional independence constraints; the parameters of a Bayesian network corresponding to a DAG with latent variables are often not fully identifiable; and assumptions about latent variables of a DAG and their parameterization in the corresponding Bayesian network can have a profound impact on modeling and inference including a loss of model smoothness [32, 33, 68, 76, 90].

A more elegant approach is to use a graphical family that is stable under marginalization and conditioning. These families are usually comprised of mixed graphs which are named for the mixture of edge types that they contain: directed, bi-directed, and undirected. Maximal ancestral graphs (MAGs) make up one such family. A thorough treatment of graphical families stable under marginalization and conditioning is given by [73] and discussed in Section 3.4. The set of independence models induced by the family of MAGs is a superset of the set of independence models induced by DAGs. Accordingly, MAGs can represent all (and only) independence models obtained through marginalization and conditioning of the independence models induced by DAGs [70]. This is of interest because graphical Markov models can represent latent confounding as the marginalization of latent variables and selection effects as the conditioning of latent variables—conditioning on latent variables applies a selection effect [4, 79].

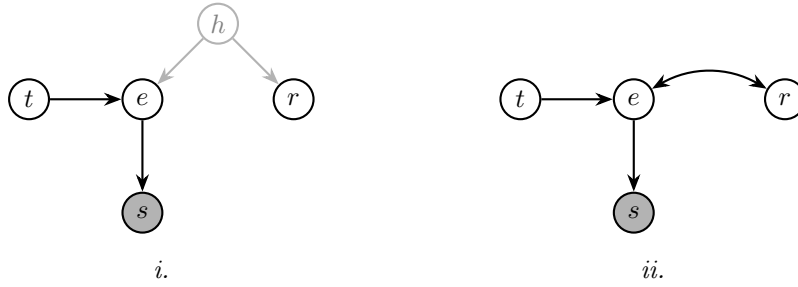


Figure 2.3: Marginalization: (i) a DAG with vertices  $\{e, h, r, s, t\}$ ; (ii) a MAG corresponding to the marginalization of  $h$ . Grayed vertices represent latent variables to be marginalized.

Figure 2.3 depicts an example of marginalization in a DAG where the grayed vertices of the DAG in (i) are the variables to be marginalized—the MAG in (ii) is the resulting graph. The marginalization of  $h$  induces a dependence between  $e$  and  $r$  which corresponds to the bi-directed edge between them. Generally, latent confounding is represented with bi-directed edges.

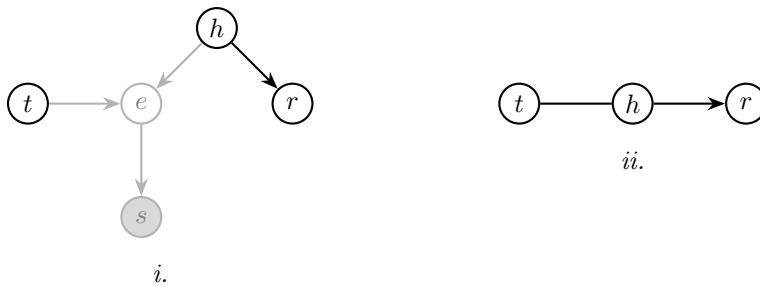


Figure 2.4: Conditioning: (i) a DAG with vertices  $\{e, h, r, s, t\}$ ; (ii) a MAG corresponding to the marginalization of  $e$  and conditioning of  $s$ . Grayed vertices represent latent variables to be marginalized and colored vertices represent latent variables to be conditioned on.

Figure 2.4 depicts an example of conditioning in a DAG where the grayed vertices of the DAG in (i) are the variables to be marginalized and the colored vertices of the DAG

in (ii) are the variables to be conditioned on—the MAG in (ii) is the resulting graph. The conditioning of  $s$  induces a dependence between  $h$  and  $t$  which corresponds to the undirected edge between them. Generally, selection effects are represented with undirected edges.

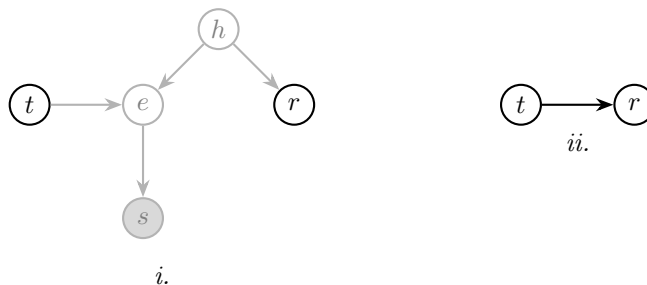


Figure 2.5: Marginalization and conditioning: (i) a DAG with vertices  $\{e, h, r, s, t\}$ ; (ii) a MAG corresponding to the marginalization of  $h$  and  $e$  and conditioning of  $s$ . Grayed vertices represent latent variables to be marginalized and colored vertices represent latent variables to be conditioned on.

Figure 2.5 depicts an example of marginalization and conditioning in a DAG where the grayed vertices of the DAG in (i) are the variables to be marginalized and the colored vertices of the DAG in (ii) are the variables to be conditioned on—the MAG in (ii) is the resulting graph. The marginalization of  $t$  and conditioning of  $s$  induces a dependence between  $r$  and  $t$  which corresponds to the directed edge between them. The MAG in (ii) is statistically and causally valid, however, the causal interpretation of the edges of a MAG is slightly different from the causal interpretation of the edges of a DAG. The MAG in (ii) expresses that the variable for treatment is either a causal ancestor of the variable for recovery or a causal ancestor of a selection variable. In actuality, treatment is an ancestor of the selection variable. The general causal interpretation of MAGs is given in Section 3.3.

Ancestral graph Markov models are graphical Markov models whose independence models are induced by MAGs. Similar to Bayesian networks, Ancestral graph Markov models can sometimes provide researchers with a means to calculate the effects of intervention without the need for experimentation [62, 92]. Additionally, ancestral graph Markov models are

equipped with the  $m$ -separation criterion [70, 66] and the heads and tails factorization criterion [67, 30]. The heads and tails factorization criterion consists of multiple factorizations for marginal densities admitted by a probability measure and only applies to ancestral graph Markov models whose independence models are induced by directed MAGs—MAGs with no undirected edges. The factorization criterion can be extended to all ancestral graph Markov models by factoring the part of the model corresponding to the undirected section of the MAG using the factorization criterion for undirected graph Markov models. These characterizations may be exploited for modeling and inference, but the system of factorizations given by the heads and tails factorization criterion does not readily admit a closed-form objective function for model selection—a closed-form objective function for the model selection of MAGs is a key topic discussed in this dissertation. MAGs and their properties are discussed in Section 3.3.

Graphs are not the only mathematical object used to encode conditional independence. Imsetal Markov models use structural imsets, short for integer-valued multiset, rather than graphs to encode the conditional independence statements represented in a probability measure. Structural imsets are equipped with an analogue to graphical separation criteria and a product formula which can be used as a factorization criterion. Additionally, the family of structural imsets induces a richer set of independence models [83]. Unfortunately, they lack an intuitive interpretation and as a consequence their literature is largely theoretical. Structural imsets and their properties are discussed in Section 3.5.

$$\begin{array}{rcccccccc}
 T & & \emptyset & \{h\} & \{r\} & \{t\} & \{h,r\} & \{h,t\} & \{r,t\} & \{h,r,t\} \\
 u(T) & & \left[ \begin{array}{cccccccc}
 0 & 1 & 0 & 0 & -1 & -1 & 0 & 1
 \end{array} \right]^T
 \end{array}$$

Figure 2.6: A structural imset which induced the same independence model as the MAG in Figure 2.4 (ii).

Figure 2.6 depicts a structural imset which induced the same independence model as the

MAG in Figure 2.4 (ii). The inset is a column vector whose elements correspond to subsets of variables, but may also be thought of as a function  $u : \mathcal{P}(\{h, r, t\}) \rightarrow \mathbb{Z}$  mapping the power set of  $\{h, r, t\}$  to the integers. This representation has theoretical merits, but does not lend itself to an intuitive interpretation, causal or otherwise. Nevertheless, structural imsets have been successfully applied as a framework for DAG learning [37, 84, 85, 86]. To our knowledge, an analogous application for learning MAGs does not exist and is a key topic discussed in this dissertation.

This dissertation introduces inducing sets as a new perspective for reasoning about ancestral graph Markov models. Using this new perspective, we give an alternative representation for MAGs called  $m$ -connecting sets and provide a novel factorization grounded in the theory of structural imsets. Accordingly, we utilize preexisting theoretical machinery from the literature of MAGs graphs and structural imsets and form new connections between them in the process. To demonstrate the effectiveness of this new perspective, we show how the factorization admits a closed-form estimate of the posterior probability of a model; this allows ancestral graph Markov models to be compared, ranked, and averaged. Ultimately, we develop and evaluate the ancestral probability (AP) procedure for computing the posterior probabilities of ancestral relations among pairs of variables.

## 2.2 Outline

This dissertation is organized as follows. Chapter 3 introduces general background information, concepts useful for the study of ancestral graphs, and alternative independence models. Chapter 4 introduces inducing sets and  $m$ -connecting sets as a special case of inducing sets. Additionally, we review related prior work and we prove that the independence models induced by MAGs may be characterized by  $m$ -connecting sets and their factorization. Chapter 5 discusses curved exponential families and derives conditions under which Lee and Hastie probability measures are curved exponential families subject to an independence model induced by a directed MAG. Chapter 6 develops and evaluates a probabilistic score and the ancestral probability (AP) procedure, which performs Bayesian local causal



discovery on directed MAGs. An implementation of the AP algorithm is run on synthetically generated data and a real data set measuring airborne pollutants, cardiovascular health, and respiratory health. Lastly, the dissertation closes with Chapter 7 which summarizes and discusses the main results and provides suggestions for future work.

### 3.0 Background and Related Work

Throughout this dissertation we use the following conventions: upper case symbols, such as  $A$  and  $B$ , denote sets; juxtapositions of upper case letters, such as  $AB = A \cup B$ , denote unions; and lower case symbols, such as  $a$  and  $b$ , denote set elements or singletons. Occasionally in figures and subscripts the juxtaposition of lower case letters, such as  $ab = \{a, b\}$ , denote sets. With a few exceptions that will be noted later, upper case letter in a sans-serif font, such as  $\mathbf{A}$  and  $\mathbf{B}$  denote sets of sets.

The symbol  $V$  denotes a non-empty set of variables—or a set of vertices in the graphical context—that indexes a non-empty finite collection of random variables  $(X_a)_{a \in V}$  with sample spaces  $(\mathcal{X}_a)_{a \in V}$ . These spaces may be finite discrete spaces or finite-dimensional continuous spaces. Given a subset  $A \subseteq V$ , define  $X_A \equiv (X_a)_{a \in A}$  and  $\mathcal{X}_A \equiv \times_{a \in A}(\mathcal{X}_a)$ . Furthermore, denote the fixed elements of  $\mathcal{X}_A$  by  $x_A$ . Lastly, let  $X_V \equiv X$ ,  $\mathcal{X}_V \equiv \mathcal{X}$ , and  $x_V \equiv x$ .

The following symbols are reserved for sets of numbers:  $\mathbb{R}$  denotes the real numbers,  $\mathbb{Q}$  denotes the rational numbers, and  $\mathbb{Z}$  denotes the integers. Furthermore,  $\mathbb{Q}_+$  denotes the non-negative rational numbers, and  $\mathbb{Z}_+$  denotes the non-negative integers. The symbol is reserved for  $\mathbb{S}_{++}^{|n|}$  is the set of  $|n| \times |n|$  symmetric positive definite matrices. The symbol  $\emptyset$  is reserved for the empty set and the symbol  $\mathcal{P}$  is reserved for the power set. Furthermore, the subset of the power set bounded by  $l, u \in \mathbb{Z}_+$  ( $l \leq u$ ) is defined as follows:

$$\mathcal{P}_l^u(V) \equiv \{T \subseteq V ; \quad l \leq |T| \leq u\}.$$

Lastly, let  $\mathcal{P}_l(V) \equiv \mathcal{P}_l^{|V|}(V)$  and  $\mathcal{P}^u(V) \equiv \mathcal{P}_0^u(V)$ .

### 3.1 Conditional Independence

Central to this dissertation are mathematical objects that represent sets of conditional independence statements, called independence models. Conditional independence usually refers to probabilistic conditional independence, that is, conditional independence state-

ments that hold in a probability measure. In this dissertation we use the term conditional independence statement more generally, for instance, conditional independence statements that hold in a graph correspond to separations in that graph; see Section 3.3.2. Mathematical objects that induce independence models include but are not limited to probability measures, mixed graphs, and structural imsets. Let the symbol  $\mathcal{O}$  denote an abstract mathematical object that represents conditional independence statements.

**Definition** (*conditional independence statement*). Let  $V$  be a non-empty set of variables with disjoint subsets  $A, B, C \subseteq V$ . A *conditional independence statement* over  $V$  is a statement of the form “ $A$  is conditionally independent of  $B$  given  $C$ .” Every conditional independence statement over  $V$  corresponds to a disjoint triple of the form  $\langle A, B \mid C \rangle$  and should be understood with respect to a mathematical object. For a mathematical object  $\mathcal{O}$  over  $V$ , if  $\langle A, B \mid C \rangle$  is represented in  $\mathcal{O}$ , then we write  $A \perp\!\!\!\perp B \mid C [\mathcal{O}]$ .

The punctuation of a triple anticipates the intended role for each set. The two former components are independent sets while the third component, written after the vertical bar, is the conditioning set. The corresponding conditional independence statement is *elementary* when the two former sets are singletons and *semi-elementary* otherwise. The set of all disjoint triples over  $V$  is denoted by  $\mathcal{T}(V)$ . Formally, an independence model is defined as follows.

**Definition** (*independence model*). Let  $V$  be a non-empty set of variables and  $\mathcal{O}$  be a mathematical object over  $V$ . The *independence model*  $\mathcal{J}(\mathcal{O})$  induced by  $\mathcal{O}$  is a set of disjoint triples defined as follows:

$$\mathcal{J}(\mathcal{O}) \equiv \{ \langle A, B \mid C \rangle \in \mathcal{T}(V) ; \quad A \perp\!\!\!\perp B \mid C [\mathcal{O}] \}.$$

Let  $V$  be a non-empty set of variables and  $\mathcal{O}$  be a mathematical object over  $V$ . Classes of independence models may be characterized axiomatically as follows. The independence model  $\mathcal{J}(\mathcal{O})$  is called a *semi-graphoid* whenever conditions (*i - v*) hold for every collection of disjoint sets  $A, B, C, D \subseteq V$ :

- i.* triviality       $A \perp\!\!\!\perp \emptyset \mid C [\mathcal{O}]$ ;
- ii.* symmetry       $A \perp\!\!\!\perp B \mid C [\mathcal{O}] \Rightarrow B \perp\!\!\!\perp A \mid C [\mathcal{O}]$ ;

- iii. decomposition  $A \perp\!\!\!\perp BD \mid C [\mathcal{O}] \Rightarrow A \perp\!\!\!\perp D \mid C [\mathcal{O}]$ ;
- iv. weak union  $A \perp\!\!\!\perp BD \mid C [\mathcal{O}] \Rightarrow A \perp\!\!\!\perp B \mid CD [\mathcal{O}]$ ;
- v. contraction  $A \perp\!\!\!\perp B \mid CD [\mathcal{O}]$  and  $A \perp\!\!\!\perp D \mid C [\mathcal{O}] \Rightarrow A \perp\!\!\!\perp BD \mid C [\mathcal{O}]$ .

Furthermore,  $\mathcal{J}(\mathcal{O})$  is called a *graphoid* whenever conditions (i - vi) hold for every collection of disjoint sets  $A, B, C, D \subseteq V$ :

- vi. intersection  $A \perp\!\!\!\perp B \mid CD [\mathcal{O}]$  and  $A \perp\!\!\!\perp D \mid BC [\mathcal{O}] \Rightarrow A \perp\!\!\!\perp BD \mid C [\mathcal{O}]$ .

Lastly,  $\mathcal{J}(\mathcal{O})$  is called a *compositional graphoid* whenever conditions (i - vii) hold for every collection of disjoint sets  $A, B, C, D \subseteq V$ :

- vii. composition  $A \perp\!\!\!\perp B \mid C [\mathcal{O}]$  and  $A \perp\!\!\!\perp D \mid C [\mathcal{O}] \Rightarrow A \perp\!\!\!\perp BD \mid C [\mathcal{O}]$ .

### 3.1.1 Probabilistic Conditional Independence

The most common independence models are induced by probability measures. Let  $V$  be a non-empty set of variables with disjoint subsets  $A, B, C \subseteq V$ . Furthermore, let  $X$  be a collection of random variables indexed by  $V$  with probability measure  $P$  dominated by  $\sigma$ -finite product measure  $\nu$ . We say  $\langle A, B \mid C \rangle$  is represented in  $P$  and write  $A \perp\!\!\!\perp B \mid C [P]$  if for every measurable subset  $T \subseteq \mathcal{X}_A$ :

$$P(X_A \in T \mid X_{BC} = x_{BC}) = P(X_A \in T \mid X_C = x_C) \quad \text{for } P\text{-a.e. } x \in \mathcal{X}. \quad (3.1)$$

In Equation 3.1,  $P(X_A \in T \mid X_{BC})$  does not depend on the value of  $B$ . Intuitively, this conveys that  $B$  provides no additional information about  $A$  when the value of  $C$  is known. Probabilistic conditional independence is a mathematical formalization of this notion of irrelevance [17, 46]. If  $P$  admits density  $f(x)$  with respect to  $\nu$ , then we may define the following equivalent definitions of conditional independence:

$$A \perp\!\!\!\perp B \mid C [P] \Leftrightarrow f_{A|BC}(x) = f_{A|C}(x) \quad \text{for } P\text{-a.e. } x \in \mathcal{X} \quad (3.2)$$

and for some real-valued functions  $g : \mathcal{X}_{AC} \rightarrow \mathbb{R}$  and  $h : \mathcal{X}_{BC} \rightarrow \mathbb{R}$

$$A \perp\!\!\!\perp B \mid C [P] \Leftrightarrow f_{ABC}(x) = g(x)h(x) \quad \text{for } P\text{-a.e. } x \in \mathcal{X}. \quad (3.3)$$

The independence model induced by  $P$  is denoted by  $\mathcal{J}(P)$ . Furthermore, every independence model defined by a probability measure is a semi-graphoid.

**Proposition 3.1.1** (Lemma 2.1 [83]). *Let  $P$  be a probability measure. The induced independence model  $\mathcal{J}(P)$  is a semi-graphoid.*

## 3.2 Partially Ordered Sets

The notion of a partially ordered set provides a principled way to order the vertices of an ancestral graph and is required to define the Möbius inversion. Ancestral graphs are discussed in Section 3.3 and the importance of the Möbius inversion becomes apparent when we are able to understand  $\log f$  as a linear combination of interaction information rates; see Chapter 4 for details. Unless otherwise specified, the symbol  $\mathbf{P}$  denotes a finite partially ordered set. Furthermore, the elements of  $\mathbf{P}$  may be sets—hence our choice of notation. In this dissertation all partially ordered sets are finite.

**Definition** (*partial order*). A partial order is a binary relation  $\leq$  over a set  $\mathbf{P}$  such that  $\leq$  is reflexive, antisymmetric, and transitive. That is, for every collection of mathematical objects  $A, B, C \in \mathbf{P}$ :

- i.* reflexivity       $A \leq A$ ;
- ii.* antisymmetry     $A \leq B$  and  $B \leq A \Rightarrow A = B$ ;
- iii.* transitivity     $A \leq B$  and  $B \leq C \Rightarrow A \leq C$ .

**Definition** (*partially ordered set*). A partially ordered set, *poset* for short, is a set  $\mathbf{P}$  with a partial order  $\leq$ . A pair of mathematical objects  $A, B \in \mathbf{P}$  are *comparable* if  $A \leq B$  or  $B \leq A$  and *incomparable* otherwise. If every pair of elements is comparable, then  $\leq$  is a *total order* and  $\mathbf{P}$  is a *totally ordered set*.

The canonical poset used throughout this dissertation is defined by the power set of a non-empty set of variables  $V$  ordered by inclusion:

$$A \leq B \iff A \subseteq B \quad \text{for all } A, B \subseteq V.$$

Figure 3.1 depicts the Hasse diagram for the poset  $P = \mathcal{P}(\{a, b, c\})$  ordered by inclusion. Vertices represent the elements of  $P$  where vertices appearing higher in the diagram have greater cardinality than vertices appearing lower in the diagram and edges connect sets to their maximal subsets—or minimal supersets.

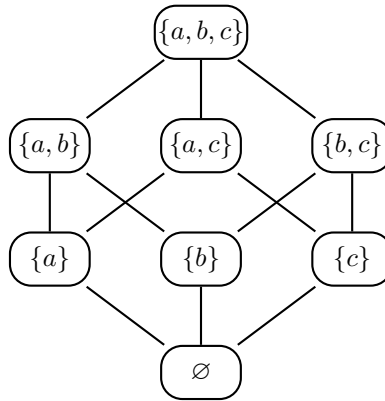


Figure 3.1: The Hasse diagram for a poset  $P = \mathcal{P}(\{a, b, c\})$  ordered by inclusion.

Let  $P$  be a poset with partial order  $\leq$  and consider a pair of mathematical objects  $A, B \in P$ . The join of  $A$  and  $B$ , denoted  $A \vee B$ , is their supremum. Similarly, the meet of  $A$  and  $B$ , denoted  $A \wedge B$ , is their infimum. In general, the join and meet of a pair of mathematical objects might not exist. Figure 3.1 illustrates the concepts of join and meet. In the poset:

- $\{a, b, c\}$  and  $\{a, c\}$  have join  $\{a, b, c\} \vee \{a, c\} = \{a, b, c\}$  and meet  $\{a, b, c\} \wedge \{a, c\} = \{a, c\}$ ;
- $\{a, b\}$  and  $\{b, c\}$  have join  $\{a, b\} \vee \{b, c\} = \{a, b, c\}$  and meet  $\{a, b\} \wedge \{b, c\} = \{b\}$ ;
- $\{a\}$  and  $\{c\}$  have join  $\{a\} \vee \{c\} = \{a, c\}$  and meet  $\{a\} \wedge \{c\} = \emptyset$ .

In the poset defined by the power set of a non-empty set of variables ordered by inclusion, join and meet behave identically to union and intersection respectively.

In general a Hasse diagram graphically represents a finite posets where vertices correspond to elements of the poset where vertices appearing higher in the diagram appear later

in the partial order. Edges connect vertices to their maximal non-trivial join—or minimal non-trivial meet.

**Definition** (*lattice*). Let  $\mathbf{P}$  be a poset with partial order  $\leq$ . If every pair of elements  $a, b \in \mathbf{P}$  has a unique join  $a \vee b \in \mathbf{P}$  and meet  $a \wedge b \in \mathbf{P}$ , then  $\mathbf{P}$  is a lattice.

The poset ordered by inclusion in Figure 3.1 illustrates the concept of a lattice. Furthermore, any totally ordered set is a lattice. Let  $\mathbf{P}$  be a lattice with partial order  $\leq$ . We adopt the notation for ceiling and floor to denote the join and meet of a subset  $A \subseteq \mathbf{P}$  in a lattice:

$$\lceil A \rceil_{\leq} \equiv \bigvee_{a \in A} a \qquad \lfloor A \rfloor_{\leq} \equiv \bigwedge_{a \in A} a$$

If  $\leq$  is a total order, then these operations return the first and last elements of the set respectively. If the partial order is not specified, we adopt the order for the canonical poset. Figure 3.1 illustrates the concepts of ceiling and floor:

- if  $A = \{\{a\}, \{b\}, \{c\}\}$ , then  $\lceil A \rceil = \{a, b, c\}$  and  $\lfloor A \rfloor = \emptyset$ ;
- if  $A = \{\{a\}, \{a, b\}, \{a, b, c\}\}$ , then  $\lceil A \rceil = \{a, b, c\}$  and  $\lfloor A \rfloor = \{a\}$ .

### 3.2.1 Möbius Inversion

Two useful functions for analyzing a poset  $\mathbf{P}$  are the zeta function and the Möbius function. Let  $V$  be a non-empty set of variables and  $\mathbf{P} = \mathcal{P}(V)$  be a poset ordered by inclusion. The zeta function  $\zeta_{\mathbf{P}} : \mathbf{P} \times \mathbf{P} \rightarrow \{0, 1\}$  is defined as follows:

$$\zeta_{\mathbf{P}}(B, A) = \begin{cases} 0 & B \not\subseteq A; \\ 1 & B \subseteq A. \end{cases}$$

The Möbius function  $\mu_{\mathbf{P}} : \mathbf{P} \times \mathbf{P} \rightarrow \mathbb{Z}$  is defined as follows:

$$\mu_{\mathbf{P}}(B, A) = \begin{cases} 0 & B \not\subseteq A; \\ -1^{|A \setminus B|} & B \subseteq A. \end{cases}$$

These functions may be thought of as matrices because the posets we consider are finite. Abusing notation, we interpret  $\zeta_{\mathbf{P}}$  and  $\mu_{\mathbf{P}}$  as matrices where the first and second arguments

of these functions act as the row and column indices respectively. Under this interpretation, the Möbius function is the inverse of the zeta function in the sense that  $\mu_{\mathcal{P}} = \zeta_{\mathcal{P}}^{-1}$ .

$$\begin{array}{c}
 \zeta_{\mathcal{P}} \\
 \emptyset \\
 \{a\} \\
 \{b\} \\
 \{c\} \\
 \{a, b\} \\
 \{a, c\} \\
 \{b, c\} \\
 \{a, b, c\}
 \end{array}
 \begin{array}{c}
 \emptyset \quad \{a\} \quad \{b\} \quad \{c\} \quad \{a, b\} \quad \{a, c\} \quad \{b, c\} \quad \{a, b, c\} \\
 \left[ \begin{array}{cccccccc}
 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\
 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \\
 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\
 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\
 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\
 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
 \end{array} \right]
 \end{array}$$

Figure 3.2: The zeta function of a poset  $\mathcal{P} = \mathcal{P}(\{a, b, c\})$  ordered by inclusion—the first and second arguments of the zeta function act as row and column indices respectively.

Figure 3.2 depicts the zeta function  $\zeta_{\mathcal{P}}$  as a matrix for the poset  $\mathcal{P}$  depicted in Figure 3.1. Notice that the matrix is invertible—it is an upper triangular matrix with non-zero entries on the main diagonal. In general, the rows and columns of the matrix corresponding to the zeta function of a poset can be rearranged in this manner. Accordingly, the matrix corresponding to the zeta function is invertible.



$$\begin{array}{c}
\mu_{\mathbf{P}} \\
\emptyset \\
\{a\} \\
\{b\} \\
\{c\} \\
\{a, b\} \\
\{a, c\} \\
\{b, c\} \\
\{a, b, c\}
\end{array}
\begin{array}{c}
\emptyset \\
\{a\} \\
\{b\} \\
\{c\} \\
\{a, b\} \\
\{a, c\} \\
\{b, c\} \\
\{a, b, c\}
\end{array}
\begin{bmatrix}
1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 \\
0 & 1 & 0 & 0 & -1 & -1 & 0 & 1 \\
0 & 0 & 1 & 0 & -1 & 0 & -1 & 1 \\
0 & 0 & 0 & 1 & 0 & -1 & -1 & 1 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}$$

Figure 3.3: The Möbius function of a poset  $\mathbf{P} = \mathcal{P}(\{a, b, c\})$  ordered by inclusion—the first and second arguments of the Möbius function act as row and column indices respectively.

Figure 3.3 depicts the Möbius function  $\mu_{\mathbf{P}}$  as a matrix for the poset  $\mathbf{P}$  depicted in Figure 3.1. Again, notice that  $\mu_{\mathbf{P}}$  is invertible—it is an upper triangular matrix with non-zero entries on the main diagonal. We encourage the reader to check that the matrices depicted in Figures 3.2 and 3.3 are indeed inverses of each other. This relation holds in general and provides an intuition for the so called Möbius inversion. In what follows, we provide two Characterizations of the Möbius inversion—we will use both later in this document.

**Proposition 3.2.1** (Proposition 2 [71]). *Let  $\mathbf{P}$  be a poset and  $g : \mathbf{P} \rightarrow \mathbb{R}$  and  $h : \mathbf{P} \rightarrow \mathbb{R}$  be real-valued functions. The following expressions imply each other:*

- i.  $g(A) = \sum_{B \in \mathbf{P}} h(B) \mu_{\mathbf{P}}(B, A)$  for all  $A \in \mathbf{P}$ ;
- ii.  $h(A) = \sum_{B \in \mathbf{P}} g(B) \zeta_{\mathbf{P}}(B, A)$  for all  $A \in \mathbf{P}$ .

*Alternatively, if we abuse notation and treat  $g$  and  $h$  as column vectors, then the Möbius inversion states that  $g = \mu_{\mathbf{P}}h \Leftrightarrow h = \zeta_{\mathbf{P}}g$ . If  $V$  is a non-empty set of variables and  $\mathbf{P} = \mathcal{P}(V)$  is a poset ordered by inclusion, then the Möbius inversion simplifies to the following equivalent*

statements:

- i.  $g(A) = \sum_{B \subseteq A} (-1)^{|A \setminus B|} h(B)$  for all  $A \subseteq V$ ;
- ii.  $h(A) = \sum_{B \subseteq A} g(B)$  for all  $A \subseteq V$ .

**Corollary 3.2.1** (Corollary 1 [71]). *Let  $\mathbf{P}$  be a poset and  $g : \mathbf{P} \rightarrow \mathbb{R}$  and  $h : \mathbf{P} \rightarrow \mathbb{R}$  be real-valued functions. The following expressions imply each other:*

- i.  $g(A) = \sum_{B \in \mathbf{P}} \mu_{\mathbf{P}}(A, B) h(B)$  for all  $A \in \mathbf{P}$ ;
- ii.  $h(A) = \sum_{B \in \mathbf{P}} \zeta_{\mathbf{P}}(A, B) g(B)$  for all  $A \in \mathbf{P}$ .

Alternatively, if we abuse notation and view  $g$  and  $h$  as column vectors then the corollary states that  $g = \mu_{\mathbf{P}}^{\top} h \Leftrightarrow h = \zeta_{\mathbf{P}}^{\top} g$ . If  $V$  is a non-empty set of variables and  $\mathbf{P} = \mathcal{P}(V)$  is a poset ordered by inclusion, then the corollary simplifies to the following equivalent statements:

- i.  $g(A) = \sum_{B \subseteq V (A \subseteq B)} (-1)^{|B \setminus A|} h(B)$  for all  $A \subseteq V$ ;
- ii.  $h(A) = \sum_{B \subseteq V (A \subseteq B)} g(B)$  for all  $A \subseteq V$ .

| $T$           | $g(T)$ |                                    | $T$           | $h(T)$ |
|---------------|--------|------------------------------------|---------------|--------|
| $\emptyset$   | 0      |                                    | $\emptyset$   | 0      |
| $\{a\}$       | 0      | $\xrightarrow{\zeta_{\mathbf{P}}}$ | $\{a\}$       | 0      |
| $\{b\}$       | 0      |                                    | $\{b\}$       | 0      |
| $\{c\}$       | 1      |                                    | $\{c\}$       | 0      |
| $\{a, b\}$    | 0      |                                    | $\{a, b\}$    | 1      |
| $\{a, c\}$    | -1     |                                    | $\{a, c\}$    | 0      |
| $\{b, c\}$    | -1     | $\xleftarrow{\mu_{\mathbf{P}}}$    | $\{b, c\}$    | 0      |
| $\{a, b, c\}$ | 1      |                                    | $\{a, b, c\}$ | 1      |

Figure 3.4: An application of the zeta and Möbius functions of a poset  $\mathbf{P} = \mathcal{P}(\{a, b, c\})$  ordered by inclusion.

Figure 3.4 depicts an application of the Möbius inversion with respect to a poset  $\mathbf{P} = \mathcal{P}(\{a, b, c\})$  ordered by inclusion. If  $g : \mathbf{P} \rightarrow \mathbb{R}$  and  $h : \mathbf{P} \rightarrow \mathbb{R}$  are real-valued functions satisfying Proposition 3.2.1, then the zeta function depicted in Figure 3.2 applied to  $g$  results in  $h$  and the Möbius function depicted in Figure 3.3 applied to  $h$  results in  $g$ ; Figure 3.4 gives an example.

### 3.3 Ancestral Graphs

A common theme throughout this dissertation is the use of mixed graphs as independence models. This section introduces several families of mixed graphs, including directed acyclic graphs, acyclic directed mixed graphs, and maximal ancestral graphs.

#### 3.3.1 Preliminaries

**Definition** (*mixed graph*). A *mixed graph*  $\mathcal{G} = (V, E)$  is an ordered pair consisting of a vertex set and an edge set respectively. The edge set contains a mixture of directed, bi-directed, and undirected edges which connect pairs of vertices in the vertex set such that no pair of vertices is connected by more than one edge of the same type.

**Definition** (*characteristics of mixed graphs*). A few characteristics used to further refine the definition of a mixed graph are defined as follows:

- a mixed graph is *loopless* if no edge connects a vertex to itself;
- a mixed graph has *multiple edges* if more than one edge connects any pair of vertices;
- a mixed graph is *simple* if it is loopless and does not have multiple edges;
- a mixed graph is *directed* if it does not contain any undirected edges;
- a mixed graph is *acyclic* if it does not contain any *directed cycles*—a sequence of commonly oriented edges that starts and ends with the same vertex.

As a point of clarification, a *directed graph* is a mixed graph that only contains directed edges, whereas a directed mixed graph can additionally contain bi-directed edges.

**Definition** (*paths in mixed graphs*). Let  $\mathcal{G} = (V, E)$  be a mixed graph. The notion of a *path* and a few related concepts are defined as follows:

- a *path*  $\pi = \langle v_1, \dots, v_m \rangle$  is a sequence of  $m > 1$  distinct vertices where an edge connects  $v_i$  and  $v_{i+1}$  for all  $1 \leq i < m$ ;
- the *endpoints* of a path  $\pi = \langle v_1, \dots, v_m \rangle$  are the first and last vertices  $\{v_1, v_m\}$ ;
- a *triple* is a path  $\pi = \langle v_1, v_2, v_3 \rangle$  with three vertices and is *unshielded* if no edge connects its endpoints  $v_1$  and  $v_3$ ;
- a *collider* on  $\pi = \langle v_1, \dots, v_m \rangle$  ( $m \geq 3$ ) is a vertex  $v_i$  ( $1 < i < m$ ) such that:

$$v_{i-1} \left\{ \begin{array}{c} \rightarrow \\ \rightarrow \\ \leftrightarrow \end{array} \right\} v_i \left\{ \begin{array}{c} \leftarrow \\ \leftrightarrow \\ \leftrightarrow \end{array} \right\} v_{i+1}$$

and is *unshielded* if no edge connects  $v_{i-1}$  and  $v_{i+1}$ .

Paths are sometimes defined as sequences of distinct edges linked by shared endpoints, however, in this dissertation, the notion of a path is only considered within simple mixed graphs where the two definitions are equivalent.

A *directed acyclic graph* (DAG) is a simple directed graph that is acyclic. The family of DAGs is of primary importance because it is both a subfamily and a constructor of mixed graphs. Section 3.4.2 details how DAGs construct mixed graphs through a process called latent projection. The two most prevalent families of mixed graphs that can be constructed by a latent projection process are *acyclic directed mixed graphs* (ADMGs) and *maximal ancestral graph* (MAGs). ADMGs are relatively easy to understand syntactically, while MAGs are generally more convenient to work with theoretically. The families of ADMGs and directed MAGs are equivalent with respect to representing conditional independence statements. The family of directed MAGs is a subfamily of ADMGs so results on ADMGs apply to directed MAGs, but not the other way around. Accordingly, prior work on both families will be referenced throughout this dissertation, but MAGs will be the primary family of mixed graphs discussed.

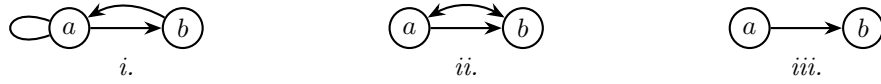


Figure 3.5: Mixed graphs with vertices  $\{a, b\}$ : (i) a mixed graph with a loop  $a - a$  and multiple edges  $a \xleftrightarrow{\leftarrow} b$ ; (ii) a acyclic directed loopless mixed graph with multiple edges  $a \xleftrightarrow{\leftarrow} b$ ; (iii) a simple acyclic directed graph.

Figure 3.5 illustrates several characteristics of mixed graphs. The mixed graphs in (ii) and (iii) are loopless and the mixed graph in (iii) is simple. Furthermore, the mixed graph in (i) contains a directed cycle  $a \rightarrow b \rightarrow a$ , the mixed graph in (ii) is an ADMG, and the mixed graph in (iii) is a DAG. Note that the multiple edges  $a \xleftrightarrow{\leftarrow} b$  and the bi-directed edge  $a \leftrightarrow b$  are not semantically equivalent. All families of mixed graphs discussed within this dissertation are loopless. Accordingly, from this point on, the terms mixed graph and loopless mixed graph will be used synonymously.

We utilize standard familial terminology from the vernacular of graphical models. Let  $\mathcal{G} = (V, E)$  be a mixed graph. For a vertex  $a \in V$ :

$$\text{pa}_{\mathcal{G}}(a) \equiv \{b ; b \rightarrow a \text{ in } \mathcal{G}\}$$

$$\text{ch}_{\mathcal{G}}(a) \equiv \{b ; b \leftarrow a \text{ in } \mathcal{G}\}$$

$$\text{sp}_{\mathcal{G}}(a) \equiv \{b ; b \leftrightarrow a \text{ in } \mathcal{G}\}$$

$$\text{neg}(a) \equiv \{b ; b - a \text{ in } \mathcal{G}\}$$

are the *parents*, *children*, *spouses*, and *neighbors* of  $a$  respectively. If any of the above edges are present in  $\mathcal{G}$ , then  $a$  and  $b$  are *adjacent*. Similarly:

$$\text{an}_{\mathcal{G}}(a) \equiv \{b ; b \rightarrow \cdots \rightarrow a \text{ in } \mathcal{G} \text{ or } a = b\}$$

$$\text{de}_{\mathcal{G}}(a) \equiv \{b ; b \leftarrow \cdots \leftarrow a \text{ in } \mathcal{G} \text{ or } a = b\}$$

$$\text{dis}_{\mathcal{G}}(a) \equiv \{b ; b \leftrightarrow \cdots \leftrightarrow a \text{ in } \mathcal{G} \text{ or } a = b\}$$

$$\text{ant}_{\mathcal{G}}(a) \equiv \{b ; \left. \begin{array}{l} b \rightarrow \cdots \rightarrow a \\ b - \cdots \rightarrow a \\ b - \cdots - a \end{array} \right\} \text{ in } \mathcal{G} \text{ or } a = b\}$$

are the *ancestors*, *descendants*, *district*, and *anterior vertices* of  $a$  respectively. These functions are applied to sets disjunctively, that is, applying one to a set of vertices is the union of the operation applied to each vertex in the set. For example, a set of vertices  $A \subseteq V$  has the following parents and ancestors:

$$\text{pa}_{\mathcal{G}}(A) \equiv \bigcup_{a \in A} \text{pa}_{\mathcal{G}}(a) \quad \text{an}_{\mathcal{G}}(A) \equiv \bigcup_{a \in A} \text{an}_{\mathcal{G}}(a).$$

We use inclusive definitions of these functions:  $a \in \text{an}_{\mathcal{G}}(A)$ ,  $a \in \text{deg}_{\mathcal{G}}(A)$ , and  $a \in \text{dis}_{\mathcal{G}}(A)$  for all  $a \in A$ . These operators are not always defined this way—we define them as such for theoretical convenience. Notably, the definitions for parents, children, spouses, and neighbors are not inclusive, however, having inclusive versions will be useful later. We define the inclusive versions of these functions as follows:  $\text{pa}_{\mathcal{G}}^+$ ,  $\text{ch}_{\mathcal{G}}^+$ ,  $\text{sp}_{\mathcal{G}}^+$ ,  $\text{ne}_{\mathcal{G}}^+$ .

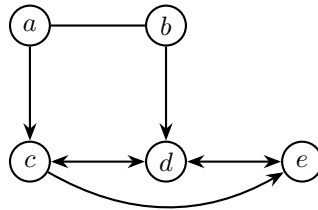


Figure 3.6: A mixed graph with vertices  $\{a, b, c, d, e\}$ .

Figure 3.6 illustrates concepts of parents, children, neighbors, and spouses. In the graph, the non-trivial relations are as follows:

- $c$  has parents  $\{a\}$ ,  $d$  has parents  $\{b\}$ , and  $e$  has parents  $\{c\}$ ;
- $a$  has children  $\{c\}$ ,  $b$  has children  $\{d\}$ , and  $c$  has children  $\{e\}$ ;
- $a$  has neighbors  $\{b\}$ , and  $b$  has neighbors  $\{a\}$ ;
- $c$  has spouses  $\{d\}$ ,  $d$  has spouses  $\{c, e\}$ , and  $e$  has spouses  $\{d\}$ .

Similarly, Figure 3.6 illustrates concepts of ancestors, descendants, and districts. In the graph, the non-trivial relations are as follows:

- $c$  has ancestors  $\{a, c\}$ ,  $d$  has ancestors  $\{b, d\}$ , and  $e$  has ancestors  $\{a, c, e\}$ ;
- $a$  has descendants  $\{a, c, e\}$ ,  $b$  has descendants  $\{b, d\}$ , and  $c$  has descendants  $\{c, e\}$ ;
- $a$  has anterior vertices  $\{a, b\}$ ,  $b$  has anterior vertices  $\{a, b\}$ ,  $c$  has anterior vertices  $\{a, b, c\}$ ,  $d$  has anterior vertices  $\{a, b, d\}$ , and  $e$  has anterior vertices  $\{a, b, c, e\}$ ;
- $\{a\}$ ,  $\{b\}$ , and  $\{c, d, e\}$  form districts.

We now have a sufficient set of graphical concepts to define the ancestral graphs and are one step closer to defining MAGs.

**Definition** (*ancestral graph*). Let  $\mathcal{G} = (V, E)$  be a simple mixed graph.  $\mathcal{G}$  is *ancestral* if for all vertices  $a \in V$ :

- i.*  $\text{ch}_{\mathcal{G}}(a) \cap \text{an}_{\mathcal{G}}(a) = \emptyset$ ;
- ii.*  $\text{sp}_{\mathcal{G}}(a) \cap \text{an}_{\mathcal{G}}(a) = \emptyset$ ;
- iii.*  $\text{pa}_{\mathcal{G}}(a) \cup \text{sp}_{\mathcal{G}}(a) \neq \emptyset \Rightarrow \text{ne}_{\mathcal{G}}(a) = \emptyset$ .

Criteria (*i*) states that ancestral graphs cannot have directed cycles and criteria (*ii*) states that ancestral graphs cannot have *almost-directed cycles*—a sequence of commonly oriented edges that starts and ends with vertices connected by a bi-directed edge. Criteria (*iii*) states that ancestral graphs cannot have a directed arrowhead pointed into a vertex that is connected to another vertex with an undirected edge. Accordingly, ancestral graphs have clearly defined directed and undirected parts. This notion can be made rigorous using the graphical concept of a subgraph.

Two important graphical concepts used throughout this dissertation are anterior and ancestral sets.

**Definition** (*anterior set*). Let  $\mathcal{G} = (V, E)$  be a mixed graph containing a set  $A \subseteq V$ .  $A$  is *anterior* if  $\text{ant}_{\mathcal{G}}(A) = A$ , in other words,  $A$  contains all its own anterior vertices. The set of all anterior sets in  $\mathcal{G}$  is denoted by  $\mathcal{A}(\mathcal{G})$ .

**Definition** (*ancestral set*). Let  $\mathcal{G} = (V, E)$  be a mixed graph containing a set  $A \subseteq V$ .  $A$  is *ancestral* if  $\text{an}_{\mathcal{G}}(A) = A$ , in other words,  $A$  contains all its own ancestors. Notably, if  $\mathcal{G}$  is

directed, then  $\mathcal{A}(\mathcal{G})$  is the set of all ancestral sets in  $\mathcal{G}$ .

**Definition** (*subgraph of mixed graphs*). Let  $\mathcal{G} = (V, E)$  and  $\mathcal{G}' = (V', E')$  be mixed graphs. If  $V' \subseteq V$  and  $E' \subseteq E$ , then  $\mathcal{G}'$  is a *subgraph* of  $\mathcal{G}$ —of particular interest:

- the *directed subgraph* of  $\mathcal{G}$ , denoted by  $\text{dir}(\mathcal{G}) = (V', E')$  where  $V' = \{a \in V ; \text{pa}_{\mathcal{G}}(a) \cup \text{ch}_{\mathcal{G}}(a) \cup \text{sp}_{\mathcal{G}}(a) \neq \emptyset\}$  and  $E' = \{e \in E ; e \text{ is a directed or bi-directed edge}\}$
- the *undirected subgraph* of  $\mathcal{G}$ , denoted by  $\text{un}(\mathcal{G}) = (V', E')$  where  $V' = \{a \in V ; \text{pa}_{\mathcal{G}}(a) \cup \text{sp}_{\mathcal{G}}(a) = \emptyset\}$  and  $E' = \{e \in E ; e \text{ is an undirected edge}\}$ ;
- the *induced subgraph* of  $\mathcal{G}$  over  $A \subseteq V$ , denoted by  $\mathcal{G}_A = (A, E')$  where  $E' = \{e \in E ; e \text{ connects two members of } A\}$ .

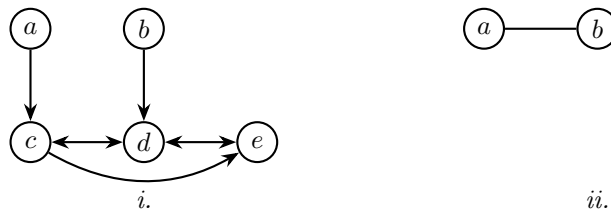


Figure 3.7: Subgraphs of the graph in Figure 3.6: (i) the directed subgraph; (ii) the undirected subgraph.

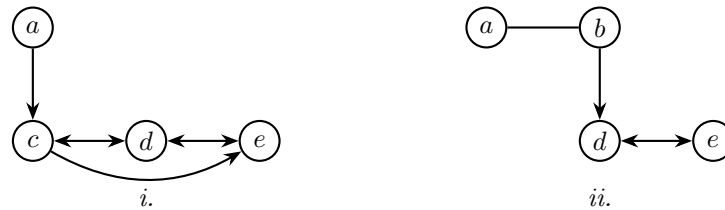


Figure 3.8: Induced subgraphs of the graph in Figure 3.6: (i) the induced subgraph over  $\{a, c, d, e\}$ ; (ii) the induced subgraph over  $\{a, b, d, e\}$ .



**Proposition 3.3.1** (Proposition 3.5 [70]). *Let  $\mathcal{G}$  be an ancestral graph. If  $\mathcal{G}'$  is a subgraph of  $\mathcal{G}$ , then  $\mathcal{G}'$  is an ancestral graph.*

As noted earlier, DAGs are not stable under marginalization and conditioning, however, ancestral graphs are stable under marginalization and conditioning. For any DAG with latent confounding and selection effects, there is an ancestral graph over the measured variables alone that represents the conditional independence and ancestral relations entailed by the original DAG; in the case of a causal DAG, the ancestral relations are causal. The edges of a causal ancestral graph may be interpreted causally as follows:

- $a \rightarrow b$  means that  $a$  is a cause of  $b$  or some selection variable, but  $b$  is not a cause of  $a$  or any selection variable;
- $a \leftrightarrow b$  means that  $a$  is not a cause of  $b$  or any selection variable, and  $b$  is not a cause of  $a$  or any selection variable;
- $a - b$  means that  $a$  is a cause of  $b$  or some selection variable, and  $b$  is a cause of  $a$  or some selection variable.

### 3.3.2 Graphical Conditional Independence

Graphical separation criteria define the notion of graphical conditional independence. In this dissertation, we use the so called  $m$ -separation criterion for mixed graphs, which naturally extends the well known  $d$ -separation criterion for directed graphs [58, 70, 74]. Given a DAG with latent confounding and selection effects, inducing paths characterize when two vertices cannot be not graphically separated conditioned on any set of vertices that corresponds to a set of measured variables. Throughout this dissertation, the symbols  $L$  and  $S$  denote sets of latent confounding and selection effects (and their corresponding vertices) respectively.

**Definition** (*inducing path*). Let  $\mathcal{G} = (V, E)$  be an ancestral graph containing vertices  $a, b \in V$  ( $a \neq b$ ) and disjoint sets  $L, S \subseteq V \setminus \{a, b\}$ . A path  $\pi$  between  $a$  and  $b$  is *inducing* relative to  $\langle L, S \rangle$  if the following hold:

- i.* every non-endpoint on  $\pi$  is a member of  $L$  or a collider;
- ii.* every collider on  $\pi$  is an ancestor of  $a, b$ , or  $s \in S$ .

If  $L = S = \emptyset$ , then  $\pi$  is a *primitive inducing path*.

Looking ahead, Figure 3.10 gives an example of a primitive inducing path. The path  $\langle a, c, d, b \rangle$  is primitively inducing in both graphs, but  $a$  and  $b$  are only adjacent in (ii). In the literature, inducing paths have only been defined for ancestral graphs, but it is likely the case that they can be extended to all families of mixed graphs discussed in section 3.4.

In Section 3.4, we review how a DAG with latent confounding and selection effects may be represented as a loopless mixed graph derived by the marginalization and conditioning of that DAG. In the case of a loopless mixed graph, graphical conditional independence is characterized by  $m$ -connecting paths and  $m$ -separation.

**Definition** ( *$m$ -connecting path*). Let  $\mathcal{G} = (V, E)$  be a mixed graph containing vertices  $a, b \in V$  ( $a \neq b$ ) and a subset  $C \subseteq V \setminus \{a, b\}$ . A path  $\pi$  between  $a$  and  $b$  is  *$m$ -connecting* relative to  $C$  if the following hold:

- i.* every non-collider on  $\pi$  is not a member of  $C$ ;
- ii.* every collider on  $\pi$  is an ancestor of  $a$ ,  $b$ , or  $c \in C$ .

**Definition** ( *$m$ -separation*). Let  $\mathcal{G} = (V, E)$  be an mixed graph containing disjoint sets  $A, B, C \subseteq V$ . If for every  $a \in A$  and  $b \in B$  no  $m$ -connecting path exists between  $a$  and  $b$  relative to  $C$ , then  $A$  and  $B$  are  *$m$ -separated* by  $C$ .

Let  $\mathcal{G} = (V, E)$  be a mixed graph containing disjoint sets  $A, B, C \subseteq V$ . We say  $\langle A, B \mid C \rangle$  is represented in  $\mathcal{G}$  by  $m$ separation and write  $A \perp\!\!\!\perp_m B \mid C [\mathcal{G}]$  if  $A$  and  $B$  are  $m$ -separated by  $C$ . The independence model induced by  $\mathcal{G}$  is denoted  $\mathcal{J}_m(\mathcal{G})$ .

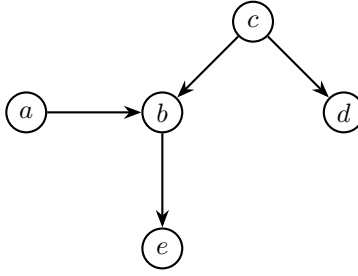


Figure 3.9: An ancestral graph with vertices  $\{a, b, c, d, e\}$ .

Figure 3.9 illustrates the concepts of inducing paths,  $m$ -connecting paths, and  $m$ -separation.

In the graph:

- $\langle a, b, c, d \rangle$  is an inducing path relative to  $\langle \{c\}, \{b\} \rangle$  and relative to  $\langle \{c\}, \{e\} \rangle$ ;
- $\langle a, b, c, d \rangle$  is an  $m$ -connecting path relative to  $\{b\}$  and relative to  $\{e\}$ ;
- $\langle a, b, c, d \rangle$  is not  $m$ -connecting relative to  $\{c\}$ ,  $\{b, c\}$ , or  $\{c, e\}$  because  $a$  and  $d$  are  $m$ -separated by  $\{c\}$ ,  $\{b, c\}$ , and  $\{c, e\}$  respectively.

Additionally,  $m$ -connecting and inducing paths in ancestral graphs are related by the following proposition.

**Proposition 3.3.2** (Theorem 4.2 [70]). *Let  $\mathcal{G} = (V, E)$  be an ancestral graph containing vertices  $a, b \in V$  ( $a \neq b$ ) and disjoint sets  $L, S \subset V \setminus \{a, b\}$ . The following are equivalent:*

- i. there exists an inducing path between  $a$  and  $b$  relative to  $\langle L, S \rangle$ ;*
- ii.  $a$  and  $b$  are not  $m$ -separated by  $C$  for all  $S \subset C \subseteq V \setminus L$  ( $a, b \notin C$ ).*

Occasionally, it is useful to have an alternative separation criterion for the simplification of proofs. Accordingly, we define the augmented graph and  $m^*$ -separation criterion for ancestral graphs.

**Definition** (*collider-connecting path*). Let  $\mathcal{G} = (V, E)$  be a mixed graph containing vertices  $a, b \in V$ . A path  $\pi$  between  $a$  and  $b$  is a *collider-connecting path* if every non-endpoint vertex on  $\pi$  is a collider.

Let  $\mathcal{G} = (V, E)$  be a mixed graph containing a vertex  $a \in V$ . The non-trivial collider-connecting vertices of  $a$  are the vertices connected to  $a$  by collider-connecting paths. Let  $\mathcal{G} = (V, E)$  be a mixed graph. For a vertex  $a \in V$ ,

$$\text{col}_{\mathcal{G}}(a) \equiv \text{ne}_{\mathcal{G}}(a) \cup \text{pa}_{\mathcal{G}}^+(\text{dis}_{\mathcal{G}}(\text{ch}_{\mathcal{G}}^+(a)))$$

are the *collider-connecting* vertices of  $a$ . We define this function to be conjunctive when applied to sets, that is, by definition applying the collider-connecting function to a set of vertices is the intersection of the operation applied to each vertex in the set. For example, a set of vertices  $A \subseteq V$  has collider-connecting vertices:

$$\text{col}_{\mathcal{G}}(A) \equiv \bigcap_{a \in A} \text{col}_{\mathcal{G}}(a).$$

**Definition** (*augmented graph*). Let  $\mathcal{G} = (V, E)$  be a mixed graph. The augmented graph, denoted  $\mathcal{G}' = \text{aug}(\mathcal{G})$ , is the undirected graph over the same vertices such that  $\text{ne}_{\mathcal{G}'}(a) = \text{col}_{\mathcal{G}}(a)$  for all  $a \in V$ .

**Definition** ( *$m^*$ -separation*). Let  $\mathcal{G} = (V, E)$  be an ancestral graph containing disjoint sets  $A, B, C \subseteq V$  and  $D = \text{ant}_{\mathcal{G}}(ABC)$ . If for every  $a \in A$  and  $b \in B$  no  $m$ -connecting path exists between  $a$  and  $b$  relative to  $C$  in  $\text{aug}(\mathcal{G}_D)$ , then  $A$  and  $B$  are  *$m^*$ -separated* by  $C$  in  $\mathcal{G}$ .

Let  $\mathcal{G} = (V, E)$  be a mixed graph containing disjoint sets  $A, B, C \subseteq V$ . We say  $\langle A, B \mid C \rangle$  is represented in  $\mathcal{G}$  by  $m^*$ -separation and write  $A \perp\!\!\!\perp_{m^*} B \mid C [\mathcal{G}]$  if  $A$  and  $B$  are  $m^*$ -separated by  $C$ . The independence model induced by  $\mathcal{G}$  is denoted  $\mathcal{J}_{m^*}(\mathcal{G})$ .

**Theorem 3.3.1** (Theorem 3.18 [70]). *If  $\mathcal{G}$  is an ancestral graph, then  $\mathcal{J}_{m^*}(\mathcal{G}) = \mathcal{J}_m(\mathcal{G})$ .*

Since the two separation criterion are equivalent we drop the identifying subscript in the relevant notation. The following corollary is a direct consequence of the equivalence of  $m^*$ -separation and  $m$ -separation.

**Corollary 3.3.1.** *If  $\mathcal{G}$  is an ancestral graph, then  $\mathcal{J}(\mathcal{G}) = \bigcup_{A \in \mathcal{A}(\mathcal{G})} \mathcal{J}(\text{aug}(\mathcal{G}_A))$ .*

*Proof.* This directly follows from the definition of  $m^*$ -separation and Theorem 3.3.1.  $\square$

Lastly, we note that an induced independence model defined by a mixed graph and  $m$ -separation, including ancestral graphs, is compositional graphoid.

**Proposition 3.3.3** (Theorem 1 [74]). *If  $\mathcal{G}$  is a mixed graph, then the induced independence model  $\mathcal{J}(\mathcal{G})$  is a compositional graphoid.*

### 3.3.3 Markov Properties

Formally, ancestral graph Markov models are characterized by the  $m$ -separation criterion in conjunction with the global Markov property.

**Definition** (*global Markov property*). Let  $\mathcal{G} = (V, E)$  be a mixed graph and  $P$  be a probability measure over  $V$ .  $P$  satisfies the *global Markov property* for  $\mathcal{G}$  if the following holds for all disjoint triples  $\langle A, B \mid C \rangle \in \mathcal{T}(V)$ :

$$A \perp\!\!\!\perp B \mid C [\mathcal{G}] \quad \Rightarrow \quad A \perp\!\!\!\perp B \mid C [P].$$

Alternatively,  $P$  satisfies the *global Markov property* for  $\mathcal{G}$  if:

$$\mathcal{J}(\mathcal{G}) \subseteq \mathcal{J}(P).$$

However, it is often the case that many of the conditional independence statements characterized by the global Markov property are redundant—implied by the semi-graphoid axiom and other conditional independence statements. Accordingly, for many graphical families, the global Markov property is often reduced to simpler Markov properties, such as the ordered local Markov property for ADMGs. In what follows, we introduce concepts needed to define the ordered local Markov property.

**Definition** (*collider-connecting set*). Let  $\mathcal{G} = (V, E)$  be a mixed graph containing a set  $A \subseteq V$ .  $A$  is *collider-connecting* if  $A \subseteq \text{col}_{\mathcal{G}}(A)$ . That is, there exists a collider path between  $a$  and  $b$  for all  $a, b \in A$  ( $a \neq b$ ).

Let  $\mathcal{G} = (V, E)$  be a mixed graph containing a vertex  $b \in V$ . The set of collider-connecting vertices for  $b$  has special property:

$$b \perp\!\!\!\perp a \mid \text{col}_{\mathcal{G}}(b) \setminus b [\mathcal{G}] \quad \text{for all } a \in V \setminus \text{col}_{\mathcal{G}}(b)$$

That is  $\text{col}_{\mathcal{G}}(b) \setminus b$  is the set that renders  $b$  independent of all other vertices in the

graph. In many cases, this special property is what allows simplified Markov properties to be constructed. In general this set is called a Markov blanket and the set consisting of a vertex and its Markov blanket is called a closure. Accordingly, the Markov blanket and closure for ADMGs are defined as follows:

$$\text{mb}_{\mathcal{G}}(b) \equiv \text{col}_{\mathcal{G}}(b) \setminus b \quad \text{cl}_{\mathcal{G}}(b) \equiv \text{col}_{\mathcal{G}}(b)$$

The global Markov property can also be simplified by using the concept of a consistent order.

**Definition** (*consistent order*). Let  $\mathcal{G} = (V, E)$  be an ADMG. A total order  $\leq$  over  $V$  is *consistent* with  $\mathcal{G}$  if:

$$a \leq b \quad \Rightarrow \quad b \notin \text{an}_{\mathcal{G}}(a) \setminus a \quad \text{for all } a, b \in V.$$

**Definition** (*preceding vertices*). Let  $\mathcal{G} = (V, E)$  be an ADMG containing a vertex  $b \in V$  and  $\leq$  be a total order consistent with  $\mathcal{G}$ . The preceding vertices of  $b$  with respect to  $\leq$  are defined as follows:

$$\text{pre}_{\mathcal{G}}^{\leq}(b) \equiv \{a \in V ; \quad a \leq b\}.$$

The concepts of a Markov blanket and a closure can be redefined with respect to a consistent order which directly leads to the ordered local Markov property.

$$\text{mb}_{\mathcal{G}}^{\leq}(b) \equiv \text{mb}_{\mathcal{G}}(b) \cap \text{pre}_{\mathcal{G}}^{\leq}(b) \quad \text{cl}_{\mathcal{G}}^{\leq}(b) \equiv \text{cl}_{\mathcal{G}}(b) \cap \text{pre}_{\mathcal{G}}^{\leq}(b)$$

**Definition** (*ordered local Markov property*). Let  $\mathcal{G} = (V, E)$  be an ADMG,  $\leq$  be a total order consistent with  $\mathcal{G}$ , and  $P$  be a probability measure over  $V$ . If for every vertex  $b \in V$  and ancestral set  $A \in \mathcal{A}(\mathcal{G})$  where  $b \in A \subseteq \text{pre}_{\mathcal{G}}^{\leq}(b)$ :

$$b \perp\!\!\!\perp A \setminus \text{cl}_{\mathcal{G}_A}^{\leq}(b) \mid \text{mb}_{\mathcal{G}_A}^{\leq}(b) [P]$$

then  $P$  satisfies the *ordered local Markov property* for  $\mathcal{G}$  with respect to  $\leq$ .

**Theorem 3.3.2** (Theorem 2 [66]). *Let  $\mathcal{G} = (V, E)$  be an ADMG,  $\leq$  be a total order consistent with  $\mathcal{G}$ , and  $P$  be a probability measure over  $V$ . The following are equivalent:*

- i.  $P$  satisfies the global Markov property for  $\mathcal{G}$ ;
- ii.  $P$  satisfies the ordered local Markov property for  $\mathcal{G}$  with respect to  $\leq$ .

Lastly we introduce the augmented pairwise Markov property for ancestral graphs. This criterion extends the pairwise Markov property for undirected graphs using graph augmentation; see Lauritzen [46] for more details.

**Definition** (*augmented pairwise Markov property*). Let  $\mathcal{G} = (V, E)$  be an ancestral graph and  $P$  be a probability measure over  $V$ . If for every anterior set  $A \in \mathcal{A}(\mathcal{G})$  and pair of vertices  $a, b \in A$  where  $a \notin \text{ne}_{\text{aug}(\mathcal{G}_A)}^+(b)$ :

$$a \perp\!\!\!\perp b \mid A \setminus \{a, b\} [P]$$

then  $P$  satisfies the *augmented pairwise Markov property* for  $\mathcal{G}$ .

Richardson and Spirtes introduced the pairwise Markov property for MAGs which also extends the pairwise Markov property for undirected graphs [70]. Sadeghi showed that their pairwise Markov property is equivalent to the global Markov property for compositional graphoids [74]. We show that the augmented pairwise Markov property is equivalent to the global Markov property for graphoids using a classic result for undirected graphs.

**Theorem 3.3.3** (Theorem 1 [60]). *Let  $\mathcal{G} = (V, E)$  be an undirected graph and  $P$  be a probability measure over  $V$ . If  $\mathcal{J}(P)$  is a graphoid, then the following are equivalent:*

- i.  $a \perp\!\!\!\perp b \mid V \setminus \{a, b\} [P]$  for all  $a, b \in V$  ( $a \notin \text{ne}_{\mathcal{G}}^+(b)$ );
- ii.  $A \perp\!\!\!\perp B \mid C [P]$  for all  $\langle A, B \mid C \rangle \in \mathcal{J}(\mathcal{G})$ .

**Theorem 3.3.4.** *Let  $\mathcal{G} = (V, E)$  be an ancestral graph and  $P$  be a probability measure over  $V$ . If  $\mathcal{J}(P)$  is a graphoid, then the following are equivalent:*

- i.  $P$  satisfies the global Markov property for  $\mathcal{G}$ ;
- ii.  $P$  satisfies the augmented pairwise Markov property.

*Proof.* ( $i \Rightarrow ii$ ): Let  $A \in \mathcal{A}(\mathcal{G})$  be an anterior set and  $a, b \in A$  ( $a \neq b$ ). By Corollary 3.3.1

and the antecedent:

$$\begin{aligned} a \perp\!\!\!\perp b \mid A \setminus \{a, b\} [\text{aug}(\mathcal{G}_A)] &\Rightarrow a \perp\!\!\!\perp b \mid A \setminus \{a, b\} [\mathcal{G}] \\ &\Rightarrow a \perp\!\!\!\perp b \mid A \setminus \{a, b\} [P]. \end{aligned}$$

( $i \Leftarrow ii$ ): Let  $A, B, C \subseteq V$  be disjoint sets and  $D = \text{ant}_{\mathcal{G}}(ABC)$ . By the antecedent:

$$a \perp\!\!\!\perp b \mid D \setminus \{a, b\} [\text{aug}(\mathcal{G}_D)] \Rightarrow a \perp\!\!\!\perp b \mid D \setminus \{a, b\} [P] \quad \text{for all } a, b \in D (a \neq b).$$

Accordingly, by Corollary 3.3.1 and Theorem 3.3.3:

$$\begin{aligned} A \perp\!\!\!\perp B \mid C [\mathcal{G}] &\Rightarrow A \perp\!\!\!\perp B \mid C [\text{aug}(\mathcal{G}_D)] \\ &\Rightarrow A \perp\!\!\!\perp B \mid C [P]. \end{aligned}$$

□

### 3.3.4 Maximality

**Definition** (*maximal*). Let  $\mathcal{G} = (V, E)$  be an ancestral graph.  $\mathcal{G}$  is *maximal* if for all  $a, b \in V$  ( $a \neq b$ ) the following are equivalent:

- i.*  $a$  and  $b$  are adjacent;
- ii.* there exists a primitive inducing path between  $a$  and  $b$ ;
- iii.*  $a$  and  $b$  are not  $m$ -separated by  $C$  for all  $C \subseteq V \setminus \{a, b\}$ .

Proposition 3.3.2 implies that (*ii*) and (*iii*) are equivalent; they are included here to provide alternative definitions of maximal.

A maximal ancestral graph (MAG) is an ancestral graph that is maximal. MAGs are maximal in the sense that no additional edges can be added to the graph without changing the independence model. Furthermore, any non-maximal ancestral graph can be made maximal by adding bi-directed edges. Intuitively, the definition of maximality for ancestral graphs in (*iii*) may be applied to other families of mixed graphs which utilize  $m$ -separation.



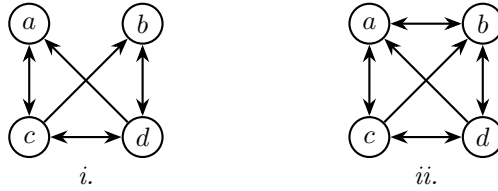


Figure 3.10: Ancestral graphs with vertices  $\{a, b, c, d\}$ : (i) a non-maximal ancestral graph; (ii) a maximal ancestral graph.

Figure 3.10 illustrates the concept of maximality. The ancestral graph in (i) depicts a graph that is not maximal and the ancestral graph in (ii) depicts a graph that is maximal. In general, the presence of a bi-directed edge in a MAG corresponds to one or more latent confounders on a path between the endpoints of the bi-directed edge. However, it does not necessarily mean that there is a latent confounder between the endpoints. For example, in (i) and (ii) there could be a latent confounder between  $a$  and  $c$ . In this case the bi-directed edge between  $a$  and  $b$  in (ii) could be induced exclusively by the confounded path between  $a$  and  $b$  mediated by  $c$ .

**Theorem 3.3.5** (Theorem 5.1 [70]). *Let  $\mathcal{G} = (V, E)$  be an ancestral graph. Then there exists a unique maximal ancestral graph formed by adding bi-directed edges to  $\mathcal{G}$  such that the independence model does not change.*

Accordingly, every DAG is maximal and the family of DAGs is a subset of the family of MAGs. Additionally, transforming an ancestral graph into a MAG does not affect the ancestral relations—only bi-directed edges are added. In Chapter 4 we work with MAGs rather than ancestral graphs to develop the theory in this dissertation because they are theoretically simpler and retain the statistical and causal properties of the corresponding ancestral graphs.

**Proposition 3.3.4.** *Let  $\mathcal{G} = (V, E)$  be a MAG:*

- *the directed subgraph  $\text{dir}(\mathcal{G})$  is a MAG;*

- the undirected subgraph  $\text{un}(\mathcal{G})$  is a MAG;
- the induced subgraph  $\mathcal{G}_A$  is a MAG for all anterior sets  $A \in \mathcal{A}(\mathcal{G})$ .

*Proof.* By Proposition 3.3.1, subgraphs of  $\mathcal{G}$  are ancestral. The proposition is proven by first showing that  $\mathcal{G}_A$  is maximal and then noting that  $\text{dir}(\mathcal{G})$  and  $\text{un}(\mathcal{G})$  are induced subgraphs of MAGs.

Suppose there is a primitive inducing path  $\pi$  in  $\mathcal{G}_A$  such that the endpoint are not adjacent. By the definition of induced subgraph, the endpoint are also not adjacent in  $\mathcal{G}$ . Furthermore, since any path in  $\mathcal{G}_A$  exists in  $\mathcal{G}$ ,  $\pi$  is also a primitive inducing in  $\mathcal{G}$ . This is a contradiction because  $\mathcal{G}$  is maximal. Accordingly,  $\mathcal{G}_A$  is maximal.

In the case of the directed subgraph  $\text{dir}(\mathcal{G})$ , consider the subgraph of  $\mathcal{G}$  where the undirected edges have been removed  $\mathcal{G}' = (V, E')$ . Notably,  $\text{dir}(\mathcal{G})$  is an induced subgraph of  $\mathcal{G}'$ . Suppose there is a primitive inducing path  $\pi$  in  $\mathcal{G}'$  such that the endpoint are not adjacent. By the definition of primitive inducing path, every non-endpoint on  $\pi$  is a collider. Furthermore, since removing an undirected edge can only destroy non-colliders,  $\pi$  is also primitively inducing in  $\mathcal{G}$ . This is a contradiction because  $\mathcal{G}$  is maximal. Accordingly,  $\mathcal{G}'$  are  $\text{dir}(\mathcal{G})$  are maximal. In the case of the undirected subgraph  $\text{un}(\mathcal{G})$ ,  $\text{un}(\mathcal{G})$  is an induced subgraph of  $\mathcal{G}$ . Accordingly,  $\text{un}(\mathcal{G})$  is maximal.  $\square$

### 3.3.5 Factorization

For a probability measure  $P$ , the global Markov property implies that the conditional independence statements represented in a graph are represented in  $P$ . Equivalently, some graphical families admit well-known factorizations that algebraically imply that the conditional independence statements represented in a graph are represented in  $P$ . For instance, DAGs provide a well known recursive factorization.

Let  $\mathcal{G} = (V, E)$  be a DAG. Furthermore, let  $X$  be a collection of random variables indexed by  $V$  with probability measure  $P$  that admits density  $f(x)$  with respect to dominating  $\sigma$ -finite product measure  $\nu$ .  $P$  satisfies the global Markov property with respect to  $\mathcal{G}$  if and only if

$$f(x) = \prod_{v \in V} f_{v|\text{pa}_{\mathcal{G}}(v)}(x) \quad \text{for } \nu\text{-a.e. } x \in \mathcal{X}.$$

A similar factorization was developed by Evans and Richardson for ADMGs [30, 67]. However, the factorization developed by Evans and Richardson requires multiple equations. In Chapter 4 we develop an alternative to Evans and Richardson’s factorization that only requires a single equation.

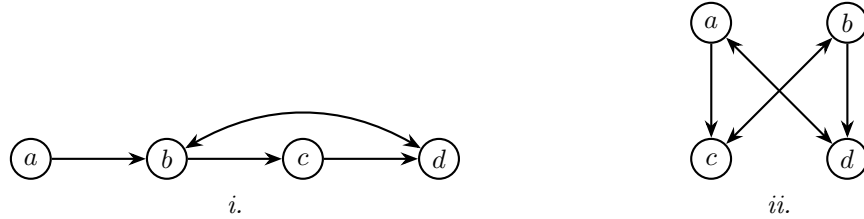


Figure 3.11: ADMGs with vertices  $\{a, b, c, d\}$ .

In order to state Richardson’s factorization criterion for ADMGs, we must first introduce a few additional concepts. Let  $\mathcal{G} = (V, E)$  be a mixed graph. For a vertex  $a \in V$  and a subset  $B \subseteq V$

$$\text{bar}_{\mathcal{G}}(B) \equiv \{b \in B ; B \cap \text{deg}_{\mathcal{G}}(b) = b\}$$

is the *barren subset* of  $B$ .

**Definition** (*barren set*). Let  $\mathcal{G} = (V, E)$  be an ADMG containing a set  $B \subseteq V$ .  $B$  is barren if  $B = \text{bar}_{\mathcal{G}}(B)$ . That is,  $B$  is barren if it does not contain any non-trivial descendants.

Richardson’s factorization criterion for ADMGs utilizes a partition function that partitions the variables into sets called heads. The factorization criterion is a product over conditional density terms comprised of heads conditioned on their corresponding tails.

**Definition** (*head*). Let  $\mathcal{G} = (V, E)$  be an ADMG containing a set  $H \subseteq V$  ( $H \neq \emptyset$ ).  $H$  is a head if it is barren in  $\mathcal{G}$  and contained within a single district of  $\mathcal{G}_{\text{ang}(H)}$ . The set of all heads in  $\mathcal{G}$  is denoted by  $\mathcal{H}(\mathcal{G})$ .

**Definition** (*tail*). Let  $\mathcal{G} = (V, E)$  be an ADMG. For a head  $H \in \mathcal{H}(\mathcal{G})$ , the tail of  $H$  is the set

$$\text{tail}_{\mathcal{G}}(H) \equiv T \setminus H \cup \text{pa}_{\mathcal{G}}(T) \quad \text{where } T = \text{dis}_{\mathcal{G}_{\text{ang}(H)}}(H).$$

Let  $\mathcal{G} = (V, E)$  be an ADMG and  $\leq$  be the partial order

$$H \leq H' \Leftrightarrow H \subseteq \text{an}_{\mathcal{G}}(H') \quad \text{for all } H, H' \in \mathcal{H}(\mathcal{G}).$$

Heads partition the variables with the help of two functions:  $\Pi_{\mathcal{G}} : \mathcal{P}(V) \rightarrow \mathcal{P}(\mathcal{H}(\mathcal{G}))$  which is such that  $\Pi_{\mathcal{G}}(A)$  returns the set of heads that are subsets of  $A$  and maximal with respect to  $\leq$ ; and  $\Psi_{\mathcal{G}} : \mathcal{P}(V) \rightarrow \mathcal{P}(V)$  which is such that  $\Psi_{\mathcal{G}}(A)$  returns the elements of  $A$  which are not contained in a set in  $\Pi_{\mathcal{G}}(A)$ :

$$\begin{aligned} \Pi_{\mathcal{G}}(A) &\equiv \{H \in \mathcal{H}(\mathcal{G}) ; \quad H \subseteq A \text{ and } H \not\leq H' \text{ for all } H' \subseteq A (H \neq H')\}; \\ \Psi_{\mathcal{G}}(A) &\equiv A \setminus \bigcup_{B \in \Pi_{\mathcal{G}}(A)} B. \end{aligned}$$

For a subset  $A \subseteq V$ , recursively define the partition function:

$$[A]_{\mathcal{G}} \equiv \begin{cases} \emptyset & A = \emptyset; \\ \Pi_{\mathcal{G}}(A) \cup [\Psi(A)]_{\mathcal{G}} & A \neq \emptyset, \end{cases}$$

where square brackets denote the partition function. The partition function removes maximal sets from  $A$ , and is recursively applied again to what remains.

**Theorem 3.3.6** (Theorem 4.12 [30]). *Let  $\mathcal{G} = (V, E)$  be an ADMG. Furthermore, let  $X$  be a collection of random variables indexed by  $V$  with probability measure  $P$  that admits density  $f(x)$  with respect to dominating  $\sigma$ -finite product measure  $\nu$ .  $P$  satisfies the global Markov property with respect to  $\mathcal{G}$  if and only if for every ancestral set  $A \in \mathcal{A}(\mathcal{G})$ ,*

$$f_A(x) = \prod_{H \in [A]_{\mathcal{G}}} f_{H|\text{tail}_{\mathcal{G}}(H)}(x) \quad \text{for } \nu\text{-a.e. } x \in \mathcal{X}.$$

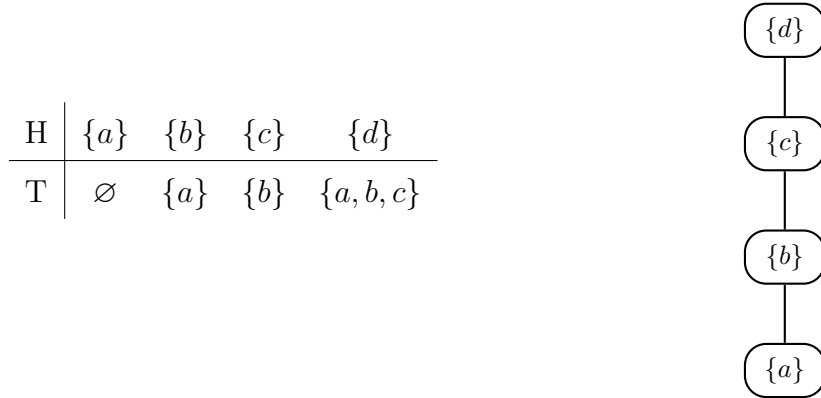


Figure 3.12: The heads and tails for the ADMG illustrated in Figure 3.11 (i) and the Hasse diagram for the corresponding poset over the ADMG's heads.

Figure 3.12 depicts the heads and tails for the ADMG illustrated in Figure 3.11 (i) and the posets and partial order. Accordingly, a probability measure obeys the global Markov property with respect to the graph if and only if it factors as:

$$\begin{aligned}
 f_{abcd}(x) &= f_{d|abc}(x) f_{c|b}(x) f_{b|a}(x) f_a(x) && \text{for } \nu\text{-a.e. } x \in \mathcal{X} \\
 f_{abc}(x) &= f_{c|b}(x) f_{b|a}(x) f_a(x) && \text{for } \nu\text{-a.e. } x \in \mathcal{X} \\
 f_{ab}(x) &= f_{b|a}(x) f_a(x) && \text{for } \nu\text{-a.e. } x \in \mathcal{X}
 \end{aligned}$$

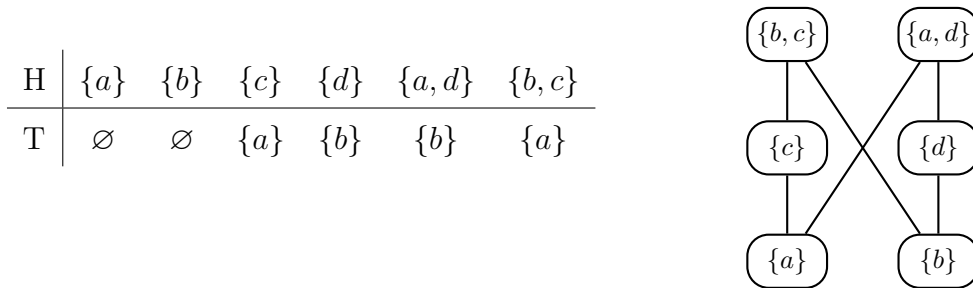


Figure 3.13: The heads and tails for the ADMG illustrated in Figure 3.11 (ii) and the Hasse diagram for the corresponding poset over the ADMG's heads.

Figure 3.13 depicts the heads and tails for the ADMG illustrated in Figure 3.11 (ii) and the posets and partial order. Accordingly, a probability measure obeys the global Markov property with respect to the graph if and only if it factors as:

$$\begin{aligned}
f_{abcd}(x) &= f_{ad|b}(x) f_{bc|a}(x) && \text{for } \nu\text{-a.e. } x \in \mathcal{X}; \\
f_{abc}(x) &= f_{bc|a}(x) f_a(x) && \text{for } \nu\text{-a.e. } x \in \mathcal{X}; \\
f_{abd}(x) &= f_{ad|b}(x) f_b(x) && \text{for } \nu\text{-a.e. } x \in \mathcal{X}; \\
f_{ab}(x) &= f_a(x) f_b(x) && \text{for } \nu\text{-a.e. } x \in \mathcal{X}; \\
f_{ac}(x) &= f_{c|a}(x) f_a(x) && \text{for } \nu\text{-a.e. } x \in \mathcal{X}; \\
f_{bd}(x) &= f_{d|b}(x) f_b(x) && \text{for } \nu\text{-a.e. } x \in \mathcal{X}.
\end{aligned}$$

Note that both the factorization characterized by Evans and Richardson and the factorization presented in this proposal are equivalent to the global Markov property and therefore equivalent to each other. The key difference is that the factorization characterized by Evans and Richardson requires an equation for every non-empty ancestral subset of variables, while the factorization presented in this proposal only requires a single equation.

### 3.3.6 Markov Equivalence

Multiple graphs representing the same independence model is made rigorous by the notion of Markov equivalence.

**Definition** (*Markov equivalence*). Let  $\mathcal{G} = (V, E)$  and  $\mathcal{G}' = (V, E')$  be mixed graphs.  $\mathcal{G}$  and  $\mathcal{G}'$  are *Markov equivalent* if  $\mathcal{J}(\mathcal{G}) = \mathcal{J}(\mathcal{G}')$ :

$$A \perp\!\!\!\perp B \mid C [\mathcal{G}] \iff A \perp\!\!\!\perp B \mid C [\mathcal{G}'] \quad \text{for all } \langle A, B \mid C \rangle \in \mathcal{T}(V).$$

As noted above, there exists a unique MAG for every ancestral graph with the same independence model. Accordingly, Markov equivalence is usually discussed in terms of MAGs rather than ancestral graphs. Furthermore, the set of MAGs that form a Markov equivalence class may be graphically summarized using a maximally informative partial ancestral graph

(PAG). A maximally informative PAG is not a mixed graph, but a graph that summarizes a set of mixed graphs. In addition to the standard set of edges used by mixed graphs, maximally informative PAGs also include edges with circle edge marks to denote ambiguity—the edge mark varies among the summarized graphs.

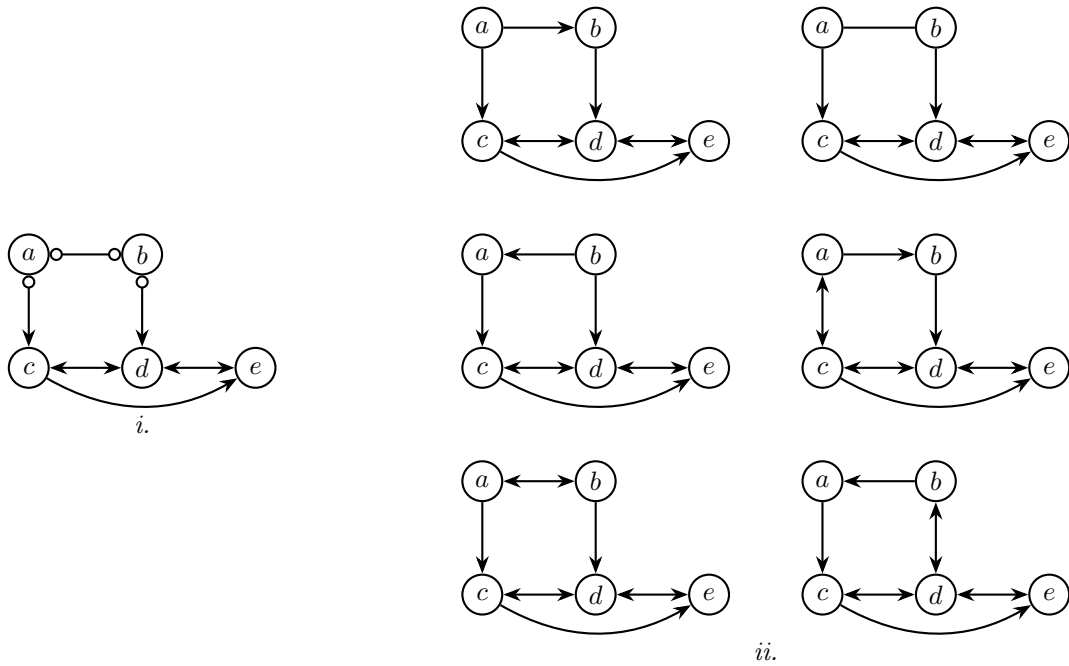


Figure 3.14: A Markov equivalence class of MAGs with vertices  $\{a, b, c, d, e\}$ : (i) a maximally informative PAG; (ii) a set of Markov equivalent MAGs.

**Definition** (*maximally informative partial ancestral graph*). A maximally informative PAG is a graph used to summarize the Markov equivalence class of a MAG and contains at most one of six possible edge types  $\{\rightarrow, \leftrightarrow, -, \circ-\circ, \circ\rightarrow, \rightarrow\circ\}$  between every pair of vertices.

If  $\mathcal{G}$  is a MAG, then the maximally informative PAG  $[\mathcal{G}]$  for  $\mathcal{G}$  is a graph with the same adjacencies as  $\mathcal{G}$ . Furthermore, every non-circle edge mark in  $[\mathcal{G}]$  occurs in every member of  $\mathcal{G}$ 's Markov equivalence class and every circle edge mark in  $[\mathcal{G}]$  corresponds to an edge mark that varies among the members of  $\mathcal{G}$ 's Markov equivalence class.

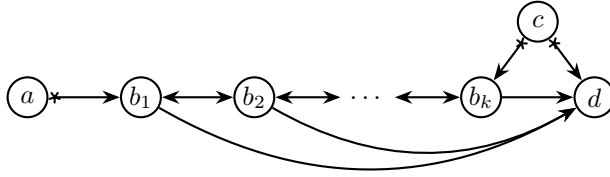


Figure 3.15: The general form of a discriminating path.

The concept of a discriminating path partly characterizes whether two MAGs belong to the same Markov equivalence class. Figure 3.15 depicts the general form of a discriminating path, where asterisks are used to denote edge marks that may either be an arrowhead or a tail.

**Definition** (*discriminating path*). Let  $\mathcal{G} = (V, E)$  be a MAG with a path  $\pi = \langle a, b_1, \dots, b_k, c, d \rangle$  ( $k \geq 1$ ). We say  $\pi$  is a discriminating path for  $c$  if:

- i.*  $a$  and  $d$  are not adjacent;
- ii.*  $b_i$  is a collider on  $\pi$  for all  $1 \leq i \leq k$ ;
- iii.*  $b_i$  is a parent of  $d$  for all  $1 \leq i \leq k$ .

**Theorem 3.3.7** (Theorem 1 [81]). *Let  $\mathcal{G}$  and  $\mathcal{G}'$  be MAGs.  $\mathcal{G}$  and  $\mathcal{G}'$  are Markov equivalent if and only if:*

- i.*  $\mathcal{G}$  and  $\mathcal{G}'$  have the same adjacencies;
- ii.*  $\mathcal{G}$  and  $\mathcal{G}'$  have the same unshielded colliders;
- iii.* if  $\pi = \langle a, b_1, \dots, b_k, c, d \rangle$  ( $k \geq 1$ ) is a discriminating path for  $c$  in  $\mathcal{G}$  and  $\mathcal{G}'$ , then  $c$  is a collider on  $\pi$  in  $\mathcal{G}$  if and only if it is a collider on  $\pi$  in  $\mathcal{G}'$ .

**Definition** (*parametrizing sets*). The parametrizing set of  $\mathcal{G}$ , denoted by  $\mathcal{S}(\mathcal{G})$  is defined as follows:

$$\mathcal{S}(\mathcal{G}) \equiv \{HT ; \quad H \in \mathcal{H}(\mathcal{G}) \text{ and } T \subseteq \text{tail}_{\mathcal{G}}(H)\}.$$

This definition is extended from directed MAGs to all MAGs by adding all *cliques* of the *undirected subgraph* to the set. The undirected subgraph is the graph with the same vertices



where all directed and bi-directed edges have been removed. A clique is a complete subset of the graph, that is, every vertex in the subset is connected to every other vertex in the subset. The following results hold:

**Proposition 3.3.5** (Proposition 3.3 [38]). *Let  $\mathcal{G} = (V, E)$  be a MAG containing a set  $N \subseteq V$ .  $N \notin \mathcal{S}(\mathcal{G})$  if and only if there exist  $a, b \in N$  ( $a \neq b$ ) and  $C \subseteq V$  ( $N \subseteq C$ ) such that  $a$  and  $b$  are  $m$ -separated by  $C \setminus \{a, b\}$ .*

**Proposition 3.3.6** (Proposition 3.4 [38]). *For a MAG  $\mathcal{G}$ , we have*

- i. any two vertices  $a$  and  $b$  are adjacent in  $\mathcal{G}$  if and only if  $\{a, b\} \in \mathcal{S}(\mathcal{G})$ ;*
- ii. for any unshielded triple  $\langle a, b, c \rangle$  in  $\mathcal{G}$ ,  $\{a, b, c\} \in \mathcal{S}(\mathcal{G})$  if and only if  $b$  is a collider on the triple  $\langle a, b, c \rangle$ ;*
- iii. if  $\pi$  forms a discriminating path for  $b$  with endpoints  $a$  and  $c$  in  $\mathcal{G}$  then  $\{a, b, c\} \in \mathcal{S}(\mathcal{G})$  if and only if  $b$  is a collider on  $\pi$ .*

**Theorem 3.3.8** (Theorem 3.2 [38]). *Let  $\mathcal{G}$  and  $\mathcal{G}'$  be MAGs.  $\mathcal{G}$  and  $\mathcal{G}'$  are Markov equivalent if and only if  $\mathcal{S}(\mathcal{G}) = \mathcal{S}(\mathcal{G}')$ .*

Hu and Evans refine the set of parametrizing sets by specifying a subset that is particularly useful for efficient calculation of Markov equivalence.

$$\tilde{\mathcal{S}}(\mathcal{G}) \equiv \{T \in \mathcal{S}(\mathcal{G}) ; \quad 1 \leq |\mathcal{P}_2^2(T) \cap \mathcal{S}(\mathcal{G})| \leq 2 \leq |T| \leq 3\}$$

**Corollary 3.3.2** (Corollary 3.2.1 [38]). *Let  $\mathcal{G}$  and  $\mathcal{G}'$  be MAGs.  $\mathcal{G}$  and  $\mathcal{G}'$  are Markov equivalent if and only if  $\tilde{\mathcal{S}}(\mathcal{G}) = \tilde{\mathcal{S}}(\mathcal{G}')$ .*

### 3.4 Stable Mixed Graphs

Suppose that the causal relationships of a system of variables can be correctly represented by a DAG. If only some variables are measured and others are latent or measured selection bias, then the system of variables can be represented by the marginalization and conditioning of the DAG respectively. Accordingly, we often refer to the marginalization set as  $L$  and the conditioning set as  $S$ . In some cases, we consider marginals and conditionals of the graph

for theoretical purposes. That is, when we refer to a latent or selection variable, we may be referring to a variable that has been marginalized or conditioned on.

Families of stable mixed graphs are families that are closed under this process of marginalization and conditioning. Since marginalization and conditioning can correspond to latent confounding and selection effects, these families of graphs are quite useful for modeling. If a graphical family is not stable under marginalization and conditioning, then dealing with latent confounding and selection effects can be more difficult; see the example in Chapter 2.

DAGs make up an important family of graphs. In particular, Bayesian networks, which are graphical Markov models that use DAGs, have been applied with much success across many domains. However, when a subset of variables in a DAG are latent, independence models induced by DAGs are generally insufficient to encode the complete set of conditional independence statements represented in the probability measure of a Markov model. Latent confounding variables and selection bias are treated as marginalization and conditioning respectively. Accordingly, this shortcoming manifests statistically as a lack of stability under marginalization and conditioning.

In this section, we discuss previous works on mixed graphs that capture the modified independence structure of a DAG after marginalization over unobserved variables and conditioning on selection variables using the  $m$ -separation criterion. These include ribbonless, summary, and ancestral graphs. Ribbonless graphs were introduced in order to straightforwardly deal with the problem of finding a superset of the family of DAGs that is stable under marginalization and conditioning while summary graphs extend ADMGs to include undirected edges.

**Definition** (*summary graph*). Let  $\mathcal{G} = (V, E)$  be a mixed graph.  $\mathcal{G}$  is a summary graph if for every  $a \in V$ :

- i.*  $\text{ch}_{\mathcal{G}}(a) \cap \text{an}_{\mathcal{G}}(a) = \emptyset$ ;
- ii.*  $\text{pa}_{\mathcal{G}}(a) \cup \text{sp}_{\mathcal{G}}(a) \neq \emptyset \Rightarrow \text{ne}_{\mathcal{G}}(a) = \emptyset$ .

The family of summary graphs extends the family of ancestral graphs. In particular, summary graphs are loopless rather than simple—summary graphs can contain multiple edges. Additionally, criterion (*ii*) of ancestral graphs has been removed—summary graphs

can contain almost directed cycles. Figure 3.16 illustrates an example of a summary graph that is not an ancestral graph.

**Definition** (*ribbonless graph* [73]). Let  $\mathcal{G} = (V, E)$  be a mixed graph.  $\mathcal{G}$  is a ribbonless graph if for every triple  $\langle a, b, c \rangle$  in  $\mathcal{G}$  where:

$$\left\{ \begin{array}{l} a \rightarrow b \leftarrow c \\ a \leftrightarrow b \leftrightarrow c \\ a \rightarrow b \leftrightarrow c \end{array} \right\} \text{ in } \mathcal{G} \text{ and } \left\{ \begin{array}{l} a - c \\ a \leftrightarrow c \\ a \rightarrow c \end{array} \right\} \text{ not in } \mathcal{G};$$

for all vertices  $d \in \text{de}_{\mathcal{G}}(b)$ :

- i.*  $\text{ch}_{\mathcal{G}}(d) \cap \text{an}_{\mathcal{G}}(d) = \emptyset$ ;
- ii.*  $\text{ne}_{\mathcal{G}}(d) = \emptyset$ .

The family of ribbonless graphs extends the family of a summary graphs. In particular, the criteria of summary graphs are only required hold for descendants of colliders with a special form. Figure 3.16 illustrates an example of a ribbonless graph that is not a summary graph.

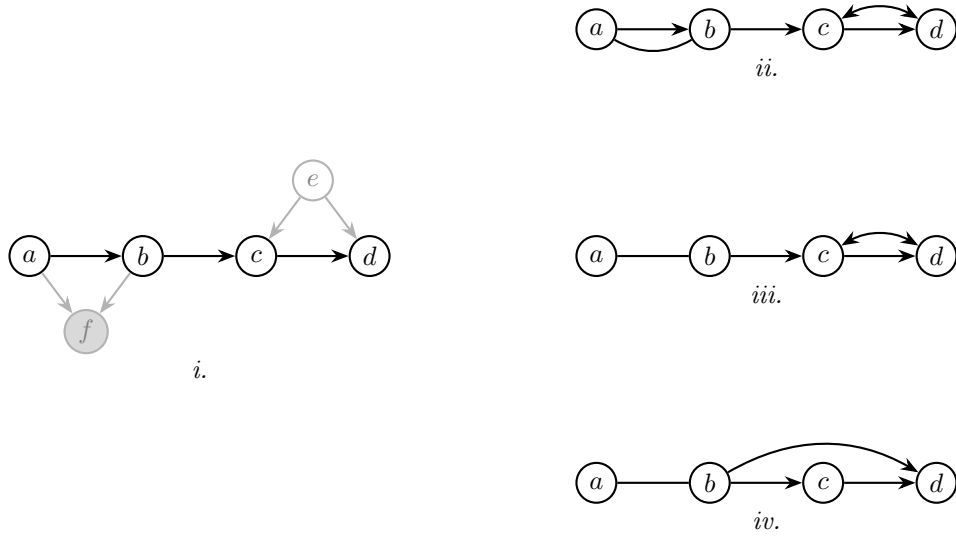


Figure 3.16: Stable mixed graphs: (i) a DAG with latent and selection variables; (ii) the projected ribbonless graph; (iii) the projected summary graph; (iv) the projected ancestral graph. All graphs encode the same independence model over the measured variables using  $m$ -separation.

Accordingly, the graphical families discussed in this dissertation form a hierarchy. This hierarchy is further expanded through the application of “directed” and “maximal” modifiers.

- RG     Ribbonless Graph;
- SG     Summary Graph;
- AnG    Ancestral Graph;
- MAG    Maximal Ancestral Graph;
- UG     Undirected Graph.
- ADMG   Acyclic Directed Mixed Graph;
- DAnG   Directed Ancestral Graph;
- DMAG   Directed Maximal Ancestral Graph;
- DAG     Directed Acyclic Graph;

Ribbonless, summary, and ancestral graphs are stable under marginalization and conditioning and their directed counterparts are stable under marginalization; see the top right of Figure 3.17.

In what follows, we use  $\mathcal{F}$  to denote a family of graphs. Furthermore, we use  $\mathcal{F}(V)$  to

denote a family of graphs over vertex set  $V$ .

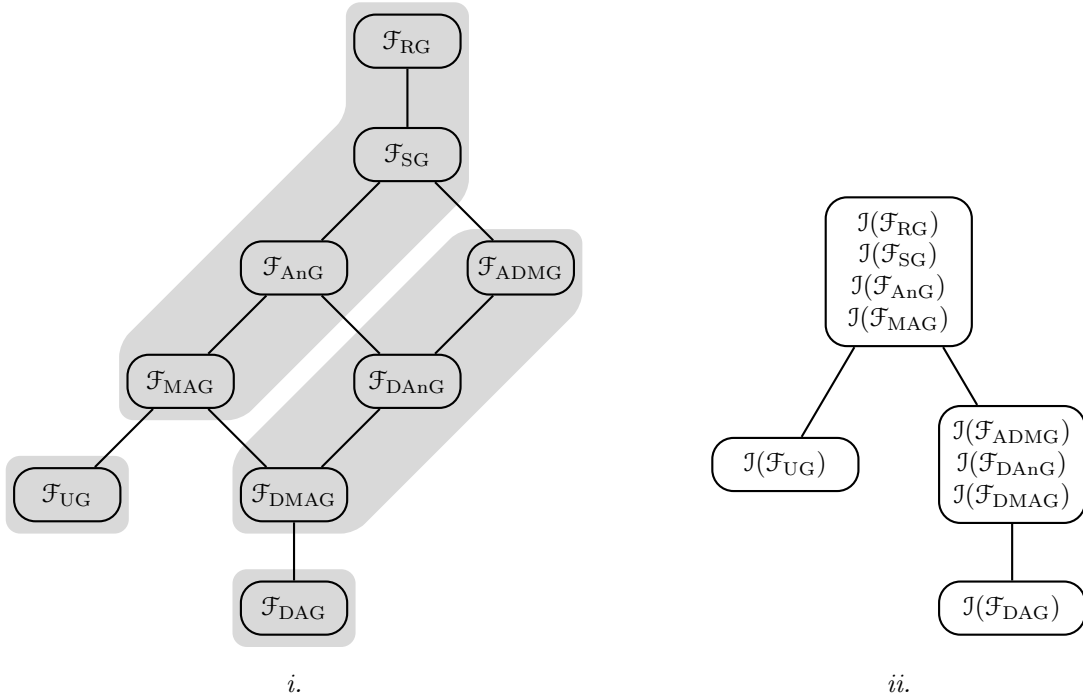


Figure 3.17: Hasse diagrams for posets of graphical families: (i) families of stable mixed graphs and DAGs ordered by inclusion; (ii) independence models of the families of stable mixed graphs and DAGs ordered by inclusion.

Figure 3.17 (i) depicts a Hasse diagram for a poset of graphical families ordered by inclusion—the colored sections indicate families that induce the same independence models as before. Figure 3.17 (ii) depicts a Hasse diagram for the poset of independence models induced by the families of graphs ordered by inclusion.

### 3.4.1 Marginalization and Conditioning

Let  $\mathcal{J}$  be an independence model over a non-empty set of variables  $V$  with a subset  $L \subseteq V$ . The resulting independence model after marginalizing  $L$  out of  $\mathcal{J}$ , denoted  $\alpha(\mathcal{J}, L, \emptyset)$ , is the

subset of disjoint triples that do not involve any members of  $L$ :

$$\alpha(\mathcal{J}, L, \emptyset) \equiv \{ \langle A, B \mid C \rangle \in \mathcal{J}(V \setminus L) ; \quad \langle A, B \mid C \rangle \in \mathcal{J} \}.$$

If  $\mathcal{J}$  captures the conditional independence statements represented in a probability measure  $P$ , then  $\alpha(\mathcal{J}, L, \emptyset)$  returns the set of conditional independence relations after marginalizing  $L$  out of  $P$ . The symbol  $L$  is used because latent variables represent one context in which marginalization may occur.

While the various families of stable mixed graphs are all stable under marginalization and conditioning, they were developed for different purposes. In this dissertation we will identify their differences based on the causal relationships and amount of information that they can represent. Since the maximal modifier primarily exists for statistical convenience and the directed modifier is used for cases where there is no conditioning, we discuss the families of ribbonless graphs, summary graphs, and ancestral graphs

In general, families of stable mixed graphs use the various edges types of mixed graphs as follows: directed edges identify dependence due to causal ancestry; bi-directed identify dependence due to marginalization or latent confounding; and undirected edges identify dependence due to conditioning or selection effects. The families of stable mixed graphs differ in how they resolve conflicts of multiple sources of dependence. Figure 3.16 provides a visual aid for the following comparison.

The family of ribbonless graphs is the most general family of stable mixed graphs. Ribbonless graphs include all edges that apply to a given pair of vertices. Accordingly, ribbonless graphs can have up to three edges (directed, bi-directed, and undirected) between a pair of vertices. For this reason, they are able to encode constraints beyond conditional independence constraints, however, to our knowledge, the extent of these constraints has not been studied. Note that ribbonless graphs can encode any form of constraint encoded by summary graphs. An algorithm to construct ribbonless graphs by latent projection is detailed in Algorithm 8.

The family of summary graphs lies between ribbonless graphs and ancestral graphs in terms of complexity. Summary graphs give priority to undirected edges and include all edges that apply otherwise for a given pair of vertices. Accordingly, summary graphs can have up

to two edges (directed and bi-directed) between a pair of vertices. For this reason they are able to encode constraints beyond conditional independence constraints. These constraints have been studied in some detail [29, 76, 90]. An algorithm to construct summary graphs by latent projection is detailed in Algorithm 9.

The family of ancestral graphs is the simplest family of stable mixed graphs. Ancestral graphs give first priority to undirected edges, second priority to directed edges, and third priority to bi-directed edges for a given pair of vertices. Accordingly, ancestral graphs can have up to a single edge between a pair of vertices. Due to their simplicity, ancestral graphs only represent condition independence constraints. An algorithm to construct MAGs by latent projection is detailed in Algorithm 10.

Let  $\mathcal{J}$  be an independence model over a non-empty set of variables  $V$  with a subset  $S \subseteq V$ . The resulting independence model after conditioning  $\mathcal{J}$  on  $S$ , denoted  $\alpha(\mathcal{J}, \emptyset, S)$ , is the subset of disjoint triples defined as follows:

$$\alpha(\mathcal{J}, \emptyset, S) \equiv \{\langle A, B \mid C \rangle \in \mathcal{T}(V \setminus S) ; \langle A, B \mid CS \rangle \in \mathcal{J}\}.$$

If  $\mathcal{J}$  captures the conditional independence statements represented in a probability measure  $P$ , then  $\alpha(\mathcal{J}, \emptyset, S)$  returns the set of conditional independence relations after conditioning  $P$  on  $S$ . The symbol  $S$  is used because selection bias represent one context in which conditioning may occur.

Combining these definitions, we obtain:

$$\alpha(\mathcal{J}, L, S) \equiv \{\langle A, B \mid C \rangle \in \mathcal{T}(V \setminus LS) ; \langle A, B \mid CS \rangle \in \mathcal{J}\}.$$

If  $\mathcal{J}$  captures the conditional independence statements represented in a probability measure  $P$ , then  $\alpha(\mathcal{J}, L, S)$  returns the set of conditional independence relations after marginalizing  $L$  out of  $P$  and conditioning  $P$  on  $S$ .

### 3.4.2 Latent Projections

We may apply the marginalization and conditioning operations directly to graphs using the concept of latent projection. Although the concept of latent projection was introduced

by Pearl and Verma [61], Sadeghi provides the most complete treatment of latent projection [73]. Consider a family of graphs  $\mathcal{F}$ . If for every graph  $\mathcal{G} = (V, E) \in \mathcal{F}$  and disjoint sets  $L, S \subseteq V$  there is a graph  $\mathcal{G}' \in \mathcal{F}$  such that  $\mathcal{I}(\mathcal{G}') = \alpha(\mathcal{I}(\mathcal{G}), L, \emptyset)$ , then  $\mathcal{F}$  is stable under marginalization, and if there is a graph  $\mathcal{G}' \in \mathcal{F}$  such that  $\mathcal{I}(\mathcal{G}') = \alpha(\mathcal{I}(\mathcal{G}), \emptyset, S)$ , then  $\mathcal{F}$  is stable under conditioning. Furthermore, we call  $\mathcal{F}$  stable under marginalization and conditioning if there is a graph  $\mathcal{G}'$  such that  $\mathcal{I}(\mathcal{G}') = \alpha(\mathcal{I}(\mathcal{G}), L, S)$ . Below, we define an algorithm for the latent projections of ancestral graphs. Additional algorithms for the latent projections of ribbonless, summary, and ancestral graphs are provided in Appendix B.1

Let  $\mathcal{G} = (V, E)$  be a MAG such that  $V$  contains disjoint subsets  $L, S \subset V$ . The resulting graph after marginalizing  $L$  out of  $\mathcal{G}$  and conditioning  $\mathcal{G}$  on  $S$ , denoted  $\alpha_{AG}(\mathcal{G}; L, S)$ , is a graph over the set of vertices  $V \setminus LS$ , and edges specified as follows: For all distinct vertices  $a, b \in V \setminus LS$  where there exists an inducing path between  $a$  and  $b$  relative to  $\langle L, S \rangle$

$$\text{if } \left\{ \begin{array}{l} a \in \text{ant}_{\mathcal{G}}(b \cup S) \text{ and } b \notin \text{ant}_{\mathcal{G}}(a \cup S) \\ a \notin \text{ant}_{\mathcal{G}}(b \cup S) \text{ and } b \notin \text{ant}_{\mathcal{G}}(a \cup S) \\ a \in \text{ant}_{\mathcal{G}}(b \cup S) \text{ and } b \in \text{ant}_{\mathcal{G}}(a \cup S) \end{array} \right\} \text{ then } \left\{ \begin{array}{l} a \rightarrow b \\ a \leftrightarrow b \\ a - b \end{array} \right\} \text{ in } \alpha_{AG}(\mathcal{G}; L, S).$$

That is,  $\alpha_{AG}(\mathcal{G}; L, S)$  is a graph containing vertices  $V \setminus LS$  and edges between vertices that are  $m$ -connecting in  $\mathcal{G}$  given all subsets containing the members of  $S$  and no members of  $L$ . Furthermore, an edge between two distinct vertices  $a, b \in V \setminus LS$  will have an arrowhead at  $a$  if and only if  $a$  is not an ancestor of  $b$  or  $s \in S$  in  $\mathcal{G}$ , and a tail otherwise.

Richardson and Spirtes showed that latent projection has several nice properties.

**Theorem 3.4.1** (Theorem 4.18 [70]). *If  $\mathcal{G} = (V, E)$  is a MAG containing disjoint sets  $L, S \subseteq V$ , then:*

$$\alpha(\mathcal{I}(\mathcal{G}); L, S) = \mathcal{I}(\alpha_{AG}(\mathcal{G}; L, S))$$

In words, the independence model corresponding to the transformed graph is the independence model obtained by marginalizing and conditioning the independence model of the original graph. Additionally, the latent projection procedure defined by Richardson and Spirtes and has several nice properties.

**Corollary 3.4.1** (Corollary 4.19 [70]). *If  $\mathcal{G} = (V, E)$  is a MAG containing disjoint sets*



$L, S \subseteq V$ , then  $\alpha_{AG}(\mathcal{G}; L, S)$  is a MAG.

**Theorem 3.4.2** (Theorem 4.18 [70]). *If  $\mathcal{G} = (V, E)$  is a MAG containing disjoint sets  $L_1, L_2, S_1, S_2 \subseteq V$ , then:*

$$\alpha_{AG}(\alpha_{AG}(\mathcal{G}; L_1, S_1); L_2, S_2) = \alpha_{AG}(\mathcal{G}; L_1 \cup L_2, S_1 \cup S_2)$$

Furthermore, the family of directed MAGs represents DAG under marginalization, that is, directed MAGs are capable of representing latent confounding.

**Proposition 3.4.1** (Proposition 4.13 [70]). *If  $\mathcal{G}$  is an ancestral graph which contains no undirected edges, then neither does  $\alpha(\mathcal{G}, L, \emptyset)$ .*

Let  $\mathcal{G} = (V, E)$  be a MAG containing an anterior set  $A \in \mathcal{A}(\mathcal{G})$ . Next we note the induced subgraph  $\mathcal{G}_A$  and the latent projection  $\alpha(\mathcal{G}; V \setminus A, \emptyset)$  are related—namely that they are the same. First, note two useful results about the anterior relationships in ancestral graphs and their marginals.

**Corollary 3.4.2** (Corollary 3.10 [70]). *Let  $\mathcal{G} = (V, E)$  and  $\mathcal{G}' = (V, E')$  are ancestral graphs with the same adjacencies. If for all  $a, b \in V$ , adjacent in  $\mathcal{G}$  and  $\mathcal{G}'$ ,  $a \in \text{ant}_{\mathcal{G}}(b) \Leftrightarrow a \in \text{ant}_{\mathcal{G}'}(b)$ , then  $\mathcal{G} = \mathcal{G}'$ .*

**Corollary 3.4.3** (Corollary 4.8 [70]). *In an ancestral graph  $\mathcal{G} = (V, E)$  if  $a \in V \setminus L$  then  $\text{ant}_{\mathcal{G}}(a) \setminus L = \text{ant}_{\alpha(\mathcal{G}; L, \emptyset)}(a)$ .*

We now show that induced subgraphs on anterior sets are the marginals over the same vertices.

**Proposition 3.4.2.** *Let  $\mathcal{G} = (V, E)$  be a MAG containing sets  $A, L \subseteq V$  that partition  $V$ . If  $A \in \mathcal{A}(\mathcal{G})$  is an anterior set, then:*

$$\mathcal{G}_A = \alpha(\mathcal{G}; L, \emptyset).$$

*Proof.* By construction, for all  $a, b \in A$  adjacent in  $\mathcal{G}_A$  and  $\mathcal{G}$ ,  $a \in \text{ant}_{\mathcal{G}_A}(b)$  if and only if  $a \in \text{ant}_{\mathcal{G}}(b)$ . By Corollary 3.4.3, for all  $a, b \in A$  adjacent in  $\mathcal{G}_A$  and  $\alpha(\mathcal{G}; L, \emptyset)$ ,  $a \in \text{ant}_{\mathcal{G}_A}(b)$  if and only if  $a \in \text{ant}_{\alpha(\mathcal{G}; L, \emptyset)}(b)$ . What remains to be shown is that they have the same adjacencies.

Take two arbitrary vertices that are not latent. We need to show that they are adjacent in  $\mathcal{G}_A$  if and only if they are adjacent in  $\alpha(\mathcal{G}; L, \emptyset)$ . By definition, there is an edge in  $\alpha(\mathcal{G}; L, \emptyset)$  if and only if there is an inducing path in  $\mathcal{G}$  with respect to  $\langle L, \emptyset \rangle$ . Therefore, we show that there is an edge in  $\mathcal{G}_A$  if and only if there is an inducing path in  $\mathcal{G}$  with respect to  $\langle L, \emptyset \rangle$ .

In other words, there is a primitive inducing path in  $\mathcal{G}_A$  if and only if there is an inducing path in  $\mathcal{G}$  with respect to  $\langle L, \emptyset \rangle$ .

Every (primitive inducing) path in  $\mathcal{G}_A$  is in  $\mathcal{G}$  by construction. Since these paths do not include  $L$ , they are inducing in  $\mathcal{G}$  with respect to  $\langle L, \emptyset \rangle$ .

Suppose that there is an inducing path with respect to  $\langle L, \emptyset \rangle$  in  $\mathcal{G}$  that is not a primitive inducing path in  $\mathcal{G}_A$ . Then there is a non-collider in  $L$  on the path. Since  $L$  is a non-collider, it is anterior to either an endpoint or a collider on the path. Since collider on the path are ancestors of the endpoints by definition, the vertex must be anterior to an endpoint. This is a contradiction because  $a, b \in A \in \mathcal{A}(\mathcal{G})$ .

By Corollary 3.4.2,  $\mathcal{G}_A = \alpha(\mathcal{G}; L, \emptyset)$ .

□

### 3.5 Alternative Independence Models

In this dissertation, we consider several mathematical objects apart from probability measures and graphs that induce independence models. In this section, we discuss integer-valued multisets, multiinformation, and supermodular functions as alternative mathematical objects that induce independence models. In this section, we introduce these objects and their relevant properties. In particular, this work relies heavily on the theory of integer-valued multisets or *imset* for short; see Studený [83] for more details.

**Definition** (*integer-valued multiset*). Let  $V$  be a non-empty set of variables. An *integer-valued multiset* over  $V$  is an integer-valued function  $u: \mathcal{P}(V) \rightarrow \mathbb{Z}$  or, alternatively, an element of  $\mathbb{Z}^{\mathcal{P}(V)}$ .

Basic operations with imsets—summation, subtraction, and multiplication by an integer—

are defined coordinate-wise. Besides basic operations with imsets, an operation of a scalar product of a real-valued function  $m : \mathcal{P}(V) \rightarrow \mathbb{R}$  and an imset  $u$  over  $V$  defined by

$$u^\top m \equiv \sum_{T \in \mathcal{P}(V)} u(T) m(T)$$

is used. A simple example of an imset is the *identifier* of a set  $A \subseteq V$  denoted by  $\delta_A$  and defined as follows:

$$\delta_A(T) \equiv \begin{cases} 1 & T = A; \\ 0 & T \subseteq V, T \neq A. \end{cases}$$

We generalize the concept of the identifier to sets of sets. The identifier of a set of sets  $\mathbf{A} \subseteq \mathcal{P}(V)$  is denoted by  $\delta_{\mathbf{A}}$  and defined as follows:

$$\delta_{\mathbf{A}}(T) \equiv \begin{cases} 1 & T \in \mathbf{A}; \\ 0 & T \subseteq V, T \notin \mathbf{A}. \end{cases}$$

### 3.5.1 Elementary and Semi-elementary Imsets

Elementary and semi-elementary conditional independence statements can be expressed as imsets of the same name. This becomes clear in the following sections on supermodular functions and structural imsets.

**Definition** (*elementary imset*). Let  $V$  be a non-empty set of variables and  $\langle a, b \mid C \rangle \in \mathcal{J}(V)$  be a disjoint triple over  $V$ . The corresponding *elementary imset* over  $V$  is an imset defined by the formula:

$$u_{\langle a, b \mid C \rangle} \equiv \delta_{ab \cup C} + \delta_C - \delta_{a \cup C} - \delta_{b \cup C}.$$

$$\begin{array}{rcccccccc}
T & \emptyset & \{a\} & \{b\} & \{c\} & \{a,b\} & \{a,c\} & \{b,c\} & \{a,b,c\} \\
u_{\langle a,b|c \rangle}(T) & \left[ \begin{array}{cccccccc}
0 & 0 & 0 & 1 & 0 & -1 & -1 & 1
\end{array} \right]^T
\end{array}$$

Figure 3.18: An elementary imset:  $u_{\langle a,b|c \rangle}$ .

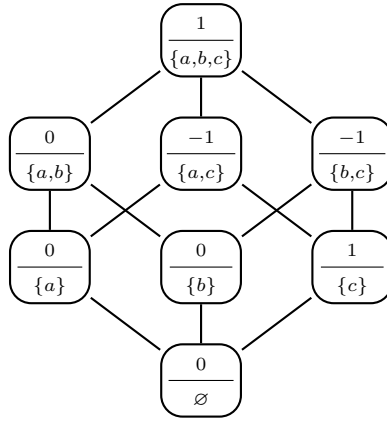


Figure 3.19: The Hasse diagram for an elementary imset:  $u_{\langle a,b|c \rangle}$ .

**Definition** (*semi-elementary imset*). Let  $V$  be a non-empty set of variables and  $\langle A, B \mid C \rangle \in \mathcal{T}(V)$  be a disjoint triple over  $V$ . The corresponding *semi-elementary imset*  $u_{\langle A,B|C \rangle}$  is defined by the formula:

$$u_{\langle A,B|C \rangle} \equiv \delta_{ABC} + \delta_C - \delta_{AC} - \delta_{BC}.$$

**Proposition 3.5.1** (Proposition 4.2 [83]). *Every semi-elementary imset is a linear combination of elementary imsets with non-negative integer coefficients.*

### 3.5.2 Multiinformation

Supermodular functions, in particular the multiinformation of a probability measure, are essential concepts for the theory of imsetal Markov models as they connect semi-elementary imsets to probabilistic conditional independence.

**Definition** (*supermodular function*). Let  $V$  be a non-empty set of variables. A function  $m : \mathcal{P}(V) \rightarrow \mathbb{R}$  is a *supermodular function* over  $V$  if

$$m(A \cup B) + m(A \cap B) \geq m(A) + m(B) \quad \text{for all } A, B \subseteq V.$$

**Definition** (*multiinformation*). Let  $V$  be a non-empty set of variables containing a subset  $A \subseteq V$ . Furthermore, let  $X$  be a collection of random variables indexed by  $V$  with probability measure  $P$  that admits density  $f(x)$  with respect to dominating  $\sigma$ -finite product measure  $\nu$ . The *multiinformation* of  $P$  is a real-valued function  $m_P : \mathcal{P}(V) \rightarrow [0, \infty)$  that is the relative entropy of  $P$  with respect to the product of its one-dimensional marginals:

$$m_P(A) \equiv \begin{cases} \int_{x \in \mathcal{X}_A} \log \left[ \frac{f_A(x)}{\prod_{a \in A} f_a(x)} \right] dP(x) & A \neq \emptyset; \\ 0 & A = \emptyset. \end{cases}$$

In the field of information theory, the above integral is an instance of Kullback-Liebler divergence or relative entropy. Other terms for multiinformation in the literature include *total correlation*, *dependency tightness*, and *entaxy* [83]. The following corollary gives a nice intuition for elementary and semi-elementary imsets can be used in conjunction with multiinformation to define probabilistic conditional independence.

**Proposition 3.5.2** (Corollary 2.2 [83]). *Let  $V$  be a non-empty set of variables and  $P$  be a probability measure over  $V$ . If  $P$  has finite multiinformation  $m_P$ , then  $m_P$  is a non-negative supermodular function that satisfies*

$$m_P(A) = 0 \quad \text{whenever } A \subseteq V \text{ } (|A| \leq 1).$$

*That is,*

$$m_P(ABC) + m_P(C) - m_P(AC) - m_P(BC) \geq 0 \quad \text{for all } \langle A, B \mid C \rangle \in \mathcal{J}(V).$$

These two conditions imply  $m_P(A) \leq m_P(B)$  whenever  $A \subseteq B \subseteq V$ . Moreover, for every  $\langle A, B \mid C \rangle \in \mathcal{T}(V)$

$$m_P(ABC) + m_P(C) - m_P(AC) - m_P(BC) = 0 \quad \Leftrightarrow \quad A \perp\!\!\!\perp B \mid C [P].$$

### 3.5.3 Structural Imsets as Independence Models

**Definition** (*structural imset*). Let  $V$  be a non-empty set of variables and  $u$  be an imset over  $V$ . The imset  $u$  is *structural* if it is a linear combination of elementary imsets with non-negative rational coefficients:

$$u \equiv \sum_{\langle A, B \mid C \rangle \in \mathcal{T}(V)} k_{\langle A, B \mid C \rangle} u_{\langle A, B \mid C \rangle} \quad \text{for some } k_{\langle A, B \mid C \rangle} \in \mathbb{Q}_+.$$

One says that a disjoint triple  $\langle A, B \mid C \rangle \in \mathcal{T}(V)$  is represented in a structural imset  $u$  over  $V$  and writes  $A \perp\!\!\!\perp B \mid C [u]$  if there exists  $k \in \mathbb{Q}_+$  such that  $u - k u_{\langle A, B \mid C \rangle}$  is a structural imset over  $V$ . The class of represented triples then defines the (conditional independence) model induced by  $u$ :

$$\mathcal{J}(u) \equiv \{ \langle A, B \mid C \rangle \in \mathcal{T}(V) ; \quad A \perp\!\!\!\perp B \mid C [u] \}.$$

Unlike the previously discussed families of mixed graphs which induce compositional graphoid independence models, structural imsets induce semi-graphoid independence models.

**Proposition 3.5.3** (Lemma 4.6 [83]). *A structural imset over  $V$  induces a semi-graphoid over  $V$ .*

The primary advantage of structural imsets is their representation power. In fact, structural imsets can represent the independence model of any probability measure with finite multiinformation [83]. Structural imsets are closely related to supermodular functions.

**Proposition 3.5.4** (Proposition 5.1 [83]). *Let  $V$  be a non-empty set of variables. A function  $m : \mathcal{P}(V) \rightarrow \mathbb{R}$  is supermodular if and only if any of the following three conditions holds:*

- i.  $u^\top m \geq 0$  for every structural imset  $u$  over  $V$ ;*
- ii.  $u^\top m \geq 0$  for every semi-elementary imset  $u$  over  $V$ ;*

iii.  $u^\top m \geq 0$  for every elementary imset  $u$  over  $V$ .

We now give a factorization using of structural imsets.

**Theorem 3.5.1** (Theorem 4.1 [83]). *Let  $V$  be a non-empty set of variables and  $u$  be a structural imset over  $V$ . Furthermore, let  $X$  be a collection of random variables indexed by  $V$  with probability measure  $P$  that admits density  $f(x)$  with respect to dominating  $\sigma$ -finite product measure  $\nu$ . If  $P$  has finite multiinformation  $m_P$ , then the following are equivalent:*

- i.  $\log f(x) = \log f(x) - \sum_{T \in \mathcal{P}(V)} u(T) \log f_T(x)$  for  $P$ -a.e.  $x \in \mathcal{X}$ ;
- ii.  $u^\top m_P = 0$ ;
- iii.  $A \perp\!\!\!\perp B \mid C [u] \Rightarrow A \perp\!\!\!\perp B \mid C [P]$  for every  $\langle A, B \mid C \rangle \in \mathcal{J}(V)$ .

Of course, along with their representation power comes complexity that makes practical use difficult. For that purpose, standard and characteristic imsets were developed.

### 3.5.4 Characteristic Imsets as Independence Models

Let  $V$  be a non-empty set of variables and  $\mathbf{P} = \mathcal{P}(V)$  be a poset ordered by inclusion.

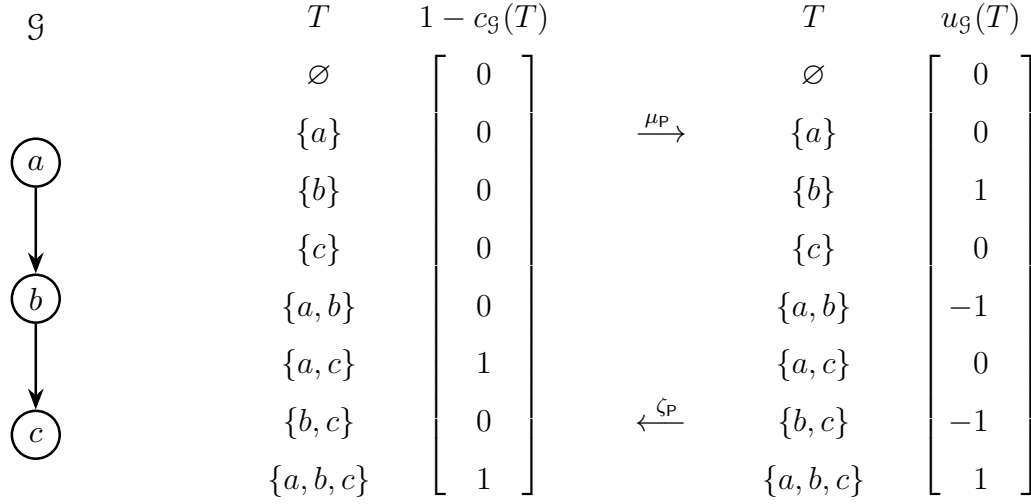


Figure 3.20: A DAG with vertices  $\{a, b, c\}$  and an application of the zeta and Möbius function of a poset  $\mathcal{P} = \mathcal{P}(V)$  ordered by inclusion as a transition between the standard and characteristic imsets of the DAG.

**Definition** (*standard imset*). Let  $\mathcal{G} = (V, E)$  be a DAG. The corresponding standard imset  $u_{\mathcal{G}}$  over  $V$  is defined as follows:

$$u_{\mathcal{G}} \equiv \delta_V - \delta_{\emptyset} + \sum_{a \in V} (\delta_{\text{pa}_{\mathcal{G}}(a)} - \delta_{\text{pa}_{\mathcal{G}}^+(a)}).$$

**Proposition 3.5.5** (Lemma 7.1 [83]). *Let  $\mathcal{G} = (V, E)$  be a DAG with standard imset  $u_{\mathcal{G}}$ .  $u_{\mathcal{G}}$  is a structural imset where  $\mathcal{J}(u_{\mathcal{G}}) = \mathcal{J}(\mathcal{G})$ .*

**Definition** (*characteristic imset*). Let  $\mathcal{G} = (V, E)$  be a DAG with standard imset  $u_{\mathcal{G}}$ . The corresponding characteristic imset is defined as follows:

$$c_{\mathcal{G}}(A) \equiv 1 - \sum_{A \subseteq T \subseteq V} u_{\mathcal{G}}(T) \quad \text{for all } A \subseteq V \ (|A| \geq 2).$$

Note that characteristic imsets are not defined on the empty set or singletons. However, if we let  $c_{\mathcal{G}}(A) = 1$  for all  $A \subseteq V$  ( $|A| \leq 1$ ), then by the Möbius inversion:



- i.  $u_{\mathcal{G}}(A) = \sum_{B \subseteq V (A \subseteq B)} (-1)^{|B \setminus A|} (1 - c_{\mathcal{G}}(B))$  for all  $A \subseteq V$ ;
- ii.  $1 - c_{\mathcal{G}}(A) = \sum_{B \subseteq V (A \subseteq B)} u_{\mathcal{G}}(B)$  for all  $A \subseteq V$ .

Accordingly, the following corollary follows from Theorem 3.5.1, Proposition 3.5.5, and Corollary 3.2.1.

**Corollary 3.5.1.** *Let  $\mathcal{G} = (V, E)$  be a DAG with standard imset  $u_{\mathcal{G}}$  and characteristic imset  $c_{\mathcal{G}}$ . Let  $\mathbf{P} = \mathcal{P}(V)$  be a poset ordered by inclusion. Furthermore, let  $X$  be a collection of random variables indexed by  $V$  with probability measure  $P$  that admits density  $f(x)$  with respect to dominating  $\sigma$ -finite product measure  $\nu$ . If  $P$  has finite multiinformation  $m_P$ , then the following are equivalent:*

- i.  $\log f(x) = \sum_{T \in \mathcal{P}(V)} \mu_{\mathbf{P}} c_{\mathcal{G}}(T) \log f_T(x)$  for  $P$ -a.e.  $x \in \mathcal{X}$ ;
- ii.  $u_{\mathcal{G}}^{\top} m_P = 0$ ;
- iii.  $A \perp\!\!\!\perp B \mid C [\mathcal{G}] \Rightarrow A \perp\!\!\!\perp B \mid C [P]$  for every  $\langle A, B \mid C \rangle \in \mathcal{T}(V)$ .

See [37, 87] for more details.

**Proposition 3.5.6** (Theorem 1 [37, 87]). *Let  $\mathcal{G} = (V, E)$  be a DAG and  $\leq$  be a total order consistent with  $\mathcal{G}$ . For all  $A \subseteq V$  ( $|A| \geq 2$ ):*

- i.  $c_{\mathcal{G}}(A) \in \{0, 1\}$ ;
- ii.  $c_{\mathcal{G}}(A) = 1 \Leftrightarrow A \subseteq \text{pa}_{\mathcal{G}}^+(\lceil A \rceil_{\leq})$ .

It follows that two DAGs  $\mathcal{G}$  and  $\mathcal{G}'$  are Markov equivalent if and only if  $c_{\mathcal{G}} = c_{\mathcal{G}'}$ .

There has been extensive work toward applying imsets to the problem of DAG learning [37, 83, 84, 85, 86, 87]. However, imsets have not been applied to learning maximal ancestral graphs. We explore this topic in Chapter 6.

## 4.0 Inducing Sets

In this chapter we introduce a new perspective for reasoning about ancestral graph Markov models, which is the primary contribution of this dissertation. Accordingly, we define the novel concept of an inducing set and the concept of an  $m$ -connecting set as a special case. While this chapter primarily focuses on ancestral graphs, especially those that are maximal, many of the forthcoming results may be applied to any family of stable mixed graphs. As we have seen earlier, all families of stable mixed graphs induce the same family of independence models; the focus on ancestral graphs is largely for theoretical convenience.

**Definition** (*inducing set*). Let  $\mathcal{G} = (V, E)$  be an ancestral graph containing disjoint sets  $M, L, S \subseteq V$  ( $M \neq \emptyset$ ).  $M$  is an inducing set relative to  $\langle L, S \rangle$  for  $\mathcal{G}$  if one of the following hold:

- i.*  $M$  is a singleton;
- ii.* there exists an inducing path between  $a$  and  $b$  relative to  $\langle L, MS \setminus \{a, b\} \rangle$  for all  $a, b \in M$  ( $a \neq b$ ).

If  $L = S = \emptyset$ , then  $M$  is a *primitive inducing set*.

Proposition 3.3.2 allows us to equivalently define a primitive inducing set in terms of  $m$ -connecting paths. Therefore, we adopt the term  *$m$ -connecting set* in place of primitive inducing set.

**Definition** ( *$m$ -connecting set*). Let  $\mathcal{G} = (V, E)$  be an ancestral graph containing a set  $M \subseteq V$  ( $M \neq \emptyset$ ).  $M$  is an  $m$ -connecting set for  $\mathcal{G}$  if one of the following hold:

- i.*  $M$  is a singleton;
- ii.* there exists an inducing path between  $a$  and  $b$  relative to  $\langle \emptyset, M \setminus \{a, b\} \rangle$  for all  $a, b \in M$  ( $a \neq b$ );
- iii.*  $a$  and  $b$  are not  $m$ -separated by  $C$  for all  $a, b \in M$  ( $a \neq b$ ) and all  $M \subseteq C \subseteq V$  ( $a, b \notin C$ ).

Proposition 3.3.2 implies that *(ii)* and *(iii)* are equivalent; they are included here to provide alternative definitions of  *$m$ -connecting set*.

Note that the concept of an  $m$ -connecting set can be extended to any family of stable mixed graphs using (iii). Let  $\mathcal{G} = (V, E)$  be a MAG. The set of all  $m$ -connecting sets for  $\mathcal{G}$  is denoted by  $\mathcal{M}(\mathcal{G})$ . Furthermore, the set of *non- $m$ -connecting sets* for  $\mathcal{G}$  are defined as the complement excluding the empty set and denoted by  $\mathcal{N}(\mathcal{G}) = \mathcal{P}_1(V) \setminus \mathcal{M}(\mathcal{G})$ .

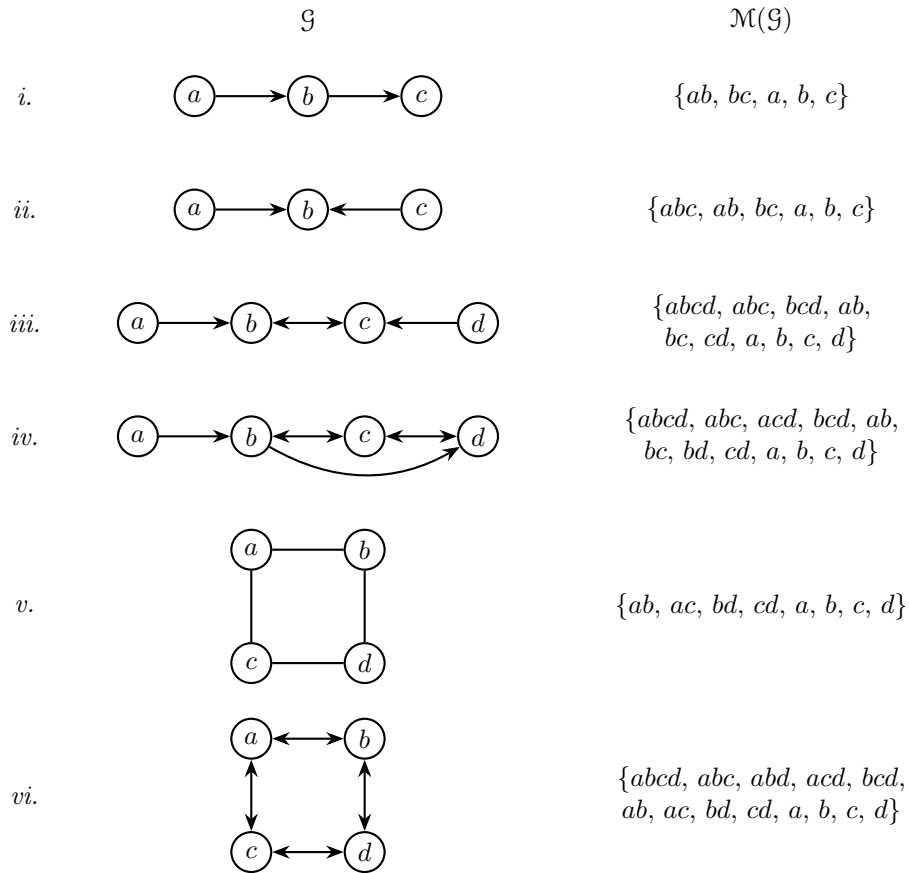


Figure 4.1: An illustration of various MAGs  $\mathcal{G}$  and their corresponding  $m$ -connecting sets  $\mathcal{M}(\mathcal{G})$ .

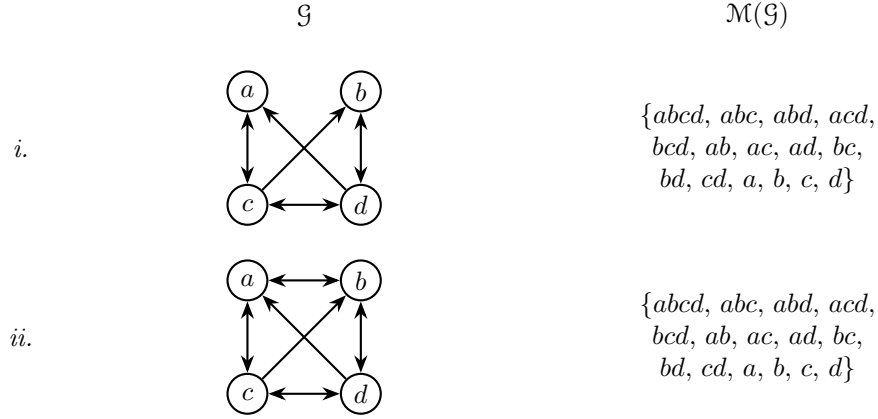


Figure 4.2: A comparison of two Markov equivalent ancestral graphs that are (i) not maximal and (ii) maximal, along with their corresponding  $m$ -connecting sets  $\mathcal{M}(\mathcal{G})$ ; their  $m$ -connecting sets are identical.

We make the following connection between  $m$ -connecting sets and collider-connecting sets. Lemma 4.0.1 shows that the set of maximal  $m$ -connecting sets and maximal collider-connecting sets are the same.

**Lemma 4.0.1.** *Let  $\mathcal{G} = (V, E)$  be a MAG containing a set  $M \subseteq V$ . For the following conditions  $i \Rightarrow ii$ :*

- i.  $M$  is an  $m$ -connecting set for  $\mathcal{G}$ ;
- ii.  $M$  is a collider-connecting set.

Furthermore, the following are equivalent:

- iii.  $M$  is a maximal  $m$ -connecting set for  $\mathcal{G}$ ;
- iv.  $M$  is a maximal collider-connecting set.

*Proof.* ( $i \Rightarrow ii$ ): If  $M$  is  $m$ -connecting for  $\mathcal{G}$ , then suppose by way of contradiction that  $M$  is not collider-connecting.  $M$  not collider-connecting: There exist vertices  $a, b \in M$  ( $a \neq b$ ) such that  $a$  and  $b$  are not collider-connecting.  $M$   $m$ -connecting for  $\mathcal{G}$ : inducing path  $\pi_{ab}$  between  $a$  and  $b$  relative to  $\langle L = \emptyset, M \setminus \{a, b\} \rangle$ .  $a$  and  $b$  are not collider-connecting:  $\pi_{ab}$  is

not collider-connecting. There exists a non-collider  $v \in V$  on  $\pi_{ab}$  such that  $v \notin L$ ; this is a contradiction. It follows that  $M$  is collider-connecting.

(iii  $\Rightarrow$  ii): This directly follows from the facts that every maximal  $m$ -connecting set is  $m$ -connecting and that every  $m$ -connecting set is collider-connecting.

(i  $\Leftarrow$  iv): If  $M$  is a maximal collider-connecting set, then for all  $a, b \in M$  ( $a \neq b$ ) there exists a collider-connecting path  $\pi_{ab}$  between  $a$  and  $b$  such that every vertex on  $\pi_{ab}$  is a member of  $M$ . It follows that every  $\pi_{ab}$  is inducing relative to  $\langle \emptyset, M \setminus \{a, b\} \rangle$ . Therefore,  $M$  is  $m$ -connecting for  $\mathcal{G}$ .

(iii  $\Leftrightarrow$  iv): We have that if  $M$  is a maximal  $m$ -connecting set for  $\mathcal{G}$ , then  $M$  is collider-connecting and that if  $M$  is a maximal collider-connecting set, then  $M$  is a maximal  $m$ -connecting set.

If  $M$  is a maximal  $m$ -connecting set, then  $M$  is a collider-connecting set. Suppose by way of contradiction that  $M$  is not a maximal collider-connecting set. It follows that there is a proper maximal collider-connecting superset of  $M$ . But every maximal collider-connecting set is  $m$ -connecting, so the super set is also  $m$ -connecting; this is a contradiction.

If  $M$  is a maximal collider-connecting set, then  $M$  is an  $m$ -connecting set. Suppose by way of contradiction that  $M$  is not a maximal  $m$ -connecting set. It follows that there is a proper maximal  $m$ -connecting superset of  $M$ . But every maximal  $m$ -connecting set is collider-connecting, so the super set is also collider-connecting; this is a contradiction.

Accordingly,  $M$  is a maximal  $m$ -connecting set if and only if  $M$  is a maximal collider-connecting set. □

## 4.1 Equivalence

In this section, we show that  $m$ -connecting sets may be used as an alternative representation of Markov equivalence for ancestral graphs. It follows that  $m$ -connecting sets equivalently characterize the independence models of ancestral graphs. We also show how

these sets relate to characteristic imsets and parametrizing sets.

#### 4.1.1 Characterization of Markov Equivalence

Theorem 3.3.7 characterizes Markov equivalence using adjacencies, unshielded colliders, and colliders at the end of discriminating paths. Accordingly, the forthcoming three lemmas address each of these points. Lemma 4.1.1 details the relation between  $m$ -connecting sets and adjacencies, Lemma 4.1.2 details the relation between  $m$ -connecting sets and unshielded colliders, and Lemma 4.1.3 details the relation between  $m$ -connecting sets and the colliders at the end of discriminating paths.

**Lemma 4.1.1.** *Let  $\mathcal{G} = (V, E)$  be a MAG containing vertices  $a, b \in V$ . The following are equivalent:*

- i.  $a$  and  $b$  are adjacent;*
- ii.  $M_{ab} \equiv \{a, b\}$  is  $m$ -connecting.*

*Proof.* ( $i \Rightarrow ii$ ): If  $a$  and  $b$  are adjacent, then there is a primitive inducing path between  $a$  and  $b$  because  $\mathcal{G}$  is maximal. Therefore,  $M_{ab}$  is  $m$ -connecting.

( $i \Leftarrow ii$ ): If  $M_{ab}$  is  $m$ -connecting set, then there is a primitive inducing path between  $a$  and  $b$ . Therefore,  $a$  and  $b$  are adjacent because  $\mathcal{G}$  is maximal. □

**Lemma 4.1.2.** *Let  $\mathcal{G} = (V, E)$  be a MAG with an unshielded triple  $\langle a, b, c \rangle$ . The following are equivalent:*

- i.  $b$  is a collider on  $\langle a, b, c \rangle$ ;*
- ii.  $M_{abc} \equiv \{a, b, c\}$  is  $m$ -connecting.*

*Proof.* ( $i \Rightarrow ii$ ): If  $b$  is a collider on  $\langle a, b, c \rangle$ , then:

- $\langle a, b \rangle$  is an inducing path between  $a$  and  $b$  relative to  $\langle \emptyset, c \rangle$ :  $a * \rightarrow b$ ;
- $\langle a, b, c \rangle$  is an inducing path between  $a$  and  $c$  relative to  $\langle \emptyset, b \rangle$ :  $a * \rightarrow b \leftarrow * c$ ;
- $\langle b, c \rangle$  is an inducing path between  $b$  and  $c$  relative to  $\langle \emptyset, a \rangle$ :  $b \leftarrow * c$ .

Therefore,  $M_{abc}$  is  $m$ -connecting.

( $i \Leftarrow ii$ ): If  $M_{abc}$  is  $m$ -connecting but  $a$  and  $c$  are not adjacent, then there exists an inducing path  $\pi$  between  $a$  and  $c$  relative to  $\langle \emptyset, b \rangle$  that is not inducing relative to  $\langle \emptyset, \emptyset \rangle$ . Accordingly, every collider on  $\pi$  is an ancestor of  $a$ ,  $b$ , or  $c$ . However, there exists a collider  $v \in V$  on  $\pi$  that is not an ancestor of  $a$  or  $c$ , otherwise,  $\pi$  would be inducing relative to  $\langle \emptyset, \emptyset \rangle$ . It follows that  $v$  is an ancestor of  $b$  and that  $b$  is not an ancestor of  $a$  or  $c$ ; if  $b$  was an ancestor of  $a$  or  $c$ , then  $v$  would also be an ancestor of  $a$  or  $c$ . Therefore,  $b$  is a collider on  $\langle a, b, c \rangle$ .  $\square$

**Lemma 4.1.3.** *Let  $\mathcal{G} = (V, E)$  be a MAG with a discriminating path  $\langle a, b_1, \dots, b_k, c, d \rangle$  ( $k \geq 1$ ) for  $c$ . The following are equivalent:*

- i.  $c$  is a collider on  $\langle b_k, c, d \rangle$ ;*
- ii.  $M_{acd} \equiv \{a, c, d\}$  is  $m$ -connecting.*

*Proof.* ( $i \Rightarrow ii$ ): If  $c$  is a collider on  $\langle b_k, c, d \rangle$ , then:

- $\langle a, b_1, \dots, b_k, c, d \rangle$  is an inducing path between  $a$  and  $d$  relative to  $\langle \emptyset, c \rangle$ :  
 $a * \rightarrow b_1 \leftrightarrow \dots \leftrightarrow b_k \leftrightarrow c \leftrightarrow d$  where  $b_i \rightarrow d$  for all  $1 \leq i \leq k$ —every collider on the path is an ancestor of  $\{c, d\}$ ;
- $\langle a, b_1, \dots, b_k, c \rangle$  is an inducing path between  $a$  and  $c$  relative to  $\langle \emptyset, d \rangle$ :  
 $a * \rightarrow b_1 \leftrightarrow \dots \leftrightarrow b_k \leftrightarrow c$  where  $b_i \rightarrow d$  for all  $1 \leq i \leq k$ —every collider on the path is an ancestor of  $d$ ;
- $\langle c, d \rangle$  is an inducing path between  $c$  and  $d$  relative to  $\langle \emptyset, a \rangle$ :  $c \leftrightarrow d$ .

Therefore,  $M_{acd}$  is  $m$ -connecting.

( $i \Leftarrow ii$ ): If  $M_{acd}$  is  $m$ -connecting but  $a$  and  $d$  are not adjacent, then there exists an inducing path  $\pi$  between  $a$  and  $d$  relative to  $\langle \emptyset, c \rangle$  that is not inducing relative to  $\langle \emptyset, \emptyset \rangle$ . Accordingly, every collider on  $\pi$  is an ancestor of  $a$ ,  $c$ , or  $d$ . However, there exists a collider  $v \in V$  on  $\pi$  that is not an ancestor of  $a$  or  $d$ , otherwise  $\pi$  would be inducing relative to  $\langle \emptyset, \emptyset \rangle$ . It follows that  $v$  is an ancestor of  $c$  and that  $c$  is not an ancestor of  $a$  or  $d$ ; if  $c$  was an ancestor of  $a$  or  $d$ , then  $v$  would also be an ancestor of  $a$  or  $d$ . Similarly,  $c$  is not an ancestor of  $b_k$  since  $b_k$  is a parent of  $d$ . Therefore,  $c$  is a collider on  $\langle b_k, c, d \rangle$ .  $\square$

Accordingly, in conjunction with Theorem 3.3.7, the preceding three lemmas may be used to characterize Markov equivalence.

**Theorem 4.1.1.** *Let  $\mathcal{G} = (V, E)$  and  $\mathcal{G}' = (V, E')$  be MAGs. The following are equivalent:*

- i.  $\mathcal{G}$  and  $\mathcal{G}'$  are Markov equivalent;*
- ii.  $\mathcal{G}$  and  $\mathcal{G}'$  have the same  $m$ -connecting sets.*

*Proof.* ( $i \Rightarrow ii$ ): If  $\mathcal{G}$  and  $\mathcal{G}'$  are Markov equivalent, then they have the same  $m$ -separations. It follows from the definition of  $m$ -connecting set (*iv*) that  $\mathcal{G}$  and  $\mathcal{G}'$  have the same  $m$ -connecting sets.

( $i \Leftarrow ii$ ): Lemma 4.1.1 implies that  $\mathcal{G}$  and  $\mathcal{G}'$  have the same adjacencies and, accordingly, the same unshielded triples. Lemma 4.1.2 implies that  $\mathcal{G}$  and  $\mathcal{G}'$  have the same unshielded colliders. Lemma 4.1.3 implies that if  $\pi$  forms a discriminating path for  $b$  in  $\mathcal{G}$  and  $\mathcal{G}'$ , then  $b$  is a collider on  $\pi$  in  $\mathcal{G}$  if and only if it is a collider on  $\pi$  in  $\mathcal{G}'$ . Theorem 3.3.7 implies that  $\mathcal{G}$  and  $\mathcal{G}'$  are Markov equivalent. □

An interesting takeaway is that the induced independence model of a MAG may be characterized by its  $m$ -connecting sets of cardinality two and three. Additionally, the sets of cardinality three can be further refined to those that have at least one and at most two subsets of cardinality two that are  $m$ -connecting. This is an important result used for quickly defining Markov equivalence with parametrizing sets [38]. This characterization of equivalence may be straightforwardly extended to any family of stable mixed graphs by noting that all families of stable mixed graphs induce the same family of independence models [73, 74]. Since the  $m$ -connecting sets of a graph can be defined directly from the induced independence model of the graph the result is immediate.

## 4.2 Relation to Other Work

In this section, we discuss how the ideas presented in this dissertation relate to previous works. Similar ideas have been explore independently by other authors.



### 4.2.1 Parametrizing Sets and Characteristic Imsets

Hu and Evan’s work on parametrizing sets [38] is closely related our work on  $m$ -connecting sets. These sets were developed concurrently with this work and published during the synthesis of this dissertation. Parametrizing sets are identical to  $m$ -connecting sets, but are defined using the heads and tails of an ADMG explicitly for the purpose of characterizing Markov equivalence. Additionally, for those familiar with the work of Hu and Evans, Lemmas 4.1.1, 4.1.2, and 4.1.3 achieve the same result as Proposition 3.4 in [38].

**Proposition 4.2.1.** *Let  $\mathcal{G} = (V, E)$  be a MAG containing a set  $M \subseteq V$ . The following are equivalent:*

- i.  $M$  is  $m$ -connecting for  $\mathcal{G}$ ;*
- ii.  $M$  is a parametrizing set of  $\mathcal{G}$ .*

*Proof.* The proof directly follows from Proposition 3.3.5. □

Studený et al.’s work on characteristic imsets [87, 37] is closely related to our work on  $m$ -connecting sets. These imsets have only been defined for independence models induced by DAGs, but for these models  $m$ -connecting sets and characteristic imsets are nearly identical. To facilitate this comparison, note that a set of sets can be represented by an identifier imset for that set of sets. With this, the only difference is that characteristic imsets are not defined for singletons; singletons are trivially  $m$ -connecting.

**Proposition 4.2.2.** *Let  $\mathcal{G} = (V, E)$  be a DAG containing a set  $M \subseteq V$  ( $|M| \geq 2$ ) and  $\leq$  be a total order consistent with  $\mathcal{G}$ . The following are equivalent:*

- i.  $M$  is an  $m$ -connecting set for  $\mathcal{G}$ ;*
- ii. the characteristic imset  $c_{\mathcal{G}}(M) = 1$ .*

*Proof.* ( $i \Rightarrow ii$ ): If  $M$  is  $m$ -connecting, then by Lemma 4.0.1  $M$  is a collider-connecting set. In a DAG this is only possible if  $M \setminus \lceil M \rceil_{\leq} \subseteq \text{pa}_{\mathcal{G}}(\lceil M \rceil_{\leq})$ . Therefore, by Proposition 3.5.6  $c_{\mathcal{G}}(M) = 1$ .

( $i \Leftarrow ii$ ): If  $c_{\mathcal{G}}(M) = 1$ , then by Proposition 3.5.6  $M \setminus \lceil M \rceil_{\leq} \subseteq \text{pa}_{\mathcal{G}}(\lceil M \rceil_{\leq})$ . Accordingly, there exist  $m$ -connecting paths between the members of  $M \setminus \lceil M \rceil_{\leq}$  and  $\lceil M \rceil_{\leq}$ . Further-

more, there exist  $m$ -connecting paths between all members of  $M \setminus \lceil M \rceil_{\leq}$  relative to  $\lceil M \rceil_{\leq}$ . Therefore,  $M \in \mathcal{M}(\mathcal{G})$ .  $\square$

#### 4.2.2 The Causal Inference Algorithm

The causal inference (CI) algorithm recovers a PAG  $\mathcal{G}$  that represents a Markov equivalence class of MAGs by querying a conditional independence oracle  $\mathcal{J}$  [79]; the CI algorithm is detailed in Appendix B.2. Algorithm 1 outlines a modified version of the CI algorithm that replaces the queries to a conditional independence oracle with queries to an  $m$ -connecting set oracle  $\mathcal{M}$ . This modification directly follows from Lemmas 4.1.1, 4.1.2, and 4.1.3. Algorithm 1 provides a procedure to reconstruct a MAG up to its Markov equivalence class from its  $m$ -connecting sets.

---

#### Algorithm 1: CAUSAL INFERENCE FROM M-CONNECTING SETS CIM( $\mathcal{M}$ )

---

**Input:**  $m$ -connecting sets:  $\mathcal{M}$   
**Output:** partial ancestral graph:  $\mathcal{G}$

- 1 Let  $\mathcal{G} = (V, E)$  where  $E = \{a \circ\text{-} \circ b \mid a, b \in V\}$  ;
- 2 **foreach** edge  $a \circ\text{-} \circ b \in E$  **do**
- 3     **if**  $\{a, b\} \notin \mathcal{M}$  **then**
- 4         Remove  $a \circ\text{-} \circ b$  from  $E$  ;
- 5     **end**
- 6 **end**
- 7 **foreach** unshielded triple  $\langle a, b, c \rangle$  in  $\mathcal{G}$  **do**
- 8     Rule 0: If  $\{a, b, c\} \in \mathcal{M}$ , then orient it as a collider  $a * \rightarrow b \leftarrow * c$  ;
- 9 **end**
- 10 **repeat**
- 11     Rule 1: If  $a * \rightarrow b \circ\text{-} * c$ , and  $a$  and  $c$  are not adjacent, then orient the triple as  $a * \rightarrow b \rightarrow c$  ;
- 12     Rule 2: If  $a \rightarrow b * \rightarrow c$  or  $a * \rightarrow b \rightarrow c$ , and  $a * \text{-} \circ c$ , then orient  $a * \text{-} \circ c$  as  $a \circ \rightarrow c$  ;
- 13     Rule 3: If  $a * \rightarrow b \leftarrow * c$ ,  $a * \text{-} \circ d \circ\text{-} * c$ ,  $a$  and  $c$  are not adjacent, and  $d * \text{-} \circ b$ , then orient  $d * \text{-} \circ b$  as  $d * \rightarrow b$  ;
- 14     Rule 4: If  $\langle a, \dots, b, c, d \rangle$  is a discriminating path from  $a$  to  $d$  for  $c$  and  $c \circ\text{-} * d$ , then: if  $\{a, c, d\} \notin \mathcal{M}$ , then orient  $c \circ\text{-} * d$  as  $c \rightarrow d$ ; otherwise orient the triple  $\langle b, c, d \rangle$  as  $b \leftrightarrow c \leftrightarrow d$  ;
- 15 **until** Rules 1 - 4 no longer apply;

---

### 4.3 Factorization

In this section we present one of the main results of this dissertation: a factorization criterion for the log density of a probability measure. The factorization criterion is derived from the  $m$ -connecting sets of a directed MAG for a probability measure and is equivalent to the probability measure satisfying the global Markov property with respect to that MAG. The general proof strategy uses an algorithm to construct the primary and secondary imsets out of the non- $m$ -connecting sets; see Algorithm 3. Applying the Möbius inversion to the primary imset yields a structural imset that induces the same independence model as the directed MAG. The secondary imset is incorporated into the factorization as an adjustment term. Ultimately we show: (i) the factorization criterion holding implies that the dot product of the structural imset with the multiinformation of the probability measure equals zero; (ii) the dot product of the structural imset with the multiinformation of the probability measure equaling zero implies that the global Markov property holds; and (iii) the global Markov property holding implies that the factorization criterion holds.

To facilitate the forthcoming discussion, we define several new terms. Let  $V$  be a non-empty set of variables. Furthermore, let  $X$  be a collection of random variables indexed by  $V$  with probability measure  $P$  that admits density  $f(x)$  with respect to dominating  $\sigma$ -finite product measure  $\nu$ . We define a function  $\phi_A : \mathcal{X}_A \rightarrow \mathbb{R}$  as a linear combination of log density terms motivated by the Möbius inversion.

$$\phi_A(x) = \sum_{B \subseteq A} (-1)^{|A \setminus B|} \log f_B(x) \qquad \log f_A(x) = \sum_{B \subseteq A} \phi_B(x)$$

The expectation of  $\phi_A(x)$  with respect to  $P$  has been previously studied in the field of information theory by several researchers including McGill, who coined the term interaction information [53]. Accordingly, we call  $\phi_A(x)$  the interaction information rate.

We provide an analogous term for a non-empty set of sets. Let  $\mathbf{A} \subseteq \mathcal{P}(V)$  be a set of sets:

$$\phi_{\mathbf{A}}(x) \equiv \sum_{T \in \mathbf{A}} \phi_T(x)$$

We define the following case for sets of sets and shorthand for the corresponding  $\phi$  term. Let

$A, B \subseteq V$  ( $A \neq \emptyset$ ) be disjoint sets:

$$M_{A|B} \equiv \bigcup_{\substack{T \subseteq AB \\ A \subseteq T}} \{T\} \quad \phi_{A|B}(x) \equiv \phi_{M_{A|B}}(x) \quad \delta_{A|B} \equiv \delta_{M_{A|B}}$$

Similar to above, we call  $\phi_{A|B}(x)$  the conditional interaction information rate.

$$\begin{aligned} \phi_{A|B}(x) &= \sum_{\substack{T \subseteq AB \\ A \subseteq T}} \phi_T(x) \\ &= \sum_{\substack{T \subseteq AB \\ B \subseteq T}} (-1)^{|AB \setminus T|} \log f_T(x) \end{aligned}$$

Another case is when the set of sets corresponds to a semi-elementary imset that has been transformed by the Möbius inversion. Let  $A, B, C \subseteq V$  ( $AB \neq \emptyset$ ) be disjoint sets.

$$N_{A,B|C} \equiv \bigcup_{\substack{T \subseteq ABC \\ T \not\subseteq AC \\ T \not\subseteq BC}} \{T\} \quad \phi_{A,B|C}(x) \equiv \phi_{N_{A,B|C}}(x) \quad \delta_{A,B|C} \equiv \delta_{N_{A,B|C}}$$

The expectation of  $\phi_{A,B|C}(x)$  with respect to  $P$  is the well-known information theoretic concept of mutual information. Accordingly, we call  $\phi_{A,B|C}(x)$  the mutual information rate.

The mutual information rate corresponds to the imsets constructed by Algorithm 3. Additionally, these terms are closely related to conditional independence. Let  $A, B, C \subseteq V$  ( $AB \neq \emptyset$ ) be disjoint sets.

$$\begin{aligned} \phi_{A,B|C}(x) &= \sum_{\substack{T \subseteq ABC \\ T \not\subseteq AC \\ T \not\subseteq BC}} \phi_T(x) \\ &= \sum_{T \subseteq ABC} \phi_T(x) + \sum_{T \subseteq C} \phi_T(x) - \sum_{T \subseteq AC} \phi_T(x) - \sum_{T \subseteq BC} \phi_T(x) \\ &= \log f_{ABC}(x) + \log f_C(x) - \log f_{AC}(x) - \log f_{BC}(x). \end{aligned}$$

$$A \perp\!\!\!\perp B \mid C [P] \Leftrightarrow \phi_{A,B|C}(x) = 0 \quad \text{for } P\text{-a.e. } x \in \mathcal{X} \quad (4.1)$$

This relation can be expressed more generally using imsets. Let  $\mathbf{P} = \mathcal{P}(V)$  be a poset ordered by inclusion and note that  $\zeta_{\mathbf{P}}\delta_A = \sum_{T \subseteq A} \delta_T$ .

$$\begin{aligned}
\delta_{A,B|C} &= \sum_{\substack{T \subseteq ABC \\ T \not\subseteq AC \\ T \not\subseteq BC}} \delta_T \\
&= \sum_{T \subseteq ABC} \delta_T + \sum_{T \subseteq C} \delta_T - \sum_{T \subseteq AC} \delta_T - \sum_{T \subseteq BC} \delta_T \\
&= \zeta_{\mathbf{P}}\delta_{ABC} + \zeta_{\mathbf{P}}\delta_C - \zeta_{\mathbf{P}}\delta_{AC} - \zeta_{\mathbf{P}}\delta_{BC} \\
&= \zeta_{\mathbf{P}}[\delta_{ABC} + \delta_C - \delta_{AC} - \delta_{BC}] \\
&= \zeta_{\mathbf{P}}u_{\langle A,B|C \rangle}.
\end{aligned}$$

Accordingly,

$$u_{\langle A,B|C \rangle} = \mu_{\mathbf{P}}\delta_{A,B|C}.$$

$$A \perp\!\!\!\perp B \mid C [P] \quad \Leftrightarrow \quad (\mu_{\mathbf{P}}\delta_{A,B|C})^\top m_P = 0 \quad (4.2)$$

The non- $m$ -connecting set terms constructed by the Algorithm 3 are exactly the non- $m$ -connecting sets for a directed MAG, and we use their correspondence to conditional independence in a probability measure to show the equivalence between the factorization and the global Markov property.

### 4.3.1 Preliminaries

To facilitate the forthcoming proofs, we introduce the concept of constrained subsets.

**Definition** (*constrained subsets*). Let  $V$  be a non-empty set of variables containing sets  $A, B \subseteq V$ . Let  $\mathbf{R} \subseteq \mathcal{P}(V)$  be a set of sets. The subset operator applied to  $A$  with respect to  $B$  constrained by  $\mathbf{R}$ , denoted by  $A \subseteq_{\mathbf{R}} B$ , is the conjunction:

- i.*  $A \subseteq B$ ;
- ii.*  $A \in \mathbf{R}$ .

Let  $b \in V$  be a variable. The subset operator applied to  $A$  with respect to  $B$  constrained by  $b$ , denoted by  $A \subseteq^b B$ , is the conjunction:

- i.  $A \subseteq B$ ;
- ii.  $b \in A$ .

The subset operator applied to  $A$  with respect to  $B$  constrained by  $R$  and  $b$ , denoted by  $A \subseteq_R^b B$ , is the conjunction:

- i.  $A \subseteq B$ ;
- ii.  $b \in A \in R$ .

Additionally, a *maximal constrained subset*, denoted by  $A \in [B]_R^b$ , is a maximal set satisfying  $A \subseteq_R^b B$ .

Proposition 4.3.1 shows that the induced subgraph of a MAG over an anterior set is a MAG and induces an independence subset over the shared variables.

**Proposition 4.3.1.** *Let  $\mathcal{G} = (V, E)$  be a MAG containing a set  $A \subseteq V$ . If  $A$  is an anterior set, then the induced subgraph  $\mathcal{G}_A$  is a MAG and:*

$$\mathcal{J}(\mathcal{G}_A) = \{\langle A, B \mid C \rangle \in \mathcal{T}(A) ; \quad \langle A, B \mid C \rangle \in \mathcal{J}(\mathcal{G})\}.$$

*Proof.* By Proposition 3.3.4,  $\mathcal{G}_A$  is a MAG and by Proposition 3.4.2  $\mathcal{J}(\mathcal{G}_A) = \{\langle A, B \mid C \rangle \in \mathcal{T}(A) ; \quad \langle A, B \mid C \rangle \in \mathcal{J}(\mathcal{G})\}$ . □

**Corollary 4.3.1.** *Let  $\mathcal{G} = (V, E)$  be a directed MAG containing a set  $A \subseteq V$ . If  $A$  is an ancestral set, then the induced subgraph  $\mathcal{G}_A$  is a directed MAG and:*

$$\mathcal{J}(\mathcal{G}_A) = \{\langle A, B \mid C \rangle \in \mathcal{T}(A) ; \quad \langle A, B \mid C \rangle \in \mathcal{J}(\mathcal{G})\}.$$

*Proof.* The proof immediately follows from Propositions 3.4.1 and 4.3.1. □

Lemma 4.3.1 shows that the  $m$ -connecting sets of the induced subgraph of a MAG over an anterior set is the induced set of  $m$ -connecting sets. That is, for an ancestral subset  $A \subseteq V$ ,  $\mathcal{M}(\mathcal{G}_A)$  is the set of  $m$ -connecting sets containing every  $m$ -connecting set present in  $\mathcal{M}(\mathcal{G})$  over the members of  $A$ .

**Lemma 4.3.1.** *Let  $\mathcal{G} = (V, E)$  be a MAG containing an anterior subset  $A \subseteq V$ . If  $M \subseteq A$ , then the following are equivalent:*

- i.  $M$  is  $m$ -connecting for  $\mathcal{G}$ ;
- ii.  $M$  is  $m$ -connecting for  $\mathcal{G}_A$ .

*Proof.* The proof immediately follows from Proposition 3.4.2 and the definitions of marginalization and  $m$ -connecting set.  $\square$

Lemma 4.3.2 shows that barren vertices have a unique maximal collider-connecting set.

**Lemma 4.3.2.** *Let  $\mathcal{G} = (V, E)$  be a directed MAG containing a vertex  $b \in \text{bar}_{\mathcal{G}}(V)$ . There is exactly one maximal collider-connecting set containing  $b$ .*

*Proof.* Let  $A, C \in [V]_{\text{col}_{\mathcal{G}}(b)}^b$  be maximal collider-connecting sets containing  $b$  and note that  $|A| = |C|$ . If  $|A| = |C| = 1$ , then  $A = C = \{b\}$  and there is exactly one maximal collider-connecting set.

If  $|A| = |C| > 1$ , then for all  $a \in A$  ( $a \neq b$ ) and all  $c \in C$  ( $c \neq b$ ), there exists a collider path  $\pi_{ab}$  between  $a$  and  $b$  and a collider path  $\pi_{bc}$  between  $b$  and  $c$ . In what follows, we show that  $a$  and  $c$  are collider-connecting; if  $a = c$  this is trivial.

Construct a path  $\pi_{ac}$  as follows. Traverse  $\pi_{ab}$  from  $a$  to  $b$  until reaching a vertex  $v \in V$  such that  $v$  is on  $\pi_{bc}$ . Let  $\pi_{av}$  be the subpath of  $\pi_{ab}$  between  $a$  and  $v$ . Similarly, traverse  $\pi_{bc}$  from  $v$  to  $c$ . Let  $\pi_{vc}$  be the subpath of  $\pi_{bc}$  between  $v$  and  $c$ . Then  $\pi_{ac}$  is the path formed by concatenating  $\pi_{av}$  and  $\pi_{vc}$ .

If  $v = b$ , then  $v$  is a collider on  $\pi_{ac}$  since  $b \in \text{bar}_{\mathcal{G}}(V)$ . If  $v \neq b$ , then  $v$  is a collider on  $\pi_{ab}$  and  $\pi_{bc}$ . It follows that  $v$  is a collider on  $\pi_{ac}$ . Therefore  $a$  and  $c$  are collider-connecting. Since every  $a \in A$  and  $c \in C$  are collider-connecting,  $A = C$  and there is exactly one maximal collider-connecting set containing  $v$ .  $\square$

Corollary 4.3.2 shows that barren vertices have a unique maximal  $m$ -connecting set. It is worth noting that the unique maximal  $m$ -connecting set of a vertex is also the unique maximal collider-connecting set for that vertex.

**Corollary 4.3.2.** *Let  $\mathcal{G} = (V, E)$  be a directed MAG containing a vertex  $b \in \text{bar}_{\mathcal{G}}(V)$ . There is exactly one maximal  $m$ -connecting set containing  $b$ .*

*Proof.* The proof immediately follows from Lemmas 4.0.1 and 4.3.2.  $\square$

Lemma 4.3.3 shows that  $m$ -connecting sets may be characterized by the existence of inducing paths between a barren vertex and the other vertices in the set.

**Lemma 4.3.3.** *Let  $\mathcal{G} = (V, E)$  be a directed MAG containing a set  $M \subseteq V$  and let  $L = V \setminus M$ . If  $b \in \text{bar}_{\mathcal{G}}(M)$ , then the following are equivalent:*

- i. there exists an inducing path between  $a$  and  $b$  relative to  $\langle L, M \setminus \{a, b\} \rangle$  for all  $a \in M \setminus b$ ;*
- ii.  $M$  is  $m$ -connecting for  $\mathcal{G}$ .*

*Proof.* ( $i \Rightarrow ii$ ): Suppose by way of contradiction that  $M$  is not  $m$ -connecting for  $\mathcal{G}$ . Then there exists  $a, c \in M \setminus b$  ( $a \neq c$ ) such that there is no inducing path between  $a$  and  $c$  relative to  $\langle L, M \setminus \{a, c\} \rangle$ . However, there exists an inducing path  $\pi_{ab}$  between  $a$  and  $b$  relative to  $\langle L, M \setminus \{a, b\} \rangle$  and an inducing path  $\pi_{bc}$  between  $b$  and  $c$  relative to  $\langle L, M \setminus \{b, c\} \rangle$ . Construct the path  $\pi_{ac}$  by traversing  $\pi_{ab}$  from  $a$  to  $b$  until reaching some  $d \in \pi_{bc}$  then traversing  $\pi_{bc}$  from  $d$  to  $c$ .

Note the status of every non-endpoint vertex on  $\pi_{ac}$ . In particular, check if each non-endpoint vertex is a non-collider on  $\pi_{ac}$  and member of  $L$ , a collider on  $\pi_{ac}$  and an ancestor of  $M$ , or neither. By construction, every non-endpoint vertex on  $\pi_{ac}$  has the same status as on  $\pi_{ab}$  and  $\pi_{bc}$  except for  $d$ . Therefore, all non-endpoint vertices other than  $d$  satisfy the criteria required for  $\pi_{ac}$  to be inducing relative to  $\langle L, M \setminus \{a, c\} \rangle$ .

Accordingly, we consider the possible scenarios for  $d$ . If  $d = b$ , then  $d$  is a collider on  $\pi_{ac}$  and a trivial ancestor of  $M$  since  $b \in \text{bar}_{\mathcal{G}}(M)$ . If  $d \neq b$  is a non-collider on  $\pi_{ac}$ , then  $d$  is a non-collider on  $\pi_{ab}$  or  $\pi_{bc}$  and  $d \in L$ . If  $d \neq b$  is a collider on  $\pi_{ac}$  and  $d$  is a collider on  $\pi_{ab}$  or  $\pi_{bc}$ , then  $d \in \text{ang}_{\mathcal{G}}(M)$ . If  $d$  is a collider on  $\pi_{ac}$  and a non-collider on  $\pi_{ab}$  and  $\pi_{bc}$ , then  $d$  is an ancestor of  $a, c$ , or a collider on  $\pi_{ac}$ ; accordingly  $d \in \text{ang}_{\mathcal{G}}(M)$ .

Therefore,  $d$  satisfies the criteria required for  $\pi_{ac}$  to be inducing relative to  $\langle L, M \setminus \{a, c\} \rangle$ . The path  $\pi_{ac}$  is inducing for  $\langle L, M \setminus \{a, c\} \rangle$ ; this is a contradiction.

( $i \Leftarrow ii$ ): This is trivial by the definition of  $m$ -connecting set. □

Algorithm 3 uses a helper algorithm to construct pairs of  $m$ -connecting and non- $m$ -connecting sets; see Algorithm 2.



---

**Algorithm 2:** PAIRS( $\mathcal{G}, b$ )

---

**Input:** directed MAG:  $\mathcal{G} = (V, E)$ , barren vertex:  $b \in \text{bar}_{\mathcal{G}}(V)$

**Output:** ordered lists:  $\mathbf{M}^{\mathcal{G},b}$ ,  $\mathbf{N}^{\mathcal{G},b}$

1 Initialize ordered lists  $\mathbf{M}^{\mathcal{G},b} = \langle \rangle$  and  $\mathbf{N}^{\mathcal{G},b} = \langle \rangle$ ;

2 Let  $\mathbf{R} = \{N ; N \subseteq_{\mathcal{N}(\mathcal{G})}^b V\}$  ;

3 **repeat**

4     Pick  $N \in [V]_{\mathbf{R}}^b$  and  $M \in [N]_{\mathcal{M}(\mathcal{G})}^b$  ;

5     Append  $N$  to  $\mathbf{N}^{\mathcal{G},b}$  and  $M$  to  $\mathbf{M}^{\mathcal{G},b}$  ;

6     **foreach**  $T \subseteq N$  **do**

7         **if**  $b \in T$  and  $T \not\subseteq M$  **then**

8             Remove  $T$  from  $\mathbf{R}$  ;

9         **end**

10     **end**

11 **until**  $\mathbf{R} = \emptyset$ ;

---

Algorithm 2 requires several new concepts. Accordingly, we define the following notation. Let  $\mathcal{G} = (V, E)$  be a directed MAG,  $\mathbf{P} = \mathcal{P}(V)$  be a poset ordered by inclusion, and  $b \in \text{bar}_{\mathcal{G}}(V)$  be a barren vertex. Additionally, we use  $\mathbf{M}^{\mathcal{G},b}$  and  $\mathbf{N}^{\mathcal{G},b}$  to define the ordered lists output by Algorithm 2. These are ordered lists of  $m$ -connecting, non- $m$ -connecting sets respectively:

$$\mathbf{M}^{\mathcal{G},b} = \langle M_1^{\mathcal{G},b}, \dots, M_n^{\mathcal{G},b} \rangle \quad \mathbf{N}^{\mathcal{G},b} = \langle N_1^{\mathcal{G},b}, \dots, N_n^{\mathcal{G},b} \rangle$$

where  $n = |\mathbf{M}^{\mathcal{G},b}|$ .

Additionally, we define the restricted universe of sets with respect to  $N_i^{\mathcal{G},b}$  and  $b$ :

$$\mathbf{U}_i^{\mathcal{G},b} \equiv \bigcup_{\substack{T \subseteq N_i^{\mathcal{G},b} \\ b \in T}} \{T\}.$$

We simplify notation and use  $\mathbf{N}_{i,i}^{\mathcal{G},b}$  to define sets of sets that corresponds to the conditional independence statement  $b \perp\!\!\!\perp N_i^{\mathcal{G},b} \setminus M_i^{\mathcal{G},b} \mid M_i^{\mathcal{G},b} \setminus b$ . Let  $A = b$ ,  $B = N_i^{\mathcal{G},b} \setminus M_i^{\mathcal{G},b}$ , and

$C = M_i^{\mathcal{G},b} \setminus b$ , then

$$\begin{aligned}
ABC &= (N_i^{\mathcal{G},b} \setminus M_i^{\mathcal{G},b}) \cup (M_i^{\mathcal{G},b} \setminus b) \cup b \\
&= N_i^{\mathcal{G},b} \\
AC &= (M_i^{\mathcal{G},b} \setminus b) \cup b \\
&= M_i^{\mathcal{G},b} \\
BC &= (N_i^{\mathcal{G},b} \setminus M_i^{\mathcal{G},b}) \cup (M_i^{\mathcal{G},b} \setminus b) \\
&= N_i^{\mathcal{G},b} \setminus b.
\end{aligned}$$

Therefore, define

$$\mathbf{N}_{i,i}^{\mathcal{G},b} \equiv \bigcup_{\substack{T \in \mathbf{U}_i^{\mathcal{G},b} \\ T \not\subseteq M_i^{\mathcal{G},b}}} \{T\} = \bigcup_{\substack{T \subseteq N_i^{\mathcal{G},b} \\ b \in T \\ T \not\subseteq M_i^{\mathcal{G},b}}} \{T\} = \bigcup_{\substack{T \subseteq ABC \\ T \not\subseteq AC \\ T \not\subseteq BC}} \{T\}.$$

Accordingly

$$\delta_{A,B|C} = \delta_{\mathbf{N}_{i,i}^{\mathcal{G},b}} \quad \text{and} \quad u_{\langle A,B|C \rangle} = \mu_{\mathbf{P}} \delta_{\mathbf{N}_{i,i}^{\mathcal{G},b}}.$$

Lemma 4.3.4 states that the non- $m$ -connecting sets constructed at each step of Algorithm 3 are the non- $m$ -connecting sets containing  $b$  that have not yet been accounted for in  $\mathcal{N}(\mathcal{G})$ .

**Lemma 4.3.4.** *Let  $\mathcal{G} = (V, E)$  be a directed MAG containing vertex  $b \in V$  with preceding vertices  $R = \text{pre}_{\mathcal{G}}^{\leq}(b)$  and ancestral set  $A \in \mathcal{A}(\mathcal{G})$  such that  $b \in A \subseteq R$ . If  $\mathbf{M}^{\mathcal{G}_A,b}, \mathbf{N}^{\mathcal{G}_A,b} = \text{PAIRS}(\mathcal{G}_A, b)$  are the ordered lists constructed by Algorithm 2, then*

$$\bigcup_{i=1}^n \mathbf{N}_{i,i}^{\mathcal{G}_A,b} = \{T \subseteq_{\mathcal{N}(\mathcal{G})}^b A\}.$$

*Proof.* By Corollary 4.3.1  $\mathcal{G}_A$  is a directed MAG and by Lemma 4.3.1  $\{T \subseteq_{\mathcal{N}(\mathcal{G}),b} A\} = \{T \subseteq_{\mathcal{N}(\mathcal{G}_A),b} A\}$ . Let  $T \subseteq_{\mathcal{N}(\mathcal{G}_A),b} A$  be a non- $m$ -connecting subset of  $A$  containing  $b$  and suppose by way of contradiction that  $T \notin N_{i,i}^{\mathcal{G}_A,b}$  for any  $1 \leq i \leq n$ .

Note that  $T \subseteq N_i^{\mathcal{G}_A,b}$  for some  $1 \leq i \leq n$  since  $N_1^{\mathcal{G}_A,b} = A$ . Pick  $i$  such that  $T \subseteq N_i^{\mathcal{G}_A,b}$ . If  $T \not\subseteq M_i^{\mathcal{G}_A,b}$ , then  $T \in N_{i,i}^{\mathcal{G}_A,b}$ ; this is a contradiction. Otherwise, there exists  $T \subseteq N_j^{\mathcal{G}_A,b} \subset M_i^{\mathcal{G}_A,b}$  for some  $N_j^{\mathcal{G}_A,b} \in \mathbf{N}^{\mathcal{G}_A,b}$  by maximality. Repeat this logic until  $T \in N_{j,j}^{\mathcal{G}_A,b}$  or  $M_j^{\mathcal{G}_A,b}$  has no maximal non- $m$ -connecting subsets; the latter is a contradiction.

Thus, there exists  $1 \leq i \leq n$  such that  $T \in N_{i,i}^{\mathcal{G}_A,b}$  and  $\bigcup_{i=1}^n N_{i,i}^{\mathcal{G}_A,b} = \{T \subseteq_{\mathcal{N}(\mathcal{G}),b} A\}$ .  $\square$

Let  $\mathcal{G} = (V, E)$  be a directed MAG and  $\leq$  be a total order consistent with  $\mathcal{G}$ . Accordingly, the set of sets  $\mathcal{N}(\mathcal{G})$  is the set of all non- $m$ -connecting sets for  $\mathcal{G}$  and the imset  $\delta_{\mathcal{N}(\mathcal{G})}$  is the identifier of  $\mathcal{N}(\mathcal{G})$ . Algorithm 3 characterizes  $\delta_{\mathcal{N}(\mathcal{G})}$  as a linear combination of imsets whose Möbius inversions are semi-elementary imsets. Two new imsets are subsequently constructed from the linear combination imsets—one is the sum of the (absolute) positive components and the other is the sum of the (absolute) negative components. We call these imsets the primary and secondary imsets respectively. It follows that the Möbius inversion of the newly constructed imsets are structural imsets and induce semi-graphoids. Notably, the primary imset induces the same independence model as  $\mathcal{G}$  but is not part of the factorization, while the secondary imset induces a strict independence subset but is part of the factorization.

Algorithm 3 begins by defining a set  $R$  as the set of all variables  $V$ . As the algorithm loops, variables are removed one at a time and  $R$  contains the remaining variables. A vertex  $b$  is selected to be removed from the remaining vertices  $R$  where  $b$  is the last vertex according to  $\leq$ . Algorithm 2 is called to construct ordered lists  $\mathbf{M}^{\mathcal{G},b}$  and  $\mathbf{N}^{\mathcal{G},b}$ . The ordered lists  $\mathbf{M}^{\mathcal{G},b}$  and  $\mathbf{N}^{\mathcal{G},b}$  contain  $m$ -connecting and non- $m$ -connecting sets respectively; all sets contain  $b$ . By Corollary 4.3.2, each set  $N_i^{\mathcal{G},b}$  in  $\mathbf{N}^{\mathcal{G},b}$  has exactly one unique maximal  $m$ -connecting subset  $M_i^{\mathcal{G},b}$  that contains  $b$ . Accordingly, we construct pairs of non- $m$ -connecting and  $m$ -connecting sets by adding  $M_i^{\mathcal{G},b}$  and  $N_i^{\mathcal{G},b}$  terms to  $\mathbf{M}^{\mathcal{G},b}$  and  $\mathbf{N}^{\mathcal{G},b}$  respectively where each  $N_i^{\mathcal{G},b}$  is paired with the corresponding  $M_i^{\mathcal{G},b}$ .

In each loop on Algorithm 2, we pick a maximal non- $m$ -connecting set  $N_i^{\mathcal{G},b}$  that contains  $b$  from  $R$  and the previously described pairing process is repeated. All subsets of  $N_i^{\mathcal{G},b}$  and supersets of  $M_i^{\mathcal{G},b}$  are removed from  $R$ .

New  $m$ -connecting and non- $m$ -connecting sets are added to  $\mathbf{M}^{\mathcal{G},b}$  and  $\mathbf{N}^{\mathcal{G},b}$  respectively using this process until  $R$  does not contain any more sets. At this point, the pairs for  $b$  and  $R$  have been extracted and Algorithm 2 returns  $\mathbf{M}^{\mathcal{G},b}$  and  $\mathbf{N}^{\mathcal{G},b}$ . In general the  $N_i^{\mathcal{G},b}$  terms are subsets of vertices containing  $b$  and the  $M_i^{\mathcal{G},b}$  terms are the closure of  $b$  within the corresponding  $N_i^{\mathcal{G},b}$ , that is,  $b \perp\!\!\!\perp N_i^{\mathcal{G},b} \setminus M_i^{\mathcal{G},b} \mid M_i^{\mathcal{G},b} \setminus b [\mathcal{G}_{N_i^{\mathcal{G},b}}]$ —this is shown in Lemma 4.3.5.

Additionally, Lemma 4.3.1 implies  $b \perp\!\!\!\perp N_i^{\mathcal{G},b} \setminus M_i^{\mathcal{G},b} \mid M_i^{\mathcal{G},b} \setminus b [\mathcal{G}]$ . These conditional independence statements are represented by the  $\mathbf{N}_{i,i}^{\mathcal{G},b}$  imsets and by Lemma 4.3.4 their union

is equivalent to the non- $m$ -connecting sets of  $\mathcal{G}_R$  that contain  $b$ , that is,  $\delta_{\{T \in \mathcal{N}(\mathcal{G}_R) ; b \in T\}}$ . Using the principle of inclusion and exclusion, we define the union in terms of the sum of positive and negative intersection terms represented by the  $\mathbf{N}_{J,K}^{\mathcal{G},b}$  imsets. These positive and negative terms reflect the conditional independence statements used in the definition of the ordered local Markov property. Once these components have been accounted for in the imsets,  $b$  is removed from  $R$  and the process of constructing pairs begins again with a new  $b$  and  $R$ . When  $R = \emptyset$ , the algorithm is complete.

---

**Algorithm 3:** NON-M-CONNECTING SETS AS IMSETS  $\text{NSI}(\mathcal{G}, \leq)$ 

---

**Input:** directed MAG:  $\mathcal{G} = (V, E)$ , total order consistent with  $\mathcal{G}: \leq$   
**Output:** imsets:  $u_{\mathcal{N}(\mathcal{G})}^{\leq,+}$ ,  $u_{\mathcal{N}(\mathcal{G})}^{\leq,-}$

- 1 Initialize imsets  $u_{\mathcal{N}(\mathcal{G})}^{\leq,+}: \mathcal{P}(V) \rightarrow 0$  and  $u_{\mathcal{N}(\mathcal{G})}^{\leq,-}: \mathcal{P}(V) \rightarrow 0$  ;
- 2 Let  $R = V$  ;
- 3 **repeat**
- 4   Let  $b = \lceil R \rceil_{\leq}$  ;
- 5   Let  $\mathbf{M}^{\mathcal{G}_R, b}, \mathbf{N}^{\mathcal{G}_R, b} = \text{PAIRS}(\mathcal{G}_R, b)$  ;
- 6   Initialize lists  $\mathbf{A} = \langle \rangle$  and  $\mathbf{B} = \langle \rangle$  ;
- 7   **foreach**  $J \subseteq \{1, \dots, |\mathbf{M}^{\mathcal{G}_R, b}|\}$  **do**
- 8     **foreach**  $K \subseteq J$  where  $K \neq \emptyset$  **do**
- 9       **if**  $\mathbf{N}_{J,K}^{\mathcal{G}_R, b} \neq \emptyset$  **then**
- 10          **if**  $|J \setminus K| \bmod 2 = 0$  and  $\mathbf{N}_{J,K}^{\mathcal{G}_R, b} \notin \mathbf{B}$  **then**
- 11            Append  $\mathbf{N}_{J,K}^{\mathcal{G}_R, b}$  to  $\mathbf{A}$  ;
- 12          **else if**  $|J \setminus K| \bmod 2 = 0$  and  $\mathbf{N}_{J,K}^{\mathcal{G}_R, b} \in \mathbf{B}$  **then**
- 13            Remove  $\mathbf{N}_{J,K}^{\mathcal{G}_R, b}$  from  $\mathbf{B}$  ;
- 14          **else if**  $|J \setminus K| \bmod 2 = 1$  and  $\mathbf{N}_{J,K}^{\mathcal{G}_R, b} \notin \mathbf{A}$  **then**
- 15            Append  $\mathbf{N}_{J,K}^{\mathcal{G}_R, b}$  to  $\mathbf{B}$  ;
- 16          **else if**  $|J \setminus K| \bmod 2 = 1$  and  $\mathbf{N}_{J,K}^{\mathcal{G}_R, b} \in \mathbf{A}$  **then**
- 17            Remove  $\mathbf{N}_{J,K}^{\mathcal{G}_R, b}$  from  $\mathbf{A}$  ;
- 18        **end**
- 19     **end**
- 20    **end**
- 21    **end**
- 22    **foreach**  $\mathbf{N} \in \mathbf{A}$  **do**
- 23      $u_{\mathcal{N}(\mathcal{G})}^{\leq,+} = u_{\mathcal{N}(\mathcal{G})}^{\leq,+} + \delta_{\mathbf{N}}$  ;
- 24    **end**
- 25    **foreach**  $\mathbf{N} \in \mathbf{B}$  **do**
- 26      $u_{\mathcal{N}(\mathcal{G})}^{\leq,-} = u_{\mathcal{N}(\mathcal{G})}^{\leq,-} + \delta_{\mathbf{N}}$  ;
- 27    **end**
- 28    Remove  $b$  from  $R$  ;
- 29 **until**  $R = \emptyset$  ;

---

Algorithms 3 require several new concepts. Accordingly, we define the following notation. Let  $\mathcal{G} = (V, E)$  be a directed MAG,  $\mathbf{P} = \mathcal{P}(V)$  be a poset ordered by inclusion, and  $b \in \text{bar}_{\mathcal{G}}(V)$  be a barren vertex. Additionally, we use  $\mathbf{M}^{\mathcal{G}, b}$  and  $\mathbf{N}^{\mathcal{G}, b}$  to define the ordered lists

output by Algorithm 2. We expand this notation to intersection terms as follows:

$$M_K^{\mathcal{G},b} \equiv \bigcap_{k \in K} M_k^{\mathcal{G},b} \quad N_J^{\mathcal{G},b} \equiv \bigcap_{j \in J} N_j^{\mathcal{G},b} \quad \mathbf{U}_J^{\mathcal{G},b} \equiv \bigcup_{\substack{T \subseteq N_J^{\mathcal{G},b} \\ b \in T}} \{T\} \quad M_{J,K}^{\mathcal{G},b} \equiv M_K^{\mathcal{G},b} \cap N_J^{\mathcal{G},b}.$$

We simplify notation and use  $\mathbf{N}_{J,K}^{\mathcal{G},b}$  to define sets of sets which correspond to the conditional independence statement  $b \perp\!\!\!\perp N_J^{\mathcal{G},r} \setminus M_{J,K}^{\mathcal{G},b} \mid M_{J,K}^{\mathcal{G},b} \setminus b$ . If  $A = b$ ,  $B = N_J^{\mathcal{G},b} \setminus M_{J,K}^{\mathcal{G},b}$ , and  $C = M_{J,K}^{\mathcal{G},b} \setminus b$ , then

$$\begin{aligned} ABC &= (N_J^{\mathcal{G},b} \setminus M_{J,K}^{\mathcal{G},b}) \cup (M_{J,K}^{\mathcal{G},b} \setminus b) \cup b \\ &= N_J^{\mathcal{G},b} \\ AC &= (M_{J,K}^{\mathcal{G},b} \setminus b) \cup b \\ &= M_{J,K}^{\mathcal{G},b} \\ BC &= (N_J^{\mathcal{G},b} \setminus M_{J,K}^{\mathcal{G},b}) \cup (M_{J,K}^{\mathcal{G},b} \setminus b) \\ &= N_J^{\mathcal{G},b} \setminus b. \end{aligned}$$

Therefore, define

$$\mathbf{N}_{J,K}^{\mathcal{G},b} \equiv \bigcup_{\substack{T \in \mathbf{U}_J^{\mathcal{G},b} \\ T \not\subseteq M_{J,K}^{\mathcal{G},b}}} \{T\} = \bigcup_{\substack{T \subseteq N_J^{\mathcal{G},b} \\ b \in T \\ T \not\subseteq M_{J,K}^{\mathcal{G},b}}} \{T\} = \bigcup_{\substack{T \subseteq ABC \\ T \not\subseteq AC \\ T \not\subseteq BC}} \{T\}.$$

Accordingly

$$\delta_{A,B|C} = \delta_{\mathbf{N}_{J,K}^{\mathcal{G},b}} = \delta_{b, N_J^{\mathcal{G},b} \setminus M_{J,K}^{\mathcal{G},b} \mid M_{J,K}^{\mathcal{G},b} \setminus b} \quad \text{and} \quad \mu_{\mathbf{P}} \delta_{\mathbf{N}_{J,K}^{\mathcal{G},b}} = u_{\langle b, N_J^{\mathcal{G},b} \setminus M_{J,K}^{\mathcal{G},b} \mid M_{J,K}^{\mathcal{G},b} \setminus b \rangle}.$$

Now we show that the output of Algorithm 3 characterize the set identifier for the non- $m$ -connecting sets; Appendix B.3 shows that Algorithm 3 does not necessarily give the most efficient solution.

**Definition** (*inclusion/exclusion for imsets* [91]). Let  $V$  be a non-empty set of variables and  $N_1, \dots, N_n \subseteq \mathcal{P}(V)$  be  $n$  sets of sets. The concept of inclusion/exclusion is extended to imsets as follows:

$$\delta_{\bigcup_{i=1}^n N_i} = \sum_{\substack{J \subseteq \{1, \dots, n\} \\ J \neq \emptyset}} (-1)^{|J|-1} \delta_{\bigcap_{j \in J} N_j}.$$

Several applications of De Morgan's laws gives an alternative form as follows:

$$\delta_{\bigcap_{i=1}^n N_i} = \sum_{\substack{J \subseteq \{1, \dots, n\} \\ J \neq \emptyset}} (-1)^{|J|-1} \delta_{\bigcup_{j \in J} N_j}.$$

**Proposition 4.3.2.** *If  $\mathcal{G} = (V, E)$  be a directed MAG,  $\leq$  be a total order consistent with  $\mathcal{G}$ ,  $\mathcal{P} = \mathcal{P}(V)$  be a poset ordered by inclusion, then:*

$$\delta_{\mathcal{N}(\mathcal{G}_R)} = \sum_{b \in V} \delta_{\{T \subseteq_{\mathcal{N}(\mathcal{G}_R)}^b \mathcal{G}_R\}} = u_{\mathcal{N}(\mathcal{G}_R)}^{\leq, +} - u_{\mathcal{N}(\mathcal{G}_R)}^{\leq, -}$$

where  $R = \text{pre}_{\mathcal{G}}^{\leq}(b)$  for all  $b \in V$ .

*Proof.* Let  $\mathcal{G} = (V, E)$  be a directed MAG,  $\leq$  be a total order consistent with  $\mathcal{G}$ ,  $\mathcal{P} = \mathcal{P}(V)$  be a poset ordered by inclusion, and  $R_b = \text{pre}_{\mathcal{G}}^{\leq}(b)$ . Furthermore, let  $\mathbf{M}^{\mathcal{G}_R, b}, \mathbf{N}^{\mathcal{G}_R, b} = \text{PAIRS}(\mathcal{G}_R, b)$ :

$$\mathbf{M}^{\mathcal{G}_R, b} = \langle M_1^{\mathcal{G}_R, b}, \dots, M_{n_b}^{\mathcal{G}_R, b} \rangle \quad \mathbf{N}^{\mathcal{G}_R, b} = \langle N_1^{\mathcal{G}_R, b}, \dots, N_{n_b}^{\mathcal{G}_R, b} \rangle$$

where  $n_b = |\mathbf{M}^{\mathcal{G}_R, b}|$ .

$$\begin{aligned} \delta_{\{T \subseteq_{\mathcal{N}(\mathcal{G}_R)}^b R\}} &= \delta_{\bigcup_{i=1}^{n_b} \mathbf{N}_{i,i}^{\mathcal{G}_R, b}} && \text{(Lemma 4.3.4)} \\ &= \delta \left[ \bigcup_{i=1}^{n_b} \left[ \bigcup_{\substack{T \in \mathbf{U}_i^{\mathcal{G}_R, b} \\ T \not\subseteq M_i^{\mathcal{G}_R, b}}} \{T\} \right] \right] \\ &= \sum_{\substack{J \subseteq \{1, \dots, n_b\} \\ J \neq \emptyset}} (-1)^{|J|-1} \delta \left[ \bigcap_{j \in J} \left[ \bigcup_{\substack{T \in \mathbf{U}_j^{\mathcal{G}_R, b} \\ T \not\subseteq M_j^{\mathcal{G}_R, b}}} \{T\} \right] \right] && \text{(inclusion/exclusion)} \\ &= \sum_{\substack{J \subseteq \{1, \dots, n_b\} \\ J \neq \emptyset}} (-1)^{|J|-1} \delta \left[ \bigcap_{j \in J} \left[ \bigcup_{\substack{T \in \mathbf{U}_j^{\mathcal{G}_R, b} \\ T \not\subseteq M_j^{\mathcal{G}_R, b}}} \{T\} \right] \right] && (\mathbf{U}_j^{\mathcal{G}_R, b} \rightarrow \mathbf{U}_j^{\mathcal{G}_R, b}) \\ &= \sum_{\substack{J \subseteq \{1, \dots, n_b\} \\ J \neq \emptyset}} (-1)^{|J|-1} \sum_{\substack{K \subseteq J \\ K \neq \emptyset}} (-1)^{|K|-1} \delta \left[ \bigcup_{k \in K} \left[ \bigcup_{\substack{T \in \mathbf{U}_k^{\mathcal{G}_R, b} \\ T \not\subseteq M_k^{\mathcal{G}_R, b}}} \{T\} \right] \right] && \text{(inclusion/exclusion)} \end{aligned}$$

$$\begin{aligned}
&= \sum_{\substack{J \subseteq \{1, \dots, n_b\} \\ J \neq \emptyset}} \sum_{\substack{K \subseteq J \\ K \neq \emptyset}} (-1)^{|J \setminus K|} \delta \left[ \bigcup_{k \in K} \left[ \bigcup_{T \in \mathcal{U}_J^{\mathcal{G}_R, b} \{T\}} \right] \right. \\
&\quad \left. T \not\subseteq M_k^{\mathcal{G}_R, b} \right] \\
&= \sum_{\substack{J \subseteq \{1, \dots, n_b\} \\ J \neq \emptyset}} \sum_{\substack{K \subseteq J \\ K \neq \emptyset}} (-1)^{|J \setminus K|} \delta \left[ \bigcup_{k \in K} \left[ \mathcal{U}_J^{\mathcal{G}_R, b} \setminus \bigcup_{T \in \mathcal{U}_J^{\mathcal{G}_R, b} \{T\}} \right] \right. \\
&\quad \left. T \subseteq M_k^{\mathcal{G}_R, b} \right] \quad (\text{complement}) \\
&= \sum_{\substack{J \subseteq \{1, \dots, n_b\} \\ J \neq \emptyset}} \sum_{\substack{K \subseteq J \\ K \neq \emptyset}} (-1)^{|J \setminus K|} \delta \left[ \mathcal{U}_J^{\mathcal{G}_R, b} \setminus \left[ \bigcap_{k \in K} \bigcup_{T \in \mathcal{U}_J^{\mathcal{G}_R, b} \{T\}} \right] \right. \\
&\quad \left. T \subseteq M_k^{\mathcal{G}_R, b} \right] \quad (\text{De Morgan's law}) \\
&= \sum_{\substack{J \subseteq \{1, \dots, n_b\} \\ J \neq \emptyset}} \sum_{\substack{K \subseteq J \\ K \neq \emptyset}} (-1)^{|J \setminus K|} \delta \left[ \mathcal{U}_J^{\mathcal{G}_R, b} \setminus \left[ \bigcap_{k \in K} \bigcup_{T \in \mathcal{U}_J^{\mathcal{G}_R, b} \{T\}} \right] \right. \\
&\quad \left. T \subseteq M_{J, K}^{\mathcal{G}_R, b} \right] \quad (M_k^{\mathcal{G}_R, b} \rightarrow M_{J, K}^{\mathcal{G}_R, b}) \\
&= \sum_{\substack{J \subseteq \{1, \dots, n_b\} \\ J \neq \emptyset}} \sum_{\substack{K \subseteq J \\ K \neq \emptyset}} (-1)^{|J \setminus K|} \delta \left[ \mathcal{U}_J^{\mathcal{G}_R, b} \setminus \left[ \bigcup_{T \in \mathcal{U}_J^{\mathcal{G}_R, b} \{T\}} \right] \right. \\
&\quad \left. T \subseteq M_{J, K}^{\mathcal{G}_R, b} \right] \\
&= \sum_{\substack{J \subseteq \{1, \dots, n_b\} \\ J \neq \emptyset}} \sum_{\substack{K \subseteq J \\ K \neq \emptyset}} (-1)^{|J \setminus K|} \delta \bigcup_{T \in \mathcal{U}_J^{\mathcal{G}_R, b} \{T\}} \delta_{T \not\subseteq M_{J, K}^{\mathcal{G}_R, b}} \quad (\text{complement}) \\
&= \sum_{\substack{J \subseteq \{1, \dots, n_b\} \\ J \neq \emptyset}} \sum_{\substack{K \subseteq J \\ K \neq \emptyset}} (-1)^{|J \setminus K|} \delta_{\mathcal{N}_{J, K}^{\mathcal{G}_R, b}}
\end{aligned}$$

Accordingly,

$$\begin{aligned}
\delta_{\mathcal{N}(\mathcal{G}_R)} &= \sum_{b \in V} \delta_{\{T \subseteq \mathcal{N}(\mathcal{G}_R) \mathcal{G}_R\}} \\
&= \sum_{b \in V} \sum_{\substack{J \subseteq \{1, \dots, n_b\} \\ J \neq \emptyset}} \sum_{\substack{K \subseteq J \\ K \neq \emptyset}} (-1)^{|J \setminus K|} \delta_{\mathcal{N}_{J, K}^{\mathcal{G}_R, b}} \\
&= u_{\mathcal{N}(\mathcal{G}_R)}^{\leq, +} - u_{\mathcal{N}(\mathcal{G}_R)}^{\leq, -}
\end{aligned}$$

where  $R = \text{pre}_{\vec{g}}^{\leq}(b)$  for all  $b \in V$ . □

In what follows, we give an illustrative example of Algorithm 3. Figure 4.3 depicts a directed MAG  $\mathcal{G} = (V, E)$ , its  $m$ -connecting sets  $\mathcal{M}(\mathcal{G})$ , and its non- $m$ -connecting sets  $\mathcal{N}(\mathcal{G})$ . Consider the total order  $\leq$  over  $V$  such that  $e \leq a \leq d \leq b \leq c$ .



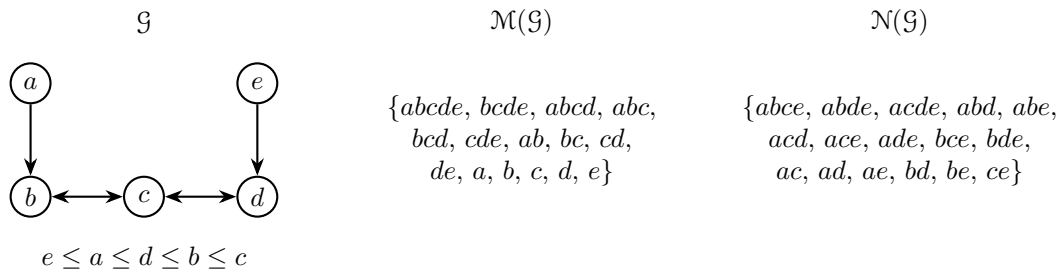
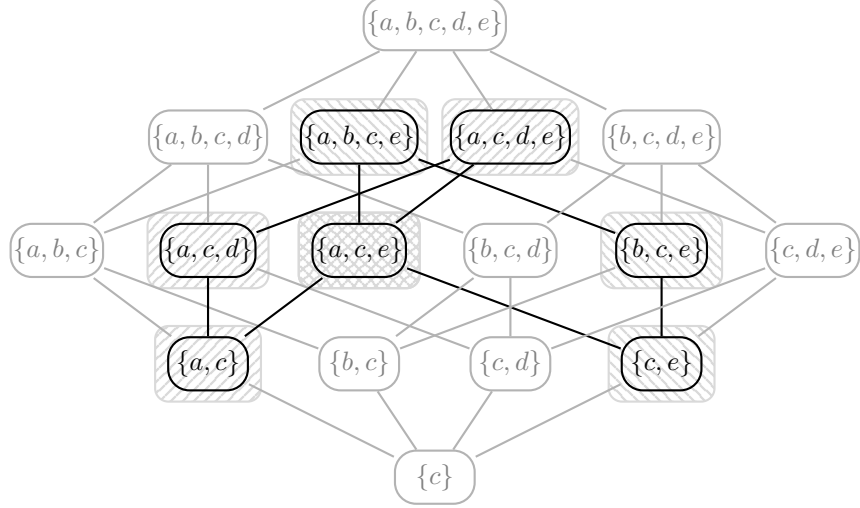


Figure 4.3: A directed MAG with vertices  $\{a, b, c, d, e\}$  and the corresponding  $m$ -connecting and non- $m$ -connecting sets for the directed MAG.

Run  $\text{PAIRS}(\mathcal{G}_{abcde}, c)$  to construct ordered lists  $\mathbf{N}^{\mathcal{G}_{abcde}, c} = \langle \{a, b, c, e\}, \{a, c, d, e\} \rangle$  and  $\mathbf{M}^{\mathcal{G}_{abcde}, c} = \langle \{a, b, c\}, \{c, d, e\} \rangle$ .

$$\begin{aligned} \mathbf{N}^{\mathcal{G}_{abcde,c}} &= \langle \{a, b, c, e\}, \{a, c, d, e\} \rangle \\ \mathbf{M}^{\mathcal{G}_{abcde,c}} &= \langle \{a, b, c\}, \{c, d, e\} \rangle \end{aligned}$$



$$\begin{aligned} \mathbf{N}_{1,1}^{\mathcal{G}_{abcde,c}} &= \{N \in \mathcal{N}(\mathcal{G}) ; \delta_{c,e|ab}(N) = 1\} = \{abce, ace, bce, ce\} \\ \mathbf{N}_{2,2}^{\mathcal{G}_{abcde,c}} &= \{N \in \mathcal{N}(\mathcal{G}) ; \delta_{c,a|de}(N) = 1\} = \{acde, ace, acd, ac\} \end{aligned}$$

Figure 4.4: A visualization of  $\text{PAIRS}(\mathcal{G}_{abcde,c})$  applied to the directed MAG in Figure 4.3 and the corresponding base conditional terms.

The intersection terms are as follows—these terms correspond to intersections over members of  $\mathbf{N}^{\mathcal{G}_{abcde,c}}$  indexed by the loop on line 7 of Algorithm 3.

Intersection Terms:

$$\begin{aligned} N_1^{\mathcal{G}_{abcde,c}} &= \{a, b, c, e\} & M_1^{\mathcal{G}_{abcde,c}} &= \{a, b, c\} \\ N_2^{\mathcal{G}_{abcde,c}} &= \{a, c, d, e\} & M_2^{\mathcal{G}_{abcde,c}} &= \{c, d, e\} \\ N_{12}^{\mathcal{G}_{abcde,c}} &= \{a, c, e\} & M_{12}^{\mathcal{G}_{abcde,c}} &= \{c\} \end{aligned}$$

The conditional terms are as follows—these terms correspond to those appended and removed on lines 11, 13, 15, and 17.

Conditional Terms:

$$\begin{aligned}
\mathbf{N}_{1,1}^{\mathcal{G}_{abcde},c} &= \{N \in \mathcal{N}(\mathcal{G}) ; \quad \delta_{c,e|ab}(N) = 1\} \\
\mathbf{N}_{2,2}^{\mathcal{G}_{abcde},c} &= \{N \in \mathcal{N}(\mathcal{G}) ; \quad \delta_{c,a|cd}(N) = 1\} \\
\mathbf{N}_{12,1}^{\mathcal{G}_{abcde},c} &= \{N \in \mathcal{N}(\mathcal{G}) ; \quad \delta_{c,e|a}(N) = 1\} \\
\mathbf{N}_{12,2}^{\mathcal{G}_{abcde},c} &= \{N \in \mathcal{N}(\mathcal{G}) ; \quad \delta_{c,a|e}(N) = 1\} \\
\mathbf{N}_{12,12}^{\mathcal{G}_{abcde},c} &= \{N \in \mathcal{N}(\mathcal{G}) ; \quad \delta_{c,ae}(N) = 1\}
\end{aligned}$$

The positive and negative conditional terms are as follows—the positive terms are on the left and correspond to the list  $A$  in Algorithm 3 and the negative terms are on the right and correspond to the list  $B$  in Algorithm 3.

Positive Conditional Terms:

Negative Conditional Terms:

$$\begin{aligned}
\mathbf{N}_{1,1}^{\mathcal{G}_{abcde},c} &= \{N \in \mathcal{N}(\mathcal{G}) ; \quad \delta_{c,e|ab}(N) = 1\} & \mathbf{N}_{12,1}^{\mathcal{G}_{abcde},c} &= \{N \in \mathcal{N}(\mathcal{G}) ; \quad \delta_{c,e|a}(N) = 1\} \\
\mathbf{N}_{2,2}^{\mathcal{G}_{abcde},c} &= \{N \in \mathcal{N}(\mathcal{G}) ; \quad \delta_{c,a|de}(N) = 1\} & \mathbf{N}_{12,2}^{\mathcal{G}_{abcde},c} &= \{N \in \mathcal{N}(\mathcal{G}) ; \quad \delta_{c,a|e}(N) = 1\} \\
\mathbf{N}_{12,12}^{\mathcal{G}_{abcde},c} &= \{N \in \mathcal{N}(\mathcal{G}) ; \quad \delta_{c,ae}(N) = 1\} & &
\end{aligned}$$

Accordingly, the non- $m$ -connecting set terms added on lines 23 and 26 of Algorithm 3 are as follows—these imsets represent all non- $m$ -connecting subsets of  $\{a, b, c, d, e\}$  that contain  $c$ .

$$\begin{aligned}
\delta_{\{T \subseteq_{\mathcal{N}(\mathcal{G})}^c \{a,b,c,d,e\}\}} &= \delta_{c,e|ab} + \delta_{c,a|de} + \delta_{c,ae} - \delta_{c,e|a} - \delta_{c,a|e} \\
&= [\delta_{abce} + \delta_{ace} + \delta_{bce} + \delta_{ce}] + [\delta_{acde} + \delta_{acd} + \delta_{ace} + \delta_{ac}] \\
&\quad + [\delta_{ace} + \delta_{ac} + \delta_{ce}] - [\delta_{ace} + \delta_{ce}] - [\delta_{ace} + \delta_{ac}]
\end{aligned}$$

Run  $\text{PAIRS}(\mathcal{G}_{abcde}, b)$  to construct ordered lists  $\mathbf{N}^{\mathcal{G}_{abcde},b} = \langle \{a, b, d, e\} \rangle$  and  $\mathbf{M}^{\mathcal{G}_{abcde},b} = \langle \{a, b\} \rangle$ .

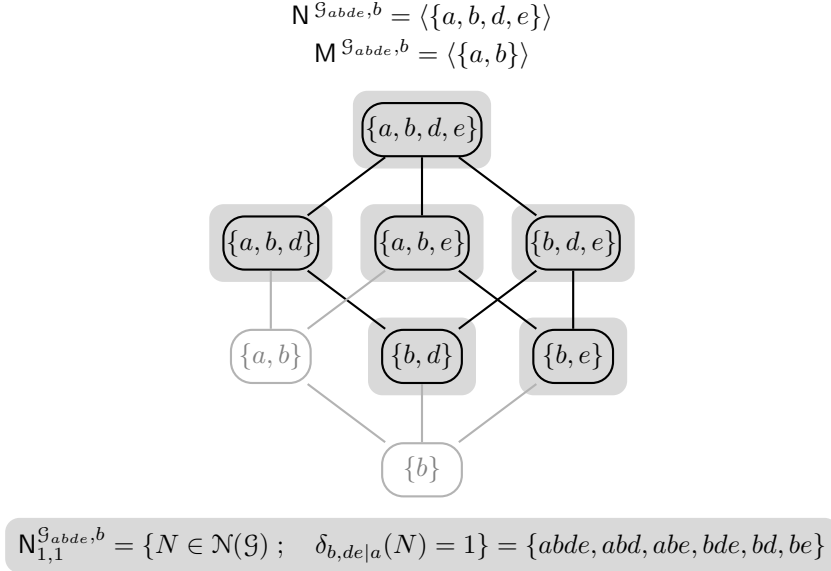


Figure 4.5: A visualization of  $\text{PAIRS}(\mathcal{G}_{abde}, b)$  applied to the directed MAG in Figure 4.3 and the corresponding base conditional terms.

The intersection terms are as follows—these terms correspond to intersections over members of  $\mathbf{N}^{\mathcal{G}_{abde},b}$  indexed by the loop on line 7 of Algorithm 3.

| Intersection Terms:   | Positive Conditional Terms:   |
|---|---|
| $N_1^{\mathcal{G}_{abde},b} = \{a, b, d, e\} \quad M_1^{\mathcal{G}_{abde},b} = \{a, b\}$ | $\mathbf{N}_{1,1}^{\mathcal{G}_{abde},b} = \{N \in \mathcal{N}(\mathcal{G}) ; \delta_{b,de a}(N) = 1\}$ |

Accordingly, the non- $m$ -connecting set terms added on lines 23 and 26 of Algorithm 3 are as follows—these imsets represent all non- $m$ -connecting subsets of  $\{a, b, d, e\}$  that contain  $b$ .

$$\begin{aligned} \delta_{\{T \subseteq_{\mathcal{N}(\mathcal{G})}^b \{a, b, d, e\}\}} &= \delta_{b,de|a} \\ &= [\delta_{abde} + \delta_{abd} + \delta_{abe} + \delta_{bde} + \delta_{bd} + \delta_{be}] \end{aligned}$$

Run  $\text{PAIRS}(\mathcal{G}_{ade}, d)$  to construct ordered lists  $\mathbf{N}^{\mathcal{G}_{ade},d} = \langle \{a, d, e\} \rangle$  and  $\mathbf{M}^{\mathcal{G}_{ade},d} = \langle \{d, e\} \rangle$ .

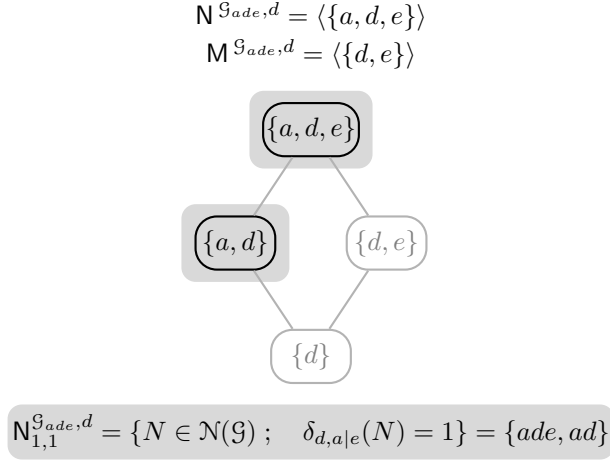


Figure 4.6: A visualization of  $\text{PAIRS}(\mathcal{G}_{ade}, d)$  applied to the directed MAG in Figure 4.3 and the corresponding base conditional terms.

The intersection terms are as follows—these terms correspond to intersections over members of  $\mathbf{N}^{\mathcal{G}_{ade},d}$  indexed by the loop on line 7 of Algorithm 3.

|   |  |
|---|--|
| <p>Intersection Terms:</p> $N_1^{\mathcal{G}_{ade},d} = \{a, d, e\} \quad M_1^{\mathcal{G}_{ade},d} = \{d, e\}$ | <p>Positive Conditional Terms:</p> $\mathbf{N}_{1,1}^{\mathcal{G}_{ade},d} = \{N \in \mathcal{N}(\mathcal{G}) ; \delta_{d,a e}(N) = 1\}$ |
|---|--|

Accordingly, the non- $m$ -connecting set terms added on lines 23 and 26 of Algorithm 3 are as follows—these insets represent all non- $m$ -connecting subsets of  $\{a, d, e\}$  that contain  $d$ .

$$\begin{aligned} \delta_{\{T \subseteq_{\mathcal{N}(\mathcal{G})}^d \{a, d, e\}\}} &= \delta_{d,a|e} \\ &= [\delta_{ade} + \delta_{ad}] \end{aligned}$$

Run  $\text{PAIRS}(\mathcal{G}_{ae}, a)$  to construct ordered lists  $\mathbf{N}^{\mathcal{G}_{ae},a} = \langle \{a, e\} \rangle$  and  $\mathbf{M}^{\mathcal{G}_{ae},a} = \langle \{a\} \rangle$ .

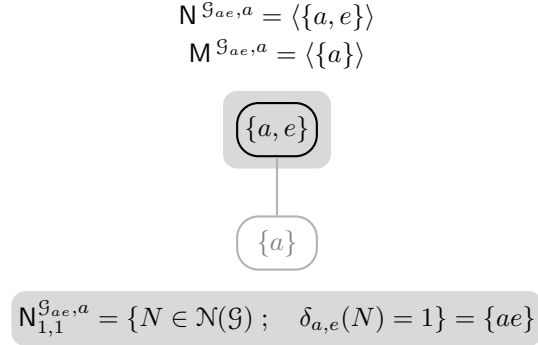


Figure 4.7: A visualization of  $\text{PAIRS}(\mathcal{G}_{ae}, a)$  applied to the directed MAG in Figure 4.3 and the corresponding base conditional terms.

The intersection terms are as follows—these terms correspond to intersections over members of  $\mathbf{N}^{\mathcal{G}_{ae},a}$  indexed by the loop on line 7 of Algorithm 3.

|   |   |
|---|---|
| <p>Intersection Terms:</p> $N_1^{\mathcal{G}_{ae},a} = \{a, e\} \quad M_1^{\mathcal{G}_{ae},a} = \{a\}$ | <p>Positive Conditional Terms:</p> $\mathbf{N}_{1,1}^{\mathcal{G}_{ae},a} = \{N \in \mathcal{N}(\mathcal{G}) ; \delta_{a,e}(N) = 1\}$ |
|---|---|

Accordingly, the non- $m$ -connecting set terms added on lines 23 and 26 of Algorithm 3 are as follows—these insets represent all non- $m$ -connecting subsets of  $\{a, e\}$  that contain  $a$ .

$$\delta_{\{T \subseteq_{\mathcal{N}(\mathcal{G})}^a \{a, e\}\}} = \delta_{ae}$$

Run  $\text{PAIRS}(\mathcal{G}_e, e)$  to construct ordered lists  $\mathbf{N}^{\mathcal{G}_e,e} = \langle \rangle$  and  $\mathbf{M}^{\mathcal{G}_e,e} = \langle \rangle$ .

$$\begin{aligned}
\mathbf{N}^{\mathcal{G}_e, e} &= \langle \rangle \\
\mathbf{M}^{\mathcal{G}_e, e} &= \langle \rangle \\
&\quad \textcircled{\{e\}}
\end{aligned}$$

Figure 4.8: A visualization of  $\text{PAIRS}(\mathcal{G}_e, e)$  applied to the directed MAG in Figure 4.3 and the corresponding base conditional terms.

There are no intersection terms. Accordingly, the non- $m$ -connecting set terms added on lines 23 and 26 of Algorithm 3 are as follows (there are none)—these imsets represent all non- $m$ -connecting subsets of  $\{e\}$  that contain  $e$ .

Combining the results from all the iterations of the procedure, we get

$$\begin{aligned}
u_{\mathbf{N}(\mathcal{G})}^{\leq, +} &= \delta_{c, e|ab} + \delta_{c, a|de} + \delta_{c, ae} + \delta_{b, de|a} + \delta_{d, a|e} + \delta_{a, e} \\
u_{\mathbf{N}(\mathcal{G})}^{\leq, -} &= \delta_{c, e|a} + \delta_{c, a|e}
\end{aligned}$$

or

$$\begin{aligned}
u_{\mathbf{N}(\mathcal{G})}^{\leq, +} &= [\delta_{abce} + \delta_{ace} + \delta_{bce} + \delta_{ce}] + [\delta_{acde} + \delta_{acd} + \delta_{ace} + \delta_{ac}] + [\delta_{ace} + \delta_{ac} + \delta_{ce}] \\
&\quad + [\delta_{abde} + \delta_{abd} + \delta_{abe} + \delta_{bde} + \delta_{bd} + \delta_{be}] + [\delta_{ade} + \delta_{ad}] + \delta_{ae} \\
u_{\mathbf{N}(\mathcal{G})}^{\leq, -} &= [\delta_{ace} + \delta_{ce}] + [\delta_{ace} + \delta_{ac}]
\end{aligned}$$

where the linear combination contains all the non- $m$ -connecting set terms.

Let  $V$  be a non-empty set of variables and  $\mathbf{P} = \mathcal{P}(V)$  be a poset ordered by inclusion.

Applying the Möbius inversion, we get

$$\begin{aligned}
\mu_{\mathbf{P}} u_{\mathbf{N}(\mathcal{G})}^{\leq, +} &= u_{\langle c, e|ab \rangle} + u_{\langle c, a|de \rangle} + u_{\langle c, ae \rangle} + u_{\langle b, de|a \rangle} + u_{\langle d, a|e \rangle} + u_{\langle a, e \rangle} \\
\mu_{\mathbf{P}} u_{\mathbf{N}(\mathcal{G})}^{\leq, -} &= u_{\langle c, e|a \rangle} + u_{\langle c, a|e \rangle}
\end{aligned}$$

or

$$\begin{aligned}\mu_{\mathcal{P}u_{\mathcal{N}(\mathcal{G})}^{\leq,+}} &= [\delta_{abce} + \delta_{ab} - \delta_{abc} - \delta_{abe}] + [\delta_{acde} + \delta_{de} - \delta_{ade} - \delta_{cde}] + [\delta_{ace} - \delta_{ae} - \delta_c] \\ &\quad + [\delta_{abde} + \delta_a - \delta_{ade} - \delta_{ab}] + [\delta_{ade} + \delta_e - \delta_{ae} + \delta_{de}] + [\delta_{ae} - \delta_a - \delta_e] \\ \mu_{\mathcal{P}u_{\mathcal{N}(\mathcal{G})}^{\leq,-}} &= [\delta_{ace} + \delta_a - \delta_{ac} - \delta_{ae}] + [\delta_{ace} + \delta_e - \delta_{ae} - \delta_{ce}]\end{aligned}$$

Clearly  $\mu_{\mathcal{P}u_{\mathcal{N}(\mathcal{G})}^{\leq,+}}$  and  $\mu_{\mathcal{P}u_{\mathcal{N}(\mathcal{G})}^{\leq,-}}$  are structural insets.

### 4.3.2 Factorization Implies Markov

In this section, we provide the necessary lemmas to prove that if the factorization presented in Section 4.3.4 holds, then the global Markov property holds. However, in order to do so we first introduce the concept of a minimal latent set. The minimal latent set is defined as follows. Let  $\mathcal{G} = (V, E)$  be an ADMG such that  $A \in \mathcal{A}(\mathcal{G})$  is an ancestral set and  $b = [A]_{\leq}$  with preceding vertices  $R = \text{pre}_{\mathcal{G}}^{\leq}(b)$ :

$$\text{ml}_{\mathcal{G}}^{\leq}(A) \equiv \text{sp}_{\mathcal{G}_R}(\text{dis}_{\mathcal{G}_A}(b)) \setminus \text{dis}_{\mathcal{G}_A}(b).$$

Let  $L = V \setminus A$  be the set of latent variables. Intuitively,  $\text{ml}_{\mathcal{G}}^{\leq}(A)$  defines the minimal subset of latent vertices  $L^{\min} \subseteq L$  for which every member is automatically added to the ordered Markov blanket and order closure when added to  $A$ :

$$\begin{aligned}\text{mb}_{\mathcal{G}_{l \cup A}}^{\leq}(b) &\equiv l \cup \text{mb}_{\mathcal{G}_A}^{\leq}(b) \quad \text{for all } l \in L^{\min}; \\ \text{cl}_{\mathcal{G}_{l \cup A}}^{\leq}(b) &\equiv l \cup \text{cl}_{\mathcal{G}_A}^{\leq}(b) \quad \text{for all } l \in L^{\min}.\end{aligned}$$

This concept was originally introduced by Richardson to construct maximal ancestral sets and is made rigorous in Lemma 5 of [66]. These sets were used to simplify the set of conditional independence statements required to characterize independence models induced by ADMGs [66].



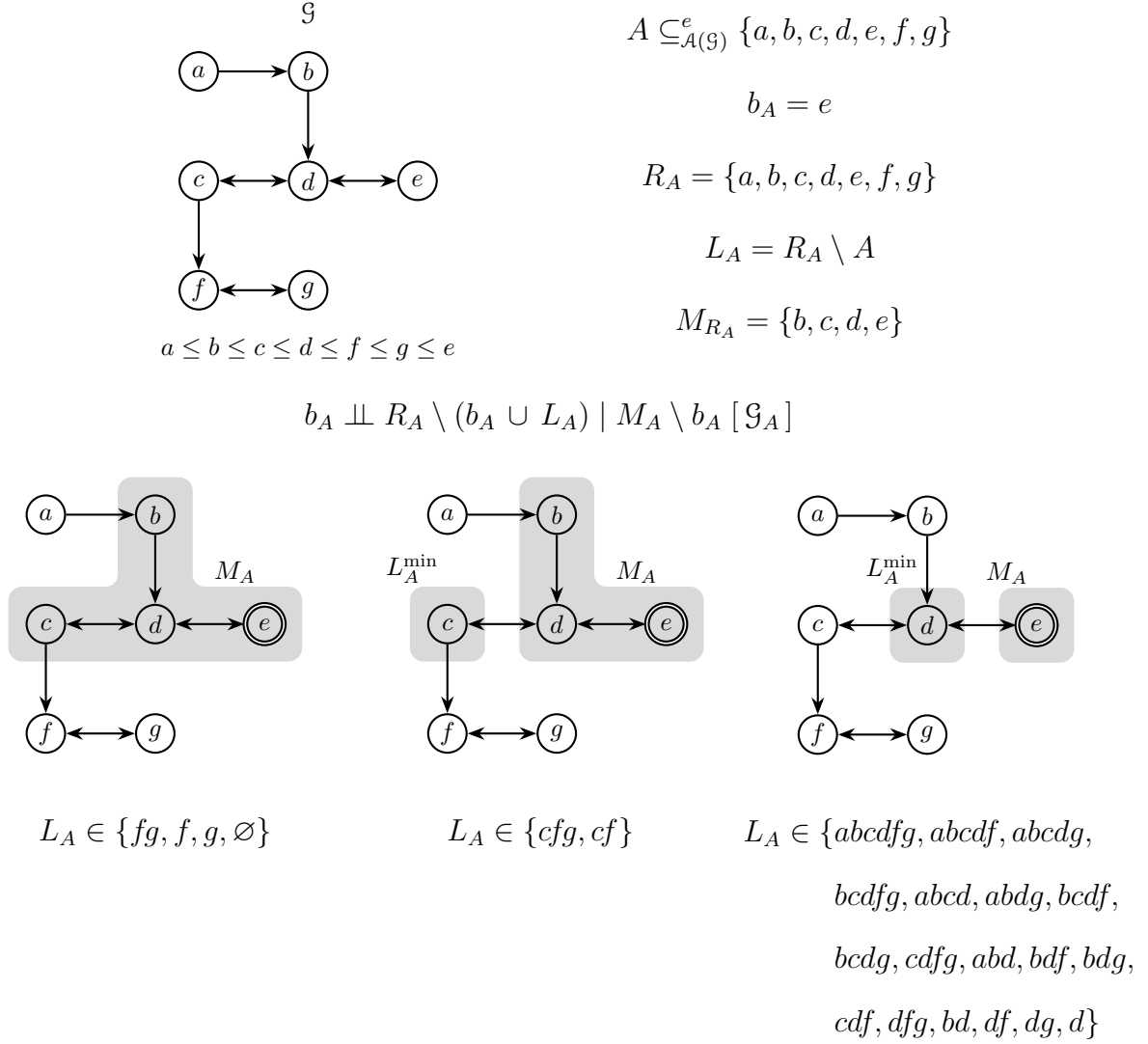


Figure 4.9: An illustration of the minimal latent sets.

Figure 4.9 illustrates the minimal latent set for an ADMG  $\mathcal{G} = (V, E)$  and ancestral set  $e \in A \in \mathcal{A}(\mathcal{G})$ . Let  $b_A = [A]_{\leq}$  with preceding vertices  $R_A = \text{pre}_{\mathcal{G}}^{\leq}(b_A)$ ,  $L_A = R_A \setminus A$ ,  $M_A = \text{col}_{\mathcal{G}_A}(b_A)$ , and  $L_A^{\min} = \text{ml}_{\mathcal{G}}^{\leq}(A)$ . All possible sets for  $L_A$  are listed and partitioned by  $M_A$  at the bottom of the figure. In particular,  $M_A$  is the closure of  $b_A$  with respect to  $A$ . The minimal latent set  $L_A^{\min}$  is the minimal subset of  $L_A$  intersected with  $M_{R_A} = \text{col}_{\mathcal{G}_{R_A}}(b_A)$  for each partition. Note that  $L_A^{\min}$  need not be one of the possible sets of  $L_A$ ; see Figure 4.9

when  $L_A^{\min} = \{c\}$  and  $L_A \in \{\{c, f, g\}, \{c, f\}\}$ .

Lemma 4.3.5 uses the concept of a minimal latent set to extract conditional independence statements from a directed MAG.

**Lemma 4.3.5.** *Let  $\mathcal{G} = (V, E)$  be a directed MAG,  $\mathbf{P} = \mathcal{P}(V)$  be the poset ordered by inclusion,  $\leq$  be a total order consistent with  $\mathcal{G}$ , and  $u_{\mathcal{N}(\mathcal{G})}^{\leq,+}, u_{\mathcal{N}(\mathcal{G})}^{\leq,-} = \text{NSI}(\mathcal{G}, \leq)$  be the imsets constructed by Algorithm 3. If  $\mathcal{G}$  contains a vertex  $b \in V$  with preceding vertices  $R = \text{pre}_{\mathcal{G}}^{\leq}(b)$ , then for  $M = \text{col}_{\mathcal{G}_R}(b)$ :*

$$b \perp\!\!\!\perp R \setminus M \mid M \setminus b [\mu_{\mathbf{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq,+}].$$

Furthermore, if  $\mathcal{G}$  contains an ancestral set  $A \in \mathcal{A}(\mathcal{G})$  such that  $b \in A \subseteq R$ , then for  $L = \text{ml}_{\mathcal{G}}^{\leq}(A)$ ,  $N = M \setminus L$ , and  $M_A = \text{col}_{\mathcal{G}_A}(b)$ :

$$b \perp\!\!\!\perp N \setminus M_A \mid M_A \setminus b [\mu_{\mathbf{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq,+}].$$

*Proof.* By Corollary 4.3.1  $\mathcal{G}_R$  is a directed MAG.

By Proposition 3.5.3 the independence model induced by a structural imset is a semi-graphoid. Accordingly, we may apply the semi-graphoid axioms.

Let  $\mathbf{M}^{\mathcal{G}_R, b}$  and  $\mathbf{N}^{\mathcal{G}_R, b}$  be the ordered lists constructed by Algorithm 2 and let  $n$  be their cardinality. The structural imset  $\mu_{\mathbf{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq,+}$  is constructed as the sum over a set of semi-elementary imsets including the semi-elementary imsets defined as  $\mu_{\mathbf{P}} \delta_{N_{i,i}^{\mathcal{G}_R, b}}$  for  $1 \leq i \leq n$ . Accordingly

$$b \perp\!\!\!\perp N_i^{\mathcal{G}_R, b} \setminus M_i^{\mathcal{G}_R, b} \mid M_i^{\mathcal{G}_R, b} \setminus b [\mu_{\mathbf{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq,+}].$$

By construction  $N_1^{\mathcal{G}_R, b} = R$  and by Lemma 4.0.1  $M_1^{\mathcal{G}_R, b} = M$ , therefore

$$b \perp\!\!\!\perp R \setminus M \mid M \setminus b [\mu_{\mathbf{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq,+}].$$

Note  $M_A \subseteq N \subseteq M$  (because  $N \subseteq M$  only removes latent vertices). By Lemma 4.0.1  $M_A \in [N]_{\mathcal{M}(\mathcal{G}_R)}^b$ .

If  $L = \emptyset$ , then  $M_A = N = M$  (because  $M_A = M$ ). Accordingly  $N \setminus M_A = \emptyset$  and by semi-graphoid axiom of triviality

$$b \perp\!\!\!\perp N \setminus M_A \mid M_A \setminus b [\mu_{\mathbf{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq,+}].$$

If  $L \neq \emptyset$ , then assume by way of contradiction that  $N$  is not a maximal non- $m$ -connecting superset of  $M_A$ . But if we add a member of  $L$  to  $N$ , then we change  $M_A$ . Therefore,  $N$  is maximal and  $N_i^{\mathcal{G}_{R,b}} = N$  and  $M_i^{\mathcal{G}_{R,b}} = M_A$  for some  $1 \leq i \leq n$ . Accordingly

$$b \perp\!\!\!\perp N \setminus M_A \mid M_A \setminus b [\mu_{\mathcal{P}u_{\mathcal{N}(\mathcal{G})}^{\leq,+}}].$$

□

We now extend the ideas of Lemma 4.3.5 to incorporate the conditional independence statements required by the ordered local Markov property. Let  $\mathcal{G} = (V, E)$  be a directed MAG,  $\leq$  be a total order consistent with  $\mathcal{G}$ , and  $A \in \mathcal{A}(\mathcal{G})$  be an ancestral set where  $b_A = [A]_{\leq}$ ,  $R_A = \text{pre}_{\mathcal{G}}^{\leq}(b_A)$ . Let  $M_A = \text{col}_{\mathcal{G}}(b_A)$ ,  $M_{R_A} = \text{col}_{\mathcal{G}_{R_A}}(b_A)$ ,  $L_A^{\min} = \text{ml}_{\mathcal{G}}^{\leq}(A)$  and  $N_A = M_{R_A} \setminus L_A^{\min}$ . Furthermore, let  $L_A = R_A \setminus A$  be the latent set with respect to  $A$ .

Let  $B_A = b_A \cup L_A^{\min}$  be the union of the barren vertex  $b_A$  in  $\mathcal{G}_{R_A}$  with the minimal latent set  $L_A^{\min}$ . Let  $C_A = M_{R_A} \setminus B_A$  be the ordered Markov blanket of the barren vertex  $b_A$  excluding the set of minimal latent set  $L_A^{\min}$ . Let  $D_A = \text{de}_{\mathcal{G}_{R_A}}(L_A^{\min}) \setminus L_A^{\min}$  be the proper descendants of the set of minimal latent set  $L_A^{\min}$  contained in the set of preceding variables  $R_A$  with respect to the barren vertex  $b_A$  and the total order  $\leq$ . Let  $F_A = R_A \setminus M_{R_A} D_A$  be the variables in the set of preceding variables  $R_A$  with respect to the barren vertex  $b_A$  and the total order  $\leq$  that have not already been assigned to a set. Accordingly,  $B_A$ ,  $C_A$ ,  $D_A$ , and  $F_A$  partition  $R_A$ . Ultimately, we show that  $B_A \perp\!\!\!\perp F_A \mid C_A [u_{\mathcal{P}u_{\mathcal{N}(\mathcal{G})}^{\leq,+}}]$ .

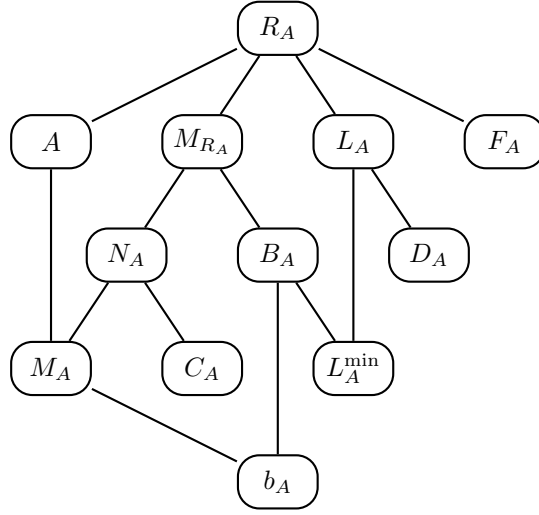


Figure 4.10: The Hasse diagram for the poset over sets ordered by inclusion.

Intuitively, the sets in Figure 4.10 are:

|  |   |
|--|---|
| $A \in \mathcal{A}(\mathcal{G})$                                       | an ancestral set;   |
| $b_A = [A]_{\leq}$   | the last vertex in $A$ with respect to $\leq$ ;                 |
| $R_A = \text{pre}_{\mathcal{G}}^{\leq}(b_A)$                           | the preceding vertices of $b_A$ with respect to $\leq$ ;        |
| $M_A = \text{col}_{\mathcal{G}_A}(b_A)$                                | the maximal $m$ -connecting set with respect to $A$ and $b$ ;   |
| $M_{R_A} = \text{col}_{\mathcal{G}_{R_A}}(b_A)$                        | the maximal $m$ -connecting set with respect to $R_A$ and $b$ ; |
| $L_A = R_A \setminus A$  | the latent set with respect to $A$ ;                            |
| $L_A^{\min} = \text{ml}_{\mathcal{G}}^{\leq}(A)$                       | the minimal latent set with respect to $A$ and $\leq$ ;         |
| $N_A = M_{R_A} \setminus L_A^{\min}$                                   | the maximal non- $m$ -connecting subset of $M_{R_A}$ ;          |
| $B_A = b_A \cup L_A^{\min}$  | the independent set containing $b$ ;                            |
| $C_A = M_{R_A} \setminus B_A$  | the conditioning set;   |
| $D_A = \text{de}_{\mathcal{G}_{R_A}}(L_A^{\min}) \setminus L_A^{\min}$ | the set to be dropped;  |
| $F_A = R_A \setminus M_{R_A} D_A$                                      | the independent set not containing $b_A$ .                      |

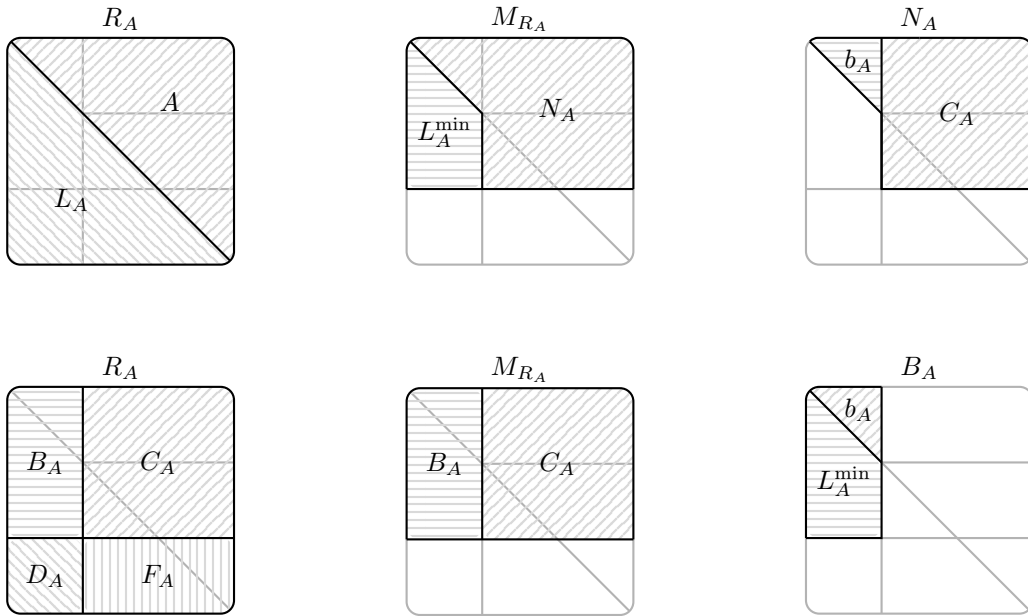


Figure 4.11: An illustration of how various sets interact and partition each other.

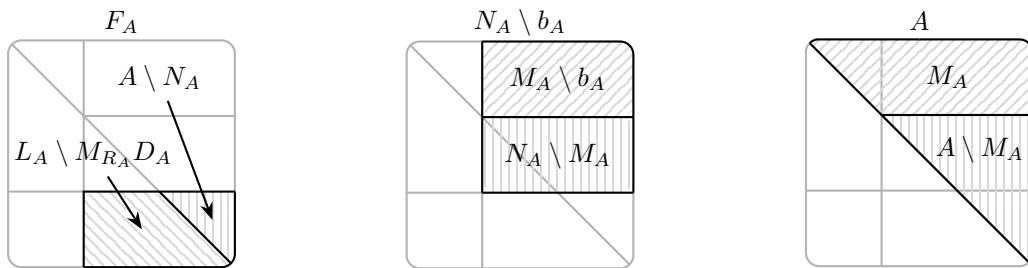


Figure 4.12: An illustration of how various sets interact and partition each other.

$C_A = N_A \setminus b_A$  because  $N_A \subseteq M_{R_A}$ :

$$\begin{aligned}
C_A &= M_{R_A} \setminus B_A \\
&= M_{R_A} \setminus (b_A \cup L_A^{\min}) && (B = b_A \cup L_A^{\min}) \\
&= (M_{R_A} \setminus b_A) \cap (M_{R_A} \setminus L_A^{\min}) && (\text{distributive property}) \\
&= M_{R_A} \setminus b_A \cap N_A && (N_A = M_{R_A} \setminus L_A^{\min}) \\
&= N_A \setminus b_A && (M_{R_A} \cap N_A = N_A)
\end{aligned}$$

$R_A = AL_A$  because  $A \subseteq R_A$ :

$$\begin{aligned}
R_A &= (R_A \cap A) \cup L_A && (L_A = R_A \setminus A) \\
&= AL_A && (R_A \cap A = A)
\end{aligned}$$

$M_{R_A} = N_A L_A^{\min}$  because  $L_A^{\min} = \text{ml}_g^{\leq}(A) \subseteq \text{col}_{g_{R_A}}(b_A) = M_{R_A}$ :

$$\begin{aligned}
M_{R_A} &= N_A \cup (M_{R_A} \cap L_A^{\min}) && (N_A = M_{R_A} \setminus L_A^{\min}) \\
&= N_A L_A^{\min} && (M_{R_A} \cap L_A^{\min} = L_A^{\min})
\end{aligned}$$

In order to facilitate the forthcoming proof we define a few alternative relations.  $F_A = (A \setminus N_A) \cup (L_A \setminus M_{R_A} D_A)$  because  $L_A \cap A = \emptyset$  and  $L_A^{\min} D_A \subseteq L_A$ . Note that  $D_A \subseteq L_A$  because  $D_A = \text{deg}(L_A^{\min})$  and  $A$  is an ancestral set.

$$\begin{aligned}
F_A &= R_A \setminus M_{R_A} D_A \\
&= R_A \setminus (M_{R_A} D_A \cup (L_A \cap A)) && (L_A \cap A = \emptyset) \\
&= R_A \setminus (M_{R_A} D_A L_A \cap M_{R_A} D_A A) && (\text{distributive property}) \\
&= AL_A \setminus (N_A L_A^{\min} D_A L_A \cap M_{R_A} D_A A) && (\text{change notation}) \\
&= AL_A \setminus (N_A L_A \cap M_{R_A} D_A A) && (L_A^{\min} D_A \subseteq L_A) \\
&= (AL_A \setminus N_A L_A) \cup (AL_A \setminus M_{R_A} D_A A) && (\text{distributive property}) \\
&= (A \setminus N_A) \cup (L_A \setminus M_{R_A} D_A) && (\text{simplify differences})
\end{aligned}$$

$N_A \setminus b_A = (N_A \setminus M_A) \cup (M_A \setminus b_A)$  because  $b_A \in M_A \subseteq N_A$ . Note that  $M_A \subseteq N_A$  because

$M_A \subseteq M_{R_A}$  and  $M_A \cap L_A = \emptyset$ .

$$\begin{aligned}
N_A \setminus b_A &= N_A \setminus (M_A \cap b_A) && (b_A = M_A \cap b_A) \\
&= N_A \setminus ((M_A \cap b_A) \cup (M_A \cap (N_A \setminus M_A))) && (M_A \cap (N_A \setminus M_A) = \emptyset) \\
&= N_A \setminus (M_A \cap (b_A \cup (N_A \setminus M_A))) && (\text{distributive property}) \\
&= (N_A \setminus M_A) \cup (N_A \setminus (b_A \cup (N_A \setminus M_A))) && (\text{distributive property}) \\
&= (N_A \setminus M_A) \cup ((N_A \setminus b_A) \cap (N_A \setminus (N_A \setminus M_A))) && (\text{distributive property}) \\
&= (N_A \setminus M_A) \cup ((N_A \setminus b_A) \cap (N_A \cap M_A)) && (N_A \setminus (N_A \setminus M_A) = N_A \cap M_A) \\
&= (N_A \setminus M_A) \cup ((N_A \setminus b_A) \cap M_A) && (N_A \cap M_A = M_A) \\
&= (N_A \setminus M_A) \cup (M_A \setminus b_A) && ((N_A \setminus b_A) \cap M_A = M_A \setminus b_A)
\end{aligned}$$

$A \setminus M_A \subseteq (A \setminus N_A) \cup (N_A \setminus M_A)$  because  $M_A \subseteq N_A$ :

$$\begin{aligned}
A \setminus M_A &= A \setminus (N_A \cap M_A) && (M_A = N_A \cap M_A) \\
&= A \setminus ((N_A \cap M_A) \cup (N_A \cap (A \setminus N_A))) && (N_A \cap (A \setminus N_A) = \emptyset) \\
&= A \setminus (N_A \cap (M_A \cup (A \setminus N_A))) && (\text{distributive property}) \\
&= (A \setminus N_A) \cup (A \setminus (M_A \cup (A \setminus N_A))) && (\text{distributive property}) \\
&= (A \setminus N_A) \cup ((A \setminus M_A) \cap (A \setminus (A \setminus N_A))) && (\text{distributive property}) \\
&= (A \setminus N_A) \cup ((A \setminus M_A) \cap (A \cap N_A)) && (A \setminus (A \setminus N_A) = A \cap N_A) \\
&\subseteq (A \setminus N_A) \cup ((A \setminus M_A) \cap N_A) && (A \cap N_A \subseteq N_A) \\
&= (A \setminus N_A) \cup (N_A \setminus M_A) && ((A \setminus M_A) \cap N_A = N_A \setminus M_A)
\end{aligned}$$

Algorithm 4 outlines a generalized process to extract conditional independence statements from a directed MAG. The conditional independence statements are used to construct a structural imset whose induced independence model is a subset of the induced independence model of the graph and a subset of the independence model induced by the output of Algorithm 3. Furthermore, the conditional independence statements required by the ordered local Markov property are represented in the constructed imset. This is a key result for the formulation of the factorization presented in Section 4.3.4.

---

**Algorithm 4: ORDERED LOCAL MARKOV PROPERTY OLMP( $\mathcal{G}, \leq, A$ )**

---

**Input:** directed MAG:  $\mathcal{G} = (V, E)$ , total order consistent with  $\mathcal{G}: \leq$ ,  
ancestral set:  $A \in \mathcal{A}(\mathcal{G})$

**Output:** structural inset:  $u_A$

- 1 Let  $b_A = \lceil A \rceil_{\leq}$ ,  $R_A = \text{pre}_{\mathcal{G}}^{\leq}(b_A)$ ,  $M_{R_A} = \text{col}_{\mathcal{G}_{R_A}}(b_A)$ ,  $L_A^{\min} = \text{ml}_{\mathcal{G}}^{\leq}(A)$ ,  
 $N_A = M_{R_A} \setminus L_A^{\min}$  ;
- 2 Let  $B_A = b_A \cup L_A^{\min}$ ,  $C_A = M_{R_A} \setminus B_A$ ,  $D_A = \text{deg}_{\mathcal{G}_{R_A}}(L_A^{\min}) \setminus L_A^{\min}$ ,  
 $F_A = R_A \setminus M_{R_A} D_A$  ;
- 3 Initialize inset  $u_A : \mathcal{P}(V) \rightarrow 0$  ;
- 4 Let  $i = 1$ ,  $r_i^A = \lfloor B_A \rfloor_{\leq}$ ,  $R_i^A = \text{pre}_{\mathcal{G}}^{\leq}(r_i^A)$  ;
- 5 **repeat**
- 6   Let  $B_i^A = B_A \cap R_i^A$ ,  $C_i^A = C_A \cap R_i^A$ ,  $D_i^A = D_A \cap R_i^A$ ,  $F_i^A = F_A \cap R_i^A$  ;
- 7   **if**  $r_i^A \in \text{dis}_{\mathcal{G}_{R_A}}(b_A)$  **then**
- 8     Let  $M_i^A = \text{col}_{\mathcal{G}_{R_i^A}}(r_i^A)$  ;
- 9      $u_A = u_A + u_{\langle r_i^A, R_i^A \setminus M_i^A | M_i^A \setminus r_i^A \rangle}$  // Lemma 4.3.5 ;
- 10      $u_A = u_A + u_{\langle r_i^A, F_i^A | B_i^A C_i^A \setminus r_i^A \rangle}$  // decomposition and weak union ;
- 11     **if**  $r_i^A \in C_A$  **then**
- 12        $u_A = u_A + u_{\langle r_i^A \cup B_i^A, F_i^A | C_{i-1}^A \rangle}$  // contraction ;
- 13     **end**
- 14   **else if**  $r_i^A \in C_A F_A$  **then**
- 15     Let  $A' = R_i^A \setminus D_A$  ;
- 16      $u_A = u_A + \text{OLMP}(\mathcal{G}, \leq, A')$  // recursive call ;
- 17      $u_A = u_A + u_{\langle B_i^A, r_i^A | C_i^A F_i^A \setminus r_i^A \rangle}$  // decomposition and weak union ;
- 18     **if**  $r_i^A \in C_A$  **then**
- 19        $u_A = u_A + u_{\langle B_i^A, r_i^A \cup F_i^A | C_{i-1}^A \rangle}$  // contraction ;
- 20     **end**
- 21   **end**
- 22    $u_A = u_A + u_{\langle B_i^A, F_i^A | C_i^A \rangle}$  // weak union or contraction ;
- 23   **if**  $r_i^A \neq b_A$  **then**
- 24     Let  $i = i + 1$ ,  $r_i^A = \lfloor R_A \setminus R_{i-1}^A \rfloor_{\leq}$ ,  $R_i^A = \text{pre}_{\mathcal{G}}^{\leq}(r_i^A)$  ;
- 25   **end**
- 26 **until**  $r_i^A = b_A$  ;
- 27 Let  $M_A = \text{col}_{\mathcal{G}_A}(b_A)$  ;
- 28  $u_A = u_A + u_{\langle b_A, A \setminus N_A | N_A \setminus b_A \rangle}$  // decomposition ;
- 29  $u_A = u_A + u_{\langle b_A, N_A \setminus M_A | M_A \setminus b_A \rangle}$  // Lemma 4.3.5 ;
- 30  $u_A = u_A + u_{\langle b_A, A \setminus M_A | M_A \setminus b_A \rangle}$  // contraction ;

---

Applications of the symmetry semi-graphoid axiom are not noted in the algorithm.



In the following series of figures, we give an illustrative example of the steps of Algorithm 4. Let  $\mathcal{G} = (V, E)$  be a directed MAG,  $\leq$  be a total order consistent with  $\mathcal{G}$ , and  $A_1, A_2 \in \mathcal{A}(\mathcal{G})$  be ancestral sets. Additionally, let  $\mathcal{P} = \mathcal{P}(V)$  be the poset ordered by inclusion. We construct the structural imset  $u_{A_1}$  by adding semi-elementary imsets to  $u_{A_1}$  throughout Algorithm 4. Note that  $u_{A_1}$  is guaranteed to be structural since it is constructed as a linear combination of semi-elementary imsets to  $u_{A_1}$  with positive integral coefficients.

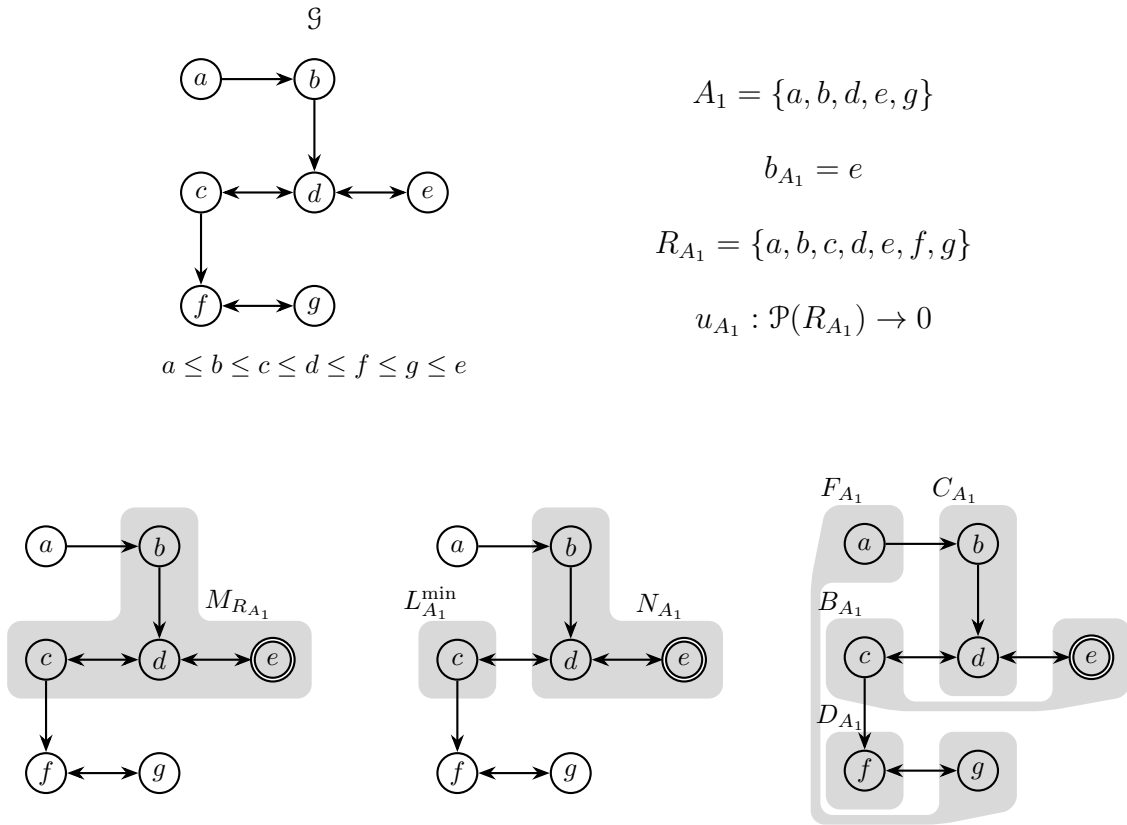


Figure 4.13: An illustration of the setup of  $\text{OLMP}(\mathcal{G}, \leq, A_1)$  (step  $i$ ).

Figure 4.13 initializes many of the sets used throughout the example for  $\text{OLMP}(\mathcal{G}, \leq, A_1)$ .

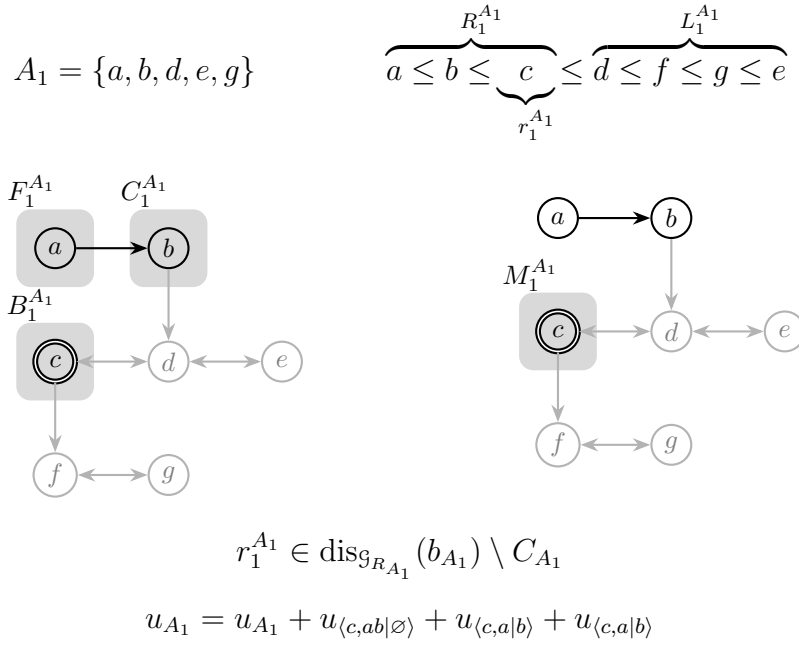


Figure 4.14: An illustration of  $\text{OLMP}(\mathcal{G}, \leq, A_1)$  (step *ii*).

Figure 4.14 depicts the first step of  $\text{OLMP}(\mathcal{G}, \leq, A_1)$  where  $r_1^{A_1} = c$ ,  $R_1^{A_1} = \{a, b, c\}$ , and  $M_1^{A_1} = \{c\}$ . Note that  $r_1^{A_1} \in \text{dis}_{\mathcal{G}_{R_{A_1}}}(b_{A_1}) \setminus C_{A_1}$ . Semi-elementary imsets corresponding to the following conditional independence statements are added to  $u_{A_1}$ :

line 9 :  $r_1^{A_1} \perp\!\!\!\perp R_1^{A_1} \setminus M_1^{A_1} \mid M_1^{A_1} \setminus r_1^{A_1} [\mu_{\mathcal{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq, +}]$   
line 10 :  $r_1^{A_1} \perp\!\!\!\perp F_1^{A_1} \mid B_1^{A_1} C_1^{A_1} [\mu_{\mathcal{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq, +}]$   
line 22 :  $B_1^{A_1} \perp\!\!\!\perp F_1^{A_1} \mid C_1^{A_1} [\mu_{\mathcal{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq, +}]$

Instantiating the sets:

line 9 :  $c \perp\!\!\!\perp ab \mid \emptyset [\mu_{\mathcal{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq, +}]$  (Lemma 4.3.5)  
line 10 :  $c \perp\!\!\!\perp a \mid b [\mu_{\mathcal{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq, +}]$  (weak union—(line 9))  
line 22 :  $c \perp\!\!\!\perp a \mid b [\mu_{\mathcal{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq, +}]$  (line 10)

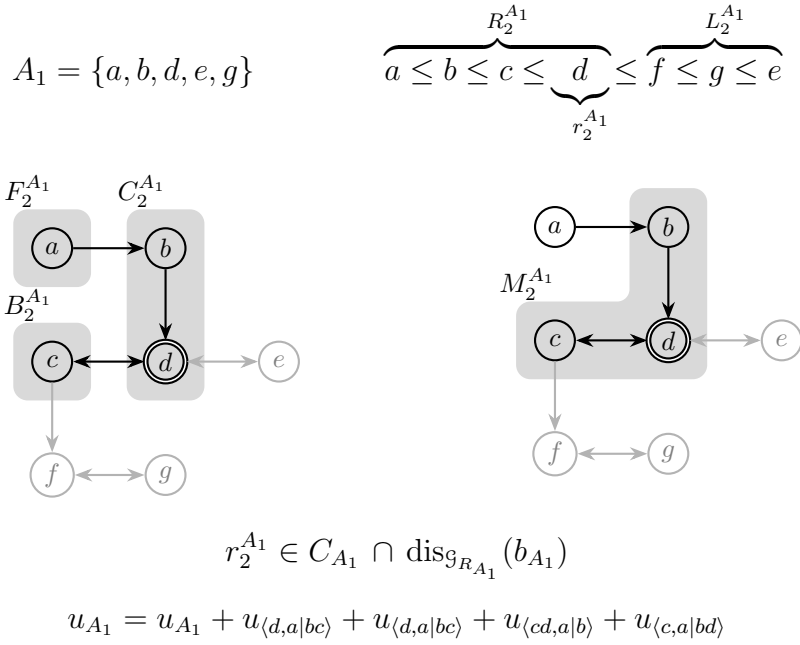


Figure 4.15: An illustration of  $\text{OLMP}(\mathcal{G}, \leq, A_1)$  (step *iii*).

Figure 4.15 depicts the step of  $\text{OLMP}(\mathcal{G}, \leq, A_1)$  where  $r_2^{A_1} = d$ ,  $R_2^{A_1} = \{a, b, c, d\}$ , and  $M_2^{A_1} = \{b, c, d\}$ . Note that  $r_2^{A_1} \in C_{A_1} \cap \text{dis}_{\mathcal{G}_{R_{A_1}}}(b_{A_1})$  and  $C_1^{A_1} = C_2^{A_1} \setminus r_2^{A_1}$ . Semi-elementary imsets corresponding to the following conditional independence statements are added to  $u_{A_1}$ :

- line 9 :  $r_2^{A_1} \perp\!\!\!\perp R_2^{A_1} \setminus M_2^{A_1} \mid M_2^{A_1} \setminus r_2^{A_1} [\mu_{\mathcal{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq,+}]$
- line 10 :  $r_2^{A_1} \perp\!\!\!\perp F_2^{A_1} \mid B_2^{A_1} C_2^{A_1} [\mu_{\mathcal{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq,+}]$
- line 12 :  $r_2^{A_1} \cup B_2^{A_1} \perp\!\!\!\perp F_2^{A_1} \mid C_1^{A_1} [\mu_{\mathcal{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq,+}]$
- line 22 :  $B_2^{A_1} \perp\!\!\!\perp F_2^{A_1} \mid C_2^{A_1} [\mu_{\mathcal{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq,+}]$

Instantiating the sets:

- line 9 :  $d \perp\!\!\!\perp a \mid bc \left[ \mu_{\mathcal{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq,+} \right]$  (Lemma 4.3.5)
- line 10 :  $d \perp\!\!\!\perp a \mid bc \left[ \mu_{\mathcal{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq,+} \right]$  (line 9)
- line 12a :  $c \perp\!\!\!\perp a \mid b \left[ \mu_{\mathcal{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq,+} \right]$  (previous step—(step *ii*))
- line 12b :  $cd \perp\!\!\!\perp a \mid b \left[ \mu_{\mathcal{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq,+} \right]$  (contraction—(line 10 + line 12a))
- line 22 :  $c \perp\!\!\!\perp a \mid bd \left[ \mu_{\mathcal{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq,+} \right]$  (weak union—(line 12b))

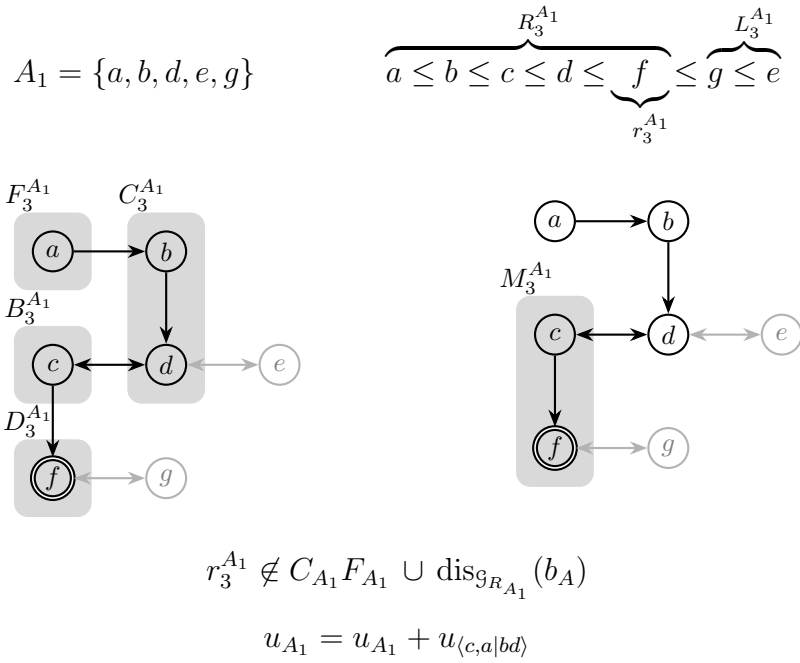


Figure 4.16: An illustration of  $\text{OLMP}(\mathcal{G}, \leq, A_1)$  (step *iv*).

Figure 4.16 depicts the step of  $\text{OLMP}(\mathcal{G}, \leq, A_1)$  where  $r_3^{A_1} = f$ ,  $R_3^{A_1} = \{a, b, c, d, f\}$ , and  $M_3^{A_1} = \{c, f\}$ . Note that  $r_3^{A_1} \notin C_{A_1} F_{A_1} \cup \text{dis}_{\mathcal{G}_{R_{A_1}}}(b_A)$ . A semi-elementary imset corresponding to the following conditional independence statement is added to  $u_{A_1}$ :

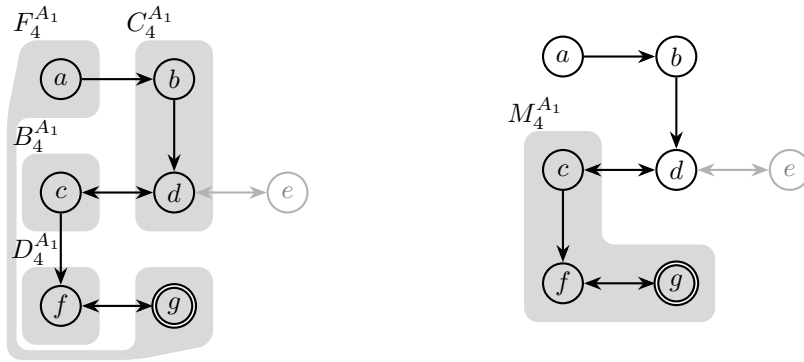
$$\text{line 22 : } B_3^{A_1} \perp\!\!\!\perp F_3^{A_1} \mid C_3^{A_1} \left[ \mu_{\mathcal{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq,+} \right]$$

Instantiating the sets:

line 22a :  $c \perp\!\!\!\perp a \mid bd [\mu_{\mathcal{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq,+}]$  (previous step—(step *iii*))

line 22b :  $c \perp\!\!\!\perp a \mid bd [\mu_{\mathcal{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq,+}]$  (line 22a)

$$A_1 = \{a, b, d, e, g\} \quad \overbrace{a \leq b \leq c \leq d \leq f \leq g}^{R_4^{A_1}} \leq \underbrace{e}_{L_4^{A_1}} \leq \underbrace{g}_{r_4^{A_1}}$$



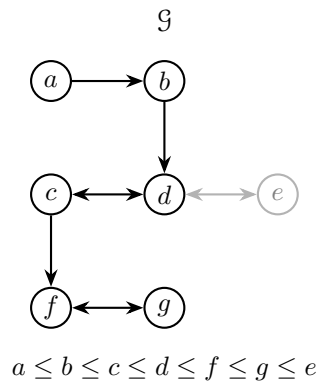
$$r_4^{A_1} \in F_{A_1} \setminus \text{dis}_{\mathcal{G}_{R_{A_1}}}(b_{A_1})$$

$$A_2 = \{a, b, c, d, g\}$$

$$u_{A_1} = u_{A_1} + \text{OLMP}(\mathcal{G}, \leq, A_2)$$

Figure 4.17: An illustration of  $\text{OLMP}(\mathcal{G}, \leq, A_1)$  (step *v*).

Figure 4.17 depicts the step of  $\text{OLMP}(\mathcal{G}, \leq, A_1)$  where  $r_4^{A_1} = g$ ,  $R_4^{A_1} = \{a, b, c, d, f, g\}$ , and  $M_4^{A_1} = \{c, f, g\}$ . Note that  $r_4^{A_1} \in F_{A_1} \setminus \text{dis}_{\mathcal{G}_{R_{A_1}}}(b_{A_1})$ . Algorithm 4 performs a recursive call with  $A_2 = \{a, b, c, d, g\}$ .



$$A_2 = \{a, b, c, d, g\}$$

$$b_{A_2} = g$$

$$R_{A_2} = \{a, b, c, d, f, g\}$$

$$u_{A_2} : \mathcal{P}(R_{A_2}) \rightarrow 0$$

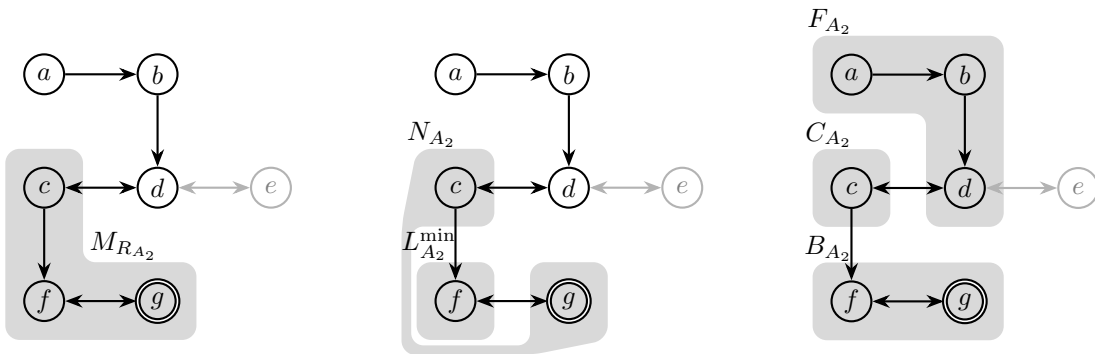


Figure 4.18: An illustration of the setup of  $\text{OLMP}(\mathcal{G}, \leq, A_2)$  (step *vi*).

Figure 4.18 initializes many of the sets used throughout the example for  $\text{OLMP}(\mathcal{G}, \leq, A_2)$ .

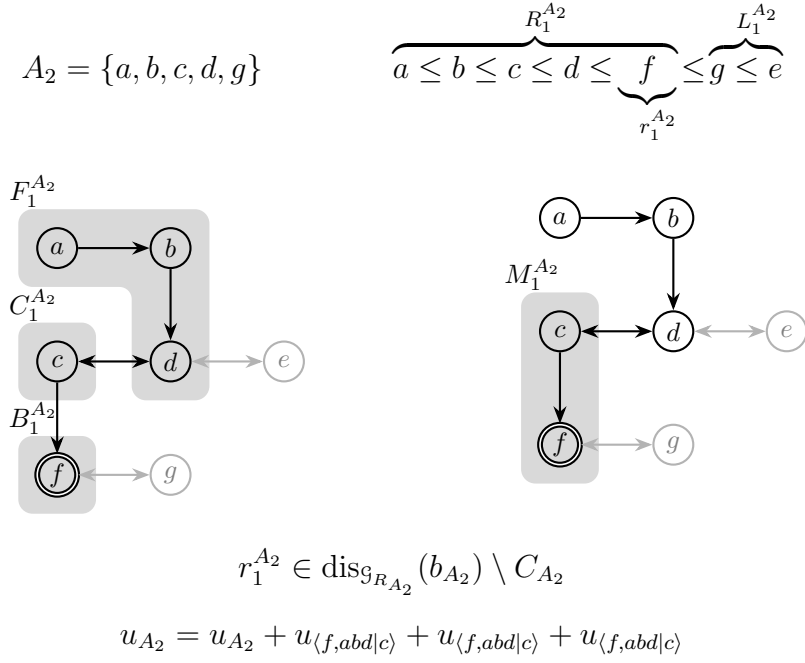


Figure 4.19: An illustration of  $\text{OLMP}(\mathcal{G}, \leq, A_2)$  (step *vii*).

Figure 4.19 depicts the first step of  $\text{OLMP}(\mathcal{G}, \leq, A_2)$  where  $r_1^{A_2} = f$ ,  $R_1^{A_2} = \{a, b, c, d, f\}$ , and  $M_1^{A_2} = \{c, f\}$ . Note that  $r_1^{A_2} \in \text{dis}_{\mathcal{G}_{R_{A_2}}}(b_{A_2}) \setminus C_{A_2}$ . Semi-elementary insets corresponding to the following conditional independence statements are added to  $u_{A_2}$ :

$$\text{line 9 : } r_1^{A_2} \perp\!\!\!\perp R_1^{A_2} \setminus M_1^{A_2} \mid M_1^{A_2} \setminus r_1^{A_2} [\mu_{\mathcal{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq, +}]$$

$$\text{line 10 : } r_1^{A_2} \perp\!\!\!\perp F_1^{A_2} \mid B_1^{A_2} C_1^{A_2} [\mu_{\mathcal{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq, +}]$$

$$\text{line 22 : } B_1^{A_2} \perp\!\!\!\perp F_1^{A_2} \mid C_1^{A_2} [\mu_{\mathcal{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq, +}]$$

Instantiating the sets:

$$\text{line 9 : } f \perp\!\!\!\perp abd \mid c [\mu_{\mathcal{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq, +}] \quad (\text{Lemma 4.3.5})$$

$$\text{line 10 : } f \perp\!\!\!\perp abd \mid c [\mu_{\mathcal{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq, +}] \quad (\text{line 9})$$

$$\text{line 22 : } f \perp\!\!\!\perp abd \mid c [\mu_{\mathcal{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq, +}] \quad (\text{line 10})$$

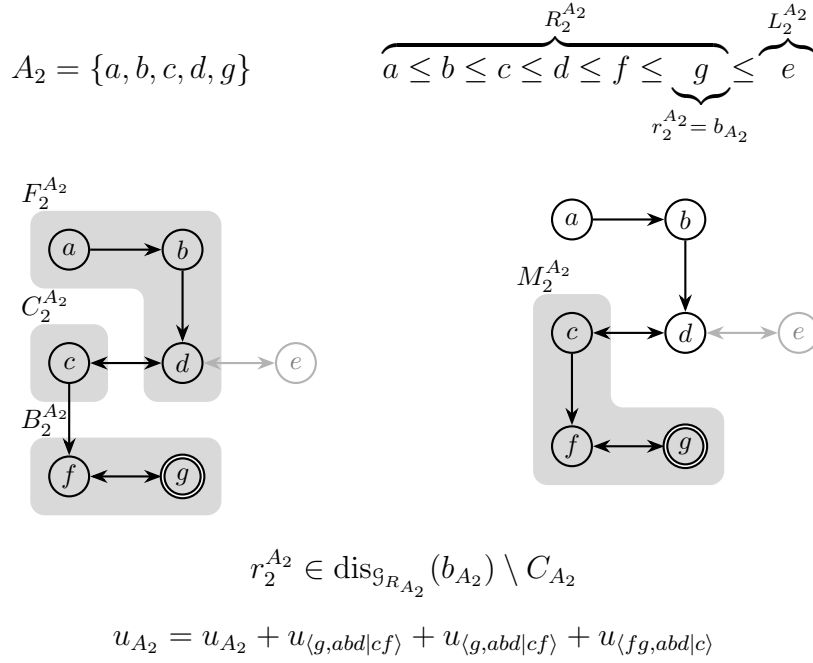


Figure 4.20: An illustration of  $\text{OLMP}(\mathcal{G}, \leq, A_2)$  (step *viii*).

Figure 4.20 depicts the step of  $\text{OLMP}(\mathcal{G}, \leq, A_2)$  where  $r_2^{A_2} = g$ ,  $R_2^{A_2} = \{a, b, c, d, f, g\}$ , and  $M_2^{A_2} = \{c, f, g\}$ . Note that  $r_2^{A_2} \in \text{dis}_{\mathcal{G}_{R_{A_2}}}(b_{A_2}) \setminus C_{A_2}$ . Semi-elementary insets corresponding to the following conditional independence statements are added to  $u_{A_2}$ :

line 9 :  $r_2^{A_2} \perp\!\!\!\perp R_2^{A_2} \setminus M_2^{A_2} \mid M_2^{A_2} \setminus r_2^{A_2} [\mu_{\mathcal{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq,+}]$

line 10 :  $r_2^{A_2} \perp\!\!\!\perp F_2^{A_2} \mid B_2^{A_2} C_2^{A_2} [\mu_{\mathcal{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq,+}]$

line 22 :  $B_2^{A_2} \perp\!\!\!\perp F_2^{A_2} \mid C_2^{A_2} [\mu_{\mathcal{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq,+}]$

Instantiating the sets:

line 9 :  $g \perp\!\!\!\perp abd \mid cf [\mu_{\mathcal{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq,+}]$  (Lemma 4.3.5)

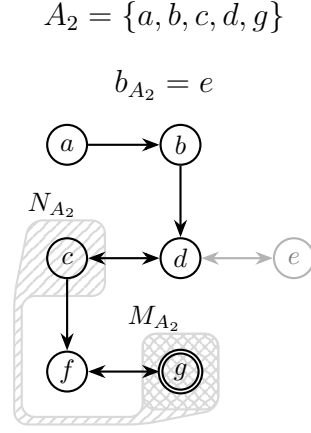
line 10 :  $g \perp\!\!\!\perp abd \mid cf [\mu_{\mathcal{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq,+}]$ . (line 9)

line 22a :  $f \perp\!\!\!\perp abd \mid c [\mu_{\mathcal{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq,+}]$  (previous step—(step *vii*))

line 22b :  $fg \perp\!\!\!\perp abd \mid c [\mu_{\mathcal{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq,+}]$  (contraction—(line 10 + line 22a))



Since  $R_2 = b_{A_2}$ , the main loop of  $\text{OLMP}(\mathcal{G}, \leq, A_2)$  is done.



$$u_{A_2} = u_{A_2} + u_{\langle g, abd|c \rangle} + u_{\langle g, c|\emptyset \rangle} + u_{\langle g, abcd|\emptyset \rangle}$$

Figure 4.21: An illustration of  $\text{OLMP}(\mathcal{G}, \leq, A_2)$  (step *ix*).

Figure 4.21 depicts the step of  $\text{OLMP}(\mathcal{G}, \leq, A_2)$  after completing the main loop. Semi-elementary imsets corresponding to the following conditional independence statements are added to  $u_{A_2}$ :

$$\text{line 28 : } b_{A_2} \perp\!\!\!\perp A_2 \setminus N_{A_2} \mid N_{A_2} \setminus b_{A_2} [\mu_{\mathbb{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq,+}]$$

$$\text{line 29 : } b_{A_2} \perp\!\!\!\perp N_{A_2} \setminus M_{A_2} \mid M_{A_2} \setminus b_{A_2} [\mu_{\mathbb{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq,+}]$$

$$\text{line 30 : } b_{A_2} \perp\!\!\!\perp A_2 \setminus M_{A_2} \mid M_{A_2} \setminus b_{A_2} [\mu_{\mathbb{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq,+}]$$

Instantiating the sets:

$$\text{line 28a : } fg \perp\!\!\!\perp abd \mid c [\mu_{\mathbb{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq,+}] \quad (\text{previous step—(step } viii))$$

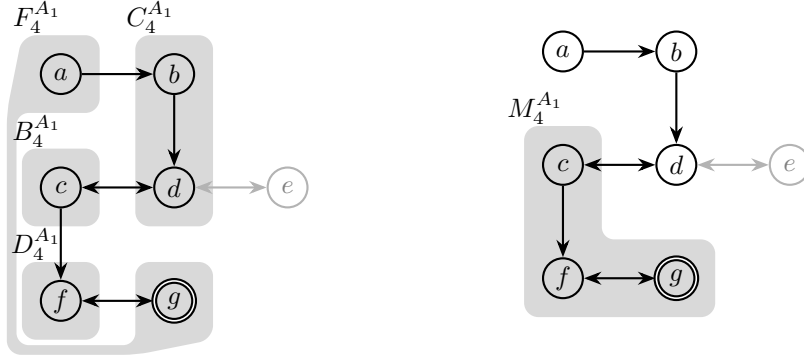
$$\text{line 28b : } g \perp\!\!\!\perp abd \mid c [\mu_{\mathbb{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq,+}] \quad (\text{decomposition—(line 28a)})$$

$$\text{line 29 : } g \perp\!\!\!\perp c \mid \emptyset [\mu_{\mathbb{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq,+}] \quad (\text{Lemma 4.3.5})$$

$$\text{line 30 : } g \perp\!\!\!\perp abcd \mid \emptyset [\mu_{\mathbb{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq,+}] \quad (\text{contraction—(line 28b + line 29)})$$

$$A_1 = \{a, b, d, e, g\} \quad \overbrace{a \leq b \leq c \leq d \leq f \leq g}^{R_4^{A_1}} \leq \overbrace{e}^{L_4^{A_1}}$$

$r_4^{A_1}$



$$r_4^{A_1} \in F_{A_1} \setminus \text{dis}_{\mathcal{G}_{R_{A_1}}}(b_{A_1})$$

$$u_{A_1} = u_{A_1} + \text{OLMP}(\mathcal{G}, \leq, A_2)$$

$$u_{A_1} = u_{A_1} + u_{\langle c, g | abd \rangle} + u_{\langle c, ag | bd \rangle}$$

Figure 4.22: An illustration of  $\text{OLMP}(\mathcal{G}, \leq, A_1)$  (step  $x$ ).

Figure 4.22 depicts the step of  $\text{OLMP}(\mathcal{G}, \leq, A_1)$  where  $r_4^{A_1} = g$ ,  $R_4^{A_1} = \{a, b, c, d, f, g\}$ , and  $M_4^{A_1} = \{c, f, g\}$ . Algorithm 4 returns to this step after a recursive call. Note that  $r_4^{A_1} \in F_{A_1} \setminus \text{dis}_{\mathcal{G}_{R_{A_1}}}(b_{A_1})$ . Semi-elementary insets corresponding to the following conditional independence statements are added to  $u_{A_1}$ :

$$\text{line 17 : } B_4^{A_1} \perp\!\!\!\perp r_4^{A_1} \mid C_4^{A_1} F_4^{A_1} \setminus r_4^{A_1} [\mu_{\mathcal{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq, +}]$$

$$\text{line 22 : } B_4^{A_1} \perp\!\!\!\perp F_4^{A_1} \mid C_4^{A_1} [\mu_{\mathcal{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq, +}]$$

Instantiating the sets:

|  |                                     |
|--|-------------------------------------|
| line 17a : $abcd \perp\!\!\!\perp g \mid \emptyset [\mu_{\mathcal{P}}u_{\mathcal{N}(g)}^{\leq,+}]$ | (recursive call—(step $ix$ ))       |
| line 17b : $c \perp\!\!\!\perp g \mid abd [\mu_{\mathcal{P}}u_{\mathcal{N}(g)}^{\leq,+}]$          | (weak union—(line 17a))             |
| line 22a : $c \perp\!\!\!\perp a \mid bd [\mu_{\mathcal{P}}u_{\mathcal{N}(g)}^{\leq,+}]$           | (previous step—(step $iv$ ))        |
| line 22b : $c \perp\!\!\!\perp ag \mid bd [\mu_{\mathcal{P}}u_{\mathcal{N}(g)}^{\leq,+}]$          | (contraction—(line 17b + line 22a)) |

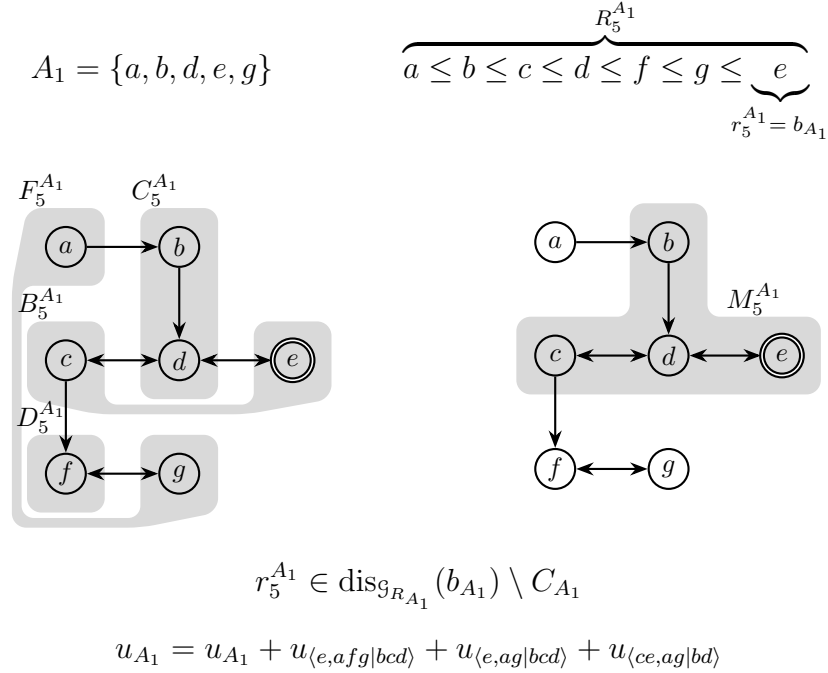


Figure 4.23: An illustration of  $\text{OLMP}(\mathcal{G}, \leq, A_1)$  (step  $xi$ ).

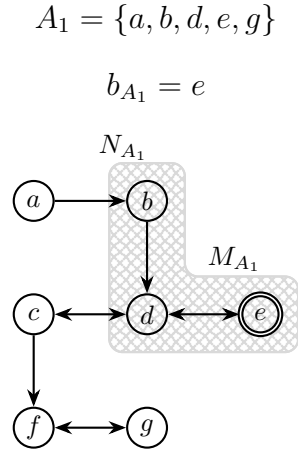
Figure 4.23 depicts the step of  $\text{OLMP}(\mathcal{G}, \leq, A_1)$  where  $r_5^{A_1} = e$ ,  $R_5^{A_1} = \{a, b, c, d, e, f, g\}$ , and  $M_5^{A_1} = \{b, c, d, e\}$ . Note that  $r_5^{A_1} \in \text{dis}_{\mathcal{G}_{R_{A_1}}}(b_{A_1}) \setminus C_{A_1}$ . Semi-elementary imsets corresponding to the following conditional independence statements are added to  $u_{A_1}$ :

|   |
|---|
| line 9 : $r_5^{A_1} \perp\!\!\!\perp R_5^{A_1} \setminus M_5^{A_1} \mid M_5^{A_1} \setminus r_5^{A_1} [\mu_{\mathcal{P}}u_{\mathcal{N}(g)}^{\leq,+}]$ |
| line 10 : $r_5^{A_1} \perp\!\!\!\perp F_5^{A_1} \mid B_5^{A_1} C_5^{A_1} \setminus r_5^{A_1} [\mu_{\mathcal{P}}u_{\mathcal{N}(g)}^{\leq,+}]$          |
| line 22 : $B_5^{A_1} \perp\!\!\!\perp F_5^{A_1} \mid C_5^{A_1} [\mu_{\mathcal{P}}u_{\mathcal{N}(g)}^{\leq,+}]$  |

Instantiating the sets:

- line 9 :  $e \perp\!\!\!\perp afg \mid bcd [\mu_{\mathcal{P}}u_{\mathcal{N}(\mathcal{G})}^{\leq,+}]$  (Lemma 4.3.5)  
line 10 :  $e \perp\!\!\!\perp ag \mid bcd [\mu_{\mathcal{P}}u_{\mathcal{N}(\mathcal{G})}^{\leq,+}]$  (decomposition—(line 9))  
line 22a :  $c \perp\!\!\!\perp ag \mid bd [\mu_{\mathcal{P}}u_{\mathcal{N}(\mathcal{G})}^{\leq,+}]$  (previous step—(step  $x$ ))  
line 22b :  $ce \perp\!\!\!\perp ag \mid bd [\mu_{\mathcal{P}}u_{\mathcal{N}(\mathcal{G})}^{\leq,+}]$  (contraction—(line 10 + line 22a))

Since  $r_5 = b_{A_1}$ , the main loop of  $\text{OLMP}(\mathcal{G}, \leq, A_1)$  is done.



$$u_{A_1} = u_{A_1} + u_{\langle e, ag \mid bd \rangle} + u_{\langle e, \emptyset \mid bd \rangle} + u_{\langle e, ag \mid bd \rangle}$$

Figure 4.24: An illustration of  $\text{OLMP}(\mathcal{G}, \leq, A_1)$  (step  $xii$ ).

Figure 4.24 depicts the step of  $\text{OLMP}(\mathcal{G}, \leq, A_1)$  after completing the main loop. Semi-elementary imsets corresponding to the following conditional independence statements are added to  $u_{A_1}$ :

- line 28 :  $b_{A_1} \perp\!\!\!\perp A_1 \setminus N_{A_1} \mid N_{A_1} \setminus b_{A_1} [\mu_{\mathcal{P}}u_{\mathcal{N}(\mathcal{G})}^{\leq,+}]$   
line 29 :  $b_{A_1} \perp\!\!\!\perp N_{A_1} \setminus M_{A_1} \mid M_{A_1} \setminus b_{A_1} [\mu_{\mathcal{P}}u_{\mathcal{N}(\mathcal{G})}^{\leq,+}]$   
line 30 :  $b_{A_1} \perp\!\!\!\perp A_1 \setminus M_{A_1} \mid M_{A_1} \setminus b_{A_1} [\mu_{\mathcal{P}}u_{\mathcal{N}(\mathcal{G})}^{\leq,+}]$

Instantiating the sets:

$$\begin{array}{ll}
\text{line 28a : } ce \perp\!\!\!\perp ag \mid bd [\mu_{\mathbf{P}}u_{\mathcal{N}(\mathcal{G})}^{\leq,+}] & (\text{previous step—(step } xi)) \\
\text{line 28b : } e \perp\!\!\!\perp ag \mid bd [\mu_{\mathbf{P}}u_{\mathcal{N}(\mathcal{G})}^{\leq,+}] & (\text{decomposition—(line 28a)}) \\
\text{line 29 : } e \perp\!\!\!\perp \emptyset \mid bd [\mu_{\mathbf{P}}u_{\mathcal{N}(\mathcal{G})}^{\leq,+}] & (\text{Lemma 4.3.5}) \\
\text{line 30 : } e \perp\!\!\!\perp ag \mid bd [\mu_{\mathbf{P}}u_{\mathcal{N}(\mathcal{G})}^{\leq,+}] & (\text{line 28b})
\end{array}$$

**Lemma 4.3.6.** *Let  $\mathcal{G} = (V, E)$  be a directed MAG and  $\leq$  be a total order consistent with  $\mathcal{G}$ . Let  $A \in \mathcal{A}(\mathcal{G})$  and  $b = \lceil A \rceil_{\leq}$  with preceding vertices  $R = \text{pre}_{\mathcal{G}}^{\leq}(b)$ . Let  $r \in R$  and  $R' = \text{pre}_{\mathcal{G}}^{\leq}(r)$ . If  $r \in \text{dis}_{\mathcal{G}_R}(b)$ , then:*

$$\text{col}_{\mathcal{G}_{R'}}(r) \subseteq \text{col}_{\mathcal{G}_R}(b).$$

*If  $r \notin \text{dis}_{\mathcal{G}_R}(b)$ , and  $B = b \cup \text{ml}_{\mathcal{G}}^{\leq}(A)$ :*

$$\text{col}_{\mathcal{G}_{R'}}(r) \cap B = \emptyset.$$

*Proof.* Note that  $\mathcal{G}_{R'}$  is a subgraph of  $\mathcal{G}_R$  so any vertices and paths in  $\mathcal{G}_{R'}$  are in  $\mathcal{G}_R$ . Pick vertex  $a \in \text{col}_{\mathcal{G}_{R'}}(r)$  and path  $\pi_{ar}$  in  $\mathcal{G}_{R'}$  between  $a$  and  $r$  such that  $\pi_{ar}$  is a collider-connecting path. Furthermore,  $r = \lceil R' \rceil_{\leq}$  so  $\pi_{ar}$  must have an arrowhead directed into  $r$ .

For the first statement, we show that  $a \in \text{col}_{\mathcal{G}_R}(b)$ . Since  $r \in \text{dis}_{\mathcal{G}_R}(b)$ , there is a path  $\pi_{br}$  in  $\mathcal{G}_R$  between  $b$  and  $r$  consisting entirely of bi-directed edges. Accordingly, the composition of  $\pi_{ar}$  from  $a$  to  $r$  with  $\pi_{br}$  from  $r$  to  $b$  is a collider-connecting path between  $a$  and  $b$  in  $\mathcal{G}_R$ . It follows that  $\text{col}_{\mathcal{G}_{R'}}(r) \subseteq \text{col}_{\mathcal{G}_R}(b)$ .

For the second statement, we show that if  $a \in B$ , then  $r \in \text{dis}_{\mathcal{G}_R}(b)$ ; this is the contrapositive statement. Since  $a \in B$ , there is a path  $\pi_{ab}$  in  $\mathcal{G}_R$  between  $a$  and  $b$  consisting entirely of bi-directed edges. Accordingly, the composition of  $\pi_{ab}$  from  $b$  to  $a$  with  $\pi_{ar}$  from  $a$  to  $r$  is a collider-connecting path between  $b$  and  $r$  in  $\mathcal{G}_R$ . It follows that  $\text{col}_{\mathcal{G}_{R'}}(r) \cap B = \emptyset$   $\square$

**Lemma 4.3.7.** *Let  $\mathcal{G} = (V, E)$  be a directed MAG,  $\mathbf{P} = \mathcal{P}(V)$  be the poset ordered by inclusion,  $\leq$  be a total order consistent with  $\mathcal{G}$ , and  $u_{\mathcal{N}(\mathcal{G})}^{\leq,+}, u_{\mathcal{N}(\mathcal{G})}^{\leq,-} = \text{NSI}(\mathcal{G}, \leq)$  be the imsets constructed by Algorithm 3. If  $A \in \mathcal{A}(\mathcal{G})$  is an ancestral set and  $u_A = \text{OLMP}(\mathcal{G}, \leq, A)$  is the imset constructed by Algorithm 4, then  $u_A$  is a structural imset and  $\mathcal{J}(u_A) \subseteq \mathcal{J}(\mu_{\mathbf{P}}u_{\mathcal{N}(\mathcal{G})}^{\leq,+})$ .*

Furthermore, if  $b_A = \lceil A \rceil_{\leq}$  and  $M_A = \text{col}_{\mathfrak{g}_A}(b_A)$ , then:

$$b_A \perp\!\!\!\perp A \setminus M_A \mid M_A \setminus b_A [u_A].$$

*Proof.* Since  $u_A$  is defined as the sum of semi-elementary imsets,  $u_A$  is a structural imset. Additionally,  $b_A \perp\!\!\!\perp A \setminus M_A \mid M_A \setminus b_A [u_A]$  by line 30. Consider the recursive call on line 16: Note that  $A'$  is ancestral because it is defined as an ancestral set minus a set that contains all of its descendants. Let  $b_{A'} = \lceil A' \rceil_{\leq}$  and  $R_{A'} = \text{pre}_{\mathfrak{g}}^{\leq}(b_{A'})$  and note that  $R_{A'} \subset R_A$ . Accordingly, each time the algorithm is called recursively, the set of preceding variables is smaller. Since these sets are finite, Algorithm 4 is guaranteed to terminate.

We show that the conditional independence statement represented by semi-elementary imset added to  $u_A$  are either represented in  $\mu_{\mathcal{P}} u_{\mathcal{N}(g)}^{\leq,+}$  by Lemma 4.3.5 or implied by preexisting conditional independence statements represented in  $u$ . Accordingly,  $\mathcal{J}(u_A) \subseteq \mathcal{J}(\mu_{\mathcal{P}} u_{\mathcal{N}(g)}^{\leq,+})$ .

Let  $R_A = \text{pre}_{\mathfrak{g}}^{\leq}(b_A)$ ,  $M_{R_A} = \text{col}_{\mathfrak{g}_{R_A}}(b_A)$ ,  $L_A^{\min} = \text{ml}_{\mathfrak{g}}^{\leq}(A)$ ,  $N_A = M_{R_A} \setminus L_A^{\min}$ , and  $L_A = R_A \setminus A$ . Let  $B_A = b_A \cup L_A^{\min}$ ,  $C_A = M_{R_A} \setminus B_A$ ,  $D_A = \text{de}_{\mathfrak{g}_{R_A}}(L_A^{\min}) \setminus L_A^{\min}$ ,  $F_A = R_A \setminus M_{R_A} D_A$ .

We proceed by induction. For the base case, let  $r_1^A = \lceil B_A \rceil_{\leq}$ ,  $R_1^A = \text{pre}_{\mathfrak{g}}^{\leq}(r_1^A)$ , and  $M_1^A = \text{col}_{\mathfrak{g}_{R_1^A}}(r_1^A)$ . Let  $B_1^A = B_A \cap R_1^A$ ,  $C_1^A = C_A \cap R_1^A$ ,  $D_1^A = D_A \cap R_1^A$ , and  $F_1^A = F_A \cap R_1^A$  be the sets constrained to the set of variables  $R_1^A$ . Note that  $r_1^A \in \text{dis}_{\mathfrak{g}_{R_1^A}}(b_A)$ . By Lemma 4.3.5,

$$r_1^A \perp\!\!\!\perp R_1^A \setminus M_1^A \mid M_1^A \setminus r_1^A [\mu_{\mathcal{P}} u_{\mathcal{N}(g)}^{\leq,+}].$$

Thus line 9 is satisfied. By changing notation,

$$r_1^A \perp\!\!\!\perp B_1^A C_1^A D_1^A F_1^A \setminus M_1^A \mid M_1^A \setminus r_1^A [\mu_{\mathcal{P}} u_{\mathcal{N}(g)}^{\leq,+}].$$

By Lemma 4.3.6  $M_1^A \subseteq B_1^A C_1^A$ . By the decomposition and weak union semi-graphoid axioms,

$$r_1^A \perp\!\!\!\perp F_1^A \mid B_1^A C_1^A \setminus r_1^A [\mu_{\mathcal{P}} u_{\mathcal{N}(g)}^{\leq,+}].$$

Thus line 10 is satisfied. Noting that  $B_1^A = r_1^A$ ,

$$B_1^A \perp\!\!\!\perp F_1^A \mid C_1^A [\mu_{\mathcal{P}} u_{\mathcal{N}(g)}^{\leq,+}].$$

Thus line 22 is satisfied.

Let  $r_i^A = [R_A \setminus R_{i-1}^A]_{\leq}$ ,  $R_i^A = \text{pre}_{\mathfrak{G}}^{\leq}(r_i^A)$ , and  $M_i^A = \text{col}_{\mathfrak{G}_{R_i^A}}(r_i^A)$ . Let  $B_i^A = B_A \cap R_i^A$ ,  $C_i^A = C_A \cap R_i^A$ ,  $D_i^A = D_A \cap R_i^A$ , and  $F_i^A = F_A \cap R_i^A$  be the sets constrained to the set of variables  $R_i^A$ . By the inductive hypothesis:

$$B_{i-1}^A \perp\!\!\!\perp F_{i-1}^A \mid C_{i-1}^A [\mu_{\mathfrak{P}u_{\mathcal{N}(\mathfrak{G})}}^{\leq,+}].$$

If  $r_i^A \in \text{dis}_{\mathfrak{G}_{R_i^A}}(b_A)$ , then by Lemma 4.3.5,

$$r_i^A \perp\!\!\!\perp R_i^A \setminus M_{R_i^A} \mid M_{R_i^A} \setminus r_i^A [\mu_{\mathfrak{P}u_{\mathcal{N}(\mathfrak{G})}}^{\leq,+}].$$

Thus line 9 is satisfied. By changing notation,

$$r_i^A \perp\!\!\!\perp B_i^A C_i^A D_i^A F_i^A \setminus M_i^A \mid M_i^A \setminus r_i^A [\mu_{\mathfrak{P}u_{\mathcal{N}(\mathfrak{G})}}^{\leq,+}].$$

By Lemma 4.3.6  $M_i^A \subseteq B_i^A C_i^A$ . By the decomposition and weak union semi-graphoid axioms,

$$r_i^A \perp\!\!\!\perp F_i^A \mid B_i^A C_i^A \setminus r_i^A [\mu_{\mathfrak{P}u_{\mathcal{N}(\mathfrak{G})}}^{\leq,+}].$$

Thus line 10 is satisfied.

If  $r_i^A \in B_A$ , then  $B_i^A = r_i^A \cup B_{i-1}^A$ ,  $C_i^A = C_{i-1}^A$ ,  $D_i^A = D_{i-1}^A$ , and  $F_i^A = F_{i-1}^A$ . By changing notation,

$$r_i^A \perp\!\!\!\perp F_i^A \mid B_{i-1}^A C_i^A [\mu_{\mathfrak{P}u_{\mathcal{N}(\mathfrak{G})}}^{\leq,+}].$$

By changing the notation of the inductive hypothesis,

$$B_{i-1}^A \perp\!\!\!\perp F_i^A \mid C_i^A [\mu_{\mathfrak{P}u_{\mathcal{N}(\mathfrak{G})}}^{\leq,+}].$$

By the symmetry and contraction semi-graphoid axioms,

$$B_i^A \perp\!\!\!\perp F_i^A \mid C_i^A [\mu_{\mathfrak{P}u_{\mathcal{N}(\mathfrak{G})}}^{\leq,+}].$$

Thus line 22 is satisfied.

If  $r_i^A \in C_A$ , then  $B_i^A = B_{i-1}^A$ ,  $C_i^A = r_i^A \cup C_{i-1}^A$ ,  $D_i^A = D_{i-1}^A$ , and  $F_i^A = F_{i-1}^A$ . By changing notation,

$$r_i^A \perp\!\!\!\perp F_i^A \mid B_i^A C_{i-1}^A [\mu_{\mathcal{P}u_{\mathcal{N}(g)}}^{\leq,+}].$$

By changing the notation of the inductive hypothesis,

$$B_i^A \perp\!\!\!\perp F_i^A \mid C_{i-1}^A [\mu_{\mathcal{P}u_{\mathcal{N}(g)}}^{\leq,+}].$$

By the symmetry and contraction semi-graphoid axioms,

$$r_i^A \cup B_i^A \perp\!\!\!\perp F_i^A \mid C_{i-1}^A [\mu_{\mathcal{P}u_{\mathcal{N}(g)}}^{\leq,+}].$$

Thus line 12 satisfies the lemma. By the symmetry and weak union semi-graphoid axioms,

$$B_i^A \perp\!\!\!\perp F_i^A \mid C_i^A [\mu_{\mathcal{P}u_{\mathcal{N}(g)}}^{\leq,+}].$$

Thus line 22 is satisfied.

Else if  $r_i^A \in C_A F_A$ , then let  $A' = R_i^A \setminus D_A$ ,  $b_{A'} = [A']_{\leq}$ , and  $M_{A'} = \text{col}_{g_{A'}}(b_{A'})$ . Note that  $A'$  is ancestral because it is defined as an ancestral set minus a set that contains all of its descendants. Note that  $b_{A'} = r_i^A$ . Since we show that all other lines are satisfied and Algorithm 4 terminates, lines 16 is satisfied. Accordingly,

$$b_{A'} \perp\!\!\!\perp A' \setminus M_{A'} \mid M_{A'} \setminus b_{A'} [\mu_{\mathcal{P}u_{\mathcal{N}(g)}}^{\leq,+}].$$

By changing notation,

$$r_i^A \perp\!\!\!\perp B_i^A C_i^A F_i^A \setminus M_{A'} \mid M_{A'} \setminus r_i^A [\mu_{\mathcal{P}u_{\mathcal{N}(g)}}^{\leq,+}].$$

By Lemma 4.3.6  $M_{A'} \subseteq C_i^A F_i^A$ . By the symmetry, decomposition, and weak union semi-graphoid axioms,

$$B_i^A \perp\!\!\!\perp r_i^A \mid C_i^A F_i^A \setminus r_i^A [\mu_{\mathcal{P}u_{\mathcal{N}(g)}}^{\leq,+}].$$

Thus line 17 is satisfied.

If  $r_i^A \in F_A$ , then  $B_i^A = B_{i-1}^A$ ,  $C_i^A = C_{i-1}^A$ ,  $D_i^A = D_{i-1}^A$ , and  $F_i^A = r_i^A \cup F_{i-1}^A$ . By changing



notation,

$$B_i^A \perp\!\!\!\perp r_i^A \mid C_i^A F_{i-1}^A [\mu_{\mathbf{P}} u_{\mathcal{N}(g)}^{\leq,+}].$$

By changing the notation of the inductive hypothesis,

$$B_i^A \perp\!\!\!\perp F_{i-1}^A \mid C_i^A [\mu_{\mathbf{P}} u_{\mathcal{N}(g)}^{\leq,+}].$$

By the contraction semi-graphoid axiom,

$$B_i^A \perp\!\!\!\perp F_i^A \mid C_i^A [\mu_{\mathbf{P}} u_{\mathcal{N}(g)}^{\leq,+}].$$

Thus line 22 is satisfied.

If  $r_i^A \in C_A$ , then  $B_i^A = B_{i-1}^A$ ,  $C_i^A = r_i^A \cup C_{i-1}^A$ ,  $D_i^A = D_{i-1}^A$ , and  $F_i^A = F_{i-1}^A$ . By changing notation,

$$B_i^A \perp\!\!\!\perp r_i^A \mid C_{i-1}^A F_i^A [\mu_{\mathbf{P}} u_{\mathcal{N}(g)}^{\leq,+}].$$

By changing the notation of the inductive hypothesis,

$$B_i^A \perp\!\!\!\perp F_i^A \mid C_{i-1}^A [\mu_{\mathbf{P}} u_{\mathcal{N}(g)}^{\leq,+}].$$

By the contraction semi-graphoid axiom,

$$B_i^A \perp\!\!\!\perp r_i^A \cup F_i^A \mid C_{i-1}^A [\mu_{\mathbf{P}} u_{\mathcal{N}(g)}^{\leq,+}].$$

Thus line 19 is satisfied. By the weak union semi-graphoid axiom,

$$B_i^A \perp\!\!\!\perp F_i^A \mid C_i^A [\mu_{\mathbf{P}} u_{\mathcal{N}(g)}^{\leq,+}].$$

Thus line 22 is satisfied.

If  $r_i^A \in D_A$ , then  $B_i^A = B_{i-1}^A$ ,  $C_i^A = C_{i-1}^A$ ,  $D_i^A = r_i^A \cup D_{i-1}^A$ , and  $F_i^A = F_{i-1}^A$ . By changing the notation of the inductive hypothesis,

$$B_i^A \perp\!\!\!\perp F_i^A \mid C_i^A [\mu_{\mathbf{P}} u_{\mathcal{N}(g)}^{\leq,+}].$$

Thus line 22 is satisfied.

Accordingly,

$$B_A \perp\!\!\!\perp F_A \mid C_A [\mu_{\mathbf{P}}u_{\mathcal{N}(\mathcal{G})}^{\leq,+}].$$

Note that  $B_A = b_A \cup L_A^{\min}$ ,  $C_A = N_A \setminus b_A$ , and  $F_A \subseteq (A \setminus N_A) \cup (L_A \setminus M_{R_A} D_A)$ . By changing notation and the decomposition semi-graphoid axiom,

$$b \cup L_A^{\min} \perp\!\!\!\perp (A \setminus N_A) \cup (L_A \setminus M_{R_A} D_A) \mid N_A \setminus b_A [\mu_{\mathbf{P}}u_{\mathcal{N}(\mathcal{G})}^{\leq,+}].$$

By the symmetry and decomposition semi-graphoid axioms,

$$b_A \perp\!\!\!\perp A \setminus N_A \mid N_A \setminus b_A [\mu_{\mathbf{P}}u_{\mathcal{N}(\mathcal{G})}^{\leq,+}].$$

Thus line 28 is satisfied. Note that  $N_A \setminus b_A = (N_A \setminus M_A) \cup (M_A \setminus b_A)$  because  $b_A \in M_A \subseteq N_A$ .

By expanding notation,

$$b_A \perp\!\!\!\perp A \setminus N_A \mid (N_A \setminus M_A) \cup (M_A \setminus b) [\mu_{\mathbf{P}}u_{\mathcal{N}(\mathcal{G})}^{\leq,+}]$$

By Lemma 4.3.5,

$$b_A \perp\!\!\!\perp N_A \setminus M_A \mid M_A \setminus b_A [\mu_{\mathbf{P}}u_{\mathcal{N}(\mathcal{G})}^{\leq,+}].$$

Thus line 29 is satisfied. By the contraction and decomposition semi-graphoid axioms,

$$b_A \perp\!\!\!\perp (A \setminus N_A) \cup (N_A \setminus M_A) \mid M_A \setminus b_A [\mu_{\mathbf{P}}u_{\mathcal{N}(\mathcal{G})}^{\leq,+}].$$

Note that  $A \setminus M_A \subseteq (A \setminus N_A) \cup (N_A \setminus M_A)$  because  $M_A \subseteq N_A$ . By the decomposition semi-graphoid axiom,

$$b_A \perp\!\!\!\perp A \setminus M_A \mid M_A \setminus b_A [\mu_{\mathbf{P}}u_{\mathcal{N}(\mathcal{G})}^{\leq,+}].$$

Thus line 30 is satisfied. □

**Corollary 4.3.3.** *Let  $\mathcal{G} = (V, E)$  be a directed MAG,  $\mathbf{P} = \mathcal{P}(V)$  be the poset ordered by inclusion, and  $\leq$  be a total order consistent with  $\mathcal{G}$ . If  $b \in V$  and  $A \in \mathcal{A}(\mathcal{G})$  such that  $b \in A \subseteq \text{pre}_{\mathcal{G}}^{\leq}(b)$ , then*

$$b \perp\!\!\!\perp A \setminus \text{cl}_{\mathcal{G}_A}(b) \mid \text{mb}_{\mathcal{G}_A}(b) [\mu_{\mathbf{P}}u_{\mathcal{N}(\mathcal{G})}^{\leq,+}]$$

where  $\mu_{\mathbf{P}}u_{\mathcal{N}(\mathcal{G})}^{\leq,+}$  is the structural imset derived by applying the Möbius inversion to the primary imset constructed by Algorithm 3.

*Proof.* The proof follows from the definitions of Markov blanket and closure and above lemmas and corollaries.  $\square$

### 4.3.3 Markov Implies Factorization

In this section, we provide the necessary lemmas to prove that if the global Markov property holds, then the factorization presented in Section 4.3.4 holds. The intuition for Lemma 4.3.8 is given by the ordered local Markov property. In what follows,  $b$  is a barren vertex and  $M \setminus b$  is its Markov blanket with respect to the set  $N$ .

**Lemma 4.3.8.** *Let  $\mathcal{G} = (V, E)$  be a directed MAG containing a set  $N \subseteq V$  ( $N \neq \emptyset$ ). If  $b \in \text{bar}_{\mathcal{G}}(N)$  and  $M \in [N]_{\mathcal{M}(\mathcal{G})}^b$ , then:*

$$b \perp\!\!\!\perp N \setminus M \mid M \setminus b [\mathcal{G}].$$

*Proof.* By Proposition 3.3.3 the induced independence model  $\mathcal{J}(\mathcal{G})$  is a compositional graphoid. Accordingly, graphoid axioms (*i - vi*) may be applied. Consider the cases where  $N$  is  $m$ -connecting and not  $m$ -connecting.

If  $N$  is  $m$ -connecting, then by maximally  $M = N$ . By the triviality graphoid axiom

$$b \perp\!\!\!\perp N \setminus M \mid M \setminus b [\mathcal{G}].$$

If  $N$  is not  $m$ -connecting, then  $M \subset N$ . Pick a vertex  $a \in N \setminus M$  and let  $N_a = M \cup a$ . By maximally  $N_a$  is non- $m$ -connecting. By Lemma 4.3.3, since  $b \in \text{bar}_{\mathcal{G}}(N)$ , no inducing path exists between  $a$  and  $b$  relative to  $\langle V \setminus N_a, M \setminus b \rangle$ . By Proposition 3.3.2 if no inducing path exists between  $a$  and  $b$  relative to  $\langle V \setminus N_a, M \setminus b \rangle$ , then  $a$  and  $b$  are  $m$ -separated by  $C$  for some  $M \setminus b \subseteq C \subseteq N_a$  ( $a, b \notin C$ ). According

$$b \perp\!\!\!\perp a \mid M \setminus b [\mathcal{G}] \quad \text{for all } a \in N \setminus M.$$

By the composition graphoid axiom

$$b \perp\!\!\!\perp N \setminus M \mid M \setminus b [\mathcal{G}].$$

$\square$

**Lemma 4.3.9.** *Let  $\mathcal{G} = (V, E)$  be a directed MAG and  $\leq$  be a total order consistent with  $\mathcal{G}$ . Furthermore, let  $X$  be a collection of random variables indexed by  $V$  with probability measure  $P$  that admits density  $f(x)$  with respect to dominating  $\sigma$ -finite product measure  $\nu$ . If  $P$  has finite multiinformation  $m_P$  and satisfies the global Markov property for  $\mathcal{G}$ , then:*

$$\sum_{N \in \mathcal{P}(V)} u_{N(\mathcal{G})}^{\leq,+}(N) \phi_N(x) = 0 \quad \text{for } P\text{-a.e } x \in \mathcal{X}$$

and

$$\sum_{N \in \mathcal{P}(V)} u_{N(\mathcal{G})}^{\leq,-}(N) \phi_N(x) = 0 \quad \text{for } P\text{-a.e } x \in \mathcal{X}$$

where  $u_{N(\mathcal{G})}^{\leq,+}$  and  $u_{N(\mathcal{G})}^{\leq,-}$  are the imsets constructed by Algorithm 3.

*Proof.* Pick a variable  $b \in V$  and let  $R = \text{pre}_{\mathcal{G}}^{\leq}(b)$ . By Corollary 4.3.1 the induced subgraph  $\mathcal{G}_R$  is a directed MAG and  $\mathcal{J}(\mathcal{G}_R) \subseteq \mathcal{J}(\mathcal{G})$ . By Proposition 3.3.3 the induced independence model  $\mathcal{J}(\mathcal{G}_R)$  is a compositional graphoid. Accordingly, graphoid axioms (*i* - *vi*) may be applied. Run  $\text{PAIR}(\mathcal{G}_R, b) = N^{\mathcal{G}_R, b}, M^{\mathcal{G}_R, b}, n$ . Pick  $J \subseteq \{1, \dots, n\}$  and let  $M \in [N_J^{\mathcal{G}, b}]_{M(\mathcal{G})}^b$ . Note that

$$M \subseteq N_J^{\mathcal{G}, b} \subseteq N_i^{\mathcal{G}, b} \quad \text{for all } i \in J.$$

By maximality, since  $M \subseteq N_i^{\mathcal{G}, b}$  and  $M_i^{\mathcal{G}, b} \in [N_i^{\mathcal{G}, b}]_{M(\mathcal{G})}^b$  for all  $i \in J$ ,

$$M \subseteq M_i^{\mathcal{G}, b} \quad \text{for all } i \in J.$$

Accordingly,

$$M \subseteq M_{J, K}^{\mathcal{G}, b} \quad \text{for all } K \subseteq J.$$

By Lemma 4.3.8

$$b \perp\!\!\!\perp N_J^{\mathcal{G}, b} \setminus M \mid M \setminus b [\mathcal{G}_R].$$

By the weak union graphoid axiom

$$b \perp\!\!\!\perp N_J^{\mathcal{G}, b} \setminus M_{J, K}^{\mathcal{G}, b} \mid M_{J, K}^{\mathcal{G}, b} \setminus b [\mathcal{G}_R] \quad \text{for all } K \subseteq J.$$

Therefore, since  $\mathcal{J}(\mathcal{G}_R) \subseteq \mathcal{J}(\mathcal{G})$  and  $P$  satisfies the global Markov property for  $\mathcal{G}$ ,

$$b \perp\!\!\!\perp N_J^{\mathcal{G}, b} \setminus M_{J, K}^{\mathcal{G}, b} \mid M_{J, K}^{\mathcal{G}, b} \setminus b [P] \quad \text{for all } K \subseteq J.$$

By Equation 3.2, since the insets constructed by Algorithm 3 are constructed as sums over  $\mathbf{N}_{J,K}^{\mathcal{G},b}$  terms:

$$\sum_{N \in \mathcal{P}(V)} u_{\mathcal{N}(\mathcal{G})}^{\leq,+}(N) \phi_N(x) = 0 \quad \text{for } P\text{-a.e } x \in \mathcal{X}$$

and

$$\sum_{N \in \mathcal{P}(V)} u_{\mathcal{N}(\mathcal{G})}^{\leq,-}(N) \phi_N(x) = 0 \quad \text{for } P\text{-a.e } x \in \mathcal{X}.$$

□

#### 4.3.4 Formalization and Alternatives

In this section, we present the factorization and several alternatives. Notably, while the factorizations presented in this chapter are defined from probability measures with finite multiinformation, a similar proof could be given for positive measures.

**Theorem 4.3.1.** *Let  $\mathcal{G} = (V, E)$  be a directed MAG,  $\mathbf{P} = \mathcal{P}(V)$  be a poset ordered by inclusion, and  $\leq$  be a total order consistent with  $\mathcal{G}$ . Furthermore, let  $X$  be a collection of random variables indexed by  $V$  with probability measure  $P$  that admits density  $f(x)$  with respect to dominating  $\sigma$ -finite product measure  $\nu$ . If  $P$  has finite multiinformation  $m_P$ , then the following are equivalent:*

- i.*  $\log f(x) = \sum_{M \in \mathcal{M}(\mathcal{G})} \phi_M(x) - \sum_{N \in \mathcal{P}(V)} u_{\mathcal{N}(\mathcal{G})}^{\leq,-}(N) \phi_N(x) \quad \text{for } P\text{-a.e. } x \in \mathcal{X};$
- ii.*  $(\mu_P u_{\mathcal{N}(\mathcal{G})}^{\leq,+})^\top m_P = 0;$
- iii.*  $A \perp\!\!\!\perp B \mid C [\mathcal{G}] \Rightarrow A \perp\!\!\!\perp B \mid C [P] \quad \text{for every } \langle A, B \mid C \rangle \in \mathcal{T}(V);$

where  $u_{\mathcal{N}(\mathcal{G})}^{\leq,+}$  and  $u_{\mathcal{N}(\mathcal{G})}^{\leq,-}$  are the insets constructed by Algorithm 3.

*Proof.* (*i*  $\Rightarrow$  *ii*): By the Möbius inversion

$$\begin{aligned} \log f(x) &= \sum_{T \in \mathcal{P}(V)} \phi_T(x) \\ &= \sum_{M \in \mathcal{M}(\mathcal{G})} \phi_M(x) + \sum_{N \in \mathcal{P}(V)} \delta_{\mathcal{N}(\mathcal{G})}(N) \phi_N(x) \\ &= \sum_{M \in \mathcal{M}(\mathcal{G})} \phi_M(x) + \sum_{N \in \mathcal{P}(V)} u_{\mathcal{N}(\mathcal{G})}^{\leq,+}(N) \phi_N(x) - \sum_{N \in \mathcal{P}(V)} u_{\mathcal{N}(\mathcal{G})}^{\leq,-}(N) \phi_N(x). \end{aligned}$$

By the antecedent

$$\log f(x) = \sum_{M \in \mathcal{M}(\mathcal{G})} \phi_M(x) - \sum_{N \in \mathcal{P}(V)} u_{\mathcal{N}(\mathcal{G})}^{\leq, -}(N) \phi_N(x) \quad \text{for } P\text{-a.e. } x \in \mathcal{X}.$$

Therefore

$$\sum_{N \in \mathcal{P}(V)} u_{\mathcal{N}(\mathcal{G})}^{\leq, +}(N) \phi_N(x) = 0 \quad \text{for } P\text{-a.e. } x \in \mathcal{X}.$$

By integrating with respect to  $P$

$$\begin{aligned} \int_{x \in \mathcal{X}} \sum_{N \in \mathcal{P}(V)} u_{\mathcal{N}(\mathcal{G})}^{\leq, +}(N) \phi_N(x) \, dP(x) &= \sum_{N \in \mathcal{P}(V)} u_{\mathcal{N}(\mathcal{G})}^{\leq, +}(N) \int_{x \in \mathcal{X}} \phi_N(x) \, dP(x) \\ &= \sum_{N \in \mathcal{P}(V)} u_{\mathcal{N}(\mathcal{G})}^{\leq, +}(N) \sum_{T \subseteq N} \mu_P(N, T) \int_{x \in \mathcal{X}} \log f_T(x) \, dP(x) \\ &= \sum_{N \in \mathcal{P}(V)} u_{\mathcal{N}(\mathcal{G})}^{\leq, +}(N) \sum_{T \subseteq N} \mu_P(N, T) m_P(T) \\ &= (\mu_P u_{\mathcal{N}(\mathcal{G})}^{\leq, +})^\top m_P. \end{aligned}$$

Accordingly

$$(\mu_P u_{\mathcal{N}(\mathcal{G})}^{\leq, +})^\top m_P = 0.$$

(*ii*  $\Rightarrow$  *iii*): By Corollary 4.3.3, if  $b \in V$  and  $A \in \mathcal{A}(\mathcal{G})$  such that  $b \in A \subseteq \text{pre}_{\mathcal{G}}^{\leq}(b)$ , then

$$b \perp\!\!\!\perp A \setminus \text{cl}_{\mathcal{G}_A}(b) \mid \text{mb}_{\mathcal{G}_A}(b) [\mu_P u_{\mathcal{N}(\mathcal{G})}^{\leq, +}].$$

By Theorem 3.5.1 and the antecedent, if  $b \in V$  and  $A \in \mathcal{A}(\mathcal{G})$  such that  $b \in A \subseteq \text{pre}_{\mathcal{G}}^{\leq}(b)$ , then

$$b \perp\!\!\!\perp A \setminus \text{cl}_{\mathcal{G}_A}(b) \mid \text{mb}_{\mathcal{G}_A}(b) [P].$$

By Theorem 3.3.2

$$A \perp\!\!\!\perp B \mid C [\mathcal{G}] \quad \Rightarrow \quad A \perp\!\!\!\perp B \mid C [P] \quad \text{for every } \langle A, B \mid C \rangle \in \mathcal{T}(V).$$

( $i \Leftarrow iii$ ): By the Möbius inversion and Lemma 4.3.9

$$\begin{aligned}
\log f(x) &= \sum_{T \in \mathcal{P}(V)} \phi_T(x) \\
&= \sum_{M \in \mathcal{M}(\mathcal{G})} \phi_M(x) + \sum_{N \in \mathcal{N}(\mathcal{G})} \phi_N(x) \\
&= \sum_{M \in \mathcal{M}(\mathcal{G})} \phi_M(x) + \sum_{N \in \mathcal{N}(\mathcal{G})} u_{\mathcal{N}(\mathcal{G})}^{\leq,+}(N) \phi_N(x) - \sum_{N \in \mathcal{N}(\mathcal{G})} u_{\mathcal{N}(\mathcal{G})}^{\leq,-}(N) \phi_N(x) \\
&= \sum_{M \in \mathcal{M}(\mathcal{G})} \phi_M(x) - \sum_{N \in \mathcal{N}(\mathcal{G})} u_{\mathcal{N}(\mathcal{G})}^{\leq,-}(N) \phi_N(x) \quad \text{for } P\text{-a.e. } x \in \mathcal{X}.
\end{aligned}$$

□

In general, we refer to  $i$  of Theorem 3.5.1 as the factorization which we may alternatively characterize and with a structural imset:

$$\begin{aligned}
\log f(x) &= \sum_{M \in \mathcal{M}(\mathcal{G})} \phi_M(x) - \sum_{N \in \mathcal{N}(\mathcal{G})} u_{\mathcal{N}(\mathcal{G})}^{\leq,-}(N) \phi_N(x) \\
&= \sum_{N \in \mathcal{P}(V)} (\delta_{\mathcal{M}(\mathcal{G})}(N) - u_{\mathcal{N}(\mathcal{G})}^{\leq,-}(N)) \phi_N(x) \\
&= \sum_{N \in \mathcal{P}(V)} (1 - u_{\mathcal{N}(\mathcal{G})}^{\leq,+}(N)) \phi_N(x) \\
&= \log f(x) - \sum_{N \in \mathcal{N}(\mathcal{G})} u_{\mathcal{N}(\mathcal{G})}^{\leq,+}(N) \phi_N(x)
\end{aligned}$$

and with heads and tails:

$$\begin{aligned}
\log f(x) &= \sum_{M \in \mathcal{M}(\mathcal{G})} \phi_M(x) - \sum_{N \in \mathcal{N}(\mathcal{G})} u_{\mathcal{N}(\mathcal{G})}^{\leq,-}(N) \phi_N(x) \\
&= \sum_{H \in \mathcal{H}(\mathcal{G})} \phi_{H|\text{tail}_{\mathcal{G}}(H)}(x) - \sum_{N \in \mathcal{N}(\mathcal{G})} u_{\mathcal{N}(\mathcal{G})}^{\leq,-}(N) \phi_N(x)
\end{aligned}$$

and with conditional densities:

$$\begin{aligned}
\log f(x) &= \sum_{M \in \mathcal{M}(\mathcal{G})} \phi_M(x) - \sum_{N \in \mathcal{N}(\mathcal{G})} u_{\mathcal{N}(\mathcal{G})}^{\leq, -}(N) \phi_N(x) \\
&= \sum_{b \in V} \sum_{\substack{N \subseteq \text{cl}_{\mathcal{G}}^{\leq}(b) \\ b \in N}} (\delta_{\mathcal{M}(\mathcal{G})}(N) - u_{\mathcal{N}(\mathcal{G})}^{\leq, -}(N)) \phi_N(x) \\
&= \sum_{b \in V} \sum_{\substack{N \subseteq \text{cl}_{\mathcal{G}}^{\leq}(b) \\ b \in N}} (1 - u_{\mathcal{N}(\mathcal{G})}^{\leq, +}(N)) \phi_N(x) \\
&= \sum_{b \in V} \left[ \log f_{b|\text{mb}_{\mathcal{G}}^{\leq}(b)}(x) - \sum_{\substack{N \subseteq \text{cl}_{\mathcal{G}}^{\leq}(b) \\ b \in N}} u_{\mathcal{N}(\mathcal{G})}^{\leq, +}(N) \phi_N(x) \right]
\end{aligned}$$

The last alternative has a nice intuition using a special case of unfaithful DAGs [11] which we call dominating DAGs.

**Definition** (*dominating DAG*). Let  $\mathcal{G} = (V, E)$  be a directed MAG with consistent total order  $\leq$ . The dominating DAG  $\mathcal{G}' = \text{dom}(\mathcal{G}, \leq)$  is the DAG over the same vertices such that  $\text{pa}_{\mathcal{G}'}(b) = \text{mb}_{\mathcal{G}}^{\leq}(b)$  for all  $b \in V$ .

Accordingly, the last alternative can be expressed as an adjusted version of the recursive factorization for the dominating DAG:

$$\log f(x) = \sum_{b \in V} \left[ \log f_{b|\text{pa}_{\text{dom}(\mathcal{G}, \leq)}(b)}(x) - \sum_{\substack{N \subseteq \text{pa}_{\text{dom}(\mathcal{G}, \leq)}^+(b) \\ b \in N}} u_{\mathcal{N}(\mathcal{G})}^{\leq, +}(N) \phi_N(x) \right].$$

Appendix B.4 show that while it may be the case that the factorization does not need an adjustment term, our current proof strategy is insufficient.

### 4.3.5 Worked-out Examples

The first equality is the originally posed factorization, the second equality is the second alternative, and the third equality is the third alternative. The last line of the first two equalities makes up the adjustment term and the last line of the third equality includes terms required to construct a Markov DAG factorization in addition to the adjustment term.



Figure 4.3 where  $e \leq a \leq d \leq b \leq c$ :

$$\begin{aligned}
\log f(x) &= \phi_{abcde}(x) + \phi_{bcde}(x) + \phi_{abcd}(x) + \phi_{abc}(x) + \phi_{bcd}(x) + \phi_{cde}(x) + \phi_{ab}(x) \\
&\quad + \phi_{bc}(x) + \phi_{cd}(x) + \phi_{de}(x) + \phi_a(x) + \phi_b(x) + \phi_c(x) + \phi_d(x) + \phi_e(x) \\
&\quad - \phi_{ace}(x) - \phi_{ce}(x) - \phi_{ace}(x) - \phi_{ac}(x) \\
&= \phi_{bcd|ae}(x) + \phi_{bc|a}(x) + \phi_{cd|e}(x) + \phi_{b|a}(x) + \phi_{d|e}(x) + \phi_a(x) + \phi_c(x) + \phi_e(x) \\
&\quad - \phi_{c,e|a}(x) - \phi_{c,a|e}(x) \\
&= \log f_{c|abde}(x) + \log f_{b|a}(x) + \log f_{d|e}(x) + \log f_a(x) + \log f_e(x) \\
&\quad - \phi_{c,e|ab}(x) - \phi_{c,a|de}(x) - \phi_{c,ae}(x)
\end{aligned}$$

For directed MAGs with five vertices or fewer, no adjustment term is needed if the correct total order is chosen. That means that we can simplify the factorization for small graphs; this is worked out exhaustively in Appendix C. However, it also illuminates the fact that the factorization gives different decompositions for different total orders. Theorem 4.3.1 implies that if the result holds for any total order consistent with  $\mathcal{G}$ , then the result must hold for all total orders consistent with  $\mathcal{G}$ .

Figure 3.11 (i) where  $a \leq b \leq c \leq d$ :

$$\begin{aligned}
\log f(x) &= \phi_{abcd}(x) + \phi_{abd}(x) + \phi_{acd}(x) + \phi_{bcd}(x) + \phi_{ab}(x) + \phi_{ad}(x) \\
&\quad + \phi_{bc}(x) + \phi_{bd}(x) + \phi_{cd}(x) + \phi_a(x) + \phi_b(x) + \phi_c(x) + \phi_d(x) \\
&= \phi_{d|abc}(x) + \phi_{c|b}(x) + \phi_{b|a}(x) + \phi_a(x) \\
&= \log f_{d|abc}(x) + \log f_{c|b}(x) + \log f_{b|a}(x) + \log f_a(x)
\end{aligned}$$

Figure 3.11 (ii) where  $a \leq b \leq c \leq d$ :

$$\begin{aligned}
\log f(x) &= \phi_{abc}(x) + \phi_{abd}(x) + \phi_{ac}(x) + \phi_{ad}(x) + \phi_{bc}(x) \\
&\quad + \phi_{bd}(x) + \phi_a(x) + \phi_b(x) + \phi_c(x) + \phi_d(x) \\
&= \phi_{ad|b}(x) + \phi_{bc|a}(x) + \phi_{d|b}(x) + \phi_{c|a}(x) + \phi_b(x) + \phi_a(x) \\
&= \log f_{d|ab}(x) + \log f_{c|ab}(x) + \log f_b(x) + \log f_a(x)
\end{aligned}$$

Figure 4.1 (i) where  $a \leq b \leq c$ :

$$\begin{aligned}
 \log f(x) &= \phi_{ab}(x) + \phi_{bc}(x) + \phi_a(x) + \phi_b(x) + \phi_c(x) \\
 &= \phi_{c|b}(x) + \phi_{b|a}(x) + \phi_a(x) \\
 &= \log f_{c|b}(x) + \log f_{b|a}(x) + \log f_a(x)
 \end{aligned}$$

Figure 4.1 (ii) where  $a \leq c \leq b$ :

$$\begin{aligned}
 \log f(x) &= \phi_{abc}(x) + \phi_{ab}(x) + \phi_{bc}(x) + \phi_a(x) + \phi_b(x) + \phi_c(x) \\
 &= \phi_{b|ac}(x) + \phi_c(x) + \phi_a(x) \\
 &= \log f_{b|ac}(x) + \log f_c(x) + \log f_a(x)
 \end{aligned}$$

Figure 4.1 (iii) where  $a \leq d \leq b \leq c$ :

$$\begin{aligned}
 \log f(x) &= \phi_{abcd}(x) + \phi_{abc}(x) + \phi_{bcd}(x) + \phi_{ab}(x) + \phi_{bc}(x) \\
 &\quad + \phi_{cd}(x) + \phi_a(x) + \phi_b(x) + \phi_c(x) + \phi_d(x) \\
 &= \phi_{bc|ad}(x) + \phi_{c|d}(x) + \phi_{b|a}(x) + \phi_d(x) + \phi_a(x) \\
 &= \log f_{c|abd}(x) + \log f_{b|a}(x) + \log f_d(x) + \log f_a(x) \\
 &\quad - \phi_{a,c|d}(x)
 \end{aligned}$$

Figure 4.1 (iv) where  $a \leq c \leq b \leq d$ :

$$\begin{aligned}
 \log f(x) &= \phi_{abcd}(x) + \phi_{abc}(x) + \phi_{acd}(x) + \phi_{bcd}(x) + \phi_{ab}(x) + \phi_{bc}(x) \\
 &\quad + \phi_{bd}(x) + \phi_{cd}(x) + \phi_a(x) + \phi_b(x) + \phi_c(x) + \phi_d(x) \\
 &= \phi_{cd|ab}(x) + \phi_{bc|a}(x) + \phi_{d|b}(x) + \phi_{b|a}(x) + \phi_c(x) + \phi_a(x) \\
 &= \log f_{d|abc}(x) + \log f_{b|ac}(x) + \log f_c(x) + \log f_a(x) \\
 &\quad - \phi_{a,d|b}(x)
 \end{aligned}$$

Figure 4.1 (vi) where  $a \leq b \leq c \leq d$ :

$$\begin{aligned}
\log f(x) &= \phi_{abcd}(x) + \phi_{abc}(x) + \phi_{abd}(x) + \phi_{acd}(x) + \phi_{bcd}(x) + \phi_{ab}(x) \\
&\quad + \phi_{ac}(x) + \phi_{bd}(x) + \phi_{cd}(x) + \phi_a(x) + \phi_b(x) + \phi_c(x) + \phi_d(x) \\
&= \phi_{abcd}(x) + \phi_{abc}(x) + \phi_{abd}(x) + \phi_{acd}(x) + \phi_{bcd}(x) + \phi_{ab}(x) \\
&\quad + \phi_{ac}(x) + \phi_{bd}(x) + \phi_{cd}(x) + \phi_a(x) + \phi_b(x) + \phi_c(x) + \phi_d(x) \\
&= \log f_{d|abc}(x) + \log f_{c|ab}(x) + \log f_b(x) + \log f_a(x) \\
&\quad - \phi_{c,d}(x)
\end{aligned}$$

In the following chapter we discuss curved exponential families and apply the factorization to these families in order to develop probabilistic score for learning ancestral relationships.

## 5.0 MAG Curved Exponential Families

In this chapter we discuss exponential families whose independence models are described by MAGs—ancestral graph Markov models. Exponential families are attributed to Darrois, Koopman, and Pitman, who independently published the following defining theoretical result. The Darrois-Koopman-Pitman theorem states that a probability measure belongs to an exponential family if and only if the dimension of the sufficient statistic for data drawn from that probability measure is independent of the sample size of the data [1, 6]. Another defining theoretical result for exponential families was discovered in Bayesian statistics and states that a probability measure belongs to an exponential family if and only if that probability measure has a conjugate prior [34]. For these reasons, exponential families have found wide application in probabilistic graphical models [45].

**Definition** (*exponential family*). An *exponential family* is a family of probability measures that admit densities with respect to  $\sigma$ -finite measure  $\nu$  proportional to:

$$f(x \mid \theta) \propto \exp [\theta^\top t(x) - \psi(\theta)]$$

where  $\theta \in \Theta \equiv \{\theta \in \mathbb{R}^k ; \int_{x \in \mathcal{X}} \exp [\theta^\top t(x)] d\nu(x) < \infty\}$  is the natural parameter of dimension  $k$ ,  $t(x)$  is the sufficient statistic, and  $\psi(\theta) \equiv \int_{x \in \mathcal{X}} \exp [\theta^\top t(x)] d\nu(x)$  is the cumulant function.

When the natural parameter space is an open set, the exponential family is regular. Furthermore, a minimal exponential family is an exponential family where the components of the sufficient statistics  $t(x)$  are linearly independent. We are interested in minimal regular exponential families whose natural parameter spaces are constrained to smooth manifolds—curved exponential families.

**Definition** (*curved exponential family*). A *curved exponential family* is a minimal regular exponential family whose natural parameter space is constrained to a manifold characterized by a smooth bijective function called a *diffeomorphism*  $\Phi : \Theta \rightarrow \mathbb{R}^{k-m} \times \mathbb{R}^m$  for  $1 \leq m \leq k$ .

Accordingly, a curved exponential family is defined as follows:

$$\mathcal{F}_{\text{EF}} \equiv \{P_\theta \in \mathcal{F}_{\text{EF}} ; \quad \Phi(\theta)^\top = [\eta, C]\}$$

where  $C$  is constant. For more details about exponential families and curved exponential families see [5, 43].

Let  $P_\theta$  be an exponential family with natural parameter space  $\Theta$ . One way to constrain the natural parameter space of an exponential family is to restrict the members of the family to probability measures whose induced independence models are subsets of a prespecified independence model:

$$\mathcal{F}_{\text{EF}}(\mathcal{J}) \equiv \{P_\theta \in \mathcal{F}_{\text{EF}} ; \quad \mathcal{J} = \mathcal{J}(P_\theta)\}.$$

If the predefined independence model is induced by a MAG  $\mathcal{G}$ , then the result is a family of ancestral graph Markov models:

$$\mathcal{F}_{\text{EF}}(\mathcal{G}) \equiv \{P_\theta \in \mathcal{F}_{\text{EF}} ; \quad \mathcal{J}(\mathcal{G}) = \mathcal{J}(P_\theta)\}.$$

We denote the parameter space constrained by an independence model  $\mathcal{J}(\mathcal{O})$  as  $\Theta_{\mathcal{O}}$ . However, not all exponential families constrained by independence models induced by MAGs are curved exponential families. In this chapter, we discuss curved exponential families constrained by the independence models induced by MAGs. The families discussed include the following:

- CG    Conditional Gaussian;
- M    Multinomial;
- LH    Lee and Hastie;
- G    Gaussian.

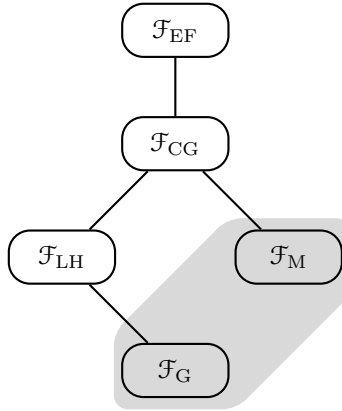


Figure 5.1: The Hasse diagram for the poset over families of probability measures ordered by inclusion.

Figure 5.1 depicts a Hasse diagram for a poset of families of probability measures ordered by inclusion—the colored section indicates families that are known to be curved exponential families when restricted by a MAG and the uncolored section indicates families that require additional restrictions to be curved exponential families. In particular, diffeomorphisms were given for Gaussian probability measures by Spirites et al. [82, 70] and for multinomial probability measures by Evans and Richardson [30].

In the forthcoming sections, we discuss conditional Gaussian, Gaussian, and Lee and Hastie probability measures respectively. Notably, conditional Gaussian and Lee and Hastie probability measures constrained by independence models induced by MAGs have not been shown to be curved exponential families. We provide an additional condition for MAGs such that Lee and Hastie probability measures constrained by independence models induced by MAGs satisfying the condition are curved exponential families. An analogous proof for conditional Gaussian is outside the scope of this dissertation.

## 5.1 Conditional Gaussian Probability Measures

The family of conditional Gaussian probability measures is the most general exponential family of probability measures discussed in this dissertation. In fact, the other families we discuss are subfamilies of conditional Gaussian probability measures. Conditional Gaussian probability measures were studied in detail by Lauritzen [46] and model mixtures of continuous and discrete variables where the conditional distribution of the continuous variables given the discrete variables is Gaussian. Following Lauritzen's notation, we use  $\Gamma$  to denote continuous variables and  $\Delta$  to denote discrete variables.

Let  $V$  be a non-empty set of variables partitioned by sets  $\Gamma, \Delta \in V$  which denote the continuous and discrete variables respectively. Let  $g(x_\Delta) : x_\Delta \rightarrow \mathbb{R}$ ,  $h(x_\Delta) : x_\Delta \rightarrow \mathbb{R}^{|\Gamma|}$ , and  $K(x_\Delta) : x_\Delta \rightarrow \mathbb{S}_{++}^{|\Gamma|}$ . A conditional Gaussian probability measure is a probability measure whose density has the following form:

$$f(x | \theta) \propto \exp \left[ g(x_\Delta) + h(x_\Delta)^\top x_\Gamma - \frac{1}{2} x_\Gamma^\top K(x_\Delta) x_\Gamma \right].$$

Furthermore, if  $K(x_\Delta)$  is constant, then the probability measure is a homogeneous conditional Gaussian probability measure. Let

$$g^*(x_\Delta) = g(x_\Delta) + \frac{1}{2} h(x_\Delta)^\top K(x_\Delta) h(x_\Delta) \quad \xi_\Gamma(x_\Delta) = K(x_\Delta)^\top h(x_\Delta).$$

The density of a conditional Gaussian probability measure can be put in a form where in the cases of only continuous or discrete variables, the probability measure is Gaussian or multinomial respectively

$$f(x | \theta) \propto \left[ g^*(x_\Delta) - \frac{1}{2} (x_\Gamma - \xi_\Gamma(x_\Delta))^\top K(x_\Delta) (x_\Gamma - \xi_\Gamma(x_\Delta)) \right].$$

For more details about conditional Gaussian probability measures see [46].

Notably, our prior assumption about multiinformation hold for this family of probability measures.

**Proposition 5.1.1** (Corollary 4.1 [83]). *Let  $V$  be a non-empty set of variables and  $X$  be a non-empty collection of random variables indexed by  $V$  with probability measure  $P_\theta$  dominated*

by  $\sigma$ -finite product measure  $\nu$ . If  $P_\theta$  is a conditional Gaussian probability measure over  $V$ , then  $P_\theta$  has finite multiinformation.

However, current theoretical results cannot guarantee that families of conditional Gaussian probability measures restricted by the independence models of MAGs are curved exponential families. This is due to the fact that conditional Gaussian probability measure are not closed under marginalization.

### 5.1.1 Conditional Gaussian Marginalization Condition

Unfortunately, while conditional Gaussian probability measures are closed under conditioning, they are not closed under marginalization. Lauritzen accounted for this by defining the concept of a weak marginal. Intuitively, a weak marginal is the conditional Gaussian probability measure over a marginal set of variables that is as close as possible to the actual marginal.

**Definition** (*weak marginal*). Let  $V$  be a non-empty set of variables containing a set  $A \subseteq V$ . Furthermore, let  $X$  be a collection of random variables indexed by  $V$  with conditional Gaussian probability measure  $P_\theta$  that admits density  $f(x | \theta)$  with respect to dominating  $\sigma$ -finite product measure  $\nu$ . Lastly, denote the density of the weak marginal with respect to  $A$  as  $f_{[A]}$ . As mentioned above, the weak marginal is “close” to the actual marginal in the following sense. If  $f_A$  is a conditional Gaussian density, then  $f_{[A]} = f_A$ , otherwise  $f_{[A]}$  is the conditional Gaussian distribution that minimizes:

$$\inf_{\theta \in \Theta} \int_{x \in \mathcal{X}} \log \left[ \frac{f_A(x | \theta)}{f_{[A]}(x | \theta)} \right] dP_\theta(x).$$

In the field of information theory, the above integral is the Kullback-Liebler divergence or relative entropy of the weak marginal with respect to the actual marginal. To be clear,  $f_A(x | \theta)$  is the actual marginal of a conditional Gaussian density but not necessarily a conditional Gaussian density. On the other hand,  $f_{[A]}(x | \theta)$  is the weak marginal of a conditional Gaussian density and a conditional Gaussian density.

It turns out that conditional Gaussian probability measures are closed under marginalization subject to the following condition.



**Definition** (*conditional Gaussian marginalization condition*). Let  $V$  be a non-empty set of variables partitioned by sets  $\Gamma, \Delta \subseteq V$  which denote the continuous and discrete variables respectively. Furthermore, let  $P$  be a conditional Gaussian probability measure that admits density  $f$  with respect to  $\sigma$ -finite product measure  $\nu$ . If  $A, L \subseteq V$  such that  $L = V \setminus A$ , then  $f_A$  is conditional Gaussian if and only if:

$$A \cap \Gamma \perp\!\!\!\perp L \cap \Delta \mid A \cap \Delta [P].$$

That is, the continuous variables in the margin are independent of the marginalized discrete variables given the discrete variables in the margin. Accordingly, this condition characterizes when an actual marginal and a weak marginal are equivalent [31, 46].

We call this the conditional Gaussian MAG condition (CGMC) because conditional Gaussian probability measures whose induced independence models are subsets of MAGs that satisfy this condition are curved exponential families. This result is not proven here, but should be straightforward to prove using the analogous proofs for multinomial and Gaussian probability measures. Notably, there are MAGs that do not satisfy this condition that are Markov equivalent to a MAG that does satisfy this condition

If  $P$  is a conditional Gaussian probability measure whose independence model is restricted by  $\mathcal{I}(\mathcal{G})$ , then the following is a necessary and sufficient condition for  $P$  to satisfy the conditional Gaussian marginalization condition with respect to the ancestral set of the directed subgraph.

**Definition** (*conditional Gaussian MAG condition*). Let  $\mathcal{G} = (V, E)$  be a MAG whose vertices are partitioned by sets  $\Gamma, \Delta \subseteq V$  which denote the continuous and discrete vertices respectively. If there exists  $\mathcal{G}' \in [\mathcal{G}]$  and  $\mathcal{G}'' = \text{dir}(\mathcal{G}')$  such that:

$$HT \not\subseteq \Delta \quad \Rightarrow \quad H \subseteq \Gamma$$

for all  $H \in \mathcal{H}(\mathcal{G}'')$  and  $T = \text{tail}_{\mathcal{G}''}(H)$ , then  $\mathcal{G}$  satisfies the conditional Gaussian MAG condition (CGMC).

**Proposition 5.1.2.** *Let  $\mathcal{G} = (V, E)$  be a MAG whose vertices are partitioned by sets  $\Gamma, \Delta \subseteq V$  which denote the continuous and discrete vertices respectively. If  $\text{dir}(\mathcal{G}) = \mathcal{G}' = (V', E')$ ,*

then following are equivalent:

- i.  $HT \not\subseteq \Delta \Rightarrow H \subseteq \Gamma$  for all  $H \in \mathcal{H}(\mathcal{G}')$  and  $T = \text{tail}_{\mathcal{G}'}(H)$ ;
- ii.  $A \cap \Gamma \perp\!\!\!\perp L \cap \Delta \mid A \cap \Delta [\mathcal{G}']$  for all  $A \in \mathcal{A}(\mathcal{G}')$  and  $L = V' \setminus A$ .

*Proof.* ( $i \Rightarrow ii$ ):

By the antecedent, the descendants of continuous variables must also be continuous and the districts are either completely continuous or completely discrete.

Let  $A \in \mathcal{A}(\mathcal{G})$  and  $L = V' \setminus A$ . If  $A \cap \Gamma = \emptyset$ , then  $A \cap \Gamma \perp\!\!\!\perp L \cap \Delta \mid A \cap \Delta [\mathcal{G}]$  by the triviality semi-graphoid axiom.

If  $A \cap \Gamma \neq \emptyset$ , then let  $a \in A \cap \Gamma$  and  $B = V' \setminus \text{deg}_{\mathcal{G}'}(L \cap \Gamma)$ . By definition  $a \perp\!\!\!\perp B \setminus \text{cl}_{\mathcal{G}'_B}(a) \mid \text{mb}_{\mathcal{G}'_B}(a) [\mathcal{G}'_B]$ .

The district must be continuous by the antecedent:  $\text{dis}_{\mathcal{G}'_B}(a) \subseteq A \cap \Gamma$ . Since  $A$  is ancestral, all the districts parents must be in  $A$  as well. All the district's descendants must be continuous by the antecedent and accordingly, in  $B$  if not latent. If not latent, then in  $A$  and because  $A$  is ancestral, their parents are in  $A$ . Therefore,  $\text{mb}_{\mathcal{G}'_B}(a) \subseteq A \setminus a$ .

By the antecedent and since the continuous descendants are continuous,  $L \cap \Delta \subseteq B$ . Since  $\text{mb}_{\mathcal{G}'_B}(a) \subseteq A \setminus a$  and  $a \in A$ ,  $L \cap \Delta \subseteq B \setminus \text{cl}_{\mathcal{G}'_B}(a)$ .

By the weak union semi-graphoid axiom,  $a \perp\!\!\!\perp L \cap \Delta \mid A \setminus a [\mathcal{G}'_B]$ . By Proposition 3.4.2,  $\mathcal{J}(\mathcal{G}'_B) \subseteq \mathcal{J}(\mathcal{G}')$ . By the intersection semi-graphoid axiom,  $A \cap \Gamma \perp\!\!\!\perp L \cap \Delta \mid A \cap \Delta [\mathcal{G}]$ .

( $i \Leftarrow ii$ ):

Assume by way of contradiction that there exists  $H \in \mathcal{H}(\mathcal{G}')$  such that  $HT \cap \Gamma \neq \emptyset$  and  $H \cap \Delta \neq \emptyset$  where  $T = \text{tail}_{\mathcal{G}'}(H)$  and  $A \cap \Gamma \perp\!\!\!\perp L \cap \Delta \mid A \cap \Delta [\mathcal{G}']$  for all  $A \in \mathcal{A}(\mathcal{G}')$ . Let  $B = H \cap \Delta$  and  $A = \text{an}_{\mathcal{G}'}(HT \setminus B)$ .  $a \in HT \cap \Gamma$  and  $b \in B$  such that  $a$  and  $b$  are adjacent in  $\mathcal{G}'$ . Therefore,  $a \not\perp\!\!\!\perp b \mid A \cap \Delta [\mathcal{G}]$  by maximally. Accordingly,  $A \cap \Gamma \not\perp\!\!\!\perp L \cap \Delta \mid A \cap \Delta [\mathcal{G}']$ . This is a contradiction.  $\square$

We conjecture that the CGMC defines exactly the set of MAGs that describe the independence models of conditional Gaussian probability measures. That is, conditional Gaussian probability measures cannot represent the same set of conditional independence statements

as a directed MAG that does not satisfy the marginalization condition—it is parametrically impossible.

**Conjecture 5.1.1.** *Let  $\mathcal{G} = (V, E)$  be a MAG whose variables are partitioned by sets  $\Gamma, \Delta \subseteq V$  which denote the continuous and discrete variables respectively and let  $P_\theta$  be a conditional Gaussian probability measure. If  $\mathcal{G}$  does not satisfy the CGMC and  $\mathcal{J}(P_\theta) \subseteq \mathcal{J}(\mathcal{G})$ , then there exists  $a, b \in \Gamma$  and  $C \in V \setminus \{a, b\}$  where  $\text{deg}_{\mathcal{G}}(a) \cap \text{deg}_{\mathcal{G}}(b) \cap \Delta \neq \emptyset$  such that:*

- i.  $a \perp\!\!\!\perp b \mid C [\mathcal{G}]$ ;*
- ii.  $a \not\perp\!\!\!\perp b \mid C [P_\theta]$ .*

## 5.2 Gaussian Probability Measures

### 5.2.1 Gaussian Parameterization

In this section, we detail the parameterization of MAGs for Gaussian probability measures, discussed in detail by Richardson and Spirtes [70]. Let  $\mathcal{G} = (V, E)$  be a MAG. Define  $D = \text{ch}_{\mathcal{G}}(V) \cup \text{sp}_{\mathcal{G}}(V)$  and  $U = V \setminus D$  as the directed and undirected vertices of  $\mathcal{G}$  respectively.

- Define  $\Lambda(\mathcal{G}) \subseteq \mathbb{S}_{++}^{|U|}$  to be the set of matrices  $(\Lambda)_{ab} = \lambda_{ab}$  that satisfy:

$$(\Lambda)_{ab} \equiv \lambda_{ab} \in \begin{cases} \mathbb{R} & a \in \text{ne}_{\mathcal{G}}^+(b); \\ \{0\} & \text{otherwise.} \end{cases}$$

- Define  $\Omega(\mathcal{G}) \subseteq \mathbb{S}_{++}^{|D|}$  to be the set of matrices  $(\Omega)_{ab} = \omega_{ab}$  that satisfy:

$$(\Omega)_{ab} \equiv \omega_{ab} \in \begin{cases} \mathbb{R} & a \in \text{sp}_{\mathcal{G}}^+(b); \\ \{0\} & \text{otherwise.} \end{cases}$$

- Define  $\mathbf{B}(\mathcal{G}) \subseteq \mathbb{R}^{|V| \times |V|}$  to be the set of matrices  $(B)_{ab} = \beta_{ab}$  that satisfy:

$$(B)_{ab} \equiv \beta_{ab} \in \begin{cases} \mathbb{R} & a \in \text{ch}_{\mathcal{G}}(b); \\ \{0\} & \text{otherwise.} \end{cases}$$

- Define  $\boldsymbol{\mu}(\mathcal{G}) \equiv \mathbb{R}^{|V|}$  to be the set of real numbers.

The parameterization of  $\mathcal{G}$  given by the diffeomorphism  $\Phi_{\mathcal{G}}^{-1} : \Lambda \times \Omega \times \mathbf{B} \times \boldsymbol{\mu} \rightarrow \Theta_{\mathcal{G}}$  is defined as follows:

$$\Phi_{\mathcal{G}}^{-1}(\Lambda, \Omega, B, \mu) = \begin{bmatrix} \mu^{\top} K \\ -\frac{1}{2} \text{vec}(K)^{\top} \end{bmatrix}$$

where

$$K = (I - B)^{\top} \begin{bmatrix} \Lambda^{-1} & 0 \\ 0 & \Omega \end{bmatrix}^{-1} (I - B)$$

and  $I$  is the  $|V| \times |V|$  identity matrix. Accordingly, the family of Gaussian MAG probability measures is a curved exponential family characterized by the inverse diffeomorphism  $\Phi_{\mathcal{G}}$ .

**Proposition 5.2.1** (Theorem 8.23 [70]). *For a MAG  $\mathcal{G} = (V, E)$ ,  $\mathcal{F}_{\mathcal{G}}(\mathcal{G})$  is a curved exponential family, with dimension  $2|V| + |E|$ .*

Furthermore, these curved exponential families correspond exactly to the independence models induced by the corresponding directed MAG.

**Proposition 5.2.2** (Theorem 8.14 [70]). *Let  $\mathcal{G}$  be a MAG. If  $\mathcal{F}_{\mathcal{G}}(\mathcal{G})$  is the family of Gaussian probability measures parameterized by  $\mathcal{G}$  and  $\mathcal{F}_{\mathcal{G}}(\mathcal{J}(\mathcal{G}))$  is the family of Gaussian probability measures constrained by  $\mathcal{J}(\mathcal{G})$ , then*

$$\mathcal{F}_{\mathcal{G}}(\mathcal{G}) = \mathcal{F}_{\mathcal{G}}(\mathcal{J}(\mathcal{G})).$$

Conveniently, all the parameters used to define the diffeomorphism have meaningful interpretations:

- $K$  and  $\mu$  are the precision matrix mean vector respectively;
- $\beta_{ab}$  corresponds to the coefficient of  $b$  in the regression of  $a$  on its parents  $\text{pa}_{\mathcal{G}}(a)$ ;
- $\omega_{ab}$  corresponds to the covariance between the residuals of  $a$  regressed on its parents

$\text{pa}_{\mathcal{G}}(a)$  and the residuals of  $b$  regressed on its parents  $\text{pa}_{\mathcal{G}}(b)$ ;

- $\lambda_{ab}$  corresponds to an edge potential in  $\text{un}(\mathcal{G})$ .

In Section 5.3, we repurpose Richardson and Spirtes parameterization for the family of Lee and Hastie probability measures.

### 5.3 Lee and Hastie Probability Measures

In this section, we consider a subfamily of conditional Gaussian probability measures first characterized by Lee and Hastie [47]—Accordingly, we call these measures Lee and Hastie probability measures. Raghu et al. provide a summary of methods developed to learn Markov equivalence classes of MAGs [63] on data generated from Lee and Hastie probability measures.

The family of Lee and Hastie probability measures is the special case of the family of homogeneous conditional Gaussian probability measures. The covariance matrix is constant for different values of the discrete variables and the discrete variables factorize as a pairwise discrete Markov random field (MRF). When the independence model of a probability measure is described by an undirected graph, that model is called a Markov random field; see [45] for details on pairwise MRFs.

We give a diffeomorphism to show that Lee and Hastie probability measures whose independence models are restricted by MAGs are curved exponential families. However, we first describe a transformation to facilitate the discussion of the diffeomorphism and provide an additional condition for Lee and Hastie probability measures to be curved exponential families.

#### 5.3.1 Binary Transformation

We show that the family of Lee and Hastie probability measures is an exponential family using the following transformation. Let  $V$  be a non-empty set of variables partitioned by sets  $\Gamma, \Delta \subseteq V$  which denote the continuous and discrete variables respectively. Furthermore, let

$X$  be a collection of random variables indexed by  $V$  with conditional Gaussian probability measure  $P_\theta$  that admits density  $f(x | \theta)$  with respect to dominating  $\sigma$ -finite product measure  $\nu$ . Accordingly,  $\mathcal{X} = \mathcal{X}_\Gamma \times \mathcal{X}_\Delta$  where  $\mathcal{X}_\Gamma \subseteq \mathbb{R}^{|\Gamma|}$  and  $\mathcal{X}_\Delta \subseteq \mathbb{Z}_+^{|\Delta|}$ .

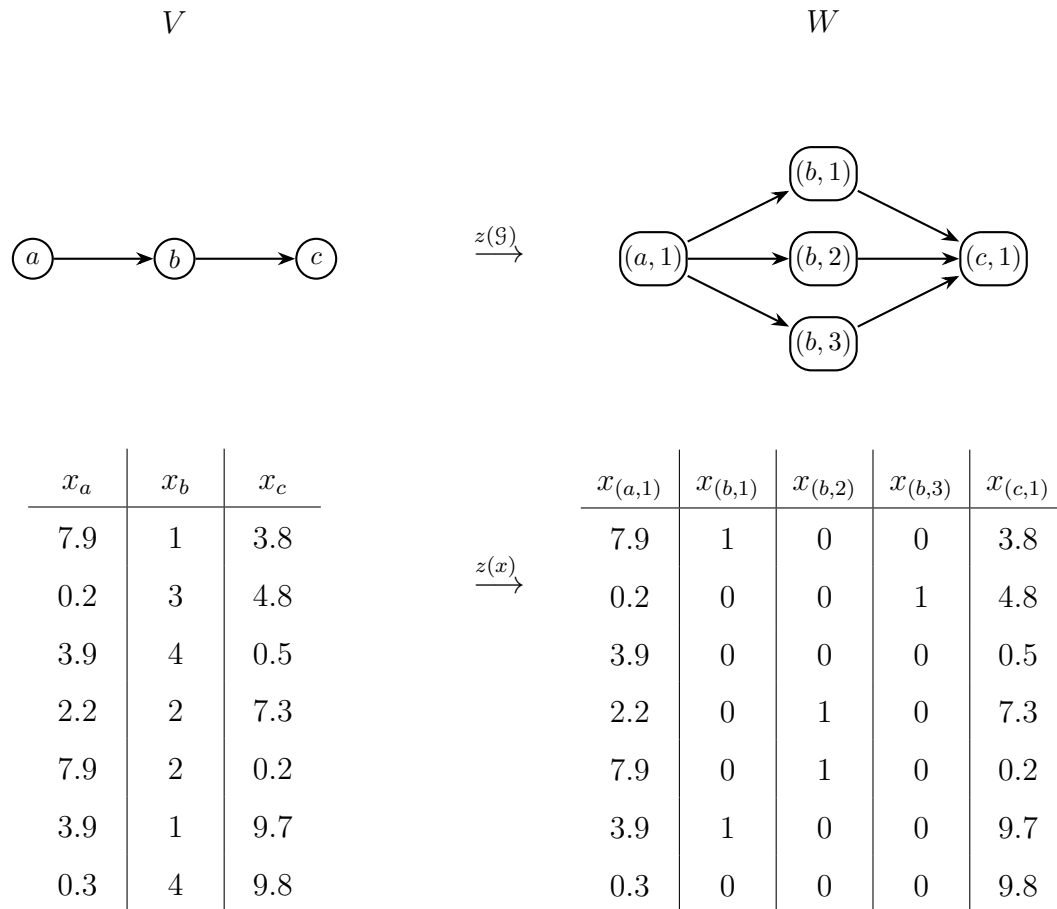


Figure 5.2: An illustration of the binary transformation

Define the binary transformation of  $V$  with respect to  $A \subseteq V$  as follows:

$$z_A(V) \equiv W_A = \{a \in A \times \mathbb{Z}_+ ; \quad a_2 \leq |a_1|\}$$

where subscripts are used to index the first or second part of a transformed variable and  $|w_1|$  equals the number of non-redundant categories for discrete random variables and one for continuous variables. Define the binary transformation of  $x$  with respect to  $A$  as a function

$$z_A : \mathcal{X} \rightarrow \mathbb{R}^{|W_A \cap \Gamma|} \times \{0, 1\}^{|W_A \cap \Delta|};$$

$$(z_A(x))_a \equiv z_a = \begin{cases} x_{a_1} & a_1 \in \Gamma \\ \delta_{x_{a_1}, a_2} & a_1 \in \Delta \end{cases}$$

where  $\delta_{x_{a_1}, a_2}$  is the Kronecker delta. Additionally, we define a corresponding transformation for directed MAGs.

---

**Algorithm 5:** BINARY TRANSFORMATION  $z(\mathcal{G})$

---

**Input:** MAG:  $\mathcal{G} = (V, E)$   
**Output:** MAG:  $\mathcal{G}' = (W, F)$

- 1  $W = \{a \in V \times \mathbb{Z}_+ ; a_2 \leq |a_1|\}$  ;
- 2  $F = \emptyset$  ;
- 3 **foreach**  $a, b \in W$  ( $a \neq b$ ) **do**
- 4     **if**  $a_1 \leftarrow b_1$  in  $\mathcal{G}$  **then**
- 5         Add  $a \leftarrow b$  to  $F$ ;
- 6     **else if**  $a_1 \leftrightarrow b_1$  in  $\mathcal{G}$  **then**
- 7         Add  $a \leftrightarrow b$  to  $F$ ;
- 8     **else if**  $a_1 - b_1$  in  $\mathcal{G}$  **then**
- 9         Add  $a - b$  to  $F$ ;
- 10    **end**
- 11 **end**

---

We show that the transformed graph is a directed MAG which has the same conditional independence relationships.

**Proposition 5.3.1.** *Let  $\mathcal{G} = (V, E)$  be a MAG. If  $z(\mathcal{G}) = (W, F)$  is the transformed graph, then  $z(\mathcal{G})$  is a MAG and*

$$A \perp\!\!\!\perp B \mid C [\mathcal{G}] \iff W_A \perp\!\!\!\perp W_B \mid W_C [z(\mathcal{G})]$$

*Proof.* By construction,  $z(\mathcal{G})$  is a ancestral graph. We consider the contrapositive for the double implication.

If  $A \not\perp\!\!\!\perp B \mid C [\mathcal{G}]$ , then there is an  $m$ -connecting path  $\pi$  between  $a \in A$  and  $b \in B$  relative to  $C$  in  $\mathcal{G}$ . Construct  $\pi'$  in  $z(\mathcal{G})$  by replacing each vertex  $v \in \pi$  with  $w \in z(v)$ . By construction,  $\pi'$  is an  $m$ -connecting path between  $a' \in W_A$  and  $b' \in W_B$  relative to  $W_C$  in

$z(\mathcal{G})$ . Accordingly  $W_A \not\perp\!\!\!\perp W_B \mid W_C [z(\mathcal{G})]$ .

If  $W_A \not\perp\!\!\!\perp W_B \mid W_C [\mathcal{G}]$ , then there is an  $m$ -connecting path  $\pi'$  between  $a \in W_A$  and  $b \in W_B$  relative to  $W_C$  in  $z(\mathcal{G})$ . Construct  $\pi'$  in  $\mathcal{G}$  by replacing each vertex  $w \in \pi$  with  $v = w_1$ . If any  $w \in \pi$  appears more than once, then remove all vertices the between the first and last occurrence of  $w$  and the last occurrence of  $w$ . By construction,  $\pi'$  is an  $m$ -connecting path between  $a' \in A$  and  $b' \in B$  relative to  $C$  in  $\mathcal{G}$ . Accordingly  $A \not\perp\!\!\!\perp B \mid C [\mathcal{G}]$ .

Accordingly,  $z(\mathcal{G})$  is an ancestral graph and  $A \perp\!\!\!\perp B \mid C [\mathcal{G}] \Leftrightarrow W_A \perp\!\!\!\perp W_B \mid W_C [z(\mathcal{G})]$ ; maximality follows from the maximality of  $\mathcal{G}$ .

□

### 5.3.2 Lee and Hastie MAG condition

Lee and Hastie probability measures require an additional condition for MAGs such that Lee and Hastie probability measures constrained by independence models induced by MAGs satisfying the condition are curved exponential families.

**Definition** (*Lee and Hastie MAG condition*). Let  $V$  be a non-empty set of variables partitioned by sets  $\Gamma, \Delta \subseteq V$  which denote the continuous and discrete variables respectively. Furthermore, let  $A, B, C \subseteq V$  be disjoint sets and  $D \subseteq \Delta \setminus AB$ .  $\mathcal{G}$  satisfies the Lee and Hastie MAG condition (LHMC) if:

- i.  $\mathcal{G}$  satisfies the CGMC;
- ii.  $A \perp\!\!\!\perp B \mid C [\mathcal{G}] \Rightarrow A \perp\!\!\!\perp B \mid CD [\mathcal{G}]$  for all  $\langle A, B \mid C \rangle \in \mathcal{T}(V)$  and  $D \subseteq \Delta \setminus AB$ .

If conditioning on a set of discrete variables induces dependence between two other sets of variables, then marginalizing the same set of discrete variables will result in a mixture of two or more marginal Lee and Hastie densities and induce the same dependency.

An intuition for this comes from the similarity between Lee and Hastie and Gaussian probability measures. Figure 5.3 illustrates the case where two continuous variables cause a discrete variable that otherwise have no relation. In this case we would expect to see that the continuous variables are marginally independent, however, this is not the case. The light gray points give the marginal of the Gaussian probability measure with the same parameterization. Accordingly, the Lee and Hastie probability measure appears similar to a Gaussian



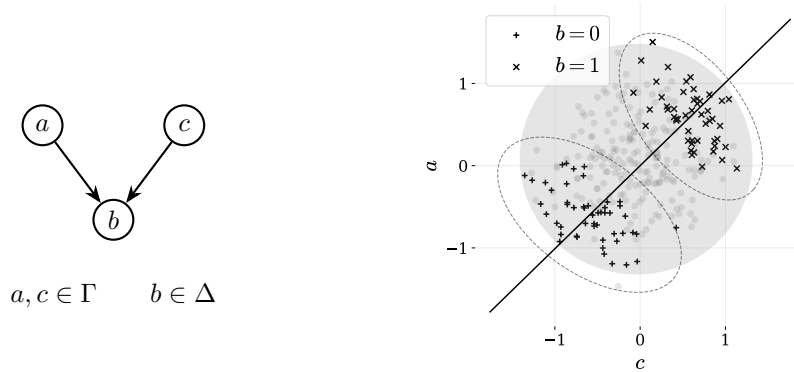


Figure 5.3: Lee and Hastie probability measures and violations of the marginalization condition. The contours give three standard deviations and the solid black line gives the first principal component.

probability measure subject to a selection effect. This selection effect induces a marginal dependence between the two continuous parents, which would otherwise be marginally independent.

Accordingly, the LHMC ensures that these induced dependencies do not occur. Graphically, this implies that the discrete variables are contained within the undirected subgraph of MAGs.

### 5.3.3 Lee and Hastie Parameterization

Let  $\mathcal{G} = (V, E)$  be a directed MAG satisfying the LHMC whose variables are partitioned by sets  $\Gamma, \Delta \subseteq V$  which denote the continuous and discrete variables respectively, and let  $z(\mathcal{G}) = (W, F)$  be the transformed directed MAG. Define  $D = \text{ch}_{z(\mathcal{G})}(W) \cup \text{sp}_{z(\mathcal{G})}(W)$  and  $U = W \setminus D$  as the directed and undirected vertices of  $z(\mathcal{G})$  respectively. Furthermore,  $X$  be a collection of random variables indexed by  $V$ . We redefine Richardson and Spirtes parameterization of  $\mathcal{G}$  for Lee and Hastie probability measures as follows:

- Define  $\Lambda(\mathcal{G}) \subseteq \mathbb{S}_{++}^{|U|}$  to be the set of matrices  $(\Lambda)_{ab} \equiv \lambda_{ab}$  that satisfy:

$$(\Lambda)_{ab} \equiv \lambda_{ab} \in \begin{cases} \mathbb{R} & a = b \text{ and } a_1 \in \Gamma \text{ or } a_1 \in \text{neg}(b_1); \\ \{1\} & a = b \text{ and } a_1 \in \Delta; \\ \{0\} & \text{otherwise.} \end{cases}$$

- Define  $\Omega(\mathcal{G}) \subseteq \mathbb{S}_{++}^{|D|}$  to be the set of matrices  $(\Omega)_{ab} \equiv \omega_{ab}$  that satisfy:

$$(\Omega)_{ab} \equiv \omega_{ab} \in \begin{cases} \mathbb{R} & a_1 \in \text{sp}_{\mathcal{G}}^+(b_1); \\ \{0\} & \text{otherwise.} \end{cases}$$

- Define  $\mathbf{B}(\mathcal{G}) \subseteq \mathbb{R}^{|W| \times |W|}$  to be the set of matrices  $(B)_{ab} \equiv \beta_{ab}$  that satisfy:

$$(B)_{ab} \equiv \beta_{ab} \in \begin{cases} \mathbb{R} & a_1 \in \text{ch}_{\mathcal{G}}(b_1); \\ \{0\} & \text{otherwise.} \end{cases}$$

- Define  $\mu(\mathcal{G}) \equiv \mathbb{R}^{|W|}$  to be the set of real numbers.

The parameterization of  $\mathcal{G}$  given by the diffeomorphism  $\Phi_{\mathcal{G}}^{-1} : \Lambda \times \Omega \times \mathbf{B} \times \mu \rightarrow \Theta_{\mathcal{G}}$  is defined as follows:

$$\Phi_{\mathcal{G}}^{-1}(\Lambda, \Omega, B, \mu) \equiv \begin{bmatrix} \mu^{\top} K \\ -\frac{1}{2} \text{vec}(K)^{\top} \end{bmatrix}$$

where

$$K \equiv (I - B)^{\top} \begin{bmatrix} \Lambda^{-1} & 0 \\ 0 & \Omega \end{bmatrix}^{-1} (I - B)$$

and  $I$  is the  $|W| \times |W|$  identity matrix. A parameterization is maximal if for all  $a, b \in V$ :

$$\begin{aligned} a_1 \in \text{ne}_{\mathcal{G}}^+(b_1) &\Rightarrow \lambda_{ab} \neq 0; \\ a_1 \in \text{sp}_{\mathcal{G}}^+(b_1) &\Rightarrow \omega_{ab} \neq 0; \\ a_1 \in \text{ch}_{\mathcal{G}}(b_1) &\Rightarrow \beta_{ab} \neq 0. \end{aligned}$$

We show the family of Lee and Hastie MAG probability measures is a curved exponential

family characterized by the inverse diffeomorphism  $\Phi_g$ . The parameterization given by Lee and Hastie differs from the parameterization given here and we have not verified that the two are equivalent. Notably, the parameterization given by Lee and Hastie uses more parameters, so it is possible that they describe a more general family of probability measures. However, the parameterization given by Lee and Hastie is not minimal, so it is also possible that our parameterization is a minimal characterization of the same family of probability measures.

### 5.3.4 Lee and Hastie as Curved Exponential Families

Let  $z \equiv z(x)$  and  $z_A \equiv z_A(x)$ . Lee and Hastie probability measures form an exponential family as follows:

$$\theta^\top \equiv \left[ \mu^\top K \quad -\frac{1}{2} \text{vec}(K)^\top \right] \quad t(x) \equiv \begin{bmatrix} z \\ \text{vec}(zz^\top) \end{bmatrix}$$

where  $\text{vec}(A)$  is the vectorization of matrix  $A$  into a column vector. In what follows, we show that Lee and Hastie probability measures are conditional Gaussian probability measures. We abuse notation and use the following shorthand  $\mu_A = \mu_{W_A}$  and  $K_{AB} = K_{W_A W_B}$  for all  $A, B \in V$ .

$$\begin{aligned} \theta^\top t(x) &= \mu^\top K z - \frac{1}{2} \text{vec}(K)^\top \text{vec}(zz^\top) \\ &= \mu^\top K z - \frac{1}{2} z^\top K z \\ &= z_\Gamma^\top K_{\Gamma\Gamma} \mu_\Gamma + z_\Gamma^\top K_{\Gamma\Delta} \mu_\Delta + \mu_\Gamma^\top K_{\Gamma\Delta} z_\Delta + \mu_\Delta^\top K_{\Delta\Delta} z_\Delta - \frac{1}{2} z_\Gamma^\top K_{\Gamma\Gamma} z_\Gamma - z_\Gamma^\top K_{\Gamma\Delta} z_\Delta - \frac{1}{2} z_\Delta^\top K_{\Delta\Delta} z_\Delta \\ &= \mu_\Gamma^\top K_{\Gamma\Delta} z_\Delta - \frac{1}{2} z_\Delta^\top K_{\Delta\Delta} z_\Delta + \mu_\Delta^\top K_{\Delta\Delta} z_\Delta + (K_{\Gamma\Gamma} \mu_\Gamma + K_{\Gamma\Delta} (\mu_\Delta - z_\Delta))^\top z_\Gamma - \frac{1}{2} z_\Gamma^\top K_{\Gamma\Gamma} z_\Gamma \\ &= g(z_\Delta) + h(z_\Delta)^\top z_\Gamma - \frac{1}{2} z_\Gamma^\top K_{\Gamma\Gamma} z_\Gamma \end{aligned}$$

Accordingly,

$$g(z_\Delta) = \mu_\Gamma^\top K_{\Gamma\Delta} z_\Delta - \frac{1}{2} z_\Delta^\top K_{\Delta\Delta} z_\Delta + \mu_\Delta^\top K_{\Delta\Delta} z_\Delta \quad h(z_\Delta) = K_{\Gamma\Gamma} \mu_\Gamma + K_{\Gamma\Delta} (\mu_\Delta - z_\Delta).$$

Therefore

$$\begin{aligned}
g^*(z_\Delta) &= g(z_\Delta) + \frac{1}{2}h(z_\Delta)^\top K_{\Gamma\Gamma}^{-1}h(z_\Delta) \\
&= \mu_\Gamma^\top K_{\Gamma\Delta}z_\Delta - \frac{1}{2}z_\Delta^\top K_{\Delta\Delta}z_\Delta + \mu_\Delta^\top K_{\Delta\Delta}z_\Delta \\
&\quad + \frac{1}{2}(K_{\Gamma\Gamma}\mu_\Gamma + K_{\Gamma\Delta}(\mu_\Delta - z_\Delta))^\top K_{\Gamma\Gamma}^{-1}(K_{\Gamma\Gamma}\mu_\Gamma + K_{\Gamma\Delta}(\mu_\Delta - z_\Delta)) \\
&= -\frac{1}{2}z_\Delta^\top K_{\Delta\Delta}z_\Delta + \mu_\Delta^\top K_{\Delta\Delta}z_\Delta + \frac{1}{2}\mu_\Gamma^\top K_{\Gamma\Gamma}\mu_\Gamma + \mu_\Gamma^\top K_{\Gamma\Delta}\mu_\Delta \\
&\quad + \frac{1}{2}\mu_\Delta^\top K_{\Delta\Gamma}K_{\Gamma\Gamma}^{-1}K_{\Gamma\Delta}\mu_\Delta - \mu_\Delta^\top K_{\Delta\Gamma}K_{\Gamma\Gamma}^{-1}K_{\Gamma\Delta}z_\Delta + \frac{1}{2}z_\Delta^\top K_{\Delta\Gamma}K_{\Gamma\Gamma}^{-1}K_{\Gamma\Delta}z_\Delta \\
&= -\frac{1}{2}z_\Delta^\top (K_{\Delta\Delta} - K_{\Delta\Gamma}K_{\Gamma\Gamma}^{-1}K_{\Gamma\Delta})z_\Delta + \mu_\Delta^\top (K_{\Delta\Delta} - K_{\Delta\Gamma}K_{\Gamma\Gamma}^{-1}K_{\Gamma\Delta})z_\Delta \\
&\quad + \frac{1}{2}\mu_\Gamma^\top K_{\Gamma\Gamma}\mu_\Gamma + \mu_\Gamma^\top K_{\Gamma\Delta}\mu_\Delta + \frac{1}{2}\mu_\Delta^\top K_{\Delta\Delta}\mu_\Delta - \frac{1}{2}\mu_\Delta^\top (K_{\Delta\Delta} - K_{\Delta\Gamma}K_{\Gamma\Gamma}^{-1}K_{\Gamma\Delta})\mu_\Delta \\
&= -\frac{1}{2}z_\Delta^\top \Sigma_{\Delta\Delta}^{-1}z_\Delta + \mu_\Delta^\top \Sigma_{\Delta\Delta}^{-1}z_\Delta + \frac{1}{2}(\mu_\Gamma^\top K \mu - \mu_\Delta^\top \Sigma_{\Delta\Delta}^{-1}\mu_\Delta)
\end{aligned}$$

where  $\Sigma$  is the covariance matrix and the transformation  $\Sigma_{AA}^{-1} = K_{AA} - K_{AB}K_{BB}^{-1}K_{BA}$  is given in Bishop [10] and

$$\begin{aligned}
\xi_\Gamma(z_\Delta) &= K_{\Gamma\Gamma}^{-1}h(z_\Delta) \\
&= \mu_\Gamma + K_{\Gamma\Gamma}^{-1}K_{\Gamma\Delta}(\mu_\Delta - z_\Delta).
\end{aligned}$$

Accordingly, a Lee and Hastie density is proportional to the following:

$$f(x \mid \theta) \propto \exp \left[ \underbrace{g^*(z_\Delta)}_{\text{MRF}} - \frac{1}{2} \underbrace{(z_\Gamma - \xi_\Gamma(z_\Delta))^\top K_{\Gamma\Gamma}(z_\Gamma - \xi_\Gamma(z_\Delta))}_{\text{mean shifted Gaussian}} \right]$$

Notably, the pairwise edge potentials and the pairwise MRF should be associated with the undirected augmented graph rather than the original directed MAG.

**Proposition 5.3.2.** *Let  $\mathcal{G} = (V, E)$  be a directed MAG satisfying the LHMC. If  $\Phi_{\mathcal{G}}^{-1} : \Lambda \times \Omega \times \mathbf{B} \times \boldsymbol{\mu} \rightarrow \Theta_{\mathcal{G}}$  is the diffeomorphism corresponding to the parameterization of  $\mathcal{G}$ , then  $\Phi_{\mathcal{G}}^{-1}$  is a bijection.*

*Proof.*  $\Phi_{\mathcal{G}}^{-1}$  is surjective by construction, therefore we show that  $\Phi_{\mathcal{G}}^{-1}$  is injective.

If  $D = \text{ch}_{z(\mathcal{G})}(W) \cup \text{sp}_{z(\mathcal{G})}(W)$  and  $U = W \setminus D$  are the directed and undirected vertices

of  $z(\mathcal{G})$  respectively, then:

$$f(x | \theta) \propto \exp \left[ -\frac{1}{2} z_U^\top \Sigma_{UU}^{-1} z_U + \mu_U^\top \Sigma_{UU}^{-1} z_U + \frac{1}{2} (\mu^\top K \mu - \mu_U^\top \Sigma_{UU}^{-1} \mu_U) - \frac{1}{2} (z_D - \xi_D(z_U))^\top K_{DD} (z_D - \xi_D(z_U)) \right].$$

Notably,  $D \subseteq z(\Gamma)$  and  $\Sigma_{UU}^{-1} = \Lambda$  and the directed part of a Lee and Hastie probability measure is Gaussian. Accordingly, we check the undirected non-constant part of Lee and Hastie probability measures. If  $A = U \cap \Delta$  and  $B = U \cap \Gamma$ , then:

$$\begin{aligned} -\frac{1}{2} z_U^\top \Sigma_{UU}^{-1} z_U + \mu_U^\top \Sigma_{UU}^{-1} z_U &= -\frac{1}{2} z_U^\top \Lambda z_U + \mu_U^\top \Lambda z_U \\ &= -\frac{1}{2} z_A^\top \Lambda_{AA} z_A - z_A^\top \Lambda_{AB} z_B - \frac{1}{2} z_B^\top \Lambda_{BB} z_B \\ &\quad + \mu_A^\top \Lambda_{AA} z_A + \mu_A^\top \Lambda_{AB} z_B + \mu_B^\top \Lambda_{BA} z_A + \mu_B^\top \Lambda_{BB} z_B \\ &= \sum_{a \in A} \sum_{a' \in A} -\frac{1}{2} \lambda_{aa'} z_{a'} z_a - \sum_{a \in A} \sum_{b \in B} \lambda_{ab} z_a z_b + \sum_{b \in B} \sum_{b' \in B} -\frac{1}{2} \lambda_{bb'} z_b z_{b'} \\ &\quad + \sum_{a \in A} \left[ \sum_{a' \in A \setminus z_{a_1}} \lambda_{aa'} \mu_{a'} + \sum_{b \in B} \lambda_{ab} \mu_b \right] z_a + \sum_{b \in B} \left[ \sum_{a \in A} \lambda_{ab} \mu_a + \sum_{b' \in B} \lambda_{bb'} \mu_{b'} \right] z_b \\ &= \sum_{a \in A} \sum_{a' \in A \setminus z_{a_1}} -\frac{1}{2} \lambda_{aa'} z_{a'} z_a - \sum_{a \in A} \sum_{b \in B} \lambda_{ab} z_a z_b + \sum_{b \in B} \sum_{b' \in B} -\frac{1}{2} \lambda_{bb'} z_b z_{b'} \\ &\quad + \sum_{a \in A} \left[ \sum_{a' \in A \setminus z_{a_1}} \lambda_{aa'} \mu_{a'} + \sum_{b \in B} \lambda_{ab} \mu_b + \mu_a - \frac{1}{2} \right] z_a + \sum_{b \in B} \left[ \sum_{a \in A} \lambda_{ab} \mu_a + \sum_{b' \in B} \lambda_{bb'} \mu_{b'} \right] z_b. \end{aligned}$$

The first three terms are edge potentials, where the lambda terms are non-zero if and only if the two corresponding vertices are adjacent. Accordingly, every unique value of the lambda terms results in a different probability measure. The last two terms are vertex potentials. Therefore, for fixed lambda terms, every unique value of the mu terms results in a different probability measure. It follows that the diffeomorphism is a bijection.

The Lee and Hastie family of probability measures uses the same parameterization as the Gaussian family of probability measures after transforming the variables. Thus, by applying the transformation to both the variables and the graph, we have that Lee and Hastie directed MAG probability measures are curved exponential families.  $\square$

**Corollary 5.3.1.** *If  $\mathcal{G} = (V, E)$  is a directed MAG and  $z(\mathcal{G}) = (W, F)$  be the transformed*

directed MAG, then  $\mathcal{F}_{\text{LH}}(\mathcal{G})$  is a curved exponential family with dimension  $|W| + |\Gamma| + |F|$ .

*Proof.* The proof follows from Proposition 5.2.1 because Lee and Hastie exponential Families and Gaussian exponential families use the same diffeomorphism.  $\square$

**Proposition 5.3.3** (Proposition 3.1 [46]). *Let  $X$  be a collection of random variables and  $P_\theta$  be a probability measure on  $X$  that admits density  $f(x)$  with respect to dominating  $\sigma$ -finite product measure  $\nu$ . If  $f(x) > 0$  is positive for all  $x \in \mathcal{X}$ , then  $\mathcal{J}(P_\theta)$  is a graphoid.*

**Lemma 5.3.1.** *Let  $X$  be a collection of random variables and  $P_\theta$  be a probability measure on  $X$  that admits density  $f(x)$  with respect to dominating  $\sigma$ -finite product measure  $\nu$ . If  $P_\theta$  is a Lee and Hastie probability measure, then  $\mathcal{J}(P_\theta)$  is a graphoid.*

*Proof.* This directly follows from Proposition 5.3.3 and the well-known fact that the density admitted by a Gaussian probability measure is positive.  $\square$

**Proposition 5.3.4** (Lemma 8.17 [70]). *If  $K$  is a precision matrix parameterized by a MAG  $\mathcal{G} = (V, E)$ , and  $a, b \in V$  are not adjacent in  $\text{aug}(\mathcal{G})$  then  $(K)_{ab} = 0$ .*

**Lemma 5.3.2.** *Let  $\mathcal{G} = (V, E)$  be a directed MAG and  $P_\theta$  be a Lee and Hastie probability measure over  $V$ . If  $\theta$  is maximal parameterization with respect to  $\mathcal{G}$ , then for all disjoint  $A, B, C \subseteq V$  where  $\text{ang}_{\mathcal{G}}(ABC) = V$*

$$A \perp\!\!\!\perp B \mid C [\mathcal{G}] \quad \Leftrightarrow \quad A \perp\!\!\!\perp B \mid C [P_\theta].$$

*Proof.* Let  $\mathcal{G} = (V, E)$  be a directed MAG and  $X$  be a collection of random variables indexed by  $V$  with probability measure  $P_\theta$  that admits density  $f(x \mid \theta)$  with respect to dominating  $\sigma$ -finite product measure  $\nu$ . We note that  $\mathcal{G}$  and  $z(\mathcal{G})$  have the same conditional independence relationships by Proposition 5.3.1 and apply the Proposition 5.3.4. Let  $A, B, C \subseteq V$  be three sets that partition  $V$ . Let  $D = AB$  and define  $r = z - \mu$  as the residual of  $z$  given the mean. We use the shorthand  $K_{W_A W_A} = K_{AA}$ .

$$\log f(x \mid \theta) \propto -\frac{1}{2} r_C^\top K_{CC} r_C - r_C^\top K_{CD} r_D - \frac{1}{2} r_D^\top K_{DD} r_D$$

$$\begin{aligned}
r_C^\top K_{CC} r_C &= \sum_{c \in C} \sum_{c' \in C} K_{cc'} r_c r_{c'} \\
r_C^\top K_{CD} r_D &= \sum_{c \in C} \sum_{d \in D} K_{cd} r_c r_d \\
&= \sum_{a \in A} \sum_{c \in C} K_{ac} r_a r_c + \sum_{b \in B} \sum_{c \in C} K_{bc} r_b r_c \\
r_D^\top K_{DD} r_D &= \sum_{d \in D} \sum_{d' \in D} K_{dd'} r_d r_{d'} \\
&= \sum_{a \in A} \sum_{a' \in A} K_{aa'} r_a r_{a'} + \sum_{a \in A} \sum_{b \in B} K_{ab} r_a r_b + \sum_{b \in B} \sum_{b' \in B} K_{bb'} r_b r_{b'}
\end{aligned}$$

$$\begin{aligned}
\log f(x \mid \theta) &\propto -\frac{1}{2} \sum_{a \in A} \sum_{a' \in A} K_{aa'} r_a r_{a'} - \sum_{a \in A} \sum_{c \in C} K_{ac} r_a r_c \\
&\quad - \frac{1}{2} \sum_{b \in B} \sum_{b' \in B} K_{bb'} r_b r_{b'} - \sum_{b \in B} \sum_{c \in C} K_{bc} r_b r_c \\
&\quad - \frac{1}{2} \sum_{c \in C} \sum_{c' \in C} K_{cc'} r_c r_{c'} - \frac{1}{2} \sum_{a \in A} \sum_{b \in B} K_{ab} r_a r_b
\end{aligned}$$

Let  $a \in A$  and  $b \in B$  be variables. From the equation above, we see that  $f(x \mid \theta)$  can be split into a function of  $\{a\} \cup C$  and a function of  $\{b\} \cup C$  if and only if  $K_{a,b} = 0$ . This occurs if and only if  $a$  and  $b$  are not adjacent in  $\text{aug}(z(\mathcal{G}))$ . Furthermore,  $f(x \mid \theta)$  can be split into a function of  $\{a\} \cup C$  and a function of  $\{b\} \cup C$  if and only if  $a \perp\!\!\!\perp b \mid C [P_\theta]$ .

Accordingly,  $a \perp\!\!\!\perp b \mid C [P_\theta]$  if and only if  $K_{a,b} = 0$ .  $\square$

**Theorem 5.3.1.** *Let  $\mathcal{G}$  be a directed MAG satisfying the LHMC. If  $\mathcal{F}_{\text{LH}}(\mathcal{G})$  is the family of Lee and Hastie probability measures parameterized by  $\mathcal{G}$  and  $F_{\text{LH}}(\mathcal{J}(\mathcal{G}))$  is the family of Lee and Hastie probability measures constrained by  $\mathcal{J}(\mathcal{G})$ , then*

$$\mathcal{F}_{\text{LH}}(\mathcal{G}) = \mathcal{F}_{\text{LH}}(\mathcal{J}(\mathcal{G})).$$

*Proof.* This follows from the LHMC, Theorem 3.3.4, and Lemmas 5.3.1 and 5.3.2.  $\square$

## 6.0 Scoring Criterion and Applications

In this chapter we discuss an application of the factorization derived in the Section 4.3. In particular, we formulate a consistent probabilistic score with a closed-form solution for exponential families whose independence models are described by directed MAGs—directed ancestral graph Markov models. The families discussed in this dissertation are subfamilies of conditional Gaussian probability measures and include the families of Gaussian probability measures and multinomial probability measures.

The consistent probabilistic score developed in this chapter is formulated by employing an approximation of the maximum log-likelihood with respect to a directed MAG in the well known Bayesian information criterion (BIC). Notably, the BIC using the exact maximum log-likelihood with respect to a directed MAG also provides a consistent probabilistic score, however, the resulting score does not always have a closed-form solution for the families of probability measures considered in this dissertation. Furthermore, calculation of the exact maximum log-likelihood with respect to a MAG requires the development of family specific solvers—solvers have been developed for Gaussian and multinomial directed ancestral graph Markov models [20, 30]. In contrast, the approximate maximum log-likelihood calculation developed in this chapter only requires knowledge of the unconstrained probability density. We compare the ability of the exact and approximate probabilistic scores to recover the correct Markov equivalence class for Gaussian directed ancestral graph Markov models and report run times.

Historically, methods that optimize a score for directed MAG Markov equivalence class recovery have not seen much development due to theoretical complications. Instead, directed MAG Markov equivalence class recovery has been done by the fast causal inference (FCI) algorithm and its variants. These methods rely on a series of conditional independence tests in order to learn a Markov equivalence class; this approach readily handles latent variables. Accordingly, there is an abundance of FCI variants in the literature [57, 63, 80, 93]. However, in these approaches, errors made by conditional independence tests can propagate, compound, and result in poor overall performance. Furthermore, these methods give no



indication of how much better the best Markov equivalence class is compared to the next best Markov equivalence class [78]. These issues are non-existent in methods that optimize a score, which motivates their development.

Indeed, in the past five years there has been an influx of methods capable of learning directed MAGs by optimizing a score. These methods include: a method that searches over causal orders [8], a continuous optimization method [9], an integer programming method [12], a method that scores conditional independence statements [41, 42], steepest ascent hill climbing methods [56, 88], and a method that uses an independence-based subroutine to determine local structures [89]. The majority of these methods use the exact score described above and are therefore candidates for the approximate score. By switching out the exact score for the approximate score, these methods gain flexibility and computational efficiency. Additionally, Appendix B.5 shows the similarity between one of these methods and our score. We compare the probabilistic scores against the FCI algorithm and two of its variants to compare the performance of a score based approach to a constraint based approach.

Ultimately, we design a local causal discovery algorithm called the ancestral probability (AP) procedure, which estimates the posterior probabilities of ancestral relationships using the probabilistic score developed in this chapter. The idea of local causal discovery, originally formulated by Cooper as a constraint based approach and later extended to score based methods by Mani [15, 52], focus on local subsets of variables in order to efficiently target specific causal relationships. We evaluate the AP procedure on synthetically generated data and a real data set measuring airborne pollutants, cardiovascular health, and respiratory health.

## 6.1 Asymptotic Behavior of Directed MAG Curved Exponential Families

In this section, we formulate a consistent probabilistic score with a closed-form for curved exponential families whose independence models are described by directed MAGs. We investigate the theoretical and empirical asymptotic behavior of curved exponential families subject to the parametric constraints of independence models induced by directed MAG. We

compare the ability of the probabilistic score to recover the correct Markov equivalence class against the well-known FCI algorithm and two of its variants.

### 6.1.1 Theoretical Evaluation

Let  $\mathcal{G} = (V, E)$  be a directed MAG. Furthermore, let  $X$  be a collection of random variables indexed by  $V$  with probability measure  $P_\theta$  that admits density  $f(x | \theta)$  with respect to dominating  $\sigma$ -finite product measure  $\nu$ . Throughout this section, let  $P_\theta$  belong to a curved exponential family with parameter space  $\Theta$  and  $x^1, \dots, x^n \stackrel{iid}{\sim} f(x | \theta)$ . It will be useful to review preexisting theoretical results. Define log-likelihood as follows:

$$\ell(\hat{\theta} | x^1, \dots, x^n) \equiv \sum_{i=1}^n \log f(x^i | \hat{\theta})$$

where  $\hat{\theta} \in \Theta$ . Berks proved strong consistency for the maximum likelihood parameter estimates of exponential families under mild regularity conditions [7]. If  $\theta \in \Theta_{\mathcal{G}}$ , then:

$$\hat{\theta}_{\mathcal{G},n}^{\text{mle}} \xrightarrow{\text{a.s.}} \theta.$$

Therefore, by the continuous mapping theorem:

$$\ell(\hat{\theta}_{\mathcal{G},n}^{\text{mle}} | x^1, \dots, x^n) \xrightarrow{\text{a.s.}} \ell(\theta | x^1, \dots, x^n).$$

Haughton provides a computationally efficient approximation of marginal likelihood for curved exponential families called the Bayesian information criterion (BIC) using the maximum likelihood and a parameter penalty [36, 75]:

$$\text{BIC}(\mathcal{G}, x^1, \dots, x^n) \equiv \ell(\hat{\theta}_{\mathcal{G},n}^{\text{mle}} | x^1, \dots, x^n) - \frac{|\Theta_{\mathcal{G}}|}{2} \log(n)$$

and approximates the log marginal likelihood up to a constant under mild regularity conditions:

$$\begin{aligned} \log \Pr(x^1, \dots, x^n | \mathcal{G}) &= \log \int_{\theta \in \Theta_{\mathcal{G}}} \prod_{i=1}^n f(x^i | \theta) d\nu(\theta) \\ &= \text{BIC}(\mathcal{G}, x^1, \dots, x^n) + O_p(1). \end{aligned}$$

Notably,  $P_\theta$  satisfies the global Markov condition with respect to a directed MAG  $\mathcal{G}$  if and only if  $\theta \in \Theta_{\mathcal{G}}$ . The BIC is a consistent score for model selection.

**Proposition 6.1.1** (Proposition 1.2 [36]). *Let  $\mathcal{G} = (V, E)$  and  $\mathcal{G}' = (V, E')$  be directed MAGs. Furthermore, let  $X$  be a collection of random variables indexed by  $V$  with probability measure  $P_\theta$  that admits density  $f(x | \theta)$  with respect to dominating  $\sigma$ -finite product measure  $\nu$ . If  $x^1, \dots, x^n \stackrel{iid}{\sim} f(x | \theta)$  and either  $\theta \in (\Theta_{\mathcal{G}'} \setminus \Theta_{\mathcal{G}})$  or  $(\Theta_{\mathcal{G}} \cap \Theta_{\mathcal{G}'})$  with  $|\Theta_{\mathcal{G}'}| < |\Theta_{\mathcal{G}}|$ , then:*

$$\lim_{n \rightarrow \infty} \Pr(\text{BIC}(\mathcal{G}, x^1, \dots, x^n) < \text{BIC}(\mathcal{G}', x^1, \dots, x^n)) = 1$$

where  $\Theta_{\mathcal{G}'} \subseteq \Theta_{\mathcal{G}}$  if  $\mathcal{J}(\mathcal{G}) \subseteq \mathcal{J}(\mathcal{G}')$ .

The BIC has a closed-form solution for the curved exponential families when the parameter space is constrained by an independence model induced by a DAG. Unfortunately, this is not always the case when the parameter space is constrained by an independence model induced by a directed MAG. We develop an approximation for the BIC that has a closed-form solution in both cases.

Let  $\mathcal{G}' = \text{dom}(\mathcal{G}, \leq)$  and define an approximate log-likelihood using the factorization with respect to  $\mathcal{G}$  and  $\leq$ :

$$\hat{\ell}_{\mathcal{G}}^{\leq}(\hat{\theta} | x^1, \dots, x^n) \equiv \sum_{b \in V} \left[ \sum_{i=1}^n \log f_{b|\text{pa}_{\mathcal{G}'}(b)}(x^i | \hat{\theta}) - \sum_{\substack{N \subseteq \text{pa}_{\mathcal{G}'}^+(b) \\ b \in N}} u_{\mathcal{N}(\mathcal{G})}^{\leq,+}(N) \sum_{i=1}^n \phi_N(x^i | \hat{\theta}) \right]$$

Accordingly, we define the following score which approximates the BIC:

$$\widehat{\text{BIC}}(\mathcal{G}, \leq, x^1, \dots, x^n) \equiv \hat{\ell}_{\mathcal{G}}^{\leq}(\hat{\theta}_{\mathcal{G}',n}^{\text{mle}} | x^1, \dots, x^n) - \frac{|\Theta_{\mathcal{G}}|}{2} \log(n).$$

which simplifies to BIC if  $\mathcal{G}$  is a DAG and has nice asymptotic properties.

**Proposition 6.1.2.** *Let  $\mathcal{G} = (V, E)$  and  $\mathcal{G}' = (V, E')$  be directed MAGs and  $\leq$  and  $\leq'$  be total orders consistent with  $\mathcal{G}$  and  $\mathcal{G}'$  respectively. Furthermore, let  $X$  be a collection of random variables indexed by  $V$  with probability measure  $P_\theta$  that admits density  $f(x | \theta)$  with respect to dominating  $\sigma$ -finite product measure  $\nu$ .*

If  $x^1, \dots, x^n \stackrel{iid}{\sim} f(x | \theta)$  and  $\theta \in \Theta_{\mathcal{G}'} \setminus \Theta_{\mathcal{G}}$ , then:

$$\lim_{n \rightarrow \infty} \frac{1}{n} \left| \hat{\text{BIC}}(\mathcal{G}', \leq', x^1, \dots, x^n) - \hat{\text{BIC}}(\mathcal{G}, \leq, x^1, \dots, x^n) \right| = (\mu_{\mathcal{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq, +})^\top m_{P_\theta}.$$

If  $x^1, \dots, x^n \stackrel{iid}{\sim} f(x | \theta)$  and  $\theta \in \Theta_{\mathcal{G}} \cap \Theta_{\mathcal{G}'}$  with  $|\Theta_{\mathcal{G}}| < |\Theta_{\mathcal{G}'}|$ , then:

$$\lim_{n \rightarrow \infty} \frac{1}{\log n} \left| \hat{\text{BIC}}(\mathcal{G}', \leq', x^1, \dots, x^n) - \hat{\text{BIC}}(\mathcal{G}, \leq, x^1, \dots, x^n) \right| = \frac{|\Theta_{\mathcal{G}}| - |\Theta_{\mathcal{G}'}|}{2}.$$

*Proof.* Let  $\mathcal{G}'' = \text{dom}(\mathcal{G}, \leq)$ . If  $x^1, \dots, x^n \stackrel{iid}{\sim} f(x | \theta)$  and  $\theta \in \Theta_{\mathcal{G}'} \setminus \Theta_{\mathcal{G}}$ , then:

$$\begin{aligned} & \lim_{n \rightarrow \infty} \frac{1}{n} \left| \hat{\text{BIC}}(\mathcal{G}', \leq', x^1, \dots, x^n) - \hat{\text{BIC}}(\mathcal{G}, \leq, x^1, \dots, x^n) \right| \\ &= \lim_{n \rightarrow \infty} \frac{1}{n} \ell(\hat{\theta}_{\mathcal{G}', n}^{\text{mle}} | x^1, \dots, x^n) - \lim_{n \rightarrow \infty} \frac{1}{n} \hat{\ell}_{\mathcal{G}}^{\leq}(\hat{\theta}_{\mathcal{G}', n}^{\text{mle}} | x^1, \dots, x^n) - \lim_{n \rightarrow \infty} \frac{|\Theta_{\mathcal{G}'}| - |\Theta_{\mathcal{G}}|}{2n} \log(n) \\ &= (\mu_{\mathcal{P}} \delta_{\mathcal{P}(V)})^\top m_{P_\theta} - (\mu_{\mathcal{P}} \delta_{\mathcal{M}(\mathcal{G})} - \mu_{\mathcal{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq, -})^\top m_{P_\theta} \\ &= (\mu_{\mathcal{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq, +})^\top m_{P_\theta}. \end{aligned}$$

If  $x^1, \dots, x^n \stackrel{iid}{\sim} f(x | \theta)$  and  $\theta \in \Theta_{\mathcal{G}} \cap \Theta_{\mathcal{G}'}$  with  $|\Theta_{\mathcal{G}}| < |\Theta_{\mathcal{G}'}|$ , then:

$$\begin{aligned} & \lim_{n \rightarrow \infty} \frac{1}{\log n} \left| \hat{\text{BIC}}(\mathcal{G}', \leq', x^1, \dots, x^n) - \hat{\text{BIC}}(\mathcal{G}, \leq, x^1, \dots, x^n) \right| \\ &= \lim_{n \rightarrow \infty} \frac{1}{\log n} \left[ \ell(\hat{\theta}_{\mathcal{G}', n}^{\text{mle}} | x^1, \dots, x^n) - \hat{\ell}_{\mathcal{G}}^{\leq}(\hat{\theta}_{\mathcal{G}', n}^{\text{mle}} | x^1, \dots, x^n) \right] - \frac{|\Theta_{\mathcal{G}'}| - |\Theta_{\mathcal{G}}|}{2} \\ &= \frac{|\Theta_{\mathcal{G}}| - |\Theta_{\mathcal{G}'}|}{2}. \end{aligned}$$

□

The approximate BIC is a consistent score for model selection.

**Corollary 6.1.1.** *Let  $\mathcal{G} = (V, E)$  and  $\mathcal{G}' = (V, E')$  be directed MAGs and  $\leq$  and  $\leq'$  be total orders consistent with  $\mathcal{G}$  and  $\mathcal{G}'$  respectively. Furthermore, let  $X$  be a collection of random variables indexed by  $V$  with probability measure  $P_\theta$  that admits density  $f(x | \theta)$  with respect to dominating  $\sigma$ -finite product measure  $\nu$ . If  $x^1, \dots, x^n \stackrel{iid}{\sim} f(x | \theta)$  and either  $\theta \in (\Theta_{\mathcal{G}'} \setminus \Theta_{\mathcal{G}})$  or  $(\Theta_{\mathcal{G}} \cap \Theta_{\mathcal{G}'})$  with  $|\Theta_{\mathcal{G}'}| < |\Theta_{\mathcal{G}}|$ , then:*

$$\lim_{n \rightarrow \infty} \Pr(\hat{\text{BIC}}(\mathcal{G}, \leq, x^1, \dots, x^n) < \hat{\text{BIC}}(\mathcal{G}', \leq', x^1, \dots, x^n)) = 1$$

where  $\Theta_{\mathcal{G}'} \subseteq \Theta_{\mathcal{G}}$  if  $\mathcal{J}(\mathcal{G}) \subseteq \mathcal{J}(\mathcal{G}')$ .

*Proof.* The proof directly follows from Proposition 6.1.2. □

In what follows, we reason about the asymptotic properties of the approximate BIC and its relation to the marginal likelihood. We first note the following result.

**Proposition 6.1.3** (Theorem 1 [18]). *Let  $\mathcal{G} = (V, E)$  be a directed MAG. Furthermore, let  $X$  be a collection of random variables indexed by  $V$  with probability measure  $P_\theta$  that admits density  $f(x | \theta)$  with respect to dominating  $\sigma$ -finite product measure  $\nu$ . If  $x^1, \dots, x^n \stackrel{iid}{\sim} f(x | \theta)$ , then:*

$$\lim_{n \rightarrow \infty} \log \frac{\int_{\hat{\theta} \in \Theta_{\mathcal{G}}} \prod_{i=1}^n f(x^i | \hat{\theta}) d\nu(\theta)}{\prod_{i=1}^n f(x^i | \theta)} = - \lim_{n \rightarrow \infty} n \inf_{\hat{\theta} \in \Theta_{\mathcal{G}}} \int_{x \in \mathcal{X}} \log \left[ \frac{f(x | \theta)}{f(x | \hat{\theta})} \right] dP_\theta(x) + O_p(n^{\frac{1}{2}})$$

where  $\inf_{\hat{\theta} \in \Theta_{\mathcal{G}}} \int_{x \in \mathcal{X}} \log \left[ \frac{f(x | \theta)}{f(x | \hat{\theta})} \right] dP_\theta(x) = 0$  if and only if  $\theta \in \Theta_{\mathcal{G}}$ .

**Theorem 6.1.1.** *Let  $\mathcal{G} = (V, E)$  and  $\mathcal{G}' = (V, E')$  be directed MAGs and  $\leq$  and  $\leq'$  be total orders consistent with  $\mathcal{G}$  and  $\mathcal{G}'$  respectively. Furthermore, let  $X$  be a collection of random variables indexed by  $V$  with probability measure  $P_\theta$  that admits density  $f(x | \theta)$  with respect to dominating  $\sigma$ -finite product measure  $\nu$ .*

*If  $x^1, \dots, x^n \stackrel{iid}{\sim} f(x | \theta)$  and  $\theta \in \Theta_{\mathcal{G}}$ , then:*

$$\hat{\text{BIC}}(\mathcal{G}, \leq, x^1, \dots, x^n) = \text{BIC}(\mathcal{G}, x^1, \dots, x^n) \text{ almost surely.}$$

*If  $x^1, \dots, x^n \stackrel{iid}{\sim} f(x | \theta)$  and  $\theta \in \Theta_{\mathcal{G}'} \setminus \Theta_{\mathcal{G}}$ , then:*

$$\lim_{n \rightarrow \infty} \frac{\left| \Pr(x^1, \dots, x^n | \mathcal{G}) - \exp \hat{\text{BIC}}(\mathcal{G}, \leq, x^1, \dots, x^n) \right|}{\exp \hat{\text{BIC}}(\mathcal{G}', \leq', x^1, \dots, x^n)} = O_p(\exp -n)$$

*Proof.* If  $\theta \in \Theta_{\mathcal{G}}$ , then  $\hat{\text{BIC}}(\mathcal{G}, \leq, x^1, \dots, x^n) = \text{BIC}(\mathcal{G}, x^1, \dots, x^n)$  almost surely by the continuous mapping theorem and strong consistency of the maximum likelihood estimate. If  $\theta \in \Theta_{\mathcal{G}'} \setminus \Theta_{\mathcal{G}}$ , then:

$$\begin{aligned} & \lim_{n \rightarrow \infty} \log \frac{\left| \Pr(x^1, \dots, x^n | \mathcal{G}) - \exp \hat{\text{BIC}}(\mathcal{G}, \leq, x^1, \dots, x^n) \right|}{\exp \hat{\text{BIC}}(\mathcal{G}', \leq', x^1, \dots, x^n)} \\ & \leq \lim_{n \rightarrow \infty} \log \frac{\max \left[ \Pr(x^1, \dots, x^n | \mathcal{G}), \exp \hat{\text{BIC}}(\mathcal{G}, \leq, x^1, \dots, x^n) \right]}{\exp \hat{\text{BIC}}(\mathcal{G}', \leq', x^1, \dots, x^n)}. \end{aligned}$$

If  $\exp \hat{\text{BIC}}(\mathcal{G}, \leq, x^1, \dots, x^n) > \Pr(x^1, \dots, x^n \mid \mathcal{G})$ , then the results directly follows from Proposition 6.1.2. If  $\Pr(x^1, \dots, x^n \mid \mathcal{G}) > \exp \hat{\text{BIC}}(\mathcal{G}, \leq, x^1, \dots, x^n)$ , then the result directly follows from Proposition 6.1.3.  $\square$

Accordingly, the BIC and approximate BIC are equal almost surely when the probability measure is Markov with respect to the directed MAG. Furthermore, the difference between the log marginal likelihood and approximate BIC tends towards zero exponentially when the probability measure is not Markov with respect to the directed MAG relative to the approximate BIC for a directed MAG that is Markov with respect to the probability measure.

### 6.1.2 Empirical Evaluation

We supplement the theoretical evaluation with an empirical evaluation on synthetic data simulated from Gaussian densities as follows:

---

**Algorithm 6:** SIMULATE( $\mathcal{G}, n$ )

---

**Input:** directed MAG:  $\mathcal{G}$ , number of instances:  $n$   
**Output:** data:  $x^1, \dots, x^n$

**1 repeat**

**2**  $\Omega = (\omega_{ab})$  where  $\omega_{ab} \sim \begin{cases} \text{Uniform}[-0.7, -0.3] \cup [0.3, 0.7] & \text{if } a \leftrightarrow b \text{ in } \mathcal{G} \\ \text{Uniform}[1.0, 3.0] & \text{if } i = j \text{ in } \mathcal{G} \\ 0 & \text{otherwise} \end{cases}$  ;

**3 until**  $\Omega$  is positive-definite;

**4**  $B = (\beta_{ab})$  where  $\beta_{ab} \sim \begin{cases} \text{Uniform}[-0.7, -0.3] \cup [0.3, 0.7] & \text{if } a \leftarrow b \text{ in } \mathcal{G} \\ 0 & \text{otherwise} \end{cases}$  ;

**5**  $\Sigma = (I - B)^{-1} \Omega (I - B)^{-\top}$  ;

**6**  $x^1, \dots, x^n \sim \text{Gaussian}(0, \Sigma, n)$  ;

---

We compare our log-likelihood approximation against the maximum log-likelihood. The maximum log-likelihood is calculated using an R implementation of the iterative conditional fitting (ICF) procedure: <https://CRAN.R-project.org/package=ggm> [20]. Notably, ICF optimizes the likelihood for curved exponential families constrained MAG independence models, however, this space is not guaranteed to be convex. Accordingly, ICF does not necessarily converge to the MLE—in practice rarely converges to something other than then

MLE [21]. Furthermore, we are using a general implementation of ICF and not one that was designed to be efficient in this scenario. For comparison purposes, the approximate and exact negative log-likelihoods are shifted so that their smallest values are equal to 1. Notably, the smallest log-likelihoods always correspond to the saturated model, which is the same for both the approximate and exact methods. Accordingly, both methods are shifted by the same amount. The shifted negative log-likelihoods are compared on a log scale and each equivalence class is marked according to whether or not it is Markov with respect to the probability measure. Additional comparisons are given in Appendix B.6.

We use  $\hat{\text{BIC}}$  to exhaustively rank all directed MAG Markov equivalence classes. Histograms show the distribution of the MEC of the data generating graph in the ranking. That is, each bin represents the number of times the MEC of the true graph ranked according to the number associated with the bin. Notably, there are 248 possible positions in the ranking for the four-variable case and 24,259 possible positions in the ranking for the five-variable case. Accordingly, we enumerate all possible positions in the ranking on a log scale. Histograms for the exact BIC are given in Appendix B.7.

Finally, three causal discovery algorithms, FCI [80, 93], FCI max [63], and GFCE [57], were applied to the same data with several standard parameter settings for comparison. The reported number for each algorithm is the proportion of times that the Markov equivalence class of the true graph was returned; the numbers may be directly compared to the first bin of the corresponding histogram.

- FCI is a two stage search algorithm that attempts to recover the maximally informative PAG for a directed MAG from data using tests of conditional independence. The first stage starts with a completely connected graph and uses tests to determine which adjacencies to remove from the PAG. The second stage uses tests to determine invariant edge marks among the graphs in the equivalence class and orients them in the PAG [80, 93]. See Algorithm 11 for details. We use Fisher’s Z test with alpha levels of 0.01 and 0.001 for testing conditional independence.
- $\text{FCI}_{\text{max}}$  uses a maximum probability-based search technique in the edge orientation stage of FCI to determine which conditioning sets of variables are most likely to provide correct conditional independence statements. This approach has been shown to improve

performance, but requires significantly more tests [63]. We use Fisher’s Z test with alpha levels of 0.01 and 0.001 for testing conditional independence.

- GFCI optimizes a probabilistic score over DAGs using a greedy hill climbing approach and then runs FCI using the maximal DAG as a starting point rather than a completely connected graph [57]. We use BIC as a probabilistic score and Fisher’s Z test with alpha levels of 0.01 and 0.001 for testing conditional independence.  $\text{GFCI}_1$  uses standard BIC and  $\text{GFCI}_2$  using a variant of BIC where the parameter penalty has been doubled.

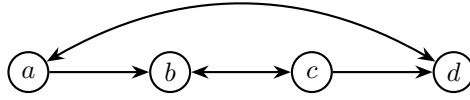
For all experiments, we simulate data sets of 500, 5,000 and 50,000 instances. We run experiments for 7 prespecified graphs, 4 of which have 4 vertices and 3 of which have 5 vertices, and random graphs. The random graph cases include graphs with 4 vertices and between 0 and 3 edges, graphs with 4 vertices and between 4 and 6 edge, graphs with 5 vertices and between 0 and 5 edges, and graphs with 5 vertices and between 6 and 10 edges. For each case, we run 1,000 repetitions. Each repetition has a unique parameterization and in the random graph cases, have unique graphs as well—barring random repeats. All experiments were run on a system with the following hardware:

- Memory: 7.7 GiB
- Processor: Intel<sup>®</sup> Core<sup>™</sup> i5-5200U CPU @ 2.20GHz  $\times$  4

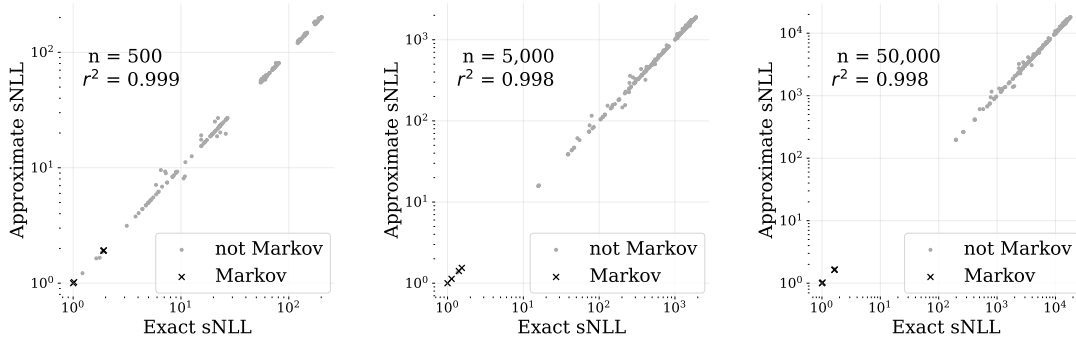
We perform paired z-tests to give an indication for whether the differences in performance are real. Statistical significance at an alpha level of 0.05 is reported as either an overline if  $\hat{\text{BIC}}$  is better or an underline if the alternative method is better. Note that there are no reported cases where  $\hat{\text{BIC}}$  and BIC are statistically significant at an alpha level of 0.05. In general, we find that the approximation for BIC performs well with low sample sizes and performs favorably compared to the other algorithms.



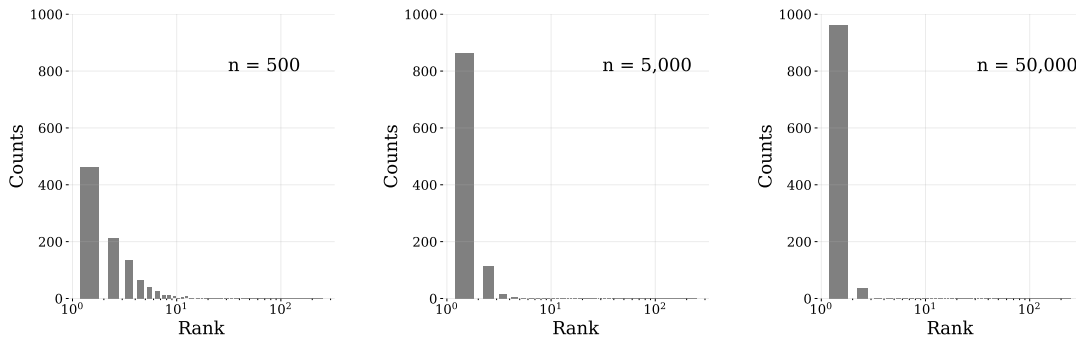
### Directed MAG



### Negative Log-likelihood



### MEC Recovery

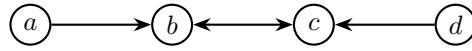


|                 | BIC   | $\hat{BIC}$ | FCI          |              | $FCI_{max}$  |              | GFCI <sub>1</sub> |              | GFCI <sub>2</sub> |              |
|-----------------|-------|-------------|--------------|--------------|--------------|--------------|-------------------|--------------|-------------------|--------------|
| $\alpha$ -level | -     | -           | 0.01         | 0.001        | 0.01         | 0.001        | 0.01              | 0.001        | 0.01              | 0.001        |
| n = 500         | 0.472 | 0.464       | <u>0.087</u> | <u>0.02</u>  | <u>0.265</u> | <u>0.202</u> | <u>0.617</u>      | <u>0.617</u> | 0.471             | 0.471        |
| n = 5,000       | 0.866 | 0.864       | <u>0.669</u> | <u>0.584</u> | <u>0.784</u> | <u>0.773</u> | <u>0.927</u>      | <u>0.929</u> | <u>0.926</u>      | <u>0.926</u> |
| n = 50,000      | 0.962 | 0.961       | <u>0.866</u> | <u>0.864</u> | <u>0.921</u> | <u>0.935</u> | <u>0.981</u>      | <u>0.981</u> | <u>0.979</u>      | <u>0.979</u> |

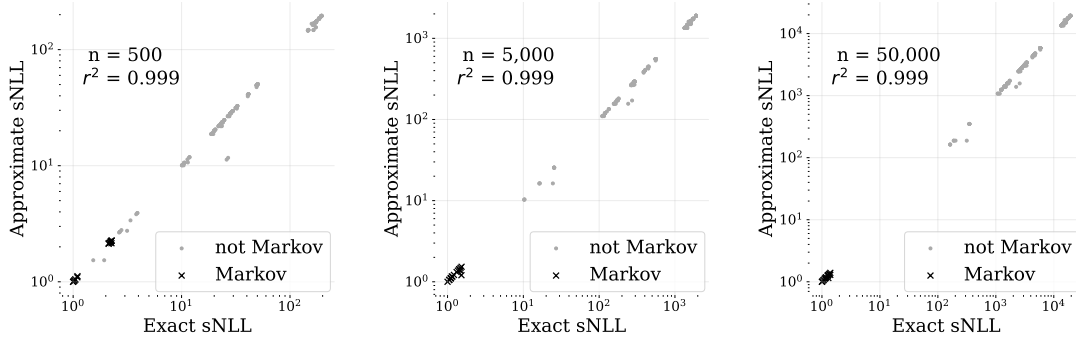
Figure 6.1: An evaluation of the approximate log-likelihood and BIC for the specified directed MAG with  $n = \{500, 5,000, 50,000\}$ . The approximate and exact shifted negative log-likelihoods are compared for a random parameterization. The approximate BIC ranking of the data generating MEC amongst all MECs is shown using histograms. The rate of recovery for the data generating MEC given by the highest scoring approximate BIC score is compared against several other state-of-the-art algorithms. Statistical significance at an alpha level of 0.05 is reported as either an overline if  $\hat{BIC}$  is better or an underline if the alternative method is better.

Figure 6.1 the prespecified graph is the simplest example showing that there may be no total order over the districts of a MAG. Notably, the prespecified graph is Markov equivalent to a DAG which perhaps explains the performance of GFCI—GFCI reduces to a state-of-the-art score based procedure in this case. The approximate log-likelihood closely aligns with the exact log-likelihood with clear separation of Markov versus not Markov as  $n \rightarrow \infty$ ; the approximate BIC performs nearly identically to BIC and consistently ranks the correct MEC in the top 10 with the ranking converging to a point-mass in the first bin as  $n \rightarrow \infty$ ; the top ranked approximate BIC model performs worse than GFCI, but better than the other methods in MEC recovery.

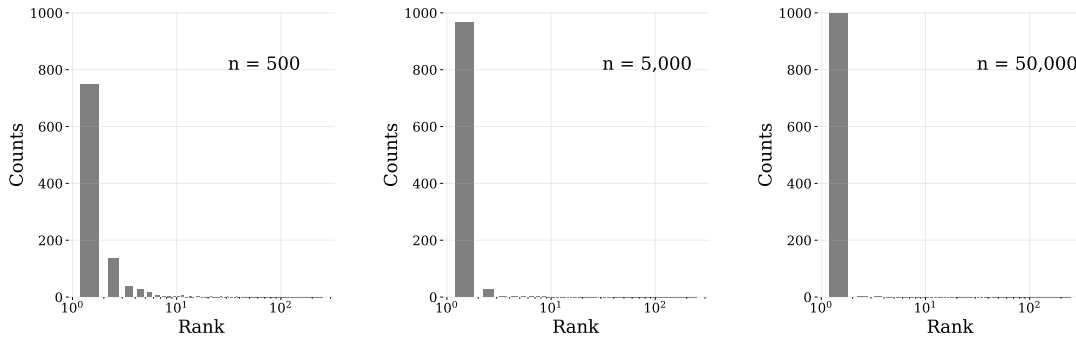
Directed MAG



Negative Log-likelihood



MEC Recovery



|                 | BIC   | $\hat{BIC}$ | FCI          |              | $FCI_{max}$  |              | GFCI <sub>1</sub> |              | GFCI <sub>2</sub> |              |
|-----------------|-------|-------------|--------------|--------------|--------------|--------------|-------------------|--------------|-------------------|--------------|
| $\alpha$ -level | -     | -           | 0.01         | 0.001        | 0.01         | 0.001        | 0.01              | 0.001        | 0.01              | 0.001        |
| n = 500         | 0.749 | 0.749       | <u>0.862</u> | 0.755        | <u>0.698</u> | <u>0.621</u> | <u>0.37</u>       | <u>0.345</u> | <u>0.16</u>       | <u>0.161</u> |
| n = 5,000       | 0.967 | 0.966       | <u>0.986</u> | <u>0.998</u> | 0.966        | <u>0.978</u> | <u>0.922</u>      | <u>0.927</u> | <u>0.829</u>      | <u>0.83</u>  |
| n = 50,000      | 0.997 | 0.997       | <u>0.988</u> | 1.0          | <u>0.988</u> | 1.0          | 0.995             | 1.0          | 0.994             | 0.999        |

Figure 6.2: An evaluation of the approximate log-likelihood and BIC for the specified directed MAG with  $n = \{500, 5,000, 50,000\}$ . Statistical significance at an alpha level of 0.05 is reported as either an overline if  $\hat{BIC}$  is better or an underline if the alternative method is better.

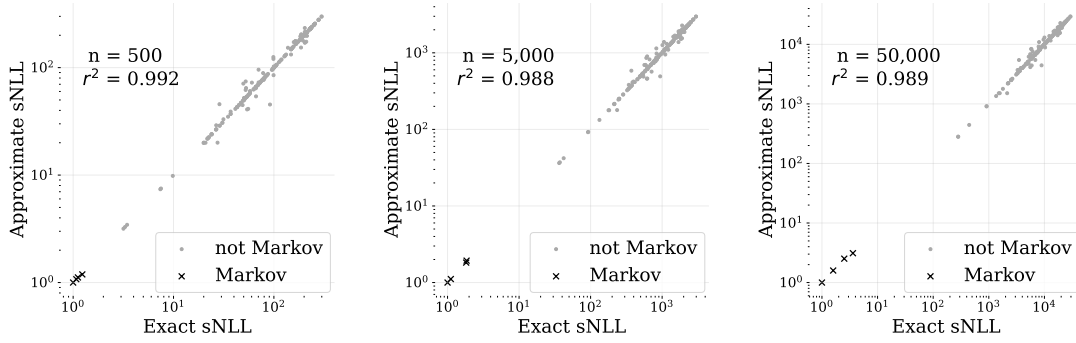
Figure 6.2 the prespecified graph is a MAG from the simplest MEC that does not contain a DAG in graphs with four vertices. The approximate log-likelihood closely aligns with the

exact log-likelihood with clear separation of Markov versus not Markov as  $n \rightarrow \infty$ ; the approximate BIC performs nearly identically to BIC and consistently ranks the correct MEC in the top 10 with the ranking converging to a point-mass in the first bin as  $n \rightarrow \infty$ ; the top ranked approximate BIC model performs worse than FCI, about the same as  $\text{FCI}_{\max}$ , and better than GFCI with low sample sizes and about the same otherwise in MEC recovery.

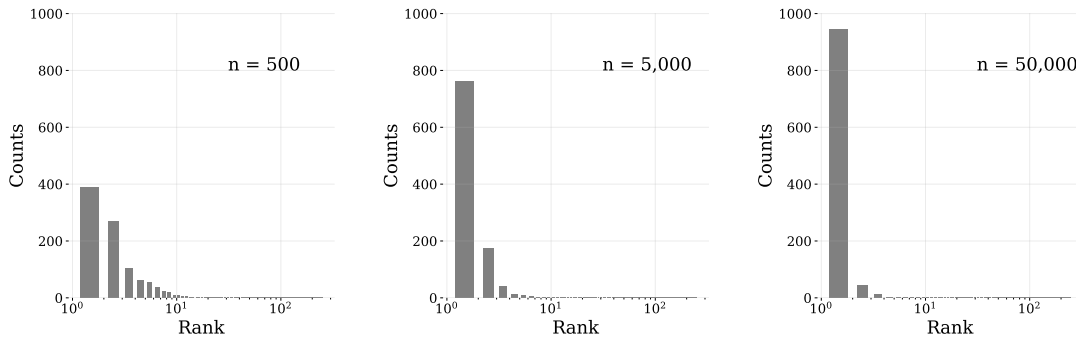
### Directed MAG



### Negative Log-likelihood



### MEC Recovery



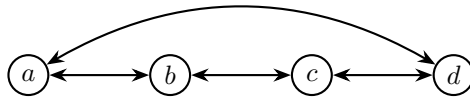
|                 | BIC   | $\hat{\text{BIC}}$ | FCI          |              | $\text{FCI}_{\max}$ |              | GFCI <sub>1</sub> |              | GFCI <sub>2</sub> |              |
|-----------------|-------|--------------------|--------------|--------------|---------------------|--------------|-------------------|--------------|-------------------|--------------|
| $\alpha$ -level | -     | -                  | 0.01         | 0.001        | 0.01                | 0.001        | 0.01              | 0.001        | 0.01              | 0.001        |
| n = 500         | 0.392 | 0.39               | <u>0.037</u> | <u>0.012</u> | <u>0.038</u>        | <u>0.012</u> | <u>0.081</u>      | <u>0.038</u> | <u>0.041</u>      | <u>0.024</u> |
| n = 5,000       | 0.764 | 0.763              | <u>0.348</u> | <u>0.247</u> | <u>0.348</u>        | <u>0.247</u> | <u>0.549</u>      | <u>0.49</u>  | <u>0.499</u>      | <u>0.466</u> |
| n = 50,000      | 0.941 | 0.943              | <u>0.832</u> | <u>0.787</u> | <u>0.832</u>        | <u>0.787</u> | <u>0.861</u>      | <u>0.862</u> | <u>0.844</u>      | <u>0.843</u> |

Figure 6.3: An evaluation of the approximate log-likelihood and BIC for the specified directed MAG with  $n = \{500, 5,000, 50,000\}$ . Statistical significance at an alpha level of 0.05 is reported as either an overline if  $\hat{\text{BIC}}$  is better or an underline if the alternative method is better.

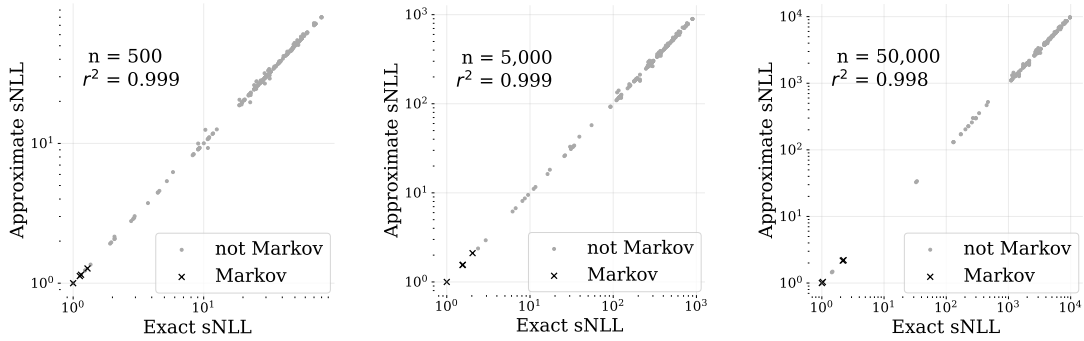
Figure 6.3 the prespecified graph is the simplest example of a discriminating path in

graphs with four vertices. The approximate log-likelihood closely aligns with the exact log-likelihood with clear separation of Markov versus not Markov as  $n \rightarrow \infty$ ; the approximate BIC performs nearly identically to BIC and consistently ranks the correct MEC in the top 10 with the ranking converging to a point-mass in the first bin as  $n \rightarrow \infty$ ; the top ranked approximate BIC model performs better than the other methods in MEC recovery.

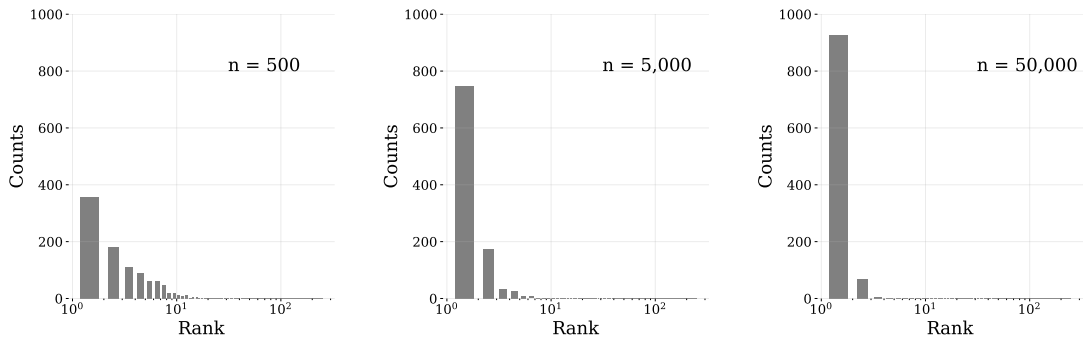
### Directed MAG



### Negative Log-likelihood



### MEC Recovery

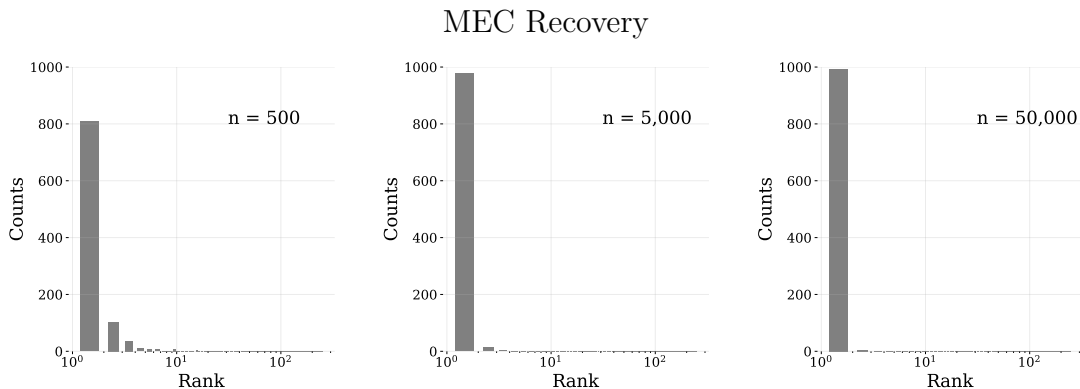


|                 | BIC   | $\hat{BIC}$ | FCI          |              | $FCI_{max}$  |              | GFCI <sub>1</sub> |              | GFCI <sub>2</sub> |              |
|-----------------|-------|-------------|--------------|--------------|--------------|--------------|-------------------|--------------|-------------------|--------------|
| $\alpha$ -level | -     | -           | 0.01         | 0.001        | 0.01         | 0.001        | 0.01              | 0.001        | 0.01              | 0.001        |
| n = 500         | 0.354 | 0.356       | <u>0.843</u> | <u>0.676</u> | <u>0.337</u> | <u>0.281</u> | <u>0.285</u>      | <u>0.253</u> | <u>0.172</u>      | <u>0.174</u> |
| n = 5,000       | 0.746 | 0.749       | <u>0.978</u> | <u>0.997</u> | <u>0.744</u> | 0.75         | <u>0.568</u>      | <u>0.578</u> | <u>0.501</u>      | <u>0.51</u>  |
| n = 50,000      | 0.928 | 0.927       | <u>0.985</u> | <u>0.996</u> | <u>0.92</u>  | 0.927        | <u>0.775</u>      | <u>0.779</u> | <u>0.711</u>      | <u>0.715</u> |

Figure 6.4: An evaluation of the approximate log-likelihood and BIC for the specified directed MAG with  $n = \{500, 5,000, 50,000\}$ . Statistical significance at an alpha level of 0.05 is reported as either an overline if  $\hat{BIC}$  is better or an underline if the alternative method is better.

In general, FCI tends to overestimate colliders and GFCI tends to underestimate colliders. Figure 6.4 the prespecified graph is a bi-directed four-cycle, which perhaps explains the poor performance of GFCI. The approximate log-likelihood closely aligns with the exact log-likelihood with clear separation of Markov versus not Markov as  $n \rightarrow \infty$ ; the approximate BIC performs nearly identically to BIC and consistently ranks the correct MEC in the top 10 with the ranking converging to a point-mass in the first bin as  $n \rightarrow \infty$ ; the top ranked approximate BIC model performs worse than FCI, about the same as  $\text{FCI}_{\max}$ , and better than GFCI in MEC recovery.

Random Directed MAGs with  $|V| = 4$  and  $|E| \in [0, 3]$



|                 | BIC   | $\hat{\text{BIC}}$ | FCI          |              | $\text{FCI}_{\max}$ |              | GFCI <sub>1</sub> |              | GFCI <sub>2</sub> |              |
|-----------------|-------|--------------------|--------------|--------------|---------------------|--------------|-------------------|--------------|-------------------|--------------|
| $\alpha$ -level | -     | -                  | 0.01         | 0.001        | 0.01                | 0.001        | 0.01              | 0.001        | 0.01              | 0.001        |
| $n = 500$       | 0.81  | 0.809              | <u>0.694</u> | <u>0.564</u> | 0.793               | <u>0.732</u> | <u>0.78</u>       | <u>0.777</u> | <u>0.704</u>      | <u>0.704</u> |
| $n = 5,000$     | 0.977 | 0.977              | 0.975        | 0.983        | <u>0.968</u>        | <u>0.985</u> | 0.973             | 0.978        | <u>0.947</u>      | <u>0.947</u> |
| $n = 50,000$    | 0.993 | 0.993              | <u>0.977</u> | <u>0.997</u> | <u>0.977</u>        | 0.996        | 0.993             | 0.994        | 0.996             | 0.997        |

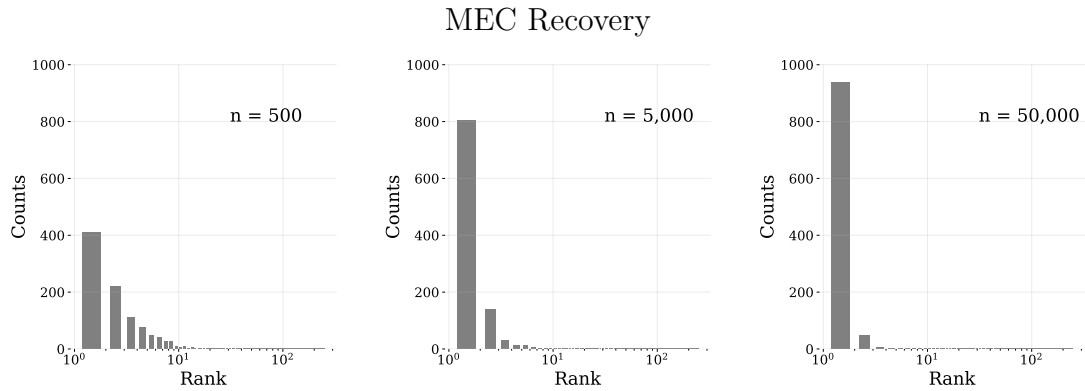
Figure 6.5: An evaluation of the approximate BIC for random directed MAGs with specified edge ranges and  $n = \{500, 5,000, 50,000\}$ . Statistical significance at an alpha level of 0.05 is reported as either an overline if  $\hat{\text{BIC}}$  is better or an underline if the alternative method is better.

Figure 6.5 the approximate BIC performs nearly identically to BIC and consistently ranks the correct MEC in the top 10 with the ranking converging to a point-mass in the



first bin as  $n \rightarrow \infty$ ; the top ranked approximate BIC model performs better than the other methods with low sample size and about the same otherwise in MEC recovery.

Random Directed MAGs with  $|V| = 4$  and  $|E| \in [4, 6]$



|                 | BIC   | $\hat{BIC}$ | FCI          |              | FCI <sub>max</sub> |              | GFCI <sub>1</sub> |              | GFCI <sub>2</sub> |              |
|-----------------|-------|-------------|--------------|--------------|--------------------|--------------|-------------------|--------------|-------------------|--------------|
| $\alpha$ -level | -     | -           | 0.01         | 0.001        | 0.01               | 0.001        | 0.01              | 0.001        | 0.01              | 0.001        |
| n = 500         | 0.41  | 0.41        | <u>0.237</u> | <u>0.125</u> | <u>0.305</u>       | <u>0.226</u> | <u>0.153</u>      | <u>0.194</u> | <u>0.238</u>      | <u>0.147</u> |
| n = 5,000       | 0.801 | 0.803       | <u>0.693</u> | <u>0.624</u> | <u>0.707</u>       | <u>0.662</u> | <u>0.582</u>      | <u>0.637</u> | <u>0.666</u>      | <u>0.574</u> |
| n = 50,000      | 0.941 | 0.939       | <u>0.875</u> | <u>0.861</u> | <u>0.873</u>       | <u>0.867</u> | <u>0.851</u>      | <u>0.863</u> | <u>0.864</u>      | <u>0.851</u> |

Figure 6.6: An evaluation of the approximate BIC for random directed MAGs with specified edge ranges and  $n = \{500, 5,000, 50,000\}$ . Statistical significance at an alpha level of 0.05 is reported as either an overline if  $\hat{BIC}$  is better or an underline if the alternative method is better.

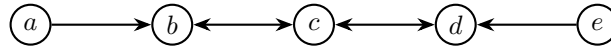
Figure 6.6 the approximate BIC performs nearly identically to BIC and consistently ranks the correct MEC in the top 10 with the ranking converging to a point-mass in the first bin as  $n \rightarrow \infty$ ; the top ranked approximate BIC model performs better than the other methods in MEC recovery.

| sample size | BIC         |             |             | $\hat{\text{BIC}}$ |            |            |
|-------------|-------------|-------------|-------------|--------------------|------------|------------|
|             | n = 500     | n = 5,000   | n = 50,000  | n = 500            | n = 5,000  | n = 50,000 |
| Figure 6.1  | 0.65 (0.18) | 0.65 (0.17) | 0.65 (0.18) | 0.01 (0.0)         | 0.01 (0.0) | 0.01 (0.0) |
| Figure 6.2  | 0.57 (0.04) | 0.57 (0.04) | 0.56 (0.04) | 0.01 (0.0)         | 0.01 (0.0) | 0.01 (0.0) |
| Figure 6.3  | 0.58 (0.05) | 0.58 (0.05) | 0.58 (0.05) | 0.01 (0.0)         | 0.01 (0.0) | 0.01 (0.0) |
| Figure 6.4  | 0.56 (0.04) | 0.56 (0.04) | 0.56 (0.04) | 0.01 (0.0)         | 0.01 (0.0) | 0.01 (0.0) |
| Figure 6.5  | 0.55 (0.03) | 0.54 (0.03) | 0.54 (0.03) | 0.01 (0.0)         | 0.01 (0.0) | 0.01 (0.0) |
| Figure 6.6  | 0.57 (0.04) | 0.57 (0.04) | 0.57 (0.03) | 0.01 (0.0)         | 0.01 (0.0) | 0.01 (0.0) |

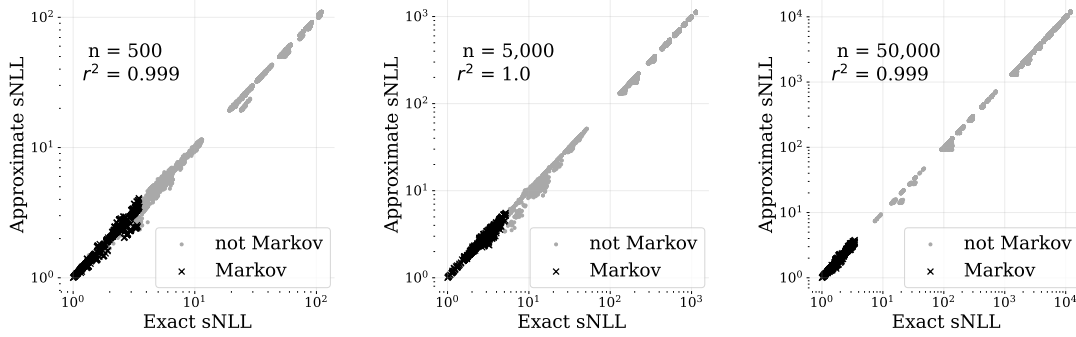
Table 6.1: Mean run time for graphs (std in parentheses) with 4 vertices in seconds with two decimal places of precision for BIC and  $\hat{\text{BIC}}$ . Statistical significance at an alpha level of 0.05 is reported as either an overline if  $\hat{\text{BIC}}$  is better or an underline if the alternative method is better.

Takes approximately 5% of the run time or two orders of magnitude. As a point of reference, the time to compute the sample covariance in these experiments generally took between 2 and 5 milliseconds.

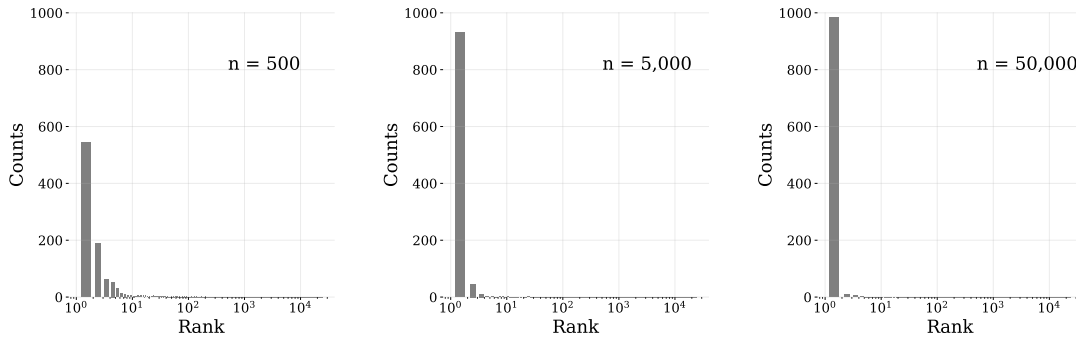
Directed MAG



Negative Log-likelihood



MEC Recovery



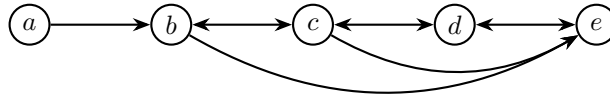
|                 | BIC  | $\hat{B}IC$ |       | FCI          |              | $FCI_{max}$  |              | GFCI <sub>1</sub> |              | GFCI <sub>2</sub> |              |
|-----------------|------|-------------|-------|--------------|--------------|--------------|--------------|-------------------|--------------|-------------------|--------------|
| repetitions     | 100  | 100         | 1,000 | 1,000        |              | 1,000        |              | 1,000             |              | 1,000             |              |
| $\alpha$ -level | -    | -           | -     | 0.01         | 0.001        | 0.01         | 0.001        | 0.01              | 0.001        | 0.01              | 0.001        |
| n = 500         | 0.49 | 0.47        | 0.544 | <u>0.836</u> | <u>0.702</u> | <u>0.499</u> | <u>0.444</u> | <u>0.223</u>      | <u>0.213</u> | <u>0.08</u>       | <u>0.08</u>  |
| n = 5,000       | 0.91 | 0.93        | 0.932 | <u>0.978</u> | <u>0.999</u> | 0.933        | <u>0.948</u> | <u>0.779</u>      | <u>0.786</u> | <u>0.575</u>      | <u>0.576</u> |
| n = 50,000      | 1.0  | 1.0         | 0.985 | <u>0.975</u> | <u>0.998</u> | <u>0.975</u> | <u>0.998</u> | 0.99              | 0.991        | 0.977             | 0.977        |

Figure 6.7: An evaluation of the approximate log-likelihood and BIC for the specified directed MAG with  $n = \{500, 5,000, 50,000\}$ . Statistical significance at an alpha level of 0.05 is reported as either an overline if  $\hat{B}IC$  is better or an underline if the alternative method is better.

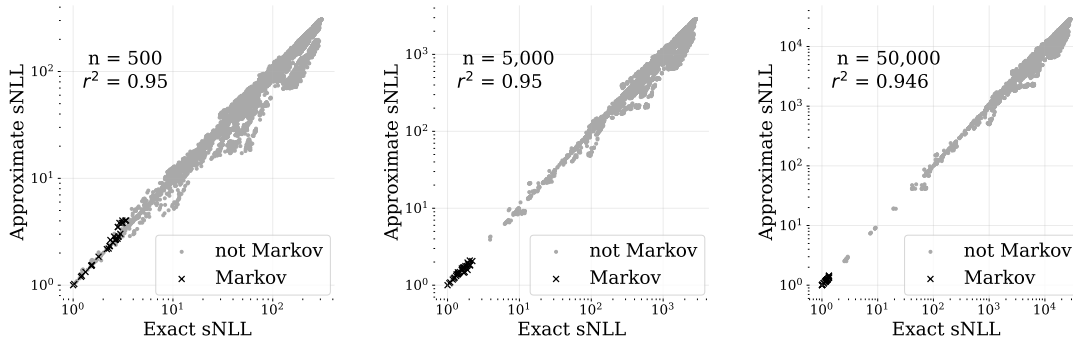
Figure 6.7 the prespecified graph is a MAG from a MEC with five vertices that does not

contain a DAG. The approximate log-likelihood closely aligns with the exact log-likelihood with clear separation of Markov versus not Markov as  $n \rightarrow \infty$ ; the approximate BIC performs nearly identically to BIC and consistently ranks the correct MEC in the top 10 with the ranking converging to a point-mass in the first bin as  $n \rightarrow \infty$ ; the top ranked approximate BIC model performs worse than FCI, about the same as  $\text{FCI}_{\max}$ , and better than GFCI with low sample sizes and about the same otherwise in MEC recovery. In this case, FCI does well because it is general biased towards bi-directed edges.

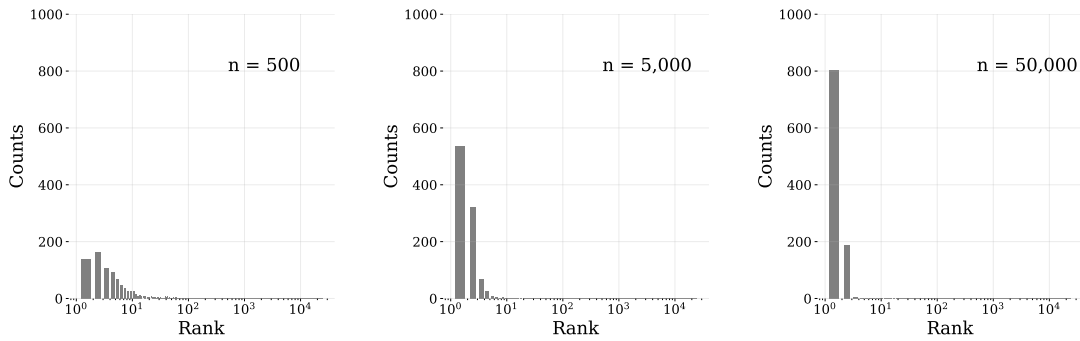
Directed MAG



Negative Log-likelihood



MEC Recovery

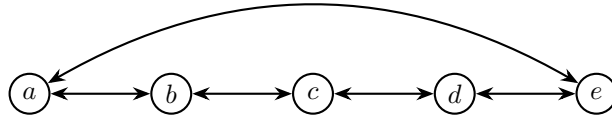


|                 | BIC  | $\hat{B}IC$ |       | FCI          |              | $FCI_{max}$  |              | GFCI <sub>1</sub> |              | GFCI <sub>2</sub> |              |
|-----------------|------|-------------|-------|--------------|--------------|--------------|--------------|-------------------|--------------|-------------------|--------------|
| repetitions     | 100  | 100         | 1,000 | 1,000        |              | 1,000        |              | 1,000             |              | 1,000             |              |
| $\alpha$ -level | -    | -           | -     | 0.01         | 0.001        | 0.01         | 0.001        | 0.01              | 0.001        | 0.01              | 0.001        |
| n = 500         | 0.12 | 0.14        | 0.138 | <u>0.006</u> | <u>0.0</u>   | <u>0.005</u> | <u>0.0</u>   | <u>0.0</u>        | <u>0.0</u>   | <u>0.0</u>        | <u>0.0</u>   |
| n = 5,000       | 0.53 | 0.54        | 0.538 | <u>0.052</u> | <u>0.028</u> | <u>0.048</u> | <u>0.028</u> | <u>0.041</u>      | <u>0.029</u> | <u>0.032</u>      | <u>0.02</u>  |
| n = 50,000      | 0.77 | 0.79        | 0.804 | <u>0.307</u> | <u>0.23</u>  | <u>0.296</u> | <u>0.221</u> | <u>0.304</u>      | <u>0.247</u> | <u>0.28</u>       | <u>0.221</u> |

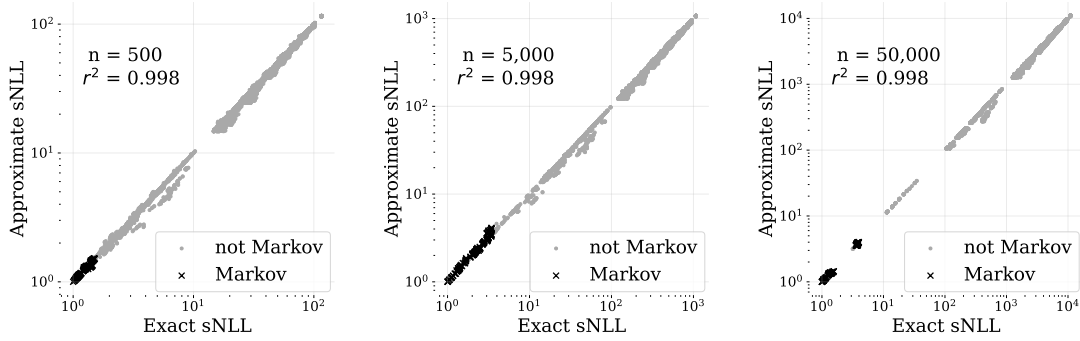
Figure 6.8: An evaluation of the approximate log-likelihood and BIC for the specified directed MAG with  $n = \{500, 5,000, 50,000\}$ . Statistical significance at an alpha level of 0.05 is reported as either an overline if  $\hat{B}IC$  is better or an underline if the alternative method is better.

Figure 6.8 the prespecified graph contains a discriminating path of length five in graphs with five vertices. The approximate log-likelihood closely aligns with the exact log-likelihood with poor separation of Markov versus not Markov, but tending towards good separation as  $n \rightarrow \infty$ ; the approximate BIC performs nearly identically to BIC and consistently ranks the correct MEC in the top 100 with the ranking converging to a point-mass in the first bin as  $n \rightarrow \infty$ ; the top ranked approximate BIC model performs better than the other methods in MEC

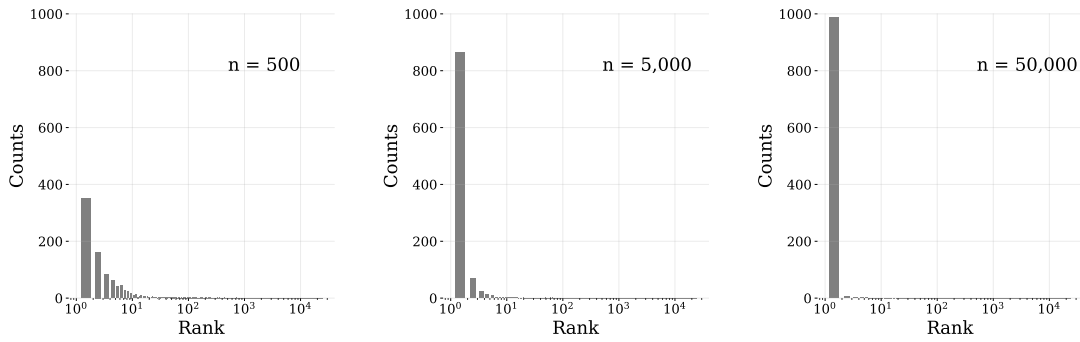
## Directed MAG



### Negative Log-likelihood



### MEC Recovery

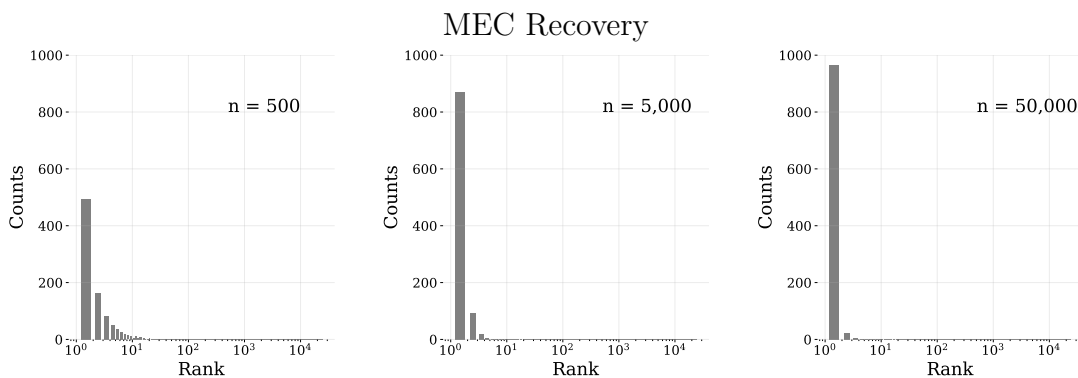


|                 | BIC  |      | $\hat{B}IC$ | FCI          |              | $FCI_{max}$  |              | GF $CI_1$    |              | GF $CI_2$    |              |
|-----------------|------|------|-------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| repetitions     | 100  | 100  | 1,000       | 1,000        |              | 1,000        |              | 1,000        |              | 1,000        |              |
| $\alpha$ -level | -    | -    | -           | 0.01         | 0.001        | 0.01         | 0.001        | 0.01         | 0.001        | 0.01         | 0.001        |
| n = 500         | 0.33 | 0.31 | 0.35        | <u>0.813</u> | <u>0.6</u>   | <u>0.326</u> | <u>0.261</u> | <u>0.03</u>  | <u>0.028</u> | <u>0.006</u> | <u>0.006</u> |
| n = 5,000       | 0.86 | 0.86 | 0.864       | <u>0.966</u> | <u>0.998</u> | 0.858        | 0.875        | <u>0.358</u> | <u>0.363</u> | <u>0.168</u> | <u>0.171</u> |
| n = 50,000      | 0.99 | 0.99 | 0.988       | <u>0.973</u> | <u>0.998</u> | <u>0.971</u> | <u>0.995</u> | <u>0.833</u> | <u>0.839</u> | <u>0.745</u> | <u>0.751</u> |

Figure 6.9: An evaluation of the approximate log-likelihood and BIC for the specified directed MAG with  $n = \{500, 5,000, 50,000\}$ . Statistical significance at an alpha level of 0.05 is reported as either an overline if  $\hat{B}IC$  is better or an underline if the alternative method is better.

Figure 6.9 the prespecified graph is a bi-directed five-cycle which perhaps explains the poor performance of GFCI. The approximate log-likelihood closely aligns with the exact log-likelihood with clear separation of Markov versus not Markov as  $n \rightarrow \infty$ ; the approximate BIC performs nearly identically to BIC and consistently ranks the correct MEC in the top 10 with the ranking converging to a point-mass in the first bin as  $n \rightarrow \infty$ ; the top ranked approximate BIC model performs worse than FCI, about the same as  $\text{FCI}_{\max}$ , and better than GFCI in MEC recovery.

Random Directed MAGs with  $|V| = 5$  and  $|E| \in [0, 5]$



|                 | BIC  | $\hat{\text{BIC}}$ |       | FCI          |              | $\text{FCI}_{\max}$ |              | GFCI <sub>1</sub> |              | GFCI <sub>2</sub> |              |
|-----------------|------|--------------------|-------|--------------|--------------|---------------------|--------------|-------------------|--------------|-------------------|--------------|
| repetitions     | 100  | 100                | 1,000 | 1,000        |              | 1,000               |              | 1,000             |              | 1,000             |              |
| $\alpha$ -level | -    | -                  | -     | 0.01         | 0.001        | 0.01                | 0.001        | 0.01              | 0.001        | 0.01              | 0.001        |
| n = 500         | 0.57 | 0.58               | 0.495 | <u>0.273</u> | <u>0.16</u>  | <u>0.398</u>        | <u>0.338</u> | <u>0.349</u>      | <u>0.324</u> | <u>0.258</u>      | <u>0.252</u> |
| n = 5,000       | 0.9  | 0.88               | 0.871 | <u>0.776</u> | <u>0.741</u> | <u>0.788</u>        | <u>0.78</u>  | <u>0.741</u>      | <u>0.732</u> | <u>0.676</u>      | <u>0.671</u> |
| n = 50,000      | 0.99 | 0.99               | 0.966 | <u>0.919</u> | <u>0.915</u> | <u>0.92</u>         | <u>0.919</u> | <u>0.904</u>      | <u>0.9</u>   | <u>0.893</u>      | <u>0.888</u> |

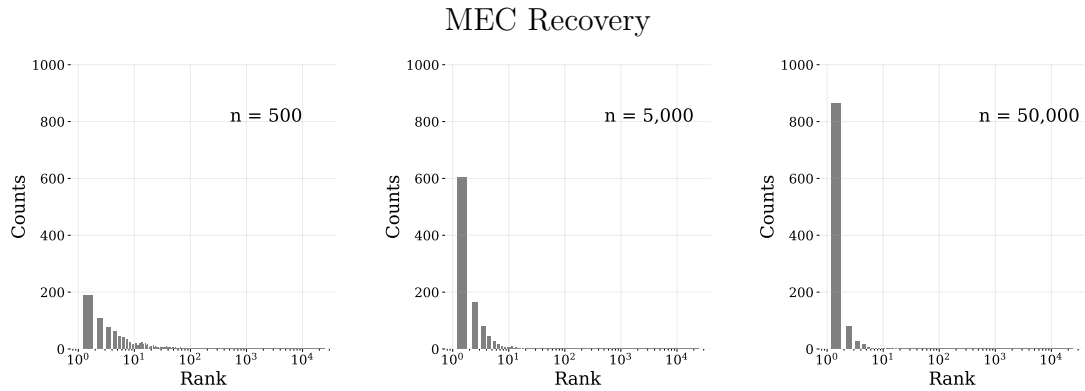
Figure 6.10: An evaluation of the approximate BIC for random directed MAGs with specified edge ranges and  $n = \{500, 5,000, 50,000\}$ . Statistical significance at an alpha level of 0.05 is reported as either an overline if  $\hat{\text{BIC}}$  is better or an underline if the alternative method is better.

Figure 6.10 the approximate BIC performs nearly identically to BIC and consistently ranks the correct MEC in the top 10 with the ranking converging to a point-mass in the



first bin as  $n \rightarrow \infty$ ; the top ranked approximate BIC model performs better than the other methods in MEC recovery.

Random Directed MAGs with  $|V| = 5$  and  $|E| \in [6, 10]$



|                 | BIC  | $\hat{BIC}$ |       | FCI          |              | FCI <sub>max</sub> |              | GF <sub>1</sub> |              | GF <sub>2</sub> |              |
|-----------------|------|-------------|-------|--------------|--------------|--------------------|--------------|-----------------|--------------|-----------------|--------------|
| repetitions     | 100  | 100         | 1,000 | 1,000        |              | 1,000              |              | 1,000           |              | 1,000           |              |
| $\alpha$ -level | -    | -           | -     | 0.01         | 0.001        | 0.01               | 0.001        | 0.01            | 0.001        | 0.01            | 0.001        |
| n = 500         | 0.24 | 0.24        | 0.188 | <u>0.026</u> | <u>0.009</u> | <u>0.072</u>       | <u>0.031</u> | <u>0.035</u>    | <u>0.021</u> | <u>0.022</u>    | <u>0.019</u> |
| n = 5,000       | 0.61 | 0.59        | 0.603 | <u>0.311</u> | <u>0.243</u> | <u>0.354</u>       | <u>0.319</u> | <u>0.297</u>    | <u>0.261</u> | <u>0.238</u>    | <u>0.216</u> |
| n = 50,000      | 0.9  | 0.91        | 0.865 | <u>0.665</u> | <u>0.617</u> | <u>0.687</u>       | <u>0.671</u> | <u>0.641</u>    | <u>0.622</u> | <u>0.595</u>    | <u>0.579</u> |

Figure 6.11: An evaluation of the approximate BIC for random directed MAGs with specified edge ranges and  $n = \{500, 5,000, 50,000\}$ . Statistical significance at an alpha level of 0.05 is reported as either an overline if  $\hat{BIC}$  is better or an underline if the alternative method is better.

Figure 6.11 the approximate BIC performs nearly identically to BIC and consistently ranks the correct MEC in the top 100 with the ranking converging to a point-mass in the first bin as  $n \rightarrow \infty$ ; the top ranked approximate BIC model performs better than the other methods in MEC recovery.

| sample size | BIC         |             |             | $\hat{\text{BIC}}$ |             |             |
|-------------|-------------|-------------|-------------|--------------------|-------------|-------------|
|             | n = 500     | n = 5,000   | n = 50,000  | n = 500            | n = 5,000   | n = 50,000  |
| Figure 6.7  | 74.6 (5.8)  | 72.8 (5.7)  | 71.9 (5.8)  | 0.64 (0.02)        | 0.63 (0.01) | 0.63 (0.01) |
| Figure 6.8  | 81.3 (8.5)  | 80.9 (8.1)  | 80.8 (8.4)  | 0.64 (0.06)        | 0.63 (0.06) | 0.63 (0.07) |
| Figure 6.9  | 72.1 (5.0)  | 71.1 (4.5)  | 70.7 (5.1)  | 0.67 (0.07)        | 0.64 (0.04) | 0.65 (0.06) |
| Figure 6.10 | 77.7 (7.5)  | 76.5 (7.8)  | 75.5 (7.7)  | 0.32 (0.1)         | 0.29 (0.07) | 0.28 (0.03) |
| Figure 6.11 | 80.9 (10.5) | 80.3 (10.1) | 80.3 (10.4) | 0.26 (0.01)        | 0.26 (0.01) | 0.26 (0.01) |

Table 6.2: Mean run time for graphs (std in parentheses) with 5 vertices in seconds with one decimal place of precision for BIC and two decimal places of precision for  $\hat{\text{BIC}}$  (100 reps).

Takes approximately 1% of the run time or two orders of magnitude. The time of covariance calculation generally took between 2 and 5 milliseconds.

Overall, the top ranked approximate BIC model performs best or second best and generally performs better with low samples sizes and more complicated graphs.

## 6.2 Ancestral Probabilities

In this section, we formulate and analyze a Bayesian local causal discovery algorithm. The algorithm computes the probabilities of ancestral relationships between pairs of variables among a local subset of variables. We call this algorithm the Ancestral Probabilities (AP) procedure. The AP procedure is motivated by the following situation. Suppose we have a data set  $x^1, \dots, x^n$  that contains variables  $a$  and  $b$ . We might ask: “what is the probability that  $a$  causes  $b$ ?” The AP procedure computes this probability, denoted  $a \rightarrow b$ ,

by marginalizing over all possible directed MAGS:

$$\begin{aligned}
\Pr(a \rightarrow b \mid x^1, \dots, x^n) &= \sum_{\mathcal{G}} \Pr(\mathcal{G} \mid x^1, \dots, x^n) \delta_{\text{an}_{\mathcal{G}}(b)}(a) \\
&= \sum_{\mathcal{G}} \frac{\Pr(x^1, \dots, x^n \mid \mathcal{G}) \Pr(\mathcal{G})}{\sum_{\mathcal{G}'} \Pr(x^1, \dots, x^n \mid \mathcal{G}') \Pr(\mathcal{G}')} \delta_{\text{an}_{\mathcal{G}}(b)}(a) \\
&= \frac{\sum_{\mathcal{G}} \Pr(x^1, \dots, x^n \mid \mathcal{G}) \Pr(\mathcal{G}) \delta_{\text{an}_{\mathcal{G}}(b)}(a)}{\sum_{\mathcal{G}'} \Pr(x^1, \dots, x^n \mid \mathcal{G}') \Pr(\mathcal{G}')}.
\end{aligned}$$

This formulation for the marginal probability of causal relationship first occurred in [52].

This algorithm is intractable for systems with more than five variables, however, because directed MAGs are closed under marginalization, we may calculate the probability that  $a$  causes  $b$  with any subset of the variables containing  $a$  and  $b$ . As an example, consider the trivial case where the subset of chosen variables is  $\{a, b\}$ . In this case, the algorithm would only have two calculations to make,  $a$  adjacent to  $b$  and  $a$  not adjacent to  $b$ . While we have accomplished a tremendous gain in computational efficiency, we have traded too much and the output is less interesting. Thus, this algorithm must balance computational efficiency with information loss.

We assume that the data for the local subset of variables is distributed according to a curved exponential family. Furthermore, we assume a uniform prior over MECs and directed MAGs within each MEC. We make the following standard assumptions.

**Assumption** (Causal Markov Property). If  $\mathcal{G} = (V, E)$  is the causal MAG for a collection of random variables indexed by  $V$  with probability measure  $P$ , then

$$\mathcal{J}(\mathcal{G}) \subseteq \mathcal{J}(P).$$

**Assumption** (Causal Faithfulness Property). If  $\mathcal{G} = (V, E)$  is the causal MAG for a collection of random variables indexed by  $V$  with probability measure  $P$ , then

$$\mathcal{J}(P) \subseteq \mathcal{J}(\mathcal{G}).$$

The AP procedure is then defined as follows:

---

**Algorithm 7: ANCESTRAL PROBABILITIES AP**  $(x^1, \dots, x^n, \Pr(\mathcal{G}))$ 

---

**Input:** data:  $x^1, \dots, x^n$ , prior:  $\Pr(\mathcal{G})$   
**Output:** mapping: Probs

```
1 Probs = {} ;
2 foreach  $a, b \in V$  do
3   | Probs[ $a, b$ ] = 0 ;
4 end
5 norm = 0 ;
6 foreach  $\mathcal{G} \in \mathcal{F}_{\text{DMAG}}(V)$  do
7   | Pick a total order  $\leq$  consistent with  $\mathcal{G}$  ;
8   | foreach  $a, b \in V$  do
9     | if  $a \in \text{ang}(b)$  then
10    |   | Probs[ $a, b$ ] = Probs[ $a, b$ ] +  $\exp\{\hat{\text{BIC}}(\mathcal{G}, \leq, x^1, \dots, x^n)\} \Pr(\mathcal{G})$  ;
11    |   end
12    end
13   | norm = norm +  $\exp\{\hat{\text{BIC}}(\mathcal{G}, \leq, x^1, \dots, x^n)\} \Pr(\mathcal{G})$  ;
14 end
15 foreach  $a, b \in V$  do
16   | Probs[ $a, b$ ] = Probs[ $a, b$ ]/norm ;
17 end
```

---

Additional constraints, such as the CLHMC must be applied by the user as background knowledge.

### 6.2.1 Synthetic Examples and Background Knowledge

In this section, we evaluate the effectiveness of the AP procedure with and without the background knowledge that one variable is exogenous with respect to the other variables on 1,000 synthetic data sets of 500, 5,000 and 50,000 instances. The synthetic data sets are generated by passing a directed MAG drawn uniformly from the set of all directed MAGs with four vertices to Algorithm 6 which uses an implicit order over the variables during simulation. The background knowledge is generated by noting that the first variables in the implicit order is exogenous with respect to the other variables. However, the background knowledge is not guaranteed to be helpful. Indeed, it could be the case that the designated exogenous variables are disconnected from the other variables. In this case, while the background knowledge is

correct, it is not helpful in refined the data generating MEC. See [3] for details on how background knowledge refines a MEC.

In what follows, we evaluate the correctness of the AP procedure with and without knowledge using precision recall curves and receiver operator curves. We also report the area under these curves respectively. Notably, given the true Markov equivalence class, the probability of getting an ancestral relation correct could be less than 0.5.

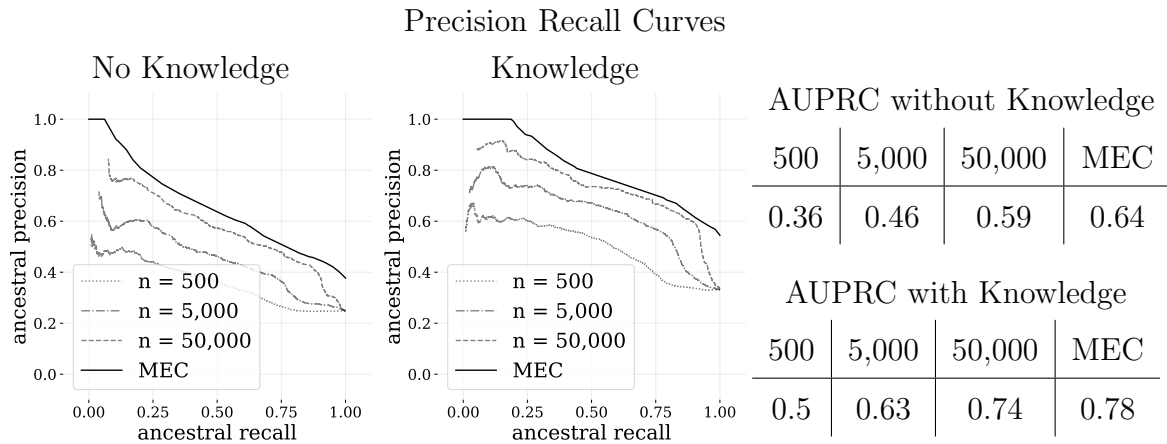


Figure 6.12: Precision recall curves for ancestral relationships with and without background knowledge.

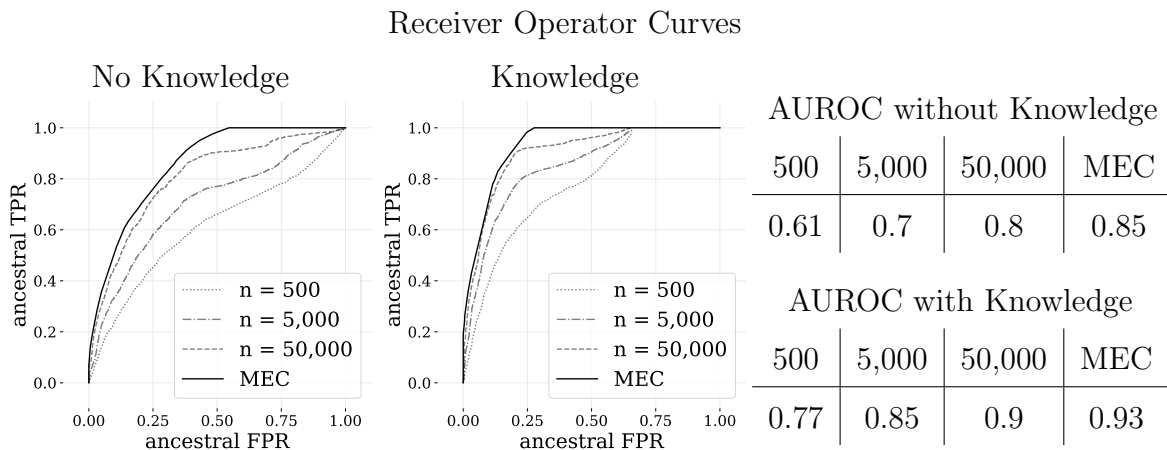


Figure 6.13: Receiver operator curves for ancestral relationships with and without background knowledge.

Figures 6.12 and 6.13 plot the precision recall and receiver operator curves respectively. In addition, the areas under the curves are tabulated. We plot the curves for  $n = \{500, 5,000, 50,000\}$ . Additionally, we plot a curve where every directed MAG in the true MEC is given equal probability as a theoretical limit on performance. We observe that as the sample size increase, we approach this theoretical limit. This behavior persists with the incorporation of background knowledge; however, there is a positive shift in the overall performance with the incorporation of background knowledge.

In what follows, we evaluate the calibration of the AP procedure with and without knowledge using calibration curves, also known as reliability diagrams [19, 55]. We also report the expected calibration error [54].

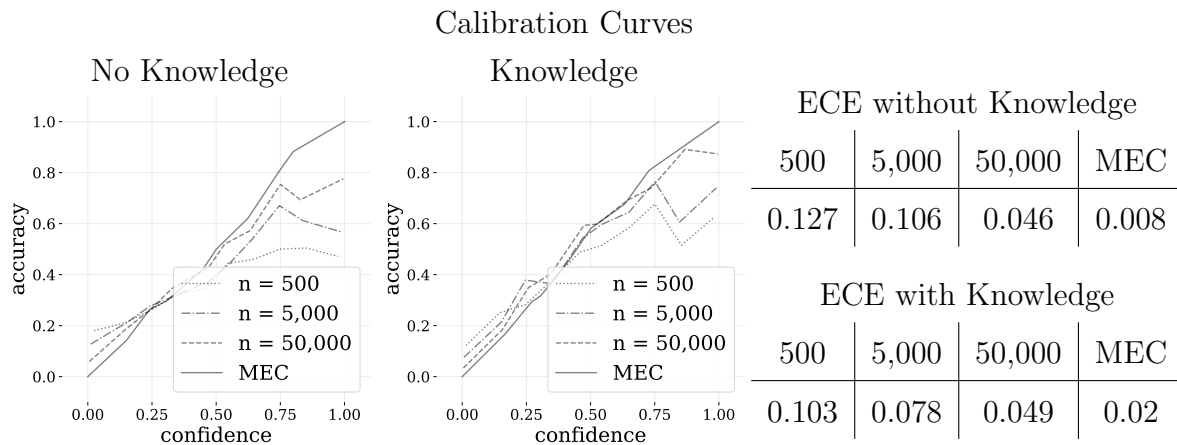


Figure 6.14: Calibration curves for ancestral relationships with and without background knowledge.

Figure 6.14 plots the calibration the AP procedure for  $n = \{500, 5,000, 50,000\}$ ; more detailed figures are given in Appendix B.8. The calibration of a probabilistic algorithm evaluate the algorithm’s ability to output meaningful probabilities. In this case, we plot the reliability curves for the ancestral probabilities output by the AP procedure. The MEC report the theoretical limit, that is, the theoretical performance given infinite data. We observe that the predicted probabilities generally underestimate the true probabilities, but as  $n \rightarrow \infty$  the predicted and true probabilities tend towards a closer correspondence. This

behavior persists with the incorporation of background knowledge.

Background knowledge helps with reliability and calibration in low sample sizes. Note that the prior used in the algorithm over directed MAGs in a MECs matches the corresponding prior used to generate the data. As long as these priors are consistent, we can expect to achieve performance similar to what was achieved in this simulation study. Notably, we do not have to assume that every directed MAG within a MEC is equally likely.

### **6.2.2 Airborne Pollutants' Effect on Health**

In this section we evaluate the AP procedure on a real data set measuring airborne pollutants, cardiovascular health, and respiratory health. We joined local air composition data from the Environmental Protection Agency (EPA) with clinical data from the University of Pittsburgh Medical Center (UPMC) at the zip code-month level.

Measurements for 160 airborne pollutants were collected from air-monitoring stations in greater Pittsburgh area in 2015. These measurements were used to curate a data set of monthly zip code averages for the airborne pollutants—the measurements of each pollutant from stations within a 10-kilometer radius of the geographical center of each zip code were averaged over the course of a month. We selected a subset of 19 pollutants to analyze based on data sufficiency and a brief literature review. These data were obtained from Drs. Chirag Patel and Chirag Lakhani.

Positive and negative cases of cardiovascular disease and respiratory disease were collected from UPMC hospitals and outpatient facilities using the ICD9/ICD10 codes from patient visits in 2015. These cases were used to curate a data set of monthly zip code averages for the occurrence of the two diseases—the presence of each disease during patient visits for patients living within each zip code were averaged over the course of a month. There were a total of 2,068,999 patient-visit records in 2015, of which 500,448 were positive for cardiovascular disease and 262,221 were positive for respiratory disease.

After joining the two data sets, the resulting data set contained variables for zip code, month, cardiovascular disease prevalence, respiratory disease prevalence, and 19 airborne pollutant averages. We applied the AP procedure to this data in order to evaluate our

algorithm and investigate the effects of airborne pollutants on cardiovascular health and respiratory health. This research was performed under the auspices of Study PRO18020279, which was approved by the University of Pittsburgh Institutional Review Board; all data were de-identified.

Limitations of this study include:

- only analyzed data from the greater Pittsburgh area in 2015;
- only analyzed data from the UPMC hospital system;
- utilized an overly broad categorization of disease;
- utilized a temporal granularity based on the month.

In the following analysis, we modeled the data using Lee and Hastie probability measures. Notably, we did not check the appropriateness of these probability measures. However, we did apply the background knowledge that month is exogenous with respect to the other variables—month is not caused by or confounded with the other variables [3]. Since month is the only discrete variable, the background knowledge entails the CLHMC. Accordingly, the graphical Markov models considered are members of curved exponential families.

Additional data wrangling was performed as follows. If the number of patients visits used to calculate disease prevalence for an instance was less than 30, the record was removed from the analysis because the estimated may be unreliable. The values for zip code and month uniquely identify an instance of the data set because the data were only collected for one year. We removed zip code from the analysis to remove the possibility of determinism. We chose to remove zip code over month because the data were all collected from the greater Pittsburgh area, so we surmised that a temporal factor would have a more significant impact on the other variables than a spatial factor.

Using the AP procedure described above with the degenerate Gaussian approximation of the Lee and Hastie model [2], we independently analyzed the data for each pollutant against the variables for month, cardiovascular disease prevalence, and respiratory disease prevalence. In addition to the background knowledge that zip code and month are exogenous with respect to the other variables, we applied the background knowledge that airborne pollution is not cause by disease prevalence. During the analysis, we performed bootstrapping at the



patient-visit level. Notably, the instances of the data set are at the zip code level and the airborne pollutant averages are unrelated to patient visits. In practice, we recalculated the prevalence of cardiovascular disease and respiratory disease by resampling the positive and negative cases of the diseases for each zip code and month. The data set was bootstrapped accordingly 1000 times in order to provide confidence intervals. The original data set was included in this analysis.

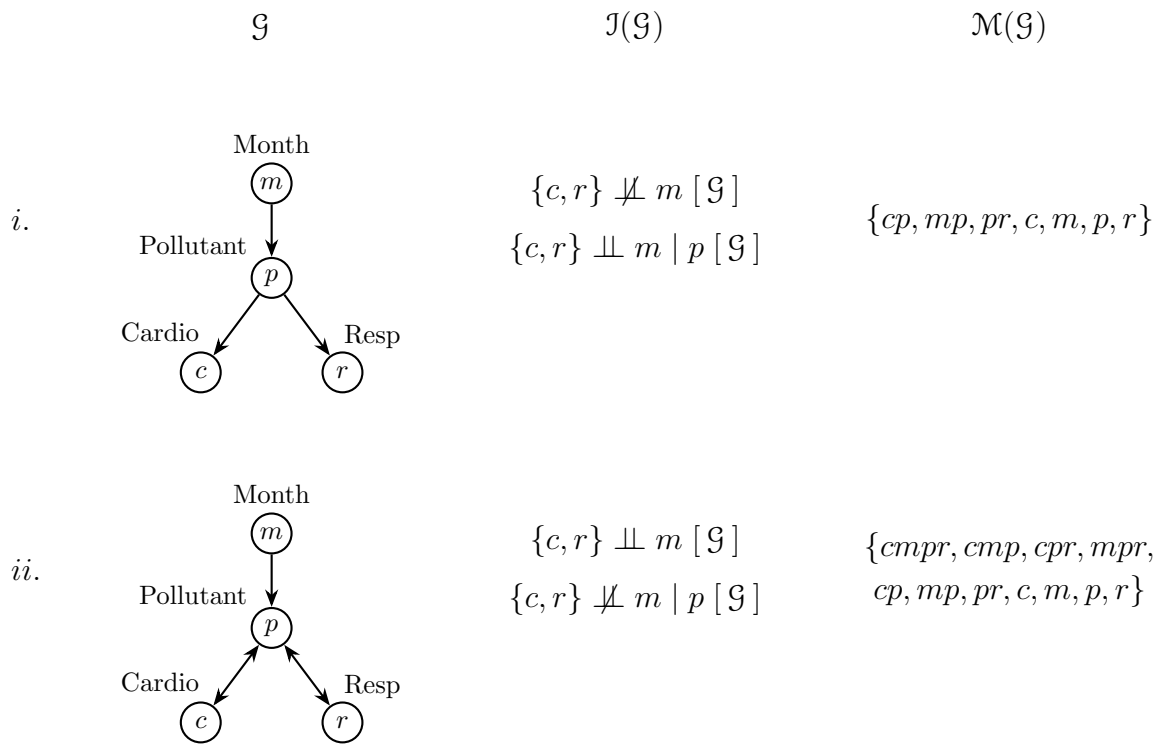


Figure 6.15: A comparison of two hypotheses for the underlying causal model: (i) the airborne pollutant is a cause of cardiovascular disease and respiratory disease; (ii) the airborne pollutant is confounded with cardiovascular disease and respiratory disease.

Figure 6.15 gives an intuition for how the AP procedure can recover causal information from the data. The figure compares two hypotheses for the underlying causal model: (i) the airborne pollutant is a cause of cardiovascular disease and respiratory disease; (ii) the airborne pollutant is confounded with cardiovascular disease and respiratory disease. A difference in represented conditional independence statements and  $m$ -connecting sets is illus-

trated. The difference in represented conditional independence statements is accounted for by the additional  $m$ -connecting sets. Accordingly, are the relationships causal or confounded boils down to if the inclusion of these additional  $m$ -connecting sets increase the approximate log-likelihood.

With two exceptions, the airborne pollutants examined are regulated by the National Ambient Air Quality Standards (NAAQS) or the National Emission Standards for Hazardous Air Pollutants (NESHAP). These standards were set by the clean air act.

With one exception, the EPA provides a comprehensive assessment of the causal relationships between the NAAQS airborne pollutants, cardiovascular disease, and respiratory disease [23, 24, 25, 26, 27, 28]. Notably, the EPA's analysis of Lead does not provide an assessment of causality for respiratory disease, so we postpone its analysis. In Tables 6.3 and 6.4, we tabulate the results of the analysis on the NAAQS airborne pollutants. In the tables, mean probabilities and 95% bootstrap confidence intervals are reported for each causal relationship along with the EPA's assessment of evidence for a short-term causal relationship using the following scale:

- 1: causal relationship;
- 2: likely to be a causal relationship;
- 3: suggestive of, but not sufficient to infer, a causal relationship;
- 4: inadequate to infer a causal relationship;
- 5: not likely to be a causal relationship.

| Air Pollutant                        | Pollutant → Cardiovascular | EPA Assessment of Causality   |
|--------------------------------------|----------------------------|-------------------------------|
| Carbon Monoxide                      | 0.931 (0.65, 1.0)          | 2: likely to be causal        |
| Nitric Oxide                         | 0.098 (0.0, 0.55)          | 3: suggestive of being causal |
| Nitrogen Dioxide                     | 0.021 (0.0, 0.14)          | 3: suggestive of being causal |
| Oxides of Nitrogen                   | 0.028 (0.0, 0.21)          | 3: suggestive of being causal |
| Ozone                                | 0.87 (0.45, 1.0)           | 3: suggestive of being causal |
| Particulate Matter 2.5 $\mu\text{m}$ | 0.057 (0.0, 0.31)          | 1: causal                     |
| Particulate Matter 10 $\mu\text{m}$  | 0.392 (0.08, 0.86)         | 3: suggestive of being causal |
| Sulfur Dioxide                       | 0.101 (0.0, 0.35)          | 4: inadequate to infer causal |

Table 6.3: NAAQS airborne pollutants and cardiovascular disease results.

| Air Pollutant                        | Pollutant → Respiratory | EPA Assessment of Causality   |
|--------------------------------------|-------------------------|-------------------------------|
| Carbon Monoxide                      | 0.041 (0.0,0.26)        | 3: suggestive of being causal |
| Nitric Oxide                         | 0.984 (0.88, 1.0)       | 1: causal                     |
| Nitrogen Dioxide                     | 0.997 (0.98, 1.0)       | 1: causal                     |
| Oxides of Nitrogen                   | 0.996 (0.98, 1.0)       | 1: causal                     |
| Ozone                                | 0.558 (0.27, 0.81)      | 1: causal                     |
| Particulate Matter 2.5 $\mu\text{m}$ | 0.002 (0.0, 0.02)       | 2: likely to be causal        |
| Particulate Matter 10 $\mu\text{m}$  | 0.187 (0.01, 0.57)      | 3: suggestive of being causal |
| Sulfur Dioxide                       | 0.592 (0.15, 0.98)      | 1: causal                     |

Table 6.4: NAAQS airborne pollutants and respiratory disease results.

The results of the causal analysis agree with many of the results of the EPA assessments of short-term causality for the NAAQS airborne pollutants. In general, the AP procedure assigned high confidence to causal relationships that the EPA assessed to be 1 or 2 and low confidence to causal relationships that the EPA assessed to be 3 or 4 with a few notable exceptions. The EPA assesses ozone as suggestive of being a cause of cardiovascular disease,

whereas our analysis yielded a high probability of 0.87 (0.45 to 1.0). These results are consistent with the EPA’s assessment for ozone, but are more confident than expected. The EPA assesses ozone and sulfur dioxide as being a cause of respiratory disease, whereas our analysis yielded modest probabilities of 0.558 (0.27 to 0.81) and 0.592 (0.15 to 0.98) respectively. These results are consistent with the EPA’s assessments of ozone and sulfur dioxide, but lack confidence. This is consistent with the earlier simulated results. The EPA assesses particulate matter 2.5  $\mu\text{m}$  as being a cause of cardiovascular disease and likely of being a cause of respiratory disease, whereas our analysis yielded near zero probabilities 0.057 (0.0, 0.31) and 0.002 (0.0 to 0.02) respectively. These results are inconsistent with the EPA’s assessments of particulate matter 2.5  $\mu\text{m}$  and suggest a confounded relationship; see Appendix B.9.

The National Emission Standards for Hazardous Air Pollutants (NESHAP) are emission standards set by the EPA for airborne pollutants associated with an increase in serious illness or death.

| Air Pollutant          | Poll $\rightarrow$ Cardio | Poll $\rightarrow$ Resp | Poll $\leftrightarrow$ Cardio | Poll $\leftrightarrow$ Resp |
|------------------------|---------------------------|-------------------------|-------------------------------|-----------------------------|
| Acrolein               | 0.0 (0.0, 0.01)           | 0.96 (0.56, 1.0)        | 1.0 (0.99, 1.0)               | 0.01 (0.0, 0.05)            |
| Arsenic                | 0.09 (0.0, 0.4)           | 0.11 (0.0, 0.36)        | 0.65 (0.04, 1.0)              | 0.39 (0.02, 0.99)           |
| Cadmium                | 0.29 (0.02, 0.77)         | 0.2 (0.04, 0.44)        | 0.34 (0.03, 0.94)             | 0.31 (0.03, 0.88)           |
| Chromium               | 0.23 (0.09, 0.41)         | 0.31 (0.11, 0.72)       | 0.44 (0.13, 0.87)             | 0.2 (0.05, 0.66)            |
| Lead PM <sub>10</sub>  | 0.1 (0.0, 0.63)           | 0.52 (0.15, 0.92)       | 0.9 (0.36, 1.0)               | 0.18 (0.02, 0.72)           |
| Lead PM <sub>2.5</sub> | 0.08 (0.0, 0.35)          | 0.12 (0.0, 0.38)        | 0.74 (0.06, 1.0)              | 0.29 (0.02, 0.98)           |
| Manganese              | 0.25 (0.01, 0.72)         | 0.27 (0.04, 0.68)       | 0.39 (0.06, 0.96)             | 0.27 (0.03, 0.87)           |
| Nickle                 | 0.21 (0.01, 0.44)         | 0.54 (0.15, 0.93)       | 0.44 (0.09, 0.96)             | 0.23 (0.03, 0.78)           |

Table 6.5: NESHAP Airborne Pollutants

The analyses for airborne acrolein, lead, and nickle supports causal relationships with respiratory disease. The analyses for airborne acrolein, arsenic, and lead supported supports confounded relationships with cardiovascular disease. The analyses for airborne cadmium, chromium, and manganese were inconclusive. For the most part, our analysis is consistent

with these listings, however, in other reports the EPA classifies Lead as a cause of cardiovascular disease. Furthermore, in the inconclusive case, the AP procedure could still find a result supporting of these listing given more data or covariates.

| Air Pollutant          | Poll $\rightarrow$ Cardio | Poll $\rightarrow$ Resp | Poll $\leftrightarrow$ Cardio | Poll $\leftrightarrow$ Resp |
|------------------------|---------------------------|-------------------------|-------------------------------|-----------------------------|
| Acetone                | 0.15 (0.03, 0.27)         | 0.38 (0.17, 0.5)        | 0.06 (0.0, 0.24)              | 0.37 (0.12, 0.5)            |
| Methyl Ethyl Ketone    | 0.33 (0.16, 0.46)         | 0.37 (0.17, 0.49)       | 0.05 (0.0, 0.19)              | 0.33 (0.09, 0.49)           |
| Methyl Isobutyl Ketone | 0.01 (0.0, 0.07)          | 0.15 (0.05, 0.4)        | 0.98 (0.83, 1.0)              | 0.03 (0.01, 0.16)           |

Table 6.6: Exceptions to NESHAP Airborne Pollutants

The analysis for airborne methyl isobutyl ketone supports a confounded relationship with cardiovascular disease, while the analyses for airborne acetone and methyl ethyl ketone were inconclusive. Furthermore, methyl isobutyl ketone is under review for delisting, while acetone and methyl ethyl ketone are currently not listed. Our analysis is consistent with these non-listings.

## 7.0 Discussion and Future Work

This dissertation introduces inducing sets as a new perspective for reasoning about ancestral graph Markov models. In particular, we derive and study  $m$ -connecting sets, which are a special case of inducing sets that provide an alternative representation for MAGs. We show that  $m$ -connecting sets admit a characterization of Markov equivalence for MAGs and a factorization criterion equivalent to the global Markov property for directed MAGs.

Using the factorization criterion, we formulate a consistent probabilistic score with a closed-form for exponential families whose independence models are described by directed MAGs—directed ancestral graph Markov models. Ultimately, we design a local causal discovery algorithm called the ancestral probability (AP) procedure, which estimates the posterior probabilities of ancestral relationships. An analysis of synthetically generated data and a real data set measuring airborne pollutants, cardiovascular health, and respiratory health suggests that score-based causal discovery can be an effective tool for real-world problems. The code for running the AP algorithm is publicly available on GitHub: <https://github.com/bja43/agMm>.

### 7.1 Discussion

This work lies at the intersection of three other bodies of work: Sadeghi and Lauritzen's work on stable mixed graphs, Studený's imsets, Richardson and Evans' work on ADMGs. In general, inducing sets provide a set-based framework for reasoning about models formed from marginalization and conditioning of DAG models. These are stable mixed graphs and their equivalence classes may be represented by  $m$ -connecting sets. Using  $m$ -connecting sets, we can derive structural imsets, a most general framework for representing independence models, that represent the same set of conditional independence statements. With these structural imsets, we formulate a factorization similar to Richardson and Evans—which lead to the development of parametrizing sets—but which only requires a single equation to be

equivalent to the global Markov property.

## 7.2 Future Work

There are many promising future research direction for inducing sets. A few that we hope to explore are as follows:

- Develop a more comprehensive theory about  $m$ -connecting sets. How can we characterize the conditional independent statements represented in a set of  $m$ -connecting sets directly from the set of sets?
- There is known redundancy in  $m$ -connecting sets, that is, we only need a subset of the  $m$ -connecting sets to fully characterize the independence model induced by a MAG. Can we develop a set of logical implications to derive the full set of  $m$ -connecting sets from the minimally sufficient subset of  $m$ -connecting sets?
- Extend the concept of inducing path to the fixing operation of Richardson et al. in order to define inducing sets with respect to a set of latent confounding variables, a set of selection variables, and a set of fixed variables. This has ramifications in intervention effect estimation, as well as extending these methods to nested Markov models [69].
- Derive a new adjustment term for the undirected part of the MAGs to extend our factorization results to general MAGs.
- The adjustment terms provide a list of parametric constraints and naturally fit into the framework of Lagrange multipliers. Can we derive an algorithm to fit the maximum likelihood estimate using Lagrange multipliers?
- Can we develop a branch and bound algorithm to extend the AP procedure to larger MAGs by considering MAGs within a factor of optimal [48]?
- Can we use  $m$ -connecting sets to generate a non-parametric score [39] for MAGs?
- Can we use  $m$ -connecting sets to construct a greedy algorithm for MEC recovery [13]? Additionally, can we extend Meek's conjecture (proved by Chickering) to MAGs?

## Appendix A List of Notation

### A.1 General Terms

- $\Rightarrow, \Leftarrow, \Leftrightarrow$  symbols for logical implication
- $\in, \notin$  symbols for set inclusion
- $\subseteq, \subset, \not\subseteq, \not\subset$  symbols for subset
- $\cup, \cap$  symbols for set union and intersection
- $\setminus$  symbol for set difference
- $\top$  symbol for matrix transpose
- $||$  symbol for set cardinality and absolute value
- $\langle \rangle$  syntax for a sequence or list
- $\{ \}$  syntax for a set
- $\text{vec}$  function for matrix vectorization

### A.2 Sets of Numbers

- $\mathbb{R}$  the set of real numbers
- $\mathbb{Q}$  the set of rational numbers
- $\mathbb{Q}_+$  the set of non-negative rational numbers
- $\mathbb{Z}$  the set of integers
- $\mathbb{Z}_+$  the set of non-negative integers
- $\mathbb{S}_{++}^n$  the set of  $n \times n$  symmetric positive definite matrices

### A.3 General Sets

- $A, B, C, \dots$  symbols for sets of sets

13



|                                  |  |    |
|----------------------------------|--|----|
| • $A, B, C, \dots$               | symbols for sets                       | 13 |
| • $a, b, c, \dots$               | symbols for set elements or singletons | 13 |
| • $\emptyset$                    | symbol for empty set                   | 13 |
| • $\mathcal{P}, \mathcal{P}_i^u$ | function for (bounded) power set       | 13 |

#### A.4 Generic Set Symbols

|            |   |
|------------|---|
| • $V$      | generic symbol for a non-empty set of variables/vertices            |
| • $L$      | generic symbol for a set of latent confounding variables/vertices   |
| • $S$      | generic symbol for a set of latent selection variables/vertices     |
| • $M$      | generic symbol for an $m$ -connecting set of variables/vertices     |
| • $N$      | generic symbol for a non- $m$ -connecting set of variables/vertices |
| • $\Gamma$ | generic symbol for the continuous variables/vertices                |
| • $\Delta$ | generic symbol for the discrete variables/vertices                  |
| • $W$      | generic symbol for the transformed variables/vertices               |

#### A.5 Probability Measures

|                 |   |    |
|-----------------|---|----|
| • $X$           | generic symbol for a collection of random variables     | 13 |
| • $\mathcal{X}$ | generic symbol for a sample space                       | 13 |
| • $x$           | generic symbol for a fixed instance of the sample space | 13 |
| • $f$           | symbol for probability density                          | 13 |
| • $f_A$         | symbol for marginal probability density                 | 13 |
| • $f_{A B}$     | symbol for conditional probability density              | 13 |
| • $P$           | symbol for probability measure                          | 15 |
| • $\nu$         | symbol for dominating measure                           | 15 |
| • $m_P$         | symbol for multiinformation of $P$                      | 56 |

## A.6 Independence Models

|                         |                                  |    |
|-------------------------|----------------------------------|----|
| • $\langle ,   \rangle$ | syntax for a disjoint triple     | 14 |
| • $\perp\!\!\!\perp$    | symbol for independence          | 14 |
| • $\mathcal{T}$         | function for disjoint triples    | 14 |
| • $\mathcal{I}$         | function for independence models | 14 |

## A.7 Partially Ordered Sets

|                        |                               |    |
|------------------------|-------------------------------|----|
| • $\leq$               | symbol for partial order      | 16 |
| • $\mathsf{P}$         | generic symbol for a poset    | 16 |
| • $\vee \wedge$        | symbols for set join and meet | 17 |
| • $\zeta_{\mathsf{P}}$ | symbol for zeta function      | 18 |
| • $\mu_{\mathsf{P}}$   | symbol for Möbius function    | 18 |

## A.8 General Graph Terms

|                   |  |     |
|-------------------|--|-----|
| • $\mathcal{G}$   | symbol for a mixed graph                                     | 22  |
| • $\pi$           | generic symbol for a path                                    | 22  |
| • $\mathcal{G}_A$ | symbol for an induced subgraph with respect to $A$           | 27  |
| • $\text{dg}$     | function for directed subgraph                               | 27  |
| • $\text{ug}$     | function for undirected subgraph                             | 27  |
| • $[\mathcal{G}]$ | function for a maximally informative partial ancestral graph | 42  |
| • $\text{dom}$    | function for dominating DAG                                  | 123 |

## A.9 Functions of Vertices

|   |  |    |
|---|--|----|
| • $\text{pa}_{\mathcal{G}}, \text{pa}_{\mathcal{G}}^+$      | function for (inclusive) parents in $\mathcal{G}$                        | 24 |
| • $\text{ch}_{\mathcal{G}}, \text{ch}_{\mathcal{G}}^+$      | function for (inclusive) children in $\mathcal{G}$                       | 24 |
| • $\text{sp}_{\mathcal{G}}, \text{sp}_{\mathcal{G}}^+$      | function for (inclusive) spouses in $\mathcal{G}$                        | 24 |
| • $\text{ne}_{\mathcal{G}}, \text{ne}_{\mathcal{G}}^+$      | function for (inclusive) neighbors in $\mathcal{G}$                      | 24 |
| • $\text{an}_{\mathcal{G}}$                                 | function for ancestors in $\mathcal{G}$                                  | 25 |
| • $\text{de}_{\mathcal{G}}$                                 | function for descendants in $\mathcal{G}$                                | 25 |
| • $\text{dis}_{\mathcal{G}}$                                | function for district in $\mathcal{G}$                                   | 25 |
| • $\text{ant}_{\mathcal{G}}$                                | function for anterior vertices in $\mathcal{G}$                          | 25 |
| • $\text{co}_{\mathcal{G}}$                                 | function for collider-connecting vertices in $\mathcal{G}$               | 31 |
| • $\text{pre}_{\mathcal{G}}^{\leq}$                         | function for preceding vertices in $\mathcal{G}$ with respect to $\leq$  | 33 |
| • $\text{mb}_{\mathcal{G}}, \text{mb}_{\mathcal{G}}^{\leq}$ | function for Markov blankets in $\mathcal{G}$ (with respect to $\leq$ )  | 33 |
| • $\text{cl}_{\mathcal{G}}, \text{cl}_{\mathcal{G}}^{\leq}$ | function for closures in $\mathcal{G}$ (with respect to $\leq$ )         | 33 |
| • $\text{ba}_{\mathcal{G}}$                                 | function for barren sets in $\mathcal{G}$                                | 38 |
| • $\text{tail}_{\mathcal{G}}$                               | function for tails in $\mathcal{G}$                                      | 38 |
| • $\text{ml}_{\mathcal{G}}^{\leq}$                          | function for minimal latent sets in $\mathcal{G}$ with respect to $\leq$ | 91 |

## A.10 Functions on Graphs

|                                      |  |    |
|--------------------------------------|--|----|
| • $\mathcal{A}$                      | function for anterior/ancestral sets       | 26 |
| • $\mathcal{H}$                      | function for heads                         | 38 |
| • $\mathcal{S}, \tilde{\mathcal{S}}$ | functions for parametrizing sets           | 44 |
| • $\mathcal{M}$                      | function for $m$ -connecting sets          | 62 |
| • $\mathcal{N}$                      | function for the non- $m$ -connecting sets | 62 |

### A.11 Evans' Partitioning Terms

|                        |  |    |
|------------------------|--|----|
| • $\Pi_{\mathcal{G}}$  | function for maximal heads   | 39 |
| • $\Psi_{\mathcal{G}}$ | function for relative complement of maximal heads                  | 39 |
| • $[\ ]_{\mathcal{G}}$ | syntax for a head partition function with respect to $\mathcal{G}$ | 39 |

### A.12 Interaction Terms

|                    |  |    |
|--------------------|--|----|
| • $\phi_A$         | symbol for interaction information rate                                | 70 |
| • $\phi_{A B}$     | symbol for conditional interaction information rate                    | 70 |
| • $\phi_{A,B C}$   | symbol for mutual information rate                                     | 70 |
| • $\delta_{A B}$   | symbol for the identifier for conditional interaction information sets | 70 |
| • $\delta_{A,B C}$ | symbol for the identifier for mutual information sets                  | 70 |

### A.13 Stable Mixed Graphs

|                 |   |    |
|-----------------|---|----|
| • $\mathcal{F}$ | symbol for a family of graphs                 | 48 |
| • $\alpha$      | function for marginalization and conditioning | 48 |

### A.14 Constrained Subsets

|                                  |   |    |
|----------------------------------|---|----|
| • $\subseteq_{\mathbb{R}}$       | symbol for subsets constrained to sets in $\mathbb{R}$                        | 72 |
| • $\subseteq^b$                  | symbol for subsets constrained to sets containing $b$                         | 72 |
| • $\subseteq_{\mathbb{R}}^b$     | symbol for subsets constrained to sets in $\mathbb{R}$ containing $b$         | 72 |
| • $\lceil \rceil_{\mathbb{R}}^b$ | symbol for maximal subsets constrained to sets in $\mathbb{R}$ containing $b$ | 72 |

## A.15 Integer-valued Multisets

|   |   |    |
|---|---|----|
| • $u$                                     | symbol for an imset                             | 53 |
| • $\delta_A$                              | symbol for the identifier for a set/set of sets | 54 |
| • $u_{\langle a,b C \rangle}$             | symbol for an elementary imset                  | 54 |
| • $u_{\langle A,B C \rangle}$             | symbol for a semi-elementary imset              | 55 |
| • $u_{\mathcal{G}}$                       | symbol for a standard imset                     | 59 |
| • $c_{\mathcal{G}}$                       | symbol for a characteristic imset               | 59 |
| • $u_{\mathcal{N}(\mathcal{G})}^{\leq,+}$ | primary imset constructed by Algorithm 3        | 80 |
| • $u_{\mathcal{N}(\mathcal{G})}^{\leq,-}$ | secondary imset constructed by Algorithm 3      | 80 |

## A.16 Non- $m$ -connecting Sets as Imsets

|  |  |    |
|--|--|----|
| • $\mathbf{M}^{\mathcal{G},b}$   | the list of $m$ -connecting sets constructed by Algorithm 2      | 76 |
| • $\mathbf{N}^{\mathcal{G},b}$   | the list of non- $m$ -connecting sets constructed by Algorithm 2 | 76 |
| • $\mathbf{M}_K^{\mathcal{G},b}, \mathbf{N}_J^{\mathcal{G},b}, \mathbf{M}_{J,K}^{\mathcal{G},b}$ | symbol for an intersection set of sets                           | 76 |
| • $\mathbf{U}_i^{\mathcal{G},b}, \mathbf{U}_J^{\mathcal{G},b}$                                   | symbol for a restricted universal set of sets                    | 76 |
| • $\mathbf{N}_{i,i}^{\mathcal{G},b}, \mathbf{N}_{J,K}^{\mathcal{G},b}$                           | symbol for a conditional set of sets                             | 76 |

## A.17 Curved Exponential Families

|                 |   |     |
|-----------------|---|-----|
| • $\theta$      | symbol for natural parameters             | 127 |
| • $\Theta$      | symbol for parameter space                | 127 |
| • $t$           | symbol for sufficient statistic           | 127 |
| • $\psi$        | symbol for cumulant function              | 127 |
| • $\Phi$        | symbol for diffeomorphism                 | 127 |
| • $\mathcal{F}$ | symbol for family of probability measures | 128 |

## A.18 Parameterization

|             |  |     |
|-------------|--|-----|
| • $z$       | symbol for binary transformation         | 137 |
| • $\Lambda$ | function for undirected edge parameters  | 134 |
| • $\Omega$  | function for bi-directed edge parameters | 134 |
| • $B$       | function for directed edge parameters    | 134 |
| • $\mu$     | function for mean parameters             | 134 |
| • $K$       | symbol for precision matrix              | 135 |
| • $\Sigma$  | symbol for covariance matrix             | 143 |

## Appendix B Additional Background, Examples, and Results

### B.1 Latent Projections

In what follows, we outline algorithms for latent projection of ribbonless graphs to ribbonless, summary, and ancestral graphs.

---

**Algorithm 8:**  $\alpha_{\text{RG}}(\mathcal{G}, L, S)$ 

---

**Input:** ribbonless graph:  $\mathcal{G} = (V, E)$ , disjoint sets  $L, S \subseteq V$

**Output:** ribbonless graph:  $\mathcal{G}'$

```
1 Set  $\mathcal{G}' = \mathcal{G}$  ;
2 foreach  $l \in L$  do
3   foreach triple  $\langle a, l, b \rangle$  in  $\mathcal{G}$  where  $a, b \in V \setminus LS$  do
4     if  $\left\{ \begin{array}{l} a \rightarrow l \rightarrow b \\ a - l \rightarrow b \\ a - l \leftrightarrow b \end{array} \right\}$  in  $\mathcal{G}$  and  $a \rightarrow b$  not in  $\mathcal{G}'$  then
5       | Add  $a \rightarrow b$  to  $\mathcal{G}'$ ;
6     end
7     if  $\left\{ \begin{array}{l} a \leftarrow l \rightarrow b \\ a \leftrightarrow l \rightarrow b \end{array} \right\}$  in  $\mathcal{G}$  and  $a \leftrightarrow b$  not in  $\mathcal{G}'$  then
8       | Add  $a \leftrightarrow b$  to  $\mathcal{G}'$ ;
9     end
10    if  $\left\{ \begin{array}{l} a \rightarrow l - b \\ a - l - b \end{array} \right\}$  in  $\mathcal{G}$  and  $a - b$  not in  $\mathcal{G}'$  then
11      | Add  $a - b$  to  $\mathcal{G}'$ ;
12    end
13  end
14  Remove  $l$  from  $\mathcal{G}'$ ;
15 end
16 foreach  $s \in S$  do
17   foreach triple  $\langle a, s', b \rangle$  in  $\mathcal{G}$  where  $a, b \in V \setminus LS$  and  $s' \in \text{ang}(S)$  do
18     if  $a \rightarrow s' \leftrightarrow b$  in  $\mathcal{G}$  and  $a \rightarrow b$  not in  $\mathcal{G}'$  then
19       | Add  $a \rightarrow b$  to  $\mathcal{G}'$ ;
20     end
21     if  $a \leftrightarrow s' \leftrightarrow b$  in  $\mathcal{G}$  and  $a \leftrightarrow b$  not in  $\mathcal{G}'$  then
22       | Add  $a \leftrightarrow b$  to  $\mathcal{G}'$ ;
23     end
24     if  $a \rightarrow s' \leftarrow b$  in  $\mathcal{G}$  and  $a - b$  not in  $\mathcal{G}'$  then
25       | Add  $a - b$  to  $\mathcal{G}'$ ;
26     end
27   end
28   Remove  $s$  from  $\mathcal{G}'$ ;
29 end
```

---



---

**Algorithm 9:**  $\alpha_{\text{SG}}(\mathcal{G}, L, S)$ 

---

**Input:** ribbonless graph:  $\mathcal{G} = (V, E)$ , disjoint sets  $L, S \subseteq V$

**Output:** summary graph:  $\mathcal{G}'$

```
1 Set  $\mathcal{G}' = \alpha_{\text{RG}}(\mathcal{G}, L, S)$  ;
2 foreach  $s' \in \text{an}_{\mathcal{G}}(S) \setminus S$  do
3   foreach  $a \in \text{pa}_{\mathcal{G}'}(s')$  do
4     Remove  $a \rightarrow s'$  from  $\mathcal{G}'$  ;
5     if  $a - s'$  not in  $\mathcal{G}'$  then
6       | Add  $a - s'$  to  $\mathcal{G}'$  ;
7     end
8   end
9   foreach  $a \in \text{sp}_{\mathcal{G}'}(s')$  do
10    Remove  $a \leftrightarrow s'$  from  $\mathcal{G}'$  ;
11    if  $a - s'$  not in  $\mathcal{G}'$  then
12      | Add  $a - s'$  to  $\mathcal{G}'$  ;
13    end
14  end
15 end
```

---

---

**Algorithm 10:**  $\alpha_{\text{AG}}(\mathcal{G}, L, S)$ 

---

**Input:** ribbonless graph:  $\mathcal{G} = (V, E)$ , disjoint sets  $L, S \subseteq V$

**Output:** ancestral graph:  $\mathcal{G}'$

```
1 Set  $\mathcal{G}' = \alpha_{\text{SG}}(\mathcal{G}, L, S)$  ;
2 foreach triple  $\langle a, b, c \rangle$  in  $\mathcal{G}'$  where  $a, c \in V \setminus LS$  and  $b \in \text{an}_{\mathcal{G}'}(c)$  do
3   if  $a \rightarrow b \leftrightarrow c$  in  $\mathcal{G}'$  and  $a \rightarrow c$  not in  $\mathcal{G}'$  then
4     | Add  $a \rightarrow c$  to  $\mathcal{G}'$  ;
5   end
6   if  $a \leftrightarrow b \leftrightarrow c$  in  $\mathcal{G}'$  and  $a \leftrightarrow c$  not in  $\mathcal{G}'$  then
7     | Add  $a \leftrightarrow c$  to  $\mathcal{G}'$  ;
8   end
9   if  $a \leftrightarrow c$  in  $\mathcal{G}'$  and  $a \in \text{an}_{\mathcal{G}'}(c)$  then
10    Remove  $a \leftrightarrow c$  from  $\mathcal{G}'$  ;
11    if  $a \rightarrow c$  not in  $\mathcal{G}'$  then
12      | Add  $a \rightarrow c$  to  $\mathcal{G}'$  ;
13    end
14  end
15 end
```

---

## B.2 The Causal Inference Algorithm

In what follows, we outline the causal inference (CI) algorithm [79]. The CI algorithm recovers a PAG  $\mathcal{G}$  that represents a Markov equivalence class of MAGs by querying a conditional independence oracle  $\mathcal{J}$

---

### Algorithm 11: CAUSAL INFERENCE CI( $\mathcal{J}$ )

---

**Input:** independence model:  $\mathcal{J}$   
**Output:** partial ancestral graph:  $\mathcal{G}$

- 1 Let  $\mathcal{G} = (V, E)$  where  $E = \{a \circ\text{-} \circ b \mid a, b \in V\}$  and initialize  $\text{Sep} = []$  ;
- 2 **foreach** edge  $a \circ\text{-} \circ b \in E$  **do**
- 3     **if** there exists  $Z \subseteq V \setminus \{a, b\}$  such that  $\langle a, b \mid Z \rangle \in \mathcal{J}$  **then**
- 4         Remove  $a \circ\text{-} \circ b$  from  $E$  ;
- 5         Append  $\langle a, b \mid Z \rangle$  and  $\langle b, a \mid Z \rangle$  to  $\text{Sep}$  ;
- 6     **end**
- 7 **end**
- 8 **foreach** unshielded triple  $\langle a, b, c \rangle$  in  $\mathcal{G}$  **do**
- 9     Rule 0: If  $\langle a, c \mid Z \cup b \rangle \notin \text{Sep}$  for all  $Z \subseteq V \setminus \{a, c\}$ , then orient it as a collider  
 $a \ast \rightarrow b \leftarrow \ast c$  ;
- 10 **end**
- 11 **repeat**
- 12     Rule 1: If  $a \ast \rightarrow b \circ\text{-} \ast c$ , and  $a$  and  $c$  are not adjacent, then orient the triple as  
 $a \ast \rightarrow b \rightarrow c$  ;
- 13     Rule 2: If  $a \rightarrow b \ast \rightarrow c$  or  $a \ast \rightarrow b \rightarrow c$ , and  $a \ast \text{-} \circ c$ , then orient  $a \ast \text{-} \circ c$  as  $a \circ \rightarrow c$  ;
- 14     Rule 3: If  $a \ast \rightarrow b \leftarrow \ast c$ ,  $a \ast \text{-} \circ d \circ\text{-} \ast c$ ,  $a$  and  $c$  are not adjacent, and  $d \ast \text{-} \circ b$ , then  
orient  $d \ast \text{-} \circ b$  as  $d \ast \rightarrow b$  ;
- 15     Rule 4: If  $\langle a, \dots, b, c, d \rangle$  is a discriminating path from  $a$  to  $d$  for  $c$  and  $c \circ\text{-} \ast d$ ,  
then: if there exists  $Z \subseteq V \setminus \{a, d\}$  such that  $\langle a, d \mid Z \cup c \rangle \in \text{Sep}$ , then orient  
 $c \circ\text{-} \ast d$  as  $c \rightarrow d$ ; otherwise orient the triple  $\langle b, c, d \rangle$  as  $b \leftrightarrow c \leftrightarrow d$  ;
- 16 **until** Rules 1 - 4 no longer apply;

---

## B.3 NSI Finds Non-minimal Solutions

The following example shows a shortcoming of Algorithm 3. If there exists a grouping of the non- $m$ -connecting set terms such that no adjustment is necessary, Algorithm 3 is not

guaranteed to find it. Furthermore, in the given example, there exists no total order for which Algorithm 3 finds such a grouping. We leave the poof of the latter to the reader.

Let  $\mathcal{G} = (V, E)$  be a directed MAG where  $V = \{a, b, c, d, e\}$  and  $E = \{a \leftrightarrow b, b \leftrightarrow c, c \leftrightarrow d, d \leftrightarrow e, e \leftrightarrow a\}$ . In other words,  $\mathcal{G}$  is a bi-directed five-cycle. Furthermore, let  $\leq$  be the total order where  $a \leq b \leq c \leq d \leq e$  and note that  $\leq$  is consistent with  $\mathcal{G}$ . The corresponding  $m$ -connecting and non- $m$ -connecting set are:

$$\begin{array}{ll} \mathcal{M}(\mathcal{G}) & \mathcal{N}(\mathcal{G}) \\ \{abcde, abcd, abce, abde, acde, & \{abd, acd, ace, bce, bde, \\ bcde, abc, abe, ade, bcd, cde, & ac, ad, bd, be, ce\} \\ ab, ae, bc, cd, de, a, b, c, d, d\} & \end{array}$$

Using imsets, we construct the following groupings of the non- $m$ -connecting sets:

$$\delta_{\mathcal{N}(\mathcal{G})} = [\delta_{abd} + \delta_{ad}] + [\delta_{acd} + \delta_{ac}] + [\delta_{ace} + \delta_{ce}] + [\delta_{bce} + \delta_{be}] + [\delta_{bde} + \delta_{bd}]$$

which simplifies to:

$$\delta_{\mathcal{N}(\mathcal{G})} = \delta_{N_{a,d|b}} + \delta_{N_{a,c|d}} + \delta_{N_{c,e|a}} + \delta_{N_{b,e|c}} + \delta_{N_{b,d|e}}$$

However, Algorithm 3 constructs the following groupings of the non- $m$ -connecting sets:

$$\begin{aligned} \delta_{\mathcal{N}(\mathcal{G})} &= [\delta_{ace} + \delta_{ce}] + [\delta_{bce} + \delta_{be} + \delta_{ce}] + [\delta_{bde} + \delta_{be}] + [\delta_{abd} + \delta_{ad} + \delta_{bd}] + [\delta_{acd} + \delta_{ad}] \\ &\quad - \delta_{ce} - \delta_{be} - \delta_{ad} \end{aligned}$$

which simplifies to:

$$\begin{aligned} \delta_{\mathcal{N}(\mathcal{G})} &= \delta_{N_{c,e|a}} + \delta_{N_{bc,e}} + \delta_{N_{b,e|d}} + \delta_{N_{ab,d}} + \delta_{N_{a,d|c}} \\ &\quad - \delta_{N_{c,e}} - \delta_{N_{b,e}} - \delta_{N_{a,d}} \end{aligned}$$

In what follows, we trace through Algorithm 3 to prove that it returns what we claim.

Run PAIRS( $\mathcal{G}, e$ ):

$$M^{\mathcal{G},e} = \{ae, e, de\} \quad N^{\mathcal{G},e} = \{ace, bce, bde\}$$

The following terms are the possible additions to the resulting imsets:

$$\begin{aligned} \delta_{\{T \subseteq N(\mathcal{G}), e\} V} &= \delta_{N_{1,1}^{\mathcal{G},e}} + \delta_{N_{2,2}^{\mathcal{G},e}} + \delta_{N_{3,3}^{\mathcal{G},e}} + \delta_{N_{12,12}^{\mathcal{G},e}} - \delta_{N_{12,1}^{\mathcal{G},e}} - \delta_{N_{12,2}^{\mathcal{G},e}} + \delta_{N_{13,13}^{\mathcal{G},e}} - \delta_{N_{13,1}^{\mathcal{G},e}} - \delta_{N_{13,3}^{\mathcal{G},e}} + \delta_{N_{23,23}^{\mathcal{G},e}} \\ &\quad - \delta_{N_{23,2}^{\mathcal{G},e}} - \delta_{N_{23,3}^{\mathcal{G},e}} + \delta_{N_{123,123}^{\mathcal{G},e}} - \delta_{N_{123,12}^{\mathcal{G},e}} - \delta_{N_{123,13}^{\mathcal{G},e}} - \delta_{N_{123,23}^{\mathcal{G},e}} + \delta_{N_{123,1}^{\mathcal{G},e}} + \delta_{N_{123,2}^{\mathcal{G},e}} + \delta_{N_{123,3}^{\mathcal{G},e}} \end{aligned}$$

Fill in the indices with actual set values:

$$\begin{aligned} \delta_{\{T \subseteq N(\mathcal{G}), e\} V} &= \delta_{N_{ac,ae}^{\mathcal{G},e}} + \delta_{N_{bce,e}^{\mathcal{G},e}} + \delta_{N_{bde,de}^{\mathcal{G},e}} + \delta_{N_{ce,e}^{\mathcal{G},e}} - \delta_{N_{ce,e}^{\mathcal{G},e}} - \delta_{N_{ce,e}^{\mathcal{G},e}} + \delta_{N_{e,e}^{\mathcal{G},e}} - \delta_{N_{e,e}^{\mathcal{G},e}} - \delta_{N_{e,e}^{\mathcal{G},e}} + \delta_{N_{be,e}^{\mathcal{G},e}} \\ &\quad - \delta_{N_{be,e}^{\mathcal{G},e}} - \delta_{N_{be,e}^{\mathcal{G},e}} + \delta_{N_{e,e}^{\mathcal{G},e}} - \delta_{N_{e,e}^{\mathcal{G},e}} - \delta_{N_{e,e}^{\mathcal{G},e}} - \delta_{N_{e,e}^{\mathcal{G},e}} + \delta_{N_{e,e}^{\mathcal{G},e}} + \delta_{N_{e,e}^{\mathcal{G},e}} + \delta_{N_{e,e}^{\mathcal{G},e}} \end{aligned}$$

Rewrite the imsets in their conditional forms:

$$\begin{aligned} \delta_{\{T \subseteq N(\mathcal{G}), e\} V} &= \delta_{N_{c,e|a}} + \delta_{N_{bc,e}} + \delta_{N_{b,e|d}} + \delta_{N_{c,e}} - \delta_{N_{c,e}} - \delta_{N_{c,e}} - \delta_{\emptyset} + \delta_{\emptyset} - \delta_{\emptyset} + \delta_{N_{b,e}} - \delta_{N_{b,e}} \\ &\quad - \delta_{N_{b,e}} + \delta_{\emptyset} - \delta_{\emptyset} - \delta_{\emptyset} - \delta_{\emptyset} + \delta_{\emptyset} + \delta_{\emptyset} + \delta_{\emptyset} \end{aligned}$$

Cancel like terms and drop empty imsets:

$$\delta_{\{T \subseteq N(\mathcal{G}), e\} V} = \delta_{N_{c,e|a}} + \delta_{N_{bc,e}} + \delta_{N_{b,e|d}} - \delta_{N_{c,e}} - \delta_{N_{b,e}}$$

Run PAIRS( $\mathcal{G}_{V \setminus e}, d$ ):

$$M^{\mathcal{G},d} = \{d, cd\} \quad N^{\mathcal{G},d} = \{abd, acd\}$$

The following terms are the possible additions to the resulting imsets:

$$\delta_{\{T \subseteq N(\mathcal{G}), e\} V \setminus e} = \delta_{N_{1,1}^{\mathcal{G},e}} + \delta_{N_{2,2}^{\mathcal{G},e}} + \delta_{N_{12,12}^{\mathcal{G},e}} - \delta_{N_{12,1}^{\mathcal{G},e}} - \delta_{N_{12,2}^{\mathcal{G},e}}$$

Fill in the indices with actual set values:

$$\delta_{\{T \subseteq N(\mathcal{G}), e\} V \setminus e} = \delta_{N_{abd,d}^{\mathcal{G},e}} + \delta_{N_{acd,cd}^{\mathcal{G},e}} + \delta_{N_{ad,d}^{\mathcal{G},e}} - \delta_{N_{ad,d}^{\mathcal{G},e}} - \delta_{N_{ad,d}^{\mathcal{G},e}}$$

Rewrite the imsets in their conditional forms:

$$\delta_{\{T \subseteq N(\mathcal{G}), e\} V \setminus e} = \delta_{N_{ab,d}} + \delta_{N_{a,d|c}} + \delta_{N_{a,d}} - \delta_{N_{a,d}} - \delta_{N_{a,d}}$$

Cancel like terms and drop empty imsets:

$$\delta_{\{T \subseteq \mathcal{N}(\mathcal{G}), e \in V \setminus e\}} = \delta_{N_{ab,d}} + \delta_{N_{a,d|c}} - \delta_{N_{a,d}}$$

Algorithm 3 completes since all the non- $m$ -connecting set are accounted for:

$$\begin{aligned} \delta_{N_{\mathcal{N}(\mathcal{G})}} &= \delta_{N_{c,e|a}} + \delta_{N_{bc,e}} + \delta_{N_{b,e|d}} + \delta_{N_{ab,d}} + \delta_{N_{a,d|c}} \\ &\quad - \delta_{N_{c,e}} - \delta_{N_{b,e}} - \delta_{N_{a,d}} \end{aligned}$$

#### B.4 Necessity of the Adjustment Term

In this section, we give an example for which our current proof strategy fails. This does not necessarily mean that the adjustment term is necessary.

Let  $\mathcal{G} = (V, E)$  be a directed MAG where  $V = \{a, b, c, d, e, f, g\}$  and  $E = \{a \rightarrow b, b \leftrightarrow c, c \leftrightarrow d, d \leftrightarrow e, e \leftrightarrow f, f \leftarrow g\}$ . In other words,  $\mathcal{G}$  is a collider chain of length seven. The  $m$ -connecting sets can be constructed using the heads and tails of the graph as follows:

$$\begin{aligned} \delta_{\mathcal{M}(\mathcal{G})} &= \delta_{bcdef|ag} + \delta_{bcde|a} + \delta_{cdef|g} + \delta_{bcd|a} + \delta_{cde} + \delta_{def|g} + \delta_{bc|a} + \delta_{cd} \\ &\quad + \delta_{de} + \delta_{ef|g} + \delta_{b|a} + \delta_c + \delta_d + \delta_e + \delta_{f|g} + \delta_a + \delta_g \end{aligned}$$

Note the following equalities:

$$\begin{aligned} \delta_{\mathcal{P}(V)} &= \delta_{\mathcal{M}(\mathcal{G})} + \delta_{\mathcal{N}(\mathcal{G})} + \delta_{\emptyset} \\ \mu_{\mathcal{P}} \delta_{\mathcal{P}(V)} &= \mu_{\mathcal{P}} \delta_{\mathcal{M}(\mathcal{G})} + \mu_{\mathcal{P}} \delta_{\mathcal{N}(\mathcal{G})} \\ \delta_{abcdefg} &= \mu_{\mathcal{P}} \delta_{\mathcal{M}(\mathcal{G})} + \mu_{\mathcal{P}} \delta_{\mathcal{N}(\mathcal{G})} \end{aligned}$$

Accordingly, the Möbius inversion of the non- $m$ -connecting sets are as follows:

$$\mu_{\mathcal{P}} \delta_{\mathcal{N}(\mathcal{G})} = \delta_{abcdefg} - \mu_{\mathcal{P}} \delta_{\mathcal{M}(\mathcal{G})}$$

Note the following Möbius inversions:

$$\mu_{\mathcal{P}} \delta_{bcdef|ag} = \delta_{abcdefg} - \delta_{abcdeg} - \delta_{abcdfg} - \delta_{abcefg} - \delta_{abdefg} - \delta_{acdefg} + \delta_{abcdg}$$

$$\begin{aligned}
& + \delta_{abceg} + \delta_{abdeg} + \delta_{acdeg} + \delta_{abcfg} + \delta_{abdfg} + \delta_{acdfg} + \delta_{abefg} + \delta_{acefg} \\
& + \delta_{adefg} - \delta_{abcg} - \delta_{abdg} - \delta_{abeg} - \delta_{abfg} - \delta_{acdg} - \delta_{aceg} - \delta_{acfg} \\
& - \delta_{adeg} - \delta{adfg} - \delta_{aefg} + \delta_{abg} + \delta_{acg} + \delta_{adg} + \delta_{aeg} + \delta_{afg} - \delta_{ag} \\
\mu_{\mathbb{P}}\delta_{bcde|a} &= \delta_{abcde} - \delta_{abcd} - \delta_{abce} - \delta_{abde} - \delta_{acde} + \delta_{abc} + \delta_{abd} \\
& + \delta_{abe} + \delta_{acd} + \delta_{ace} + \delta_{ade} - \delta_{ab} - \delta_{ac} - \delta_{ad} - \delta_{ae} + \delta_a \\
\mu_{\mathbb{P}}\delta_{cdef|g} &= \delta_{cdefg} - \delta_{cdeg} - \delta_{cdfg} - \delta_{cefg} - \delta_{defg} + \delta_{cdg} + \delta_{ceg} \\
& + \delta_{cfg} + \delta_{deg} + \delta_{dfg} + \delta_{efg} - \delta_{cg} - \delta_{dg} - \delta_{eg} - \delta_{fg} + \delta_g \\
\mu_{\mathbb{P}}\delta_{bcd|a} &= \delta_{abcd} - \delta_{abc} - \delta_{abd} - \delta_{acd} + \delta_{ab} + \delta_{ac} + \delta_{ad} - \delta_a \\
\mu_{\mathbb{P}}\delta_{cde} &= \delta_{cde} - \delta_{cd} - \delta_{ce} - \delta_{de} + \delta_c + \delta_d + \delta_e \\
\mu_{\mathbb{P}}\delta_{def|g} &= \delta_{defg} - \delta_{deg} - \delta_{dfg} - \delta_{efg} + \delta_{dg} + \delta_{eg} + \delta_{fg} - \delta_g \\
\mu_{\mathbb{P}}\delta_{bc|a} &= \delta_{abc} - \delta_{ab} - \delta_{ac} + \delta_a \\
\mu_{\mathbb{P}}\delta_{cd} &= \delta_{cd} - \delta_c - \delta_d \\
\mu_{\mathbb{P}}\delta_{de} &= \delta_{de} - \delta_d - \delta_e \\
\mu_{\mathbb{P}}\delta_{ef|g} &= \delta_{efg} - \delta_{eg} - \delta_{fg} + \delta_g \\
\mu_{\mathbb{P}}\delta_{b|a} &= \delta_{ab} - \delta_a \\
\mu_{\mathbb{P}}\delta_c &= \delta_c \\
\mu_{\mathbb{P}}\delta_d &= \delta_d \\
\mu_{\mathbb{P}}\delta_e &= \delta_e \\
\mu_{\mathbb{P}}\delta_{f|g} &= \delta_{fg} - \delta_g \\
\mu_{\mathbb{P}}\delta_a &= \delta_a \\
\mu_{\mathbb{P}}\delta_g &= \delta_g
\end{aligned}$$

The non- $m$ -connecting sets are then as follows:

$$\begin{aligned}
\mu_{\mathcal{P}}\delta_{\mathcal{N}(\mathcal{G})} = & \delta_{abcdeg} + \delta_{abcdfg} + \delta_{abcefg} + \delta_{abdefg} + \delta_{acdefg} - \delta_{abcdg} - \delta_{abceg} - \delta_{abdeg} - \delta_{acdeg} \\
& - \delta_{abcfg} - \delta_{abdfg} - \delta_{acdfg} - \delta_{abefg} - \delta_{acefg} - \delta_{adefg} + \delta_{abcg} + \delta_{abdg} + \delta_{abeg} + \delta_{abfg} \\
& + \delta_{acdg} + \delta_{aceg} + \delta_{acfg} + \delta{adeg} + \delta{adfg} + \delta{aefg} - \delta{abg} - \delta{acg} - \delta{adg} - \delta{aeg} - \delta{afg} \\
& + \delta_{ag} - \delta_{abcde} + \delta_{abcd} + \delta_{abce} + \delta_{abde} + \delta_{acde} - \delta_{abc} - \delta_{abd} - \delta_{abe} - \delta_{acd} - \delta_{ace} - \delta_{ade} \\
& + \delta_{ab} + \delta_{ac} + \delta_{ad} + \delta_{ae} - \delta_a - \delta_{cdefg} + \delta_{cdeg} + \delta_{cdfg} + \delta_{cefg} + \delta_{defg} - \delta_{cdg} - \delta_{ceg} \\
& - \delta_{cfg} - \delta_{deg} - \delta{dfg} - \delta{efg} + \delta{cg} + \delta{dg} + \delta{eg} + \delta{fg} - \delta{g} - \delta{abcd} + \delta{abc} + \delta{abd} + \delta{acd} \\
& - \delta{ab} - \delta{ac} - \delta{ad} + \delta{a} - \delta{cde} + \delta{cd} + \delta{ce} + \delta{de} - \delta{c} - \delta{d} - \delta{e} - \delta{defg} + \delta{deg} + \delta{dfg} \\
& + \delta{efg} - \delta{dg} - \delta{eg} - \delta{fg} + \delta{g} - \delta{abc} + \delta{ab} + \delta{ac} - \delta{a} - \delta{cd} + \delta{c} + \delta{d} - \delta{de} + \delta{d} \\
& + \delta{e} - \delta{efg} + \delta{eg} + \delta{fg} - \delta{g} - \delta{ab} + \delta{a} - \delta{c} - \delta{d} - \delta{e} - \delta{fg} + \delta{g} - \delta{a} - \delta{g}
\end{aligned}$$

Cancelling like terms:

$$\begin{aligned}
\mu_{\mathcal{P}}\delta_{\mathcal{N}(\mathcal{G})} = & \delta_{abcdeg} + \delta_{abcdfg} + \delta_{abcefg} + \delta_{abdefg} + \delta_{acdefg} - \delta_{abcde} - \delta_{abcdg} - \delta_{abceg} - \delta_{abcfg} - \delta_{abdeg} \\
& - \delta_{abdfg} - \delta_{abefg} - \delta_{acdeg} - \delta_{acdfg} - \delta_{acefg} - \delta_{adefg} - \delta_{cdefg} + \delta_{abcg} + \delta_{abdg} + \delta_{abeg} \\
& + \delta_{abfg} + \delta_{acdg} + \delta_{aceg} + \delta_{acfg} + \delta_{adeg} + \delta_{adfg} + \delta_{aefg} + \delta_{cdeg} + \delta_{cdfg} + \delta_{cefg} + \delta_{abce} \\
& + \delta_{abde} + \delta_{acde} - \delta_{abc} - \delta_{abe} - \delta_{abg} - \delta_{ace} - \delta_{acg} - \delta_{ade} - \delta_{adg} - \delta_{aeg} - \delta_{afg} - \delta_{cde} \\
& - \delta_{cdg} - \delta_{ceg} - \delta_{cfg} - \delta_{efg} + \delta_{ac} + \delta_{ae} + \delta_{ag} + \delta_{ce} + \delta_{cg} + \delta_{eg} - \delta_a - \delta_c - \delta_e - \delta_g
\end{aligned}$$

Let  $S_i(\mu_{\mathcal{P}}\delta_{\mathcal{N}(\mathcal{G})})$  sum the Möbius inversion of the non- $m$ -connecting sets of cardinality  $i$ :

$$S_i(\mu_{\mathcal{P}}\delta_{\mathcal{N}(\mathcal{G})}) = \sum_{\substack{T \in \mathcal{P}(V) \\ |T|=i}} \mu_{\mathcal{P}}\delta_{\mathcal{N}(\mathcal{G})}(T)$$

The sums are then as follows:

- $S_6(\mu_{\mathcal{P}}\delta_{\mathcal{N}(\mathcal{G})}) = 5$
- $S_5(\mu_{\mathcal{P}}\delta_{\mathcal{N}(\mathcal{G})}) = -12$
- $S_4(\mu_{\mathcal{P}}\delta_{\mathcal{N}(\mathcal{G})}) = 16$
- $S_3(\mu_{\mathcal{P}}\delta_{\mathcal{N}(\mathcal{G})}) = -14$
- $S_2(\mu_{\mathcal{P}}\delta_{\mathcal{N}(\mathcal{G})}) = 6$

- $S_1(\mu_{\mathcal{P}}\delta_{\mathcal{N}(\mathcal{G})}) = -4$

We need to construct a linear combination of semi-elementary imsets with positive coefficients. Consider semi-elementary imsets of the form  $u_{\langle A, B|C \rangle}$  where  $|A| = 1$ ,  $|B| = 1$ , and  $|C| = 4$  and note:

- $S_6(u_{\langle A, B|C \rangle}) = 1$
- $S_5(u_{\langle A, B|C \rangle}) = -2$
- $S_4(u_{\langle A, B|C \rangle}) = 1$
- $S_3(u_{\langle A, B|C \rangle}) = 0$
- $S_2(u_{\langle A, B|C \rangle}) = 0$
- $S_1(u_{\langle A, B|C \rangle}) = 0$

Let  $u$  be a structural imsets constructed as a linear combination of semi-elementary imsets of the considered form such that the coefficients sum to 5, then:

- $S_6(\mu_{\mathcal{P}}\delta_{\mathcal{N}(\mathcal{G})} - u) = 0$
- $S_5(\mu_{\mathcal{P}}\delta_{\mathcal{N}(\mathcal{G})} - u) = -2$
- $S_4(\mu_{\mathcal{P}}\delta_{\mathcal{N}(\mathcal{G})} - u) = 11$
- $S_3(\mu_{\mathcal{P}}\delta_{\mathcal{N}(\mathcal{G})} - u) = -14$
- $S_2(\mu_{\mathcal{P}}\delta_{\mathcal{N}(\mathcal{G})} - u) = 6$
- $S_1(\mu_{\mathcal{P}}\delta_{\mathcal{N}(\mathcal{G})} - u) = -4$

Accordingly, to deal with  $S_5(\mu_{\mathcal{P}}\delta_{\mathcal{N}(\mathcal{G})} - u) = -2$ , it is necessary to have semi-elementary imsets with negative coefficients. Note, that if we had chosen semi-elementary imsets of any other form, it would only exacerbate this issue.

## B.5 Comparison to Bayesian Scoring of Constraints

Jabbari et al. formulated Bayesian scoring of constraints (BSC) to estimate the log probability that an independence model induced by a directed MAG is induced by a probability measure. BSC sums over the log probabilities that conditional independence statements are



represented or not represented in the probability measure given that they have the same representation status in the directed MAG [41, 42]. Under Jabbari et al.'s assumptions, the sum in question returns the log probability that an independence model induced by the directed MAG is induced by the probability measure. In this section, we show that the  $m$ -connecting set factorization and BSC are equivalent to instantiations of the general imsetal factorization given by Theorem 3.5.1 up to a constant.

The general imsetal factorization is given with respect to a structural imset. Let  $V$  be a non-empty set of variables and  $X$  be a collection of random variables indexed by  $V$  with probability measure  $P$  that admits density  $f(x)$  with respect to dominating  $\sigma$ -finite product measure  $\nu$ . If  $u$  is a structural imset over  $V$ , then  $\mathcal{J}(u) \subseteq \mathcal{J}(P)$  if and only if:

$$\log f(x) = \log f(x) - \sum_{T \in \mathcal{P}(V)} u(T) \log f_T(x).$$

Let  $\mathcal{G} = (V, E)$  be a directed MAG. The  $m$ -connecting set factorization fits the form of the general imsetal factorization:

$$\begin{aligned} \log f(x) &= \sum_{M \in \mathcal{M}(\mathcal{G})} \phi_M(x) - \sum_{N \in \mathcal{N}(\mathcal{G})} u_{\mathcal{N}(\mathcal{G})}^{\leq, -}(N) \phi_N(x) \\ &= \sum_{T \in \mathcal{P}(V)} (\delta_{\mathcal{M}(\mathcal{G})}(T) - u_{\mathcal{N}(\mathcal{G})}^{\leq, -}(T)) \phi_T(x) \\ &= \sum_{T \in \mathcal{P}(V)} (1 - u_{\mathcal{N}(\mathcal{G})}^{\leq, +}(T)) \phi_T(x) \\ &= \log f(x) - \sum_{T \in \mathcal{N}(\mathcal{G})} u_{\mathcal{N}(\mathcal{G})}^{\leq, +}(T) \phi_T(x) \\ &= \log f(x) - \sum_{T \in \mathcal{P}(V)} \mu_{\mathcal{P}} u_{\mathcal{N}(\mathcal{G})}^{\leq, +}(T) \log f_T(x). \end{aligned}$$

Jabbari uses BIC to formulate the probability that a conditional independence statement is represented or not represented in a probability measure [41]:

$$\begin{aligned} \Pr(\langle a, b \mid C \rangle \in \mathcal{J}(P) \mid x^1, \dots, x^n) &\propto n^{-\frac{|\Theta_{\langle a, b \mid C \rangle}|}{2}} \prod_{i=1}^n \frac{f_{a \cup C}(x^i \mid \hat{\theta}_n^{\text{mle}}) f_{b \cup C}(x^i \mid \hat{\theta}_n^{\text{mle}})}{f_C(x^i \mid \hat{\theta}_n^{\text{mle}})}; \\ \Pr(\langle a, b \mid C \rangle \notin \mathcal{J}(P) \mid x^1, \dots, x^n) &\propto n^{-\frac{|\Theta|}{2}} \prod_{i=1}^n f_{ab \cup C}(x^i \mid \hat{\theta}_n^{\text{mle}}). \end{aligned}$$

In order to directly compare BSC to the  $m$ -connecting set factorization we only reason about the likelihood component of the score using density terms. The parameter penalty can be added back later similar to how we add a parameter penalty to the  $m$ -connecting set factorization to formulate  $\hat{\text{BIC}}$ :

$$\begin{aligned}
\text{BSC}_{\mathcal{G}}(x) &= \sum_{\langle a,b|C \rangle \in \mathcal{J}(\mathcal{G})} \log \Pr(\langle a,b|C \rangle \in \mathcal{J}(P) \mid x) + \sum_{\langle a,b|C \rangle \notin \mathcal{J}(\mathcal{G})} \log \Pr(\langle a,b|C \rangle \notin \mathcal{J}(P) \mid x) \\
&= \sum_{\langle a,b|C \rangle \in \mathcal{J}(\mathcal{G})} [\log f_{a \cup C}(x) + \log f_{b \cup C}(x) - \log f_C(x)] + \sum_{\langle a,b|C \rangle \notin \mathcal{J}(\mathcal{G})} \log f_{ab \cup C}(x) + g(x) \\
&= \sum_{\langle a,b|C \rangle \in \mathcal{J}(\mathcal{G})} [\log f_{a \cup C}(x) + \log f_{b \cup C}(x) - \log f_C(x)] + \sum_{\langle a,b|C \rangle \notin \mathcal{J}(\mathcal{G})} \log f_{ab \cup C}(x) \\
&\quad - \sum_{\langle a,b|C \rangle \in \mathcal{J}(V)} \log f_{ab \cup C}(x) + \sum_{\langle a,b|C \rangle \in \mathcal{J}(V)} \log f_{ab \cup C}(x) + g(x) \\
&= \sum_{\langle a,b|C \rangle \in \mathcal{J}(\mathcal{G})} [\log f_{a \cup C}(x) + \log f_{b \cup C}(x) - \log f_C(x) - \log f_{ab \cup C}(x)] + g'(x) \\
&= - \sum_{\langle a,b|C \rangle \in \mathcal{J}(\mathcal{G})} \phi_{a,b|C}(x) + g'(x) \\
&= - \sum_{\langle a,b|C \rangle \in \mathcal{J}(\mathcal{G})} \phi_{a,b|C}(x) + g'(x) - \log f(x) + \log f(x) \\
&= \log f(x) - \sum_{\langle a,b|C \rangle \in \mathcal{J}(\mathcal{G})} \phi_{a,b|C}(x) + g''(x)
\end{aligned}$$

where  $g(x) = \sum_{\langle a,b|C \rangle \in \mathcal{J}(V)} \log \left[ \frac{f_{a \cup C}(x) f_{b \cup C}(x)}{f_C(x)} + f_{ab \cup C}(x) \right]$  is the sum of conditional independence statement normalization terms,  $g'(x) = g(x) + \sum_{\langle a,b|C \rangle \in \mathcal{J}(V)} \log f_{ab \cup C}(x)$ , and  $g''(x) = g'(x) - \log f(x)$ . Notably  $g''(x)$  is constant with respect to  $\mathcal{G}$ . In the formula above, we sum over all elementary conditional independence statements for illustrative purposes. In practice, the sum would be restricted to the set of conditional independence statements considered by BSC.

Let  $u_{\text{BSC}_{\mathcal{G}}}$  be the structural imsets constructed from the of conditional independence statements represented in  $\mathcal{G}$  considered by BSC. BSC fits the form of the general imsetal factorization:

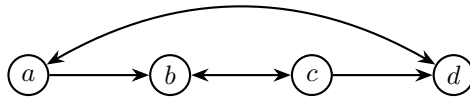
$$\text{BSC}_{\mathcal{G}}(x) = \log f(x) - \sum_{T \in \mathcal{P}(V)} u_{\text{BSC}_{\mathcal{G}}}(T) \log f_T(x) + g''(x).$$

Accordingly, both the  $m$ -connecting set factorization and BSC are equivalent to instan-

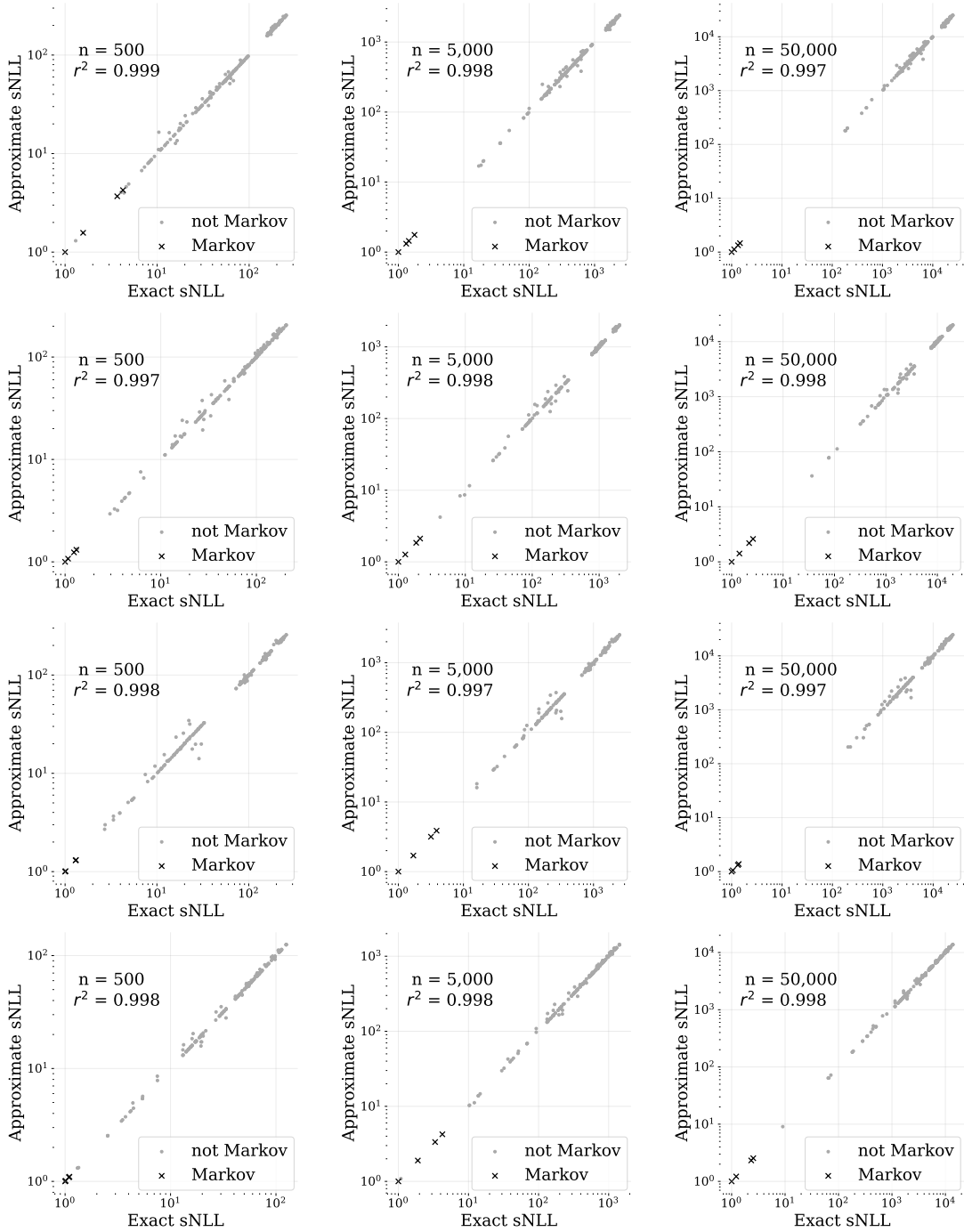
tiations of the general imsetal factorization up to a constant using structural imsets  $\mu_{\mathcal{P}} u_{\mathcal{N}(g)}^{\leq,+}$  and  $u_{\text{BSC}_g}$  respectively. Notably, the set of conditional independence statements considered by BSC to construct  $u_{\text{BSC}_g}$  are given by reruns of the FCI algorithm. Accordingly,  $u_{\text{BSC}_g}$  is likely to be constructed from redundant elementary imsets and its Möbius inversion is not guaranteed to assign a non-zero integer value to all non- $m$ -connecting sets. In contrast,  $u_{\mathcal{N}(g)}^{\leq,+}$  is constructed to reduce redundancy and guarantee that a positive integer value is assigned to all non- $m$ -connecting sets. The empirical ramifications of these properties remains to be explored.

## B.6 Shifted NLL Comparison

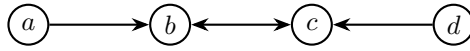
# Directed MAG



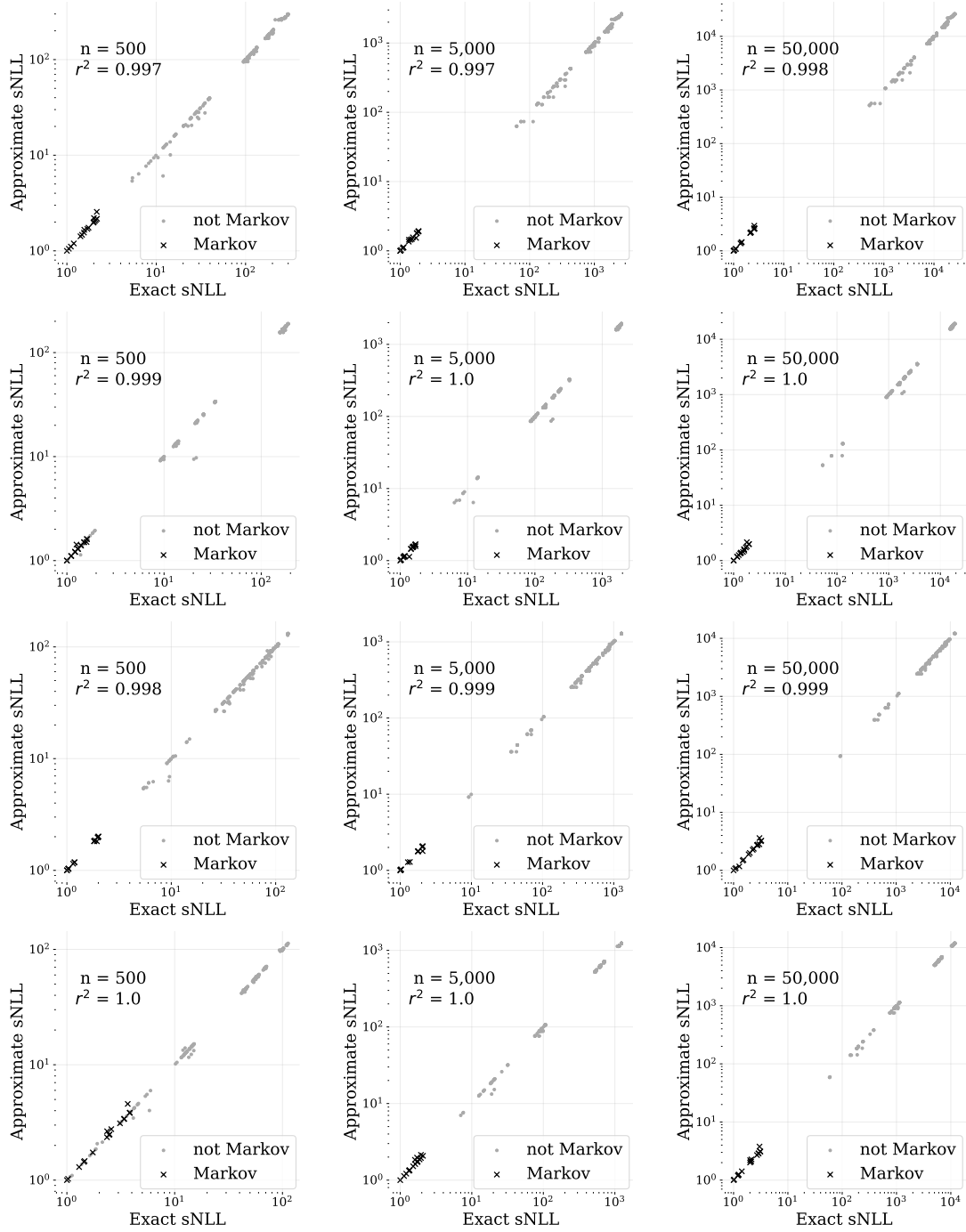
## Negative Log-likelihood



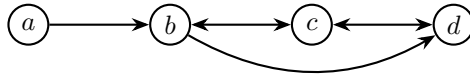
# Directed MAG



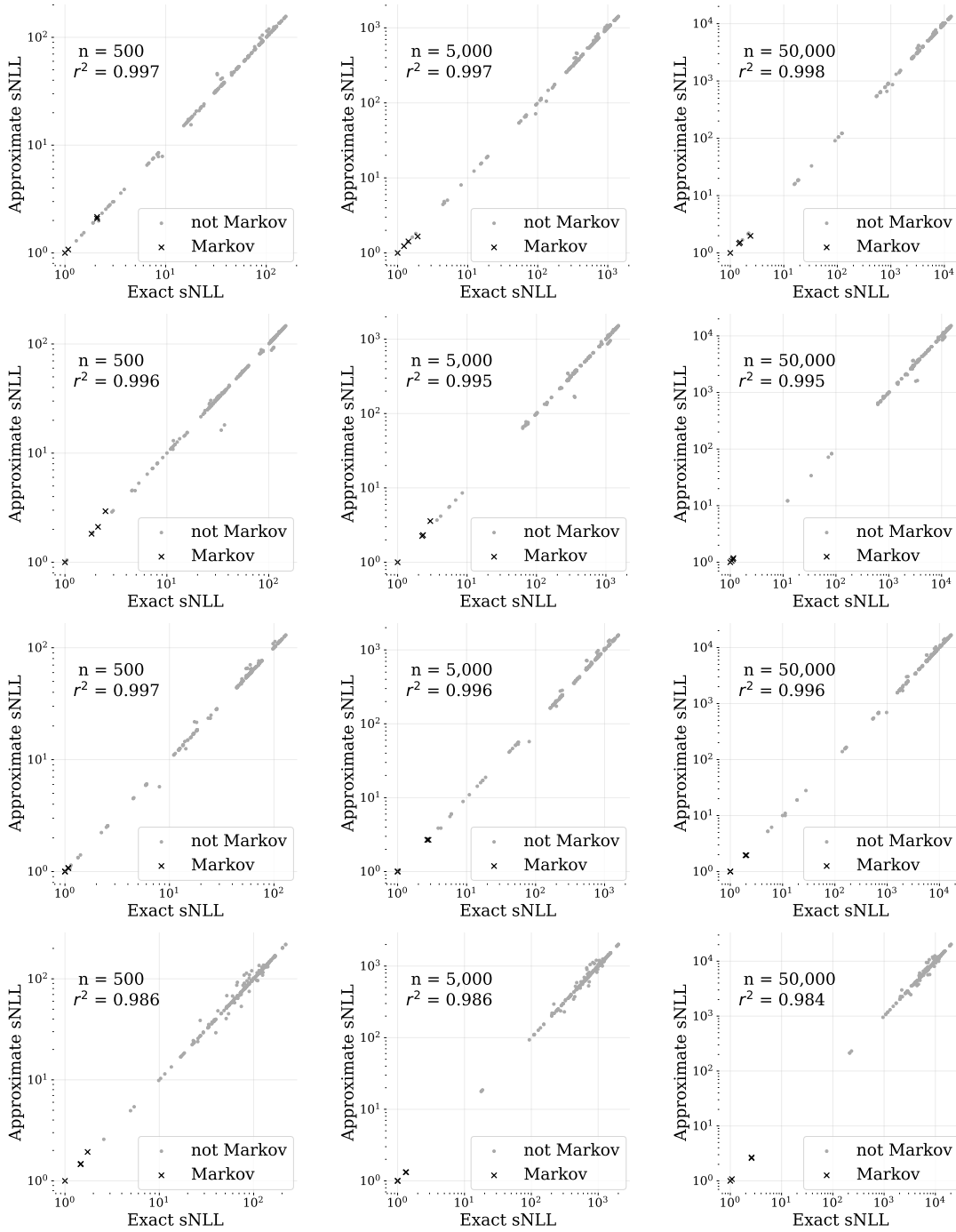
## Negative Log-likelihood



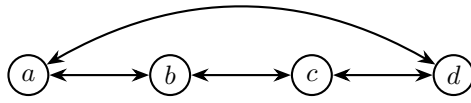
# Directed MAG



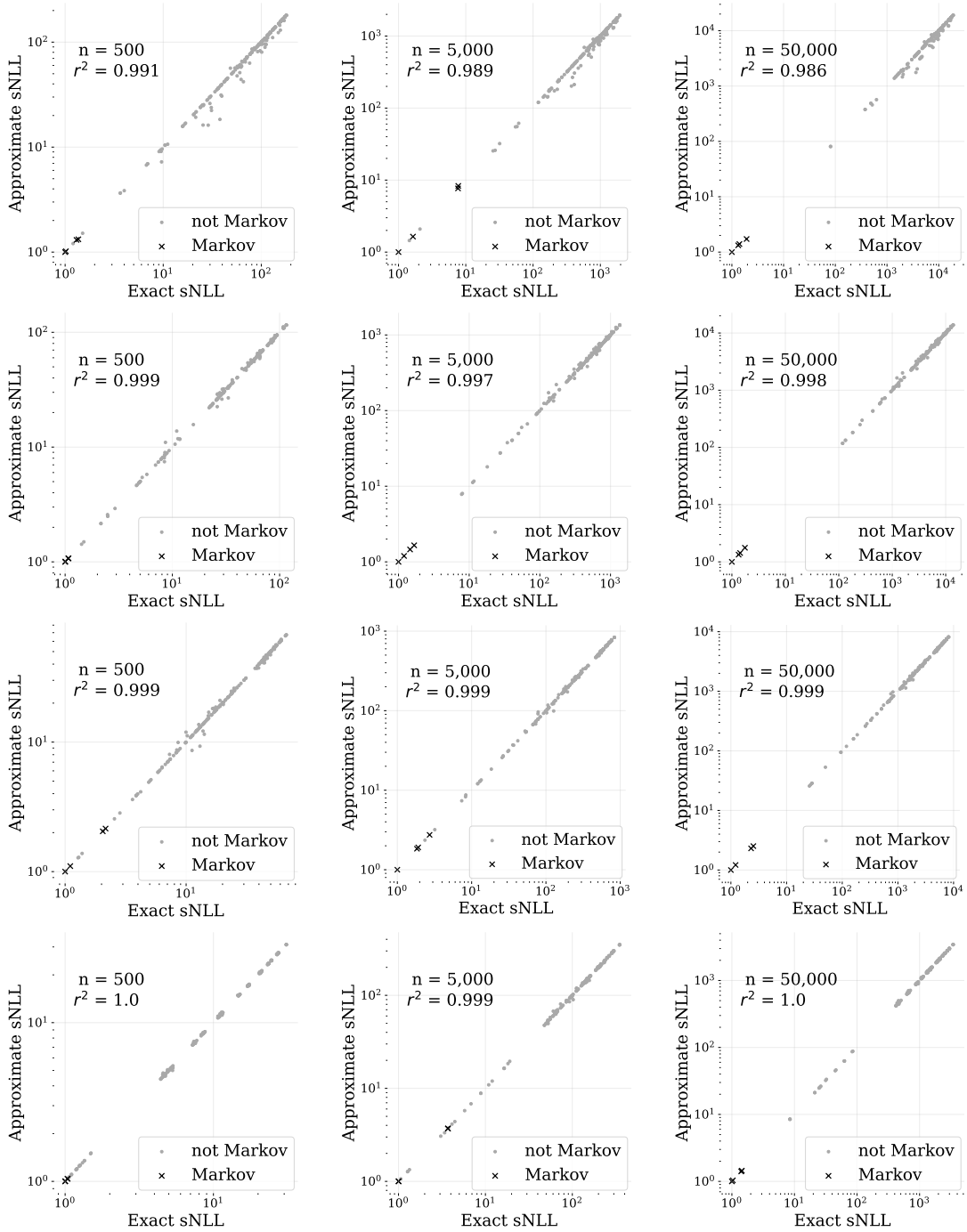
## Negative Log-likelihood



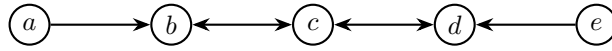
# Directed MAG



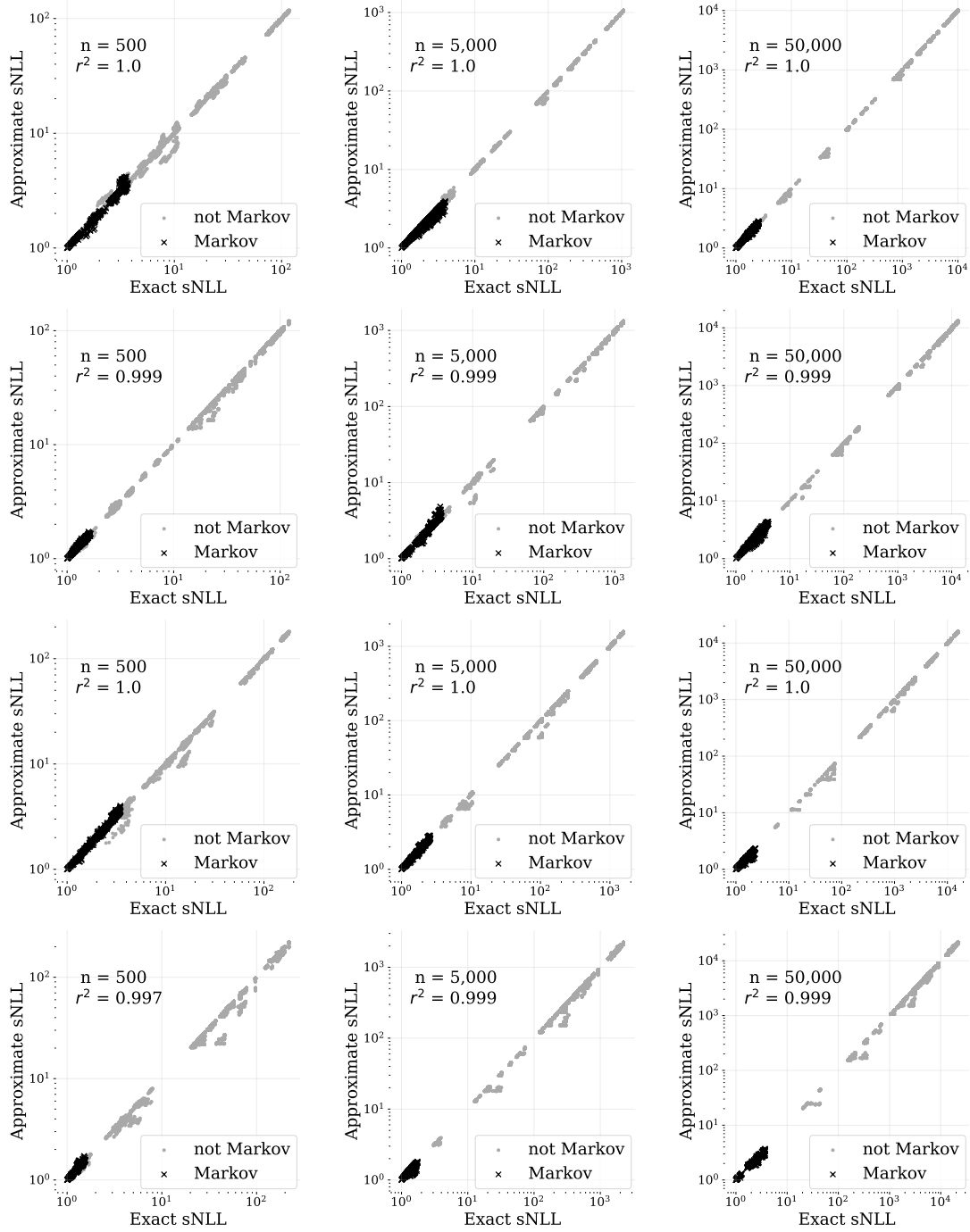
## Negative Log-likelihood



# Directed MAG

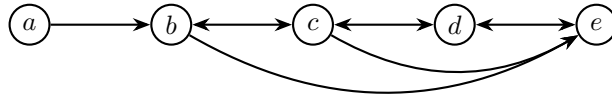


## Negative Log-likelihood

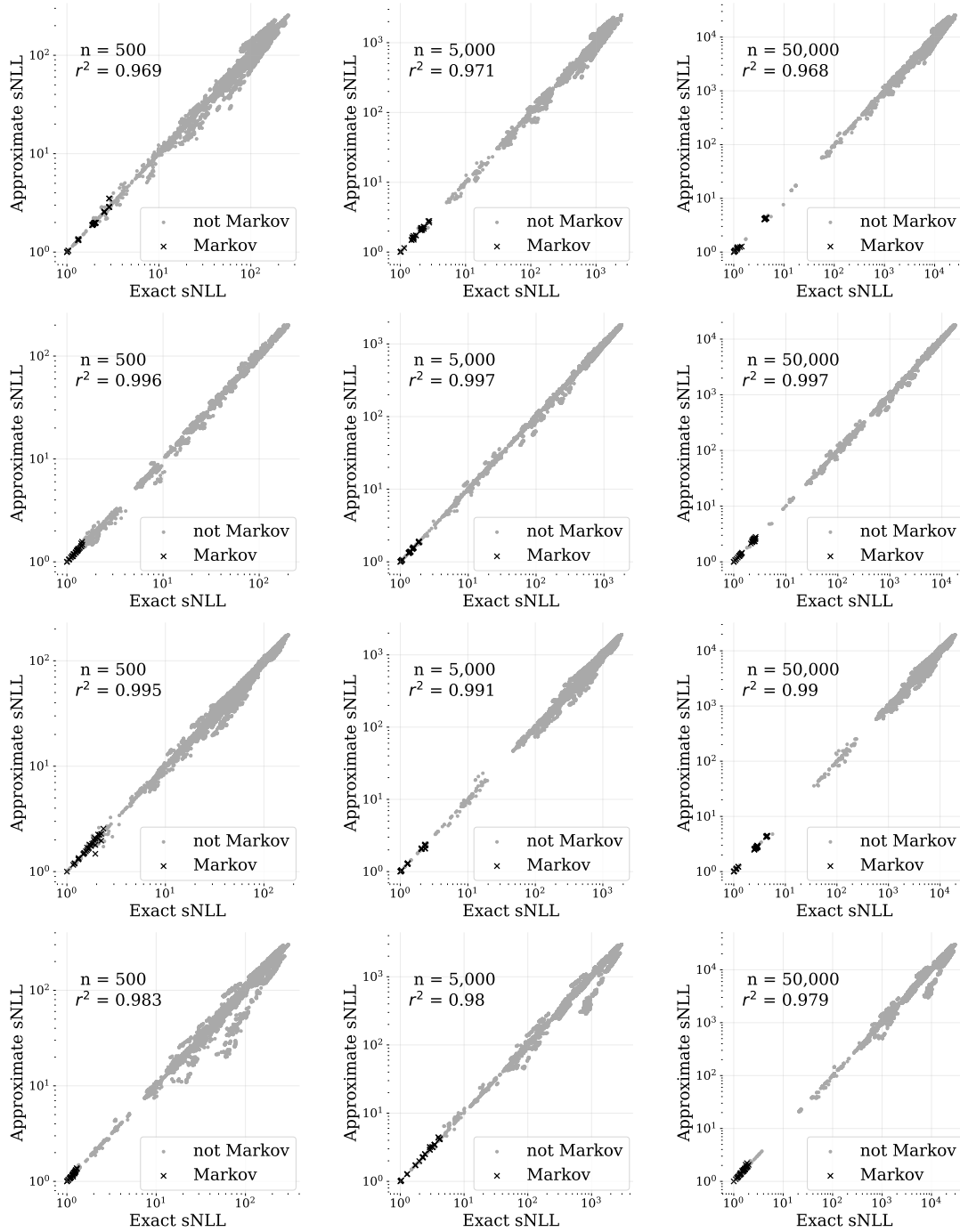




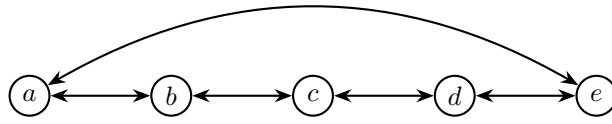
# Directed MAG



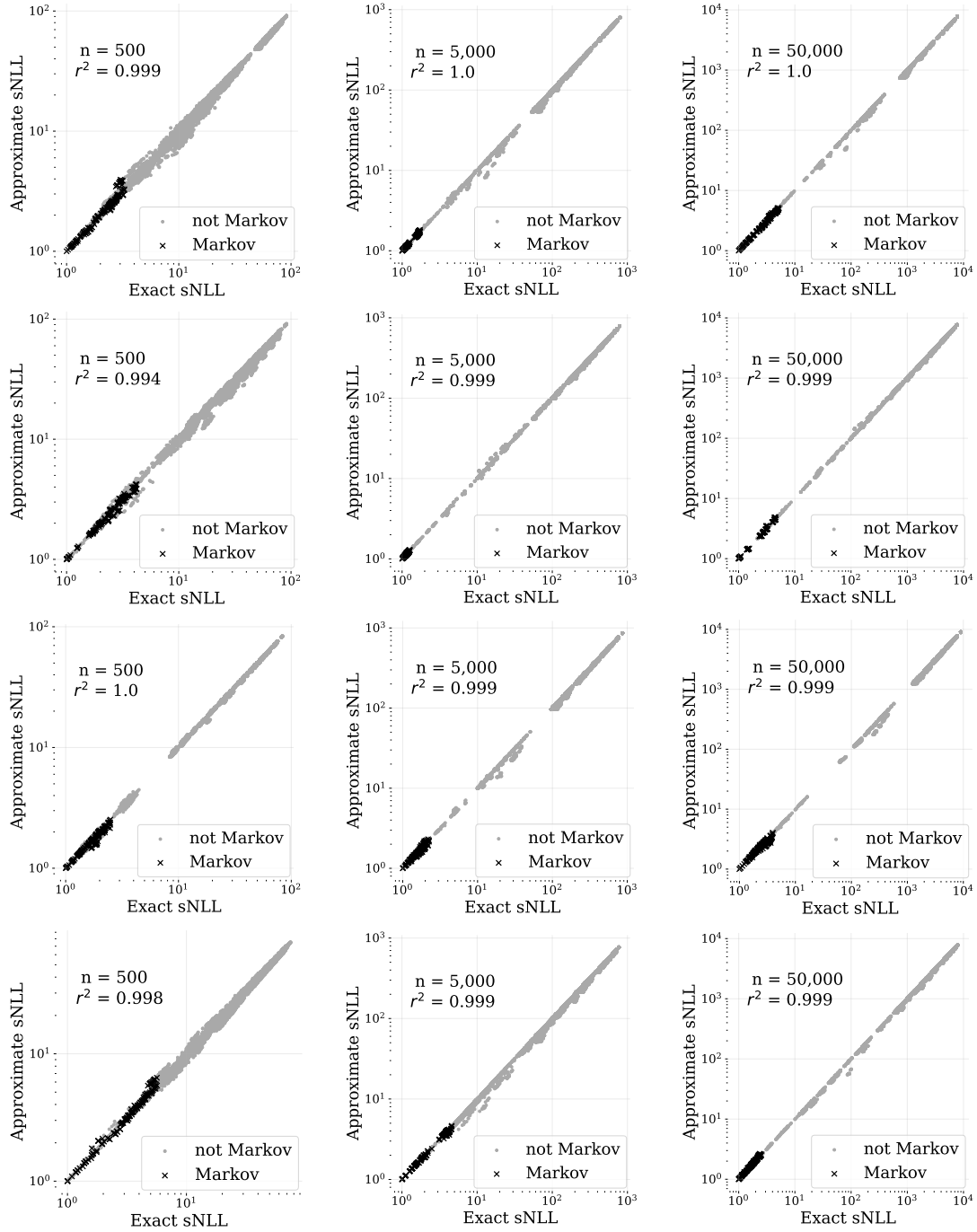
## Negative Log-likelihood



# Directed MAG

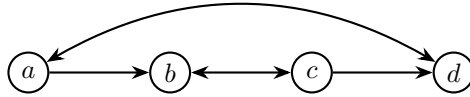


## Negative Log-likelihood

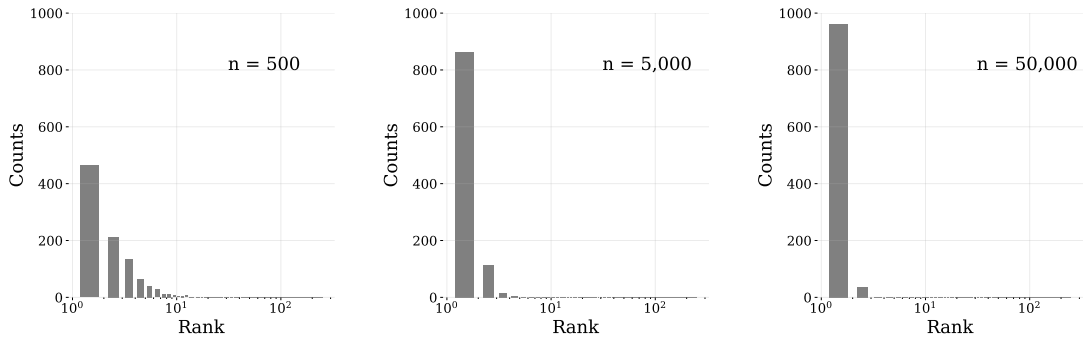


## B.7 Exact Histograms

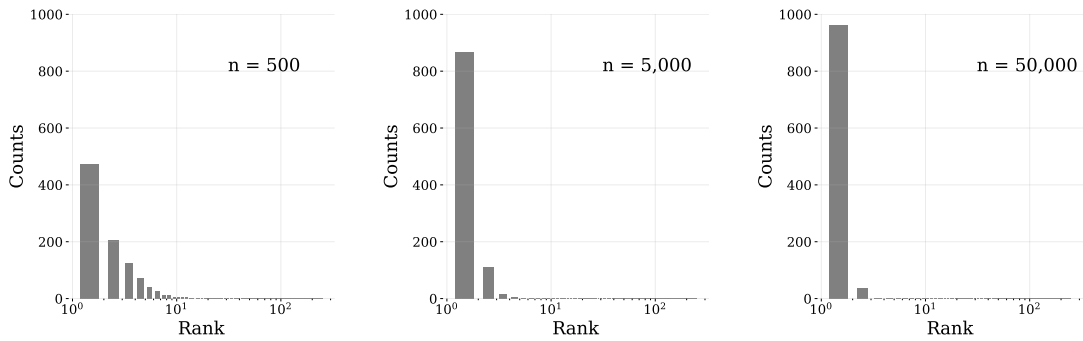
Directed MAG



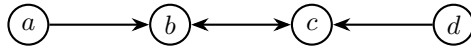
MEC Recovery by Approximate



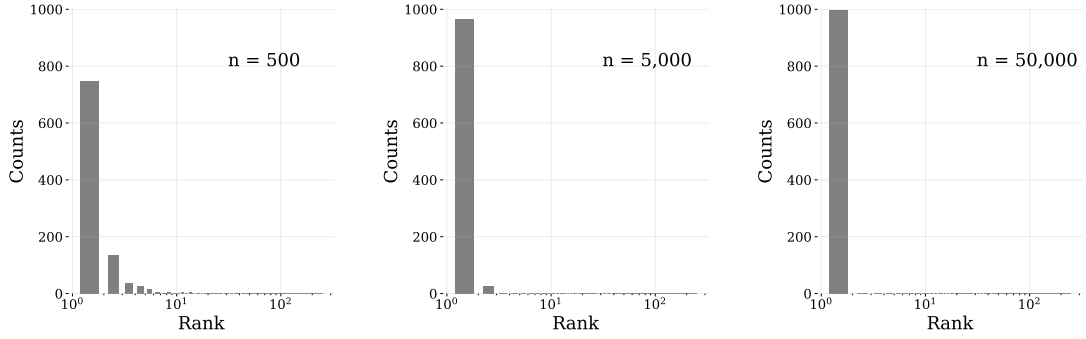
MEC Recovery by Exact



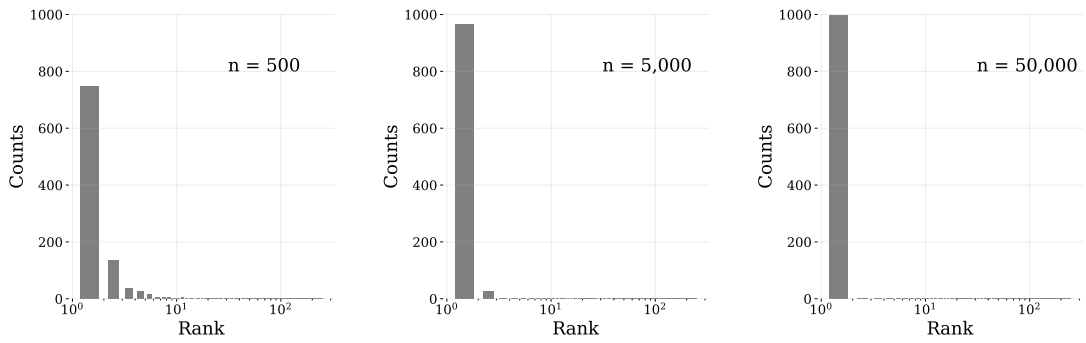
Directed MAG



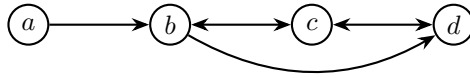
MEC Recovery by Approximate



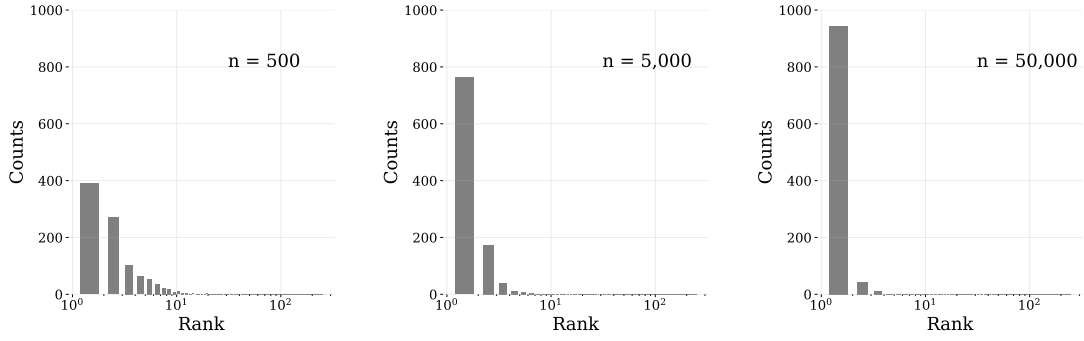
MEC Recovery by Exact



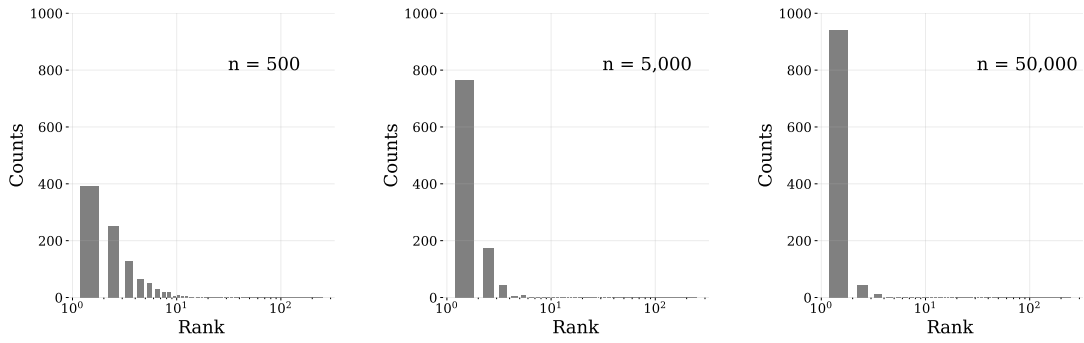
Directed MAG



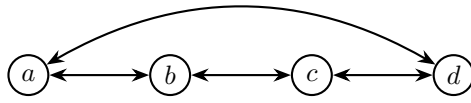
MEC Recovery by Approximate



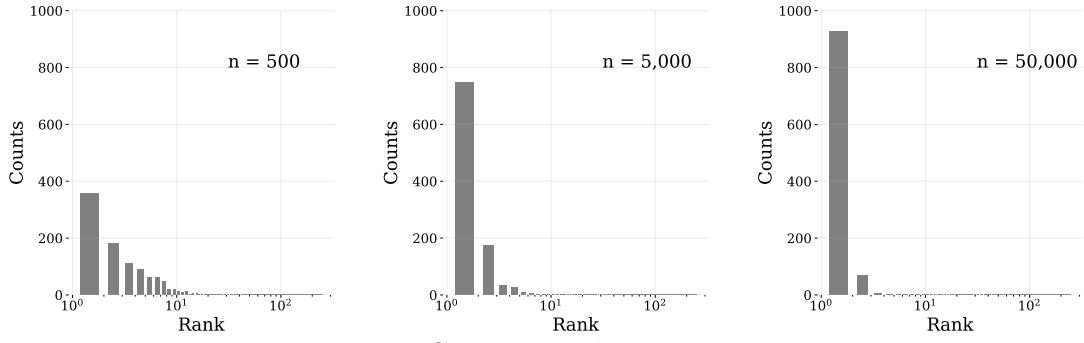
MEC Recovery by Exact



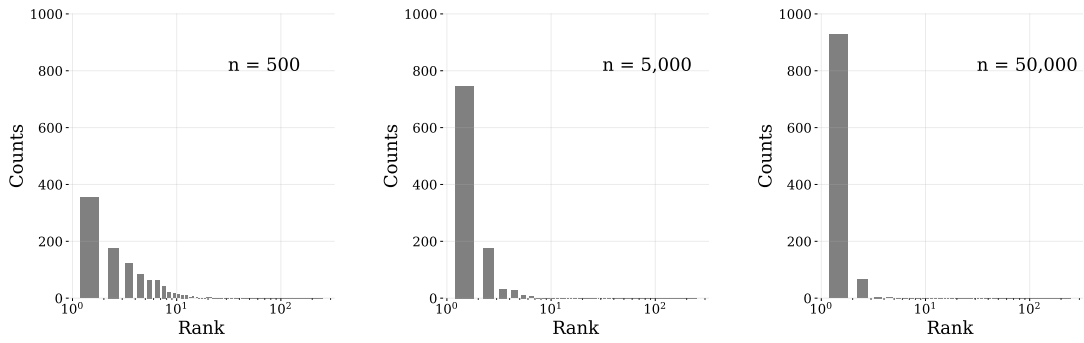
# Directed MAG



## MEC Recovery by Approximate

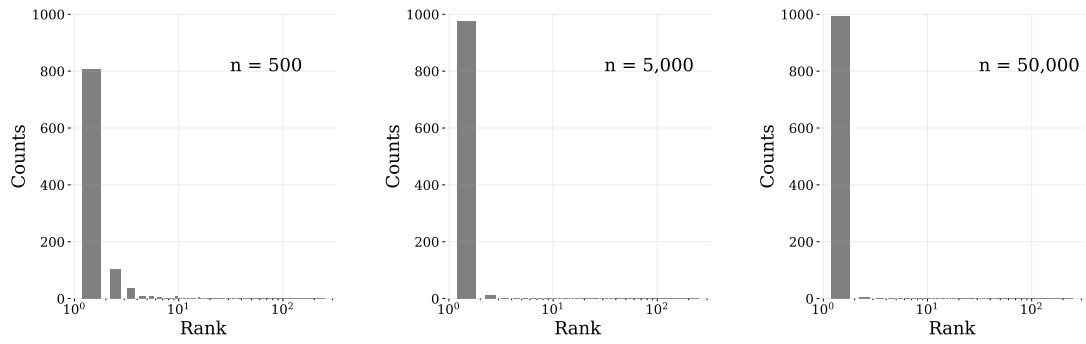


## MEC Recovery by Exact

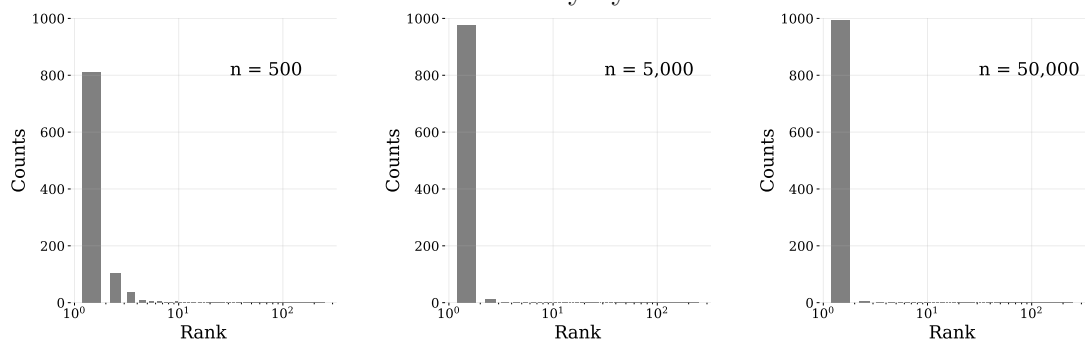


Random Directed MAGs with  $|V| = 4$  and  $|E| \in [0, 3]$

### MEC Recovery by Approximate

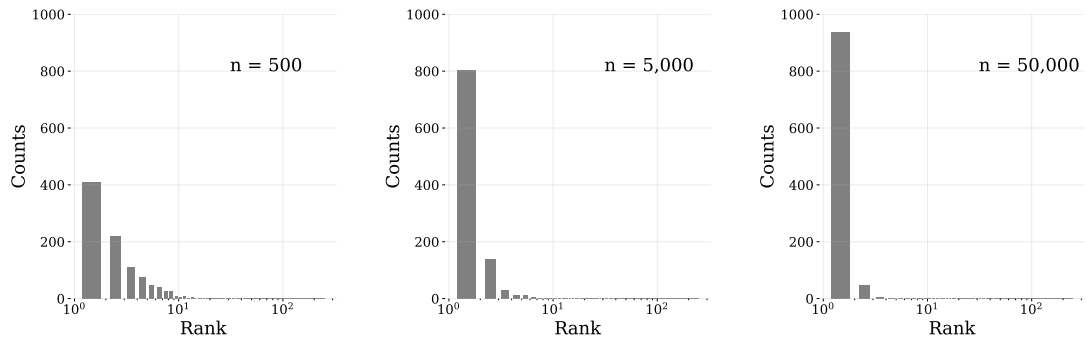


### MEC Recovery by Exact

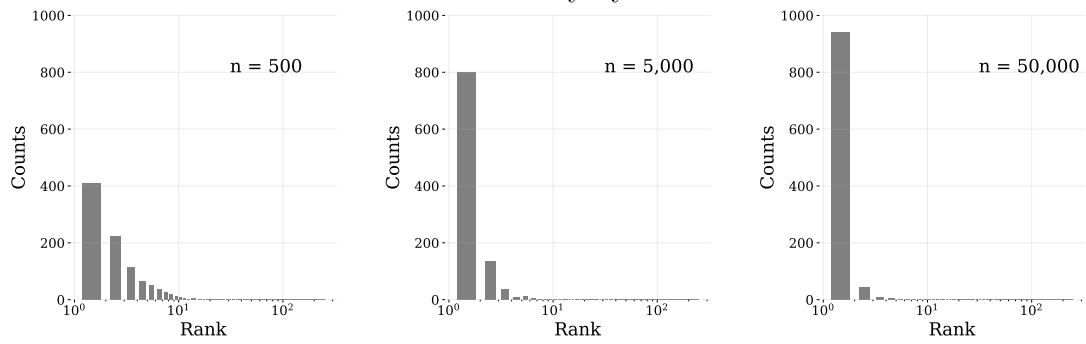


Random Directed MAGs with  $|V| = 4$  and  $|E| \in [4, 6]$

### MEC Recovery by Approximate

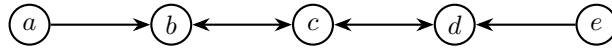


### MEC Recovery by Exact

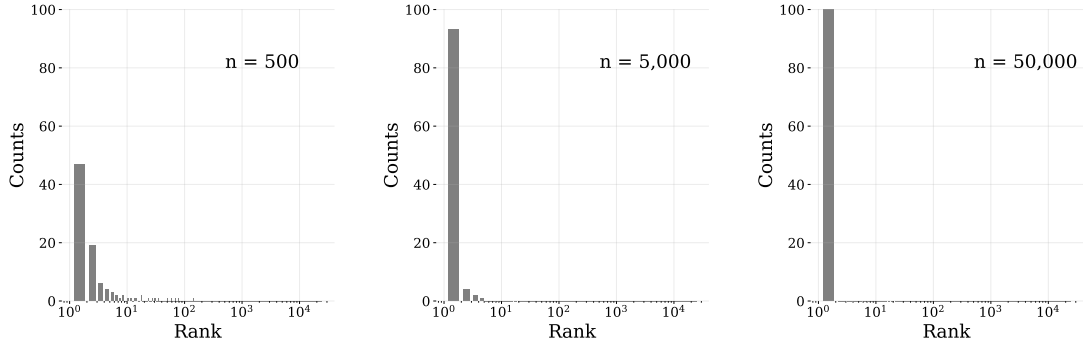




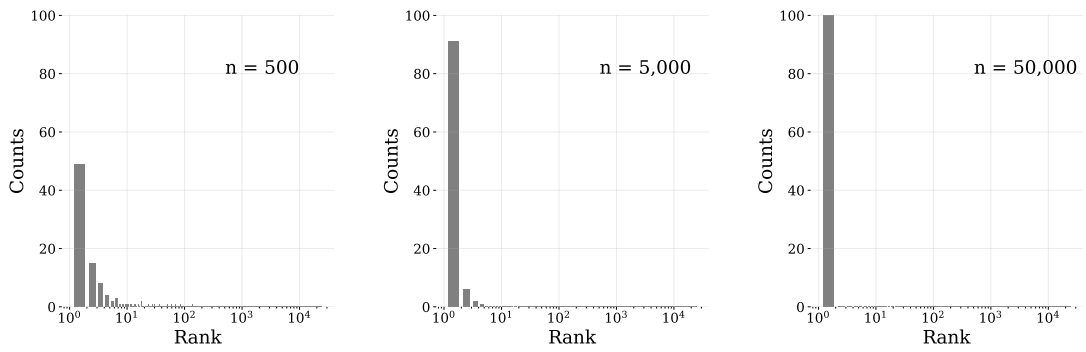
# Directed MAG



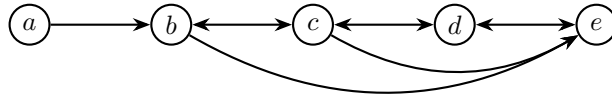
## MEC Recovery by Approximate



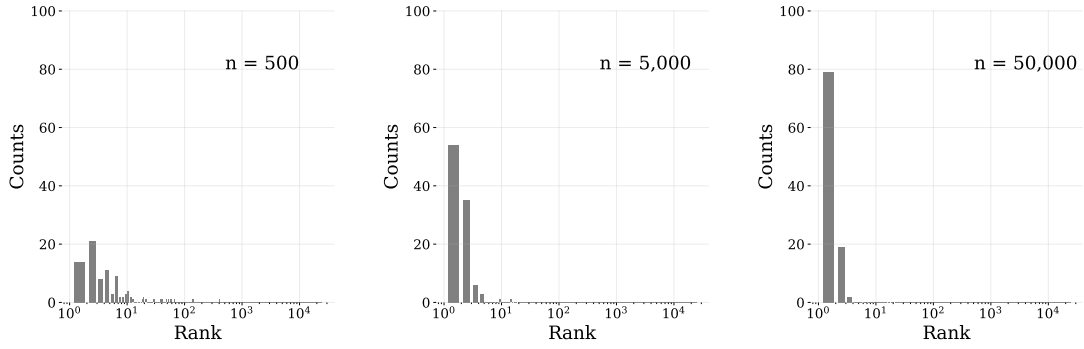
## MEC Recovery by Exact



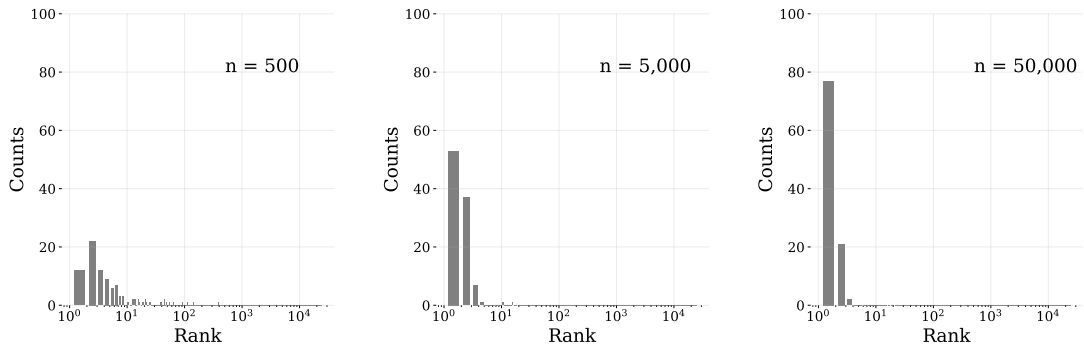
Directed MAG



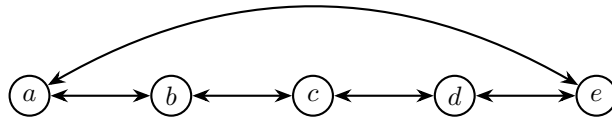
MEC Recovery by Approximate



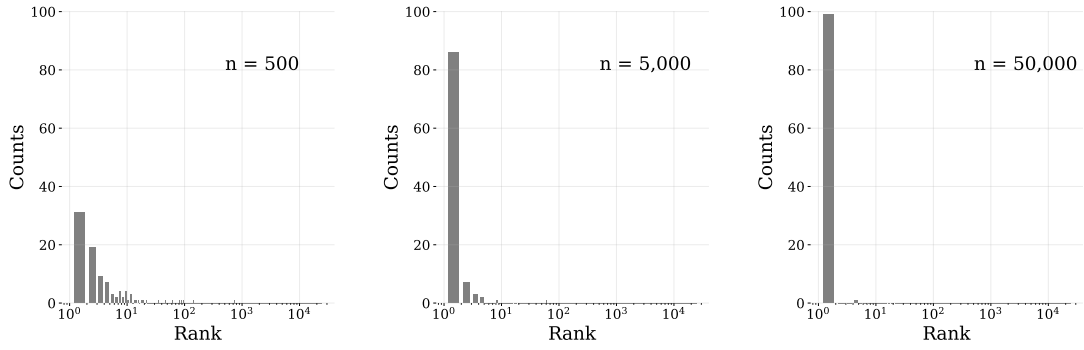
MEC Recovery by Exact



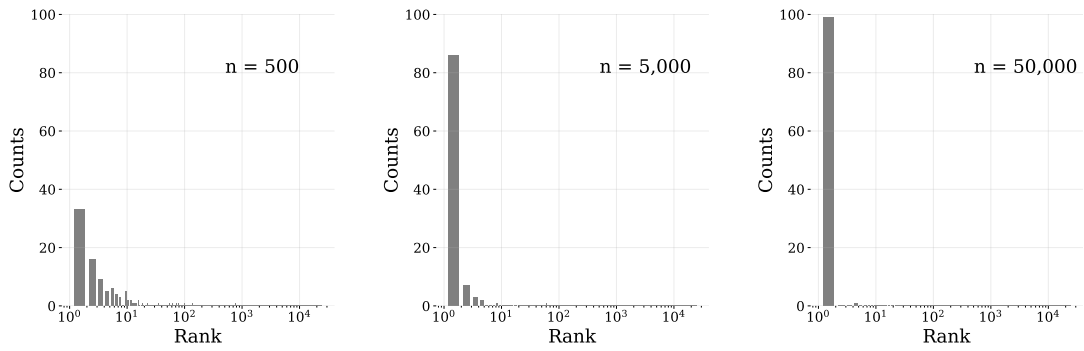
# Directed MAG



## MEC Recovery by Approximate

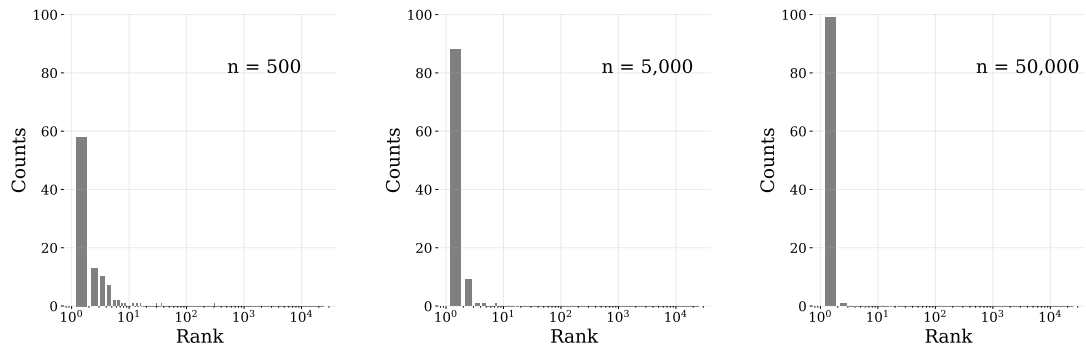


## MEC Recovery by Exact

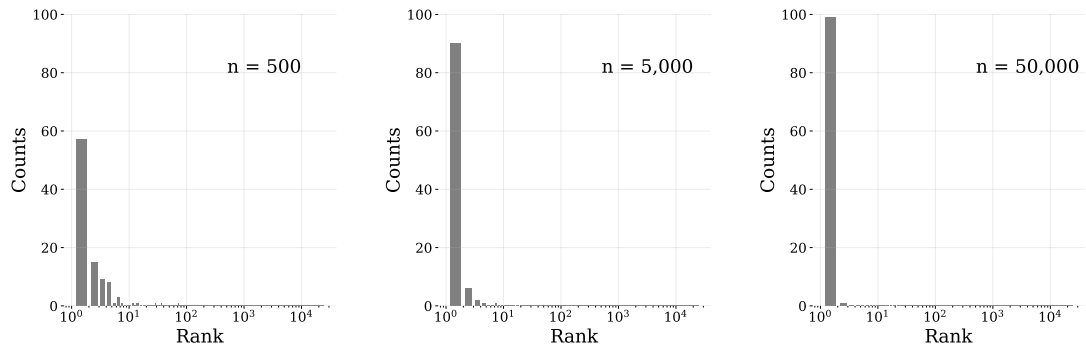


Random Directed MAGs with  $|V| = 5$  and  $|E| \in [0, 5]$

### MEC Recovery by Approximate

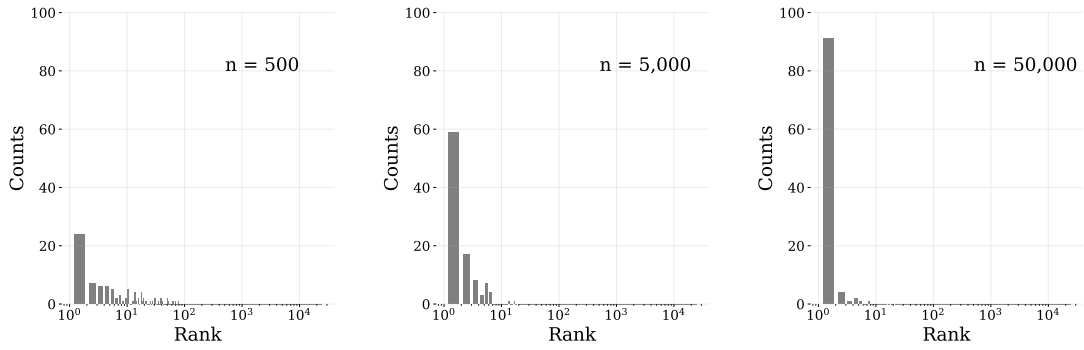


### MEC Recovery by Exact

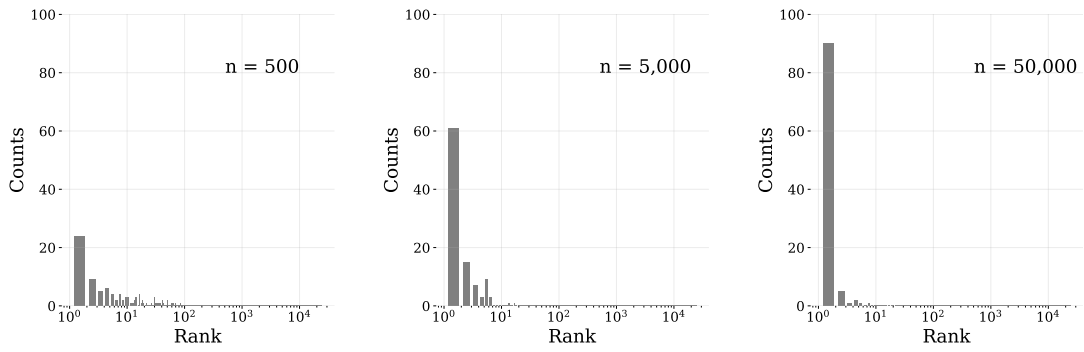


Random Directed MAGs with  $|V| = 5$  and  $|E| \in [6, 10]$

MEC Recovery by Approximate

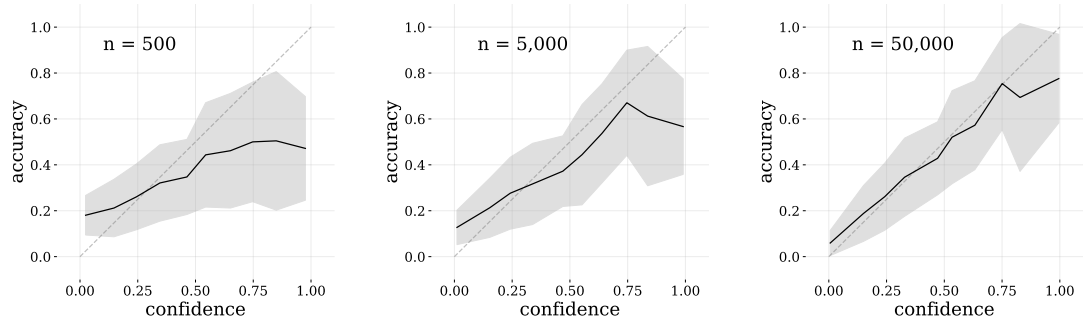


MEC Recovery by Exact

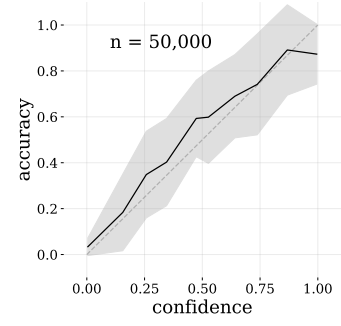
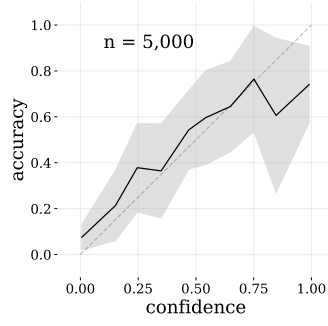
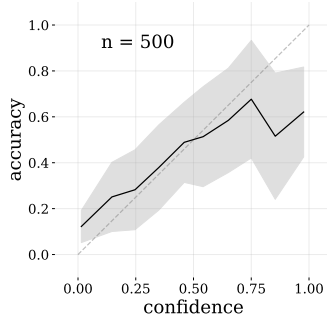


B.8 AP Calibration

Without Knowledge



## With Knowledge



## B.9 Full Airborne Pollutants Tables

| Air Pollutant                        | N   | Pollutant $\rightarrow$ Cardiovascular | Pollutant $\leftrightarrow$ Cardiovascular |
|--------------------------------------|-----|--|--|
| Acetone                              | 517 | 0.15 (0.03, 0.27)                      | 0.064 (0.0, 0.24)                          |
| Acrolein                             | 445 | 0.002 (0.0, 0.01)                      | 0.998 (0.99, 1.0)                          |
| Arsenic                              | 273 | 0.092 (0.0, 0.4)                       | 0.654 (0.04, 1.0)                          |
| Cadmium                              | 265 | 0.291 (0.02, 0.77)                     | 0.342 (0.03, 0.94)                         |
| Carbon Monoxide                      | 572 | 0.931 (0.65, 1.0)                      | 0.069 (0.0, 0.35)                          |
| Chromium                             | 273 | 0.232 (0.09, 0.41)                     | 0.438 (0.13, 0.87)                         |
| Lead PM <sub>10</sub>                | 359 | 0.1 (0.0, 0.63)                        | 0.899 (0.36, 1.0)                          |
| Lead PM <sub>2.5</sub>               | 273 | 0.084 (0.0, 0.35)                      | 0.736 (0.06, 1.0)                          |
| Manganese                            | 273 | 0.251 (0.01, 0.72)                     | 0.392 (0.06, 0.96)                         |
| Methyl Ethyl Ketone                  | 485 | 0.333 (0.16, 0.46)                     | 0.046 (0.0, 0.19)                          |
| Methyl Isobutyl Ketone               | 517 | 0.008 (0.0, 0.07)                      | 0.978 (0.83, 1.0)                          |
| Nickel                               | 273 | 0.212 (0.01, 0.44)                     | 0.443 (0.09, 0.96)                         |
| Nitric Oxide                         | 495 | 0.098 (0.0, 0.55)                      | 0.893 (0.4, 1.0)                           |
| Nitrogen Dioxide                     | 495 | 0.021 (0.0, 0.14)                      | 0.979 (0.86, 1.0)                          |
| Outdoor Temperature                  | 624 | 0.092 (0.0, 0.46)                      | 0.755 (0.07, 1.0)                          |
| Oxides of Nitrogen                   | 495 | 0.028 (0.0, 0.21)                      | 0.972 (0.79, 1.0)                          |
| Ozone                                | 327 | 0.87 (0.45, 1.0)                       | 0.094 (0.0, 0.48)                          |
| Particulate Matter 2.5 $\mu\text{m}$ | 826 | 0.057 (0.0, 0.31)                      | 0.535 (0.03, 1.0)                          |
| Particulate Matter 10 $\mu\text{m}$  | 784 | 0.392 (0.08, 0.86)                     | 0.1 (0.01, 0.34)                           |
| Sulfur Dioxide                       | 687 | 0.101 (0.0, 0.35)                      | 0.894 (0.63, 1.0)                          |

Table B1: Complete airborne pollutants and cardiovascular disease results.

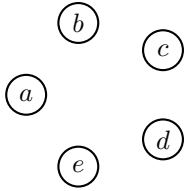
| Air Pollutant             | N   | Pollutant → Respiratory | Pollutant ↔ Respiratory |
|---------------------------|-----|-------------------------|-------------------------|
| Acetone                   | 517 | 0.382 (0.17, 0.5)       | 0.367 (0.12, 0.5)       |
| Acrolein                  | 445 | 0.955 (0.56, 1.0)       | 0.009 (0.0, 0.05)       |
| Arsenic                   | 273 | 0.109 (0.0, 0.36)       | 0.394 (0.02, 0.99)      |
| Cadmium                   | 265 | 0.202 (0.04, 0.44)      | 0.305 (0.03, 0.88)      |
| Carbon Monoxide           | 572 | 0.041 (0.0, 0.26)       | 0.959 (0.74, 1.0)       |
| Chromium                  | 273 | 0.313 (0.11, 0.72)      | 0.196 (0.05, 0.66)      |
| Lead PM <sub>10</sub>     | 359 | 0.52 (0.15, 0.92)       | 0.181 (0.02, 0.72)      |
| Lead PM <sub>2.5</sub>    | 273 | 0.119 (0.0, 0.38)       | 0.285 (0.02, 0.98)      |
| Manganese                 | 273 | 0.267 (0.04, 0.68)      | 0.272 (0.03, 0.87)      |
| Methyl Ethyl Ketone       | 485 | 0.367 (0.17, 0.49)      | 0.325 (0.09, 0.49)      |
| Methyl Isobutyl Ketone    | 517 | 0.146 (0.05, 0.4)       | 0.033 (0.01, 0.16)      |
| Nickel                    | 273 | 0.537 (0.15, 0.93)      | 0.231 (0.03, 0.78)      |
| Nitric Oxide              | 495 | 0.984 (0.88, 1.0)       | 0.003 (0.0, 0.02)       |
| Nitrogen Dioxide          | 495 | 0.997 (0.98, 1.0)       | 0.001 (0.0, 0.01)       |
| Outdoor Temperature       | 624 | 0.987 (0.98, 1.0)       | 0.013 (0.0, 0.02)       |
| Oxides of Nitrogen        | 495 | 0.996 (0.98, 1.0)       | 0.001 (0.0, 0.01)       |
| Ozone                     | 327 | 0.558 (0.27, 0.81)      | 0.094 (0.03, 0.35)      |
| Particulate Matter 2.5 μm | 826 | 0.002 (0.0, 0.02)       | 0.98 (0.74, 1.0)        |
| Particulate Matter 10 μm  | 784 | 0.187 (0.01, 0.57)      | 0.511 (0.08, 0.98)      |
| Sulfur Dioxide            | 687 | 0.592 (0.15, 0.98)      | 0.012 (0.0, 0.03)       |

Table B2: Complete airborne pollutants and respiratory disease results.

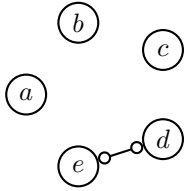


## Appendix C Factorization of Graphs with Five Vertices

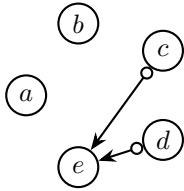
The conditional interaction information terms are similar to the terms intersection terms used to construct  $u_{\mathcal{N}(g)}^{\leq,+}$  and  $u_{\mathcal{N}(g)}^{\leq,-}$  in Algorithm 3. By expanded the conditional interaction information terms we can ensure that they are disjoint—this means that there is no need for an adjustment term in the factorization.



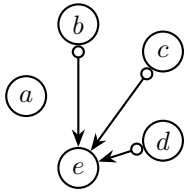
$$\log f(x) = \log f_a(x) + \log f_b(x) + \log f_c(x) + \log f_d(x) + \log f_e(x)$$



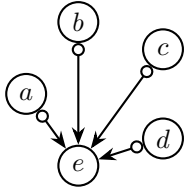
$$\log f(x) = \log f_a(x) + \log f_b(x) + \log f_c(x) + \log f_{e|d}(x) + \log f_d(x)$$



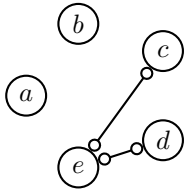
$$\log f(x) = \log f_a(x) + \log f_b(x) + \log f_{e|cd}(x) + \log f_c(x) + \log f_d(x)$$



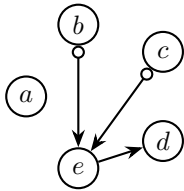
$$\log f(x) = \log f_a(x) + \log f_{e|bcd}(x) + \log f_b(x) + \log f_c(x) + \log f_d(x)$$



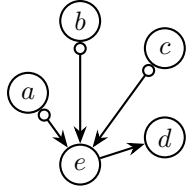
$$\log f(x) = \log f_{e|abcd}(x) + \log f_a(x) + \log f_b(x) + \log f_c(x) + \log f_d(x)$$



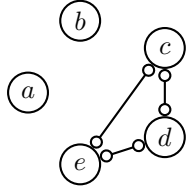
$$\log f(x) = \log f_a(x) + \log f_b(x) + \log f_{d|e}(x) + \log f_{e|c}(x) + \log f_c(x)$$



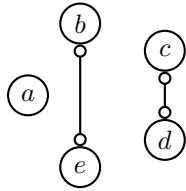
$$\log f(x) = \log f_a(x) + \log f_{d|e}(x) + \log f_{e|bc}(x) + \log f_b(x) + \log f_c(x)$$



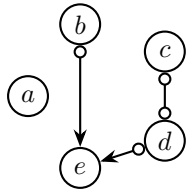
$$\log f(x) = \log f_{d|e}(x) + \log f_{e|abc}(x) + \log f_a(x) + \log f_b(x) + \log f_c(x)$$



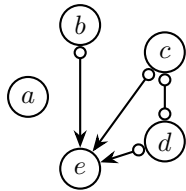
$$\log f(x) = \log f_a(x) + \log f_b(x) + \log f_{e|cd}(x) + \log f_{d|c}(x) + \log f_c(x)$$



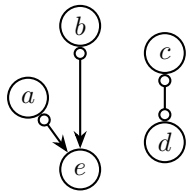
$$\log f(x) = \log f_a(x) + \log f_{d|c}(x) + \log f_c(x) + \log f_{e|b}(x) + \log f_b(x)$$



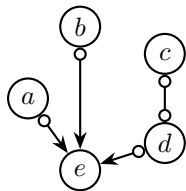
$$\log f(x) = \log f_a(x) + \log f_{e|bd}(x) + \log f_b(x) + \log f_{d|c}(x) + \log f_c(x)$$



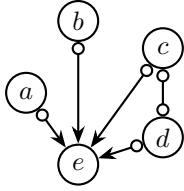
$$\log f(x) = \log f_a(x) + \log f_{e|bcd}(x) + \log f_b(x) + \log f_{d|c}(x) + \log f_c(x)$$



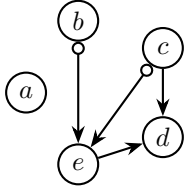
$$\log f(x) = \log f_{d|c}(x) + \log f_c(x) + \log f_{e|ab}(x) + \log f_a(x) + \log f_b(x)$$



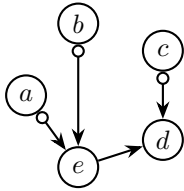
$$\log f(x) = \log f_{e|abd}(x) + \log f_a(x) + \log f_b(x) + \log f_{d|c}(x) + \log f_c(x)$$



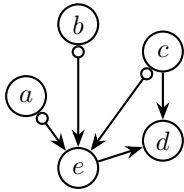
$$\log f(x) = \log f_{e|abcd}(x) + \log f_a(x) + \log f_b(x) + \log f_{d|c}(x) + \log f_c(x)$$



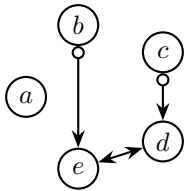
$$\log f(x) = \log f_a(x) + \log f_{d|ce}(x) + \log f_{e|bc}(x) + \log f_b(x) + \log f_c(x)$$



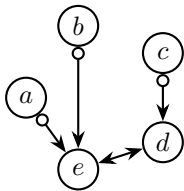
$$\log f(x) = \log f_{d|ce}(x) + \log f_c(x) + \log f_{e|ab}(x) + \log f_a(x) + \log f_b(x)$$



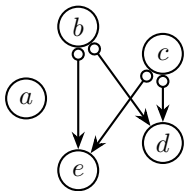
$$\log f(x) = \log f_{d|ce}(x) + \log f_{e|abc}(x) + \log f_a(x) + \log f_b(x) + \log f_c(x)$$



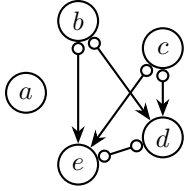
$$\log f(x) = \log f_a(x) + \log f_{d|bce}(x) + \log f_c(x) + \log f_{e|b}(x) + \log f_b(x) - \phi_{b,d|c}(x)$$



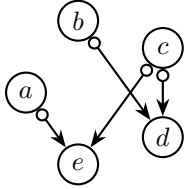
$$\log f(x) = \log f_{d|abce}(x) + \log f_c(x) + \log f_{e|ab}(x) + \log f_a(x) + \log f_b(x) - \phi_{ab,d|c}(x)$$



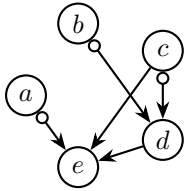
$$\log f(x) = \log f_a(x) + \log f_{d|bc}(x) + \log f_{e|bc}(x) + \log f_b(x) + \log f_c(x)$$



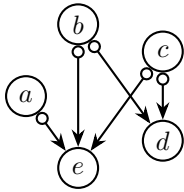
$$\log f(x) = \log f_a(x) + \log f_{e|bcd}(x) + \log f_{d|bc}(x) + \log f_b(x) + \log f_c(x)$$



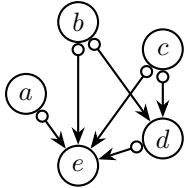
$$\log f(x) = \log f_{d|bc}(x) + \log f_b(x) + \log f_{e|ac}(x) + \log f_a(x) + \log f_c(x)$$



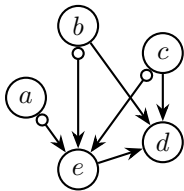
$$\log f(x) = \log f_{e|acd}(x) + \log f_a(x) + \log f_{d|bc}(x) + \log f_b(x) + \log f_c(x)$$



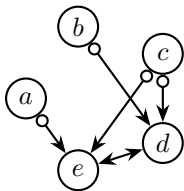
$$\log f(x) = \log f_{d|bc}(x) + \log f_{e|abc}(x) + \log f_a(x) + \log f_b(x) + \log f_c(x)$$



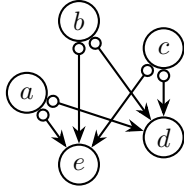
$$\log f(x) = \log f_{e|abcd}(x) + \log f_a(x) + \log f_{d|bc}(x) + \log f_b(x) + \log f_c(x)$$



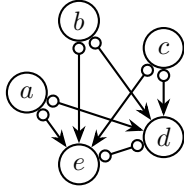
$$\log f(x) = \log f_{d|bce}(x) + \log f_{e|abc}(x) + \log f_a(x) + \log f_b(x) + \log f_c(x)$$



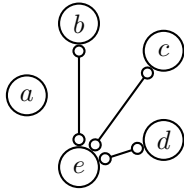
$$\log f(x) = \log f_{d|abce}(x) + \log f_b(x) + \log f_{e|ac}(x) + \log f_a(x) + \log f_c(x) - \phi_{a,d|bc}(x)$$



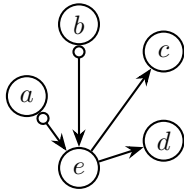
$$\log f(x) = \log f_{d|abc}(x) + \log f_{e|abc}(x) + \log f_a(x) + \log f_b(x) + \log f_c(x)$$



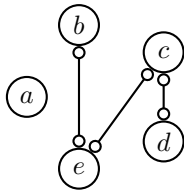
$$\log f(x) = \log f_{e|abcd}(x) + \log f_{d|abc}(x) + \log f_a(x) + \log f_b(x) + \log f_c(x)$$



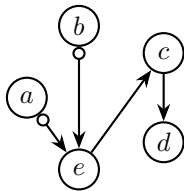
$$\log f(x) = \log f_a(x) + \log f_{c|e}(x) + \log f_{d|e}(x) + \log f_{e|b}(x) + \log f_b(x)$$



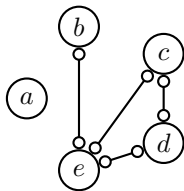
$$\log f(x) = \log f_{c|e}(x) + \log f_{d|e}(x) + \log f_{e|ab}(x) + \log f_a(x) + \log f_b(x)$$



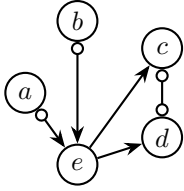
$$\log f(x) = \log f_a(x) + \log f_{d|c}(x) + \log f_{c|e}(x) + \log f_{e|b}(x) + \log f_b(x)$$



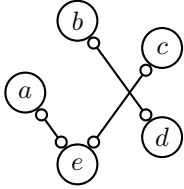
$$\log f(x) = \log f_{d|c}(x) + \log f_{c|e}(x) + \log f_{e|ab}(x) + \log f_a(x) + \log f_b(x)$$



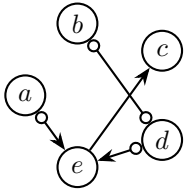
$$\log f(x) = \log f_a(x) + \log f_{d|ce}(x) + \log f_{c|e}(x) + \log f_{e|b}(x) + \log f_b(x)$$



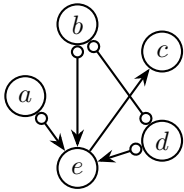
$$\log f(x) = \log f_{d|ce}(x) + \log f_{c|e}(x) + \log f_{e|ab}(x) + \log f_a(x) + \log f_b(x)$$



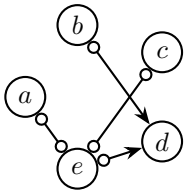
$$\log f(x) = \log f_{c|e}(x) + \log f_{d|b}(x) + \log f_b(x) + \log f_{e|a}(x) + \log f_a(x)$$



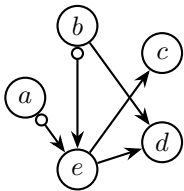
$$\log f(x) = \log f_{c|e}(x) + \log f_{e|ad}(x) + \log f_a(x) + \log f_{d|b}(x) + \log f_b(x)$$



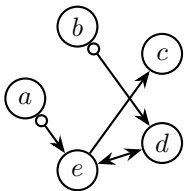
$$\log f(x) = \log f_{c|e}(x) + \log f_{e|abd}(x) + \log f_a(x) + \log f_{d|b}(x) + \log f_b(x)$$



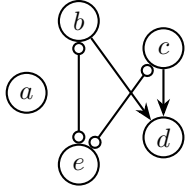
$$\log f(x) = \log f_{c|e}(x) + \log f_{d|be}(x) + \log f_b(x) + \log f_{e|a}(x) + \log f_a(x)$$



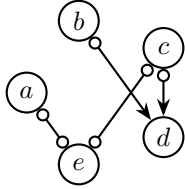
$$\log f(x) = \log f_{c|e}(x) + \log f_{d|be}(x) + \log f_{e|ab}(x) + \log f_a(x) + \log f_b(x)$$



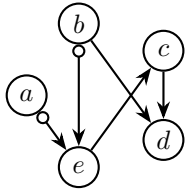
$$\log f(x) = \log f_{c|e}(x) + \log f_{d|abe}(x) + \log f_b(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{a,d|b}(x)$$



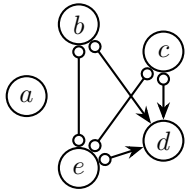
$$\log f(x) = \log f_a(x) + \log f_{d|bc}(x) + \log f_{c|e}(x) + \log f_{e|b}(x) + \log f_b(x)$$



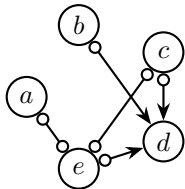
$$\log f(x) = \log f_{d|bc}(x) + \log f_b(x) + \log f_{c|e}(x) + \log f_{e|a}(x) + \log f_a(x)$$



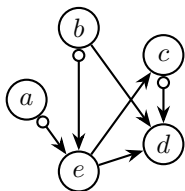
$$\log f(x) = \log f_{d|bc}(x) + \log f_{c|e}(x) + \log f_{e|ab}(x) + \log f_a(x) + \log f_b(x)$$



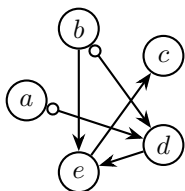
$$\log f(x) = \log f_a(x) + \log f_{d|bce}(x) + \log f_{c|e}(x) + \log f_{e|b}(x) + \log f_b(x)$$



$$\log f(x) = \log f_{d|bce}(x) + \log f_b(x) + \log f_{c|e}(x) + \log f_{e|a}(x) + \log f_a(x)$$

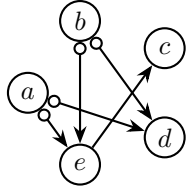


$$\log f(x) = \log f_{d|bce}(x) + \log f_{c|e}(x) + \log f_{e|ab}(x) + \log f_a(x) + \log f_b(x)$$

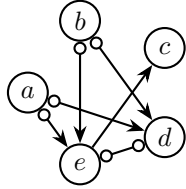


$$\log f(x) = \log f_{c|e}(x) + \log f_{e|bd}(x) + \log f_{d|ab}(x) + \log f_a(x) + \log f_b(x)$$

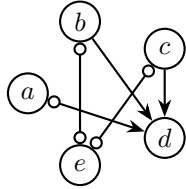




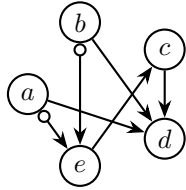
$$\log f(x) = \log f_{c|e}(x) + \log f_{d|ab}(x) + \log f_{e|ab}(x) + \log f_a(x) + \log f_b(x)$$



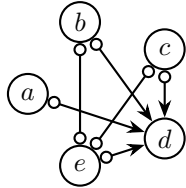
$$\log f(x) = \log f_{c|e}(x) + \log f_{e|abd}(x) + \log f_{d|ab}(x) + \log f_a(x) + \log f_b(x)$$



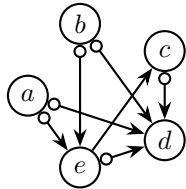
$$\log f(x) = \log f_{d|abc}(x) + \log f_a(x) + \log f_{c|e}(x) + \log f_{e|b}(x) + \log f_b(x)$$



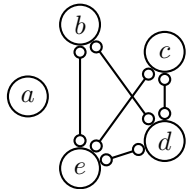
$$\log f(x) = \log f_{d|abc}(x) + \log f_{c|e}(x) + \log f_{e|ab}(x) + \log f_a(x) + \log f_b(x)$$



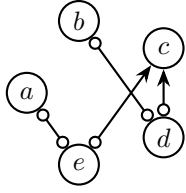
$$\log f(x) = \log f_{d|abce}(x) + \log f_a(x) + \log f_{c|e}(x) + \log f_{e|b}(x) + \log f_b(x)$$



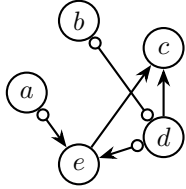
$$\log f(x) = \log f_{d|abce}(x) + \log f_{c|e}(x) + \log f_{e|ab}(x) + \log f_a(x) + \log f_b(x)$$



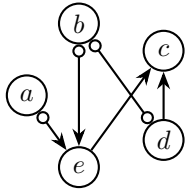
$$\log f(x) = \log f_a(x) + \log f_{c|de}(x) + \log f_{e|bd}(x) + \log f_{d|b}(x) + \log f_b(x)$$



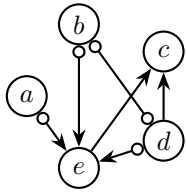
$$\log f(x) = \log f_{c|de}(x) + \log f_{d|b}(x) + \log f_b(x) + \log f_{e|a}(x) + \log f_a(x)$$



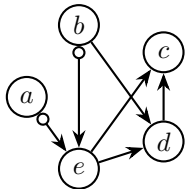
$$\log f(x) = \log f_{c|de}(x) + \log f_{e|ad}(x) + \log f_a(x) + \log f_{d|b}(x) + \log f_b(x)$$



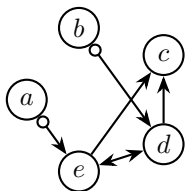
$$\log f(x) = \log f_{c|de}(x) + \log f_{d|b}(x) + \log f_{e|ab}(x) + \log f_a(x) + \log f_b(x)$$



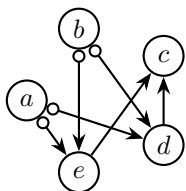
$$\log f(x) = \log f_{c|de}(x) + \log f_{e|abd}(x) + \log f_a(x) + \log f_{d|b}(x) + \log f_b(x)$$



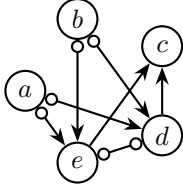
$$\log f(x) = \log f_{c|de}(x) + \log f_{d|be}(x) + \log f_{e|ab}(x) + \log f_a(x) + \log f_b(x)$$



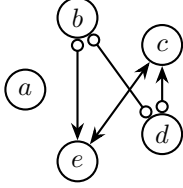
$$\log f(x) = \log f_{c|de}(x) + \log f_{d|abe}(x) + \log f_b(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{a,d|b}(x)$$



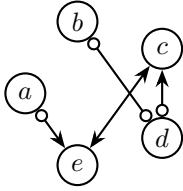
$$\log f(x) = \log f_{c|de}(x) + \log f_{d|ab}(x) + \log f_{e|ab}(x) + \log f_a(x) + \log f_b(x)$$



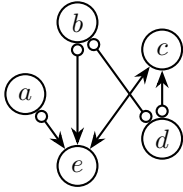
$$\log f(x) = \log f_{c|de}(x) + \log f_{e|abd}(x) + \log f_{d|ab}(x) + \log f_a(x) + \log f_b(x)$$



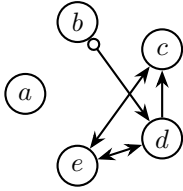
$$\log f(x) = \log f_a(x) + \log f_{c|bde}(x) + \log f_{d|b}(x) + \log f_{e|b}(x) + \log f_b(x) - \phi_{b,c|d}(x)$$



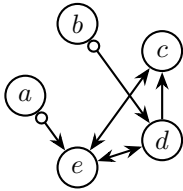
$$\log f(x) = \log f_{c|ade}(x) + \log f_{d|b}(x) + \log f_b(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{a,c|d}(x)$$



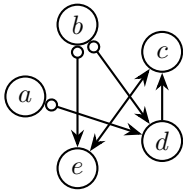
$$\log f(x) = \log f_{c|abde}(x) + \log f_{d|b}(x) + \log f_{e|ab}(x) + \log f_a(x) + \log f_b(x) - \phi_{ab,c|d}(x)$$



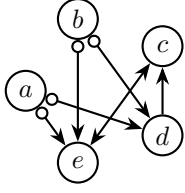
$$\log f(x) = \log f_a(x) + \log f_{c|bde}(x) + \log f_{d|be}(x) + \log f_b(x) + \log f_e(x) - \phi_{b,c|d}(x)$$



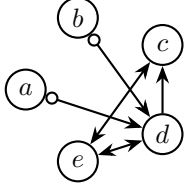
$$\log f(x) = \log f_{c|abde}(x) + \log f_{d|abe}(x) + \log f_b(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{ab,c|d}(x) - \phi_{a,d|b}(x)$$



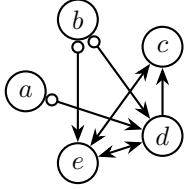
$$\log f(x) = \log f_{c|bde}(x) + \log f_{d|ab}(x) + \log f_a(x) + \log f_{e|b}(x) + \log f_b(x) - \phi_{b,c|d}(x)$$



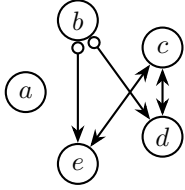
$$\log f(x) = \log f_{c|abde}(x) + \log f_{d|ab}(x) + \log f_{e|ab}(x) + \log f_a(x) + \log f_b(x) - \phi_{ab,c|d}(x)$$



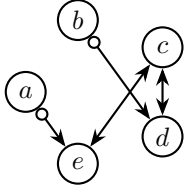
$$\log f(x) = \log f_{c|abde}(x) + \log f_{d|abe}(x) + \log f_a(x) + \log f_b(x) + \log f_e(x) - \phi_{ab,c|d}(x)$$



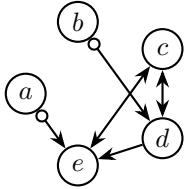
$$\log f(x) = \log f_{c|abde}(x) + \log f_{d|abe}(x) + \log f_a(x) + \log f_{e|b}(x) + \log f_b(x) - \phi_{ab,c|d}(x)$$



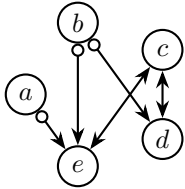
$$\log f(x) = \log f_a(x) + \log f_{c|bde}(x) + \log f_{d|b}(x) + \log f_{e|b}(x) + \log f_b(x) - \phi_{b,c}(x)$$



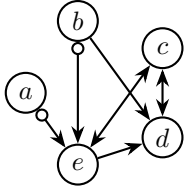
$$\log f(x) = \log f_{de|abc}(x) + \log f_a(x) + \log f_b(x) + \log f_c(x) - \phi_{a,d|bc}(x) - \phi_{b,e|ac}(x) - \phi_{d,e|ab}(x)$$



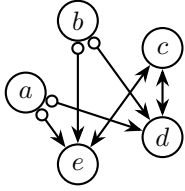
$$\log f(x) = \log f_{c|abde}(x) + \log f_{e|ad}(x) + \log f_a(x) + \log f_{d|b}(x) + \log f_b(x) - \phi_{a,c|bd}(x) - \phi_{b,c}(x)$$



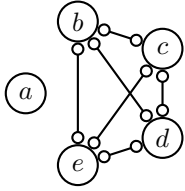
$$\log f(x) = \log f_{c|abde}(x) + \log f_{d|b}(x) + \log f_{e|ab}(x) + \log f_a(x) + \log f_b(x) - \phi_{a,c|bd}(x) - \phi_{b,c}(x)$$



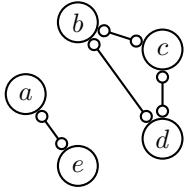
$$\log f(x) = \log f_{c|abde}(x) + \log f_{d|be}(x) + \log f_{e|ab}(x) + \log f_a(x) + \log f_b(x) - \phi_{ab,c}(x)$$



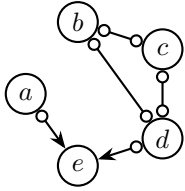
$$\log f(x) = \log f_{c|abde}(x) + \log f_{d|ab}(x) + \log f_{e|ab}(x) + \log f_a(x) + \log f_b(x) - \phi_{ab,c}(x)$$



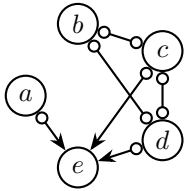
$$\log f(x) = \log f_a(x) + \log f_{e|bcd}(x) + \log f_{d|bc}(x) + \log f_{c|b}(x) + \log f_b(x)$$



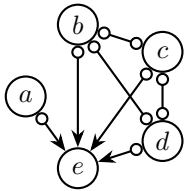
$$\log f(x) = \log f_{d|bc}(x) + \log f_{c|b}(x) + \log f_b(x) + \log f_{e|a}(x) + \log f_a(x)$$



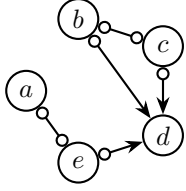
$$\log f(x) = \log f_{e|ad}(x) + \log f_a(x) + \log f_{d|bc}(x) + \log f_{c|b}(x) + \log f_b(x)$$



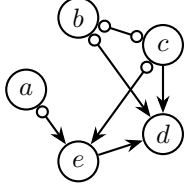
$$\log f(x) = \log f_{e|acd}(x) + \log f_a(x) + \log f_{d|bc}(x) + \log f_{c|b}(x) + \log f_b(x)$$



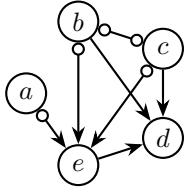
$$\log f(x) = \log f_{e|abcd}(x) + \log f_a(x) + \log f_{d|bc}(x) + \log f_{c|b}(x) + \log f_b(x)$$



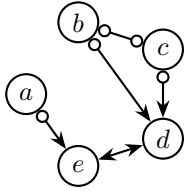
$$\log f(x) = \log f_{d|bce}(x) + \log f_{c|b}(x) + \log f_b(x) + \log f_{e|a}(x) + \log f_a(x)$$



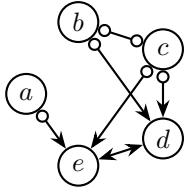
$$\log f(x) = \log f_{d|bce}(x) + \log f_{e|ac}(x) + \log f_a(x) + \log f_{c|b}(x) + \log f_b(x)$$



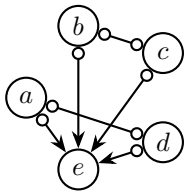
$$\log f(x) = \log f_{d|bce}(x) + \log f_{e|abc}(x) + \log f_a(x) + \log f_{c|b}(x) + \log f_b(x)$$



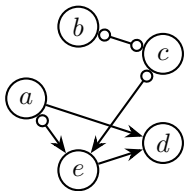
$$\log f(x) = \log f_{d|abce}(x) + \log f_{c|b}(x) + \log f_b(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{a,d|bc}(x)$$



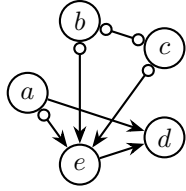
$$\log f(x) = \log f_{d|abce}(x) + \log f_{e|ac}(x) + \log f_a(x) + \log f_{c|b}(x) + \log f_b(x) - \phi_{a,d|bc}(x)$$



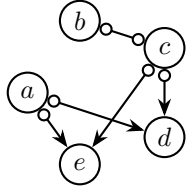
$$\log f(x) = \log f_{e|abcd}(x) + \log f_{c|b}(x) + \log f_b(x) + \log f_{d|a}(x) + \log f_a(x)$$



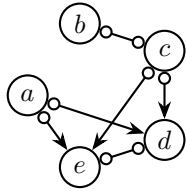
$$\log f(x) = \log f_{d|ae}(x) + \log f_{e|ac}(x) + \log f_a(x) + \log f_{c|b}(x) + \log f_b(x)$$



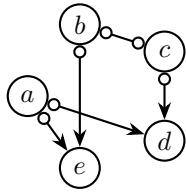
$$\log f(x) = \log f_{d|ae}(x) + \log f_{e|abc}(x) + \log f_a(x) + \log f_{c|b}(x) + \log f_b(x)$$



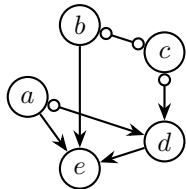
$$\log f(x) = \log f_{d|ac}(x) + \log f_{e|ac}(x) + \log f_a(x) + \log f_{c|b}(x) + \log f_b(x)$$



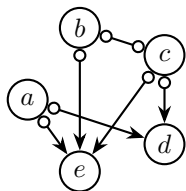
$$\log f(x) = \log f_{e|acd}(x) + \log f_{d|ac}(x) + \log f_a(x) + \log f_{c|b}(x) + \log f_b(x)$$



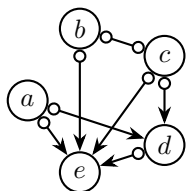
$$\log f(x) = \log f_{d|ac}(x) + \log f_{c|b}(x) + \log f_{e|ab}(x) + \log f_a(x) + \log f_b(x)$$



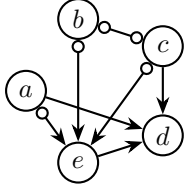
$$\log f(x) = \log f_{e|abd}(x) + \log f_{d|ac}(x) + \log f_a(x) + \log f_{c|b}(x) + \log f_b(x)$$



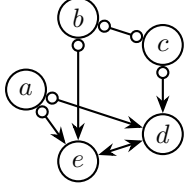
$$\log f(x) = \log f_{d|ac}(x) + \log f_{e|abc}(x) + \log f_a(x) + \log f_{c|b}(x) + \log f_b(x)$$



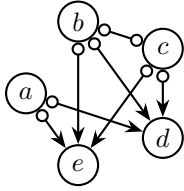
$$\log f(x) = \log f_{e|abcd}(x) + \log f_{d|ac}(x) + \log f_a(x) + \log f_{c|b}(x) + \log f_b(x)$$



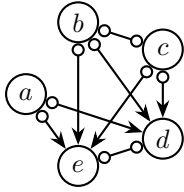
$$\log f(x) = \log f_{d|ace}(x) + \log f_{e|abc}(x) + \log f_a(x) + \log f_{c|b}(x) + \log f_b(x)$$



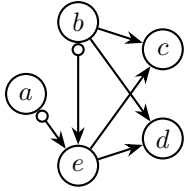
$$\log f(x) = \log f_{d|abce}(x) + \log f_{c|b}(x) + \log f_{e|ab}(x) + \log f_a(x) + \log f_b(x) - \phi_{b,d|ac}(x)$$



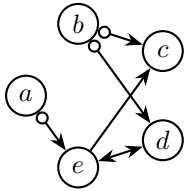
$$\log f(x) = \log f_{d|abc}(x) + \log f_{e|abc}(x) + \log f_a(x) + \log f_{c|b}(x) + \log f_b(x)$$



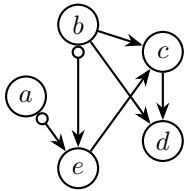
$$\log f(x) = \log f_{e|abcd}(x) + \log f_{d|abc}(x) + \log f_a(x) + \log f_{c|b}(x) + \log f_b(x)$$



$$\log f(x) = \log f_{c|be}(x) + \log f_{d|be}(x) + \log f_{e|ab}(x) + \log f_a(x) + \log f_b(x)$$

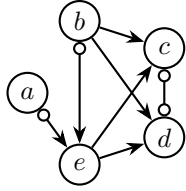


$$\log f(x) = \log f_{c|be}(x) + \log f_{d|abe}(x) + \log f_b(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{a,d|b}(x)$$

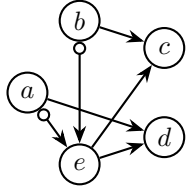


$$\log f(x) = \log f_{d|bc}(x) + \log f_{c|be}(x) + \log f_{e|ab}(x) + \log f_a(x) + \log f_b(x)$$

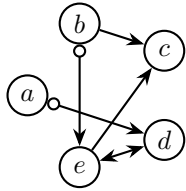




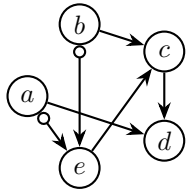
$$\log f(x) = \log f_{d|bce}(x) + \log f_{c|be}(x) + \log f_{e|ab}(x) + \log f_a(x) + \log f_b(x)$$



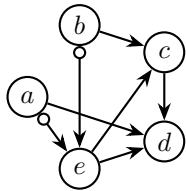
$$\log f(x) = \log f_{c|be}(x) + \log f_{d|ae}(x) + \log f_{e|ab}(x) + \log f_a(x) + \log f_b(x)$$



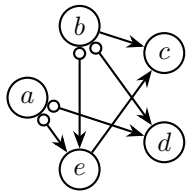
$$\log f(x) = \log f_{c|be}(x) + \log f_{d|abe}(x) + \log f_a(x) + \log f_{e|b}(x) + \log f_b(x) - \phi_{b,d|a}(x)$$



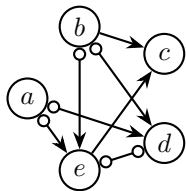
$$\log f(x) = \log f_{d|ac}(x) + \log f_{c|be}(x) + \log f_{e|ab}(x) + \log f_a(x) + \log f_b(x)$$



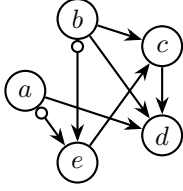
$$\log f(x) = \log f_{d|ace}(x) + \log f_{c|be}(x) + \log f_{e|ab}(x) + \log f_a(x) + \log f_b(x)$$



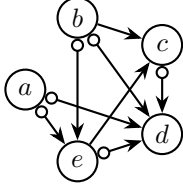
$$\log f(x) = \log f_{c|be}(x) + \log f_{d|ab}(x) + \log f_{e|ab}(x) + \log f_a(x) + \log f_b(x)$$



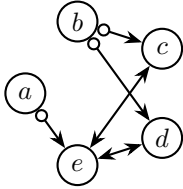
$$\log f(x) = \log f_{c|be}(x) + \log f_{e|abd}(x) + \log f_{d|ab}(x) + \log f_a(x) + \log f_b(x)$$



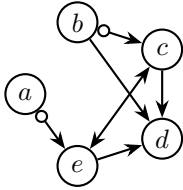
$$\log f(x) = \log f_{d|abc}(x) + \log f_{c|be}(x) + \log f_{e|ab}(x) + \log f_a(x) + \log f_b(x)$$



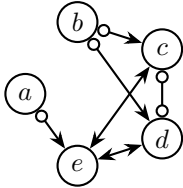
$$\log f(x) = \log f_{d|abce}(x) + \log f_{c|be}(x) + \log f_{e|ab}(x) + \log f_a(x) + \log f_b(x)$$



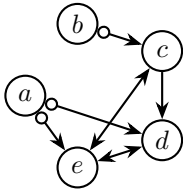
$$\log f(x) = \log f_{c|abde}(x) + \log f_{d|abe}(x) + \log f_b(x) + \log f_{e|a}(x) + \log f_a(x) \\ - \phi_{ad,c|b}(x) - \phi_{a,d|b}(x)$$



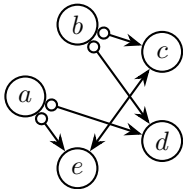
$$\log f(x) = \log f_{d|bce}(x) + \log f_{c|abe}(x) + \log f_b(x) + \log f_{e|a}(x) + \log f_a(x) \\ - \phi_{a,c|b}(x)$$



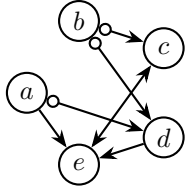
$$\log f(x) = \log f_{d|abce}(x) + \log f_{c|abe}(x) + \log f_b(x) + \log f_{e|a}(x) + \log f_a(x) \\ - \phi_{a,d|bc}(x) - \phi_{a,c|b}(x)$$



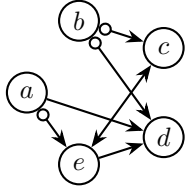
$$\log f(x) = \log f_{d|abce}(x) + \log f_{c|abe}(x) + \log f_b(x) + \log f_{e|a}(x) + \log f_a(x) \\ - \phi_{b,d|ac}(x) - \phi_{a,c|b}(x)$$



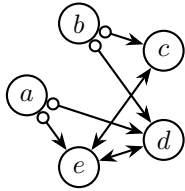
$$\log f(x) = \log f_{c|abe}(x) + \log f_{d|ab}(x) + \log f_b(x) + \log f_{e|a}(x) + \log f_a(x) \\ - \phi_{a,c|b}(x)$$



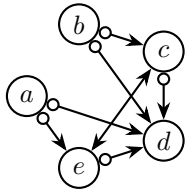
$$\log f(x) = \log f_{c|abde}(x) + \log f_{e|ad}(x) + \log f_{d|ab}(x) + \log f_a(x) + \log f_b(x) - \phi_{ad,c|b}(x)$$



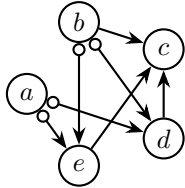
$$\log f(x) = \log f_{c|abe}(x) + \log f_{d|abe}(x) + \log f_b(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{a,c|b}(x)$$



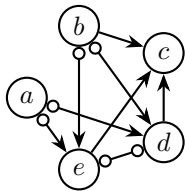
$$\log f(x) = \log f_{c|abde}(x) + \log f_{d|abe}(x) + \log f_b(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{ad,c|b}(x)$$



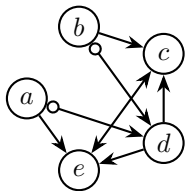
$$\log f(x) = \log f_{d|abce}(x) + \log f_{c|abe}(x) + \log f_b(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{a,c|b}(x)$$



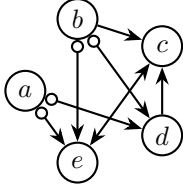
$$\log f(x) = \log f_{c|bde}(x) + \log f_{d|ab}(x) + \log f_{e|ab}(x) + \log f_a(x) + \log f_b(x)$$



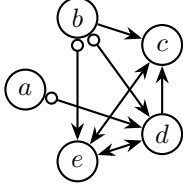
$$\log f(x) = \log f_{c|bde}(x) + \log f_{e|abd}(x) + \log f_{d|ab}(x) + \log f_a(x) + \log f_b(x)$$



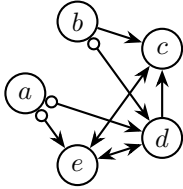
$$\log f(x) = \log f_{c|abde}(x) + \log f_{e|ad}(x) + \log f_{d|ab}(x) + \log f_a(x) + \log f_b(x) - \phi_{a,c|bd}(x)$$



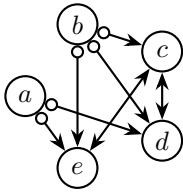
$$\log f(x) = \log f_{c|abde}(x) + \log f_{d|ab}(x) + \log f_{e|ab}(x) + \log f_a(x) + \log f_b(x) - \phi_{a,c|bd}(x)$$



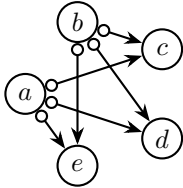
$$\log f(x) = \log f_{c|abde}(x) + \log f_{d|abe}(x) + \log f_a(x) + \log f_{e|b}(x) + \log f_b(x) - \phi_{a,c|bd}(x)$$



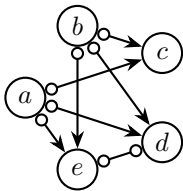
$$\log f(x) = \log f_{c|abde}(x) + \log f_{d|abe}(x) + \log f_b(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{a,c|bd}(x)$$



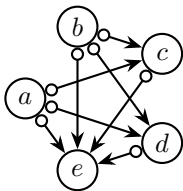
$$\log f(x) = \log f_{c|abde}(x) + \log f_{d|ab}(x) + \log f_{e|ab}(x) + \log f_a(x) + \log f_b(x) - \phi_{a,c|b}(x)$$



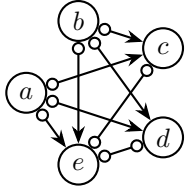
$$\log f(x) = \log f_{c|ab}(x) + \log f_{d|ab}(x) + \log f_{e|ab}(x) + \log f_a(x) + \log f_b(x)$$



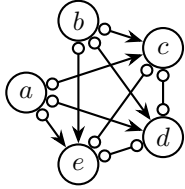
$$\log f(x) = \log f_{c|ab}(x) + \log f_{e|abd}(x) + \log f_{d|ab}(x) + \log f_a(x) + \log f_b(x)$$



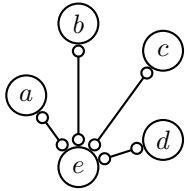
$$\log f(x) = \log f_{e|abcd}(x) + \log f_{c|ab}(x) + \log f_{d|ab}(x) + \log f_a(x) + \log f_b(x)$$



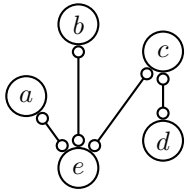
$$\log f(x) = \log f_{d|abe}(x) + \log f_{e|abc}(x) + \log f_{c|ab}(x) + \log f_a(x) + \log f_b(x)$$



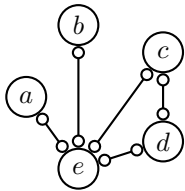
$$\log f(x) = \log f_{e|abcd}(x) + \log f_{d|abc}(x) + \log f_{c|ab}(x) + \log f_a(x) + \log f_b(x)$$



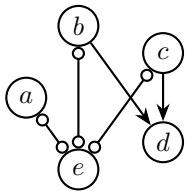
$$\log f(x) = \log f_{b|e}(x) + \log f_{c|e}(x) + \log f_{d|e}(x) + \log f_{e|a}(x) + \log f_a(x)$$



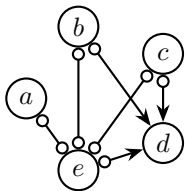
$$\log f(x) = \log f_{b|e}(x) + \log f_{d|c}(x) + \log f_{c|e}(x) + \log f_{e|a}(x) + \log f_a(x)$$



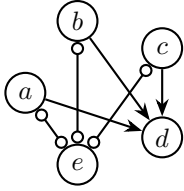
$$\log f(x) = \log f_{b|e}(x) + \log f_{d|ce}(x) + \log f_{c|e}(x) + \log f_{e|a}(x) + \log f_a(x)$$



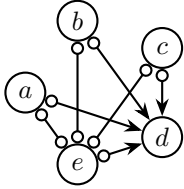
$$\log f(x) = \log f_{d|bc}(x) + \log f_{b|e}(x) + \log f_{c|e}(x) + \log f_{e|a}(x) + \log f_a(x)$$



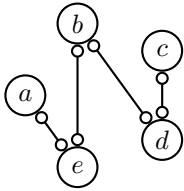
$$\log f(x) = \log f_{d|bce}(x) + \log f_{b|e}(x) + \log f_{c|e}(x) + \log f_{e|a}(x) + \log f_a(x)$$



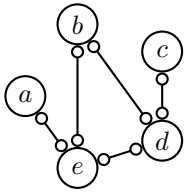
$$\log f(x) = \log f_{d|abc}(x) + \log f_{b|e}(x) + \log f_{c|e}(x) + \log f_{e|a}(x) + \log f_a(x)$$



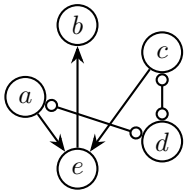
$$\log f(x) = \log f_{d|abce}(x) + \log f_{b|e}(x) + \log f_{c|e}(x) + \log f_{e|a}(x) + \log f_a(x)$$



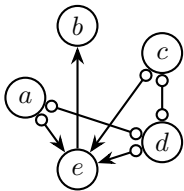
$$\log f(x) = \log f_{c|d}(x) + \log f_{d|b}(x) + \log f_{b|e}(x) + \log f_{e|a}(x) + \log f_a(x)$$



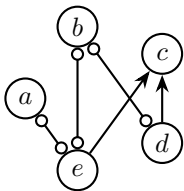
$$\log f(x) = \log f_{c|d}(x) + \log f_{d|be}(x) + \log f_{b|e}(x) + \log f_{e|a}(x) + \log f_a(x)$$



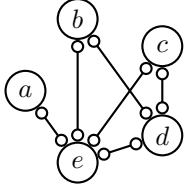
$$\log f(x) = \log f_{b|e}(x) + \log f_{e|ac}(x) + \log f_{c|d}(x) + \log f_{d|a}(x) + \log f_a(x)$$



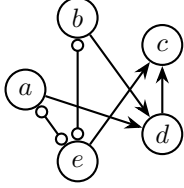
$$\log f(x) = \log f_{b|e}(x) + \log f_{e|acd}(x) + \log f_{c|d}(x) + \log f_{d|a}(x) + \log f_a(x)$$



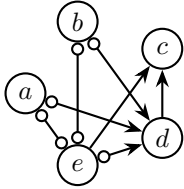
$$\log f(x) = \log f_{c|de}(x) + \log f_{d|b}(x) + \log f_{b|e}(x) + \log f_{e|a}(x) + \log f_a(x)$$



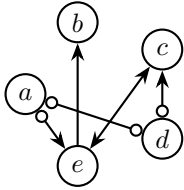
$$\log f(x) = \log f_{c|de}(x) + \log f_{d|be}(x) + \log f_{b|e}(x) + \log f_{e|a}(x) + \log f_a(x)$$



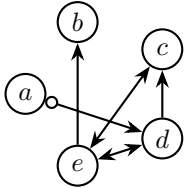
$$\log f(x) = \log f_{c|de}(x) + \log f_{d|ab}(x) + \log f_{b|e}(x) + \log f_{e|a}(x) + \log f_a(x)$$



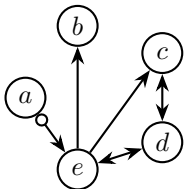
$$\log f(x) = \log f_{c|de}(x) + \log f_{d|abe}(x) + \log f_{b|e}(x) + \log f_{e|a}(x) + \log f_a(x)$$



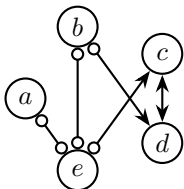
$$\log f(x) = \log f_{b|e}(x) + \log f_{c|ade}(x) + \log f_{d|a}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{a,c|d}(x)$$



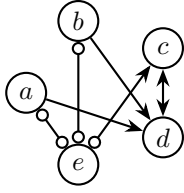
$$\log f(x) = \log f_{b|e}(x) + \log f_{c|ade}(x) + \log f_{d|ae}(x) + \log f_a(x) + \log f_e(x) - \phi_{a,c|d}(x)$$



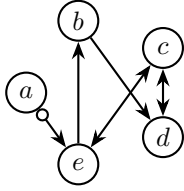
$$\log f(x) = \log f_{b|e}(x) + \log f_{c|ade}(x) + \log f_{d|ae}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{a,c|e}(x) - \phi_{a,d}(x)$$



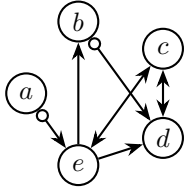
$$\log f(x) = \log f_{c|bde}(x) + \log f_{d|b}(x) + \log f_{b|e}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{b,c|e}(x)$$



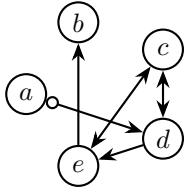
$$\log f(x) = \log f_{c|abde}(x) + \log f_{d|ab}(x) + \log f_{b|e}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{ab,c|e}(x)$$



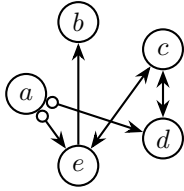
$$\log f(x) = \log f_{c|abde}(x) + \log f_{d|b}(x) + \log f_{b|e}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{b,c|ae}(x) - \phi_{a,c}(x)$$



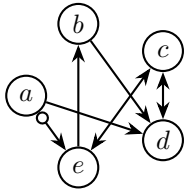
$$\log f(x) = \log f_{c|abde}(x) + \log f_{d|be}(x) + \log f_{b|e}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{b,c|ae}(x) - \phi_{a,c}(x)$$



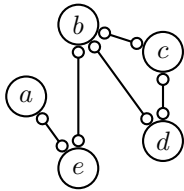
$$\log f(x) = \log f_{b|e}(x) + \log f_{c|ade}(x) + \log f_{e|d}(x) + \log f_{d|a}(x) + \log f_a(x) - \phi_{a,c}(x)$$



$$\log f(x) = \log f_{b|e}(x) + \log f_{c|ade}(x) + \log f_{d|a}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{a,c}(x)$$

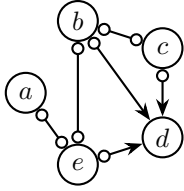


$$\log f(x) = \log f_{c|abde}(x) + \log f_{d|ab}(x) + \log f_{b|e}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{b,c|ae}(x) - \phi_{a,c}(x)$$

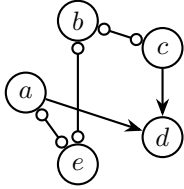


$$\log f(x) = \log f_{d|bc}(x) + \log f_{c|b}(x) + \log f_{b|e}(x) + \log f_{e|a}(x) + \log f_a(x)$$

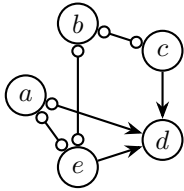




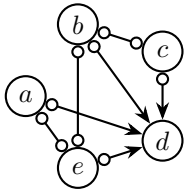
$$\log f(x) = \log f_{d|bce}(x) + \log f_{c|b}(x) + \log f_{b|e}(x) + \log f_{e|a}(x) + \log f_a(x)$$



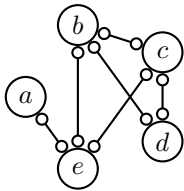
$$\log f(x) = \log f_{d|ac}(x) + \log f_{c|b}(x) + \log f_{b|e}(x) + \log f_{e|a}(x) + \log f_a(x)$$



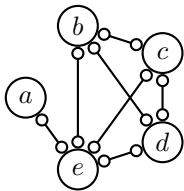
$$\log f(x) = \log f_{d|ace}(x) + \log f_{c|b}(x) + \log f_{b|e}(x) + \log f_{e|a}(x) + \log f_a(x)$$



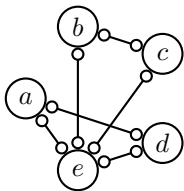
$$\log f(x) = \log f_{d|abce}(x) + \log f_{c|b}(x) + \log f_{b|e}(x) + \log f_{e|a}(x) + \log f_a(x)$$



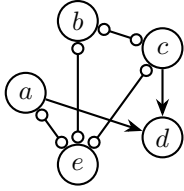
$$\log f(x) = \log f_{d|bc}(x) + \log f_{c|be}(x) + \log f_{b|e}(x) + \log f_{e|a}(x) + \log f_a(x)$$



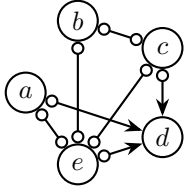
$$\log f(x) = \log f_{d|bce}(x) + \log f_{c|be}(x) + \log f_{b|e}(x) + \log f_{e|a}(x) + \log f_a(x)$$



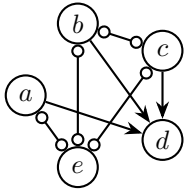
$$\log f(x) = \log f_{c|be}(x) + \log f_{b|e}(x) + \log f_{e|ad}(x) + \log f_{d|a}(x) + \log f_a(x)$$



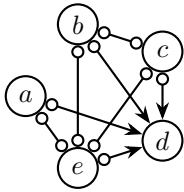
$$\log f(x) = \log f_{d|ac}(x) + \log f_{c|be}(x) + \log f_{b|e}(x) + \log f_{e|a}(x) + \log f_a(x)$$



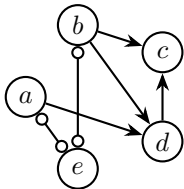
$$\log f(x) = \log f_{d|ace}(x) + \log f_{c|be}(x) + \log f_{b|e}(x) + \log f_{e|a}(x) + \log f_a(x)$$



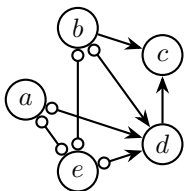
$$\log f(x) = \log f_{d|abc}(x) + \log f_{c|be}(x) + \log f_{b|e}(x) + \log f_{e|a}(x) + \log f_a(x)$$



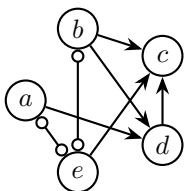
$$\log f(x) = \log f_{d|abce}(x) + \log f_{c|be}(x) + \log f_{b|e}(x) + \log f_{e|a}(x) + \log f_a(x)$$



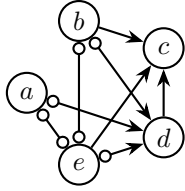
$$\log f(x) = \log f_{c|bd}(x) + \log f_{d|ab}(x) + \log f_{b|e}(x) + \log f_{e|a}(x) + \log f_a(x)$$



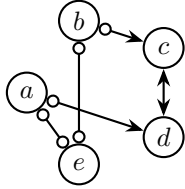
$$\log f(x) = \log f_{c|bd}(x) + \log f_{d|abe}(x) + \log f_{b|e}(x) + \log f_{e|a}(x) + \log f_a(x)$$



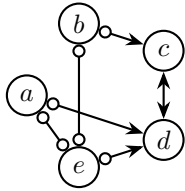
$$\log f(x) = \log f_{c|bde}(x) + \log f_{d|ab}(x) + \log f_{b|e}(x) + \log f_{e|a}(x) + \log f_a(x)$$



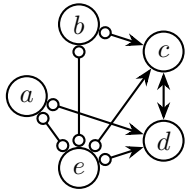
$$\log f(x) = \log f_{c|bde}(x) + \log f_{d|abe}(x) + \log f_{b|e}(x) + \log f_{e|a}(x) + \log f_a(x)$$



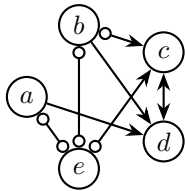
$$\log f(x) = \log f_{c|abd}(x) + \log f_{b|e}(x) + \log f_{d|a}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{a,c|b}(x)$$



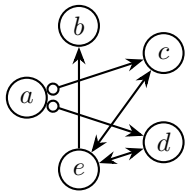
$$\log f(x) = \log f_{c|abde}(x) + \log f_{b|e}(x) + \log f_{d|ae}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{ae,c|b}(x)$$



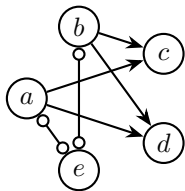
$$\log f(x) = \log f_{c|abde}(x) + \log f_{b|e}(x) + \log f_{d|ae}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{a,c|b}(x)$$



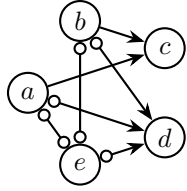
$$\log f(x) = \log f_{c|abde}(x) + \log f_{d|ab}(x) + \log f_{b|e}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{a,c|b}(x)$$



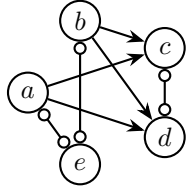
$$\log f(x) = \log f_{b|e}(x) + \log f_{c|ade}(x) + \log f_{d|ae}(x) + \log f_a(x) + \log f_e(x) - \phi_{c,d|a}(x)$$



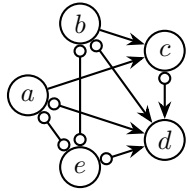
$$\log f(x) = \log f_{c|ab}(x) + \log f_{d|ab}(x) + \log f_{b|e}(x) + \log f_{e|a}(x) + \log f_a(x)$$



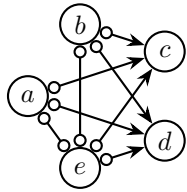
$$\log f(x) = \log f_{c|ab}(x) + \log f_{d|abe}(x) + \log f_{b|e}(x) + \log f_{e|a}(x) + \log f_a(x)$$



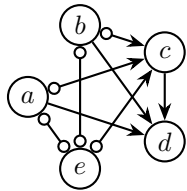
$$\log f(x) = \log f_{d|abc}(x) + \log f_{c|ab}(x) + \log f_{b|e}(x) + \log f_{e|a}(x) + \log f_a(x)$$



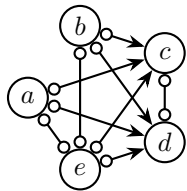
$$\log f(x) = \log f_{d|abce}(x) + \log f_{c|ab}(x) + \log f_{b|e}(x) + \log f_{e|a}(x) + \log f_a(x)$$



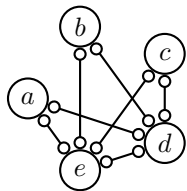
$$\log f(x) = \log f_{c|abe}(x) + \log f_{d|abe}(x) + \log f_{b|e}(x) + \log f_{e|a}(x) + \log f_a(x)$$



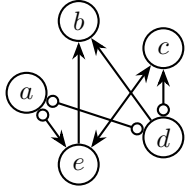
$$\log f(x) = \log f_{d|abc}(x) + \log f_{c|abe}(x) + \log f_{b|e}(x) + \log f_{e|a}(x) + \log f_a(x)$$



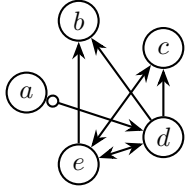
$$\log f(x) = \log f_{d|abce}(x) + \log f_{c|abe}(x) + \log f_{b|e}(x) + \log f_{e|a}(x) + \log f_a(x)$$



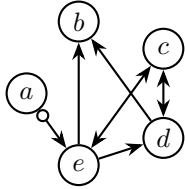
$$\log f(x) = \log f_{b|de}(x) + \log f_{c|de}(x) + \log f_{e|ad}(x) + \log f_{d|a}(x) + \log f_a(x)$$



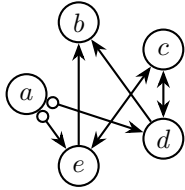
$$\log f(x) = \log f_{b|de}(x) + \log f_{c|ade}(x) + \log f_{d|a}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{a,c|d}(x)$$



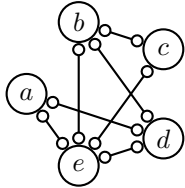
$$\log f(x) = \log f_{b|de}(x) + \log f_{c|ade}(x) + \log f_{d|ae}(x) + \log f_a(x) + \log f_e(x) - \phi_{a,c|d}(x)$$



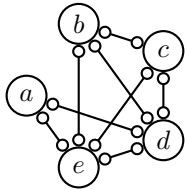
$$\log f(x) = \log f_{b|de}(x) + \log f_{c|ade}(x) + \log f_{d|e}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{a,c}(x)$$



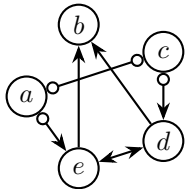
$$\log f(x) = \log f_{b|de}(x) + \log f_{c|ade}(x) + \log f_{d|a}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{a,c}(x)$$



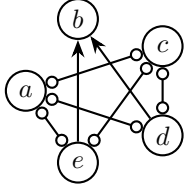
$$\log f(x) = \log f_{c|be}(x) + \log f_{b|de}(x) + \log f_{e|ad}(x) + \log f_{d|a}(x) + \log f_a(x)$$



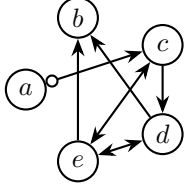
$$\log f(x) = \log f_{c|bde}(x) + \log f_{b|de}(x) + \log f_{e|ad}(x) + \log f_{d|a}(x) + \log f_a(x)$$



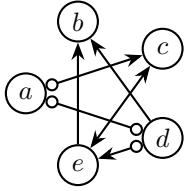
$$\log f(x) = \log f_{b|de}(x) + \log f_{d|ace}(x) + \log f_{c|a}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{a,d|c}(x)$$



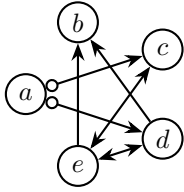
$$\log f(x) = \log f_{b|de}(x) + \log f_{d|ac}(x) + \log f_{e|ac}(x) + \log f_{c|a}(x) + \log f_a(x)$$



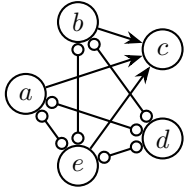
$$\log f(x) = \log f_{b|de}(x) + \log f_{d|ace}(x) + \log f_{c|ae}(x) + \log f_a(x) + \log f_e(x) - \phi_{a,d|c}(x)$$



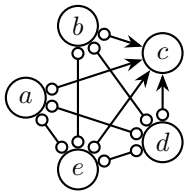
$$\log f(x) = \log f_{b|de}(x) + \log f_{c|ade}(x) + \log f_{e|d}(x) + \log f_{d|a}(x) + \log f_a(x) - \phi_{c,d|a}(x)$$



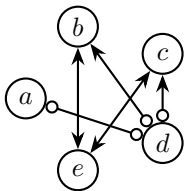
$$\log f(x) = \log f_{b|de}(x) + \log f_{c|ade}(x) + \log f_{d|ae}(x) + \log f_a(x) + \log f_e(x) - \phi_{c,d|a}(x)$$



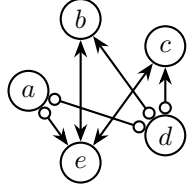
$$\log f(x) = \log f_{c|abe}(x) + \log f_{b|de}(x) + \log f_{e|ad}(x) + \log f_{d|a}(x) + \log f_a(x)$$



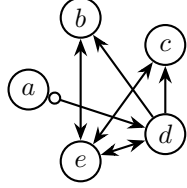
$$\log f(x) = \log f_{c|abde}(x) + \log f_{b|de}(x) + \log f_{e|ad}(x) + \log f_{d|a}(x) + \log f_a(x)$$



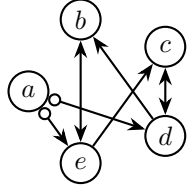
$$\log f(x) = \log f_{b|cde}(x) + \log f_{c|de}(x) + \log f_{d|a}(x) + \log f_a(x) + \log f_e(x) - \phi_{b,c|d}(x)$$



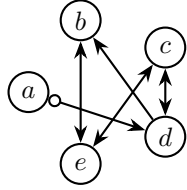
$$\log f(x) = \log f_{b|acde}(x) + \log f_{c|ade}(x) + \log f_{d|a}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{ac,b|d}(x) - \phi_{a,c|d}(x)$$



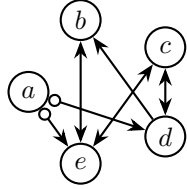
$$\log f(x) = \log f_{b|acde}(x) + \log f_{c|ade}(x) + \log f_{d|ae}(x) + \log f_a(x) + \log f_e(x) - \phi_{ac,b|d}(x) - \phi_{a,c|d}(x)$$



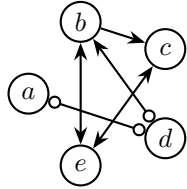
$$\log f(x) = \log f_{b|ade}(x) + \log f_{c|ade}(x) + \log f_{d|a}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{a,b|d}(x) - \phi_{a,c|e}(x)$$



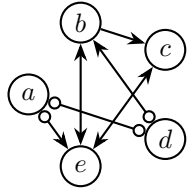
$$\log f(x) = \log f_{b|acde}(x) + \log f_{c|ade}(x) + \log f_{d|a}(x) + \log f_a(x) + \log f_e(x) - \phi_{b,c|ad}(x) - \phi_{a,b|de}(x) - \phi_{a,c|e}(x)$$



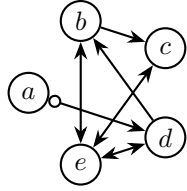
$$\log f(x) = \log f_{b|acde}(x) + \log f_{c|ade}(x) + \log f_{d|a}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{ac,b|d}(x) - \phi_{a,c}(x)$$



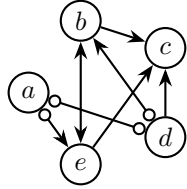
$$\log f(x) = \log f_{c|bde}(x) + \log f_{b|de}(x) + \log f_{d|a}(x) + \log f_a(x) + \log f_e(x) - \phi_{c,d|b}(x)$$



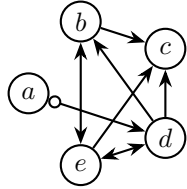
$$\log f(x) = \log f_{c|abde}(x) + \log f_{b|ade}(x) + \log f_{d|a}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{ad,c|b}(x) - \phi_{a,b|d}(x)$$



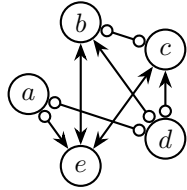
$$\log f(x) = \log f_{c|abde}(x) + \log f_{b|ade}(x) + \log f_{d|ae}(x) + \log f_a(x) + \log f_e(x) - \phi_{ad,c|b}(x) - \phi_{a,b|d}(x)$$



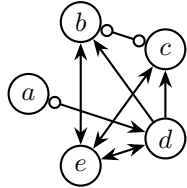
$$\log f(x) = \log f_{c|bde}(x) + \log f_{b|ade}(x) + \log f_{d|a}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{a,b|d}(x)$$



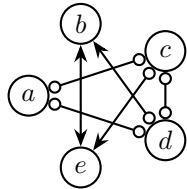
$$\log f(x) = \log f_{c|bde}(x) + \log f_{b|ade}(x) + \log f_{d|ae}(x) + \log f_a(x) + \log f_e(x) - \phi_{a,b|d}(x)$$



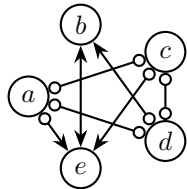
$$\log f(x) = \log f_{c|abde}(x) + \log f_{b|ade}(x) + \log f_{d|a}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{a,c|bd}(x) - \phi_{a,b|d}(x)$$



$$\log f(x) = \log f_{c|abde}(x) + \log f_{b|ade}(x) + \log f_{d|ae}(x) + \log f_a(x) + \log f_e(x) - \phi_{a,c|bd}(x) - \phi_{a,b|d}(x)$$

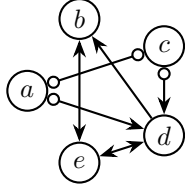


$$\log f(x) = \log f_{b|cde}(x) + \log f_{d|ac}(x) + \log f_{e|c}(x) + \log f_{c|a}(x) + \log f_a(x) - \phi_{b,c|d}(x)$$

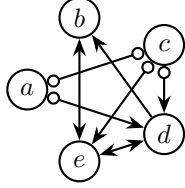


$$\log f(x) = \log f_{b|acde}(x) + \log f_{d|ac}(x) + \log f_{e|ac}(x) + \log f_{c|a}(x) + \log f_a(x) - \phi_{ac,b|d}(x)$$

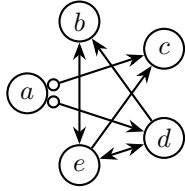




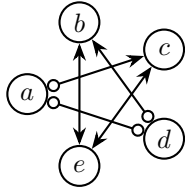
$$\log f(x) = \log f_{b|acde}(x) + \log f_{d|ace}(x) + \log f_{c|a}(x) + \log f_a(x) + \log f_e(x) - \phi_{ac,b|d}(x)$$



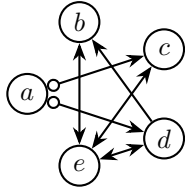
$$\log f(x) = \log f_{b|acde}(x) + \log f_{d|ace}(x) + \log f_{e|c}(x) + \log f_{c|a}(x) + \log f_a(x) - \phi_{ac,b|d}(x)$$



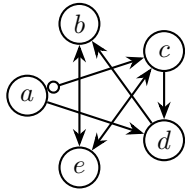
$$\log f(x) = \log f_{b|ade}(x) + \log f_{c|ae}(x) + \log f_{d|ae}(x) + \log f_a(x) + \log f_e(x) - \phi_{a,b|d}(x)$$



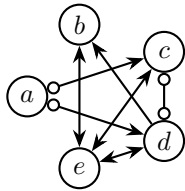
$$\log f(x) = \log f_{b|acde}(x) + \log f_{c|ae}(x) + \log f_{d|a}(x) + \log f_a(x) + \log f_e(x) - \phi_{b,c|ad}(x) - \phi_{a,b|de}(x)$$



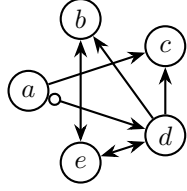
$$\log f(x) = \log f_{b|acde}(x) + \log f_{c|ade}(x) + \log f_{d|ae}(x) + \log f_a(x) + \log f_e(x) - \phi_{ac,b|d}(x) - \phi_{c,d|a}(x)$$



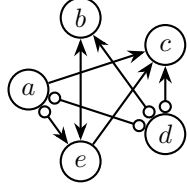
$$\log f(x) = \log f_{b|acde}(x) + \log f_{d|ac}(x) + \log f_{c|ae}(x) + \log f_a(x) + \log f_e(x) - \phi_{ac,b|d}(x)$$



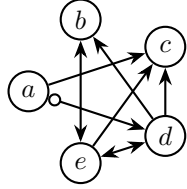
$$\log f(x) = \log f_{b|acde}(x) + \log f_{d|ace}(x) + \log f_{c|ae}(x) + \log f_a(x) + \log f_e(x) - \phi_{ac,b|d}(x)$$



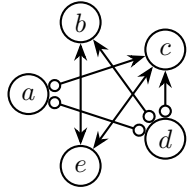
$$\log f(x) = \log f_{b|ade}(x) + \log f_{c|ad}(x) + \log f_{d|ae}(x) + \log f_a(x) + \log f_e(x) - \phi_{a,b|d}(x)$$



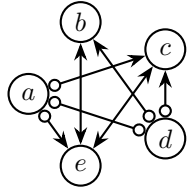
$$\log f(x) = \log f_{b|ade}(x) + \log f_{c|ade}(x) + \log f_{d|a}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{a,b|d}(x)$$



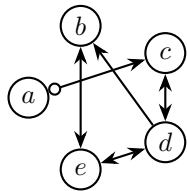
$$\log f(x) = \log f_{b|ade}(x) + \log f_{c|ade}(x) + \log f_{d|ae}(x) + \log f_a(x) + \log f_e(x) - \phi_{a,b|d}(x)$$



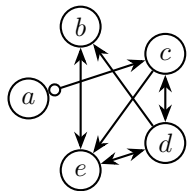
$$\log f(x) = \log f_{b|acde}(x) + \log f_{c|ade}(x) + \log f_{d|a}(x) + \log f_a(x) + \log f_e(x) - \phi_{b,c|ad}(x) - \phi_{a,b|de}(x)$$



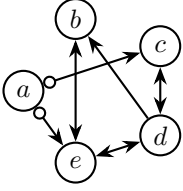
$$\log f(x) = \log f_{b|acde}(x) + \log f_{c|ade}(x) + \log f_{d|a}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{ac,b|d}(x)$$



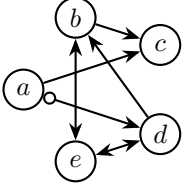
$$\log f(x) = \log f_{b|acde}(x) + \log f_{c|ade}(x) + \log f_a(x) + \log f_{d|e}(x) + \log f_e(x) - \phi_{b,c|ad}(x) - \phi_{a,b|de}(x) - \phi_{c,e|a}(x)$$



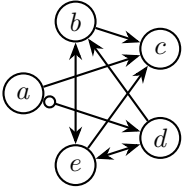
$$\log f(x) = \log f_{b|acde}(x) + \log f_{d|ace}(x) + \log f_{e|c}(x) + \log f_{c|a}(x) + \log f_a(x) - \phi_{ac,b|d}(x) - \phi_{a,d}(x)$$



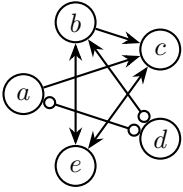
$$\log f(x) = \log f_{b|acde}(x) + \log f_{c|ade}(x) + \log f_{d|ae}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{ac,b|d}(x) - \phi_{c,e|a}(x) - \phi_{a,d}(x)$$



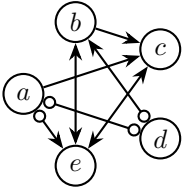
$$\log f(x) = \log f_{c|ab}(x) + \log f_{b|ade}(x) + \log f_{d|ae}(x) + \log f_a(x) + \log f_e(x) - \phi_{a,b|d}(x)$$



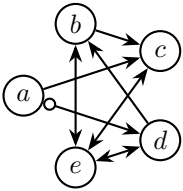
$$\log f(x) = \log f_{c|abe}(x) + \log f_{b|ade}(x) + \log f_{d|ae}(x) + \log f_a(x) + \log f_e(x) - \phi_{a,b|d}(x)$$



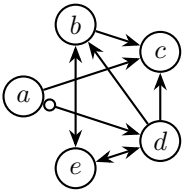
$$\log f(x) = \log f_{c|abde}(x) + \log f_{b|de}(x) + \log f_{d|a}(x) + \log f_a(x) + \log f_e(x) - \phi_{c,d|ab}(x)$$



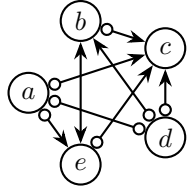
$$\log f(x) = \log f_{c|abde}(x) + \log f_{b|ade}(x) + \log f_{d|a}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{c,d|ab}(x) - \phi_{a,b|d}(x)$$



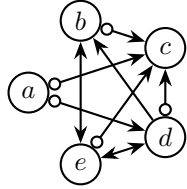
$$\log f(x) = \log f_{c|abde}(x) + \log f_{b|ade}(x) + \log f_{d|ae}(x) + \log f_a(x) + \log f_e(x) - \phi_{c,d|ab}(x) - \phi_{a,b|d}(x)$$



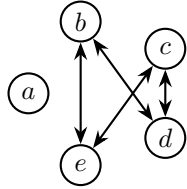
$$\log f(x) = \log f_{c|abd}(x) + \log f_{b|ade}(x) + \log f_{d|ae}(x) + \log f_a(x) + \log f_e(x) - \phi_{a,b|d}(x)$$



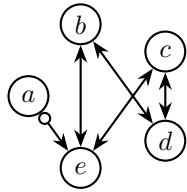
$$\log f(x) = \log f_{c|abde}(x) + \log f_{b|ade}(x) + \log f_{d|a}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{a,b|d}(x)$$



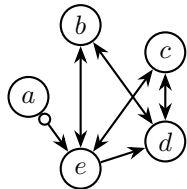
$$\log f(x) = \log f_{c|abde}(x) + \log f_{b|ade}(x) + \log f_{d|ae}(x) + \log f_a(x) + \log f_e(x) - \phi_{a,b|d}(x)$$



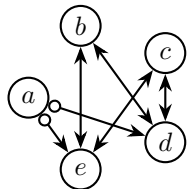
$$\log f(x) = \log f_a(x) + \log f_{b|cde}(x) + \log f_{c|de}(x) + \log f_d(x) + \log f_e(x) - \phi_{b,c}(x)$$



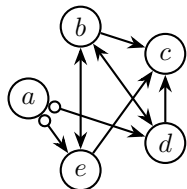
$$\log f(x) = \log f_{b|acde}(x) + \log f_{c|ade}(x) + \log f_d(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{a,b|cd}(x) - \phi_{b,c}(x) - \phi_{a,c|d}(x)$$



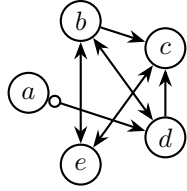
$$\log f(x) = \log f_{b|acde}(x) + \log f_{c|ade}(x) + \log f_{d|e}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{ac,b}(x) - \phi_{a,c}(x)$$



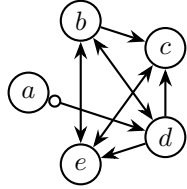
$$\log f(x) = \log f_{b|acde}(x) + \log f_{c|ade}(x) + \log f_{d|a}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{ac,b}(x) - \phi_{a,c}(x)$$



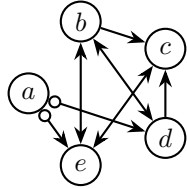
$$\log f(x) = \log f_{c|bde}(x) + \log f_{b|ade}(x) + \log f_{d|a}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{a,b}(x)$$



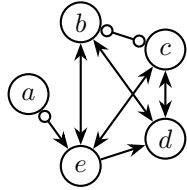
$$\log f(x) = \log f_{c|abde}(x) + \log f_{b|ade}(x) + \log f_{d|a}(x) + \log f_a(x) + \log f_e(x) - \phi_{a,c|bd}(x) - \phi_{a,b|e}(x)$$



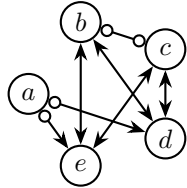
$$\log f(x) = \log f_{c|abde}(x) + \log f_{b|ade}(x) + \log f_{e|d}(x) + \log f_{d|a}(x) + \log f_a(x) - \phi_{a,c|bd}(x) - \phi_{a,b}(x)$$



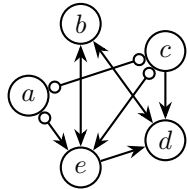
$$\log f(x) = \log f_{c|abde}(x) + \log f_{b|ade}(x) + \log f_{d|a}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{a,c|bd}(x) - \phi_{a,b}(x)$$



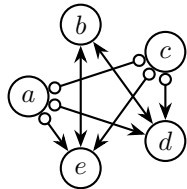
$$\log f(x) = \log f_{c|abde}(x) + \log f_{b|ade}(x) + \log f_{d|e}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{a,c|b}(x) - \phi_{a,b}(x)$$



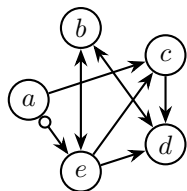
$$\log f(x) = \log f_{c|abde}(x) + \log f_{b|ade}(x) + \log f_{d|a}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{a,c|b}(x) - \phi_{a,b}(x)$$



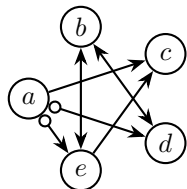
$$\log f(x) = \log f_{b|acde}(x) + \log f_{d|ce}(x) + \log f_{e|ac}(x) + \log f_{c|a}(x) + \log f_a(x) - \phi_{ac,b}(x)$$



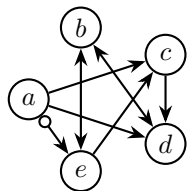
$$\log f(x) = \log f_{b|acde}(x) + \log f_{d|ac}(x) + \log f_{e|ac}(x) + \log f_{c|a}(x) + \log f_a(x) - \phi_{ac,b}(x)$$



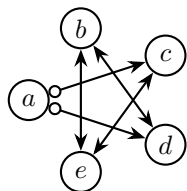
$$\log f(x) = \log f_{b|acde}(x) + \log f_{d|ce}(x) + \log f_{c|ae}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{b,c|ae}(x) - \phi_{a,b}(x)$$



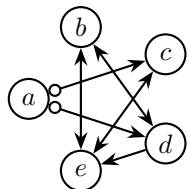
$$\log f(x) = \log f_{b|ade}(x) + \log f_{c|ae}(x) + \log f_{d|a}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{a,b}(x)$$



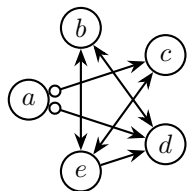
$$\log f(x) = \log f_{b|acde}(x) + \log f_{d|ac}(x) + \log f_{c|ae}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{b,c|ae}(x) - \phi_{a,b}(x)$$



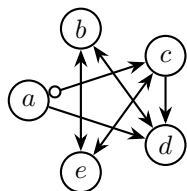
$$\log f(x) = \log f_{b|acde}(x) + \log f_{c|ae}(x) + \log f_{d|a}(x) + \log f_a(x) + \log f_e(x) - \phi_{b,c|ad}(x) - \phi_{a,b|e}(x)$$



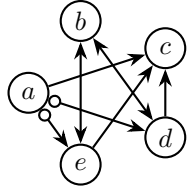
$$\log f(x) = \log f_{b|acde}(x) + \log f_{c|ade}(x) + \log f_{e|d}(x) + \log f_{d|a}(x) + \log f_a(x) - \phi_{b,c|ad}(x) - \phi_{a,b}(x) - \phi_{c,d|a}(x)$$



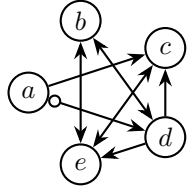
$$\log f(x) = \log f_{b|acde}(x) + \log f_{c|ae}(x) + \log f_{d|ae}(x) + \log f_a(x) + \log f_e(x) - \phi_{b,c|a}(x) - \phi_{a,b|e}(x)$$



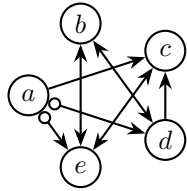
$$\log f(x) = \log f_{b|acde}(x) + \log f_{d|ac}(x) + \log f_{c|ae}(x) + \log f_a(x) + \log f_e(x) - \phi_{b,c|a}(x) - \phi_{a,b|e}(x)$$



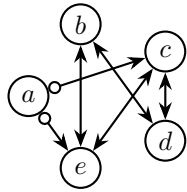
$$\log f(x) = \log f_{b|ade}(x) + \log f_{c|ade}(x) + \log f_{d|a}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{a,b}(x)$$



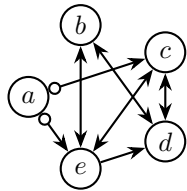
$$\log f(x) = \log f_{b|acde}(x) + \log f_{c|ade}(x) + \log f_{e|d}(x) + \log f_{d|a}(x) + \log f_a(x) - \phi_{b,c|ad}(x) - \phi_{a,b}(x)$$



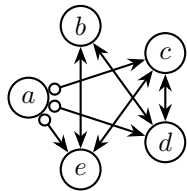
$$\log f(x) = \log f_{b|acde}(x) + \log f_{c|ade}(x) + \log f_{d|a}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{b,c|ad}(x) - \phi_{a,b}(x)$$



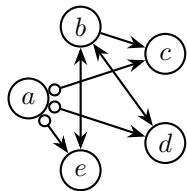
$$\log f(x) = \log f_{b|acde}(x) + \log f_{c|ade}(x) + \log f_d(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{b,c|a}(x) - \phi_{a,b|d}(x)$$



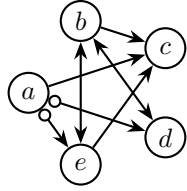
$$\log f(x) = \log f_{b|acde}(x) + \log f_{c|ade}(x) + \log f_{d|e}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{ac,b}(x)$$



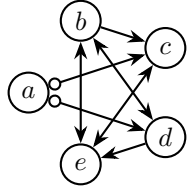
$$\log f(x) = \log f_{b|acde}(x) + \log f_{c|ade}(x) + \log f_{d|a}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{ac,b}(x)$$



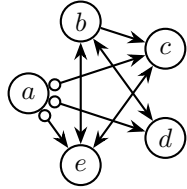
$$\log f(x) = \log f_{c|ab}(x) + \log f_{b|ade}(x) + \log f_{d|a}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{a,b}(x)$$



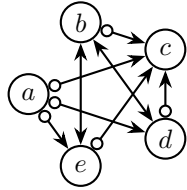
$$\log f(x) = \log f_{c|abe}(x) + \log f_{b|ade}(x) + \log f_{d|a}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{a,b}(x)$$



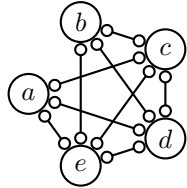
$$\log f(x) = \log f_{c|abde}(x) + \log f_{b|ade}(x) + \log f_{e|d}(x) + \log f_{d|a}(x) + \log f_a(x) - \phi_{c,d|ab}(x) - \phi_{a,b}(x)$$



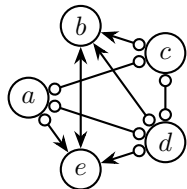
$$\log f(x) = \log f_{c|abde}(x) + \log f_{b|ade}(x) + \log f_{d|a}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{c,d|ab}(x) - \phi_{a,b}(x)$$



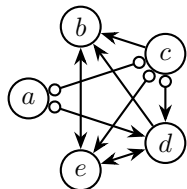
$$\log f(x) = \log f_{c|abde}(x) + \log f_{b|ade}(x) + \log f_{d|a}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{a,b}(x)$$



$$\log f(x) = \log f_{b|cde}(x) + \log f_{e|acd}(x) + \log f_{d|ac}(x) + \log f_{c|a}(x) + \log f_a(x)$$

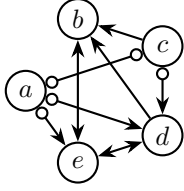


$$\log f(x) = \log f_{b|acde}(x) + \log f_{e|ad}(x) + \log f_{d|ac}(x) + \log f_{c|a}(x) + \log f_a(x) - \phi_{a,b|cd}(x)$$

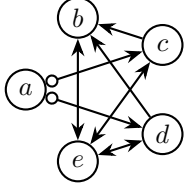


$$\log f(x) = \log f_{b|acde}(x) + \log f_{d|ace}(x) + \log f_{e|c}(x) + \log f_{c|a}(x) + \log f_a(x) - \phi_{a,b|cd}(x)$$

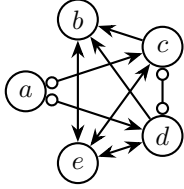




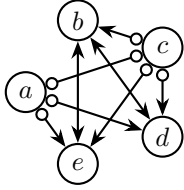
$$\log f(x) = \log f_{b|acde}(x) + \log f_{d|ace}(x) + \log f_{c|a}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{a,b|cd}(x)$$



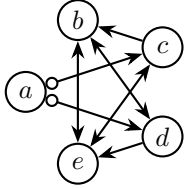
$$\log f(x) = \log f_{b|acde}(x) + \log f_{c|ade}(x) + \log f_{d|ae}(x) + \log f_a(x) + \log f_e(x) - \phi_{a,b|cd}(x) - \phi_{c,d|a}(x)$$



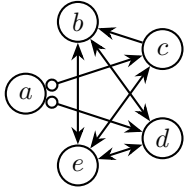
$$\log f(x) = \log f_{b|acde}(x) + \log f_{d|ace}(x) + \log f_{c|ae}(x) + \log f_a(x) + \log f_e(x) - \phi_{a,b|cd}(x)$$



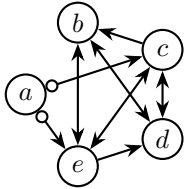
$$\log f(x) = \log f_{b|acde}(x) + \log f_{d|ac}(x) + \log f_{e|ac}(x) + \log f_{c|a}(x) + \log f_a(x) - \phi_{a,b|c}(x)$$



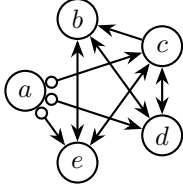
$$\log f(x) = \log f_{b|acde}(x) + \log f_{c|ade}(x) + \log f_{e|d}(x) + \log f_{d|a}(x) + \log f_a(x) - \phi_{a,b|c}(x) - \phi_{c,d|a}(x)$$



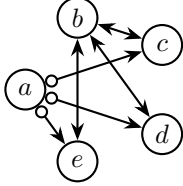
$$\log f(x) = \log f_{b|acde}(x) + \log f_{c|ade}(x) + \log f_{d|ae}(x) + \log f_a(x) + \log f_e(x) - \phi_{a,b|c}(x) - \phi_{c,d|a}(x)$$



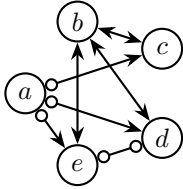
$$\log f(x) = \log f_{b|acde}(x) + \log f_{c|ade}(x) + \log f_{d|e}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{a,b|c}(x)$$



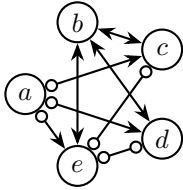
$$\log f(x) = \log f_{b|acde}(x) + \log f_{c|ade}(x) + \log f_{d|a}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{a,b|c}(x)$$



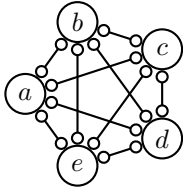
$$\log f(x) = \log f_{b|acde}(x) + \log f_{c|a}(x) + \log f_{d|a}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{a,b}(x)$$



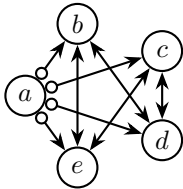
$$\log f(x) = \log f_{b|acde}(x) + \log f_{c|a}(x) + \log f_{e|ad}(x) + \log f_{d|a}(x) + \log f_a(x) - \phi_{a,b}(x)$$



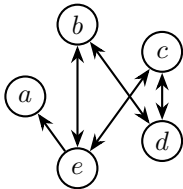
$$\log f(x) = \log f_{b|acde}(x) + \log f_{d|ae}(x) + \log f_{e|ac}(x) + \log f_{c|a}(x) + \log f_a(x) - \phi_{a,b}(x)$$



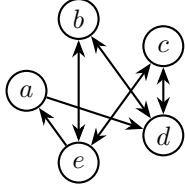
$$\log f(x) = \log f_{e|abcd}(x) + \log f_{d|abc}(x) + \log f_{c|ab}(x) + \log f_{b|a}(x) + \log f_a(x)$$



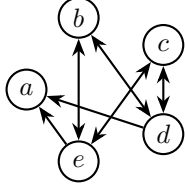
$$\log f(x) = \log f_{b|acde}(x) + \log f_{c|ade}(x) + \log f_{d|a}(x) + \log f_{e|a}(x) + \log f_a(x) - \phi_{b,c|a}(x)$$



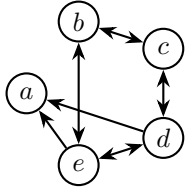
$$\log f(x) = \log f_{a|e}(x) + \log f_{b|cde}(x) + \log f_{c|de}(x) + \log f_d(x) + \log f_e(x) - \phi_{b,c}(x)$$



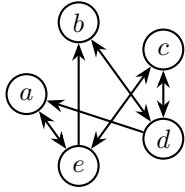
$$\log f(x) = \log f_{b|acde}(x) + \log f_{c|ade}(x) + \log f_{d|a}(x) + \log f_{a|e}(x) + \log f_e(x) - \phi_{a,b|ce}(x) - \phi_{b,c}(x) - \phi_{a,c|e}(x)$$



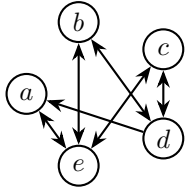
$$\log f(x) = \log f_{a|de}(x) + \log f_{b|cde}(x) + \log f_{c|de}(x) + \log f_d(x) + \log f_e(x) - \phi_{b,c}(x)$$



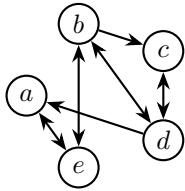
$$\log f(x) = \log f_{a|de}(x) + \log f_{b|cde}(x) + \log f_{c|de}(x) + \log f_{d|e}(x) + \log f_e(x) - \phi_{b,d}(x) - \phi_{c,e}(x)$$



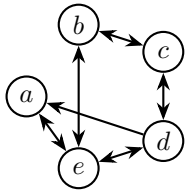
$$\log f(x) = \log f_{a|bcde}(x) + \log f_{b|cde}(x) + \log f_{c|de}(x) + \log f_d(x) + \log f_e(x) - \phi_{a,b|de}(x) - \phi_{a,c|d}(x) - \phi_{b,c|e}(x)$$



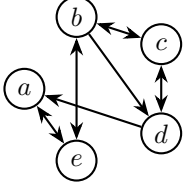
$$\log f(x) = \log f_{a|bcde}(x) + \log f_{b|cde}(x) + \log f_{c|de}(x) + \log f_d(x) + \log f_e(x) - \phi_{a,bc|d}(x) - \phi_{b,c}(x)$$



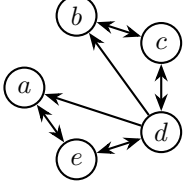
$$\log f(x) = \log f_{a|bcde}(x) + \log f_{c|bde}(x) + \log f_{b|de}(x) + \log f_d(x) + \log f_e(x) - \phi_{a,bc|d}(x) - \phi_{c,e|b}(x)$$



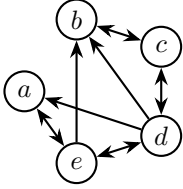
$$\log f(x) = \log f_{a|bcde}(x) + \log f_{b|cde}(x) + \log f_{c|de}(x) + \log f_{d|e}(x) + \log f_e(x) - \phi_{a,bc|d}(x) - \phi_{b,d}(x) - \phi_{c,e}(x)$$



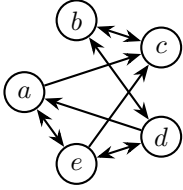
$$\log f(x) = \log f_{a|bcde}(x) + \log f_{c|bde}(x) + \log f_{d|b}(x) + \log f_{b|e}(x) + \log f_e(x) - \phi_{a,bc|d}(x) - \phi_{c,e}(x)$$



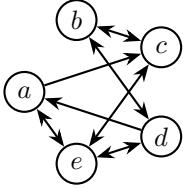
$$\log f(x) = \log f_{a|bcde}(x) + \log f_{b|cde}(x) + \log f_{c|de}(x) + \log f_{d|e}(x) + \log f_e(x) - \phi_{a,c|bd}(x) - \phi_{a,b|de}(x) - \phi_{b,e|d}(x) - \phi_{c,e}(x)$$



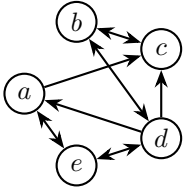
$$\log f(x) = \log f_{a|bcde}(x) + \log f_{b|cde}(x) + \log f_{c|de}(x) + \log f_{d|e}(x) + \log f_e(x) - \phi_{a,b|de}(x) - \phi_{a,c|d}(x) - \phi_{c,e}(x)$$



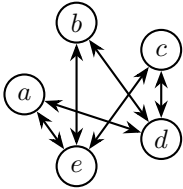
$$\log f(x) = \log f_{b|acde}(x) + \log f_{c|ae}(x) + \log f_{a|de}(x) + \log f_{d|e}(x) + \log f_e(x) - \phi_{a,b|d}(x) - \phi_{b,e}(x)$$



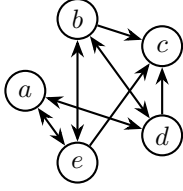
$$\log f(x) = \log f_{b|acde}(x) + \log f_{c|ade}(x) + \log f_{a|de}(x) + \log f_{d|e}(x) + \log f_e(x) - \phi_{a,b|d}(x) - \phi_{b,e}(x) - \phi_{c,d|a}(x)$$



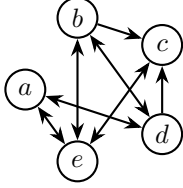
$$\log f(x) = \log f_{b|acde}(x) + \log f_{c|ad}(x) + \log f_{a|de}(x) + \log f_{d|e}(x) + \log f_e(x) - \phi_{a,b|d}(x) - \phi_{b,e}(x)$$



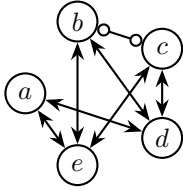
$$\log f(x) = \log f_{a|bcde}(x) + \log f_{b|cde}(x) + \log f_{c|de}(x) + \log f_d(x) + \log f_e(x) - \phi_{a,bc}(x) - \phi_{b,c}(x)$$



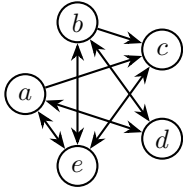
$$\log f(x) = \log f_{a|bde}(x) + \log f_{c|bde}(x) + \log f_{b|de}(x) + \log f_d(x) + \log f_e(x) - \phi_{a,b}(x)$$



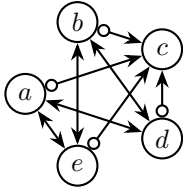
$$\log f(x) = \log f_{a|bcde}(x) + \log f_{c|bde}(x) + \log f_{b|de}(x) + \log f_d(x) + \log f_e(x) - \phi_{a,c|bd}(x) - \phi_{a,b}(x)$$



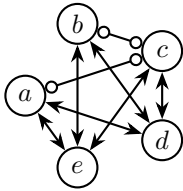
$$\log f(x) = \log f_{a|bcde}(x) + \log f_{c|bde}(x) + \log f_{b|de}(x) + \log f_d(x) + \log f_e(x) - \phi_{a,bc}(x)$$



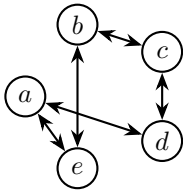
$$\log f(x) = \log f_{c|abde}(x) + \log f_{a|bde}(x) + \log f_{b|de}(x) + \log f_d(x) + \log f_e(x) - \phi_{c,d|ab}(x) - \phi_{a,b}(x)$$



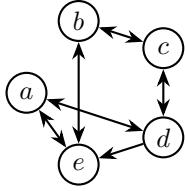
$$\log f(x) = \log f_{c|abde}(x) + \log f_{a|bde}(x) + \log f_{b|de}(x) + \log f_d(x) + \log f_e(x) - \phi_{a,b}(x)$$



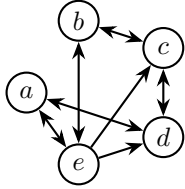
$$\log f(x) = \log f_{b|acde}(x) + \log f_{c|ade}(x) + \log f_{a|de}(x) + \log f_d(x) + \log f_e(x) - \phi_{a,b|c}(x)$$



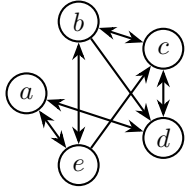
$$\log f(x) = \log f_{e|abcd}(x) + \log f_{d|abc}(x) + \log f_{c|ab}(x) + \log f_{b|a}(x) + \log f_a(x) - \phi_{a,c|b}(x) - \phi_{a,b|d}(x) - \phi_{c,e|a}(x) - \phi_{b,d|e}(x) - \phi_{d,e|c}(x)$$



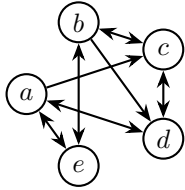
$$\log f(x) = \log f_{a|bcde}(x) + \log f_{b|cde}(x) + \log f_{c|d}(x) + \log f_{e|d}(x) + \log f_d(x) - \phi_{a,c|b}(x) - \phi_{a,b|d}(x) - \phi_{b,d}(x)$$



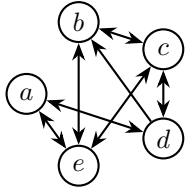
$$\log f(x) = \log f_{a|bcde}(x) + \log f_{b|cde}(x) + \log f_{c|de}(x) + \log f_{d|e}(x) + \log f_e(x) - \phi_{a,c|e}(x) - \phi_{a,b}(x) - \phi_{b,d|e}(x)$$



$$\log f(x) = \log f_{a|bcde}(x) + \log f_{c|bde}(x) + \log f_{d|b}(x) + \log f_{b|e}(x) + \log f_e(x) - \phi_{a,c|e}(x) - \phi_{a,b}(x)$$



$$\log f(x) = \log f_{c|abde}(x) + \log f_{a|bde}(x) + \log f_{d|b}(x) + \log f_{b|e}(x) + \log f_e(x) - \phi_{c,e|a}(x) - \phi_{a,b}(x)$$



$$\log f(x) = \log f_{a|bcde}(x) + \log f_{b|cde}(x) + \log f_{c|de}(x) + \log f_d(x) + \log f_e(x) - \phi_{a,b|d}(x) - \phi_{a,c}(x)$$

## Bibliography

- [1] Erling Bernhard Andersen. Sufficiency and exponential families for discrete sample spaces. *Journal of the American Statistical Association*, 65:1248–1255, 1970.
- [2] Bryan Andrews, Joseph Ramsey, and Gregory F Cooper. Learning high-dimensional directed acyclic graphs with mixed data-types. *Proceedings of Machine Learning Research*, 104:4–21, 2019.
- [3] Bryan Andrews, Peter Spirtes, and Gregory F Cooper. On the completeness of causal discovery in the presence of latent confounding with tiered background knowledge. In *International Conference on Artificial Intelligence and Statistics*, pages 4002–4011. PMLR, 2020.
- [4] Elias Bareinboim and Judea Pearl. Controlling selection bias in causal inference. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 100–108, 2012.
- [5] Ole Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*. John Wiley & Sons, 2014.
- [6] Ole Barndorff-Nielsen and Karl Pedersen. Sufficient data reduction and exponential families. *Mathematica Scandinavica*, 22:197–202, 1968.
- [7] Robert H Berk. Consistency and asymptotic normality of mle’s for exponential models. *The Annals of Mathematical Statistics*, pages 193–204, 1972.

- [8] Daniel Bernstein, Basil Saeed, Chandler Squires, and Caroline Uhler. Ordering-based causal structure learning in the presence of latent variables. In *International Conference on Artificial Intelligence and Statistics*, pages 4098–4108. PMLR, 2020.
- [9] Rohit Bhattacharya, Tushar Nagarajan, Daniel Malinsky, and Ilya Shpitser. Differentiable causal discovery under unmeasured confounding. In *International Conference on Artificial Intelligence and Statistics*, pages 2314–2322. PMLR, 2021.
- [10] Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
- [11] Remco Ronaldus Bouckaert. *Bayesian belief networks: from construction to inference*. PhD thesis, Utrecht University, 1995.
- [12] Rui Chen, Sanjeeb Dash, and Tian Gao. Integer programming for causal structure learning in the presence of latent variables. *arXiv preprint arXiv:2102.03129*, 2021.
- [13] David M Chickering. Optimal structure identification with greedy search. *The Journal of Machine Learning Research*, 3:507–554, 2002.
- [14] Diego Colombo and Marloes H Maathuis. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 15:3741–3782, 2014.
- [15] Gregory F Cooper. A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. *Data Mining and Knowledge Discovery*, 1:203–224, 1997.
- [16] Adnan Darwiche. *Modeling and Reasoning with Bayesian Networks*. Cambridge University Press, 2009.



- [17] A Philip Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41:1–15, 1979.
- [18] A Philip Dawid. Posterior model probabilities. In *Philosophy of Statistics*, pages 607–630. Elsevier, 2011.
- [19] Morris H DeGroot and Stephen E Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32:12–22, 1983.
- [20] Mathias Drton and Thomas S Richardson. Iterative conditional fitting for gaussian ancestral graph models. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 130–137, 2004.
- [21] Mathias Drton and Thomas S Richardson. Multimodality of the likelihood in the bivariate seemingly unrelated regressions model. *Biometrika*, 91:383–392, 2004.
- [22] Imme Ebert-Uphoff and Yi Deng. Causal discovery from spatio-temporal data with applications to climate science. In *Proceedings of the International Conference on Machine Learning and Applications*, pages 606–613. IEEE, 2014.
- [23] U.S. EPA. Integrated science assessment (ISA) for carbon monoxide. Technical report, U.S. Environmental Protection Agency, Washington, DC, January 2010. EPA/600/R-09/019F.
- [24] U.S. EPA. Integrated science assessment (ISA) for lead. Technical report, U.S. Environmental Protection Agency, Washington, DC, July 2013. EPA/600/R-10/075F.

- [25] U.S. EPA. Integrated science assessment (ISA) for oxides of nitrogen – health criteria. Technical report, U.S. Environmental Protection Agency, Washington, DC, January 2016. EPA/600/R-15/068.
- [26] U.S. EPA. Integrated science assessment (ISA) for sulfur oxides – health criteria. Technical report, U.S. Environmental Protection Agency, Washington, DC, December 2017. EPA/600/R-17/451.
- [27] U.S. EPA. Integrated science assessment (ISA) for particulate matter. Technical report, U.S. Environmental Protection Agency, Washington, DC, December 2019. EPA/600/R-19/188.
- [28] U.S. EPA. Integrated science assessment (ISA) for ozone and related photochemical oxidants. Technical report, U.S. Environmental Protection Agency, Washington, DC, April 2020. EPA 600/R-10/076F.
- [29] Robin J Evans. Graphs for margins of Bayesian networks. *Scandinavian Journal of Statistics*, 43:625–648, 2016.
- [30] Robin J Evans and Thomas S Richardson. Markovian acyclic directed mixed graphs for discrete data. *The Annals of Statistics*, pages 1452–1482, 2014.
- [31] Morten Frydenberg. Marginalization and collapsibility in graphical interaction models. *The Annals of Statistics*, pages 790–805, 1990.
- [32] Dan Geiger, David Heckerman, Henry King, and Christopher Meek. Stratified exponential families: Graphical models and model selection. *Annals of statistics*, pages 505–529, 2001.

- [33] Dan Geiger and Christopher Meek. Graphical models and exponential families. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 156–165, 1998.
- [34] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.
- [35] Clark Glymour and Gregory F Cooper. *Computation, Causation, and Discovery*. MIT Press, 1999.
- [36] Dominique MA Haughton. On the choice of a model to fit data from an exponential family. *The Annals of Statistics*, 16:342–355, 1988.
- [37] Raymond Hemmecke, Silvia Lindner, and Milan Studený. Characteristic imsets for learning Bayesian network structure. *The International Journal of Approximate Reasoning*, 53:1336–1349, 2012.
- [38] Zhongyi Hu and Robin J Evans. Faster algorithms for Markov equivalence. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 739–748. PMLR, 2020.
- [39] Biwei Huang, Kun Zhang, Yizhu Lin, Bernhard Schölkopf, and Clark Glymour. Generalized score functions for causal discovery. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1551–1560, 2018.
- [40] Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21:1–53, 2020.

- [41] Fattaneh Jabbari. *Instance-Specific Causal Bayesian Network Structure Learning*. PhD thesis, University of Pittsburgh, 2021.
- [42] Fattaneh Jabbari, Joseph Ramsey, Peter Spirtes, and Gregory F Cooper. Discovery of causal models that contain latent variables through Bayesian scoring of independence constraints. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 142–157. Springer, 2017.
- [43] Robert E Kass and Paul W Vos. *Geometrical Foundations of Asymptotic Inference*. John Wiley & Sons, 1997.
- [44] Harri Kiiveri, Terry P Speed, and John B Carlin. Recursive causal models. *Journal of the Australian Mathematical Society*, 36:30–52, 1984.
- [45] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.
- [46] Steffen L Lauritzen. *Graphical Models*, volume 17. Clarendon Press, 1996.
- [47] Jason D Lee and Trevor J Hastie. Learning the structure of mixed graphical models. *Journal of Computational and Graphical Statistics*, 24:230–253, 2015.
- [48] Zhenyu A Liao, Charupriya Sharma, James Cussens, and Peter van Beek. Finding all Bayesian network structures within a factor of optimal. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7892–7899, 2019.
- [49] Marloes H Maathuis, Diego Colombo, Markus Kalisch, and Peter Bühlmann. Predicting causal effects in large-scale systems from observational data. *Nature Methods*, 7:247–248, 2010.

- [50] Marloes H Maathuis, Mathias Drton, Steffen L Lauritzen, and Martin Wainwright. *Handbook of Graphical Models*. CRC Press, 2018.
- [51] Marloes H Maathuis, Markus Kalisch, and Peter Bühlmann. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37:3133–3164, 2009.
- [52] Subramani Mani and Gregory F Cooper. Causal discovery using a Bayesian local causal discovery algorithm. *Studies in Health Technology and Informatics*, 107:731–735, 2004.
- [53] William McGill. Multivariate information transmission. *Transactions of the IRE Professional Group on Information Theory*, 4:93–111, 1954.
- [54] Mahdi Pakdaman Naeini, Gregory F Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.
- [55] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the International Conference on Machine Learning*, pages 625–632, 2005.
- [56] Christopher Nowzohour, Marloes H Maathuis, Robin J Evans, and Peter Bühlmann. Distributional equivalence and structure learning for bow-free acyclic path diagrams. *Electronic Journal of Statistics*, 11:5342–5374, 2017.
- [57] Juan Miguel Ogarrio, Peter Spirtes, and Joseph Ramsey. A hybrid causal search algorithm for latent variable models. In *Conference on Probabilistic Graphical Models*, pages 368–379. PMLR, 2016.

- [58] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.
- [59] Judea Pearl. *Causality*. Cambridge University Press, 2009.
- [60] Judea Pearl and Azaria Paz. Graphoids: A graph-based logic for reasoning about relevance relations. In Ben Du Boulay, David Hogg, and Luc Steels, editors, *Advances in Artificial Intelligence-II*, pages 357–363. Elsevier, 1987.
- [61] Judea Pearl and Thomas S Verma. A statistical semantics for causation. *Statistics and Computing*, 2:91–95, 1992.
- [62] Emilija Perkovic, Johannes Textor, Markus Kalisch, and Marloes H Maathuis. Complete graphical characterization and construction of adjustment sets in markov equivalence classes of ancestral graphs. *The Journal of Machine Learning Research*, 18:8132–8193, 2017.
- [63] Vineet K Raghu, Joseph Ramsey, Alison Morris, Dimitrios V Manatakis, Peter Spirtes, Panos K Chrysanthis, Clark Glymour, and Panayiotis V Benos. Comparison of strategies for scalable causal discovery of latent variable models from mixed data. *International Journal of Data Science and Analytics*, 6:33–45, 2018.
- [64] Joseph Ramsey, Peter Spirtes, and Jiji Zhang. Adjacency-faithfulness and conservative causal inference. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 401–408, 2006.
- [65] Thomas S Richardson. Chain graphs and symmetric associations. In *Learning in Graphical Models*, pages 231–259. Springer, 1998.

- [66] Thomas S Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30:145–157, 2003.
- [67] Thomas S Richardson. A factorization criterion for acyclic directed mixed graphs. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 462–470, 2009.
- [68] Thomas S Richardson, Robin J Evans, and James M Robins. Transparent parameterizations of models for potential outcomes. *Bayesian Statistics*, 9:569–610, 2011.
- [69] Thomas S Richardson, Robin J Evans, James M Robins, and Ilya Shpitser. Nested Markov properties for acyclic directed mixed graphs. *arXiv preprint arXiv:1701.06686*, 2017.
- [70] Thomas S Richardson and Peter Spirtes. Ancestral graph Markov models. *The Annals of Statistics*, 30:962–1030, 2002.
- [71] Gian-Carlo Rota. On the foundations of combinatorial theory. I. Theory of Möbius functions. *Probability Theory and Related Fields*, 2:340–368, 1964.
- [72] Karen Sachs, Omar Perez, Dana Pe’er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308:523–529, 2005.
- [73] Kayvan Sadeghi. Stable mixed graphs. *Bernoulli*, 19:2330–2358, 2013.
- [74] Kayvan Sadeghi and Steffen L Lauritzen. Markov properties for mixed graphs. *Bernoulli*, 20:676–696, 2014.

- [75] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.
- [76] Ilya Shpitser and Judea Pearl. Dormant independence. In *Proceedings of the Conference on Artificial Intelligence and the Innovative Applications of Artificial Intelligence Conference*, pages 1081–1087. AAAI/IAAI, 2008.
- [77] Simon EF Spencer, Steven M Hill, and Sach Mukherjee. Inferring network structure from interventional time-course experiments. *The Annals of Applied Statistics*, pages 507–524, 2015.
- [78] Peter Spirtes. Introduction to causal inference. *Journal of Machine Learning Research*, 11, 2010.
- [79] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2000.
- [80] Peter Spirtes, Christopher Meek, and Thomas S Richardson. An algorithm for causal inference in the presence of latent variables and selection bias. In Clark Glymour and Gregory F Cooper, editors, *Computation, Causation, and Discovery*. MIT Press, 1999.
- [81] Peter Spirtes and Thomas S Richardson. A polynomial time algorithm for determining DAG equivalence in the presence of latent variables and selection bias. In *Proceedings of the International Workshop on Artificial Intelligence and Statistics*, pages 489–500, 1996.
- [82] Peter Spirtes, Thomas S Richardson, and Chris Meek. The dimensionality of mixed ancestral graphs. Technical report, Philosophy Department, CMU, 1997. CMU-PHIL-83.



- [83] Milan Studený. *Probabilistic Conditional Independence Structures*. Springer, 2005.
- [84] Milan Studený. Mathematical aspects of learning Bayesian networks: Bayesian quality criteria. Technical report, Institute of Information Theory and Automation, Prague, 2008.
- [85] Milan Studený and James Cussens. Towards using the chordal graph polytope in learning decomposable models. *International Journal of Approximate Reasoning*, 88:259–281, 2017.
- [86] Milan Studený and David Haws. Learning Bayesian network structure: Towards the essential graph by integer linear programming tools. *International Journal of Approximate Reasoning*, 55:1043–1071, 2014.
- [87] Milan Studený, Raymond Hemmecke, and Silvia Lindner. Characteristic imset: a simple algebraic representative of a Bayesian network structure. In *Proceedings of the European Workshop on Probabilistic Graphical Models*, pages 257–264. HIIT Publications, 2010.
- [88] Sofia Triantafillou and Ioannis Tsamardinos. Score-based vs constraint-based causal learning in the presence of confounders. In *Workshop on Causation: Foundation to Application*, pages 59–67, 2016.
- [89] Konstantinos Tsirlis, Vincenzo Lagani, Sofia Triantafillou, and Ioannis Tsamardinos. On scoring maximal ancestral graphs with the max–min hill climbing algorithm. *International Journal of Approximate Reasoning*, 102:74–85, 2018.
- [90] Thomas S Verma and Judea Pearl. Equivalence and synthesis of causal models. Technical report, Department of Computer Science, University of California, Los Angeles, 1990.

- [91] N J Wildberger. A new look at multisets. preprint, 2003.
  
- [92] Jiji Zhang. Causal reasoning with ancestral graphs. *Journal of Machine Learning Research*, 9:1437–1474, 2008.
  
- [93] Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172:1873–1896, 2008.