Markov Bases: Their Complexity and Applications to Network Analysis

By

FÉLIX ALMENDRA HERNÁNDEZ

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

MATHEMATICS

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Jesús A. De Loera, Chair

Anne Schilling

Benjamin Morris

Committee in Charge

2025

Contents

Abstract

Algebraic statistics is an emerging field that employs tools from algebraic geometry, commutative algebra, and combinatorics to address statistical problems and their applications. This interdisciplinary subject not only applies algebraic techniques to solve statistical challenges but also fosters the development of new algebraic results motivated by statistical applications. This dynamic exchange has enriched both disciplines, driving advancements in areas such as experimental design, graphical models, and parametric inference.

In 1998, Diaconis and Sturmfels made a foundational contribution to the field by introducing a Markov Chain Monte Carlo algorithm for sampling fibers of log-linear models. The algorithm's inherently algebraic nature relies on the construction of a *Markov basis*, a set of moves with origin in polynomial algebra, thereby establishing a connection between commutative algebra, probability and statistics.

Since this seminal work, there has been a surge of research into various aspects of Markov bases and their structure for specific sets of discrete exponential families. This dissertation investigates the complexity of Markov bases in general and explores their combinatorial aspects in the context of popular random network models and their applications in statistics.

In Chapter 2, we extend the classical notion of Markov basis by allowing moves to connect fibers, even when these occasionally take steps in a negative relaxation of the fiber. This concept is motivated by earlier work of Bunea and Besag in the context of the Rasch model, and Chen, Dinwoodie, Dobra, and Huber in the context of logistic regression. These studies considered alternative methods for defining irreducible Markov chains on fibers without computing a full Markov basis in certain specific cases.

Nevertheless, we show that for general log-linear models, there is no universal upper bound on the level of negative relaxation required to connect fibers. Moreover, we extend a result by De Loera and Onn, showing that Markov basis elements with arbitrarily large degrees may exist even when relaxed fibers are employed. On the other hand, we provide positive results for hierarchical models, establishing a polynomial upper bound on the size of their Graver basis in terms of certain structural parameters of the model.

In Chapter 3, we present a combinatorial description of the Markov basis for a degree-corrected variant of the Stochastic Block Model (SBM), resolving an open problem posed by Karwa et al. in

their study of goodness-of-fit tests for mixtures of log-linear models. Furthermore, we establish that the algebraic counterparts of this Markov basis constitute a Gröbner basis for the associated toric ideal under a pure lexicographical order, thereby extending work by De Loera et al. Furthermore, we analyze the Markov basis for the 0/1 restricted fibers of this model, showing that its degree increases as the number of blocks in the SBM grows.

Finally, in Chapter 4, we consider a labeled generalization of the SBM and provide a complete combinatorial description of its Markov basis. This result is directly relevant to the conditional goodness-of-fit testing framework for mixtures of exponential log-linear models introduced by Karwa et al. We also provide theoretical guarantees for the test in the frequentist setting.

Acknowledgments

I would like to express my deepest gratitude to my advisor Jesús De Loera for his constant support, insightful guidance, and the many opportunities he has provided me throughout my academic journey. His encouragement has allowed me to pursue research, attend conferences, and connect with many mathematicians, often sharing my work with great enthusiasm. I am especially grateful for the dedication and passion he has invested in my development as a mathematician.

A special thank you to Sonja Petrović for being a great mentor and collaborator. I deeply appreciate her encouragement, insights, and support, as well as the research ideas she shared on Markov bases, many of which helped shape this dissertation. I am also grateful for her dedication to fostering the algebraic statistics community. Additionally, I would like to thank Miles Bakenhus, Vishesh Karwa, Mitsunori Ogawa, and Ruriko Yoshida for their collaboration and insightful discussions.

I am very thankful to UC Davis, particularly the faculty and staff of the Mathematics Department, for their hard work and commitment to supporting students.

My journey at Davis would not have been possible without the love and support of my wife, Maddie, my greatest source of strength. Thank you for moving across the country, leaving everything behind to join me in this adventure. I will always be grateful for your love, encouragement, and understanding.

To my friends—Acadia, Luna, Sharon, Reed, Shay, Hanna, Denae, and Toño—it has been a pleasure to meet you all and to share so many wonderful memories.

Without a doubt, my mathematical path was profoundly shaped by my involvement in the Mexican Math Olympiad. I am deeply grateful to everyone who made it possible, especially the Mexico City team led by Jorge and Isabel, as well as my teammates, with whom I shared the joy of solving mathematics problems.

To my parents, my brother, and my aunt, thank you for your unconditional love and support. I owe everything to you.

Lastly, I acknowledge the financial support I have received from NSF grant DMS-1818969, NSF grant DMS-2434665, NSF TRIPODS Award CCF-1934568, and NSF Grant DMS-1929348, as well as the JMM Travel Grant, which supported one of my trips.

CHAPTER 1

Introduction

This dissertation explores the complexity of Markov bases and their applications in log-linear models. To provide the necessary context for our contributions, this chapter begins with a discussion of preliminary background material, which is presented in two sections. Section 1.1 focuses on algebraic concepts and fundamental known results, while Section 1.2 addresses statistical theory, emphasizing its connections to the algebraic framework.

In Section 1.3, we review existing results on the behavior of Markov bases and present both positive and negative findings regarding their complexity. In Section 1.4, we share our results on the study of a Markov basis used to connect spaces of graphs with a fixed degree sequence. Finally, in Section 1.5, we explore the application of Markov bases for performing goodness-of-fit tests on a labeled version of the Stochastic Block Model.

The results presented in this dissertation are based on the published works [4, 5] and an additional paper currently in progress.

Notation. The vector of all zeros is denoted by $\mathbf{0}$, with its dimension inferred from the context. For any positive integer n, we write $[n] = \{1, \ldots, n\}$. Given an integer vector $\boldsymbol{u} = (u_1, \ldots, u_n) \in \mathbb{Z}^n$, we define its positive and negative parts as $\boldsymbol{u}^+ := (\max(u_j, 0))_{j=1}^n$ and $\boldsymbol{u}^- := (\max(-u_j, 0))_{j=1}^n$, respectively. For variables $x_1, \ldots, x_n$, the monomial $\boldsymbol{x}^{\boldsymbol{u}}$ is given by $x_1^{u_1} \cdots x_n^{u_n}$. For any vector $\boldsymbol{\omega} \in \mathbb{R}^n$, $\boldsymbol{\omega}^\top$ denotes its transpose, and $\boldsymbol{\omega}^\top \boldsymbol{\nu}$ represents the inner product of $\boldsymbol{\omega}$ and $\boldsymbol{\nu}$. Given a distribution π over Ω, the notation $\pi(\boldsymbol{u}) \propto f(\boldsymbol{u})$ indicates that the distribution is proportional to $f(\boldsymbol{u})$, with the proportionality constant being the reciprocal of the normalizing factor $\sum_{\boldsymbol{u} \in \Omega} f(\boldsymbol{u})$.

1.1. Algebraic Preliminaries and an Introduction to Markov Bases

In this section, we introduce the key concept of a Markov basis along with the necessary background. The material presented here is primarily drawn from [104], an excellent resource on toric ideals and their applications. Some results presented here have been rephrased to better align with the flow of our exposition. Comprehensive bibliography on algebraic statistics include

[7, 48, 106] and the references therein. The latter two books are particularly noteworthy for being largely self-contained, offering introductory material before progressing to more advanced topics.

We begin this subsection with the following fundamental definition.

DEFINITION 1.1.1. *Let $\mathbb{K}$ be an infinite field and $A = (a_{ij}) \in \mathbb{Z}^{d\times n}$ be an integer matrix with $\boldsymbol{a}_j = (a_{1j}, \ldots, a_{dj})$ for every $j \in [n]$. Let $\varphi_A : \mathbb{K}[x_1, \ldots, x_n] \to \mathbb{K}[t_1^{\pm}, \ldots, t_d^{\pm}]$ be the homomorphism of semigroup algebras induced by the map $x_j \mapsto \boldsymbol{t}^{\boldsymbol{a}_i}$. We define the **toric ideal** of A as the kernel of this homorphism, $I_A := \ker \varphi_A$.*

Toric ideals are of significant interest from a computational perspective and have connections to various fields, including numerical semigroups [57, 95], semigroup rings [59], commutative algebra and combinatorics [19, 89, 103], algebraic geometry [32, 58], linear algebra, and polyhedral geometry [97, 109].

Our work focuses particularly on their applications in algebraic statistics via Markov bases, as well as their relevance to integer programming through Gröbner and Graver bases. Some of the well-known properties of toric ideals include the following result.

PROPOSITION 1.1.2 ([104, Corollary 4.3]). *The toric ideal I_A is generated by the binomials $\{\boldsymbol{x}^{\boldsymbol{u}^+} - \boldsymbol{x}^{\boldsymbol{u}^-} : \boldsymbol{u} \in \ker_{\mathbb{Z}} A\}$.*

Now, we recall a classic result in commutative algebra and state one of its corollaries that will be of particular interest to us.

THEOREM 1.1.3. *[Hilbert basis theorem, [31, Theorem 4]] Every ideal of $\mathbb{K}[x_1, \ldots, x_n]$ is finitely generated.*

COROLLARY 1.1.4. *The toric ideal I_A is generated by a finite subset of the set of binomials $\{\boldsymbol{x}^{\boldsymbol{u}^+} - \boldsymbol{x}^{\boldsymbol{u}^-} : \boldsymbol{u} \in \ker_{\mathbb{Z}} A\}$.*

DEFINITION 1.1.5. *Let $A \in \mathbb{Z}^{d\times n}$ be an integer matrix such that $\ker_{\mathbb{Z}} A \cap \mathbb{N}^d = \{\boldsymbol{0}\}$ and $\boldsymbol{b} \in \mathbb{N}A := \{A\boldsymbol{u} : \boldsymbol{u} \in \mathbb{N}^n\}$. We define the **b-fiber** as the set*

$$\mathcal{F}(A, \boldsymbol{b}) := \{\boldsymbol{u} \in \mathbb{N}^n : A\boldsymbol{u} = \boldsymbol{b}\}.$$

*Whenever A is clear from the context, we may suppress the dependence on the matrix A and simply use the notation $\mathcal{F}(\boldsymbol{b})$. We refer to the elements of $\mathbb{N}A$ as **margins**.*

Fibers can be understood as the set of integer points inside the parametric polytopes $P(A, \boldsymbol{b}) := \{\boldsymbol{y} \in \mathbb{R}^n : \boldsymbol{y} \geq \boldsymbol{0}, A\boldsymbol{y} = \boldsymbol{b}\}$ as we let $\boldsymbol{b}$ vary over the set of margins. Some natural problems associated with these objects include enumeration [11, 37, 74], sampling [44, 45, 75] and integer programming [29, 71, 105]. As noted earlier, we will focus on the latter two.

DEFINITION 1.1.6. *For a finite subset $\mathcal{M} \subset \mathbb{Z}^n$, called the set of **moves**, and $\mathcal{F} \subset \mathbb{N}^n$, we define the **fiber graph** induced by $\mathcal{M}$ on $\mathcal{F}$ as the graph with vertex set $\mathcal{F}$ and such that $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{F}$ form an edge if an only if $\boldsymbol{u} - \boldsymbol{v} \in \pm\mathcal{M}$. We denote this graph by $\mathcal{F}_{\mathcal{M}}$.*

DEFINITION 1.1.7. *Let $A \in \mathbb{Z}^{d\times n}$ be an integer matrix such that $\ker_{\mathbb{Z}} A \cap \mathbb{N}^n = \{\boldsymbol{0}\}$. A finite subset $\mathcal{M} \subset \ker_{\mathbb{Z}} A$ is a **Markov basis** for A if $\mathcal{F}(\boldsymbol{b})_{\mathcal{M}}$ is a connected graph for every $\boldsymbol{b} \in \mathbb{N}A$. We say that $\mathcal{M}$ is a minimal Markov basis if it does not properly contain a Markov basis.*

THEOREM 1.1.8. *[Fundamental Theorem of Markov Basis, [106, Theorem 9.2.5]] Let $A \in \mathbb{Z}^{d\times n}$ be an integer matrix such that $\ker_Z A \cap \mathbb{N}^n = \{\boldsymbol{0}\}$. Then $\mathcal{M} \subset \ker_{\mathbb{Z}} A$ is a Markov basis for A if and only if the set of binomials $\{\boldsymbol{x}^{\boldsymbol{u}^+} - \boldsymbol{x}^{\boldsymbol{u}^-} : \boldsymbol{u} \in \mathcal{M}\}$ generates the toric ideal I_A.*

This theorem was first introduced in [45] and is one of the foundational results in the field of algebraic statistics. Moreover, Theorem 1.1.3, together with Corollary 1.1.4, guarantees the existence of a Markov basis for integer matrices whose integer kernel intersects the non-negative orthant only on $\{\boldsymbol{0}\}$.

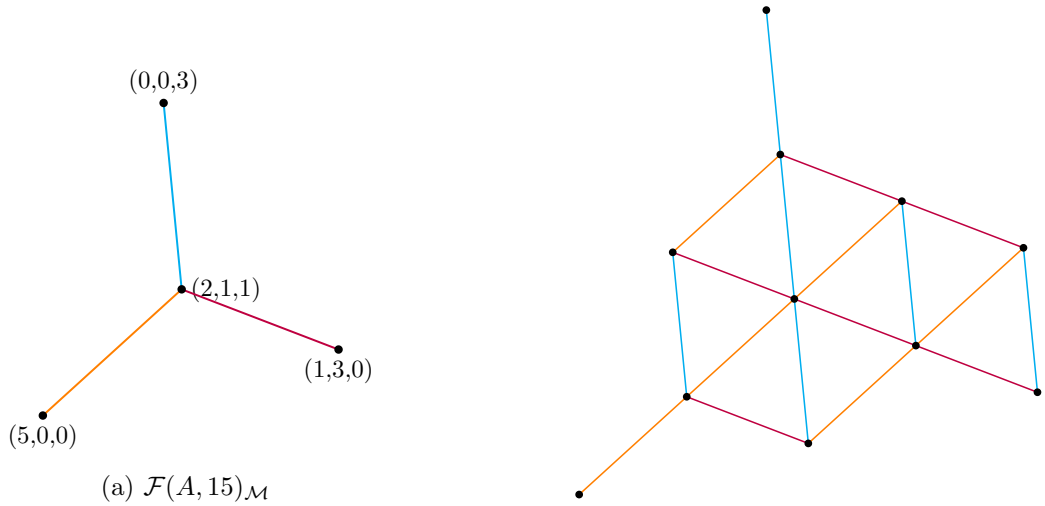
EXAMPLE 1.1.9. *For $A = (3\ 4\ 5)$ the toric ideal $I_A \subset \mathbb{K}[x_1, x_2, x_3]$ is generated by the binomials $\{x_1x_3 - x_2^2, x_1^3 - x_1x_2, x_1^2x_2 - x_3^2\}$. By the Fundamental Theorem of Markov basis this means that $\mathcal{M} = \{(1, -2, 1), (3, -1, -1), (2, 1, -2)\} \subset \ker_{\mathbb{Z}} A$ is a Markov basis for A. Figure 1.1 shows the fiber graphs $\mathcal{F}(A, b)_{\mathcal{M}}$ for margins $b = 15, 30$ induced by $\mathcal{M}$.*

Notice that all the elements in $\mathcal{M}$ are necessary to make $\mathcal{F}(A, 15)_{\mathcal{M}}$ connected. This means that $\mathcal{M}$ is minimal.

PROPOSITION 1.1.10. *Let $A \in \mathbb{Z}^{d\times n}$ be an integer matrix such that $\ker_Z A \cap \mathbb{N}^n = \{\boldsymbol{0}\}$ and let $\mathcal{M} \subset \ker_{\mathbb{Z}} A$. The following are equivalent*

(1) $\mathcal{M}$ is a Markov basis for A, i.e., $\mathcal{F}(A, \boldsymbol{b})_{\mathcal{M}}$ is connected for every $\boldsymbol{b} \in \mathbb{N}A$,

(2) $I_A = \langle \boldsymbol{x}^{\boldsymbol{u}^+} - \boldsymbol{x}^{\boldsymbol{u}^-} : \boldsymbol{u} \in \mathcal{M}\rangle$,



(a) $\mathcal{F}(A, 15)_{\mathcal{M}}$

(b) $\mathcal{F}(A, 30)_{\mathcal{M}}$

FIGURE 1.1. Fiber graphs induced by $\mathcal{M} = \{(1, -2, 1), (3, -1, -1), (2, 1, -2)\}$.

(3) For every $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{N}^n$ with $A\boldsymbol{u} = A\boldsymbol{v} = \boldsymbol{b}$, there exists a sequence $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_S \in \mathcal{M}$ such that

$$\boldsymbol{u} - \boldsymbol{v} = \sum_{s=1}^{S} \boldsymbol{m}_s \quad \text{and} \quad \boldsymbol{v} + \sum_{s=1}^{s'} \boldsymbol{m}_s \in \mathcal{F}(A, \boldsymbol{b}) \text{ for every } s' \leq S.$$

One of the most useful applications of Markov bases is their use in generating random samples from a distribution π on $\mathcal{F}(\boldsymbol{b})$ via the *Metropolis-Hastings algorithm* (see [86, Section 3.2]), as described in Algorithm 1. This approach is especially useful when the fibers are too large to be explicitly enumerated.

REMARK 1.1.11. *In most of the scenarios encountered throughout this work, we consider distributions π on $\mathcal{F}(\boldsymbol{b})$ such that $\pi(\boldsymbol{u}) \propto h(\boldsymbol{u})$ for a known function h. Consequently, during the Metropolis-Hastings step in line 6 of Algorithm 1 we can compute $\frac{h(\boldsymbol{u}_n + \boldsymbol{m})}{h(\boldsymbol{u}_n)}$ instead of $\frac{\pi(\boldsymbol{u}_n + \boldsymbol{m})}{\pi(\boldsymbol{u}_n)}$.*

THEOREM 1.1.12. *Let $\{\boldsymbol{u}_n\}_{n \geq 1}$ be the output of Algorithm 1 as $N \to \infty$. Then $\{\boldsymbol{u}_n\}_{n \geq 1}$ forms an irreducible Markov chain with stationary distribution π.*

The proof of this theorem relies on the connectivity properties inherent to Markov bases, as stated in Definition 1.1.7. For a general proof of this result, see [98, Chapter 6].

As a consequence of Theorem 1.1.8, finding a Markov basis for a matrix A is equivalent to identifying a set of generators for the toric ideal I_A. However, when the set of generators possesses

Algorithm 1: Fiber samples given a Markov basis.

Input : $\boldsymbol{u} \in \mathcal{F}(A, \boldsymbol{b})$, starting point in a fiber
$\mathcal{M}$, a Markov basis for A
π, a desired distribution on $\mathcal{F}(A, \boldsymbol{b})$
N, the number of fiber samples

Output: A sequence of points $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots$ in $\mathcal{F}(A, \boldsymbol{b})$

Set $\boldsymbol{u}_1 \leftarrow \boldsymbol{u}$;
for $n = 1, \ldots, N-1$ **do**
  Choose $\boldsymbol{m} \in \pm\mathcal{M}$ uniformly at random;
  if $\boldsymbol{u}_n + \boldsymbol{m} \geq \boldsymbol{0}$; ▷ This checks if $\boldsymbol{u}_n + \boldsymbol{m} \in \mathcal{F}(A, \boldsymbol{b})$
  then
    $\boldsymbol{u}_{n+1} \leftarrow \boldsymbol{u}_n + \boldsymbol{m}$ with probability $\min\left\{1, \frac{\pi(\boldsymbol{u}_n + \boldsymbol{m})}{\pi(\boldsymbol{u}_n)}\right\}$; ▷ Metropolis-Hastings step
  else
    $\boldsymbol{u}_{n+1} \leftarrow \boldsymbol{u}_n$;
Return sequence $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N$

additional properties, the Markov basis can be utilized not only for sampling but also for integer programming tasks, as we elaborate below.

DEFINITION 1.1.13. *A **monomial order** on $\mathbb{K}[x_1, \ldots, x_n]$ is a relation $\prec$ on the set of monomials $\mathfrak{M} = \{\boldsymbol{x}^{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \mathbb{Z}^n_{\geq 0}\}$ (or equivalently, the set $\mathbb{Z}^n_{\geq 0}$) satisfying the following:*

(i) $\prec$ is a total ordering on $\mathfrak{M}$.

(ii) If $\boldsymbol{x}^{\boldsymbol{\alpha}} \prec \boldsymbol{x}^{\boldsymbol{\beta}}$ then $\boldsymbol{x}^{\boldsymbol{\alpha}+\boldsymbol{\gamma}} \prec \boldsymbol{x}^{\boldsymbol{\beta}+\boldsymbol{\gamma}}$ for any $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma} \in \mathbb{Z}^n_{\geq 0}$.

(iii) $\prec$ is a well-ordering. This means that every nonempty subset of $\mathfrak{M}$ has a smallest element under $\prec$.

EXAMPLE 1.1.14. *The following are three examples of monomial orders that can be defined over the set of monomials $\mathfrak{M} = \{\boldsymbol{x}^{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \mathbb{Z}^n_{\geq 0}\}$.*

- *Lexicographic ($\prec_{lex}$): $\boldsymbol{x}^{\boldsymbol{\alpha}} \prec_{lex} \boldsymbol{x}^{\boldsymbol{\beta}}$ if the leftmost non-zero entry of $\boldsymbol{\beta} - \boldsymbol{\alpha}$ is positive.*
- *Graded reverse lexicographic: $\boldsymbol{x}^{\boldsymbol{\alpha}} \prec_{grevlex} \boldsymbol{x}^{\boldsymbol{\beta}}$ if $\sum_{i=1}^n \alpha_i < \sum_{i=1}^n \beta_i$, or $\sum_{i=1}^n \alpha_i = \sum_{i=1}^n \beta_i$ and the rightmost non-zero entry of $\boldsymbol{\beta} - \boldsymbol{\alpha}$ is negative.*
- *Weight order: Given a real vector $\boldsymbol{\omega} \in \mathbb{R}^n_{\geq 0}$ and an arbitrary monomial order $\prec$, we define the monomial order $\prec_{\boldsymbol{\omega}}$ as $\boldsymbol{x}^{\boldsymbol{\alpha}} \prec_{\boldsymbol{\omega}} \boldsymbol{x}^{\boldsymbol{\beta}}$ if $\boldsymbol{\omega}^\top \boldsymbol{\alpha} < \boldsymbol{\omega}^\top \boldsymbol{\beta}$ or if $\boldsymbol{\omega}^\top \boldsymbol{\alpha} = \boldsymbol{\omega}^\top \boldsymbol{\beta}$ and $\boldsymbol{\alpha} \prec \boldsymbol{\beta}$. Notice that the monomial order $\prec_{\boldsymbol{\omega}}$ extends the partial order induced by $\boldsymbol{\omega}$.*

Consider the four monomials $x_2^2, x_1x_3, x_1x_2x_3$ and $x_1x_3^2$. The monomial orders above would order these monomials as follows.

- *Lexicographic:* $x_2^2 \prec_{lex} x_1x_3 \prec_{lex} x_1x_3^2 \prec_{lex} x_1x_2x_3$
- *Graded reverse lexicographic:* $x_1x_3 \prec_{grevlex} x_2^2 \prec_{grevlex} x_1x_3^2 \prec_{grevlex} x_1x_2x_3$
- *Weight order with* $\boldsymbol{\omega} = (1, 3, 0)$ *and* $\prec = \prec_{lex}$*:* $x_1x_3 \prec_{\boldsymbol{\omega}} x_1x_3^2 \prec_{\boldsymbol{\omega}} x_1x_2x_3 \prec_{\boldsymbol{\omega}} x_2^2$

DEFINITION 1.1.15. *Given a monomial order* $\prec$ *and an ideal* $I \subset \mathbb{K}[x_1, \ldots, x_n]$*. A* ***Gröbner basis*** *for* I *is a finite collection of nonzero polynomials* $\mathcal{G} = \{g_1, \ldots, g_s\} \subset I$ *such that*

$$\langle in_\prec(g_1), \ldots, in_\prec(g_s)\rangle = \langle \{in_\prec(g) : g \in I\}\rangle,$$

where $in_\prec(g)$ *is the leading monomial in* g *with respect to the monomial order* $\prec$*. We say that* $\mathcal{G}$ *is* ***reduced*** *if every polynomial in* $\mathcal{G}$ *is monic and for each* $g \in \mathcal{G}$*, no monomial appearing in* g *lies in* $\langle in_\prec(f) : f \in \mathcal{G}\backslash\{g\}\rangle$*.*

PROPOSITION 1.1.16. *Let* $\prec$ *be a monomial order and* $I \subset \mathbb{K}[\boldsymbol{x}_1, \ldots, x_n]$ *an ideal. If* $\mathcal{G}_\prec$ *is a Gröbner basis for* I *with respect to* $\prec$*, then* $\mathcal{G}_\prec$ *generates* I *as an ideal.*

PROPOSITION 1.1.17. *Let* $I \subset \mathbb{K}[x_1, \ldots, x_n]$ *be an ideal. Then, for any monomial order,* I *has an unique reduced Gröbner basis with respect to the monomial order.*

For the reader interested in learning more about general Gröbner basis theory and its applications to computational commutative algebra we recommend the references [30, 31]and [65].

PROPOSITION 1.1.18. *Let* $\prec$ *be a monomial order and* $I \subset \mathbb{K}[x_1, \ldots, x_n]$ *be an ideal generated by binomials of the form* $\boldsymbol{x}^{\boldsymbol{u}} - \boldsymbol{x}^{\boldsymbol{v}}$ *(known as pure difference binomials). Then, any reduced Gröbner basis of* I *consists of pure difference binomials.*

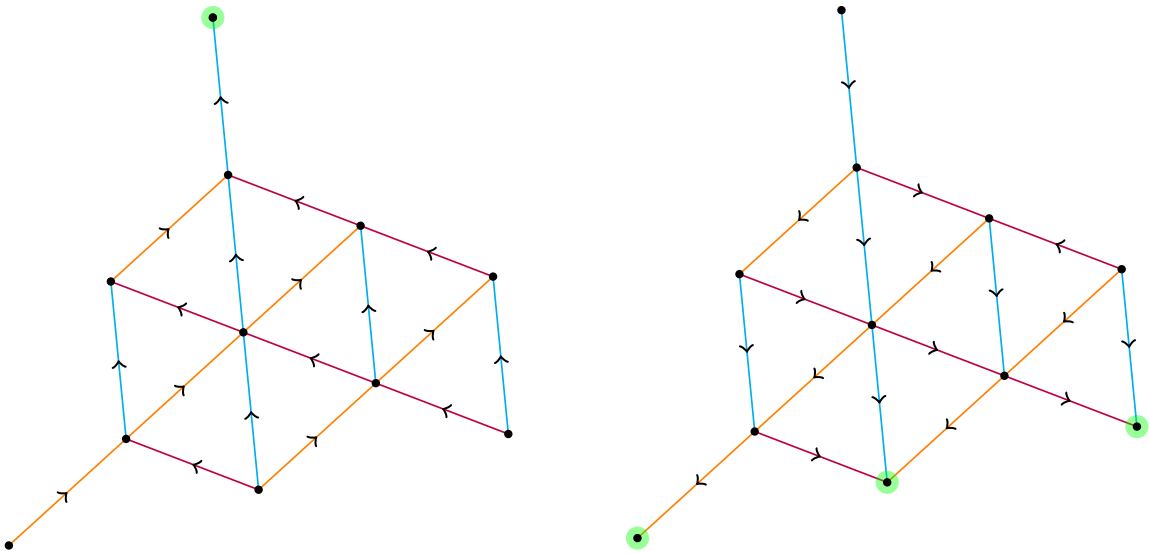
REMARK 1.1.19. *Whenever* $\{\boldsymbol{x}^{\boldsymbol{m}^+} - \boldsymbol{x}^{\boldsymbol{m}^-} : \boldsymbol{m} \in \mathcal{G}\}$ *is a Gröbner basis for a toric ideal* I_A*, it follows from Proposition 1.1.8 that* $\mathcal{G}$ *is a Markov basis for* A*.*

The previous remark implies that Gröbner basis can be used to generate random samples from a given distribution on $\mathcal{F}(A, \boldsymbol{b})$ as described in Algorithm 1. However, one of the additional applications of Gröbner basis for toric ideals relies on integer programming [29, 71, 105].

DEFINITION 1.1.20. *Let* $A \subset \mathbb{Z}^{d\times n}$ *be an integer matrix,* $\mathcal{G} \subset \ker_{\mathbb{Z}} A$ *and* $\prec$ *be a monomial order on* $\mathbb{Z}^n_{\geq 0}$*. We denote by* $\mathcal{F}(A, \boldsymbol{b})_{\mathcal{G},\prec}$ *the* ***fiber digraph*** *whose vertices correspond to* $\mathcal{F}(A, \boldsymbol{b})$ *and there is a directed edge from* $\boldsymbol{u}$ *to* $\boldsymbol{v}$ *if* $\boldsymbol{v} \prec \boldsymbol{u}$*.*

PROPOSITION 1.1.21. *Let $A \subset \mathbb{Z}^{d\times n}$ be an integer matrix, $\mathcal{G} \subset \ker_{\mathbb{Z}} A$ and $\prec$ be a monomial order on $\mathbb{Z}^n_{\geq 0}$. Then, $\mathcal{F}(A, \boldsymbol{b})_{\mathcal{G},\prec}$ has a unique sink for every $\boldsymbol{b} \in \mathbb{N}A$ if and only if the set of binomials $\{\boldsymbol{x}^{\boldsymbol{m}^+} - \boldsymbol{x}^{\boldsymbol{m}^-} : \boldsymbol{m} \in \mathcal{G}\}$ is a Gröbner basis for I_A with respect to $\prec$.*

EXAMPLE 1.1.22. *Let $A = (3\ 4\ 5)$ and $\mathcal{M} = \{(1,-2,1),(3,-1,-1),(2,1,-2)\} \subset \ker_{\mathbb{Z}} A$. Figure 1.2 illustrates the fiber digraphs $\mathcal{F}(A, 30)_{\mathcal{M},\prec}$ with two different monomial orders whose sinks have been highlighted with green. In (a), the reverse lexicographic order produces a digraph with a unique sink. In contrast, the digraph in (b) has three distinct sinks when using the weight order induced by $\boldsymbol{\omega} = (0, 1, 3)$ with ties broken using the lexicographic order.*



(a) $\mathcal{F}(A, 30)_{\mathcal{M},\prec_{grevlex}}$ (b) $\mathcal{F}(A, 30)_{\mathcal{M},\prec_{\boldsymbol{\omega}}}$

FIGURE 1.2. Fiber digraphs induced by $\mathcal{M} = \{(1,-2,1),(3,-1,-1),(2,1,-2)\}$ and different monomial orders. Sinks in each digraph are highlited with green.

In fact, $\{x^{\boldsymbol{m}^+} - \boldsymbol{x}^{\boldsymbol{m}^-} : \boldsymbol{m} \in \mathcal{M}\}$ is a Gröbner basis for I_A with respect to $\prec_{grevlex}$ but not with respect to $\prec_{\boldsymbol{\omega}}$ where $\boldsymbol{\omega} = (0, 1, 3)$.

As a consequence of Proposition 1.1.21, if $\prec_{\boldsymbol{\omega}}$ is a monomial ordering that extends the partial order induced by $\boldsymbol{\omega}$ (i.e., $\boldsymbol{\omega}^\top \boldsymbol{u} < \boldsymbol{\omega}^\top \boldsymbol{v}$ implies $\boldsymbol{x}^{\boldsymbol{u}} \prec_{\boldsymbol{\omega}} \boldsymbol{x}^{\boldsymbol{v}}$), then integer programs of the form

$$\begin{aligned} \min \quad & \boldsymbol{\omega}^\top \boldsymbol{u} \\ \text{subject to} \quad & A\boldsymbol{u} = \boldsymbol{b}, \\ & \boldsymbol{u} \geq \boldsymbol{0}; \end{aligned} \tag{1.1}$$

can be solved in polynomial time, provided we have access to a Gröbner basis of I_A with respect to $\prec_{\boldsymbol{\omega}}$. Moreover, a wider class of integer linear programs can be solved by considering generators of I_A that satisfy stricter conditions.

DEFINITION 1.1.23. *Let $A \in \mathbb{Z}^{d\times n}$ be an integer matrix. A binomial $\boldsymbol{x}^{\boldsymbol{u}^+} - \boldsymbol{x}^{\boldsymbol{u}^-} \in I_A$ is called **primitive** if there exists no other binomial $\boldsymbol{x}^{\boldsymbol{v}^+} - \boldsymbol{x}^{\boldsymbol{v}^-} \in I_A$ such that $\boldsymbol{x}^{\boldsymbol{v}^+}$ divides $\boldsymbol{x}^{\boldsymbol{u}^+}$ and $\boldsymbol{x}^{\boldsymbol{v}^-}$ divides $\boldsymbol{x}^{\boldsymbol{u}^-}$.*

DEFINITION 1.1.24. *The **Graver basis** of an integer matrix $A \in \mathbb{Z}^{d\times n}$ is defined as the set $Gr(A) := \{\boldsymbol{u} \in \mathbb{Z}^n_{\geq 0} : \boldsymbol{x}^{\boldsymbol{u}^+} - \boldsymbol{x}^{\boldsymbol{u}^-}$ is primitive$\}$.*

REMARK 1.1.25. *As shown in [35], the Graver basis $Gr(A)$ can be equivalently defined as the set of all $\sqsubseteq$-minimal elements in $\ker_{\mathbb{Z}} A \setminus \{\mathbf{0}\}$, where for any $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{N}^n$, we say $\boldsymbol{u} \sqsubseteq \boldsymbol{v}$ if and only if $u_j v_j \geq 0$ and $|u_j| \leq |v_j|$ for every $j \in [n]$.*

PROPOSITION 1.1.26. *For any reduced Gröbner basis $\{\boldsymbol{x}^{\boldsymbol{u}^+} - \boldsymbol{x}^{\boldsymbol{u}^-} : \boldsymbol{u} \in \mathcal{G}\}$ for I_A, (with respect to any monomial order), $\mathcal{G}$ is a subset of the Graver basis of A.*

The Graver basis of a matrix is finite, and an immediate implication of the previous result is that any Graver basis is also a Markov basis, i.e., $I_A = \langle \boldsymbol{x}^{\boldsymbol{u}^+} - \boldsymbol{x}^{\boldsymbol{u}^-} : \boldsymbol{u} \in Gr(A)\rangle$. In general, the Graver basis of a matrix A is significantly larger and more complex than a minimal Markov basis of that matrix, as one would expect given Proposition 1.1.29. An illustrative example involving an small matrix A is provided in Example 1.1.27.

EXAMPLE 1.1.27. *Let $A = (3\ 4\ 5)$. The Graver basis for this matrix is given by the set $Gr(A) = \{(1,-2,1),(3,-1,-1),(2,1,-2),(4,-3,0),(5,0,-3),(1,3,-3),(0,5,-4)\}$ which consists of 4 more elements than the minimal Markov basis $\mathcal{M} = \{(1,-2,1),(3,-1,-1),(2,1,-2)\} \subset Gr(A)$. In Figure 1.3 the edges induced by $\mathcal{M}$ are black while the edges induced by $Gr(A)\backslash\mathcal{M}$ are red in order to illustrate the difference between the fiber graphs induced by the different sets, $\mathcal{M}$ and $Gr(A)$.*

Another implication of the previous result is that Graver basis can be used to solve any integer linear program of the form (1.1) for any $\boldsymbol{\omega}$, provided an initial feasible point is available. Furthermore, there exist even stronger applications of Graver bases.

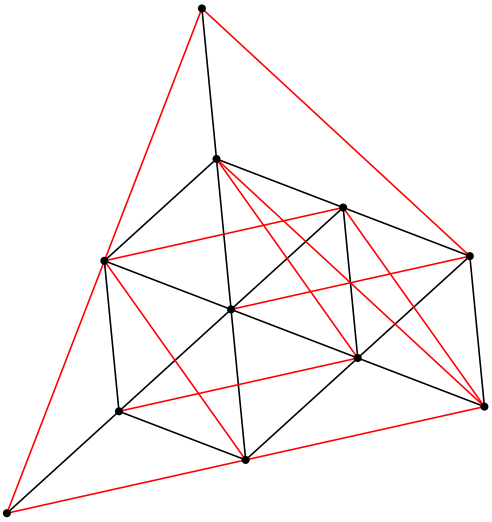


FIGURE 1.3. Fiber graph $\mathcal{F}(A, 30)_{Gr(A)}$ induced by $Gr(A)$.

DEFINITION 1.1.28. *Let $A \in \mathbb{Z}^{d\times n}$ be an integer matrix and a margin $\boldsymbol{b} \in \mathbb{N}A$. For $\boldsymbol{l}, \boldsymbol{L} \in \mathbb{Z}^n$ we define the **$(\boldsymbol{b}; \boldsymbol{l}, \boldsymbol{L})$-fiber** of A as the set*

$$\mathcal{F}(A, \boldsymbol{b}; \boldsymbol{l}, \boldsymbol{L}) := \{\boldsymbol{u} \in \mathbb{Z}^n : A\boldsymbol{u} = \boldsymbol{b}, \text{ and } l_i \leq u_i \leq L_i \text{ for every } i \in [n]\}.$$

We will omit $\boldsymbol{b}$ from both the name and the notation when this is clear from the context.

PROPOSITION 1.1.29 ([106]). *Let $A \in \mathbb{Z}^{d\times n}$ be an integer matrix and $Gr(A)$ its Graver basis. Then, for every $\boldsymbol{b} \in \mathbb{N}A$ and any $\boldsymbol{l}, \boldsymbol{L} \in \mathbb{Z}^n_{\geq 0}$, the fiber graph $\mathcal{F}(A, \boldsymbol{b}; \boldsymbol{l}, \boldsymbol{L})_{Gr(A)}$ is connected.*

PROPOSITION 1.1.30 ([60]). *Let $A \in \mathbb{Z}^{d\times n}$ be an integer matrix, $\boldsymbol{b} \in \mathbb{N}A$, $\boldsymbol{l}, \boldsymbol{L} \in \mathbb{Z}^n_{\geq 0}$ and $\boldsymbol{\omega} \in \mathbb{R}^n_{\geq 0}$. Then, for every $\boldsymbol{u} \in \mathcal{F}(A, \boldsymbol{b}; \boldsymbol{l}, \boldsymbol{L})$ non-minimal (with respect to $\boldsymbol{\omega}$), there exists $\boldsymbol{v} \in Gr(A)$ such that $\boldsymbol{u} + \boldsymbol{v} \in \mathcal{F}(A, \boldsymbol{b}; \boldsymbol{l}, \boldsymbol{L})$ and such that $\boldsymbol{\omega}^\top(\boldsymbol{u} + \boldsymbol{v}) < \boldsymbol{\omega}^\top \boldsymbol{u}$.*

In other words, the previous results implies that whenever we have access to $Gr(A)$, it is possible to solve any integer linear program of the form

$$\begin{aligned} \min \quad & \boldsymbol{\omega}^\top \boldsymbol{u} \\ \text{subject to} \quad & A\boldsymbol{u} = \boldsymbol{b}, \\ & \boldsymbol{L} \geq \boldsymbol{u} \geq \boldsymbol{l}; \end{aligned} \tag{1.2}$$

for arbitrary choices of $\boldsymbol{b} \in \mathbb{N}A, \boldsymbol{l}, \boldsymbol{L} \in \mathbb{Z}^n_{\geq 0}$ and $\boldsymbol{\omega} \in \mathbb{R}^n_{\geq 0}$. Moreover, Graver basis can be used to solve integer programs for a broader class of function called *convex separable* (see [35, Chapter 3]).

Several algorithms for computing a generating set of binomials for I_A (equivalently, a Markov basis for A) exist in the literature. The earliest such algorithms were introduced in [29, 94] and relied on computing a Gröbner basis using the Buchberger algorithm. More efficient algorithms, employing different techniques, were subsequently developed in [16, 17, 70]. The most efficient algorithm to date is the "Project and Lift" algorithm, introduced in [67], which computes a sequence of Gröbner bases for a hierarchy of projections of the integer kernel. This algorithm is implemented in the software package `4ti2` [1].

However, `4ti2`'s implementation is practical only for small matrices. For instance, [67, Section 6] presents an example of a matrix $A \in \mathbb{Z}^{48\times 64}$ of rank 37, where the Project and Lift algorithm requires approximately two days to compute its Markov basis. Moreover, [27] shows that Gröbner and Graver basis computations are strongly NP-hard in the general case.

Therefore, from an application standpoint, having access to a complete description of the Markov basis, Gröbner basis, or Graver basis for a given matrix A is highly advantageous, and the specific basis choice depends on the application at hand.

Since the publication of the seminal paper [45], there has been a surge of research examining various aspects of Markov bases and their structure for matrices associated with particular statistical models. Notable contributions include [8, 23, 39, 46, 47, 96], among others. While it is impossible to cite the entire bibliography, we recommend the comprehensive references [7, 48, 106]. These books provide a compilation of the extensive work that has been conducted in this field so far.

Before presenting some of the most prominent results, we first discuss the application of Markov bases to goodness-of-fit tests in statistical models.

1.2. Markov Bases in Probability and Statistics

Denote by $\boldsymbol{X} = (X_1, \ldots, X_m)$ a discrete random vector in the state space $\mathcal{I} = [r_1] \times \cdots \times [r_m]$ and define the joint probabilities

$$p(\boldsymbol{i}) = p(i_1, \ldots, i_m) = \mathbb{P}(X_1 = i_1, \ldots, X_m = i_m), \quad \forall \boldsymbol{i} \in \mathcal{I},$$

These form a joint probability table $\boldsymbol{p} = (p(\boldsymbol{i}) : \boldsymbol{i} \in \mathcal{I})$.

DEFINITION 1.2.1. *Let $A \in \mathbb{Z}^{d\times|\mathcal{I}|}$ be an integer matrix with $\mathbf{1} \in$ rowspan(A), whose columns are indexed by $\mathcal{I}$. Let $A(\cdot,\boldsymbol{i}) \in \mathbb{Z}^d$ denote the $\boldsymbol{i}$-th column of A, and let $\boldsymbol{h} = (h(\boldsymbol{i}) : \boldsymbol{i} \in \mathcal{I}) \in \mathbb{R}^{|\mathcal{I}|}_{>0}$ be a base measure. The* ***log-affine model*** *associated with A and $\boldsymbol{h}$ is the family of probability distributions $\mathcal{L}_{A,\boldsymbol{h}} := \{\boldsymbol{p_\theta}\}_{\boldsymbol{\theta}\in\Theta}$ such that*

$$p_{\boldsymbol{\theta}}(\boldsymbol{i}) = \frac{h(\boldsymbol{i})}{\psi(\boldsymbol{\theta})}\exp\{\eta(\boldsymbol{\theta})^\top A(\cdot,\boldsymbol{i})\} \quad \textit{for every } \boldsymbol{i}\in\mathcal{I}, \tag{1.3}$$

where $\boldsymbol{\theta} = (\theta_1,\dots,\theta_d)$ is the vector of model parameters, $\eta : \mathbb{R}^d \to \mathbb{R}^n$ is the natural parameter, and $\psi(\boldsymbol{\theta})$ is the normalizing constant. The matrix A is referred to as the ***design matrix*** *for the model $\mathcal{L}_{A,\boldsymbol{h}}$. When $\boldsymbol{h} = \mathbf{1}$, the model is called* ***log-linear*** *and is denoted simply by $\mathcal{L}_A$.*

From now on, when considering a log-affine model $\mathcal{L}_{A,\boldsymbol{h}}$, we will interchangeably refer to the Markov basis of A or the Markov basis of the model $\mathcal{L}_{A,\boldsymbol{h}}$.

REMARK 1.2.2. *One of the implications of having $\mathbf{1} \in$ rowspan(A) is that a probability vector (table) $\boldsymbol{p}$ belongs to $\mathcal{L}_{A,\boldsymbol{h}}$ if and only if $\log \boldsymbol{p}$ belongs to the affine space $\log \boldsymbol{h}$ + rowspan(A).*

EXAMPLE 1.2.3 (Independence model). *Consider a probability vector $\boldsymbol{p} = (p(i_1,i_2) : i_1 \in [r_1], i_2 \in [r_2])$, where $p(i_1,\cdot)$ and $p(\cdot,i_2)$ represent the marginal probabilities. Suppose that $\boldsymbol{p}$ belongs to the independence model $\mathcal{L}_{X\perp\!\!\!\perp Y}$, meaning that $p(i_1,i_2) = p(i_1,\cdot)p(\cdot,i_2)$ for every $(i_1,i_2) \in [r_1]\times[r_2]$. Let $\boldsymbol{\alpha} = (\alpha_{i_1} : i_1 \in [r_1])$ and $\boldsymbol{\beta} = (\beta_{i_2} : i_2 \in [r_2])$ be parameter vectors such that*

$$p(i_1,\cdot) = \frac{e^{\alpha_{i_1}}}{\psi(\boldsymbol{\alpha})} \quad \textit{and} \quad p(\cdot,i_2) = \frac{e^{\beta_{i_2}}}{\psi(\boldsymbol{\beta})} \quad \textit{for every } (i_1,i_2)\in[r_1]\times[r_2],$$

where $\psi(\boldsymbol{\alpha})$ and $\psi(\boldsymbol{\beta})$ are normalizing constants for the marginal probabilities. Then, it follows that $p(i_1,i_2) \propto \exp\{\boldsymbol{\theta}^\top A(\cdot,(i_1,i_2))\}$ for $\boldsymbol{\theta} = (\boldsymbol{\alpha},\boldsymbol{\beta})$ and a matrix $A \in \mathbb{Z}^{(r_1+r_2)\times r_1 r_2}$ such that the rows are labeled by the parameter vector $(\boldsymbol{\alpha},\boldsymbol{\beta})$ and the (i_1,i_2)-th column of A has 1s only at rows i_1, i_2 and 0s everywhere else. In other words, the independence model is log-linear with design matrix A.

For a concrete example, consider $r_1 = 2$ and $r_2 = 3$. Then, the design matrix for the model is the 5×6 matrix

$$A = \begin{array}{c} \begin{matrix} 11 & 12 & 13 & 21 & 22 & 23 \end{matrix} \\ \left(\begin{array}{cccccc} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ \hline 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{array}\right) \begin{matrix} \alpha_1 \\ \alpha_2 \\ \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{matrix} \end{array}.$$

The model representation of interest to us arises when data are arranged in a contingency table that cross-classifies items according to m categories. Specifically, consider a random sample of N independent and identically distributed vectors $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(N)} \in \mathcal{I} = [r_1] \times \cdots \times [r_m]$, and let $\boldsymbol{U} = (U(\boldsymbol{i}) : \boldsymbol{i} \in \mathcal{I})$ denote the m**-way table** of format $r_1 \times \cdots \times r_m$. Here, $U(\boldsymbol{i}) = \#\{k : \boldsymbol{X}^{(k)} = \boldsymbol{i}\}$ records the number of times $\boldsymbol{i}$ is observed. As we demonstrate below, to compute the likelihood of observing $\boldsymbol{U}$, all that is needed is the "summarized" data $A\boldsymbol{U}$.

PROPOSITION 1.2.4. *Let $\boldsymbol{p} = (p(\boldsymbol{i}) : \boldsymbol{i} \in \mathcal{I})$ be a probability vector. If $\boldsymbol{p} = \boldsymbol{p}_{\boldsymbol{\theta}} \in \mathcal{L}_{A,\boldsymbol{h}}$ and $\boldsymbol{u} \in \mathbb{Z}_{\geq 0}^{\mathcal{I}}$ satisfies $\sum_{\boldsymbol{i} \in \mathcal{I}} u(\boldsymbol{i}) = N$ and $A\boldsymbol{u} = \boldsymbol{b}$, then*

$$\mathbb{P}(\boldsymbol{U} = \boldsymbol{u} \mid \boldsymbol{\theta}) = \frac{\boldsymbol{h}^{\boldsymbol{u}} N!}{\prod_{\boldsymbol{i} \in \mathcal{I}} u(\boldsymbol{i})!} \exp\{\eta(\boldsymbol{\theta})^\top A\boldsymbol{u}\},$$

and the conditional probability $\mathbb{P}(\boldsymbol{U} = \boldsymbol{u} \mid A\boldsymbol{U} = A\boldsymbol{u}, \boldsymbol{\theta})$ does not depend on $\boldsymbol{\theta}$. Moreover,

$$\mathbb{P}(\boldsymbol{U} = \boldsymbol{u} \mid A\boldsymbol{U} = A\boldsymbol{u}, \boldsymbol{\theta}) = \frac{\boldsymbol{h}^{\boldsymbol{u}} / \big(\prod_{\boldsymbol{i} \in \mathcal{I}} u(\boldsymbol{i})!\big)}{\sum_{\boldsymbol{v} \in \mathcal{F}(A,\boldsymbol{b})} \boldsymbol{h}^{\boldsymbol{v}} / \big(\prod_{\boldsymbol{i} \in \mathcal{I}} v(\boldsymbol{i})!\big)}.$$

REMARK 1.2.5. *In the previous proposition, $\mathcal{I} = [r_1] \times \cdots \times [r_m]$, $A \in \mathbb{Z}^{d \times |\mathcal{I}|}$ is an integer matrix whose columns are indexed by the set $\mathcal{I}$, and $\boldsymbol{u} \in \mathbb{Z}_{\geq 0}^{\mathcal{I}}$ is an $r_1 \times \cdots \times r_m$-table. Hence, we implicitly assume that the matrix multiplication $A\boldsymbol{u}$ is performed by considering a vectorization of the table $\boldsymbol{u}$, with entries ordered according to the column indexing of A, which we assume to be lexicographic unless otherwise specified. To avoid introducing additional notation, we use $\boldsymbol{u}$ to refer to both the m-way table and its vectorization; the intended interpretation will usually be clear from the context.*

Given Proposition 1.2.4, when considering a log-affine model, we write the conditional probability of $\boldsymbol{U}$ given $\boldsymbol{u}$ as $\mathbb{P}(\boldsymbol{U} = \boldsymbol{u} \mid A\boldsymbol{U} = A\boldsymbol{u})$, omitting the vector of parameters $\boldsymbol{\theta}$ from the notation. The vector $A\boldsymbol{u}$ is known as the vector of **sufficient statistics** for the model $\mathcal{M}_{A,\boldsymbol{h}}$ in the statistics literature (see [22, Section 6.2]).

EXAMPLE 1.2.6 (1.2.3, continued). *In the independence model, if we consider a table of counts* $\boldsymbol{u} = (u_{i_1 i_2} : i_1 \in [r_1], i_2 \in [r_2])$, *which we think of as "vectorized," we can see that in the concrete example* $r_1 = 2, r_2 = 3$

$$A\boldsymbol{u} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} u_{11} \\ u_{12} \\ u_{13} \\ u_{21} \\ u_{22} \\ u_{23} \end{pmatrix} = \begin{pmatrix} u_{1+} \\ u_{2+} \\ u_{+1} \\ u_{+2} \\ u_{+3} \end{pmatrix},$$

where u_{i_1+} *and* u_{+i_2} *represent the* i_1*-th row sum and* i_2*-th column sum of* $\boldsymbol{u}$, *respectively.*

In the general setting, for a vector of counts $\boldsymbol{u} \in \mathbb{Z}_{\geq 0}^{r_1 \times r_2}$, *the vector of sufficient statistics* $A\boldsymbol{u}$ *consists of the row sums and column sums of* $\boldsymbol{u}$. *One can observe that* $\ker_{\mathbb{Z}} A = \{\boldsymbol{u} \in \mathbb{Z}^{r_1 \times r_2} : u_{i_1+} = 0$ *for all* i_1, *and* $u_{+i_2} = 0$ *for all* $i_2\}$. *In other words,* $\ker_{\mathbb{Z}} A$ *is the set of tables with integer entries whose row sums and column sums are 0.*

Now, consider N independent and identically distributed vectors $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(N)} \in \mathcal{I}$ sampled according to a distribution $\boldsymbol{p} = (p(\boldsymbol{i}) : \boldsymbol{i} \in \mathcal{I})$, and let $\boldsymbol{U} = (u(\boldsymbol{i}) : \boldsymbol{i} \in \mathcal{I})$ denote their vector of counts, as defined previously. Also, consider a log-affine model $\mathcal{L}_{A,\boldsymbol{h}} = \{\boldsymbol{p}_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta} \in \Theta}$. We aim to test the hypothesis

$$H_0 : \boldsymbol{p} \in \mathcal{L}_{A,\boldsymbol{h}} \quad \text{against} \quad H_1 : \boldsymbol{p} \notin \mathcal{L}_{A,\boldsymbol{h}},$$

which is commonly referred to as a **goodness-of-fit test** for the model $\mathcal{L}_{A,\boldsymbol{h}}$. See [111, Section 10.8] and the references therein for a general treatment of hypothesis testing and goodness-of-fit tests. A standard approach for reporting the results of such a hypothesis test involves computing and presenting the value of a specific test statistic called the p-value. Informally, the p-value measures the likelihood of observing $\boldsymbol{U}$ under the assumption that H_0 is true.

DEFINITION 1.2.7. *A p-value $p(\boldsymbol{U})$ is a test statistic satisfying $0 \leq p(\boldsymbol{u}) \leq 1$ for every sample point $\boldsymbol{u}$. Small values of $p(\boldsymbol{U})$ give evidence that H_1 is true. A p-value is* ***valid*** *if, for every $\boldsymbol{\theta} \in \Theta$ and every $1 \leq \alpha \leq 1$,*

$$\mathbb{P}(p(\boldsymbol{U}) \leq \alpha \mid \boldsymbol{\theta}) \leq \alpha.$$

If $p(\boldsymbol{U})$ is a valid p-value, we can construct a **level α test** based on it: the test rejects H_0 if and only if $p(\boldsymbol{U}) \leq \alpha$. The level $\alpha = .05$ is commonly used in practice.

As a consequence of Proposition 1.2.4, we know that if the **null hypothesis** H_0 is true, then the conditional distribution of $\boldsymbol{U}$ given $\boldsymbol{AU} = \boldsymbol{u}$ does not depend on $\boldsymbol{\theta}$. Consequently, if $W(\boldsymbol{U})$ is a test statistic where larger values provide evidence in favor of H_1, the test statistic $p(\boldsymbol{U})$, defined as

$$p(\boldsymbol{u}) := \mathbb{P}(W(\boldsymbol{U}) \geq W(\boldsymbol{u}) \mid A\boldsymbol{U} = A\boldsymbol{u}), \tag{1.4}$$

is a valid p-value, also known in the literature as a conditional p-value. For further details on conditional p-values, we refer the reader to [22, Subsection 8.3.4].

While we do not study test statistics $W(\boldsymbol{U})$ in detail, one natural test statistic that generalizes Fisher's exact test is the **chi-square statistic**

$$X^2(\boldsymbol{U}) := \sum_{\boldsymbol{i} \in \mathcal{I}} \frac{(u(\boldsymbol{i}) - \hat{u}(\boldsymbol{i}))^2}{\hat{u}(\boldsymbol{i})},$$

where $\hat{u}(\boldsymbol{i}) = N\hat{p}(\boldsymbol{i})$ and $\{\hat{p}(\boldsymbol{i})\}_{\boldsymbol{i} \in \mathcal{I}}$ are the maximum likelihood estimates, which can be computed using iterative proportional scaling when there are no exact formulas for them (see [106, Section 7.3]).

One useful application of Markov bases is computing (or approximating) the exact p-value, as shown by the following result, which follows from Theorem 1.1.12 and the law of large numbers for Markov chains, commonly known as the *Ergodic Theorem*. A general proof of this theorem is provided in [98, Theorem 6.63].

THEOREM 1.2.8. *Let $\{\boldsymbol{u}_n\}_{n \geq 1}$ be the output of Algorithm 1 as $N \to \infty$ with $\boldsymbol{u}_1 = \boldsymbol{u}$ and $\pi = \mathbb{P}(\cdot \mid A\boldsymbol{u})$. Let $W : \mathcal{F}(A\boldsymbol{u}) \to \mathbb{R}$ be any function. Then,*

$$\mathbb{P}\left(\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}_{W(\boldsymbol{u}_n) \geq W(\boldsymbol{u})} = \mathbb{P}(W(\boldsymbol{U}) \geq W(\boldsymbol{u}) \mid A\boldsymbol{U} = A\boldsymbol{u}) \right) = 1.$$

In practice, if $W(\boldsymbol{U})$ is a test statistic for which larger values provide evidence that H_1 is true (such as the chi-square test), we can approximate the p-value, as defined in Equation 1.4, by using a slight modification of Algorithm 1, as shown in Algorithm 2 below.

Algorithm 2: p-value approximation given a Markov basis

Input : An observed vector of counts $\boldsymbol{u} \in \mathbb{Z}^{\mathcal{I}}_{\geq 0}$ with $A\boldsymbol{u} = \boldsymbol{b}$
$\mathcal{M}$, a Markov basis for A
base measure $\boldsymbol{h} \in \mathbb{R}^{\mathcal{I}}_{>0}$
$W(\cdot)$, test statistic
N, number of fiber samples

Output: p-value as in (1.4)

1 Set $\boldsymbol{u}_1 \leftarrow \boldsymbol{u}$ and $p_1 \leftarrow 1$;
2 **for** $n = 1, \ldots, N-1$ **do**
3 Choose $\boldsymbol{m} \in \pm\mathcal{M}$ uniformly at random;
4 **if** $\boldsymbol{u}_n + \boldsymbol{m} \geq \boldsymbol{0}$; ▷ This checks if $\boldsymbol{u}_n + \boldsymbol{m} \in \mathcal{F}(A, \boldsymbol{b})$
5 **then**
6 $\boldsymbol{u}_{n+1} \leftarrow \boldsymbol{u}_n + \boldsymbol{m}$ with probability $\min\left\{1, \frac{\mathbb{P}(\boldsymbol{U}=\boldsymbol{u}_n+\boldsymbol{m} | A\boldsymbol{U}=\boldsymbol{b})}{\mathbb{P}(\boldsymbol{U}=\boldsymbol{u}_n | A\boldsymbol{U}=\boldsymbol{b})}\right\}$;
 ▷ Metropolis-Hastings step
7 **else**
8 $\boldsymbol{u}_{n+1} \leftarrow \boldsymbol{u}_n$;
9 $p_{n+1} = p_n + \mathbb{1}_{W(\boldsymbol{u}_{n+1}) \geq W(\boldsymbol{u})}$
10 Return sequence $\frac{1}{N} p_N$

REMARK 1.2.9. *Thanks to Proposition 1.2.4 the ratio computation in line 6 of Algorithm 2 can be done by computing* $\frac{\boldsymbol{h}^{\boldsymbol{u}_n+\boldsymbol{m}}\left(\prod_{i\in\mathcal{I}}(\boldsymbol{u}_n)_i!\right)}{\boldsymbol{h}^{\boldsymbol{u}_n}\left(\prod_{i\in\mathcal{I}}(\boldsymbol{u}_n+\boldsymbol{m})_i!\right)}$.

EXAMPLE 1.2.10. *Consider a hypothetical experiment designed to evaluate the effectiveness of two different pain relief drugs, A and B, on a sample of 40 individuals. Each participant rated their pain on a scale from 1 to 5, one hour after receiving the drug. The responses are summarized in Table 1.1.*

Drug	***Pain Score***					
	1	***2***	***3***	***4***	***5***	***Total***
A	*17*	*2*	*1*	*0*	*1*	*22*
B	*7*	*3*	*4*	*3*	*2*	*18*
Total	*24*	*5*	*5*	*3*	*3*	*40*

TABLE 1.1. Summary of responses in the drug effectiveness study, categorized by drug type and pain score.

In this scenario, we aim to test the independence between the type of pain relief drug and the pain score. The p-value for the observed data can be computed using Algorithm 2 through the `algstat` *package in* `R` *[77]. This computation involves the chi-square statistic and the Markov basis for the corresponding independence model, as described in Proposition 1.3.1 below. The resulting p-value for the exact test is* 11×10^{-3}*, indicating that we should reject the null hypothesis of independence.*

1.3. The Complexity of Markov Bases

In the previous section, we outlined a method for performing goodness-of-fit tests for log-linear models that relies on access to a Markov basis for the corresponding design matrix. Hierarchical log-linear models, one of the most significant classes of log-linear models in statistics, have been the focus of considerable research aimed at understanding their Markov bases. Notable contributions that describe scenarios where hierarchical models have well-behaved Markov bases include [46, 47, 72]. However, [39] demonstrated that, in general, Markov bases can be arbitrarily complex.

Motivated by this result, we introduce an extension of the Markov basis concept that enables fiber connectivity through negative fiber relaxations. These relaxations have been used in [20, 25, 113] as alternative methods for defining irreducible Markov chains on fibers with simple moves. We show that even in this setting, arbitrarily complex moves may still be required to ensure the irreducibility of the Markov chains.

On the other hand, we provide an upper bound on the size of the Graver basis for certain hierarchical models. Furthermore, it is worth noting that, driven by applications in integer and sparse integer programming, substantial effort has been devoted to understanding the complexity of Graver bases. Recent advances in integer programming have led to new and improved bounds on the maximum one-norm for Graver basis elements, which depend on the treedepth of the matrix. For further details, see [49, 82, 83] and the references therein.

We begin this section by reviewing known results on Markov bases before presenting our own contributions, starting with a description of the Markov basis for the independence model.

PROPOSITION 1.3.1. *Let* $A \in \mathbb{Z}^{(r_1+r_2)\times r_1 r_2}$ *be the design matrix of the independence model* $\mathcal{L}_{X \perp\!\!\!\perp Y}$ *as described in Example 1.2.3. For every* $(i_1, i_2) \in [r_1] \times [r_2]$ *let* $\boldsymbol{e}_{i_1 i_2} \in \mathbb{Z}^{r_1 r_2}$ *be the vector with 1 at the everywhere but at the* (i_1, i_2)*-th position. Then*

$$\mathcal{M} = \{\boldsymbol{e}_{i_1 i_2} + \boldsymbol{e}_{j_1 j_2} - \boldsymbol{e}_{i_1 j_2} - \boldsymbol{e}_{j_1 i_2} : 1 \leq i_1 < i_2 \leq r_1, 1 \leq j_1 < j_2 \leq r_2\}$$

is a minimal Markov basis for A.

The independence model is a specific example of a *hierarchical model*, and as stated in Proposition 1.3.1, its Markov basis can be described in a very compact way. More generally, as with the independence model, the sufficient statistics for hierarchical models consist of a collection of sums across higher-dimensional tables. However, we will observe that the size of the elements in a minimal Markov basis can grow significantly depending on the parameters of the model.

To formally define a hierarchical model, we need to introduce some notation, which is primarily drawn from [7].

DEFINITION 1.3.2. *Let $\mathcal{I} = [r_1] \times \cdots \times [r_m]$ and $\boldsymbol{i} = (i_1, \ldots, i_m) \in \mathcal{I}$. For any $D \subset [m]$, we define $\mathcal{I}_D := \prod_{j \in D}[r_j]$ and $\boldsymbol{i}_D := (i_j : j \in D) \in \mathcal{I}_D$. This means that up to the apropriate indices reordering we have that $\boldsymbol{i} = (\boldsymbol{i}_D, \boldsymbol{i}_{D^c})$ where $D^c = [m] \backslash D$.*

*Given an m-way table $\boldsymbol{u} = (u(\boldsymbol{i}) : \boldsymbol{i} \in \mathcal{I})$, we define the D-**margin** vector of $\boldsymbol{u}$ as the $|D|$-way table $\boldsymbol{u}_D = (u_D(\boldsymbol{i}_D) : \boldsymbol{i}_D \in \mathcal{I}_D)$ such that*

$$u_D(\boldsymbol{i}_D) := \sum_{\boldsymbol{i}_{D^c} \in \mathcal{I}_{D^c}} u(\boldsymbol{i}_D, \boldsymbol{i}_{D^c}) \quad \text{for every } \boldsymbol{i}_D \in \mathcal{I}_D.$$

In other words, $u_D(\boldsymbol{i}_D)$ represents the sum of all of the entries of $\boldsymbol{u}$ with $\boldsymbol{i}_D$ being fixed.

EXAMPLE 1.3.3. *Consider $\mathcal{I} = [r_1] \times [r_2]$ and a 2-way table $\boldsymbol{u} = (u(\boldsymbol{i}) : \boldsymbol{i} = (i_1, i_2) \in \mathcal{I}$. The table $\boldsymbol{u}$ has two 1-margins given by*

$$\boldsymbol{u}_{\{1\}} = \left(\sum_{i_2 \in [r_2]} u(i_1, i_2) : i_1 \in [r_1] \right) \quad \text{and} \quad \boldsymbol{u}_{\{2\}} = \left(\sum_{i_1 \in [r_1]} u(i_1, i_2) : i_2 \in [r_2] \right). \tag{1.5}$$

In other words, $\boldsymbol{u}_{\{1\}}$ is the vector of row sums of $\boldsymbol{u}$ and $\boldsymbol{u}_{\{2\}}$ is the vector of column sums. This means that for 2-way tables, the vector sufficient statistics of the independence model consists of the 1-margins of $\boldsymbol{u}$ as we explained in Example 1.2.6.

Now, consider $\mathcal{I} = [r_1] \times [r_2] \times [r_3]$ and let $\boldsymbol{u} = (u(\boldsymbol{i}) : \boldsymbol{i} = (i_1, i_2, i_3) \in \mathcal{I})$ be a 3-way table. The table $\boldsymbol{u}$ has three 1-margins $\boldsymbol{u}_{\{1\}}, \boldsymbol{u}_{\{2\}}$ and $\boldsymbol{u}_{\{3\}}$, also known as plane sums. Figure 1.4 provides a concrete example of a 3-way table with $r_1 = r_2 = r_3 = 3$ and Figure 1.5 provides an illustration of its 1-margins.

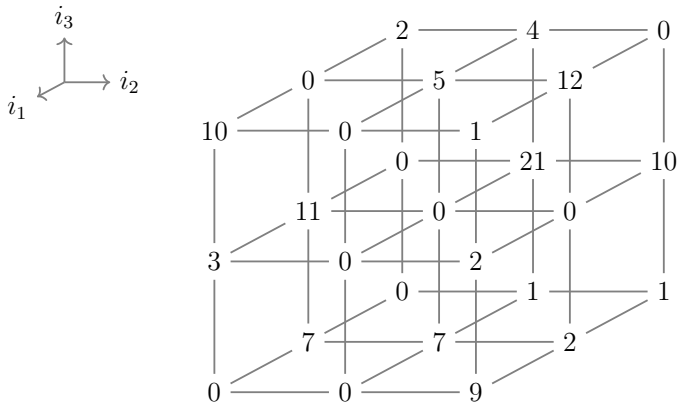
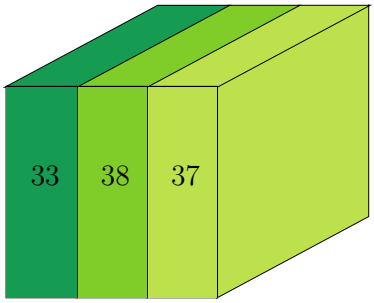
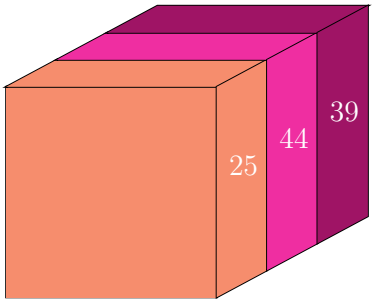


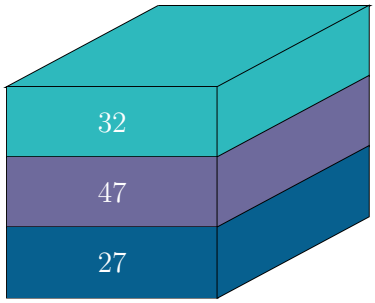
FIGURE 1.4. Illustration of a 3-way table $\boldsymbol{u}$.



(a) $\boldsymbol{u}_{\{1\}} = \big(u(1,+,+), u(2,+,+), u(3,+,+)\big)$



(b) $\boldsymbol{u}_{\{2\}} = \big(u(+,1,+), u(+,2,+), u(+,3,+)\big)$

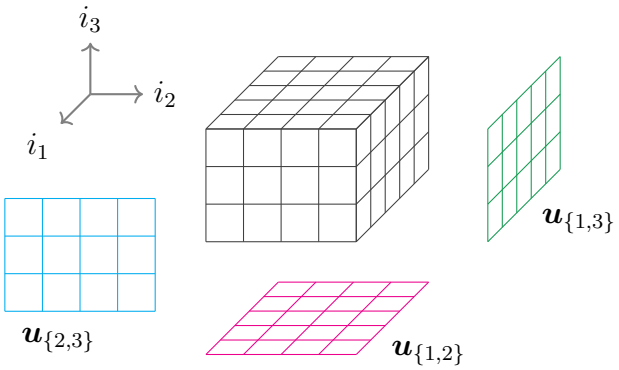


(c) $\boldsymbol{u}_{\{3\}} = \big(u(+,+,1), u(+,+,2), u(+,+,3)\big)$

FIGURE 1.5. 1-margins of the table $\boldsymbol{u}$ in Figure 1.4.

A 3*-way table* $\boldsymbol{u}$ *has three* 2*-margins,* $\boldsymbol{u}_{\{1,2\}}, \boldsymbol{u}_{\{1,3\}}$ *and* $\boldsymbol{u}_{\{1,3\}}$ *which are also known as line sums and are illustrated in Figure 1.6 for a particular example where* $r_1 = 5, r_2 = 4$ *and* $r_3 = 3$.

Hierarchical models are a class of log-linear models for which the vector of sufficient statistics is determined by the margins induced by a **simplicial complex**: a family of subsets Δ of $[m]$ such that, for every $D \in \Delta$, all subsets of D are also in Δ. The elements of the family Δ are called **faces**, and the maximal elements under inclusion are called **facets**, denoted by $\mathcal{F}(\Delta)$. Since facets



$$\boldsymbol{u}_{\{1,2\}}(i_1, i_2) = \sum_{i_3=1}^{r_3} x(i_1, i_2, i_3)$$

$$\boldsymbol{u}_{\{1,3\}}(i_1, i_3) = \sum_{i_2=1}^{r_2} x(i_1, j, i_3)$$

$$\boldsymbol{u}_{\{2,3\}}(i_2, i_3) = \sum_{i_1=1}^{r_1} x(i, i_2, i_3)$$

FIGURE 1.6. 2-margins of a 3-way table $\boldsymbol{u}$.

completely determine a simplicial complex, we describe a simplicial complex by its facets using bracket notation. For instance, $\Delta = [123][24]$ is the simplicial complex with ground set $\{1, \ldots, 4\}$, whose facets are $\{1, 2, 3\}$ and $\{2, 4\}$.

Using this informal definition, we can see that the independence model is a hierarchical model determined by the simplicial complex $\Delta = [1][2]$.

DEFINITION 1.3.4. *Let $\mathcal{I} = [r_1] \times \cdots \times [r_m]$, $\boldsymbol{r} = (r_1, \ldots, r_m)$, and let Δ be a family of subsets of $[m]$. The **hierarchical model** associated with Δ and $\boldsymbol{r}$ is the family of probability distributions*

$$\mathcal{L}_{\Delta,\boldsymbol{r}} := \left\{ \boldsymbol{p}_{\boldsymbol{\theta}} : \boldsymbol{\theta} = (\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(|D|)}) \in \prod_{D \in \Delta} \mathbb{R}^{|\mathcal{I}_D|} \right\},$$

where $\boldsymbol{\theta}^{(D)} = (\theta^{(D)}(\boldsymbol{i}_D) : \boldsymbol{i}_D \in \mathcal{I}_D)$, and such that

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{i}) = \sum_{D \in \Delta} \theta^{(D)}(\boldsymbol{i}_D).$$

In this context, the entries of $\boldsymbol{r}$ are known as the levels of the hierarchical model.

Hierarchical models are log-linear models. Specifically, they can be described through a $\prod_{D \in \Delta} |\mathcal{I}_D| \times |\mathcal{I}|$ design matrix $A_{\Delta,\boldsymbol{r}}$ associated with the simplicial complex Δ and the vector $\boldsymbol{r}$. The columns of $A_{\Delta,\boldsymbol{r}}$ are labeled by $\mathcal{I}$, typically in lexicographic order, while the rows are divided into blocks: one block per facet of Δ. The rows within each block are labeled by $\mathcal{I}_D$, also ordered lexicographically. Finally, the blocks corresponding to the different facets of Δ are ordered lexicographically.

If we let $\boldsymbol{e}_{\boldsymbol{i}_D}$ be the $|\mathcal{I}_D|$-dimensional vector such that for every $\boldsymbol{i}'_D \in \mathcal{I}_D$,

$$\boldsymbol{e}_{\boldsymbol{i}_D}(\boldsymbol{i}'_D) = \begin{cases} 1, & \text{if } \boldsymbol{i}'_D = \boldsymbol{i}_D, \\ 0, & \text{otherwise.} \end{cases}$$

It is possible to describe the $\boldsymbol{i}$-th column of A as $A(\cdot, \boldsymbol{i}) = \bigoplus_{D \in \mathcal{F}(\Delta)} \boldsymbol{e}_{\boldsymbol{i}_D}$, where $\oplus$ represents vector concatenation.

EXAMPLE 1.3.5. *Consider the simplicial complex $\Delta = [12][13][23]$. The hierarchical model associated with this simplicial complex is known as the* ***no-three-way interaction model****. A concrete example for the design matrix $\mathcal{M}_{\Delta,\boldsymbol{r}}$ of this model with levels $\boldsymbol{r} = (2,2,2)$ is given below.*

$$A_{\Delta,\boldsymbol{r}} = \begin{array}{c} \begin{array}{cccccccc} 111 & 112 & 121 & 122 & 211 & 212 & 221 & 222 \end{array} \\ \left(\begin{array}{cccccccc} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ \hline 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ \hline 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{array}\right) \end{array} \begin{array}{l} \theta^{(12)}(1,1) \\ \theta^{(12)}(1,2) \\ \theta^{(12)}(2,1) \\ \theta^{(12)}(2,2) \\ \theta^{(13)}(1,1) \\ \theta^{(13)}(1,2) \\ \theta^{(13)}(2,1) \\ \theta^{(13)}(2,2) \\ \theta^{(23)}(1,1) \\ \theta^{(23)}(1,2) \\ \theta^{(23)}(2,1) \\ \theta^{(23)}(2,2) \end{array}.$$

Notice that the column associated to $(2,1,2) \in [2] \times [2] \times [2]$ is given by

$$\begin{aligned} A_{\Delta,\boldsymbol{r}}(\cdot, 212) &= \boldsymbol{e}_{(2,1)} \oplus \boldsymbol{e}_{(2,2)} \oplus \boldsymbol{e}_{(1,2)} \\ &= (0,0,1,0) \oplus (0,0,0,1) \oplus (0,1,0,0). \end{aligned}$$

By keeping careful track of indices it can be shown that for an m-way table $\boldsymbol{u} \in \mathbb{Z}^{\mathcal{I}}_{\geq 0}$, $A_{\Delta,\boldsymbol{r}}\boldsymbol{u} = (\boldsymbol{u}_D : D \in \mathcal{F}(\Delta))$. In other words, $(\boldsymbol{u}_D : D \in \mathcal{F}(\Delta))$ is the vector of sufficient statistics for the hierarchical model $\mathcal{L}_{\Delta,\boldsymbol{r}}$. For example, the 2-margins illustrated in Figure 1.6 constitute the sufficient statistics for the no-three-way interaction model introduced in Example 1.3.5.

DEFINITION 1.3.6. *A simplicial complex Δ on $[m]$ is **reducible** with decomposition (Δ_1, S, Δ_2) and separator $S \subset [m]$ if $\Delta = \Delta_1 \cup \Delta_2$ and $\Delta_1 \cap \Delta_2 = 2^S$, where 2^S denotes the power set of S. A simplicial complex is **decomposable** if it is reducible and both Δ_1 and Δ_2 are decomposable or if they are of the form 2^R for some $R \subset [m]$.*

REMARK 1.3.7. *Decomposable models can also be defined in an alternative way (see [7, Section 8.1]). A simplicial complex Δ is **graphical** if its facets correspond to the maximal cliques of a graph G. Furthermore, Δ is decomposable if and only if G is a chordal graph.*

EXAMPLE 1.3.8. *The simplicial complex* [1][2][3] *is decomposable and corresponds to the independence model on three variables. In contrast, the simplicial complex* [12][13][23]*, associated with the no-three-way interaction model, is the simplest non-decomposable model since it is not even graphical.*

On one hand, it is known that for decomposable hierarchical models, the structure of their Markov basis is well understood thanks to work of [46, 107]. In fact, one can always find a Markov basis with moves whose one-norms equal 4, regardless of the size of the levels $r_1, \ldots, r_m$. This was subsequently used to spell out a divide-and-conquer algorithm to compute Markov bases for reducible models in [47].

On the other hand, such a bound fails to exist for even the simplest non-decomposable model: the no-three-way interaction of three discrete random variables described in Example 1.3.5. By importing powerful polyhedral geometry results into statistics, [39] showed that any minimal Markov bases of the no-three-way-interaction model on $r_1 \times r_2 \times 3$ tables can contain moves with arbitrarily large 1-norm, if r_1 and r_2 are unrestricted.

Before explicitly stating this result, let us recall that the fibers $\mathcal{F}(A, \boldsymbol{b})$ can be understood as the set of integer points of the polytope $P(A, \boldsymbol{b}) := \{\boldsymbol{y} \in \mathbb{R}^n : \boldsymbol{y} \geq \boldsymbol{0}, A\boldsymbol{y} = \boldsymbol{b}\}$. In other words, $\mathcal{F}(A, \boldsymbol{b}) = P(A, \boldsymbol{b}) \cap \mathbb{Z}^{\mathcal{I}}_{\geq 0}$. The following remarkable result shows that the no-three-way interaction model can capture the geometric structure of any polytope $P(A, \boldsymbol{b})$. Furthermore, it implies that if one is able to describe a Markov basis for the no-three-way interaction model on $r_1 \times r_2 \times 3$ tables as we let r_1, r_2 vary, then we automatically have a Markov basis for any integer matrix A.

THEOREM 1.3.9 ([38]). *For any rational matrix $A \in \mathbb{Q}^{d \times n}$ and any integer vector $\boldsymbol{b} \in \mathbb{Z}^d$, $P(A, \boldsymbol{b}) = \{\boldsymbol{y} \in \mathbb{R}^n_{\geq 0} : A\boldsymbol{y} = \boldsymbol{b}\}$ is polynomial-time representable as a slim 3-way transportation*

polytope:

$$T = \left\{ x \in \mathbb{R}_{\geq 0}^{I \times J \times 3} : \sum_{k=1}^{3} x_{i,j,k} = u_{i,j}, \sum_{j=1}^{J} x_{i,j,k} = v_{i,k}, \sum_{i=1}^{I} x_{i,j,k} = w_{j,k} \right\}.$$

For positive integers h and h', saying that a polytope $P \subset \mathbb{R}^h$ is representable as a polytope $Q \subset \mathbb{R}^{h'}$ means that there is an injection $\sigma : \{1, \ldots, h\} \to \{1, \ldots, h'\}$ such that the coordinate-erasing projection

$$\pi : \mathbb{R}^{h'} \to \mathbb{R}^h : x = (x_1, \ldots, x_{h'}) \mapsto \pi(x) = (x_{\sigma(1)}, \ldots, x_{\sigma(h)})$$

provides a bijection between Q and P and between their integer points $Q \cap \mathbb{Z}^{h'}$ and $P \cap \mathbb{Z}^h$.

As a consequence of Proposition 1.3.9, we have the following.

COROLLARY 1.3.10. *For any nonnegative integer vector $\boldsymbol{\theta} \in \mathbb{N}^\eta$, there exist positive integers r_1, r_2 such that any Markov basis for the no-three-way interaction model on $r_1 \times r_2 \times 3$ tables must contain an element whose restriction to some η entries is precisely $\boldsymbol{\theta}$. In particular, the degree and support of elements in the minimal Markov bases, as r_1 and r_2 vary, can be arbitrarily large.*

When a Markov basis is unavailable, a natural alternative is to explore simpler subsets $\mathcal{L}$ of $\ker_{\mathbb{Z}} A$, with the hope that they are sufficiently large to induce connected fiber graphs $\mathcal{F}(A, \boldsymbol{b})_{\mathcal{L}}$, leading to an irreducible Markov chain on $\mathcal{F}(A, \boldsymbol{b})$. If $\mathcal{L}$ induces disconnected fiber graphs, one could modify Algorithm 1 by temporarily stepping outside of $\mathcal{F}(A, \boldsymbol{b})$ in non-negative vectors $\boldsymbol{u}$ (still satisfying $A\boldsymbol{u} = \boldsymbol{b}$) and eventually returning to $\mathcal{F}(A, \boldsymbol{b})$, hoping to achieve an irreducible Markov chain on $\mathcal{F}(A, \boldsymbol{b})$.

This has been a direction of research for some time with many open questions, some of which are summarized in [112]. The most common subsets of the integer kernel used for this purpose are:

- *Lattice bases.* A set of vectors $\mathcal{L}$ in $\ker_{\mathbb{Z}} A$ is called a **lattice basis** if it is linearly independent and $\text{span}_{\mathbb{Z}} \mathcal{L} = \ker_{\mathbb{Z}} A$. As noted in [45], a lattice basis is typically a proper subset of a full Markov basis. While its size is determined by the rank of A and it can be computed easily using the Hermite normal form of A (see [100, Section 4.1]), it does not generally form a Markov basis. The workaround for ensuring a provably connected chain is simple: every Markov move can be expressed as a linear combination of lattice basis moves. However, the challenge is that the size of these required linear combinations is not well understood.

- *Circuits.* A vector $\boldsymbol{u} \in \ker_{\mathbb{Z}} A$ is a **circuit** if its support is minimal, meaning there is no vector $\boldsymbol{v} \in \ker_{\mathbb{Z}} A$ such that $\operatorname{supp}(\boldsymbol{v}) \subset \operatorname{supp}(\boldsymbol{u})$. The set of all such vectors is called the set of circuits of A and is denoted by $C(A)$. In certain cases, A provides a clear combinatorial description of the set of circuits.

In general, the following containment relationships hold among these subsets of $\ker_{\mathbb{Z}} A$: (1) every Markov basis contains a lattice basis, (2) the Graver basis contains any minimal Markov basis, and (3) the Graver basis contains the set of circuits.

EXAMPLE 1.3.11. *Consider the matrix* $A = (3\ 4\ 5)$ *from Example 1.1.9, and let* $\mathcal{M} = \{(1,-2,1), (3,-1,-1), (2,1,-2)\}$ *be a minimal Markov basis for* A. *Notice that both* $\mathcal{F}(A,15)$ *and* $\mathcal{F}(A,30)$ *require at least 3 moves to be connected. Since rank*$(A) = 2$*, no lattice basis of* A *forms a Markov basis in this case. Furthermore, removing an element from* $\mathcal{M}$ *recovers a lattice basis for* A*, but this induces disconnected fibers* $\mathcal{F}(A, \boldsymbol{b})$ *even for small values of* $\boldsymbol{b}$ *(cf. Figure 1.1).*

As noted in Example 1.1.27, the Graver basis for A *is given by the set* $Gr(A) = \{(1,-2,1), (3,-1,-1), (2,1,-2), (4,-3,0), (5,0,-3), (1,3,-3), (0,5,-4)\}$. *On the other hand, the set of circuits for* A *is* $C(A) = \{(0,5,-4), (5,0,-3), (0,4,-3)\}$.

There are scenarios where various bases of a model are equal, such as when circuits form a Markov basis. For example, when the design matrix A is totally unimodular, [45, Proposition 8.11] shows that the Graver basis coincides with the set of circuits of A. However, unimodularity is a strong condition, as demonstrated by Seymour's decomposition theorem (see [101, §19.4]). For more details on unimodular matrices in specific models, we refer the reader to [12, 13], which provide a comprehensive description of hierarchical models with totally unimodular design matrices.

For the no-three-way interaction model, many authors have considered a special set of moves called **basic moves**: elements of minimal 1-norm that, like a lattice basis, span $\ker_{\mathbb{Z}} A$. However, the set of basic moves is not a Markov basis for the no-three-way interaction model on $r_1 \times r_2 \times r_3$ tables when at least two of r_1, r_2, r_3 are greater than 2 (see [7, Chapter 9]). Nevertheless, this simple set of moves can still connect certain fibers, as shown in [15], which demonstrates that basic moves suffice to connect 3-way tables with *positive margins*.

Some special cases work out particularly well. For instance, [21, Proposition 3] shows that basic moves generate an irreducible chain for the no-three-way interaction model on $2 \times r_2 \times r_3$ tables if

the fiber is extended by allowing a single -1 entry at any step. Similarly, [25, Theorem 3.1] applied the non-negativity relaxation of the fibers to a logistic regression model, allowing some entries to take -1 values. Motivated by these ideas, [84] and [113] used the same approach to show that basic moves induce an irreducible Markov chain on the fibers of the no-three-way interaction model on $3 \times 3 \times r_3$ and $3 \times 4 \times r_3$ tables, while allowing temporary -1 entries. These results utilized the full descriptions of the unique minimal Markov bases presented in [9, 10]. In the following, we formalize the non-negativity relaxation approach and study its limitations.

As mentioned previously , the goal of the non-negativity relaxation approach will be to define irreducible Markov chains on the fibers $F(A, \boldsymbol{b})$ of a log-linear model with design matrix A by using a simple set of moves and allowing "temporary" steps of the chain to be taken in relaxed fibers whose elements allow for negative values in some entries.

DEFINITION 1.3.12. *Let $A \in \mathbb{Z}^{d \times n}$ be an integer matrix and $\boldsymbol{b} \in \mathbb{N}A$ a margin. For $\boldsymbol{l} \in \mathbb{Z}^n$ we define the* ***unbounded $(\boldsymbol{b}; \boldsymbol{l})$-fiber*** *of A as the set*

$$\mathcal{F}(A, \boldsymbol{b}; \boldsymbol{l}) := \{\boldsymbol{u} \in \mathbb{Z}^n : A\boldsymbol{u} = \boldsymbol{b}, \text{ and } l_i \leq u_i \text{ for every } i \in [n]\}.$$

DEFINITION 1.3.13. *Let $\mathcal{I} = [r_1] \times \cdots \times [r_m]$ and $S \subseteq \mathcal{I}$ with indicator vector $\mathbf{1}_S$, meaning that for every $\boldsymbol{i} \in \mathcal{I}$, the $\boldsymbol{i}$-th entry of $\mathbf{1}_S$ is 1 if $\boldsymbol{i} \in S$ and 0 otherwise. Let $A \in \mathbb{Z}^{d \times \mathcal{I}}$ and $q \in \mathbb{Z}_{\geq 0}$. We say that $\mathcal{M} \subset \ker_{\mathbb{Z}} A$ is a $(-q, S)$-**Markov basis** for A if for every margin $\boldsymbol{b} \in \mathbb{N}A$ and any pair $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{F}(A, \boldsymbol{b})$, there exists a choice of moves $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_K \in \mathcal{M}$ such that*

$$\boldsymbol{u} - \boldsymbol{v} = \sum_{k=1}^{K} \boldsymbol{m}_k \quad \text{and} \quad \boldsymbol{v} + \sum_{k=1}^{k'} \boldsymbol{m}_k \in \mathcal{F}(A, \boldsymbol{b}; -q\mathbf{1}_S) \quad \text{for every } k' \leq K.$$

*When $S = \mathcal{I}$, we write $\mathbf{1}$ instead of $\mathbf{1}_S$, and a $(-q, \mathcal{I})$-Markov basis is referred to simply as a $(-q)$-**Markov basis**. Notice that when $q = 0$, a $(-q, S)$-Markov basis is the same as a usual Markov basis, as in Definition 1.1.7.*

As with a usual Markov basis, there are equivalent ways to test whether a set of moves $\mathcal{M} \subset \ker_{\mathbb{Z}} A$ is a $(-q)$-Markov basis, one of which uses a particular type of ideals. Given a polynomial

$f \in \mathbb{K}[x_1, \ldots, x_{|\mathcal{I}|}]$ and $J \subset \mathbb{K}[x_1, \ldots, x_{|\mathcal{I}|}]$, we define the **ideal quotient**

$$(J : f^\infty) := \{g \in \mathbb{K}[x_1, \ldots, x_{|\mathcal{I}|}] : f^t g \in J \text{ for some } t \in \mathbb{Z}^{>0}\}.$$

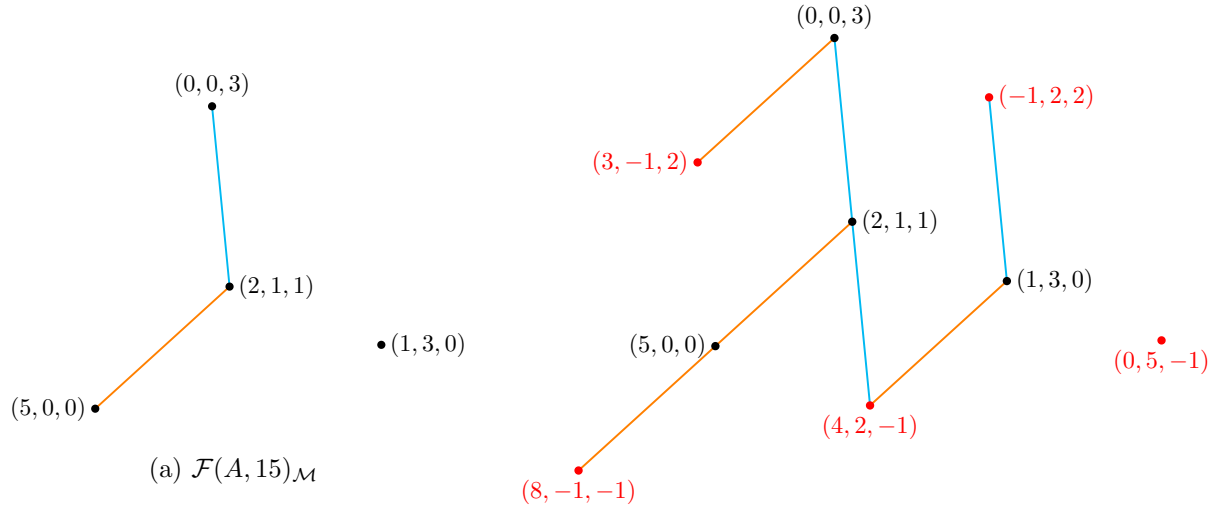
PROPOSITION 1.3.14 ([104]). *Let $A \in \mathbb{Z}^{d \times \mathcal{I}}$ and $\mathcal{M} \subset \ker_\mathbb{Z} A$. Then, the following are equivalent*

(1) there exists $q \in \mathbb{Z}_{\geq 0}$ such that $\mathcal{M}$ is a $(-q)$-Markov basis for A,

(2) $span_\mathbb{Z}(\mathcal{M}) = \ker_\mathbb{Z} A$,

(3) $(\langle \boldsymbol{x}^{\boldsymbol{u}^+} - \boldsymbol{x}^{\boldsymbol{u}^-} : \boldsymbol{u} \in \mathcal{M}\rangle : (x_1 \cdots x_{|\mathcal{I}|})^\infty) = I_A$.

EXAMPLE 1.3.15. *The set $\mathcal{M} = \{(3, -1, -2), (2, 1, -2)\}$ is a lattice basis for $A = (3\ 4\ 5)$ but not a Markov basis. On the other hand, $\mathcal{M}$ is a (-1)-Markov basis for A. In Figure 1.7, (a) illustrates the disconnected fiber graph $\mathcal{F}(A, 15)_\mathcal{M}$, while (b) depicts the fiber graph $\mathcal{F}(A, 15; -\mathbf{1})_\mathcal{M}$. This shows how one can connect the elements of $\mathcal{F}(A, 15)$ by temporarily stepping in the set $\mathcal{F}(A, 15; -\mathbf{1})$. For example, to go from $(2, 1, 1)$ to $(1, 3, 0)$ we take a step in $(4, 2, -1)$.*



(a) $\mathcal{F}(A, 15)_\mathcal{M}$

(b) $\mathcal{F}(A, 15; -\mathbf{1})_\mathcal{M}$

FIGURE 1.7. Fiber graphs induced by $\mathcal{M} = \{(3, -1, -1), (2, 1, -2)\}$ on $\mathcal{F}(A, 15)$ and $\mathcal{F}(A, 15; -\mathbf{1})$. Vertices in $\mathcal{F}(A, 15; -\mathbf{1}) \backslash \mathcal{F}(A, 15)$ are shown in red.

A small modification to Algorithm 1 yields a way to sample points of a fiber using a $(-q, S)$-Markov basis, formalized in Algorithm 3 below. Hence, by Proposition 1.3.14, for any set of moves

$\mathcal{M}$ spanning $\ker_{\mathbb{Z}} A$, there exists a sufficiently large q that allows us to define an irreducible Markov chain on $\boldsymbol{b}$-fibers $\mathcal{F}(A, \boldsymbol{b})$.

Algorithm 3: Fiber samples given a $(-q, S)$-Markov basis.

Input : $\boldsymbol{u} \in \mathcal{F}(A, \boldsymbol{b})$, starting point in a fiber
$\mathcal{M}$, a $(-q, S)$-Markov basis for A
π, a desired distribution on $\mathcal{F}(A, \boldsymbol{b})$
N, the number of fiber samples

Output: A sequence of points $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots$ in $\mathcal{F}(A, \boldsymbol{b})$

Set $\boldsymbol{u}_1 \leftarrow \boldsymbol{u}$;
Set $\boldsymbol{v} \leftarrow \boldsymbol{u}$; ▷ Auxiliary variable keeping current point in $\mathcal{F}(A, \boldsymbol{b}; -q\mathbf{1}_S)$
for $n = 1, \ldots, N-1$ **do**
  Choose $\boldsymbol{m} \in \pm\mathcal{M}$ uniformly at random;
  if $\boldsymbol{v} + \boldsymbol{m} \notin \mathcal{F}(A, \boldsymbol{b}; -q\mathbf{1}_S)$ **then**
    $\boldsymbol{u}_{n+1} \leftarrow \boldsymbol{u}_n$;
  else
    if $\boldsymbol{v} + \boldsymbol{m} \in \mathcal{F}(A, \boldsymbol{b})$ **then**
      $\boldsymbol{u}_{n+1} \leftarrow \boldsymbol{v} + \boldsymbol{m}$ and $\boldsymbol{v} \leftarrow \boldsymbol{v} + \boldsymbol{m}$ with probability $\min\left\{1, \frac{\pi(\boldsymbol{v}+\boldsymbol{m})}{\pi(\boldsymbol{u}_n)}\right\}$
    else
      $\boldsymbol{u}_{n+1} \leftarrow \boldsymbol{u}_n$;
      $\boldsymbol{v} \leftarrow \boldsymbol{v} + \boldsymbol{m}$;
Return sequence $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N$

In general, the *non-negativity relaxation approach* can be interpreted as follows:

(1) Identify an easily attainable subset $\mathcal{M} \subset \ker_{\mathbb{Z}} A$ such that $\operatorname{span}_{\mathbb{Z}}(\mathcal{M}) = \ker_{\mathbb{Z}} A$. For instance, one could compute a lattice basis $\mathcal{M}$ for $\ker_{\mathbb{Z}} A$.

(2) Find a $q > 0$ such that $\mathcal{M}$ is a $(-q, S)$-Markov basis for A for some S.

Although this strategy works in certain special situations, it cannot be applied universally without a careful analysis of the connectivity of $\mathcal{F}(A, \boldsymbol{b})_{\mathcal{M}}$. In the absence of general bounds for q, it is necessary to prove the irreducibility of the Markov chain on the fibers on a case-by-case basis, depending on the corresponding relaxation induced by a fixed value of q.

1.3.1. Our contributions

We begin this subsection with the following result whose proof is presented in Section 2.1.

THEOREM 1.3.16. *For any $N > 0$, there exists a matrix Λ_N with $||\Lambda_N||_1 = 4$ and $\mathcal{M}_N \subset \ker_{\mathbb{Z}} \Lambda_N$ such that $\operatorname{span}_{\mathbb{Z}}(\mathcal{M}_N) = \ker_{\mathbb{Z}} \Lambda_N$ but $\mathcal{M}_N$ is not a $(-q)$-Markov basis for any $q < N$.*

The conclusion of Theorem 1.3.16 is that having a set of moves that spans $\ker_{\mathbb{Z}} A$ does not ensure that a small relaxation of the fiber will suffice to construct irreducible Markov chains on the fibers of the model. However, in specific cases of the no-three-way interaction model, a simple set of moves has been shown to form a (-1)-Markov basis when r_1 and r_2 are fixed and r_3 grows, as previously discussed.

The next two results highlight the limitations of the fiber relaxation technique when applied to the no-three-way interaction model. This particular model is of interest because any fiber of any model corresponds to a fiber of an associated no-three-way interaction model, as stated in Proposition 1.3.9.

Given the results of Corollary 1.3.10, a natural question arises: are the non-negativity constraints on the entries responsible for the problematic behavior of the Markov basis for the model? In other words, we would like to investigate how large the elements of a $(-q, S)$-Markov basis can be when $r_1, r_2 > 0$ are unrestricted and $r_3 = 3$.

The following theorem suggests that translating the non-negativity constraint hyperplanes by one unit can still lead to arbitrarily complicated elements inside any minimal Markov basis when S is chosen poorly. The proof of this result is in Section 2.2.

THEOREM 1.3.17. *For any nonnegative integer vector $\boldsymbol{\theta} \in \mathbb{N}^{\eta}$, there are $r_1, r_2 \in \mathbb{Z}_{>0}$ and $S \subset [r_1] \times [r_2] \times [3]$ with $|S| = 1 + \sum_{i=1}^{\eta} \theta_i$, such that any minimal $(-1, S)$-Markov basis for the no-three-way interaction model on $r_1 \times r_2 \times 3$ tables must contain an element whose restriction to some η entries is $\boldsymbol{\theta}$ or $2\boldsymbol{\theta}$.*

We now turn to study on the effectiveness of non-negativity relaxation technique using basic moves, which for this model have received significant attention (see [15, 25, 64]).

As mentioned previously, a *basic move* for the no-three-way model on $r_1 \times r_2 \times r_3$ tables is a zero-margin table with minimal 1-norm. These basic moves can be described as 3-way tables $\boldsymbol{u} = (u_{i,j,k})$ of the form

$$u_{i,j,k} = \begin{cases} 1, & \text{if } (i,j,k) \in \{(i_1,j_1,k_1),(i_1,j_2,k_2),(i_2,j_1,k_2),(i_2,j_2,k_1)\} \\ -1, & \text{if } (i,j,k) \in \{(i_2,j_2,k_2),(i_2,j_1,k_1),(i_1,j_2,k_1),(i_1,j_1,k_2)\} \\ 0, & \text{otherwise} \end{cases}$$

for fixed indices $i_1 \neq i_2 \in [r_1]$, $j_1 \neq j_2 \in [r_2]$ and $k_1 \neq k_2 \in [r_3]$. We denote the basic move associated to these indices by $\boldsymbol{b}(i_1, i_2;\ j_1, j_2;\ k_1, k_2)$ (see Figure 1.8 for an illustration) and we denote the set of basic moves for the no-three-way interaction model on $r_1 \times r_2 \times r_3$ tables by $\mathcal{B}_{r_1,r_2,r_3}$ or simply by $\mathcal{B}$ when r_1, r_2 and r_3 are clear from the context.

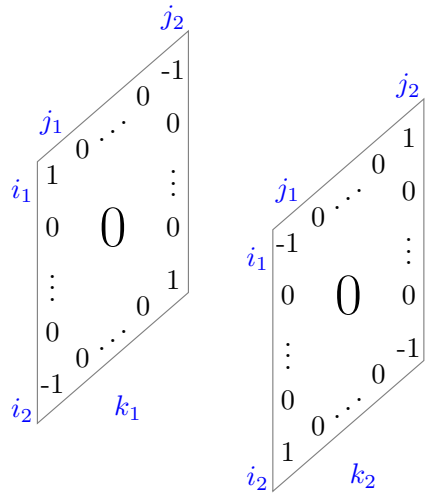


FIGURE 1.8. The basic move $\boldsymbol{b}(i_1, i_2; j_1, j_2; k_1, k_2)$

It is known that for the design matrix A of the no-three-way-interaction model for $r_1 \times r_2 \times r_3$ tables, any element in $\ker_{\mathbb{Z}} A$ can be written as a linear combination of the basic moves, i.e., $\text{span}_{\mathbb{Z}}(\mathcal{B}) = \ker_{\mathbb{Z}} A$ (see [64]). Hence, Proposition 1.3.14 guarantees the existence of a $q > 0$ such that $\mathcal{B}$ is a $(-q)$-Markov basis for A.

Although it remains an open problem whether $\mathcal{B}_{r_1,r_2,r_3}$ is a (-1)-Markov basis for the no-three-way interaction model in general, it has been established, as mentioned earlier, that for specific cases such as $2 \times r_2 \times r_3$, $3 \times 3 \times r_3$, and $4 \times 3 \times r_3$ way tables, $\mathcal{B}$ is a (-1)-Markov basis for A. However, given the complex behavior of the fibers for A described in Corollary 1.3.10, it is hard to believe that the result generalizes when fixing $r_3 = 3$ and letting r_1, r_2 be unconstrained. To address this problem, we present the following partial result with proof in Section 2.2.

PROPOSITION 1.3.18. *Let $r_1, r_2 \geq 3$ and let $S \subset [r_1] \times [r_2] \times [3]$ have an anti-staircase shape (see Definition 1.3.19). Then, for any $q > 0$ the set of basic moves is* not *a $(-q, S)$-Markov basis for no-three-way interaction model on $r_1 \times r_2 \times 3$ tables.*

DEFINITION 1.3.19. *Let $r_1, r_2 \geq 3$ and let $S \subset [r_1] \times [r_2] \times [3]$. We say that S has a staircase shape if there is a surjective function $\tau : [r_2] \to [3]$ or a surjective function $\tau' : [r_1] \to [3]$ such that*

$$S = \bigcup_{j=1}^{r_2} \{(i, j, \tau(j)) : i \in [r_1]\} \quad \text{or} \quad S = \bigcup_{i=1}^{r_1} \{(i, j, \tau'(i)) : j \in [r_2]\}.$$

We say that S has an anti-staircase shape if S is a complement of a subset of $[r_1] \times [r_2] \times [3]$ in staircase shape. As an example, the sets S corresponding to the colored cells in Figure 1.9 have a staircase shape.

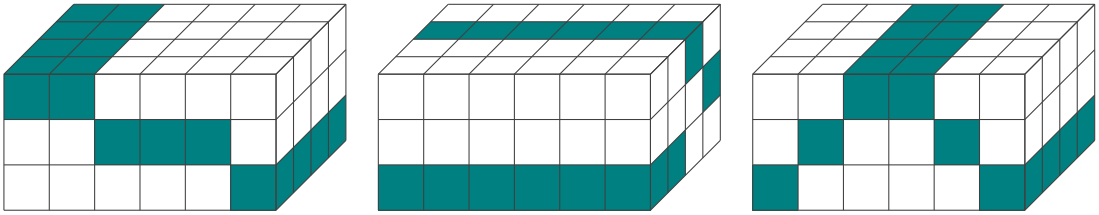


FIGURE 1.9. Subsets of $[4] \times [6] \times [3]$ with staircase shape.

Despite the previous findings, we show good complexity results for non-decomposable models as demonstrated in Corollary 1.3.23. The result builds upon two key points:

(1) often the design matrix of a hierarchical model exhibits a block structure. For instance it could be an *n-fold matrix*, defined below; and
(2) the Graver basis of an n-fold matrix solely depends on its constituent blocks.

DEFINITION 1.3.20. *Given fixed matrices $A \in \mathbb{Z}^{p \times s}$ and $B \in \mathbb{Z}^{p' \times s}$ with positive integer p, p', s, the n-**fold matrix** of the ordered pair (A, B) is defined as the $(np + p') \times sn$ matrix*

$$[A, B]^{(n)} := \begin{pmatrix} A & 0 & 0 & \cdots & 0 \\ 0 & A & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & A \\ B & B & B & \cdots & B \end{pmatrix}.$$

*We define the **type of a vector** $\boldsymbol{u} = (\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(n)}) \in \mathbb{Z}^{sn}$ as the number $|\{j : \boldsymbol{u}^{(j)} \neq \mathbf{0}\}|$ of nonzero components $\boldsymbol{u}^{(j)} \in \mathbb{N}^s$. The following result establishes a stabilization property of the Graver basis for n-fold matrices.*

PROPOSITION 1.3.21 ([72]). *Given matrices $A \in \mathbb{Z}^{p\times s}, B \in \mathbb{Z}^{p'\times s}$, there exists a constant C such that for all n, the Graver basis of $[A,B]^{(n)}$ consists of vectors of type at most C. The smallest of these constants is known as the* ***Graver complexity*** *of A, B and we denote it by $g(A,B)$. Furthermore,*

$$g(A,B) = \max_{\boldsymbol{u}\in Gr(B\cdot Gr(A))} ||\boldsymbol{u}||_1.$$

The Graver basis $Gr([A,B]^{(n)})$ for any n-fold of A, B can be obtained from the next result.

PROPOSITION 1.3.22 ([36]). *For fixed matrices $A \in \mathbb{Z}^{p\times s}$ and $B \in \mathbb{Z}^{p'\times s}$, the Graver basis $Gr([A,B]^{(n)})$ can be computed in polynomial time on n. Moreover, the size of $Gr([A,B]^{(n)})$ is bounded by $|Gr([A,B]^{(g)})|\binom{n}{g}$, where $g = g(A,B)$ is the Graver complexity of A, B.*

COROLLARY 1.3.23. *Let Δ be a simplicial complex with ground set $[m]$ and maximal faces $D_1,\dots,D_t$. Let $V \subset [m]$ be such that for every $j \in [m]$, either $V \subset D_j$ or $V \subset D_j^c$. Let $\boldsymbol{\rho} = (\rho_l)_{l\notin V}$ be fixed. Then, for any $(r_1,\dots,r_m) \in \mathbb{N}^m$ with $(r_l)_{l\notin V} = \boldsymbol{\rho}$, the size of the Graver basis $|Gr(A_\Delta)|$ is bounded by a polynomial in $\prod_{l\in V} r_l$.*

In light of Proposition 1.1.29, which states that Graver elements contain all the necessary moves for sampling restricted fibers, this has direct implications for the feasibility of sampling restricted fibers. The proof of this corollary is provided in Section 2.3. This relies on the fact that A_Δ is an $\left(\prod_{l\in V} r_l\right)$-fold matrix.

EXAMPLE 1.3.24. *Let Δ be a simplicial complex on four vertices with levels $(r_1, r_2, 2, 3)$ represented in Figure 1.10 below. The maximal faces of Δ are $D_1 = \{1,2,3\}$, $D_2 = \{1,2,4\}$, and $D_3 = \{3,4\}$, which do not correspond to the set of maximal cliques of any graph. Thus, Δ is not graphical and, consequently, not decomposable.*

The set $V = \{1,2\}$ satisfies $V \subset F_1, F_2$ and $V \subset F_3^c$. By the proof of Corollary 1.3.23, it follows that $A_\Delta = [A,B]^{(r_1r_2)}$, where $B = I_6$ and A is the design matrix of the independence model with levels $(2,3)$. Using the software `4ti2` *from [1], we compute $g(A,B) = 3$ and $|Gr([A,B]^{(3)})| = 15$. Therefore, we have $|Gr(A_\Delta)| \le 15\binom{r_1r_2}{3}$ for any r_1, r_2.*

REMARK 1.3.25. *An important assumption made in Proposition 1.3.22 is that the matrices A, B are fixed. However, it is worth noting that the Graver complexity $g(A,B)$ can become arbitrarily*

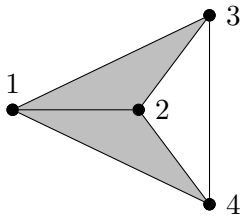


FIGURE 1.10. Non-decomposable simplicial complex Δ. The shaded regions represent the maximal faces.

large when the size of the entries in A and B, or the dimensions of A and B, vary. For example, it was recently shown in [90, Theorem 5.1] that 2×4 matrices have arbitrarily large Graver complexity when we let the entries vary. Similar results can be also found in [14].

REMARK 1.3.26. *In scenarios where it is not feasible to explicitly compute the Graver complexity $g(A,B)$ for fixed matrices $A \in \mathbb{Z}^{p\times s}$ and $B \in \mathbb{Z}^{p'\times s}$, we can rely on upper bounds. The best known upper bounds come from recent developments on sparse integer programming where the tree-depth plays an important role (see [49, 80, 90]). We would like to emphasize the significance of the block structure and sparsity within the design matrices when computing Graver bases. While these attributes have been utilized in optimization contexts [33, 34, 36, 49, 82], their application in statistics remains relatively unexplored.*

1.4. Markov Bases and Graphs with Fixed Degree Sequences

A particularly intriguing area of research focuses on the random generation of graphs with a fixed degree sequence (see [3, 24, 28, 52, 78]). The **degree sequence** of a graph $\boldsymbol{g}$ with vertices in $[n]$ is represented as the vector $d(\boldsymbol{g}) = (d_1, \ldots, d_n)$, where d_u denotes the degree of vertex u in $\boldsymbol{g}$. It has long been established that the set of all graphs with a fixed degree sequence can be connected through switches. Informally, a *switch* is an exchange of edge pairs between two graphs that preserves the degree sequence.

Notably, [66] and [63] leveraged this insight to provide a constructive solution to the *graph realization problem*, which is commonly known as the *Havel-Hakimi algorithm*. Alternatively, one can determine whether a degree sequence is graphical without constructing a corresponding graph by using the characterization given by the Erdős-Gallai theorem [51], which tests the validity of n inequalities. This characterization is closely related to the hyperplane representation of the degree

sequence polytope introduced by Koren [81]. Further results regarding this polytope can be found in [87, 91, 102].

In 1997, [78] proposed the use of the *switch Markov chain* to uniformly generate simple graphs with a fixed degree sequence. As highlighted in [52], "*the switch Markov chain can be thought of as the Markov chain of smallest possible modifications.*"

In this section, we delve into a colored generalization of the connectivity problem on spaces of graphs with a fixed degree sequence and a fixed graph statistic arising from a vertex coloring and we explore its connections with the theory of Markov bases.

From an algebraic statistics perspective, it is useful to conceptualize the space of graphs with a fixed degree sequence $\boldsymbol{d} \in \mathbb{N}^n$ as a set of vectors $\boldsymbol{g} = (g_{uv})_{u<v}$ in $\mathcal{G}_n := \prod_{i=1}^{\binom{n}{2}} \mathcal{E}$ satisfying the system of linear equations $D_n\boldsymbol{g} = \boldsymbol{d}$, where D_n is the incidence matrix of the complete graph K_n, and both the columns of D_n and the entries of $\boldsymbol{g}$ are ordered lexicographically. Here, $\mathcal{E}$ corresponds to the set of possible values that each g_{uv} can take. The set of interest $\{\boldsymbol{g} \in \mathcal{G}_n : D_n\boldsymbol{g} = \boldsymbol{d}\}$ corresponds to the $\boldsymbol{d}$-fiber $\mathcal{F}(D_n, \boldsymbol{d})$ when $\mathcal{E} = \mathbb{N}$, and corresponds to the $(\boldsymbol{d}; \boldsymbol{0}, \boldsymbol{1})$-fiber $\mathcal{F}(D_n, \boldsymbol{d}; \boldsymbol{0}, \boldsymbol{1})$ when $\mathcal{E} = \{0, 1\}$. While the first scenario considers multigraphs with a given degree sequence $\boldsymbol{d}$, the second one considers simple graphs with a given degree sequence $\boldsymbol{d}$.

REMARK 1.4.1. *For $\boldsymbol{g} = (g_{uv})_{u<v} \in \mathcal{G}_n$, g_{uv} is interpreted as the number of interactions between nodes u and v. Since we will be dealing exclusively with undirected graphs, we assume $g_{uv} = g_{vu}$ for any pair of distinct nodes.*

DEFINITION 1.4.2. *Let A be an integer matrix and $\mathcal{M} \subset \ker_{\mathbb{Z}} A$. We say that $\mathcal{M}$ is a **binary Markov basis** for A if $\mathcal{F}(A, \boldsymbol{b}; \boldsymbol{0}, \boldsymbol{1})_{\mathcal{M}}$ is connected for every $\boldsymbol{b} \in \mathbb{N}A$.*

In terms of Definition 1.4.2, our earlier discussion at the beginning of this section establishes that for any $n \in \mathbb{N}$, the set of switches constitutes both a Markov basis and a binary Markov basis for D_n. In this case, the set of switches corresponds to the elements in $\ker_{\mathbb{Z}} D_n$ of minimal 1-norm:

$$\mathcal{M}_n := \left\{\boldsymbol{e}_{uv} + \boldsymbol{e}_{u'v'} - \boldsymbol{e}_{uv'} - \boldsymbol{e}_{u'v} : u < v,\ u' < v',\ \{u, v\} \cap \{u', v'\} = \varnothing\right\},$$

where $\boldsymbol{e}_{uv}$ is the standard unit vector in $\mathbb{R}^{\binom{n}{2}}$ associated with the pair $\{u, v\}$. Figure 1.11 illustrates the move in $\mathcal{M}_n$ that consists of replacing the pair of edges $\{u, v'\}$ and $\{u', v\}$ with the pair of edges $\{u, v\}$ and $\{u', v'\}$.

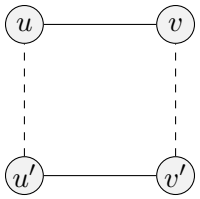


FIGURE 1.11. Switch in $\mathcal{M}_n$ replacing the pair of edges $\{u, v'\}$ and $\{u', v\}$ by the pair of edges $\{u, v\}$ and $\{u', v'\}$.

DEFINITION 1.4.3. *For a positive integer k and a k-coloring (or block assignment) $z : [n] \to [k]$, we define the* ***color sequence*** *of a graph $\boldsymbol{g} \in \mathcal{G}_n$ to be the vector $c(\boldsymbol{g}, z) := (c(z, i, j) : 1 \leq i \leq j \leq k)$ with $c(z, i, j)$ being equal to the total number of interactions in $\boldsymbol{g}$ between colors i and j. The entries of $c(\boldsymbol{g}, z)$ are ordered lexicographically with respect to the pairs (i, j) and when z is clear from the context we simply write $c(i, j)$ and $c(\boldsymbol{g})$. The* ***degree-color sequence*** *$(d(\boldsymbol{g}), c(\boldsymbol{g}))$ of $\boldsymbol{g}$ with a given k-coloring z, is the concatenation of its degree and color sequences. For simplicity, we call $(d(\boldsymbol{g}), c(\boldsymbol{g}))$ the* ***c-degree sequence*** *from now on.*

EXAMPLE 1.4.4. *For $n = 5$ and $k = 3$ let $\{\{1, 2\}, \{3, 4\}, \{5\}\}$ be the partition of $[5]$ induced by a 3-coloring z of $[5]$. ○, ◇ and ⬠ represent colors 1, 2, and 3, respectively. The c-degree sequence of the graph $\boldsymbol{g}$ illustrated in Figure 1.12 is the vector in $\mathbb{N}^{11}$ given by $(d(\boldsymbol{g}), c(\boldsymbol{g})) =$* $(4, 4, 3, 4, 7; 1, 3, 3, 0, 4, 0)$.

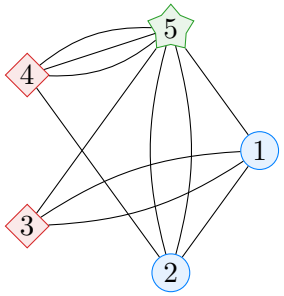


FIGURE 1.12. Graph $\boldsymbol{g}$ with vertices 1 and 2 colored blue, 3 and 4 red, and 5 green.

The c-degree sequence is also a linear graph statistic, as we now explain. For a k-coloring z of $[n]$, we define C_z as the matrix with rows labeled by the $k + \binom{k}{2} = \binom{k+1}{2}$ pairs of colors (allowing repetition) and columns labeled by the $\binom{n}{2}$ distinct pairs of vertices; with both rows and columns ordered lexicographically. Each column contains exactly one entry equal to 1 in the row corresponding to that vertex pair's color pair, with the remaining entries of the column set to zero.

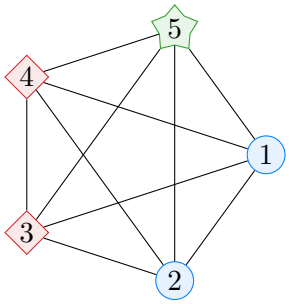
For a graph $\boldsymbol{g} \in \mathcal{G}_n$, the color sequence of $\boldsymbol{g}$ can be expressed as $c(\boldsymbol{g}) = C_z\boldsymbol{g}$. Consequently,

$$(d(\boldsymbol{g}), c(\boldsymbol{g})) = (D_n\boldsymbol{g}, C_z\boldsymbol{g}) = \mathit{DC}_{n,z}\ \boldsymbol{g},$$

where

$$\mathit{DC}_{n,z} := \begin{pmatrix} D_n \\ C_z \end{pmatrix}. \tag{1.6}$$

EXAMPLE 1.4.5. *Let $n = 5$, $k = 3$, and let z be the 3-coloring used in Example 1.4.4. The matrix $\mathit{DC}_{n,z}$ is explicitly written below alongside a depiction of K_5 on the left, which helps visualize the encoding of the matrix.*



$$\mathit{DC}_{5,z} = \begin{array}{c} \begin{matrix} 12 & 13 & 14 & 15 & 23 & 24 & 25 & 34 & 35 & 45 \end{matrix} \\ \left(\begin{array}{cccccccccc} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ \hline 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right) \end{array}$$

As we will see in more detail in Section 1.5, the c-degree sequence arises as the sufficient statistic of a random network model in which each edge in the graph appears with a probability that depends on its endpoint vertices as well as their color, which, in statistics, represents blocks or communities. While [79] discusses how to use Markov bases to extend exact tests to latent block versions of the model, they left the determination of a Markov basis for $\mathit{DC}_{n,z}$ as an open problem.

1.4.1. Our contributions

The following result, whose proof is provided in Section 3.1, presents a solution to the open problem mentioned above.

THEOREM 1.4.6. *The set of quadratic moves $\mathcal{M}_{n,z} := \{\boldsymbol{g} \in \ker_{\mathbb{Z}} \mathit{DC}_{n,z} : ||\boldsymbol{g}||_1 = 4\}$ is a Markov basis for $\mathit{DC}_{n,z}$. These are the moves in $\ker_{\mathbb{Z}} \mathit{DC}_{n,z}$ of minimal 1-norm.*

Similar to the monochromatic case, this Markov basis is equivalent to the set of smallest possible modifications. In essence, any two multigraphs with a fixed c-degree sequence can be connected by applying a sequence of c-degree-preserving switches of 4 edges at a time. A natural follow-up question is whether the connectivity in the space of multigraphs with a fixed c-degree sequence, induced by the moves in Theorem 1.4.6, is maintained when restricting to the space of simple graphs.

In contrast to the behavior observed in the monochromatic case, the 1-norm size of the moves necessary to connect spaces of simple graphs with a fixed c-degree sequence increases as the number of colors k used in the k-coloring z varies as we describe with the following three results. We present a proof of Proposition 1.4.7 in Section 3.2 while Corollary 1.4.9 and Theorem 1.4.10 are straightforward consequences of Proposition 1.4.7.

In contrast to the monochromatic case, the 1-norm size of the moves required to connect spaces of simple graphs with a fixed c-degree sequence increases as the number of colors k in the k-coloring z varies. This phenomenon is established in the following three results. Proposition 1.4.7 is proved in Section 3.2, while Corollary 1.4.9 and Theorem 1.4.10 follow directly from it.

PROPOSITION 1.4.7. *For every integer $k \geq 3$ there exists a k-coloring z of $[n]$ with $n = 2k$, and a c-degree sequence $(\boldsymbol{d}_k, \boldsymbol{c}_k) \in \mathbb{N}^{n+\binom{k+1}{2}}$ such that $\mathcal{F}_{\mathcal{DC}_{n,z}}(\boldsymbol{d}_k, \boldsymbol{c}_k; \mathbf{0}, \mathbf{1}) = \{\boldsymbol{g}, \boldsymbol{g}'\}$. Furthermore, $||\boldsymbol{g} - \boldsymbol{g}'||_1 = 2k$.*

EXAMPLE 1.4.8. *The simple graphs $\boldsymbol{g}_1, \boldsymbol{g}_2$ in Figure 1.13 represent the only two elements of the simple-graph fiber $\mathcal{F}_{A_{12,z}}(\boldsymbol{d}_6, \boldsymbol{c}_6; \mathbf{0}, \mathbf{1})$ where $z : [12] \to [6]$ is such that $z(u) \equiv u \pmod 6$ for every $u \in [12]$ and $\boldsymbol{d}_6, \boldsymbol{c}_6$ are defined as in the proof of Proposition 1.4.7 in section 3.2. The only move (up to sign) that connects this fiber is $\boldsymbol{g} = \boldsymbol{g}_1 - \boldsymbol{g}_2$.*

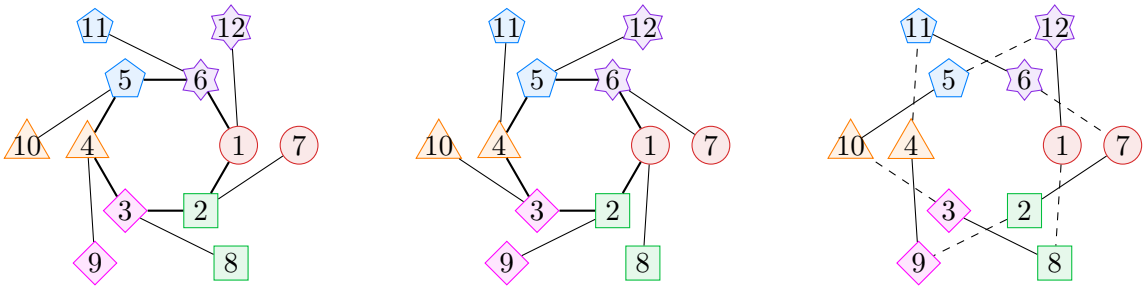


FIGURE 1.13. Simple graphs $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$ on the left and center. Markov basis move $\boldsymbol{g}_1 - \boldsymbol{g}_2$ on the right.

As an immediate consequence of Proposition 1.4.7 we have the following two results.

COROLLARY 1.4.9. *Let $n, k \in \mathbb{Z}_+$ with $k \geq 3$ and let z be a k-coloring of $[n]$. Then, any simple-graph Markov basis for $DC_{n,z}$ has an element with 1-norm equal to 2κ where κ is the number of colors used to color more than one vertex.*

THEOREM 1.4.10. *For any constant η, there exists $n, k \in \mathbb{Z}_+$ and a k-coloring z of $[n]$ such that any binary Markov basis for $DC_{n,z}$ has an element with 1-norm greater than η.*

As it has been previously mentioned, when the k-coloring z is constant, 4-edge switches are enough to connect the space of simple graphs with a fixed degree sequence $\boldsymbol{d} \in \mathbb{N}^n$ for any $\boldsymbol{d}$. In contrast, Theorem 1.4.10 shows that when we do not impose any constraints on the coloring function z, we cannot guarantee the existence of a constant η such that the set of η-edge switches induces connectivity on the space of simple graphs with fixed c-sequence $(d; c)$ for any degree-color sequence $(\boldsymbol{d}, \boldsymbol{c}) \in \mathbb{N}^{n+\binom{k+1}{2}}$.

QUESTION 1.4.11. *Given $k \in \mathbb{Z}_+$, is there a constant η_k such that for any $n \in \mathbb{Z}_+$ and any k-coloring z of $[n]$, there exists a binary Markov basis $\mathcal{B}$ for $DC_{n,z}$ such that $\max_{\boldsymbol{g} \in \mathcal{B}} ||\boldsymbol{g}||_1 \leq \eta_k$? If so, what is the minimum η_k satisfying this condition?*

For $k = 1$, the minimum constant that satisfies the conditions in Question 1.4.11 is $\eta_1 = 4$. For $k = 2$, Example 1.4.12 below demonstrates that if η_2 exists, it must be at least 8.

EXAMPLE 1.4.12. *The simple graphs $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$ in Figure 1.14 are the only two elements of the $(\mathbf{0}, \mathbf{1})$-fiber $\mathcal{F}_{DC_{n,z}}(\boldsymbol{d}, \boldsymbol{c}; \mathbf{0}, \mathbf{1})$, where z is a 2-coloring that induces the partition $\{\{1, 2, 5, 6\}, \{3, 4, 7, 8\}\}$, with $\boldsymbol{d} = (1, 6, 1, 6, 4, 3, 4, 3)$ and $\boldsymbol{c} = (3, 8, 3)$. The only move (up to sign) that connects this simple-graph fiber is $\boldsymbol{g} = \boldsymbol{g}_1 - \boldsymbol{g}_2$, whose 1-norm is 8, as illustrated in Figure 1.14. One way to connect $\boldsymbol{g}_1$ to $\boldsymbol{g}_2$ using elements from $\mathcal{M}_{n,z}$ by stepping into $\mathcal{F}_{DC_{n,z}}(\boldsymbol{d}, \boldsymbol{c}) \setminus \mathcal{F}_{DC_{n,z}}(\boldsymbol{d}, \boldsymbol{c}; \mathbf{0}, \mathbf{1})$ is depicted in Figure 1.15. The orange-highlighted edges in each graph indicate the switches performed to reach the next graph in the orange path.*

CONJECTURE 1.4.13. *For $k = 2$, η_2 exists and $\eta_2 = 8$ is the minimum constant satisfying the condition in Question 1.4.11.*

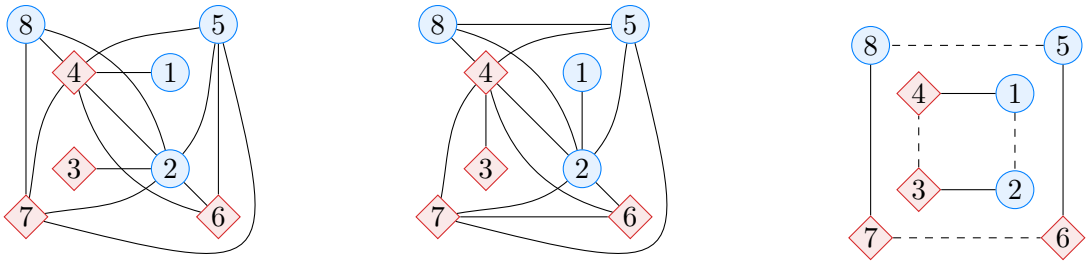


FIGURE 1.14. Simple graphs $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$ on the left and center. Markov basis move $\boldsymbol{g}_1 - \boldsymbol{g}_2$ on the right.

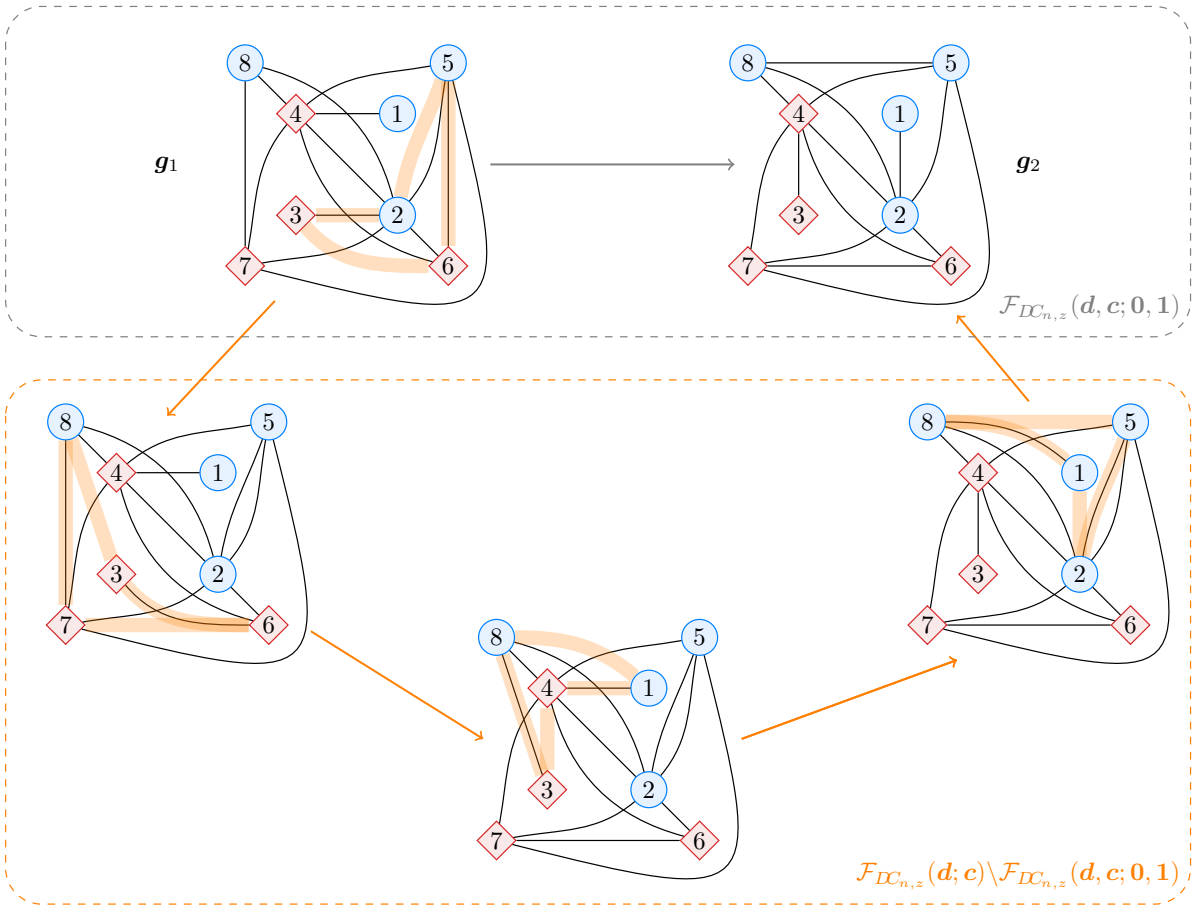


FIGURE 1.15. Simple graphs $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$ being connected with switches by leaving the $(\mathbf{0},\mathbf{1})$-fiber. The switches used in each step are highlighted in orange.

Given the result in Theorem 1.4.6, the fundamental theorem of Markov Bases implies that $I_{DC_{n,z}}$ is generated by the quadratic binomials $\{\boldsymbol{x}^{\boldsymbol{g}^+} - \boldsymbol{x}^{\boldsymbol{g}^-} : \boldsymbol{g} \in \mathcal{M}_{n,z}\}$. As discussed previously, the more conditions satisfied by these set of binomials, the more its applications.

In the monochromatic case, [40] proved that this set of quadratic binomials is a Gröbner basis for I_{D_n}, and the result was used to study triangulations and optimization of the $\boldsymbol{b}$-matching problem—the graph with the smallest cost having a given degree sequence $\boldsymbol{b}$. The final contribution of this section, whose proof is presented in Section 3.3, generalizes the quadratic Gröbner basis result from the monochromatic case to the following general case.

THEOREM 1.4.14. *There exists a monomial order $\succ$ on $\mathbb{K}[x_{uv} : 1 \leq u < v \leq n]$ such that for any k-coloring z of $[n]$, the set $\{\boldsymbol{x}^{\boldsymbol{g}^+} - \boldsymbol{x}^{\boldsymbol{g}^-} : \boldsymbol{g} \in \mathcal{M}_{n,z}\}$ is in fact a Gröbner basis for $I_{D\mathcal{C}_{n,z}}$ with respect to $\succ$.*

1.5. Goodness-of-Fit Tests for Labeled Stochastic Block Models

The study of Markov bases for the matrix $D\mathcal{C}_{n,z}$ in the previous section is primarily motivated by their applications in goodness-of-fit tests for specific random network models that share similarities with log-linear models.

In recent years, the analysis of network data has become increasingly significant across diverse fields, including the social sciences and biological studies. The foundation of probabilistic modeling for network data lies in classical random graph models, such as the Erdős-Rényi model [50]. These models provide a starting point for understanding the structural and probabilistic properties of networks.

Fienberg's approach to analyzing statistical network models bridges network science and categorical data analysis by representing graphs as contingency tables (see [55, 56]). This framework enables the application of tools from categorical data analysis to address critical challenges, such as parameter estimation and model assessment. For example, [92] introduced algebraic statistics into network analysis by studying Markov bases for the p1 model presented in [69]. For a broader perspective on the interplay between categorical data analysis, algebraic statistics, and network science, we recommend [61] and [62].

A particularly relevant class of log-linear models in this context is the *Stochastic Block Model* (SBM), which is given a contingency representation in [53]. SBMs extend classical random graph models by allowing edge probabilities to depend on the block membership of node pairs, thereby

enabling the detection and modeling of community structures within networks. Originally introduced in the social sciences by [54], SBMs have since gained prominence for their flexibility and wide applicability.

Despite significant advancements in the development of SBMs, assessing their goodness-of-fit to observed network data remains an underexplored area. Notable contributions to this field include spectral goodness-of-fit tests by [85] and graphical methods for model assessment by [73]. In this section, we focus on a general strategy for performing goodness-of-fit tests for SBMs introduced in [79]. This approach leverages Markov bases to construct Markov chains that explore the fiber of observed network data.

As in the previous section, we represent graphs with vertices on $[n]$ by a $\binom{n}{2}$-dimensional vector $\boldsymbol{g} = (g_{uv})_{u<v} \in \mathcal{G}_n = \mathcal{E}^{\binom{n}{2}}$, where for each *dyad* $\{u, v\}$ (or pair of vertices), g_{uv} takes values from the same support set $\mathcal{E}$. Here, g_{uv} is generally understood as the number of interactions between nodes u and v, and $\mathcal{E}$ imposes restrictions on these counts. For example, setting $\mathcal{E} = \{0, 1\}$ restricts g_{uv} to represent the presence (or absence) of an interaction.

Throughout this section, we assume that $g_{uv} = g_{vu}$ for every dyad, since all the graphs considered are undirected, and let $g_{uu} = 0$ since loops are not allowed. Furthermore, the function $z : [n] \to [k]$ will represent a k-coloring, which, for consistency with the SBM literature, we refer to as a *block assignment* in this section. Consequently, we refer to $B_i := z^{-1}(i)$ as the i-th block.

We say that a random graph $\boldsymbol{G} = (G_{uv})_{u<v}$ with sample space $\mathcal{G}_n = \mathcal{E}^{\binom{n}{2}}$ is drawn from a **Stochastic Block Model with block assignment** z (**SBM(z)**) if there exists a parameter vector $\boldsymbol{\theta} = (\theta_{ij} : 1 \leq i \leq j \leq k)$ (commonly referred to as the connectivity matrix in the literature) such that $\{G_{uv}\}_{u<v}$ are pairwise independent and

$$G_{uv} \sim f(\cdot, \theta_{z(u)z(v)}) := f_{\theta_{z(u)z(v)}}(\cdot), \tag{1.7}$$

where $f_{\theta_{ij}}$ is a probability distribution on $\mathcal{E}$ for each $1 \leq i \leq j \leq k$, known up to the finite-dimensional parameter θ_{ij}.

In this section, we will exclusively consider cases in which $\{f(\cdot, \theta_{ij})\}_{i \leq j}$ belong to the exponential family. Specifically, we assume that for all $\epsilon \in \mathcal{E}$ and θ_{ij},

$$f(\epsilon, \theta_{ij}) \propto h(\epsilon) \exp\{\langle \eta(\theta_{ij}), \epsilon \rangle\},$$

where $\eta(\theta_{ij})$ is the natural parameter of the family. Consequently, under the SBM(z)

$$\mathbb{P}(\boldsymbol{G} = \boldsymbol{g} \mid \boldsymbol{\theta}) \propto h(\boldsymbol{g}) \exp\{\langle \eta(\boldsymbol{\theta}), T_z(\boldsymbol{g}) \rangle\}, \tag{1.8}$$

where $\eta(\boldsymbol{\theta}) = (\eta(\theta_{ij}))_{i \leq j}$, $T_z(\boldsymbol{g}) = (T_{z,ij}(\boldsymbol{g}))_{i \leq j}$ and

$$T_{z,ij}(\boldsymbol{g}) = \begin{cases} \frac{1}{2} \sum_{u \in B_i, v \in B_j} g_{uv}, & \text{if } i = j, \\ \sum_{u \in B_i, v \in B_j} g_{uv}, & \text{otherwise.} \end{cases}$$

The vector $T_z(\boldsymbol{g})$ serves as the sufficient statistic for the SBM(z) and is linear in $\boldsymbol{g}$, which implies that the SBM(z) is a log-linear model. Consequently, the sufficient statistic can be encoded using a design matrix $A_{\text{SBM}(z)}$. When g_{uv} is a scalar representing interaction counts or a binary present/absent status, we have $\mathcal{E} \subset \mathbb{N}$. In this case, $A_{\text{SBM}(z)}$ is a $\binom{k+1}{2} \times \binom{n}{2}$ matrix with rows labeled by all possible block pairs and columns labeled by dyads, both ordered lexicographically. The column associated with the dyad uv contains a 1 in the row corresponding to the pair $z(u)z(v)$ and 0s elsewhere. In other words, $A_{\text{SBM}(z)}$ is identical to the matrix C_z in Equation (1.6) from the previous section.

As explained in Section 1.2, given an observed graph $\boldsymbol{g}_0$, the goodness-of-fit for the SBM(z) can be assessed by computing the conditional p-value:

$$p(\boldsymbol{g}_0, z) := \mathbb{P}(W_z(\boldsymbol{G}) \geq W_z(\boldsymbol{g}) \mid A_{\text{SBM}(z)}\boldsymbol{G} = A_{\text{SBM}(z)}\boldsymbol{g}), \tag{1.9}$$

where $W_z(\boldsymbol{G})$ is a test statistic such that large values indicate evidence against $\boldsymbol{g}$ being generated by an SBM(z). It is worth noting that the support of the conditional distribution in Equation (1.9) corresponds to the set $\{\boldsymbol{g} \in \mathcal{E}^{\binom{n}{2}} : A_{\text{SBM}(z)}\boldsymbol{g} = A_{\text{SBM}(z)}\boldsymbol{g}_0\}$. This set equals the fiber $\mathcal{F}(A_{\text{SBM}(z)}\boldsymbol{g}_0)$ when $\mathcal{E} = \mathbb{N}$ and corresponds to the $(\boldsymbol{l}, \boldsymbol{L})$-fiber $\mathcal{F}(A; \boldsymbol{l}, \boldsymbol{L})$ when $\mathcal{E} = [\boldsymbol{l}, \boldsymbol{L}]$.

Theorem 1.5.2 below describes a Markov basis that not only connects the $\boldsymbol{b}$-fibers for $A_{\text{SBM}(z)}$ but also connects all the $(\boldsymbol{b}; \boldsymbol{l}, \boldsymbol{L})$-fibers. This provides flexibility in choosing the support $\mathcal{E}$ for the interactions modeled by (1.7). For any of these scenarios, Algorithm 2 can be slightly modified so that the step in line 4 verifies whether the proposed graph belongs to the corresponding restricted fiber, rather than the general unrestricted $\mathcal{F}(A_{\text{SBM}(z)})$.

REMARK 1.5.1. *The matrix $DC_{n,z}$ from section 1.4 corresponds to the design matrix of a variant of the SBM, known as the β-SBM(z) for which $T_z(\boldsymbol{g}) = DC_{n,z}\,\boldsymbol{g}$ is the vector of sufficient statistics. The β-SBM(z) postulates that for a random graph $\boldsymbol{G}$, $G_{uv} \sim f\big(\cdot, \theta_{z(u)z(v)} + \beta_u + \beta_v\big)$ where $\{f\big(\cdot, \theta_{z(u)z(v)} + \beta_u + \beta_v\big) : i \leq j, 1 \leq u < v \leq n\}$ belongs to the exponential family. For this model we have a parameter β_v for each node v in addition to the block-interaction parameters θ_{ij}. When $\mathcal{E} = \mathbb{N}$ is the support for each of the random variables G_{uv}, Theorem 1.4.6 guarantees that the family of moves $\mathcal{M}_{n,z}$ can be used to test whether $\boldsymbol{g}_0 \sim \beta$-SBM(z). On the other hand, Theorem 1.4.10 implies that when $\mathcal{E} = \{0, 1\}$ (e.g., when $G_{uv} \sim$ Bernoulli$(\sigma(\theta_{z(u)z(v)+\beta_u+\beta_v}))$), performing a goodness-of-fit test for the β-SBM(z) becomes a much harder task.*

Both the SBM(z) and the β-SBM(z) belong to a more general family known as log-linear exponential random graph models (ERGMs). For reference see [62].

With more generality, one can consider graphs that are better modeled by ℓ different types of interactions between nodes. We represent such graphs with an $\ell\binom{n}{2}$ vector $\boldsymbol{g} = (\boldsymbol{g}_{uv})_{u<v}$ where $\boldsymbol{g}_{uv} = (g_{uv}^{(l)} : u < v, 1 \leq l \leq \ell)$ and $g_{uv}^{(l)}$ denotes the number of l-type interactions between nodes u and v. In this scenario, the support $\mathcal{E}$ for $\boldsymbol{g}_{uv}$ is a subset of $\mathbb{N}^\ell$ and we say that $\boldsymbol{G}$ was generated from a **Labeled-SBM(z)** with ℓ labels (**LSBM(z, ℓ)**) if there is $\boldsymbol{\theta} = (\boldsymbol{\theta}_{ij} : 1 \leq i \leq j \leq k)$ such that $\{\boldsymbol{G}_{uv}\}_{u<v}$ are pairwise independent and $\boldsymbol{G}_{uv} \sim f(\cdot, \boldsymbol{\theta}_{z(u)z(v)})$ where $\boldsymbol{\theta}_{ij}$ is a vector of parameters of the same dimension for every $1 \leq i \leq j \leq k$. This generalization also yields a distribution of the same form as in (1.8) where $T_z(\boldsymbol{g}) = (T_{z,ij}^{(l)}(\boldsymbol{g}) : 1 \leq i \leq j \leq k, 1 \leq l \leq \ell)$ and

$$T_{z,ij}^{(l)}(\boldsymbol{g}) = \begin{cases} \frac{1}{2}\sum_{u\in B_i, v\in B_j} g_{uv}^{(l)} & \text{if } i = j, \\ \sum_{u\in B_i, v\in B_j} g_{uv}^{(l)} & \text{otherwise.} \end{cases}$$

In other words, $T_{z,ij}^{(l)}$ counts the number of l-type interactions between blocks i and j. Under this setting, the linear transformation $T_z(\boldsymbol{g})$ corresponds to the design matrix $A_{\text{LSBM}(z,\ell)} := I \otimes A_{\text{SBM}(z)}$, where I is the $\ell \times \ell$ identity matrix, and $\otimes$ denotes the Kronecker product.

1.5.1. Our contributions

Due to the structured nature of the matrix $A_{\text{SBM}(z,\ell)}$, we can obtain a compact description for its Graver basis. To do so, we first introduce the following notation.

For each $u \neq v \in [n]$ and $l \in [\ell]$, let $\boldsymbol{e}_{uv}^{(l)} \in \mathbb{N}^{\ell}$ represent the vector $\boldsymbol{g} = (g_{u'v'}^{(l')} : 1 \leq u' < v' \leq n, 1 \leq l' \leq \ell)$, defined by:

$$g_{u'v'}^{(l')} = \begin{cases} 1, & \text{if } (u', v') = (u, v) \text{ and } l' = l, \\ 0, & \text{otherwise.} \end{cases}$$

THEOREM 1.5.2. *The set* $\mathcal{M}_{LSBM(z,\ell)} := \{\boldsymbol{e}_{uv}^{(l)} - \boldsymbol{e}_{u'v'}^{(l)} : l \in [\ell], z(u) = z(u'), z(v) = z(v')\}$ *is the Graver basis of* $A_{LSBM(z,\ell)}$.

Hence, by Proposition 1.1.29, we can perform a goodness-of-fit test for the LSBM(z, ℓ) by computing a natural generalization of the p-value in Equation 1.9, provided that the random vectors $\boldsymbol{G}_{uv}$ have an interval as their support. However, certain scenarios involve enforcing additional constraints on the sample space.

A natural example arises when $\boldsymbol{G}_{uv} \sim \text{Geom}(N_{uv}, \boldsymbol{\theta}_{z(u)z(v)})$, in which case $\mathcal{E} = \{\boldsymbol{g} \in \mathbb{N}^{\ell} : ||\boldsymbol{g}||_1 = N_{uv}\}$. Despite these constraints, we can demonstrate that a specific subset of $\ker_{\mathbb{Z}} A_{\text{LSBM}(z,\ell)}$ induces connected fibers $\mathcal{F}$. This subset ensures the necessary guarantees to perform a valid goodness-of-fit test for the model, even under such restricted conditions.

THEOREM 1.5.3. *The set*

$$\widetilde{\mathcal{M}}_{LSBM(z,\ell)} := \{\boldsymbol{e}_{uv}^{(l)} + \boldsymbol{e}_{u'v'}^{(l')} - \boldsymbol{e}_{uv}^{(l')} - \boldsymbol{e}_{u'v'}^{(l)} : l, l' \in [\ell], z(u) = z(u'), z(v) = z(v')\}$$

induces connected graphs $\mathcal{F}_{\widetilde{\mathcal{M}}_{LSBM(z,\ell)}}$ *for every subset* $\mathcal{F} \subset \mathbb{N}^{\ell\binom{n}{2}}$ *of the form*

$$\mathcal{F} = \mathcal{F}(A_{LSBM(z,\ell)}, \boldsymbol{b}) \cap \{\boldsymbol{g} = (\boldsymbol{g}_{uv})_{u<v} \in \mathbb{N}^{\ell\binom{n}{2}} : ||\boldsymbol{g}_{uv}||_1 = N_{uv}\},$$

where $\boldsymbol{b} \in \mathbb{N}A_{LSBM(z,\ell)}$ *and* $\boldsymbol{N} = (N_{uv})_{u<v} \in \mathbb{N}^{\binom{n}{2}}$.

We provide proofs for both of the previous results in Sections 4.1 and 4.2.

The SBM framework is amenable to three modeling assumptions. Specifically, the block assignment for each node can be: 1) fixed and known, as we have assumed so far; 2) fixed and unknown; or 3) latent, with some underlying distribution.

For scenario 3), both frequentist and Bayesian approaches are possible. In the frequentist setting, following [88], we assume the existence of a latent block assignment $z : [n] \to [k]$ where

$\{z(u)\}_{u=1}^{n} \overset{\text{i.i.d.}}{\sim} \boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)$. This implies that the LSBM(ℓ) is a mixture of exponential random graph models, and there exist true parameters $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ governing the generation of the random graph $\boldsymbol{G}$. In the Bayesian approach, one assigns a prior to both $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$.

In [79], the authors proposed a method to perform a goodness-of-fit test for scenarios 2) and 3), relying on the test developed for the fixed and known z scenario. For instance, under scenario 2), given an observed graph $\boldsymbol{g}_0$ and an unobserved true block assignment z, one can use block estimation algorithms to recover an estimated assignment $\hat{z}$ and then compute the plug-in p-value $p(\boldsymbol{g}_0, \hat{z})$. To evaluate how closely an estimator approximates the true block assignment z, we adopt the following notion from the SBM literature (see [2]).

DEFINITION 1.5.4. *The agreement between two block assignments $z, z' : [n] \to [k]$ is defined as*

$$A(z, z') = \max_{\sigma \in S_k} \frac{1}{n} \sum_{u=1}^{n} \mathbb{1}\big(\sigma(z(u)) = z'(u)\big), \tag{1.10}$$

where S_k is the set of permutations on $[k]$. Whenever $A(z, z') = 1$, there exists a permutation $\sigma \in S_k$ such that $\sigma(z(u)) = z'(u)$ for every $u \in [n]$. In this case, we write $z' = \sigma \cdot z$.

DEFINITION 1.5.5. *Let $\boldsymbol{G} \sim$ LSBM(z, ℓ). An estimator $\hat{z} = \hat{z}(\boldsymbol{G})$ is called* strongly consistent *if $\mathbb{P}(A(z, \hat{z}) = 1) = 1 - o(1)$, meaning that $\hat{z}$ is strongly consistent if $A(z, \hat{z}) = 1$ with high probability as n tends to infinity.*

The proof of the following result is presented in Section 4.2.

PROPOSITION 1.5.6. *Consider a goodness-of-fit statistic satisfying $W_{\tilde{z}}(g) = W_{\sigma \cdot \tilde{z}}(g)$ for any $\tilde{z} : [n] \to [k]$ and $\sigma \in S_k$. Let $\boldsymbol{G} \sim$ LSBM(z, ℓ) and let $\hat{z} = \hat{z}(\boldsymbol{G})$ be a strongly consistent estimator, then $\mathbb{P}(p(z, \boldsymbol{G}) = p(\hat{z}, \boldsymbol{G})) = 1 - o(1)$ as n tends to infinity.*

An example of a goodness-of-fit statistic W_z satisfying the conditions from Proposition 1.5.6 is the block-corrected chi-square statistic from [79], defined as

$$W_z(\boldsymbol{g}) = \chi^2_{BC}(\boldsymbol{g}, z) := \sum_{u=1}^{n} \sum_{i=1}^{k} \sum_{l=1}^{\ell} \frac{(m_{ui}^{(l)} - n_i \hat{\theta}_{z_u i}^{(l)})^2}{n_i \hat{\theta}_{z_u i}^{(l)}} \tag{1.11}$$

where $n_i = |B_i|$, $m_{ui}^{(l)} = \sum_{v \in B_i} g_{uv}^{(l)}$ and $\hat{\theta}_{ij}^{(l)} = \frac{T_{z,ij}^{(l)}(g)}{n_{ij}}$ is the MLE estimate for $\theta_{ij}^{(l)}$. Under the LSBM(z, ℓ), we have the expected value $\mathbb{E}[m_{ui}^{(l)}] = n_i \theta_{z_u i}$, therefore large values of $\chi^2_{BC}(g, z)$, in which we have replaced $\theta_{z_u i}^{(l)}$ with the MLE $\hat{\theta}_{z_u i}^{(l)}$, indicate lack of fit.

Parallel to goodness-of-fit testing, model selection is another crucial aspect of network analysis, which involves determining the number of communities in a network, assuming it follows an SBM. Although model selection and goodness-of-fit testing are related, the latter is a more general problem that can also aid in model selection when applied sequentially. Moreover, goodness-of-fit tests provide a way to measure the model adequacy, offering valuable insights into how well the model captures the underlying structure of the network.

Regarding scenario 3), given an observed graph $\boldsymbol{g}_0$ and an unobserved block assignment z generated from a distribution, [79] proposes the use of the p-value

$$p(\boldsymbol{g}_0) := \sum_{z \in \mathcal{Z}_{n,k}} p(\boldsymbol{g}_0, z) \mathbb{P}(z \mid \boldsymbol{g}_0), \tag{1.12}$$

where $\mathcal{Z}_{n,k}$ represents the set of all possible block assignments for n nodes and k blocks. To estimate the p-value in Equation 1.12, the key challenge is to approximate $\mathbb{P}(z \mid \boldsymbol{g}_0)$. This can be approached in two ways: In the frequentist setting, one can use model-based estimation algorithms, such as those proposed in [88] or [108], to estimate the block proportions $\boldsymbol{\pi}$. In the Bayesian setting, algorithms like those introduced in [6] or [76] can be used to directly estimate $\mathbb{P}(z \mid \boldsymbol{g}_0)$.

It is worth noting that most of the estimation algorithms mentioned are limited to the case where $\ell = 1$. Among them, only [108] supports a special case where $\ell = 3$.

CHAPTER 2

Complexity of Markov bases: Bad and Good News

In this chapter, we present the proofs of the negative and positive results concerning the complexity of Markov bases, as outlined in Subsection 1.3.1.

Theorem 1.3.16 establishes that, in general log-linear models, there is no universal upper bound on the "negative" relaxation of the fibers required to connect the original fiber.

Furthermore, Theorem 1.3.17 and Proposition 1.3.18 demonstrate that relaxing a constraint set of entries S can still result in complex elements within a Markov bases if S is chosen poorly. These findings extend the results of [39], offering a deeper understanding of the intricacies involved in these scenarios.

Finally, for hierarchical models on $r_1 \times r_2 \times \cdots \times r_m$ contigency tables, Corollary 1.3.23 provides a positive result: it shows that the size of their Graver basis is bounded above by a polynomial in a proper subset of the levels $\{r_l\}_{l=1}^m$.

2.1. Complexity of $(-q)$-Markov bases

To construct the family of parametric matrices Λ_N referenced in Theorem 1.3.16, we first introduce a family of matrices whose kernels correspond to arithmetic sequences.

For $n \geq 3$, define the $(n-2) \times n$ integer matrix

$$A_{n-2} := \begin{pmatrix} 1 & -2 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 1 & -2 & 1 \end{pmatrix}.$$

In other words, the entries of A_{n-2} are defined as

$$A_{n-2}(i,j) = \begin{cases} 1, & \text{if } j = i \text{ or } j = i+2, \\ -2, & \text{if } j = i+1, \\ 0, & \text{otherwise.} \end{cases}$$

In the proof below we will utilize an operation known as the *Lawrence lifting*, introduced by [99]. This technique, which is a special case of an n-fold matrix construction, will play a key role in our analysis and we will take advantage some of its important properties.

Proof of Theorem 1.3.16. Let us notice that the column-style Hermite normal form of A_{n-2} is given by $H = (I_{n-2}\ \mathbf{0}_{n-2}\ \mathbf{0}_{n-2})$ where I_{n-2} is the $(n-2) \times (n-2)$ identity matrix and $\mathbf{0}_{n-2}$ is the $(n-2)$-dimensional zero vector.

A simple computation shows that the $n \times n$ matrix

$$U = \begin{pmatrix} 1 & 2 & 3 & \cdots & (n-1) & -(n-2) \\ 0 & 1 & 2 & \cdots & (n-2) & -(n-3) \\ 0 & 0 & 1 & \cdots & (n-3) & -(n-4) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix},$$

is an unimodular matrix transforming A into its column-style Hermite normal form, i.e, $AU = H$. Hence, it follows that the last two columns of U provide a lattice basis for $\ker_{\mathbb{Z}}(A_{n-2})$. Let us denote by L the $n \times 2$ matrix whose column vectors are the last two columns of U. Hence, as a consequence of ([7], Proposition 16.1) we know that the column vectors of $\begin{pmatrix} L \\ -L \end{pmatrix}$ provide a lattice basis for $\Lambda(A_{n-2})$ where $\Lambda(A_{n-2}) = \begin{pmatrix} A_{n-2} & 0_{n-2} \\ I_{n-2} & I_{n-2} \end{pmatrix}$ is the Lawrence lifting of A_{n-2}, being 0_{n-2} the $(n-2) \times (n-2)$ zero matrix. Let us denote the elements of this lattice basis by z_1 and z_2.

Now, consider the n-dimensional column vector $w = (0\ 1\ 2\ \cdots\ n-1)^T$ and let

$$u = \begin{pmatrix} w \\ \mathbf{0}_n \end{pmatrix}, \quad v = \begin{pmatrix} \mathbf{0}_n \\ w \end{pmatrix}.$$

Since $\Lambda(A_{n-2})u = \Lambda(A_{n-2})v = (\mathbf{0}_{n-2}\ w)^T$, we have that u, v are in the same fiber. However, adding any of the elements $\{\pm z_1, \pm z_2\}$ to u, results in at least one coordinate smaller or equal to $-(n-2)$. Hence, the lattice basis $\mathcal{L}_{n-2} := \{z_1, z_2\}$ fails to connect u, v inside $\mathcal{F}_{-q}((\mathbf{0}_{n-2}, w)^T)$ for any $q = 0, \ldots, n-3$. Therefore, for any $N > 0$, $\Lambda_N := \Lambda(A_{N+2})$ and $\mathcal{M}_N := \mathcal{L}_{N+2}$ satisfy the properties stated in Theorem 1.3.16.

□

2.2. Complexity of $(-1, S)$-Markov bases

We begin this section by outlining some of the steps used in the proof of Theorem 1.3.9 from [38]. These steps will provide the necessary tools for the proof of Theorem 1.3.17.

Let us start with a bounded polytope $P = \{\boldsymbol{y} \in \mathbb{R}^n_{\geq 0} : A\boldsymbol{y} = \boldsymbol{b}\}$, where $A = (a_{i,j})$ is an $m \times n$ matrix. The construction of T in Theorem 1.3.9 is typically carried out in three steps (see [38]). However, for our purposes, it suffices to focus on the last two steps, which are listed below.

Step 1) Representing P as a plane-sum entry-forbidden transportation polytope T'.

Let U be an integer upper bound for the entries of P. Then, it can be proved that for some $s, h \in \mathbb{Z}^+$ and a subset $E \subset [s] \times [s] \times [h]$, P can be represented as the polytope

$$T' = \Big\{ \boldsymbol{x} \in \mathbb{R}^{s \times s \times h}_{\geq 0} : x_{i,j,k} = 0 \text{ for all } (i,j,k) \notin E \text{ and } \sum_{i,j} x_{i,j,k} = c_k, \sum_{i,k} x_{i,j,k} = b_j, \sum_{j,k} x_{i,j,k} = a_i \Big\}.$$

This representation comes with an injection $\sigma' : [n] \to [s] \times [s] \times [h]$ and its induced coordinate-erasing projection $\pi' : \mathbb{R}^{s \times s \times h} \to \mathbb{R}^n$ that provides a bijection between P and T' and between their integer points.

Observation 2.2.1. *From the description of E in [38], it follows that for a given $\boldsymbol{y} = \pi'(\boldsymbol{x}) \in P$ (where $\boldsymbol{x} \in T'$) and for any $i \in [n]$ the coordinate y_i is embedded in s_i*

distinct coordinates of $\boldsymbol{x}$*, where*

$$s_i := \max\left(\sum_{k=1}^{m}\{a_{k,i} : a_{k,i} > 0\}, \sum_{k=1}^{m}\{|a_{k,i}| : a_{k,i} < 0\}\right).$$

Moreover, the explicit set of coordinates where y_i *is embedded is given by*

$$B_i := \left\{(j, j, \kappa(j)) : j \in \left\{1 + \sum_{l<i} s_l, \ldots, \sum_{l\leq j} s_l\right\}\right\}, \tag{2.1}$$

where κ *is a function* $\kappa : [s] \to [h]$ *completely determined by the matrix* A *and* $s = \sum_{i=1}^{n} s_i$ *is the same as in the description of* T'*. Then,* $B := \bigcup_{i=1}^{n} B_i$ *is a set of indices for which the corresponding entry in* $\boldsymbol{x}$ *is equal to some* y_i*. Furthermore, the set* B *is completely determined by* A*, so the previous embedding property holds for any* $\boldsymbol{y} \in P$*.*

Notice that under the assumption that A *has nonzero columns we have that* $|B| = \sum_{i=1}^{n} s_i$*. In other words, the set* B *is completely determined by the polyhedral representation of* P*.*

Step 2) Representing the polytope T' as a slim line-sum transportation polytope T.

Given T' as in the previous step, there are r, c and $(u_{i,j}) \in \mathbb{Z}^{I\times c}, (v_{i,k}) \in \mathbb{Z}^{I\times 3}, (w_{j,k}) \in \mathbb{Z}^{J\times 3}$ such that the transportation polytope

$$\widehat{T} = \left\{\boldsymbol{x} \in \mathbb{R}_{\geq 0}^{I\times J\times 3} : \sum_{k} x_{i,j,k} = u_{i,j}, \sum_{j} x_{i,j,k} = v_{i,k}, \sum_{i} x_{i,j,k} = u_{j,k}\right\}$$

represents T'.

With this discussion we are ready to present the following proof.

Proof of 1.3.17. Given a polytope $Q = \{\boldsymbol{x} \in \mathbb{R}^k_{\geq 0} : C\boldsymbol{x} = \boldsymbol{d}\}$, and a vector $\boldsymbol{u} = (u_1, \ldots, u_k) \in \mathbb{Z}^k$ we let $Q_{\boldsymbol{u}} := \{\boldsymbol{x} \in \mathbb{R}^k : C\boldsymbol{x} = \boldsymbol{d}$ and $x_i \geq u_i$ for every $i \in [k]\}$. Also, given any $D \subset [k]$, we let $\mathbb{1}_D$ be the indicator vector of D and write $\mathbb{1}$ when $D = [k]$.

Now, consider the polytope $P = \{\boldsymbol{y} \in \mathbb{R}^{\eta+2}_{\geq 0} : y_0 + y_{\eta+1} = 1, \theta_j y_0 - y_j = 0, j = 1, \ldots \eta\}$ introduced in the proof of Corollary 1.3.10 and let $\widehat{P} := P_{-1} + \mathbb{1}$. The integer points in $\widehat{P}$ are exactly

$$\boldsymbol{y}^1 = (0, 0, \ldots, 0, 1) + \mathbb{1}, \quad \boldsymbol{y}^2 = (1, \theta_1, \ldots, \theta_\eta, 0) + \mathbb{1}, \quad \text{and}$$

$$\boldsymbol{z}^1 = (2, 2\theta_1, \ldots, 2\theta_\eta, -1) + \mathbb{1}, \quad \boldsymbol{z}^2 = (-1, -\theta_1, \ldots, -\theta_\eta, 2) + \mathbb{1}$$

By the previous discussion, there are $s, h \in \mathbb{Z}^+$ and a plane-sum entry-forbidden transportation polytope $\widehat{T}' \subset \mathbb{R}^{s\times s\times h}_{\geq 0}$ representing $\widehat{P}$. Furthermore, by 2.2.1 we know that there is a subset $B \subset [s] \times [s] \times [h]$ such that for any $\boldsymbol{x} \in \widehat{T}'$, the entries corresponding to the indices in B are all entries of $\pi'(\boldsymbol{x}) \in P$.

Let r, c and $(u_{i,j}) \in \mathbb{Z}^{I\times c}, (v_{i,k}) \in \mathbb{Z}^{I\times 3}, (w_{j,k}) \in \mathbb{Z}^{J\times 3}$ such that the transportation polytope

$$\widehat{T} = \left\{ \boldsymbol{x} \in \mathbb{R}^{I\times J\times 3}_{\geq 0} : \sum_k x_{i,j,k} = u_{i,j}, \sum_j x_{i,j,k} = v_{i,k}, \sum_i x_{i,j,k} = u_{j,k} \right\}$$

represents $\widehat{T}'$ and let $\sigma : [s] \times [s] \times [h] \to [r] \times [c] \times [3]$ be the injection given by this representation. Let $S = \sigma(B)$ and let $\boldsymbol{p}^1, \boldsymbol{p}^2, \boldsymbol{q}^1, \boldsymbol{q}^2 \in \widehat{T}$ be the integer points corresponding to $\boldsymbol{y}^1, \boldsymbol{y}^2, \boldsymbol{z}^1, \boldsymbol{z}^2$, respectively. Then, consider the following transportation polytope

$$T = \left\{ \boldsymbol{x} \in \mathbb{R}^{I\times J\times 3}_{\geq 0} : \sum_k x_{i,j,k} = u_{i,j} - (\mathbb{1}_S)_{ij+}, \sum_j x_{i,j,k} = v_{i,k} - (\mathbb{1}_S)_{i+k}, \sum_i x_{i,j,k} = u_{j,k} - (\mathbb{1}_S)_{+jk} \right\}$$

and observe that $T_{-\mathbb{1}_S} = \widehat{T} - \mathbb{1}_S$. Moreover, since $\widehat{T}$ is a representation of $\widehat{P}$ it follows that the only integer points in $T_{-\mathbb{1}_S}$ are $\boldsymbol{p}^1 - \mathbb{1}_S, \boldsymbol{p}^2 - \mathbb{1}_S, \boldsymbol{q}^1 - \mathbb{1}_S$ and $\boldsymbol{q}^2 - \mathbb{1}_S$. By construction, the first 2 of these points are non-negative and any of the differences between any of the four points has either θ or θ appearing in the restriction of some η coordinates.

To see that $|S| = 1 + \sum_{i=1}^{\eta} \theta_i$ it is enough to find $|B_i|$ using 2.1 and the defining matrix of the polytope P. □

Before providing a proof for Proposition 1.3.18 let us introduce some notation. Given $i_1 \neq i_2 \in [r_1], k_1 \neq k_2 \in [3]$ and $j' \in [r_2]$ we define the $r_1 \times r_2 \times 3$ table $\boldsymbol{b}(i_1, i_2;\ j';\ k_1, k_2)$ as follows.

$$\boldsymbol{b}(i_1, i_2;\ j';\ k_1, k_2)_{i,j,k} = \begin{cases} 1, & \text{if } (i,j,k) \in \{(i_1, j', k_1), (i_2, j', k_2)\}, \\ -1, & \text{if } (i,j,k) \in \{(i_1, j', k_2), (i_2, j', k_1)\}, \\ 0, & \text{otherwise.} \end{cases}$$

We can think of this table as the embedding of a 2-way basic move in a 3-way table. Even though $\boldsymbol{b}(i_1, i_2;\ j';\ k_1, k_2)$ has one non-zero 2-margin (so it is not a move), it will help us to describe some moves more easily. For instance, $\boldsymbol{b}(i_1, i_2;\ j_1, j_2;\ k_1, k_2)$ can be written as a sum of

two embedded $r_1 \times 3$-moves:

$$\boldsymbol{b}(i_1, i_2;\ j_1, j_2;\ k_1, k_2) = \boldsymbol{b}(i_1, i_2;\ j_1;\ k_1, k_2) + \boldsymbol{b}(i_2, i_1;\ j_2;\ k_1, k_2).$$

DEFINITION 2.2.2. *Given a $r_1 \times r_2 \times 3$ table $\boldsymbol{m}$ and a subset $S = S_{r_1} \times S_{r_2} \times S_{r_3} \subset [I] \times [J] \times [3]$. We define $\boldsymbol{m}_S$ as the restriction of $\boldsymbol{m}$ on S. Under this definition $\boldsymbol{m}_S$ is a $|S_{r_1}| \times |S_{r_2}| \times |S_{r_3}|$ table.*

PROOF OF 1.3.18. Suppose without losing generality that $\tau : [r_2] \to [3]$ is a surjective function such that $[r_1] \times [r_2] \times [3] \backslash S = \bigcup_{j=1}^{r_2} \{(i, j, \tau(j)) : i \in [r_1]\}$.

Now, we will provide a way to find an infinite family of $(-q, S)$-extended fibers for which their non-negative $r_1 \times r_2 \times 3$ tables are not connected by basic moves.

Let $\boldsymbol{m} = (m_{i,j,k})$ be a $r_1 \times r_2 \times 3$ table such that $\boldsymbol{m}_{i,j,k} = 0$ for every $(i, j, k) \notin S$ and for every $t \in [3]$ let $S_t = \{(i, j, k) \in S : \tau(j) = t\} = [r_1] \times \tau^{-1}(t) \times ([3] \backslash \{t\})$. Notice that S_1, S_2, S_3 form a partition of S. Moreover, the support of any basic move that can be added to $\boldsymbol{m}$ while preserving non-negativity constraint must be contained in some S_t. Otherwise, if $\boldsymbol{b}$ is a basic move such that $\text{supp}(\boldsymbol{b}) \cap S_t, \text{supp}(b) \cap S_{t'} \neq \emptyset$, it would follow that $\boldsymbol{m} + \boldsymbol{b}$ has a -1 for some entry in $[r_1] \times [r_2] \times [r_3] \backslash S$ by a pigeonhole principle argument.

This implies that we can connect $\boldsymbol{m}$ to another $I \times J \times 3$ table $\boldsymbol{m}'$ (with basic moves) if and only if for every $t \in [3]$ we can connect $\boldsymbol{m}_{S_t}$ and $\boldsymbol{m}'_{S_t}$ with basic moves in their respective $(-q)$-extended fiber. In particular, if $\boldsymbol{m}'$ is connected to $\boldsymbol{m}$ we must have that $\boldsymbol{m}_{S_t}$ and $\boldsymbol{m}'_{S_t}$ are in the same fiber for every $t \in [3]$. In the rest of the proof, we will build $\boldsymbol{m}$ and $\boldsymbol{m}'$ such that the latter doesn't hold.

For every $t \in [3]$ pick some $j_t \in \tau^{-1}(t)$ and let us consider the move

$$\boldsymbol{n} = \boldsymbol{b}(1, 2;\ j_1;\ 1, 2) + \boldsymbol{b}(1, 2;\ j_2;\ 2, 3) + \boldsymbol{b}(1, 2;\ j_3;\ 1, 3)$$

of degree 6. By the choice of j_1, j_2 and j_3, we know that we can add $\boldsymbol{n}$ to $\boldsymbol{m}$ while preserving the non-negativity constraint. i.e., $\boldsymbol{m}' := \boldsymbol{m} + \boldsymbol{n}$ is a non-negative table in the fiber of $\boldsymbol{m}$. Moreover, by the definition of $\boldsymbol{n}$ it follows that $\boldsymbol{m}'_{S_1} = \boldsymbol{b}(1, 2;\ j_1;\ 1, 2) + \boldsymbol{m}_{S_1}$ and therefore the 2-margins of $\boldsymbol{m}_{S_1}$ and $\boldsymbol{m}'_{S_1}$ are not the same (their ik-margins differ), contradicting our previous observations. Hence $\boldsymbol{m}$ and $\boldsymbol{m}'$ are not connected by basic moves and therefore the fiber of $\boldsymbol{m}$ is not connected by basic moves. □

2.3. Bounding the Graver Basis Size for Hierarchical Models

PROOF OF COROLLARY 1.3.23. Let Δ be a simplicial complex with ground set $[m]$ and maximal faces $D_1, \dots, D_t$. Let $V \subset [m]$ such that for every $i \in [m]$, $V \subset D_i$ or $V \subset D_i^c$. We will prove that , $A_\Delta = [A, B]^{(\eta_V)}$ where $\eta_V = \prod_{l \in V} r_l$, A is a $\sum_{j: D_j \supseteq V} \frac{\eta_j}{\eta_V} \times \frac{\eta}{\eta_V}$ matrix and B is a $\sum_{j: D_j^c \supseteq V} \eta_j \times \frac{\eta}{\eta_V}$ matrix.

First notice that the columns of A_Δ are in bijection with the entries of a $d_1 \times \cdots \times d_m$ table so we can label each column with an multi-index $\boldsymbol{i} = (i_1, \dots, i_m)$. For each multi-index $\boldsymbol{i}$ and $D \subset [m]$ we define $\boldsymbol{i}_D := (i_j)_{j \in D} \in \prod_{l \in D} [r_l]$. Observe that each row can be identified with a pair $(D_k, \boldsymbol{f})$ where $\boldsymbol{f} \in \prod_{j \in F_k} [d_j]$. Without loosing generality assume that $V = \{1, \dots, v\}$ for some $v \in [m]$ and assume that $V \subset D_1, \dots, D_s$ and $V \subset D_{s+1}^c, \dots, D_t^c$.

Now, following the construction of [72] we give a description of A_Δ as an η_V-fold matrix. Let us order the columns of A_Δ lexicographically: this order provides a partition of the columns into groups labeled by multi-indices in $\prod_{i=1}^{v} [r_i]$, i.e., the group corresponding to $\boldsymbol{i}_V$ is the set $\mathcal{C}_{\boldsymbol{i}_V} := \{(\boldsymbol{i}_V, \boldsymbol{i}') : \boldsymbol{i}' \in \prod_{l=v+1}^{m} [r_l]\}$, lexicographically ordered.

The rows will be ordered in the following way: For each $\boldsymbol{i}_V \in \prod_{l=1}^{v} [d_l]$ define the set of rows $\mathcal{R}_{\boldsymbol{i}_V} := \{(F_j, \boldsymbol{f}) : j \in [r], \boldsymbol{f}_V = i_V \text{ and } f_l \in [r_l] \text{ for every } l \notin V\}$. Furthermore, given $(F_j, \boldsymbol{f}), (F_{j'}, \boldsymbol{f}') \in \mathcal{R}_{i_V}$ we say $(F_j, \boldsymbol{f}) \prec (F_{j'}, \boldsymbol{f}')$ if $j < j'$ or if $j = j'$ and $\boldsymbol{f}$ is lexicographically smaller than $\boldsymbol{f}'$. We denote by $\mathcal{R}$ the rest of the pairs $(F, \boldsymbol{f})$ that don't belong to any $\mathcal{R}_{\boldsymbol{i}_V}$. Finally, we order the rows of A_Δ by groups $\mathcal{R}_{\boldsymbol{i}_V}$ using a lexicographic order on $\{\boldsymbol{i}_V : \boldsymbol{i} \in \prod_{j=1}^{m} [r_j]\}$ and leaving the rows $\mathcal{R}$ at the end in any order.

This order of the rows and columns provides a block description $[A, B]^{(\eta)}$ for A_Δ where $A = A_{link_\Delta(V)}$ and $B = A_{\Delta \backslash V}$ are the design matrices of the hierarchical model associated to $link_\Delta(V) := \{F \backslash V : F \supset V\}$ and $\Delta \backslash V = \{F \in \Delta : V \subseteq F^c\}$, respectively.

□

CHAPTER 3

Connecting Spaces of Graphs with Fixed Degree-color Sequence

The purpose of this chapter is to provide proofs for the contributions outlined in Section 1.4.1. Each section builds toward the proofs of the main results introduced there.

Theorem 1.4.6 establishes a quadratic Markov basis for the design matrix $D\mathcal{C}_{n,z}$ of the β-SBM, resolving an open problem posed in [79]. This result has recently been used by [18] to compute the maximum likelihood degree for the β-SBM.

Proposition 1.4.7 shows that to ensure connected fiber-graphs of simple graphs with a fixed degree-color sequence, any subset $\mathcal{B} \subset \ker_{\mathbb{Z}} D\mathcal{C}_{n,z}$ must contain elements whose 1-norm grows as k increases. Immediate consequences of this proposition include Corollary 1.4.9 and Theorem 1.4.10.

Finally, Theorem 1.4.14 demonstrates that the generators given by Theorem 1.4.6 also form a Gröbner basis for $I_{D\mathcal{C}_{n,z}}$, extending a result from [41].

3.1. A quadratic Markov basis

In this section, we focus on graphs with vertex set $[n]$, represented by vectors $\boldsymbol{g} = (g_{uv})_{1\leq u<v\leq n} \in \mathbb{N}^{\binom{n}{2}}$. We define $V(\boldsymbol{g}) = [n]$ and $E(\boldsymbol{g}) = \{\{u, v\} : g_{uv} \neq 0\}$. Since $g_{uv} \in \mathbb{N}$ for all $1 \leq u < v \leq n$, we occasionally refer to $\boldsymbol{g}$ as a multigraph.

As in Section 1.4, for a k-coloring $z : [n] \to [k]$, the vector $c(\boldsymbol{g}, z) = \big(c(z, i, j) : 1 \leq i \leq j \leq k\big)$ represents the c-degree of $\boldsymbol{g}$.

Given a graph $\boldsymbol{g} \in \mathbb{N}^{\binom{n}{2}}$, we say that an edge $uv \in E(\boldsymbol{g})$ is *positive* if $g_{uv} > 0$ and *negative* if $g_{uv} < 0$, where $|g_{uv}|$ represents the multiplicity of the edge uv. For simplicity, we may occasionally abuse notation and treat $\boldsymbol{g} \in \mathbb{Z}^{\binom{n}{2}}$ as a graph when the context allows.

For $\boldsymbol{g} \in \mathbb{Z}^{\binom{n}{2}}$ and $v \in [n]$ we define the *positive degree* and *negative degree* of v in $\boldsymbol{g}$ as

$$\deg^+_{\boldsymbol{g}}(v) := \sum_{u\in[n]:g_{uv}>0} g_{uv} \quad \text{and} \quad \deg^-_{\boldsymbol{g}}(v) := \sum_{u\in[n]:g_{uv}<0} -g_{uv}, \tag{3.1}$$

respectively. In other words, $deg^+_{\boldsymbol{g}}(u)$ and $deg^-_{\boldsymbol{g}}(u)$ are the numbers of positive and negative edges incident with u, respectively. We define the *positive degree sequence* and the *negative degree sequence*

as the vectors $d^+(\boldsymbol{g}) = (\deg^+_{\boldsymbol{g}}(1), \ldots, \deg^+_{\boldsymbol{g}}(n))$ and $d^-(\boldsymbol{g}) = (\deg^-_{\boldsymbol{g}}(1), \ldots, \deg^-_{\boldsymbol{g}}(n))$, respectively. Also, for any $1 \leq i \leq j \leq k$, let

$$c^+_{\boldsymbol{g}}(z,i,j) := \sum_{\substack{u \in z^{-1}(i),\, v \in z^{-1}(j): \\ g_{uv} > 0}} g_{uv} \quad \text{and} \quad c^-_{\boldsymbol{g}}(z,i,j) := \sum_{\substack{u \in z^{-1}(i),\, v \in z^{-1}(j): \\ g_{uv} < 0}} -g_{uv}. \tag{3.2}$$

This means that $c^+_{\boldsymbol{g}}(z,i,j)$ and $c^-_{\boldsymbol{g}}(z,i,j)$ are the number of positive and negative edges, respectively, that are connecting an i-th colored vertex with an j-th colored vertex. We define the *positive color sequence* and the *negative color sequence* as the vectors $c^+(z,\boldsymbol{g}) = \left(c^+_{\boldsymbol{g}}(z,i,j) : 1 \leq i \leq j \leq k\right)$ and $c^-(z,\boldsymbol{g}) = \left(c^-_{\boldsymbol{g}}(z,i,j) : 1 \leq i \leq j \leq k\right)$, respectively. Notice that when $\boldsymbol{g} \in \mathbb{N}^{\binom{n}{2}}$, $(d^+(\boldsymbol{g}); c^+(z,\boldsymbol{g}))$ coincides with the c-degree sequence and $(d^-(\boldsymbol{g}); c^-(z,\boldsymbol{g}))$ is a vector of zeros. When z is clear from the context we will write $c^\pm_{\boldsymbol{g}}(i,j)$ instead of $c^\pm_{\boldsymbol{g}}(z,i,j)$. We say that $\boldsymbol{g} \in \mathbb{Z}^{\binom{n}{2}}$ satisfies the *degree-balance condition* if $d^+(\boldsymbol{g}) = d^-(\boldsymbol{g})$ and the *color-balance condition with respect to* z if $c^+(z,\boldsymbol{g}) = c^-(z,\boldsymbol{g})$. When z is clear from the context, we simply say that $\boldsymbol{g}$ satisfies the color-balance condition.

Let $n \in \mathbb{Z}_+$ be a positive integer, z a k-coloring of $[n]$, and $D\mathcal{C}_{n,z}$ the matrix as defined in Equation 1.6. Notice that for any $\boldsymbol{g} \in \mathbb{Z}^{\binom{n}{2}}$,

$$D\mathcal{C}_{n,z}\,\boldsymbol{g} = (d^+(\boldsymbol{g}) - d^-(\boldsymbol{g}); c^+(\boldsymbol{g}) - c^-(\boldsymbol{g})). \tag{3.3}$$

This means that $\boldsymbol{g} \in \ker_{\mathbb{Z}} D\mathcal{C}_{n,z}$ if and only if $\boldsymbol{g}$ satisfies both the degree and the color-balance conditions. It is not hard to see that any $\boldsymbol{g} \in \mathbb{Z}^{\binom{n}{2}}$ satisfying the degree-balance condition must be a union of closed even walks whose edges in the walk alternate between positive and negative edges. In fact, when $D\mathcal{C}_{n,z}$ is regarded as the incidence graph of a 3-uniform hypergraph the elements of $\ker_{\mathbb{Z}} A$ can be understood using the notion of balanced walk on a hypergraph. These are known as *monomial walks* in the literature ([93], [110]). Hence we say that $\boldsymbol{g} \in \mathbb{Z}^{\binom{n}{2}}$ is a *monomial walk with respect to* $z : [n] \to [k]$ if $\boldsymbol{g} \in \ker_{\mathbb{Z}} D\mathcal{C}_{n,z}$. For convenience, when considering a monomial walk $\boldsymbol{g}$, sometimes we will describe it using a vertex sequence enclosed by square brackets: $[v_1, v_2, \ldots, v_{2l-1}, v_{2l}]$. This notation means that $\boldsymbol{g}$ is the element in $\ker_{\mathbb{Z}} D\mathcal{C}_{n,z}$ entrywise defined by

$$g_{uv} = \sum_{i=1}^{l} \mathbb{1}_{\{u,v\}=\{v_{2i-1},v_{2i}\}} - \mathbb{1}_{\{u,v\}=\{v_{2i},v_{2i+1}\}},$$

with $2l+1=1$ and $\mathbb{1}$ being the indicator function. The use of the different notations (either vector or brackets) will depend on the context.

EXAMPLE 3.1.1. *Let $DC_{5,z}$ the matrix of Example 1.4.5 and $\boldsymbol{g}=(0,0,-1,1,-1,2,-1,0,1,-1)\in \ker_{\mathbb{Z}} DC_{n,5}$ be the monomial walk illustrated below. One way to write $\boldsymbol{g}$ using bracket notation is $\boldsymbol{g}=[1,5,2,4,5,3,2,4]$. In this case $x^{\boldsymbol{g}^+}-x^{\boldsymbol{g}^-}=x_{15}x_{24}^2x_{35}-x_{14}x_{23}x_{25}x_{45}$.*

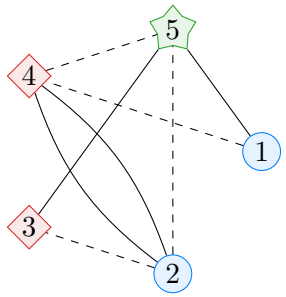


FIGURE 3.1. Monomial walk $m=[1,5,2,4,5,3,2,4]$.

As noticed previously, for any c-degree sequence $(\boldsymbol{d};\boldsymbol{c})\in\mathbb{N}^{n+\binom{k+1}{2}}$, $\mathcal{F}_{DC_{n,z}}(d;c)$ is the set of multigraphs with fixed c-degree sequence $(\boldsymbol{d};\boldsymbol{c})$. Meaning that a Markov basis for $DC_{n,z}$ is a set of moves $\mathcal{B}$ that allow us to connect any two multigraphs with a fixed degree sequence by using moves in $\mathcal{B}$. To prove Theorem 1.4.6 we use the algebraic analogue provided by Theorem 1.1.8. In other words, we show that $\mathcal{M}_{n,z}=\{\boldsymbol{g}\in\ker_{\mathbb{Z}} DC_{n,z}\colon ||\boldsymbol{g}||_1=4\}$ is a Markov basis for $DC_{n,z}$ by proving that $I_{\mathcal{M}_{n,z}}:=\langle\{x^{\boldsymbol{g}^+}-x^{\boldsymbol{g}^-}:\boldsymbol{g}\in\mathcal{M}_{n,z}\}\rangle$ is equal to $I_{DC_{n,z}}$.

To do so, we will use the combinatorial conditions of monomial walks in order to reduce binomials of degree greater than two by splitting the monomial walks of length longer than four into shorter walks. First, let us see that the toric monomial map introduced in Definition 1.1.1, whose vanishing ideal is $I_{DC_{n,z}}$, can be explicitly written as

$$\begin{aligned}\varphi_\beta:\mathbb{K}[x_{uv}:1\le u<v\le n]&\to\mathbb{K}[\{s_1,\dots,s_n\}\cup\{t_{ij}:1\le i\le j\le k\}];\\ x_{uv}&\mapsto s_us_vt_{z(u)z(v)}.\end{aligned}\tag{3.4}$$

Then, for any $\boldsymbol{g}\in\mathbb{Z}^{\binom{n}{2}}$, $\varphi_\beta(\boldsymbol{x}^{\boldsymbol{g}^+})=\varphi_\beta(\boldsymbol{x}^{\boldsymbol{g}^-})$ if and only if $DC_{n,z}\,\boldsymbol{g}^+=DC_{n,z}\,\boldsymbol{g}^-$, which means that $\boldsymbol{x}^{\boldsymbol{g}^+}-\boldsymbol{x}^{\boldsymbol{g}^-}\in\ker\varphi_\beta=I_{DC_{n,z}}$ if and only if $\boldsymbol{g}\in\ker_{\mathbb{Z}} DC_{n,z}$. In fact, as a consequence of [104, Corollary 4.3] it follows that $I_{DC_{n,z}}=\langle\boldsymbol{x}^{\boldsymbol{g}^+}-\boldsymbol{x}^{\boldsymbol{g}^-}:\boldsymbol{g}\in\ker_{\mathbb{Z}} DC_{n,z}\rangle$. This immediately implies that

$I_{\mathcal{M}_{n,z}} \subseteq I_{\mathcal{DC}_{n,z}}$. Therefore, to prove Theorem 1.4.6 it suffices to show that $I_{\mathcal{DC}_{n,z}} \subseteq I_{\mathcal{M}_{n,z}}$. To do so, we start by providing a combinatorial description of $\mathcal{M}_{n,z}$.

LEMMA 3.1.2. *$\mathcal{M}_{n,z} = \{[uvu'v'] : z(u) = z(u') \text{ or } z(v) = z(v')\}$. In other words, the elements of $\mathcal{M}_{n,z}$ are 4-cycles with at least two opposite vertices of the same color.*

PROOF. First suppose without losing generality that $\boldsymbol{g} = [uvu'v']$ with $z(u) = z(u')$. By convention this means that $uv, u'v'$ are positive edges, $vu', v'u$ are negative edges in $E(\boldsymbol{g})$. Since $z(u) = z(u')$, we have that $t_{z(u)z(v)} = t_{z(u')z(v)}$ and $t_{z(u')z(v')} = t_{z(u)z(v')}$, which means

$$\begin{aligned}\varphi_\beta(\boldsymbol{x}^{\boldsymbol{g}^+} - \boldsymbol{x}^{\boldsymbol{g}^-}) &= s_u s_v t_{z(u)z(v)} \cdot s_{u'} s_{v'} t_{z(u')z(v')} - s_u s_{v'} t_{z(u)z(v')} \cdot s_{u'} s_v t_{z(u')z(v)} \\ &= s_u s_v s_{u'} s_{v'} (t_{z(u)z(v)} t_{z(u')z(v')} - t_{z(u)z(v')} t_{z(u')z(v)}) = 0,\end{aligned}$$

implying that $\boldsymbol{x}^{\boldsymbol{g}^+} - \boldsymbol{x}^{\boldsymbol{g}^-} \in I_{\mathcal{DC}_{n,z}}$, or equivalently, $\boldsymbol{g} \in \mathcal{M}_{n,z}$.

Now, suppose that $\boldsymbol{g} \in \mathcal{M}_{n,z}$. The degree-balance condition implies that $\boldsymbol{g}$ is a 4-cycle $[uvu'v']$ with $uv, u'v'$ positive and $vu', v'u$ negative edges. Let us assume without losing generality that $z(u) \neq z(u')$. Because $\boldsymbol{g}$ has a positive edge between colors $z(u)$ and $z(v)$, the color-balance condition implies that at least one of the negative edges uv' or $u'v$ connects colors $z(u)$ and $z(v)$. This is equivalent to having $z(v') = z(v)$ or $z(u') = z(u)$. Hence, $z(v') = z(v)$ by our earlier assumptions. □

We will show that $I_{\mathcal{DC}_{n,z}} \subseteq I_{\mathcal{M}_{n,z}}$ by proving that for any $\boldsymbol{x}^{\boldsymbol{g}^+} - \boldsymbol{x}^{\boldsymbol{g}^-} \in I_{\mathcal{DC}_{n,z}}$, we can either peel off 4-cycles from $\boldsymbol{g}$ (Lemma 3.1.5 below) or, alternatively, use 4-cycles to reconnect $\boldsymbol{g}$ (Lemma 3.1.4 below). This allows us to obtain a new monomial walk from which we can peel off 4-cycles belonging to $I_{\mathcal{DC}_{n,z}}$. This process enables us to express $\boldsymbol{x}^{\boldsymbol{g}^+} - \boldsymbol{x}^{\boldsymbol{g}^-}$ as the sum of an element in $I_{\mathcal{M}_{n,z}}$ and a binomial in $I_{\mathcal{DC}_{n,z}}$ with a degree smaller than $\deg(\boldsymbol{x}^{\boldsymbol{g}^+} - \boldsymbol{x}^{\boldsymbol{g}^-})$. Before presenting the proofs, we illustrate the idea with the following example.

EXAMPLE 3.1.3. *Let $n = 8, k = 2$ and $z : [8] \to [2]$ such that $z^{-1}(1) = \{1, 4, 5, 6\}, z^{-1}(2) = \{2, 3, 7, 8\}$. Consider $f = \boldsymbol{x}^{\boldsymbol{g}^+} - \boldsymbol{x}^{\boldsymbol{g}^-} = x_{12}x_{34}x_{56}x_{78} - x_{14}x_{23}x_{58}x_{67} \in I_{\mathcal{DC}_{n,z}}$ and observe that $f = x_{34}x_{56}(x_{12}x_{78} - x_{17}x_{28}) + f'$ where $f' = \boldsymbol{x}^{\boldsymbol{g}'^+} - \boldsymbol{x}^{\boldsymbol{g}'^-} = x_{34}x_{56}x_{17}x_{28} - x_{14}x_{23}x_{58}x_{67}$. As we illustrate in the picture below, this way of rewriting f corresponds to rewriting $\boldsymbol{g}$ as a sum of a 4-cycle and a monomial walk $\boldsymbol{g}'$ with the same number of edges as $\boldsymbol{g}$. Notice that $\boldsymbol{g}$ and $\boldsymbol{g}'$ differ by a switch.*

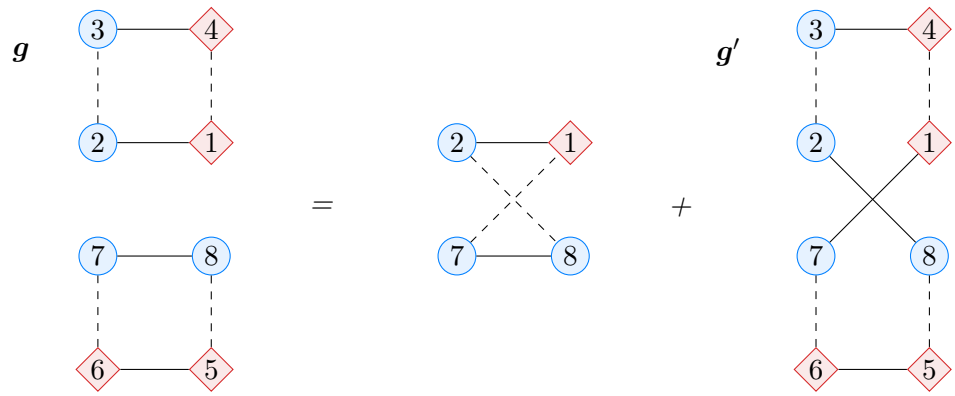


FIGURE 3.2. Reconnecting with a 4-cycle. On the left of the equality is the graph $\boldsymbol{g}$, while on the right are the 4-cycle corresponding to $x_{12}x_{78} - x_{17}x_{28}$ and $\boldsymbol{g}'$.

Now, we can peel off a 4-cycle from $\boldsymbol{g}'$ to obtain a monomial walk $\boldsymbol{g}''$ with less edges than $\boldsymbol{g}'$ (see Figure 3.3). Algebraically, this means that $f' = x_{23}x_{58}(x_{17}x_{46} - x_{14}x_{67}) + x_{46}(x_{28}x_{34}x_{56} - x_{23}x_{58}x_{17})$, where $f'' = \boldsymbol{x}^{\boldsymbol{g}''+} - \boldsymbol{x}^{\boldsymbol{g}''-} = x_{28}x_{34}x_{56} - x_{23}x_{58}x_{17}$. In fact, we can continue peeling off 4-cycles from $\boldsymbol{g}''$ in order to prove that $f \in \langle \boldsymbol{x}^{\boldsymbol{g}^+} - \boldsymbol{x}^{\boldsymbol{g}^-} : \boldsymbol{g} \in \mathcal{M}_{n,z} \rangle$.

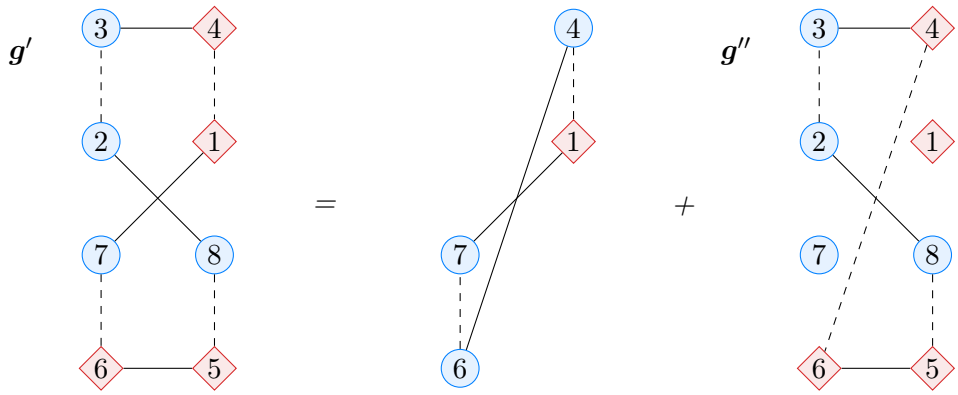


FIGURE 3.3. Peeling off a 4-cycle. On the left of the equality is the graph $\boldsymbol{g}$, while on the right are the 4-cycle corresponding to $x_{17}x_{46} - x_{14}x_{67}$ and $\boldsymbol{g}'$.

LEMMA 3.1.4. *For any $f = \boldsymbol{x}^{\boldsymbol{g}^+} - \boldsymbol{x}^{\boldsymbol{g}^-} \in I_{DC_{n,z}}$, there exists $f' \in I_{\mathcal{M}_{n,z}}$ and $f'' \in I_{DC_{n,z}}$ with $\deg(f) = \deg(f'')$ such that $f = f' + f''$, where $f'' = \boldsymbol{x}^{\boldsymbol{g}''+} - \boldsymbol{x}^{\boldsymbol{g}''-}$ and $\boldsymbol{g}''$ contains a subwalk uvw such that $z(u) = z(w)$ and uv, vw have different signs.*

PROOF. Let $f = \boldsymbol{x}^{\boldsymbol{g}^+} - \boldsymbol{x}^{\boldsymbol{g}^-} \in I_{DC_{n,z}}$ and let uv be a positive edge in $\boldsymbol{g}$. By the color-balance condition, there must exists a negative edge $u'v' \in E(\boldsymbol{g})$ such that $z(u') = z(u)$ and $z(v') = z(v)$. Since no edge in $\boldsymbol{g}$ can be positive and negative at the same time, it follows that $uv \neq u'v'$.

Moreover, if $u = u'$ then vuv' is a subwalk such that $z(v) = z(v')$ with vu positive and uv' negative so the statement would follow. Hence, we assume that $u \neq u'$ and similarly $v \neq v'$.

Now, given that each of u' and v' are adjacent to the negative edge $u'v'$, the degree-balance condition guarantees the existence of positive edges $u'\hat{w}, v'w$ in $E(\boldsymbol{g})$. Let us consider the following 2 cases:

(i) $\{w, \hat{w}\} = \{u, v\}$.

In case $w = u, \hat{w} = v$, then $uv'u'$ is a subwalk such that $z(u) = z(u')$, $uv' = wv'$ is positive and $v'u'$ is negative, so it would be enough to take $f' = 0, f'' = f$. Then, assume that $w = v, \hat{w} = u$. In such case, since u is adjacent to two positive edges, there must exists (by the degree-balance condition) a negative edge $u\hat{u} \in E(\boldsymbol{g})$. Notice that since $z(u) = z(u')$, $x_{u'\hat{u}}x_{uv'} - x_{u\hat{u}}x_{u'v'} \in I_{DC_{n,z}}$ and

$$\begin{aligned} f = \boldsymbol{x}^{\boldsymbol{g}^+} - \boldsymbol{x}^{\boldsymbol{g}^-} &= \boldsymbol{x}^{\boldsymbol{g}^+} - \boldsymbol{x}^{\boldsymbol{\alpha}}x_{u\hat{u}}x_{u'v'} \\ &= \boldsymbol{x}^{\boldsymbol{\alpha}}(x_{u'\hat{u}}x_{uv'} - x_{u\hat{u}}x_{u'v'}) + (x^{\boldsymbol{g}^+} - \boldsymbol{x}^{\boldsymbol{\alpha}}x_{u'\hat{u}}x_{uv'}), \end{aligned}$$

where, $\boldsymbol{\alpha} \in \mathbb{N}^{\binom{n}{2}}$ is such that $\boldsymbol{x}^{\boldsymbol{\alpha}}x_{u\hat{u}}x_{u'v'} = \boldsymbol{x}^{\boldsymbol{g}^-}$. Then we take $f' = \boldsymbol{x}^{\boldsymbol{\alpha}}(x_{u'\hat{u}}x_{uv'} - x_{u\hat{u}}x_{u'v'})$, $f'' = \boldsymbol{x}^{\boldsymbol{g}^+} - \boldsymbol{x}^{\boldsymbol{\alpha}}x_{u'\hat{u}}x_{uv'}$ and the subwalk vuv' satisfies that $z(v) = z(v')$, uv is positive and uv' is negative.

(ii) $\{w, \hat{w}\} \neq \{u, v\}$.

Assume without loss of generality that $w \notin \{u, v\}$. Given that $z(v) = z(v')$, it follows that $x_{uv}x_{wv'} - x_{uv'}x_{wv} \in I_{DC_{n,z}}$. Let us observe that

$$\begin{aligned} f = \boldsymbol{x}^{\boldsymbol{g}^+} - \boldsymbol{x}^{\boldsymbol{g}^-} &= \boldsymbol{x}^{\boldsymbol{\alpha}}x_{uv}x_{wv'} - \boldsymbol{x}^{\boldsymbol{g}^-} \\ &= \boldsymbol{x}^{\boldsymbol{\alpha}}(x_{uv}x_{wv'} - x_{uv'}x_{wv}) + (\boldsymbol{x}^{\boldsymbol{\alpha}}x_{uv'}x_{wv} - x^{\boldsymbol{g}^-}), \end{aligned}$$

where $\boldsymbol{\alpha} \in \mathbb{N}^{\binom{n}{2}}$ is such that $\boldsymbol{x}^{\boldsymbol{\alpha}}x_{uv}x_{wv'} = \boldsymbol{x}^{\boldsymbol{g}^+}$. In this case we set $f' = \boldsymbol{x}^{\boldsymbol{\alpha}}(x_{uv}x_{wv'} - x_{uv'}x_{wv})$, $f'' = \boldsymbol{x}^{\boldsymbol{\alpha}}x_{uv'}x_{wv} - \boldsymbol{x}^{\boldsymbol{g}^-} \in I_{DC_{n,z}}$. Notice that $\deg(f'') = \deg(f)$ and that the subwalk $uv'u'$ satisfies that $z(u) = z(u')$, uv' is positive and $v'u'$ is negative.

□

LEMMA 3.1.5. *Let* $f = \boldsymbol{x}^{\boldsymbol{g}^+} - \boldsymbol{x}^{\boldsymbol{g}^-} \in I_{D\mathcal{C}_{n,z}}$. *Suppose* $\boldsymbol{g}$ *contains a subwalk* uvw *such that* $z(u) = z(w)$ *and* uv, vw *have different signs. Then,* $f = f' + x_{uv}f''$ *for some* $f' \in I_{\mathcal{M}_{n,z}}$ *and* $f'' \in I_{D\mathcal{C}_{n,z}}$ *with* $\deg(f'') = \deg(f) - 1$.

PROOF. Suppose without loss of generality that uv is negative and vw is positive. Then the degree-balance condition guarantees the existence of a positive edge uu'. Consider the following two cases:

(i) $u' \neq w$.

Since $z(u) = z(w)$, if follows that $x_{uu'}x_{wv} - x_{uv}x_{wu'} \in I_{D\mathcal{C}_{n,z}}$. Let $\boldsymbol{\alpha}, \boldsymbol{\alpha}' \in \mathbb{N}^{\binom{n}{2}}$ be such that $\boldsymbol{x}^{\boldsymbol{g}^+} = \boldsymbol{x}^{\boldsymbol{\alpha}}x_{uu'}x_{wv}$ and $\boldsymbol{x}^{\boldsymbol{g}^-} = x_{uv}\boldsymbol{x}^{\boldsymbol{\alpha}'}$. Then, we have that

$$\begin{aligned} f = \boldsymbol{x}^{\boldsymbol{g}^+} - \boldsymbol{x}^{\boldsymbol{g}^-} &= \boldsymbol{x}^{\boldsymbol{\alpha}}x_{uu'}x_{wv} - x_{uv}\boldsymbol{x}^{\boldsymbol{\alpha}'} \\ &= \boldsymbol{x}^{\boldsymbol{\alpha}}(x_{uu'}x_{wv} - x_{uv}x_{wu'}) + x_{uv}(\boldsymbol{x}^{\boldsymbol{\alpha}}x_{wu'} - \boldsymbol{x}^{\boldsymbol{\alpha}'}). \end{aligned}$$

Let $f' = \boldsymbol{x}^{\boldsymbol{\alpha}}(x_{uu'}x_{wv} - x_{uv}x_{wu'})$ and $f'' = \boldsymbol{x}^{\boldsymbol{\alpha}}x_{wu'} - \boldsymbol{x}^{\boldsymbol{\alpha}'}$. Given that $I_{D\mathcal{C}_{n,z}}$ is prime and $f, x_{uu'}x_{wv} - x_{uv}x_{wu'}$ both belong to $I_{D\mathcal{C}_{n,z}}$ we have that $f'' \in I_{D\mathcal{C}_{n,z}}$. Furthermore, $\deg(f'') = \deg(f) - 1$.

(ii) $u' = w$.

In this case uw is a positive edge in $\boldsymbol{g}$ so by the degree-balance condition there must be a negative edge ww' with $w' \notin \{u, v\}$. This situation is analogous to the previous case since uvw is a subwalk with $z(u) = z(w)$, uv negative, vw positive and ww' a negative edge in $\boldsymbol{g}$ such that $w' \neq u$.

□

PROOF. As previously mentioned, it suffices to show that $I_{D\mathcal{C}_{n,z}} \subseteq I_{\mathcal{M}_{n,z}}$. Let us remember that $I_{D\mathcal{C}_{n,z}} = \langle \boldsymbol{x}^{\boldsymbol{g}^+} - \boldsymbol{x}^{\boldsymbol{g}^-} : \boldsymbol{g} \in \ker_{\mathbb{Z}} D\mathcal{C}_{n,z} \rangle$ and let $f = \boldsymbol{x}^{\boldsymbol{g}^+} - \boldsymbol{x}^{\boldsymbol{g}^-} \in I_{D\mathcal{C}_{n,z}}$. If $\deg(f) = 2$, then $f \in I_{\mathcal{M}_{n,z}}$ by definition. Suppose that $\deg(f) = k + 1$. By Lemma 3.1.4 and Lemma 3.1.5 we can write $f = f' + x_{uv}f''$ for some $1 \leq u < v \leq n$ where $f' \in I_{\mathcal{M}_{n,z}}$ and $\deg(f'') = k$. By induction on the degree we have that $f'' \in I_{\mathcal{M}_{n,z}}$, hence $f \in I_{\mathcal{M}_{n,z}}$. □

3.2. Restriction to Simple Graphs

In this section, we provide a proof of Proposition 1.4.7, which demonstrates that as the number of colors k increases, large moves become necessary in a binary Markov basis for $D\mathcal{C}_{n,z}$.

Proof of Proposition 1.4.7. Let $k \geq 3$, be an integer and z be the k-coloring of $[2k]$ such that $z(u) \equiv u \pmod{k}$ for every $u \in [2k]$. Let $\mathbf{3}_k$ and $\mathbf{1}_k$ be the vectors of size k with all 3's and all 1's, respectively. Consider $\boldsymbol{d}_k = (\mathbf{3}_k; \mathbf{1}_k) \in \mathbb{N}^{2k}$ and $\boldsymbol{c}_k \in \mathbb{N}^{\binom{k+1}{2}}$ be such that for every $1 \leq u \leq v \leq k$

$$c_k(u,v) = \begin{cases} 2 & \text{if } |u-v| \equiv 1 \pmod{k} \\ 0 & \text{otherwise.} \end{cases}$$

In order to prove that the simple-graph fiber $\mathcal{F}_{D\mathcal{C}_{2k,z}}(\boldsymbol{d}_k, \boldsymbol{c}_k; \mathbf{0}, \mathbf{1})$ contains only two elements we start by making the following two claims for every $\boldsymbol{g} \in \mathcal{F}_{D\mathcal{C}_{2k,z}}(\boldsymbol{d}_k, \boldsymbol{c}_k; \mathbf{0}, \mathbf{1})$.

Claim 1. If $\{u+k, v\} \in E(\boldsymbol{g})$ for $u \in [k]$, $v \in [2k]$ with $v \equiv u + \epsilon \pmod{k}$ and $\epsilon \in \{1, -1\}$; then $\{u, w\} \in E(\boldsymbol{g})$ for any $w \in [2k]$ with $w \equiv u - \epsilon \pmod{k}$.

Claim 2. The set $\{k+1, k+2, \ldots, 2k\}$ is independent in $\boldsymbol{g}$.

Assuming both claims, let $\boldsymbol{g}$ be a simple graph in $\mathcal{F}_{D\mathcal{C}_{2k,z}}(\boldsymbol{d}_k, \boldsymbol{c}_k; \mathbf{0}, \mathbf{1})$. Since $\boldsymbol{c}_k(1, v) \neq 0$ if and only if $v \in \{2, k\}$, it follows from Claim 2 that either $\{1+k, 2\}$ or $\{1+k, k\}$ is an edge in $E(\boldsymbol{g})$ but not both. We will prove that choosing one of the previous two edges determines $\boldsymbol{g}$ completely. Assume that $\{1+k, 2\}$ is an edge in $E(\boldsymbol{g})$. By Claim 1 we have that $\{1, k\}$ and $\{1, 2k\}$ are both edges in $E(\boldsymbol{g})$. Now, since $\{2k, 1\} \in E(\boldsymbol{g})$ it follows from Claim 1 that $\{k, k-1\}$ and $\{k, 2k-1\}$ are both edges in $E(\boldsymbol{g})$. Continuing with this process and repeatedly, applying Claim 1 shows that for any $u \in [k]$ and $v \in [2k]$ such that $v \equiv u - 1 \pmod{k}$, $\{u, v\} \in E(\boldsymbol{g})$. Let $E \subset E(\boldsymbol{g})$ be the set with all the edges of this form and let $\boldsymbol{g}_E$ be the subgraph of $\boldsymbol{g}$ generated by E. Then, $(\deg_{\boldsymbol{g}_E}(u))_{u \in [2k]} = (\mathbf{3}_k; \mathbf{1}_k)$. This implies that $E = E(\boldsymbol{g})$, which means $\boldsymbol{g} = \boldsymbol{g}_E$.

An analogous argument shows that if $\{1+k, k\}$ is an edge in $\boldsymbol{g}$ (instead of $\{1+k, 2\}$) then the graph $\boldsymbol{g}$ is generated by the set of edges $E' = \{\{u, v\} : u \in [k], v \in [2k] \text{ and } v \equiv u + 1 \pmod{k}\}$. This shows that the only two graphs in $\mathcal{F}_{D\mathcal{C}_{2k,z}}(\boldsymbol{d}_k, \boldsymbol{c}_k; \mathbf{0}, \mathbf{1})$ are $\boldsymbol{g}_E$ and $\boldsymbol{g}_{E'}$. Since $E \cap E' = \{\{u, v\} : u, v \in [k] \text{ and } v - u \equiv 1 \pmod{k}\}$, we conclude that $||\boldsymbol{g}_E - \boldsymbol{g}_{E'}||_1 = |E| + |E'| - 2|E \cap E'| = 2k$.

Now we prove claims 1 and 2. Suppose $\boldsymbol{g} \in \mathcal{F}_{D\mathcal{C}_{2k,z}}(\boldsymbol{d}_k, \boldsymbol{c}_k; \mathbf{0}, \mathbf{1})$ and let $\{u+k, v\} \in E(\boldsymbol{g})$ where $u \in [k]$, $v \in [2k]$ are such that $v \equiv u + \epsilon \pmod{k}$ with $\epsilon \in \{1, -1\}$. Let $w, w' \in [k]$ such

that $w \equiv u+\epsilon \pmod{k}$ and $w' \equiv u-\epsilon \pmod{k}$. Since $\boldsymbol{c}_k(u,w') = 2$, $z^{-1}(w') = \{w', w'+k\}$ and $\deg_{\boldsymbol{g}}(u+k) = 1$, it follows that $\{u, w'+k\}$ and $\{u, w'\}$ are both edges in $E(\boldsymbol{g})$. This proves Claim 1. To prove Claim 2 let $\boldsymbol{g} \in \mathcal{F}_{I\mathcal{C}_{n,z}}(\boldsymbol{d}_k; \boldsymbol{c}_k)$ and suppose that $\{k+1, k+2, \ldots, 2k\}$ is not independent. Without losing generality assume that $\{k+1, k+2\}$ is an edge in $E(\boldsymbol{g})$. Claim 1 then implies that $\{1, k\}$ and $\{1, 2k\}$ are both edges in $E(\boldsymbol{g})$. Following an argument analogous to the proof of Claim 1 we can see that for every $u \in [k]$, $\{u, v\} \in E(\boldsymbol{g})$ for any $v \in [2k]$ such that $v \equiv u-1 \pmod{k}$. In particular, this means that $\{3, 2+k\} \in E(\boldsymbol{g})$ which would imply that $\deg_{\boldsymbol{g}}(2+k) \geq 2$. By the definition of $\boldsymbol{d}_k$, this is a contradiction. Therefore, the set $\{k+1, k+2, \ldots, 2k\}$ must be independent.

□

3.3. A Quadratic Gröbner Basis

The aim of this section is to show that $\{x^{\boldsymbol{g}^+} - x^{\boldsymbol{g}^-} : \boldsymbol{g} \in \mathcal{M}_{n,z}\}$ is in fact a Gröbner basis for $I_{I\mathcal{C}_{n,z}}$ with respect to a monomial order defined below. When the k-coloring z is constant, the statement follows directly from [40, Theorem 2.1]. As a matter of fact we will use this result, stated in Proposition 3.3.1 below, as the motivation to prove Theorem 1.4.14.

To prove the main result of this section we start by introducing the monomial order $\succ$ as follows. Let us identify the set $[n]$ with the vertices of a complete graph K_n embedded in the plane in a way that the vertices form a regular n-gon, labeled clockwise from 1 to n. We define the *weight* of the variable x_{uv} as the number of edges of K_n which do not meet the edge uv. For instance, if $n = 5$, then the variables $x_{12}, x_{23}, x_{34}, x_{45}, x_{15}$ have weight 3, and the variables $x_{13}, x_{24}, x_{35}, x_{14}, x_{25}$ have weight 1. In general, the weight of a monomial $\boldsymbol{x}^{\boldsymbol{\alpha}} := \prod_{uv} x_{uv}^{\alpha_{uv}}$ is the sum of the weights of the variables x_{uv} appearing in $\boldsymbol{x}^{\boldsymbol{\alpha}}$, with multiplicity. Let $\succ$ denote any monomial order that refines the partial order on monomials specified by these weights. Given any pair of non-intersecting edges $uv, u'v'$ of K_n such that uv', vu' intersect, we have from the definition of weights that $\text{in}_{\succ}(x_{uv}x_{u'v'} - x_{uv'}x_{u'v}) = x_{uv}x_{u'v'}$.

From Lemma 3.1.2 and the definition of the order $\succ$, it follows that

$$\{\text{in}_{\succ}(x^{\boldsymbol{g}^+} - x^{\boldsymbol{g}^-}) : \boldsymbol{g} \in \mathcal{M}_{n,z}\} = \{\, x_{uv}x_{u'v'} : uv, u'v' \text{ do not intersect in the embedding of } K_n \text{ in the plane and } \{z(u), z(v)\} \cap \{z(u'), z(v')\} \neq \emptyset\}. \tag{3.5}$$

PROPOSITION 3.3.1 ([40], Theorem 2.1). *The set of binomials* $\{x^{\boldsymbol{g}^+} - x^{\boldsymbol{g}^-} : \boldsymbol{g} \in \ker_{\mathbb{Z}} D_n, ||\boldsymbol{g}||_1 = 4\}$ *is a Gröbner basis for* I_{D_n} *with respect to* $\succ$.

PROPOSITION 3.3.2. *For any monomial walk* $\boldsymbol{g} \in \ker_{\mathbb{Z}} DC_{n,z}$ *there exists a pair of non-intersecting edges* $uv, u'v'$ *in the embedding of* K_n *in the plane such that* $x_{uv}x_{u'v'}$ *divides either* $x^{\boldsymbol{g}^+}$ *or* $x^{\boldsymbol{g}^-}$.

PROOF. Let $\boldsymbol{g} \in \ker_{\mathbb{Z}} DC_{n,z}$ and let $\mathcal{M}_n = \{\boldsymbol{\omega} \in \ker_{\mathbb{Z}} D_n : ||\boldsymbol{\omega}||_1 = 4\}$. Similarly to (3.5), from the definition of $\succ$ it follows that

$$\{\text{in}_{\succ}(x^{\boldsymbol{\omega}^+} - x^{\boldsymbol{\omega}^-}) : \boldsymbol{\omega} \in \mathcal{M}_n\} = \{\, x_{uv}x_{u'v'} : uv, u'v' \text{ do not intersect in the embedding of } K_n \text{ in the plane}\}.$$

As a consequence of the containment $\ker_{\mathbb{Z}} DC_{n,z} \subseteq \ker_{\mathbb{Z}} D_n$, it follows that $x^{\boldsymbol{g}^+} - x^{\boldsymbol{g}^-} \in I_{D_n}$. Furthermore, Proposition 3.3.1 implies that $\text{in}_{\succ}(I_{D_n}) = \langle\{x^{\boldsymbol{\omega}^+} - x^{\boldsymbol{\omega}^-} : \boldsymbol{\omega} \in \mathcal{M}_n\}\rangle$. Hence, there exists $\boldsymbol{\omega} \in \mathcal{M}_n$ such that $\text{in}_{\succ}(\boldsymbol{\omega}) = x_{uv}x_{u'v'}$ divides $\text{in}_{\succ}(x^{\boldsymbol{g}^+} - x^{\boldsymbol{g}^-})$. Since uv and $u'v'$ don't intersect in the embedding of K_n in the plane, the result follows. □

Proposition 3.3.9 below extends the result mentioned above and serves as a crucial step for proving Theorem 1.4.14. To establish the proof for Proposition 3.3.9, we will introduce the following notation and lemmas.

Let $z : [n] \to [k]$ and let $q, q' \in [k]$. We define the k-coloring $z_q^{q'} : [n] \to [k]$ as

$$z_q^{q'}(i) := \begin{cases} q & \text{if } z(i) = q', \\ z(i) & \text{otherwise.} \end{cases} \tag{3.6}$$

In other words, the k-coloring $z_q^{q'}$ is obtained from the k-coloring z by re-coloring all the q'-th colored vertices with the q-th color. Then, we have the following.

LEMMA 3.3.3. *For any* k*-coloring* z *of* $[n]$ *and* $q, q' \in [k]$ *we have* $\ker_{\mathbb{Z}} A_{n,z_q^{q'}} \subseteq \ker_{\mathbb{Z}} DC_{n,z}$.

PROOF. Let $\boldsymbol{g} \in \ker_{\mathbb{Z}} DC_{n,z}$ and $q, q' \in [k]$. $\boldsymbol{g}$ satisfies the degree-condition by assumption. Hence, all we need to prove is that $\boldsymbol{g}$ satisfies the color-balance condition with respect to $z_q^{q'}$. To do so, notice that for every i, j different from q',

$$c^{\pm}_{\boldsymbol{g}}(z^{q'}_{q}, i, j) = \begin{cases} c^{\pm}_{\boldsymbol{g}}(z, i, j) & \text{if } i, j \neq q, \\ c^{\pm}_{\boldsymbol{g}}(z, i, j) + c^{\pm}_{\boldsymbol{g}}(z, q', j) & \text{if } i = q, j \neq q, \\ c^{\pm}_{\boldsymbol{g}}(z, q, q) + c^{\pm}_{\boldsymbol{g}}(z, q', q') + c^{\pm}_{\boldsymbol{g}}(z, q, q') & \text{if } i = j = q. \end{cases}$$

When either $i = q'$ or $j = q'$, $c^{\pm}_{\boldsymbol{g}}(z^{q'}_{q}, i, j) = 0$.

□

Now let $\boldsymbol{g} \in \mathbb{Z}^{\binom{n}{2}}$, $w \in [n]$, and $z : [n] \to [k]$ with $z(w) = q$. We define the *contraction of* $\boldsymbol{g}$ *with respect to* w *and* z as the vector $\sigma_w(\boldsymbol{g}) \in \mathbb{Z}^{\binom{n}{2}}$ such that for every distinct $u, v \in [n]$,

$$\sigma_w(\boldsymbol{g})_{uv} := \begin{cases} g_{uv}, & \text{if } z(u) \neq q \text{ and } z(v) \neq q, \\ \sum_{u' \in z^{-1}(q)} g_{u'v}, & \text{if } u = w \text{ and } z(v) \neq q, \\ \sum_{v' \in z^{-1}(q)} g_{uv'}, & \text{if } v = w \text{ and } z(u) \neq q, \\ 0, & \text{otherwise .} \end{cases} \tag{3.7}$$

We call $\sigma_w(\boldsymbol{g})$ simply a contraction when w and z are clear from the context.

Remark 3.3.4. *Notice that whenever u or v belong to the set $z^{-1}(q)\backslash\{w\}$ we have $\sigma_w(\boldsymbol{g})_{uv} = 0$. In other words, $z^{-1}(q)\backslash\{w\}$ is an isolated set of vertices in $\sigma_w(\boldsymbol{g})$ when regarded as a graph. This implies that for any $v \in [n]$, $S \subseteq [n]\backslash\{v\}$ and $S' \subseteq z^{-1}(q)\backslash\{w\}$*

$$\sum_{u \in S} \sigma_w(\boldsymbol{g})_{uv} = \sum_{u \in S \backslash S'} \sigma_w(\boldsymbol{g})_{uv} = \sum_{u \in S \cup S'} \sigma_w(\boldsymbol{g})_{uv}.$$

Example 3.3.5. *Consider the monomial walk $\boldsymbol{g}$ from Example 3.1.1. The contraction $\sigma_1(\boldsymbol{g})$ is shown in Figure 3.4. Notice that the reduction $\sigma_3(\sigma_1(\boldsymbol{g}))$ returns a zero-vector, or in other words, an empty graph.*

Lemma 3.3.6. *For any k-coloring z of $[n]$, any monomial walk $\boldsymbol{g} \in \ker_{\mathbb{Z}} DC_{n,z}$ and $w \in [n]$, $\sigma_w(\boldsymbol{g})$ is also a monomial walk. In other words, the map σ_w satisfies $\sigma_w(\ker_{\mathbb{Z}} DC_{n,z}) \subseteq \ker_{\mathbb{Z}} DC_{n,z}$.*

Proof. Let $\boldsymbol{g} \in \ker_{\mathbb{Z}} DC_{n,z}$ and $w \in [n]$ with $q := z(w)$. We will show the following:

(1) $\sigma_w(\boldsymbol{g})$ satisfies the degree-balance condition.

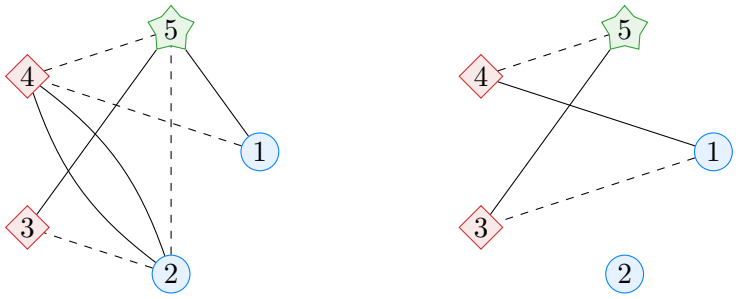


FIGURE 3.4. Monomial walk $\boldsymbol{g} = [1, 5, 2, 4, 5, 3, 2, 4]$ on the left and its contraction $\sigma_1(\boldsymbol{g}) = [1, 4, 5, 3]$ on the right.

Let $v \in [n]$. If $v \in z^{-1}(q)\backslash\{w\}$, Remark 3.3.4 implies that $\deg^{\pm}_{\sigma_w(\boldsymbol{g})}(v) = 0$, so let us assume $v \notin z^{-1}(q)\backslash\{w\}$ and consider the following two cases:

(i) $v \neq w$. In this case, by properly arranging summation indices and using Remark 3.3.4, one can see that

$$\begin{aligned}
\deg^+_{\sigma_w(\boldsymbol{g})}(v) - \deg^-_{\sigma_w(\boldsymbol{g})}(v) &= \sum_{u\neq v} \sigma_w(\boldsymbol{g})_{uv} = \sum_{\substack{u\neq v \\ u\notin z^{-1}(q)}} \sigma_w(\boldsymbol{g})_{uv} + \sum_{\substack{u\neq v \\ u\in z^{-1}(q)}} \sigma_w(\boldsymbol{g})_{uv} \\
&= \sum_{\substack{u\neq v \\ u\notin z^{-1}(q)}} g_{uv} + \sigma_w(\boldsymbol{g})_{wv} = \sum_{\substack{u\neq v \\ u\notin z^{-1}(q)}} g_{uv} + \sum_{u'\in z^{-1}(q)} g_{u'v} \\
&= \sum_{u\neq v} g_{uv} = \deg^+_{\boldsymbol{g}}(v) - \deg^-_{\boldsymbol{g}}(v) = 0.
\end{aligned}$$

(ii) $v = w$. Similarly to the previous case, by strategically rearranging summation indices and using Remark 3.3.4, we have

$$\begin{aligned}
\deg^+_{\sigma_w(\boldsymbol{g})}(v) - \deg^-_{\sigma_w(\boldsymbol{g})}(v) &= \sum_{u\neq w} \sigma_w(\boldsymbol{g})_{uw} = \sum_{u\notin z^{-1}(q)} \sigma_w(\boldsymbol{g})_{uw} \\
&= \sum_{u\notin z^{-1}(q)} \sum_{v'\in z^{-1}(q)} g_{uv'} = \sum_{\substack{i\in[k] \\ i\neq q}} \Big(\sum_{u\in z^{-1}(i)} \sum_{v'\in z^{-1}(q)} g_{uv'} \Big) \\
&= \sum_{\substack{i\in[k] \\ i\neq q}} \big(c^+_{\boldsymbol{g}}(i,q) - c^-_{\boldsymbol{g}}(i,q) \big) = 0.
\end{aligned}$$

Where the last equality holds because $\boldsymbol{g}$ satisfies the color-balance conditions.

(2) $\sigma_w(\boldsymbol{g})$ satisfies the color-balance condition.

Let $1 \leq i \leq j \leq k$. It follows from Equation (3.2) and (3.7) that if $i, j \neq q$, $c^{\pm}_{\sigma_w(\boldsymbol{g})}(i,j) = c^{\pm}_{\boldsymbol{g}}(i,j)$ and if $i = j = q$ then $c^{\pm}_{\sigma_w(\boldsymbol{g})}(i,j) = 0$. In either of these two cases we have $c^{+}_{\sigma_w(\boldsymbol{g})}(i,j) = c^{-}_{\sigma_w(\boldsymbol{g})}(i,j)$. Now, suppose that $i \neq q$ and $j = q$. In this case we have

$$c^{+}_{\sigma_w(\boldsymbol{g})}(i,j) - c^{-}_{\sigma_w(\boldsymbol{g})}(i,j) = \sum_{u \in z^{-1}(i)} \sigma_w(\boldsymbol{g})_{uw} = \sum_{u \in z^{-1}(i)} \sum_{v' \in z^{-1}(q)} g_{uv'}$$
$$= c^{+}_{\boldsymbol{g}}(i,q) - c^{-}_{\boldsymbol{g}}(i,q) = 0.$$

The case $i = q, j \neq q$ is analogous to the latter case.

□

For the rest of the section we will assume that the k-coloring z is non-decreasing which will be useful thanks to the following.

REMARK 3.3.7. *Consider the embedding of K_n in the plane and suppose the k-coloring $z : [n] \to [k]$ is non-decreasing. If $uv, u'v'$ are two non-intersecting edges with $z(v') \notin \{z(u), z(u'), z(v)\}$ then for any vertex w such that $z(w) = z(v')$, we have that $uv, u'w$ are non-intersecting edges in the embedding of K_n.*

LEMMA 3.3.8. *Let z be a k-coloring of $[n]$ with $k \geq 2$ and $\boldsymbol{g} \in \ker_{\mathbb{Z}} D\mathcal{C}_{n,z}$ be a monomial walk. Let $w \in [n]$ and uv edge in $\sigma_w(\boldsymbol{g})$. Then,*

(i) there exists a vertex $v_0 \in [n]$ such that uv_0 is an edge in $E(\boldsymbol{g})$ with same sign as uv and $z(v_0) = z(v)$,

(ii) if $u'v'$ is and edge in $E(\sigma_w(\boldsymbol{g}))$ with $u', v' \neq w$ such that $uv, u'v'$ do not intersect in the embedding of K_n in the plane, then uv_0 and $u'v'$ do not intersect in the embedding of K_n either.

PROOF. (i) Let $q = z(w)$ and uv an edge of $\sigma_w(\boldsymbol{g})$. By definition of $\sigma_w(\boldsymbol{g})$ we have that every vertex in $z^{-1}(q)\backslash\{w\}$ is isolated. Hence we have two options: either $u, v \neq w$ or $v = w$ (or $u = w$). If $u, v \neq w$, uv is also an edge of $\boldsymbol{g}$ in which case we can set $v_0 = v$. Now, assume that $v = w$ (the case $u = w$ is completely analogous).

Given that uw is an edge in $\sigma_w(\boldsymbol{g})$, we have that $\sigma_w(\boldsymbol{g})_{uw} \neq 0$. Assume without loosing generatity that $\sigma_w(\boldsymbol{g})_{uw} > 0$ (i.e., uv is positive). By definition, $\sigma_w(\boldsymbol{g})_{uw} = \sum_{v_0 \in z^{-1}(q)} g_{uv_0}$, which implies that $g_{uv_0} > 0$ for some $v_0 \in z^{-1}(q)$. This means that uv_0 is

a positive edge in $\boldsymbol{g}$ with $z(v_0) = q = z(w)$. We can apply a similar argument for when $\sigma_w(\boldsymbol{g})_{uw} < 0$.

(ii) Given that $u', v' \neq w$, it follows from the construction of $\sigma_w(\boldsymbol{g})$ that $u'v'$ is an edge of $\boldsymbol{g}$. If $v \neq w \Rightarrow uv_0 = uv$ so the result follows trivially. If $v = w$, then $z(v) \notin \{z(u), z(u'), z(v')\}$ so Remark 3.3.7 implies that uv_0 and $u'v'$ are non-intersecting edges in the embedding of K_n in the plane.

□

PROPOSITION 3.3.9. *For any monomial walk $\boldsymbol{g} \in \ker_{\mathbb{Z}} D\mathcal{C}_{n,z}$ there is a pair of non-intersecting edges $uv, u'v'$ in the embedding of K_n in the plane such that $z(u) = z(u')$ and $x_{uv}x_{u'v'}$ divides either $x^{\boldsymbol{g}^+}$ or $x^{\boldsymbol{g}^-}$.*

PROOF. Let $\boldsymbol{g} \in \ker_{\mathbb{Z}} D\mathcal{C}_{n,z}$ be a monomial walk. Notice that a monomial $x_{uv}x_{u'v'}$ divides $x^{\boldsymbol{g}^+}$ or $x^{\boldsymbol{g}^-}$ if and only if $uv, u'v'$ is pair of edges in $E(\boldsymbol{g})$ with same sign. Let $z : [n] \to [k]$ be a k-coloring assumed to be non-decreasing and for convenience suppose $z([n]) = \{1, \ldots, \kappa\}$ for some $\kappa \in \mathbb{Z}_+$. When $\kappa = 1$ (i.e., z is constant) the result follows from Proposition 3.3.2. Let us divide the proof for $\kappa \geq 2$ in the following two cases:

(a) There exists a monochromatic edge $xy \in E(\boldsymbol{g})$ with $z(x) = z(y) = \iota$.

For every $i \in [\kappa]\backslash\{\iota\}$, let $u_i := \min\{u : z(u) = i\}$, $\widetilde{\boldsymbol{g}} := \sigma_{u_\kappa}(\sigma_{u_{\kappa-1}}(\cdots \sigma_{u_2}(m)\cdots))$ and

$$w \in \operatorname{argmax}_{v\in[n]:\, z(v)\neq\iota} \deg^+_{\widetilde{\boldsymbol{g}}}(v).$$

From Proposition 3.3.2, there exists a pair of edges $uv, u'v'$ in $\sigma_w(\widetilde{\boldsymbol{g}})$ such that $uv, u'v'$ do not intersect in the embedding of K_n in the plane. Since w is the only (potentially) non-isolated vertex with respect to $\sigma_w(\widetilde{\boldsymbol{g}})$ with $z(w) \neq \iota$, we can assume w.l.o.g. that $z(u) = z(v) = z(u') = \iota$. By Lemma 3.3.8(i) there exists a vertex $v_0' \in [n]$ such that $u'v_0' \in E(\widetilde{\boldsymbol{g}})$ has same sign as $u'v'$ and $z(v_0') = z(v')$. Moreover, since $uv, u'v' \in E(\sigma_w(\widetilde{\boldsymbol{g}}))$ do not intersect in the embedding of K_n and $u, v \neq w$, it follows from Lemma 3.3.8(ii) that $uv, u'v_0'$ is also a pair of non-intersecting edges in the embedding of K_n in the plane. Notice that $uv, u'v_0'$ have same sign. Then, after applying Lemma 3.3.8 repeatedly to $\widetilde{\boldsymbol{g}}$, we will get a vertex $v_0'' \in [n]$ such that the edges $uv, u'v_0'' \in E(\boldsymbol{g})$ have the same sign, and $uv, u'v_0''$ do not intersect in the embedding of K_n in the plane. Since we assumed that

$z(u) = z(u')$, this concludes the proof of Proposition 3.3.9 under the assumptions made for this case.

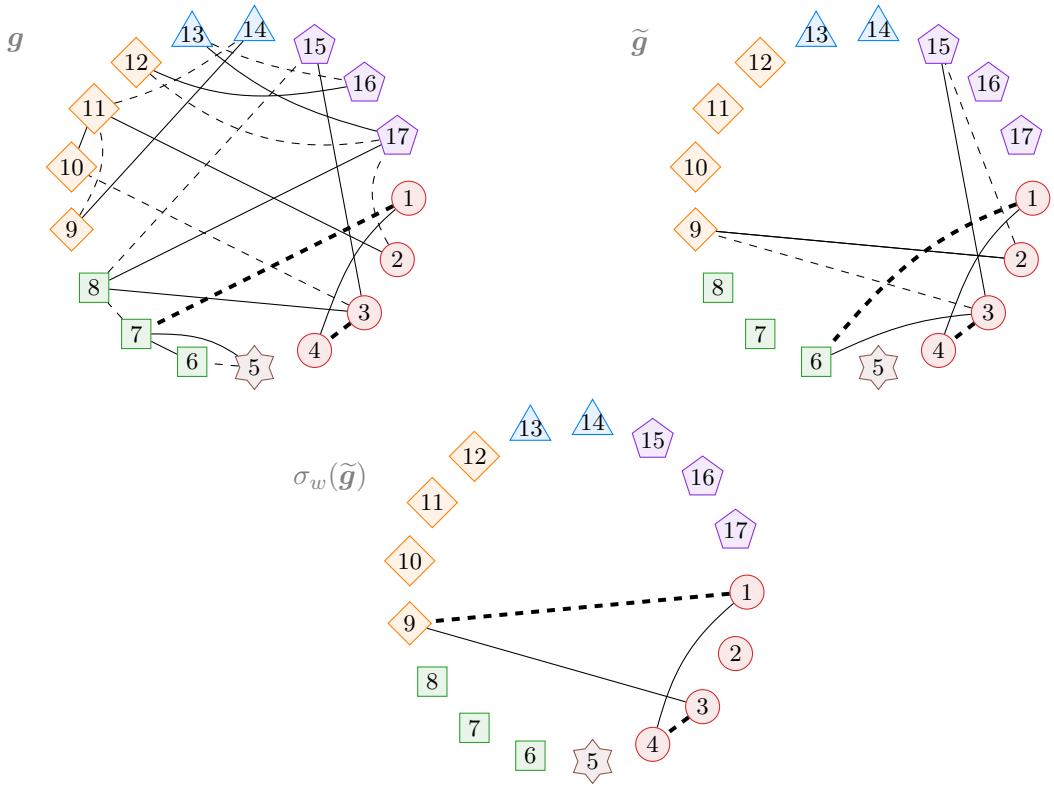


FIGURE 3.5. Illustration of Proposition 3.3.9, case (a): Recovering the pair of non-intersecting edges $\{3,4\},\{1,7\}$ in $\boldsymbol{g}$ from $\{3,4\},\{1,w\}$ in $E(\sigma_w(\widetilde{\boldsymbol{g}}))$.

(b) There are no monochromatic edges in the monomial walk $\boldsymbol{g}$.

We will prove Proposition 3.3.9 for this case by induction on κ. First, let us observe that Proposition 3.3.2 guarantees the existence of a pair of non-intersecting edges edges $uv, u'v' \in E(\boldsymbol{g})$, both with the same sign and such that do not intersect in the embedding of K_n in the plane. Now consider the following cases:

(i) Let $\kappa \leq 3$. By the assumption at the beginning of (b) we have that $z(u) \neq z(v), z(u') \neq z(v')$. From the Pigeonhole principle either $z(u) = z(u')$ or $z(u) = z(v')$. This proves our base case.

(ii) Let $\kappa \geq 4$. Assume that Proposition 3.3.9 holds for any instance of case (b) for which the size of the k-coloring's range is smaller than κ.

Let us first prove that there exists $\nu \in [n]$ such that $z(u), z(v) \notin \{\nu, \nu+1\}$ ($\nu+1 = 1$ if $\nu = n$): When $\kappa > 4$ this follows from the Pigeonhole principle. When $\kappa = 4$ we can use that z is non-decreasing to see that any pair of edges with vertex colors $1, 3$ intersects with any pair of edges with vertex colors $2, 4$. In such a case we can assume w.l.o.g that $z(u) = 1, z(v) = 2$ and set $i = 3$.

The existence of ν guarantees that for any $w \in z^{-1}(\{q, q'\})$ the contraction $\sigma_w(\boldsymbol{g})$ is non-empty. Now, let $\nu' = \nu + 1$, $q = z(\nu)$ and $q' = z(\nu')$. Let $z_q^{q'}$ be the k-coloring of $[n]$ as defined in 3.6 and notice that by Lemma 3.3.3 $\boldsymbol{g}$ is a monomial walk with respect to $z_q^{q'}$, i.e., $\boldsymbol{g} \in \ker_{\mathbb{Z}} A_{n, z_q^{q'}}$. Now, pick $w \in [n]$ such that $z_q^{q'}(w) = q$ and notice that by previous observation $\sigma_w(\boldsymbol{g})$ is non-empty. Then, either by case (a) above or by the inductive step applied to $\sigma_w(\boldsymbol{g})$ with respect to $z_q^{q'}$, there exists a pair of edges $xy, x'y' \in E(\sigma_w(\boldsymbol{g}))$ such that $z_q^{q'}(x) = z_q^{q'}(x')$ and $xy, x'y'$ do not intersect in the embedding of K_n in the plane. Since w is the only (possibly) non-isolated vertex with $z_q^{q'}(w) = q$, it follows that $x, x' \neq w$ and by definition of $z_q^{q'}$ it follows that $z(x) = z_q^{q'}(x) = z_q^{q'}(x') = z(x')$.

Notice that at least one of the vertices y, y' is different from w. Without losing generality assume $y' \neq w$. Then, by Lemma 3.3.8 there exists $y_0 \in [n]$ such that xy_0 has the same sign as xy and $xy_0, x'y'$ do not intersect in the embedding of K_n in the plane. Moreover, $z(x) = z(x')$. This finishes the prove of this case.

□

Proof of Theorem 1.4.14. Let $\succ$ be any monomial order that refines that refines the partial order specified by weights just as in the beginning of the current section. Let $\mathrm{Bin}_{\mathcal{M}_{n,z}} := \{x^{\boldsymbol{g}^+} - x^{\boldsymbol{g}^-} : \boldsymbol{g} \in \mathcal{M}_{n,z}\}$. By [104, Corollary 4.4], the set of binomials $\{x^{\boldsymbol{g}^+} - x^{\boldsymbol{g}^-} : \boldsymbol{g} \in \ker_{\mathbb{Z}} D\mathcal{C}_{n,z}\}$ contains every Gröbner basis (with respect to any monomial order) of $I_{D\mathcal{C}_{n,z}}$. Hence, to show $\mathrm{Bin}_{\mathcal{M}_{n,z}}$ is a Gröbner basis, it is enough to prove that the leading term of any binomial $x^{\boldsymbol{g}^+} - x^{\boldsymbol{g}^-} \in I_{D\mathcal{C}_{n,z}}$ is divisible by a monomial $x_{uv}x_{u'v'}$ where $uv, u'v' \in E(\boldsymbol{g})$ do not intersect in the embedding of K_n in the plane and $\{z(u), z(v)\} \cap \{z(u'), z(v')\} \neq \emptyset$. Assume that $f = x^{\boldsymbol{g}^+} - x^{\boldsymbol{g}^-} \in I_{D\mathcal{C}_{n,z}}$ with $\mathrm{in}_\succ(f) = x^{\boldsymbol{g}^+}$, is a minimal counterexample in the sense that f has minimal weight. Here the weight of a binomial is the sum of the weights of its two monomials. This means that every pair of positive edges $uv, u'v' \in E(\boldsymbol{g})$ with $z(u) = z(u')$ intersect in the embedding of K_n in the plane.

Furthermore, we can assume that every pair of negative edges $uv, u'v'$ with $z(u) = z(u')$ intersect in the embedding of K_n in the plane as well. Otherwise, we can reduce $x^{\boldsymbol{g}^-}$ modulo $\mathrm{Bin}_{\mathcal{M}_{n,z}}$ to get a counterexample of smaller weight. On the other hand, the existence of $\boldsymbol{g}$ would contradict Proposition 3.3.9. Hence no such binomial $x^{\boldsymbol{g}^+} - x^{\boldsymbol{g}^-} \in I_{D\mathcal{C}_{n,z}}$ could exist. Therefore, $\mathrm{Bin}_{\mathcal{M}_{n,z}}$ is a Gröbner basis for $I_{D\mathcal{C}_{n,z}}$ with respect to $\succ$.

□

3.4. Future Directions

The combinatorial description of the Gröbner basis in Theorem 1.4.14 has direct implications for the combinatorics of the polytope $\mathcal{P}_{D\mathcal{C}_{n,z}} := \mathrm{conv}(a_{uv} : 1 \leq u < v \leq n)$, defined as the convex hull of the column vectors a_{uv} of $D\mathcal{C}_{n,z}$. More specifically, as a consequence of [104, Theorem 8.3], the Gröbner basis of $I_{D\mathcal{C}_{n,z}}$ described in Theorem 1.4.14 induces a *unimodular* regular triangulation $\mathcal{T}_\succ$ of $\mathcal{P}_{D\mathcal{C}_{n,z}}$. Following ideas analogous to [40, Remarks 2.5], this triangulation enables the computation of the Ehrhart polynomial (which, in this case, equals the Hilbert polynomial) of $\mathcal{P}_{D\mathcal{C}_{n,z}}$. For example, for $k = 2$ and any k-coloring $z : [n] \to [2]$ with $n_i = |z^{-1}(i)|$, the Hilbert polynomial of $I_{D\mathcal{C}_{n,z}}$ is given by

$$H_{D\mathcal{C}_{n,z}}(r) = \mathrm{card}\left(r \cdot \mathcal{P}_{D\mathcal{C}_{n,z}} \cap \mathbb{Z}^{n+3}\right) = \sum_{\tau \in \mathcal{W}_{r,3}} a_\tau,$$

where $\mathcal{W}_{r,3} = \{\tau \in \mathbb{N}^3 : \sum_{1 \leq i \leq j \leq 2} \tau_{i,j} = r\}$ is the set of *weak 3-partitions* of r, and

$$\begin{aligned} a_\tau = &\binom{n_1 + 2\tau_{1,1} + 2\tau_{1,2}}{n_1 - 1}\binom{n_1 + 2\tau_{1,1} + 2\tau_{1,2}}{n_1 - 1} - n_1\binom{n_1 - 2 + \tau_{1,1} + \tau_{1,2}}{n_1 - 1}\binom{n_1 + 2\tau_{1,1} + 2\tau_{1,2}}{n_1 - 1} \\ &- n_2\binom{n_1 + 2\tau_{1,1} + 2\tau_{1,2}}{n_1 - 1}\binom{n_2 - 2 + \tau_{2,2}}{n_2 - 1} + n_1 n_2\binom{n_1 - 2 + \tau_{1,1} + \tau_{1,2}}{n_1 - 1}\binom{n_2 - 2 + \tau_{2,2}}{n_2 - 1} \end{aligned}$$

for every $\tau \in \mathcal{W}_{r,3}$. Similar formulas can be derived for $k > 2$.

When the k-coloring z is constant, $\mathcal{P}_{D\mathcal{C}_{n,z}}$ is linearly isomorphic to the *second hypersimplex* $\Delta_n(2)$. In this case, the triangulation $\mathcal{T}_\succ$ has been thoroughly described in [40]. However, a general combinatorial understanding of $\mathcal{P}_{D\mathcal{C}_{n,z}}$ and the induced triangulation $\mathcal{T}_\succ$ remains an open problem.

Additionally, previous research has focused on the study of the *degree sequence polytope*, defined as the convex hull

$$\mathcal{D}_n := \mathrm{conv}(d(\boldsymbol{g}) : \boldsymbol{g} \in \mathcal{G}_n),$$

where $\mathcal{G}_n$ is the set of simple graphs with vertex set $[n]$. For more details, see [87, 91, 102] and references therein. Two key points of interest regarding this polytope include its hyperplane representation, which can be used to recover the famous Erdős-Gallai inequalities characterizing degree sequences of simple graphs, and the vertex description of $\mathcal{D}_n$. In fact,

$$\text{Vert}(\mathcal{D}_n) = \{\boldsymbol{g} \in \mathcal{G}_n : |\mathcal{F}(d(\boldsymbol{g}); \mathbf{0}, \mathbf{1})| = 1\}.$$

The graphs constituting the vertices of $\mathcal{D}_n$ are known as *threshold graphs*, which have multiple combinatorial characterizations [87, Theorem 1.2.4]. In particular, a graph $\boldsymbol{g} \in \mathcal{G}_n$ is a threshold graph if and only if there is no $\boldsymbol{m} \in \mathcal{M}_n$ such that $m+\boldsymbol{g} \in \mathcal{G}_n$. This establishes a natural connection between threshold graphs and binary Markov bases for $\mathcal{D}_n$.

A natural extension of this idea is to provide a full hyperplane and vertex representation of the *degree-color sequence polytope*, defined as

$$\mathcal{DC}_{n,z} := \text{conv}\left((d(\boldsymbol{g}), c(\boldsymbol{g})) : \boldsymbol{g} \in \mathcal{G}_n\right),$$

given a fixed coloring $z : [n] \to [k]$. In particular, the H-representation of $\mathcal{DC}_{n,z}$ would be useful for characterizing sequences $(\boldsymbol{d}, \boldsymbol{c}) \in \mathbb{N}^n \oplus \mathbb{N}^{\binom{k+1}{2}}$ that correspond to the degree-color sequence of a simple graph under a given coloring z.

REMARK 3.4.1. *Given a coloring $z : [n] \to [k]$, there exists an injective map from $\mathcal{G}_n$ to the set of 3-regular graphs with vertex set $\{1, \ldots, n\} \sqcup \{(i, j) : 1 \leq i \leq j \leq k\}$, which sends $\boldsymbol{g} \in \mathcal{G}_n$ to the hypergraph H with edge set*

$$E(H) = \{\{u, v, (z(u), z(v))\} : \{u, v\} \in E(\boldsymbol{g})\}.$$

This implies that degree-color sequences can be interpreted as degree sequences of a family of 3-regular hypergraphs. In general, [42, 43] showed that, for fixed $r \geq 3$, determining whether a sequence of non-negative integers corresponds to the degree sequence of an r-regular hypergraph is NP-hard.

CHAPTER 4

Markov Bases for a Labeled Stochastic Block Model

In this chapter, we provide proofs for the contributions presented in Section 1.5.

Theorem 1.5.2 describes the Graver basis of the design matrix of Labeled Stochastic Block Models, which includes the classic SBM as a special case. While the moves described in this theorem allow us to connect fibers of the form $\mathcal{F}(A_{\mathrm{LSBM}(z,\ell)}; \boldsymbol{l}, \boldsymbol{L})$ for any $\boldsymbol{l}, \boldsymbol{L} \in \mathbb{N}^{\ell\binom{n}{2}}$, these moves are insufficient to connect fibers when natural constraints are imposed on the space of graphs. As discussed in Subsection 1.5.1, Theorem 1.5.3 provides a set of moves applicable in scenarios with different natural constraints on the space of graphs.

Finally, Proposition 1.5.6 establishes that when z is unknown in the Labeled SBM model, using a consistent block assignment estimator $\hat{z}$ results in a consistent plug-in p-value.

4.1. A Simple Graver basis description

Before presenting the proof of Theorem 1.5.2, we introduce some notions and a key result. We say that a 0/1 matrix A satisfies the *consecutive 1's condition* if there exists a permutation matrix P such that the 1's in each row of AP appear in consecutive positions. A 0/1 matrix A is said to be *totally unimodular* if every square submatrix has a determinant of ± 1 or 0. It follows from known results on unimodularity that if A satisfies the consecutive 1's property, then A is totally unimodular (see [68]).

Recalling the definition of a circuit from Section 1.3, the circuits of A correspond to the subset of $\ker_{\mathbb{Z}} A$ with minimal support.

Proposition 4.1.1 ([104], Proposition 8.11). *If A is a totally unimodular 0/1 matrix, then the set of circuits $C(A)$ is equal to the Graver basis $Gr(A)$.*

The proof of the following theorem relies on this proposition.

PROOF OF THEOREM 1.5.2. By definition, the design matrix of the Labeled SBM with block assignment z and ℓ interaction types has the form

$$A_{\mathrm{LSBM}(z,\ell)} = I_{\ell\times\ell} \otimes A_{\mathrm{SBM}(z)} := \underbrace{\begin{pmatrix} A_{\mathrm{SBM}(z)} & 0 & \cdots & 0 \\ 0 & A_{\mathrm{SBM}(z)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_{\mathrm{SBM}(z)} \end{pmatrix}}_{\ell \text{ column blocks}}.$$

Moreover, $A_{\mathrm{SBM}(z)}$ satisfies the consecutive 1's condition: its columns correspond to pairs of nodes, which can be reordered based on the pairs of blocks they connect according to the block assignment z. This implies that $A_{\mathrm{LSBM}(z,\ell)}$ itself has the consecutive 1's property and is therefore totally unimodular.

Furthermore, it is clear that the set

$$\begin{aligned}\mathcal{M}_{\mathrm{LSBM}(z,\ell)} :&= \{\boldsymbol{e}_{uv}^{(l)} - \boldsymbol{e}_{u'v'}^{(l)} \mid l \in [\ell],\ z(u) = z(u'),\ z(v) = z(v')\} \\ &= \{\boldsymbol{g} \in \ker_{\mathbb{Z}} A_{\mathrm{LSBM}(z,\ell)} : ||\boldsymbol{g}||_1 = 2\}\end{aligned}$$

corresponds to the elements in $A_{\mathrm{LSBM}(z,\ell)}$ of minimal support. In other words, $C(A_{\mathrm{LSBM}(z,\ell)}) = \mathcal{M}_{\mathrm{LSBM}(z,\ell)}$. As a consequence of Proposition 4.1.1, it follows that $\mathcal{M}_{\mathrm{LSBM}(z,\ell)}$ is the Graver basis of $A_{\mathrm{LSBM}(z,\ell)}$. □

4.2. Connecting Restricted Fibers and Consistency of the Plug-in p-value

PROOF OF THEOREM 1.5.3. Let z be a fixed block assignment and

$$\mathcal{F} = \mathcal{F}(A_{\mathrm{LSBM}(z,\ell)}, \boldsymbol{b}) \cap \{\boldsymbol{g} = (\boldsymbol{g}_{uv})_{u<v} \in \mathbb{N}^{\ell\binom{n}{2}} : ||\boldsymbol{g}_{uv}||_1 = N_{uv}\},$$

where $\boldsymbol{b} \in \mathbb{N}A_{\mathrm{LSBM}(z,\ell)}$ and $\boldsymbol{N} = (N_{uv})_{u<v} \in \mathbb{N}^{\binom{n}{2}}$.

Let $\boldsymbol{f} = (f_{uv}^{(l)} : u < v, l \in [\ell]), \boldsymbol{g} = (g_{uv}^{(l)} : u < v, l \in [\ell]) \in \mathcal{F}$ be different labeled graphs with the same sufficient statistic. Assume without losing generality that $g_{uv}^{(l)} > f_{uv}^{(l)}$ where $z(u) = i, z(v) = j$ and $l \in [\ell]$. Since $\sum_{l'=1}^{\ell} g_{uv}^{(l')} = N_{uv} = \sum_{l'=1}^{\ell} f_{uv}^{(l')}$ it follows that there is a $l' \in [\ell]\backslash\{l\}$ such that

$g_{uv}^{(l')} < f_{uv}^{(l')}$. Furthermore, since $A_{\mathrm{LSBM}(z,\ell)}\boldsymbol{f} = A_{\mathrm{LSBM}(z,\ell)}\boldsymbol{g}$, we have

$$\sum_{u' \in B_i, v' \in B_j} g_{u'v'}^{(l)} = \sum_{u' \in B_i, v' \in B_j} f_{u'v'}^{(l)},$$

meaning that there exists $u' \in B_i, v' \in B_j$ with $f_{u'v'}^{(l)} > g_{u'v'}^{(l)}$. Let $\boldsymbol{m} = \boldsymbol{e}_{uv}^{(l)} + \boldsymbol{e}_{u'v'}^{(l')} - \boldsymbol{e}_{uv}^{(l')} - \boldsymbol{e}_{u'v'}{}^{(l)} \in \mathbb{Z}^{L\binom{n}{2}}$ and observe that $\boldsymbol{f} + \boldsymbol{m} \in \mathbb{N}^{\ell\binom{n}{2}}$, and $\sum_{l=1}^{\ell} (\boldsymbol{f} + \boldsymbol{m})_{uv}^{(l)} = \sum_{l=1}^{\ell} f_{uv}^{(l)} = N_{uv}$. In other words, $||\boldsymbol{f} + \boldsymbol{m}|| = N_{uv}$. Furthermore, we have

$$||(\boldsymbol{f} + \boldsymbol{m}) - \boldsymbol{g}||_1 = \begin{cases} ||\boldsymbol{f} - \boldsymbol{g}||_1 - 4, & \text{if } g_{u'v'}^{(l')} > f_{u'v'}^{(l')} \\ ||\boldsymbol{f} - \boldsymbol{g}||_1 - 2, & \text{otherwise.} \end{cases}$$

By an inductive argument this shows that the set $\widetilde{\mathcal{M}}_{\mathrm{LSBM}(z,\ell)}$ described in the statement of the theorem connects $\mathcal{F}$. Since $\mathcal{F}$ was an arbitrarily picked, the statement of the theorem follows. □

Proof of Proposition 1.5.6. Let $\boldsymbol{G}^{(n)} \sim \mathrm{LSBM}(z^{(n)}, \ell)$ for every n and a fixed $\ell \in \mathbb{N}$. Then

$$\begin{aligned} \mathbb{P}(p(z^{(n)}, \boldsymbol{G}^{(n)}) = p(\hat{z}^{(n)}, G^{(n)})) & \\ &\geq P(p(z^{(n)}, \boldsymbol{G}^{(n)}) = p(\hat{z}^{(n)}, \boldsymbol{G}^{(n)}) \mid A(z^{(n)} = \hat{z}^{(n)}) = 1)\mathbb{P}(A(z^{(n)} = \hat{z}^{(n)}) = 1) \\ &= \mathbb{P}(A(z^{(n)} = \hat{z}^{(n)}) = 1). \end{aligned}$$

Where the last equality follows from the definition of the plug-in p-value, the fact that $T_z(\boldsymbol{g}) = T_z(\boldsymbol{g}') \iff T_{\sigma \cdot z}(\boldsymbol{g}) = T_{\sigma z}(\boldsymbol{g}')$ for any $\sigma \in S_k$, and the property that $\mathrm{GoF}_{\tilde{z}}(\boldsymbol{g}) = \mathrm{GoF}_{\sigma \cdot \tilde{z}}(\boldsymbol{g})$ for any $\tilde{z} \in [k]^n$ and $\sigma \in S_k$. Since $\hat{z}$ is a strongly consistent estimator, it follows that $\lim_{n\to\infty} \mathbb{P}(p(z^{(n)}, \boldsymbol{G}^{(n)}) = p(\hat{z}^{(n)}, \boldsymbol{G}^{(n)})) = 1$. □

4.3. Experimental Results and Further Questions

We illustrate the performance of the goodness-of-fit test described at the end of Section 1.5.1 for scenario 3) in the frequentist setting with the following experiment. We generated 150 graphs, each with 70 nodes, from the stochastic block model (SBM) where $z \sim \mathrm{Multinomial}(\boldsymbol{\pi})$ and $G_{uv} \sim \mathrm{Poisson}(\theta_{z(u)z(v)})$ for fixed $\boldsymbol{\pi}$ and $\boldsymbol{\theta} = (\theta_{ij} : 1 \leq i \leq j \leq 6)$. We tested the null hypothesis that an SBM with k blocks, for $k = 3, \ldots, 8$, fits the synthetic data and computed the proportion of times the test rejected the null hypothesis using an approximation of the p-value from Equation 1.12 at a

nominal level of 0.05. To approximate the p-value, we implemented Algorithm 2 from [79] with the Markov basis described in Theorem 1.5.2 and the parameter estimation algorithm from [88]. The results, shown in Table 4.1, align with expectations. As anticipated, when the synthetic networks are generated from an SBM with an underspecified number of blocks, the test rejects the null hypothesis more than 99% of the time. However, when the synthetic networks are generated from an SBM with six or more blocks, the test fails to reject the null hypothesis in most cases.

Number of blocks (k)	**Power**
3	1
4	1
5	0.99
6	0.07
7	0.02
8	0

TABLE 4.1. Power calculations for the SBM(k) with $k = 3, \ldots, 8$ and $n = 70$ nodes.

Based on multiple simulations similar to the one described above, we believe it is worthwhile to explore the following question.

QUESTION 4.3.1. *Let $k \in \mathbb{N}$ and let $z : [n] \to [k]$ be a block assignment. Let $\boldsymbol{G} \sim LSBM(z, \ell)$, and for each $q \in \{2, \ldots, n\}$, let $\hat{z}^{(q)} = \hat{z}(q, \boldsymbol{G})$ be an estimator recovered from q and $\boldsymbol{G}$ using algorithms such as those in [88, 108]. Define $p^{(q)}(\boldsymbol{G}) = p(\boldsymbol{g}, \hat{z}^{(q)})$ as the plug-in p-value from Equation 1.9, computed using the chi-square statistic from Equation 1.11. Under what conditions does $\mathbb{P}(p^{(2)}(\boldsymbol{G}) \leq \cdots \leq p^{(n)}(\boldsymbol{G}))$ approach 1?*

Assuming a positive answer to this question, the goodness-of-fit test can be used to determine the number of blocks in an observed network by applying the test sequentially. As an example, we analyzed two undirected, valued networks, where nodes represent parasitic fungal species ($n = 154$) and tree species ($n = 51$), respectively. In these cases, the edge counts g_{uv} correspond to the number of shared host species and the number of shared parasitic species, respectively. The data was obtained from the `sbm` package in `R` [26].

After sequentially applying our test to assess whether the data fits a Poisson-SBM, we obtained the results presented in Tables 4.2 and 4.3.

These results suggest that the tree species network and the fungal species network are better modeled by a Poisson-SBM with $k = 10$ and $k = 22$ blocks, respectively.

Number of Blocks	3–7	8–9	10	11	12	13	14	15
***p*-value**	0	.01	.19	.68	.93	.98	1	1

TABLE 4.2. Goodness-of-fit results for the tree species network.

Number of Blocks	3–17	18–21	22
***p*-value**	0	.01	.07

TABLE 4.3. Goodness-of-fit results for the fungal species network.

This differs from the model selection approach in [88], which uses the Integrated Classification Likelihood (ICL) criterion and suggests modeling the networks with 7 blocks for tree species and 9 blocks for fungal species. Further investigation is needed to understand the differences between the goodness-of-fit test we propose and the ICL criterion, as well as their relative strengths and limitations.

Bibliography

[1] 4ti2 team, *4ti2—a software package for algebraic, geometric and combinatorial problems on linear spaces*. Available at https://github.com/4ti2/4ti2.

[2] E. Abbe, *Community detection and stochastic block models: Recent developments*, Journal of Machine Learning Research, 18 (2018), pp. 1–86.

[3] M. Abdullah, C. Cooper, and A. Frieze, *Cover time of a random graph with given degree sequence*, in 21st International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods in the Analysis of Algorithms (AofA'10), vol. AM of Discrete Math. Theor. Comput. Sci. Proc., Assoc. Discrete Math. Theor. Comput. Sci., Nancy, 2010, pp. 1–19.

[4] F. Almendra-Hernández, J. A. De Loera, and S. Petrović, *Irreducible Markov chains on spaces of graphs with fixed degree-color sequences*, (2024). Available at arXiv:2402.09568.

[5] ———, *Markov bases: A 25 year update*, Journal of the American Statistical Association, 119 (2024), pp. 1671–1686.

[6] A. A. Amini, A. Chen, P. J. Bickel, and E. Levina, *Pseudo-likelihood methods for community detection in large sparse networks*, The Annals of Statistics, 41 (2013), pp. 2097–2122.

[7] S. Aoki, H. Hara, and A. Takemura, *Markov Bases in Algebraic Statistics*, Springer Series in Statistics, Springer New York, 2012.

[8] S. Aoki, T. Hibi, H. Ohsugi, and A. Takemura, *Markov basis and Gröbner basis of Segre-Veronese configuration for testing independence in group-wise selections*, Ann. Inst. Statist. Math., 62 (2010), pp. 299–321.

[9] S. Aoki and A. Takemura, *The list of indispensable moves of the unique minimal Markov basis for* $3 \times 4 \times k$ *and* $4 \times 4 \times 4$ *contingency tables with fixed two-dimensional marginal*, 2003.

[10] S. Aoki and A. Takemura, *Minimal basis for a connected Markov chain over* $3 \times 3 \times k$ *contingency tables with fixed two-dimensional marginals*, Australian & New Zealand Journal of Statistics, 45 (2003).

[11] A. I. Barvinok, *A polynomial time algorithm for counting integral points in polyhedra when the dimension is fixed*, Mathematics of Operations Research, 19 (1994), pp. 769–779.

[12] D. Bernstein and C. O'Neill, *Unimodular hierarchical models and their Graver bases*, Journal of Algebraic Statistics, 8 (2017).

[13] D. Bernstein and S. Sullivant, *Unimodular binary hierarchical models*, Journal of Combinatorial Theory, Series B, 123 (2017), pp. 97–125.

[14] Y. Berstein and S. Onn, *The Graver complexity of integer programming*, Annals of Combinatorics, 13 (2009), pp. 289–296.

[15] J. Besag and P. Clifford, *Generalized Monte Carlo significance tests*, Biometrika, 76 (1989), pp. 633–642.

[16] F. D. Biase and R. Urbanke, *An algorithm to calculate the kernel of certain polynomial ring homomorphisms*, Experimental Mathematics, 4 (1995), pp. 227–234.

[17] A. Bigatti, R. LaScala, and L. Robbiano, *Computing toric ideals*, Journal of Symbolic Computation, 27 (1999), pp. 351–365.

[18] C. Bortner, J. Garbett, E. Gross, C. McClain, N. Krawzik, and D. Young, *Maximum likelihood degree of the β-stochastic blockmodel*, 2024. To appear in *Algebraic Statistics*. Available at arXiv:2410.06223.

[19] W. Bruns and J. Gubeladze, *Polytopes, Rings, and K-Theory*, Springer Monographs in Mathematics, Springer, Dordrecht, 2009.

[20] F. Bunea and J. Besag, *Markov chain monte carlo in $I \times J \times K$ contingency tables*, in Fields Institute Communications, vol. 26, American Mathematical Society, Providence, Rhode Island, 2000, pp. 25–36.

[21] ———, *MCMC in $i \times j \times k$ contingency tables*, Fields Institute Communications, 26 (2000).

[22] G. Casella and R. Berger, *Statistical Inference*, Duxbury advanced series in statistics and decision sciences, Thomson Learning, 2002.

[23] H. Charalambous, A. Thoma, and M. Vladoiu, *Markov bases and generalized Lawrence liftings*, Ann. Comb., 19 (2015), pp. 661–669.

[24] S. Chatterjee, P. Diaconis, and A. Sly, *Random graphs with a given degree sequence*, Ann. Appl. Probab., 21 (2011), pp. 1400–1435.

[25] Y. Chen, I. Dinwoodie, and A. Dobra, *Lattice points, contingency tables, and sampling*, Contemporary Mathematics, 374 (2005).

[26] J. Chiquet, S. Donnet, and P. Barbillon, *sbm: Stochastic Blockmodels*, 2024. R package version 0.4.6.

[27] D. Cifuentes and S. Onn, *On the complexity of toric ideals*, 2019. Available at arXiv:1902.01484.

[28] B. Cloteaux, *Fast sequential creation of random realizations of degree sequences*, Internet Math., 12 (2016), pp. 205–219.

[29] P. Conti and C. Traverso, *Buchberger algorithm and integer programming*, in Applied Algebra, Algebraic Algorithms and Error-Correcting Codes, H. F. Mattson, T. Mora, and T. R. N. Rao, eds., Berlin, Heidelberg, 1991, Springer Berlin Heidelberg, pp. 130–139.

[30] D. Cox, J. Little, and D. O'Shea, *Using Algebraic Geometry*, Graduate texts in mathematics, Springer, 1998.

[31] D. Cox, J. Little, and D. O'Shea, *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*, Undergraduate Texts in Mathematics, Springer New York, 2008.

[32] D. Cox, J. Little, and H. Schenck, *Toric Varieties*, Graduate studies in mathematics, American Mathematical Society, 2011.

[33] J. Cslovjecsek, F. Eisenbrand, C. Hunkenschröder, L. Rohwedder, and R. Weismantel, *Block-structured integer and linear programming in strongly polynomial and near linear time*, in Proceedings of the Thirty-Second Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '21, Society for Industrial and Applied Mathematics, 2021, p. 1666–1681.

[34] J. Cslovjecsek, M. Koutecký, A. Lassota, M. Pilipczuk, and A. Polak, *Parameterized algorithms for block-structured integer programs with large entries*, 2023. To appear in SODA 2024. Available at arXiv:2311.01890.

[35] J. A. De Loera, R. Hemmecke, and M. Köppe, *Algebraic and Geometric Ideas in the Theory of Discrete Optimization*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2012.

[36] J. A. De Loera, R. Hemmecke, S. Onn, and R. Weismantel, *N-fold integer programming*, Discrete Optimization, 5 (2008), pp. 231–241.

[37] J. A. De Loera, R. Hemmecke, J. Tauzer, and R. Yoshida, *Effective lattice point counting in rational convex polytopes*, Journal of Symbolic Computation, 38 (2004), pp. 1273–1302.

[38] J. A. De Loera and S. Onn, *All linear and integer programs are slim 3-way transportation programs*, SIAM Journal on Optimization, 17 (2006), pp. 806–821.

[39] J. A. De Loera and S. Onn, *Markov bases of three-way tables are arbitrarily complicated*, Journal of Symbolic Computation, 41 (2006), pp. 173–181.

[40] J. A. De Loera, B. Sturmfels, and R. R. Thomas, *Gröbner bases and triangulations of the second hypersimplex*, Combinatorica, 15 (1995), pp. 409–424.

[41] J. A. De Loera, B. Sturmfels, and R. R. Thomas, *Gröbner bases and triangulations of the second hypersimplex*, Combinatorica, 15 (1995), pp. 409–424.

[42] A. Deza, A. Levin, S. M. Meesum, and S. Onn, *Optimization over degree sequences*, SIAM Journal on Discrete Mathematics, 32 (2018), pp. 2067–2079.

[43] A. Deza, A. Levin, S. M. Meesum, and S. Onn, *Hypergraphic degree sequences are hard*, Bulletin of EATCS, 1 (2019).

[44] P. Diaconis and A. Gangolli, *Rectangular arrays with fixed margins*, in Discrete Probability and Algorithms, D. Aldous, P. Diaconis, J. Spencer, and J. M. Steele, eds., New York, NY, 1995, Springer New York, pp. 15–41.

[45] P. Diaconis and B. Sturmfels, *Algebraic algorithms for sampling from conditional distributions*, Annals of Statistics, 26 (1998), pp. 363–397.

[46] A. Dobra, *Markov bases for decomposable graphical models*, Bernoulli, 9 (2003), pp. 1093–1108.

[47] A. Dobra and S. Sullivant, *A divide-and-conquer algorithm for generating Markov bases of multi-way tables*, Computational Statistics, 19 (2004), pp. 347–366.

[48] M. Drton, B. Sturmfels, and S. Sullivant, *Lectures on Algebraic Statistics*, vol. 39 of Oberwolfach Seminars, Springer, 2009.

[49] F. Eisenbrand, C. Hunkenschröder, K.-M. Klein, M. Koutecký, A. Levin, and S. Onn, *An algorithmic theory of integer programming*, 2022. Available at arXiv:1904.01361.

[50] P. Erdös and A. Rényi, *On the evolution of random graphs*, Princeton University Press, Princeton, 2006, pp. 38–82.

[51] P. Erdős and T. Gallai, *Gráfok előírt fokszámú pontokkal*, Matematikai Lapok, 11 (1960), p. 264–274.

[52] P. L. Erdős, C. Greenhill, T. R. Mezei, I. Miklós, D. Soltész, and L. Soukup, *The mixing time of switch markov chains: A unified approach*, European Journal of Combinatorics, 99 (2022), p. 103421.

[53] S. E. Fienberg, M. M. Meyer, and S. S. Wasserman, *Statistical analysis of multiple sociometric relations*, Journal of the American Statistical Association, 80 (1985), pp. 51–67.

[54] S. E. Fienberg and S. Wasserman, *Categorical data analysis of single sociometric relations*, Sociological Methodology, 12 (1981), p. 156.

[55] S. E. Fienberg and S. S. Wasserman, *Categorical data analysis of single sociometric relations*, Sociological Methodology, 12 (1981), pp. 156–192.

[56] ———, *Discussion of holland, p. w. and leinhardt, s. "an exponential family of probability distributions for directed graphs"*, Journal of the American Statistical Association, 76 (1981), pp. 54–57.

[57] R. Fröberg, *The frobenius number of some semigroups*, Communications in Algebra, 22 (1994), pp. 6021–6024.

[58] W. Fulton, *Introduction to Toric Varieties*, Annals of Mathematics Studies, Princeton University Press, Princeton, NJ, 1993.

[59] R. Gilmer, *Commutative Semigroup Rings*, Chicago Lectures in Mathematics, University of Chicago Press, Chicago, 1984.

[60] J. E. Graver, *On the foundations of linear and integer linear programming i*, Mathematical Programming, 9 (1975), pp. 207–226.

[61] E. Gross, V. Karwa, and S. Petrović, *Algebraic statistics, tables, and networks: The fienberg advantage*, in Statistics in the Public Interest: In Memory of Stephen E. Fienberg, Springer International Publishing, Cham, 2022, pp. 33–49.

[62] E. Gross, S. Petrović, and D. Stasi, *Goodness of fit for log-linear ergms*, arXiv preprint arXiv:2104.03167, (2024).

[63] S. L. Hakimi, *On realizability of a set of integers as degrees of the vertices of a linear graph. i*, Journal of the Society for Industrial and Applied Mathematics, 10 (1962), pp. 496–506.

[64] H. Hara, S. Aoki, and A. Takemura, *Running Markov chain without Markov basis*, in Harmony of Gröbner Bases and the Modern Industrial Society, World Scientific, 2012, pp. 46–62.

[65] B. Hassett, *Introduction to Algebraic Geometry*, Cambridge University Press, Cambridge, 2007.

[66] V. Havel, *A remark on the existence of finite graphs*, Časopis pro pěstování matematiky, 80 (1955), pp. 477–480.

[67] R. Hemmecke and P. N. Malkin, *Computing generating sets of lattice ideals and markov bases of lattices*, Journal of Symbolic Computation, 44 (2009), pp. 1463–1476.

[68] A. J. Hoffman and J. B. Kruskal, *Integral boundary points of convex polyhedra*, in 50 Years of Integer Programming 1958-2008: From the Early Years to the State-of-the-Art, M. Jünger, T. M. Liebling, D. Naddef, G. L. Nemhauser, W. R. Pulleyblank, G. Reinelt, G. Rinaldi, and L. A. Wolsey, eds., Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 49–76.

[69] P. W. Holland and S. L. Leinhardt, *An exponential family of probability distributions for directed graphs (with discussion)*, Journal of the American Statistical Association, 76 (1981), pp. 33–65.

[70] S. Hosten and B. Sturmfels, *Grin: An implementation of gröbner bases for integer programming*, in Integer Programming and Combinatorial Optimization, E. Balas and J. Clausen, eds., vol. 920 of Lecture Notes in Computer Science, Springer Verlag, 1995, pp. 267–276.

[71] S. Hosten and B. Sturmfels, *Computing the integer programming gap*, Combinatorica, 27 (2003), pp. 367–382.

[72] S. Hoşten and S. Sullivant, *A finiteness theorem for markov bases of hierarchical models*, J. Combin. Theory Ser. A, 114 (2007), pp. 311–321.

[73] D. R. Hunter, S. M. Goodreau, and M. S. Handcock, *Goodness of fit of social network model*, Journal of the American Statistical Association, 103 (2008), pp. 248–258.

[74] M. Jerrum, *Counting, Sampling and Integrating: Algorithms and Complexity*, Lectures in Mathematics. ETH Zürich, Birkhäuser Basel, 2003.

[75] M. Jerrum and A. Sinclair, *Fast uniform generation of regular graphs*, Theoretical Computer Science, 73 (1990), pp. 91–100.

[76] A. B. Junxian Geng and D. Pati, *Probabilistic community detection with unknown number of communities*, Journal of the American Statistical Association, 114 (2019), pp. 893–905.

[77] D. Kahle, L. Garcia-Puente, and R. Yoshida, *algstat: Algebraic statistics in R*, 2014. R package version 0.1.1.

[78] R. Kannan, P. Tetali, and S. S. Vempala, *Simple markov-chain algorithms for generating bipartite graphs and tournaments*, Random Struct. Algorithms, 14 (1997), pp. 293–308.

[79] V. Karwa, D. Pati, S. Petrović, L. Solus, N. Alexeev, M. Raič, D. Wilburne, R. Williams, and B. Yan, *Monte carlo goodness-of-fit tests for degree corrected and related stochastic blockmodels*, Journal of the Royal Statistical Society Series B: Statistical Methodology, 86 (2023), pp. 90–121.

[80] D. Knop, M. Pilipczuk, and M. Wrochna, *Tight complexity lower bounds for integer linear programming with few constraints*, ACM Trans. Comput. Theory, 12 (2020).

[81] M. Koren, *Extreme degree sequences of simple graphs*, Journal of Combinatorial Theory, Series B, 15 (1973), pp. 213–224.

[82] M. Koutecký, A. Levin, and S. Onn, *A Parameterized Strongly Polynomial Algorithm for Block Structured Integer Programs*, in 45th International Colloquium on Automata, Languages, and Programming (ICALP 2018), vol. 107 of Leibniz International Proceedings in Informatics (LIPIcs), 2018, pp. 85:1–85:14.

[83] M. Koutecký and S. Onn, *Sparse integer programming is fpt*, Bulletin of the European Association for Theoretical Computer Science, 134 (2021), pp. 69–71.

[84] S. Lee, *Markov chain Monte Carlo and exact conditional tests with three-way contingency tables*, Master's thesis, Naval Postgraduate School, 2018.

[85] J. Lei, *A goodness-of-fit test for stochastic block models*, The Annals of Statistics, 44 (2016), pp. 401 – 424.

[86] D. A. Levin, Y. Peres, and E. L. Wilmer, *Markov Chains and Mixing Times*, American Mathematical Society, Providence, RI, 2009.

[87] N. Mahadev and U. Peled, *Threshold Graphs and Related Topics*, Annals of Discrete Mathematics, North Holland, 1995.

[88] M. Mariadassou, S. Robin, and C. Vacher, *Uncovering latent structure in valued graphs: A variational approach*, The Annals of Applied Statistics, 4 (2010), pp. 715 – 742.

[89] E. Miller and B. Sturmfels, *Combinatorial Commutative Algebra*, vol. 227 of Graduate Texts in Mathematics, Springer, 2004.

[90] S. Onn, A. Thoma, and M. Vladoiu, *Asymptotic behavior of markov complexity*, Journal of Pure and Applied Algebra, 228 (2024), p. 107589.

[91] U. N. Peled and M. K. Srinivasan, *The polytope of degree sequences*, Linear Algebra and its Applications, 114-115 (1989), pp. 349–377. Special Issue Dedicated to Alan J. Hoffman.

[92] S. Petrović, A. Rinaldo, and S. E. Fienberg, *Algebraic statistics for a directed random graph model with reciprocation*, in Algebraic Methods in Statistics and Probability II, M. Viana and H. Wynn, eds., vol. 516 of Contemporary Mathematics, American Mathematical Society, Providence, RI, 2010, pp. 261–283.

[93] S. Petrović and D. Stasi, *Toric algebra of hypergraphs*, Journal of Algebraic Combinatorics, 39 (2014), pp. 187–208.

[94] L. Pottier, *Minimal solutions to linear diophantine systems: Bounds and algorithms*, in International Congress on Rewriting Techniques and Applications (RTA 91), vol. 488 of Lecture Notes in Computer Science, Como, Italy, 1991, Springer, pp. 162–173.

[95] J. L. Ramírez Alfonsín, *The Diophantine Frobenius Problem*, vol. 30 of Oxford Lecture Series in Mathematics and Its Applications, Oxford University Press, Oxford, 2005.

[96] F. Rapallo, *Algebraic Markov bases and MCMC for two-way contingency tables*, Scand. J. Statist., 30 (2003), pp. 385–397.

[97] E. Reyes, R. H. Villarreal, and L. Zárate, *A note on affine toric varieties*, Linear Algebra and Its Applications, 318 (2000), pp. 173–179.

[98] C. Robert and G. Casella, *Monte Carlo Statistical Methods*, Springer Texts in Statistics, Springer New York, 2013.

[99] F. Santos and B. Sturmfels, *Higher lawrence configurations*, Journal of Combinatorial Theory, Series A, 103 (2003), pp. 151–164.

[100] A. Schrijver, *Theory of Linear and Integer Programming*, Wiley-Interscience Series in Discrete Mathematics, Wiley, Chichester, 1986.

[101] ———, *Theory of linear and integer programming*, in Wiley-Interscience series in discrete mathematics and optimization, 1999.

[102] R. P. Stanley, *A zonotope associated with graphical degree sequences*, in Applied Geometry and Discrete Mathematics, vol. 4 of DIMACS Series in Discrete Mathematics and Theoretical Computer Science, American Mathematical Society, Providence, RI, 1991, pp. 555–570.

[103] ———, *Combinatorics and Commutative Algebra*, Birkhäuser Boston, 2nd ed., 1996.

[104] B. Sturmfels, *Gröbner bases and convex polytopes*, University Lecture Series, no. 8, American Mathematical Society, 1996.

[105] B. Sturmfels and R. R. Thomas, *Variation of cost functions in integer programming*, Mathematical Programming, 77 (1997), pp. 357–387.

[106] S. Sullivant, *Algebraic Statistics*, Graduate Studies in Mathematics, American Mathematical Society, 2021.

[107] A. Takken, *Monte Carlo Goodness-of-Fit Tests for Discrete Data*, PhD thesis, Dept. Statistics, Stanford University, 2000.

[108] P. B. Timothée Tabouy and J. Chiquet, *Variational inference for stochastic block models from sampled data*, Journal of the American Statistical Association, 115 (2020), pp. 455–466.

[109] R. Villarreal, *Monomial Algebras*, Monographs and Research Notes in Mathematics, Taylor & Francis, 2000.

[110] R. H. Villarreal, *Monomial Algebras*, Chapman & Hall/CRC Monographs and Research Notes in Mathematics, CRC Press, 2018.

[111] L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference*, Springer Texts in Statistics, Springer, 2004.

[112] R. Yoshida, *Open problems on connectivity of fibers with positive margins in multi-dimensional contingency tables*, J. Algebr. Stat., 1 (2010), pp. 13–26.

[113] R. Yoshida and D. Barnhill, *Connecting tables with allowing negative cell counts*, (2023). Available at arXiv:2205.07167.