ABSTRACT

HOLLERING, BENJAMIN KEITH. Computational and Combinatorial Techniques for Phylogenetic Algebraic Geometry. (Under the direction of Seth Sullivant).

Algebraic statistics uses tools from algebra, geometry, and combinatorics to study problems in statistics. This is often done by studying algebraic varieties which arise as the Zariski closure of a statistical model. In this thesis we mainly focus on statistical models and probability distributions which comes from phylogenetics.

First we study the family of exchangeable and sampling consistent distributions on rooted binary trees with labelled leaves, which we call phylogenetic trees. We begin by introducing a notion of finite sampling consistency for phylogenetic trees and show that the set of finitely sampling consistent and exchangeable distributions on n leaf phylogenetic trees is a polytope. We then study the vertices of this polytope for trees with 4 and 5 leaves. We also introduce a new semialgebraic set of exchangeable and sampling consistent models we call the multinomial model and use it to characterize the set of exchangeable and sampling consistent distributions. Using this new model, we obtain a finite de Finetti-type theorem for rooted binary trees.

Next, we discuss a new algorithm for proving that discrete parameters of parametric algebraic statistical models are identifiable. Identifiability is a particularly important property for statistical models to have since it ensures that the parameters can be reliably recovered from data. Our method uses algebraic matroids which are naturally associated to the model and by doing so avoids time consuming Gröbner basis calculations. We then use this method to solve some previously open identifiability problems from phylogenetics.

Lastly, we study the vanishing ideal of the Cavendar-Farris-Neyman (CFN) model on level-1 phylogenetic networks. We show that these ideals are multigraded and use this multigrading to break up the ideal into homogeneous pieces called *gloves*. We then give an explicit description of the quadratic polynomials in each glove which we conjecture generate the entire ideal.

Computational and Combinatorial Techniques for Phylogenetic Algebraic Geometry

by
Benjamin Keith Hollering

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Mathematics

Raleigh, North Carolina
2022

APPROVED BY:

Martin Helmer	Hoon Hong
Nathan Reading	Seth Sullivant Chair of Advisory Committee

DEDICATION

To Eleanor.

BIOGRAPHY

Benjamin Hollering grew up in Raleigh, North Carolina where he attended Panther Creek High School. He received his undergraduate degree in Mathematics from North Carolina State University in December 2016 and began his Ph.D. at NCSU in January 2017. He will continue his mathematical career as a postdoc at the Max Planck Institute for Mathematics in the Sciences in Leipzig followed by a postdoc at the Technical University of Munich. In his spare time, Ben enjoys coffee, cooking, reading, and video games.

ACKNOWLEDGEMENTS

I have many people to thank for helping me along this journey. First and foremost, I would like to thank my advisor, Seth Sullivant. He has been an excellent teacher and mentor throughout my time in graduate school. He has also given me invaluable feedback on mathematical papers, presentations, and job applications which really helped me showcase my work in the best way possible. Seth also encouraged me to present at conferences and collaborate with other researchers which helped me establish myself as a member of the algebraic statistics community. There is no doubt in my mind that working with Seth was the best decision of my career so far.

The algebraic statistics community has been exceedingly welcoming to new members and made a great effort to help up and coming researchers meet people during COVID, for which I am extremely grateful. I would especially like to thank my collaborators Carlos Améndola, Jane Coons, Joe Cummings, Chris Manon, Aida Maraj, Seth Sullivant, and Ngoc Tran. I have learned so much from them throughout my time in graduate school and I am sure I will continue to do so as a postdoc. In addition, I'd like to thank Jose Israel Rodriguez for teaching me so much about macaulay2 and computational methods in algebraic geometry. I'd also like to thank my letter writers, Elizabeth Dempster, Mathias Drton, Elizabeth Gross, Chris Manon, Seth Sullivant, and Ngoc Tran as well as my committee Martin Helmer, Hoon Hong, Nathan Reading, and Seth Sullivant, and Agnes Szanto.

I've made many great friends in graduate school who have really helped to make it a wonderful experience. I cannot possibly name everyone here, but I am grateful to all of you. I'd like to give a special thank you to Jane Coons, Joe Cummings, Aida Maraj, Cash Bortner, Joe Johnson, and Pratik Misra. Thank you all for being excellent friends and for the invaluable discussions we've had over the years. I'm especially grateful to Jane for providing me with many useful templates and fantastic advice while I was applying to jobs.

A large part of the reason I chose NCSU for graduate school was the fantastic undegraduate experience I had there. I made many of my best friends during my undergrad in both the math program and outside of it. I would like to give special thanks to Tyler Albert, Prem Shah, Zach Ledford, Shyam Patel, Cameron Johnson, and Connor Sierra. I will always remember our long walks to get coffee in the middle of the night, the conversations at the water fountain, and the late nights we had playing League of Legends.

I am incredibly grateful to my family for their ceaseless support. At every step of my academic career my parents have always been encouraging. They provided me with so many opportunities and amazing advice at every point in my life. I'd also like to thank my siblings Kate, Sam, and Will. I have always enjoyed your visits at NCSU and will miss you all so much after I move.

Finally, I would like to thank Eleanor Yeh. Since we met as undegrads she has been an amazing partner and friend. She supported me when I decided to get a Ph.D. and at every step of the way since then. I am eternally grateful for her love and support.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER

1

INTRODUCTION

In this chapter we provide background material that will be used throughout the thesis. This includes a brief introduction to ideals, varieties, polytopes, and matroids. We also discuss phylogenetic trees, networks, and Markov models which will be studied throughout this thesis.

1.1 Ideals and Varieties

In this section we give the necessary background on algebraic geometry. The objects and results discussed here will primarily be used in Chapters 3 and 4. For further information on algebraic geometry we refer the reader to [25].

Let $\mathbb{K}$ be a field and denote the polynomial ring over $\mathbb{K}$ with n indeterminates by $\mathbb{K}[p] = \mathbb{K}[p_1, \ldots, p_n]$. We will primarily focus on the case when $\mathbb{K}$ is the field of complex numbers $\mathbb{C}$. For any $a \in \mathbb{N}^n$ we will use the shorthand $p^\alpha = p_1^{\alpha_1} \ldots p_n^{\alpha_n}$.

Definition 1.1.1. An *ideal* in the polynomial ring is a subset $I \subseteq \mathbb{K}[p]$ such that

- if $f, g \in J$ then $f + g \in J$ and

- if $f \in J$ and $g \in \mathbb{K}[p]$ then $gf \in J$.

Given a subset $F \subseteq \mathbb{K}[p]$ we let $\langle F \rangle$ be the ideal generated by the polynomials in F meaning

$$\langle F \rangle := \{g_1 f_1 + \ldots + g_k f_k \mid f_i \in F, g_i \in \mathbb{K}[p]\}.$$

The *Hilbert Basis Theorem* guarantees that if J is an ideal of $\mathbb{K}[p]$, then there exists a finite subset $F = \{f_1, \ldots, f_k\} \subseteq J$ such that $J = \langle F \rangle$. Every set of polynomials F has the following natural geometric counterpart.

Definition 1.1.2. Let $F \subset \mathbb{K}[p]$. The *affine variety* defined by F is the set of points

$$\mathcal{V}(F) := \{a \in \mathbb{K}^n \mid f(a) = 0 \text{ for all } f \in F\}.$$

Note that it is also natural to associate a variety to an ideal J for the following reason. Let $F = \{f_1, \ldots, f_k\}$ be the finite set of generators of J and note that if $a \in \mathcal{V}(F)$ and $f \in J$ then $f(a) = \sum_i g_i(a) f_i(a) = \sum_i g_i(a) = 0$. So to any ideal J we can associate the variety

$$\mathcal{V}(J) = \{a \in \mathbb{K}^n \mid f(a) = 0 \text{ for all } f \in J\}.$$

Example 1.1.3. Consider the polynomial ring $R = \mathbb{C}[p_{00}, p_{01}, p_{10}, p_{11}]$ and the ideal $J = \langle p_{00}p_{11} - p_{01}p_{10} \rangle$. If we consider the points $a \in \mathbb{C}^4$ as complex 2×2 matrices with the standard indexing then the variety $\mathcal{V}(J)$ is the set of all 2×2 matrices with rank at most 1.

Just as every set of polynomials or polynomial ideal defines a variety we also get an ideal from any set of points in $\mathbb{K}^n$. Given a subset $S \subseteq \mathbb{K}^n$, the *vanishing ideal* of S is the ideal

$$\mathcal{I}(S) := \{f \in \mathbb{K}[p] \mid f(a) = 0 \text{ for all } a \in S\}.$$

For any set $S \subseteq \mathbb{K}^n$, the *Zariski closure* of S is the smallest variety containing S and is denoted by $\overline{S}$. Note that this set is simply $\overline{S} = \mathcal{V}(\mathcal{I}(S))$. We now turn to another important property of ideals and varieties.

Definition 1.1.4. An ideal J is *prime* if whenever $fg \in J$ then either $f \in J$ or $g \in J$.

Primality also has the following geometric counterpart. A variety is *irreducible* if whenever $V = V_1 \cup V_2$ then either $V_2 \subseteq V_1 = V$ or $V_1 \subseteq V_2 = V$. Otherwise V is called *reducible*. The relationship between irreducibility and primality is captured by the following proposition.

Proposition 1.1.5. *A variety V is irreducible if and only the vanishing ideal $\mathcal{I}(V)$ is a prime ideal.*

These two concepts are illustrated in the following example.

Example 1.1.6. Consider the ideal $J = \langle p_1 p_2 \rangle \subseteq \mathbb{C}[p_1, p_2]$ and the corresponding variety $V = \mathcal{V}(J) \subseteq \mathbb{C}^2$. Observe that J is not prime since $p_1, p_2 \notin J$ but of course the product $p_1 p_2 \in J$. On the other hand we can also see that V is reducible since $V = \mathcal{V}(p_1) \cup \mathcal{V}(p_2)$.

Now consider the ideal $J' = \langle p_1^2 - p_2 \rangle \subseteq \mathbb{C}[p_1, p_2]$ and its corresponding variety $V' = \mathcal{V}(J') \subseteq \mathbb{C}^2$. Note that since the polynomial $p_1^2 - p_2$ is irreducible, the ideal J' is prime and hence the variety V' is also irreducible.

Many of the ideals we work with in this thesis are not only prime but also have a natural *grading*. A polynomial ring $\mathbb{K}[p]$ is *multigraded* by a lattice $\mathcal{A}$ if it is equipped with with a semigroup homomorphism $\deg : \mathbb{N}^n \to \mathcal{A}$ which takes each monomial p^α to its degree $\deg(\alpha)$. Observe that we can apply deg to a monomial p^α by taking $\deg(p^\alpha) = \deg(\alpha)$. Also note that since deg is a semigroup homomorphism, it suffices to define deg on the set of variables $\{p_1, \dots p_n\}$. This is illustrated by Example 1.1.7.

A polynomial $f = \sum_\alpha c_\alpha p^\alpha$ is *homogeneous* with respect to deg if there is some $m \in \mathcal{A}$ such that for all $\alpha \in \mathbb{N}^n$ such that $c_\alpha \neq 0$, it holds that $\deg(p^\alpha) = m$. An ideal $J \subseteq \mathbb{K}[p]$ is homogeneous with respect to the grading deg if there exists a generating set $\{f_1, \dots, f_k\}$ for J such that each f_i is homogeneous.

Example 1.1.7. Consider the grading on $\mathbb{C}[p_{00}, p_{01}, p_{10}, p_{11}]$ given by $\deg(p_{ij}) = (1, i, j)$. This determines a multigrading on all of $\mathbb{K}[p]$ since deg is a semigroup homomorphism. This means the ring $\mathbb{C}[p]$ is multigraded by the lattice generated by
$(1,0,0), (1,0,1), (1,1,0), (1,1,1) \in \mathbb{Z}^3$.

Now we will examine the familiar polynomial $f = p_{00}p_{11} - p_{01}p_{10}$. Note that f is homogeneous with respect to this grading since

$$\deg(p_{00}p_{11}) = (1,0,0) + (1,1,1) = (2,1,1) = (1,0,1) + (1,1,0) = \deg(p_{01}p_{10}).$$

On the other hand, the polynomial $g = p_{10}^2 p_{11} - p_{00}p_{01}p_{11}$ is not homogeneous with respect to this grading since

$$\deg(p_{10}^2 p_{11}) = (3,3,1) \neq (3,1,2) = \deg(p_{00}p_{01}p_{11}).$$

If we say that a polynomial is homogeneous without respect to a pre-establish grading then we mean it is homogeneous with respect to total degree which is the grading that simply sets $\deg(p_i) = 1$ for all i. This grading is especially important since ideals that are homogeneous with respect to total degree can be used to define varieties in *projective space*.

Definition 1.1.8. Let $\mathbb{K}$ be a field. The projective space $\mathbb{P}^{n-1}$ is the set of equivalence classes of $\mathbb{K}^n \setminus \{0\}$ under the equivalence relation $\sim$ given by

$$a \sim b \iff \text{ there exists } \lambda \in \mathbb{K}^* \text{ such that } \lambda a = b.$$

Definition 1.1.9. Let $\mathcal{F} \subseteq \mathbb{K}[p]$ be a set of homogeneous polynomials. The *projective variety* defined by $\mathcal{F}$ is the set

$$\mathcal{V}(\mathcal{F}) := \{a \in \mathbb{P}^{n-1} \mid f(a) = 0 \text{ for all } f \in \mathcal{F}\}.$$

Note that the previous definition is well-defined since if f is a homogeneous polynomial and a is a representative of an equivalence class in $\mathbb{P}^n$ such that $f(a) = 0$ then for any other equivalence class representative, λa we have that $f(\lambda a) = \lambda^{\deg(f)} f(a) = 0$.

Example 1.1.10. Again consider the ideal $J = \langle p_{00}p_{11} - p_{01}p_{10} \rangle$ in the polynomial ring $\mathbb{C}[p]$. Since J is generated by a single homogeneous polynomial, it is a homogeneous ideal and thus defines a projective variety which consists of equivalence classes of matrices $a \in \mathbb{P}^3$ such that $\text{rank}(A) \leq 1$.

1.1.1 Gröbner Bases

In this subsection we develop some of the basic concepts related to Gröbner bases which are utilized in many places throughout this thesis.

Definition 1.1.11. A *term order* $<$ on $\mathbb{K}[p_1, \ldots p_n]$ is a total order on the monomials which satisfies:

1. If $p^\alpha < p^\beta$ then $p^\alpha p^\gamma < p^\beta p^\gamma$ for all α, β, γ
2. Every nonempty set of monomials has a $<$-smallest element

Definition 1.1.12. Let $f = \sum_\alpha c_\alpha p^\alpha$ be a nonzero polynomial and let $<$ be a monomial order. The *initial monomial* (or leading monomial) of f, denoted $\text{in}_<(f)$, is the largest

monomial p^α such that $c_\alpha \neq 0$. The *initial term* (or leading term) of f, denoted $\mathrm{LT}_<(f)$, is corresponding term $c_\alpha p^\alpha$.

Example 1.1.13 (Lexicographic Order)**.** Consider the term order where $p^\alpha >_{\text{lex}} p^\beta$ if the left-most nonzero entry of $(\alpha_1 - \beta_1, \ldots, \alpha_n - \beta_n)$ is positive. For example,

$$p_1 p_3 >_{\text{lex}} p_2^2 p_3 >_{\text{lex}} p_3^4$$

under the assumption that $p_1 > p_2 > p_3$. Now consider the polynomial $f = 2p_1^3 + p_1 p_2$. The initial monomial of f with respect to $<_{\text{lex}}$ is $\mathrm{in}_{<_{\text{lex}}}(f) = p_1^3$ and the leading term is $\mathrm{LT}_{<_{\text{lex}}}(f) = 2p_1^3$.

Given a term order $<$ and an ideal J the *initial ideal* of J is the ideal

$$\mathrm{in}_<(J) := \langle \mathrm{in}_<(f) \mid f \in J \rangle.$$

Definition 1.1.14. Let $J \subseteq \mathbb{K}[p]$ be an ideal and $<$ be a term order on $\mathbb{K}[p]$. Then a set $\mathcal{G}$ of polynomials is a *Gröbner basis* for the ideal J if

$$\mathrm{in}_<(J) = \langle \mathrm{in}_<(g) \mid g \in \mathcal{G} \rangle.$$

Gröbner bases are an incredibly useful tool for performing computations in algebraic geometry. In particular, they can be used to test for ideal membership and solve implicitization problems. We now describe how to compute a Gröbner basis for an ideal from a given generating set using Buchberger's algorithm. The first tool we need to do this is the following algorithm for polynomial division.

Algorithm 1: Multivariate Division Algorithm

Input : A finite set of polynomials $\mathcal{G} = \{g_1, \ldots g_k\}$, another polynomial f, and a term order $<$.

Output : A representation $f = \sum_{i=1}^k h_i g_i + r$ such that no term of r is divisible by $\mathcal{G}$.

Set $h_i = 0$ for all i and $r = f$
while *r has a term $c_\alpha p^\alpha$ divisible by a leading term of some g_i* **do**
    $h_i = h_i + \frac{c_\alpha p^\alpha}{\mathrm{in}_<(g_i)}$
    $r = r - \frac{c_\alpha p^\alpha}{\mathrm{in}_<(g_i)} g_i$
end
return *h_i and r for $i = 1, \ldots, n$*

Definition 1.1.15. The *least common multiple* of monomials p^α and p^β is

$$\mathrm{LCM}(p^\alpha, p^\beta) = \prod_{i=1}^{n} p_i^{\max\{\alpha_i, \beta_i\}}.$$

Let f_1 and f_2 be polynomials in $\mathbb{K}[p]$ and $<$ be a term order.
Let $p^{\gamma(1,2)} = \mathrm{LCM}(\mathrm{in}_<(f_1), \mathrm{in}_<(f_2))$. Then the *S-polynomial* of f and g is

$$S(f,g) := \frac{p^{\gamma(1,2)}}{\mathrm{LT}_<(f_1)} f_1 - \frac{p^{\gamma(1,2)}}{\mathrm{LT}_<(f_2)} f_2$$

Theorem 1.1.16 (Buchberger's Criterion)**.** *Let $<$ be a term order on $\mathbb{K}[p]$ and $\mathcal{G} = \{g_1, \ldots g_k\}$ be a set of polynomials. The following are equivalent:*

1. *$\mathcal{G}$ is a Gröbner basis for $\langle \mathcal{G} \rangle$.*

2. *The remainder of each S-polynomial $S(g_i, g_j)$ after division by $\mathcal{G}$ is zero.*

This criterion can be used to easily formulate an algorithm for computing a Gröbner basis for an ideal J from any generating set. This is summarized by the following algorithm.

Algorithm 2: Buchberger's Algorithm

Input : A finite set of polynomials $\mathcal{F} = \{g_1, \ldots g_k\}$ and a term order $<$.

Output : A Gröbner basis $\mathcal{G}$ for $\langle \mathcal{F} \rangle$ with respect to $<$.

```
Set G = F
while G does not satisfy Buchberger's criterion do
    Find a pair f, g ∈ G such that remainder r obtained by dividing S(f, g) by G
     is not zero
    G = G ∪ {r}
end
return G
```

We end this section with an example that illustrates all of the ideas discussed above.

Example 1.1.17. Consider the ideal $J = \langle f_1, f_2 \rangle$ where $f_1 = p_1 p_2 + 1$ and $f_2 = p_1^2 - p_2$. We will now use Buchberger's algorithm to find a Gröbner basis $\mathcal{G}$ for J with respect to the graded lexicographic term order $<$. We first set $\mathcal{G} = \{f_1, f_2\}$ and compute the S-polynomial $S(f_1, f_2)$.

Observe that $\mathrm{in}_<(f_1) = p_1 p_2$ and $\mathrm{in}_<(f_2) = p_1^2$. This means that $p^{\gamma(1,2)} = p_1^2 p_2$. So we

have that

$$S(f_1, f_2) = \frac{p_1^2 p_2}{p_1 p_2}(p_1 p_2 + 1) - \frac{p_1^2 p_2}{p_1^2}(p_1^2 - p_2) = p_1(p_1 p_2 + 1) - p_2(p_1^2 - p_2) = p_2^2 + p_1.$$

Note that both terms of $S(f_1, f_2)$ are not divisible by either $\text{in}_<(f_1)$ or $\text{in}_<(f_2)$. So the remainder of $S(f_1, f_2)$ upon division by $\{f_1, f_2\}$ is simply $f_3 = S(f_1, f_2)$. So we set $\mathcal{G} = \{f_1, f_2, f_3\}$.

We now need to compute the S-polynomials $S(f_1, f_3)$ and $S(f_2, f_3)$.
Note that $\text{in}_<(f_3) = p_2^2$. So we have that

$$S(f_1, f_3) = \frac{p_1 p_2^2}{p_1 p_2}(p_1 p_2 + 1) - \frac{p_1 p_2^2}{p_2^2}(p_1 + p_2^2) = -p_1^2 + p_2$$
$$S(f_2, f_3) = \frac{p_1^2 p_2^2}{p_1^2}(p_1^2 - p_2) - \frac{p_1^2 p_2^2}{p_2^2}(p_1 + p_2^2) = -p_1^3 - p_2^3$$

We now need to divide $S(f_1, f_3)$ and $S(f_2, f_3)$ by $\mathcal{G}$. Observe that $S(f_1, f_3) = -f_2$ so we immediately have that the remainder upon division by $\mathcal{G}$ is zero.

Recall that to divide $S(f_2, f_3)$ by $\mathcal{G}$ we want to find an expression $S(f_2, f_3) = h_1 f_1 + h_2 f_2 + h_3 f_3 + r$. We begin by setting $r = S(f_2, f_3)$, $h_i = 0$ and observe that the term $-p_1^3$ in r is divisible by $\text{in}_<(f_2) = p_1^2$. So we set

$$h_2 = 0 + \frac{-p_1^3}{p_1^2} = -p_1$$
$$r = -p_1^3 - p_2^3 - \frac{-p_1^3}{p_1^2}(p_1^2 - p_2) = -p_2^3 - p_1 p_2.$$

We now have the term $-p_2^3$ in r which is divisible by $\text{in}_<(f_3)$. This division step yields

$$h_3 = 0 + \frac{-p_2^3}{p_2^2} = -p_2$$
$$r = -p_2^3 - p_1 p_2 - \frac{-p_2^3}{p_2^2}(p_2^2 + p_1) = 0.$$

So the remainder of $S(f_2, f_3)$ upon division by $\mathcal{G}$ is also zero thus by the Buchberger criterion we have that $\mathcal{G}$ is a Gröbner basis for J.

1.1.2 Parameterized Varieties

Varieties are often given as the image of a polynomial map. This is especially common in algebraic statistics where we typically work with varieties which come from taking the Zariski closure of a parametric statistical model.

Consider a map

$$\begin{aligned} \phi: \ & \mathbb{K}^m \to \mathbb{K}^n \\ & \theta \mapsto (\phi_1(\theta), \ldots, \phi_n(\theta)) \end{aligned}$$

where the functions ϕ_i are all polynomials in $\theta = (\theta_1, \ldots \theta_m)$. The variety parameterized by ϕ is the Zariski closure of the image of ϕ and denoted as $V = \overline{\text{im}(\phi)}$. All of the parameterized varieties in this thesis will be parameterized by polynomial maps however the same tools may also be used to study rational parameterizations.

A map ϕ satisfying the above conditions is a *morphism* of affine spaces and thus it has a pullback which is a $\mathbb{K}$-algebra homomorphism. This pullback is the map

$$\begin{aligned} \phi^*: \ & \mathbb{K}[p] \to \mathbb{K}[\theta] \\ & p_i \mapsto \phi_i(\theta). \end{aligned}$$

This means that the vanishing ideal $\mathcal{I}(V)$ of the parameterized variety is actually the kernel of the map ϕ^*. Note that this implies that $\mathcal{I}(V)$ is a prime ideal which can be computed using elimination theory.

Definition 1.1.18. A term order $<$ on $\mathbb{K}[\theta_1, \ldots \theta_n, p_1, \ldots p_m]$ is an elimination order for $\theta_1 \ldots \theta_n$ if each polynomial with initial monomial in $\mathbb{K}[p_1 \ldots p_m]$ is actually contained in $\mathbb{K}[p_1, \ldots p_m]$.

Theorem 1.1.19. *Let $\phi : \mathbb{K}^n \to \mathbb{K}^m$ be a morphism of varieties with coordinate functions $(\phi_1, \ldots, \phi_n)$. Let $J \subseteq \mathbb{K}[\theta_1, \ldots \theta_n, p_1, \ldots p_m]$ be the ideal generated by the polynomials $p_i - \phi_i(\theta)$. Let $<$ be an elimination order for $\theta_1, \ldots, \theta_n$ and $\mathcal{G}$ be a Gröbner basis for J with respect to $<$ then $\mathcal{I}(\text{im}(\phi))$ is generated by $\mathcal{G} \cap \mathbb{K}[p_1, \ldots p_m]$.*

Since we have already seen how to compute Gröbner bases, the above theorem gives us an immediate algorithm for computing the vanishing ideal of the image of a polynomial map. Unfortunately, computing a Gröbner basis can be quite difficult. It is known that in the worst-case, the time complexity of computing a Gröbner basis is doubly exponential

in the number of variables, though it can often be much better in practice [5]. As the number of variables grows it often becomes untenable to actually compute a Gröbner basis and so a frequent goal of ours is to determine generating sets or other information about ideals without using Gröbner bases.

We end this section with an example of how we can use algebraic geometry to study parametric statistical models. The techniques utilized below will appear throughout this thesis.

Example 1.1.20. Suppose X_1 and X_2 are two independent binary random variables with state space $\{0,1\}$ and we are interested in studying the model M which consists of all possible joint distributions of these random variables. Let the probability that the random variable $X_j = i$ for $i = 0, 1$ be $P(X_j = i) = \theta_i^{(j)}$. In this example the $\theta_i^{(j)}$ are the parameters of our model. Since $(\theta_0^{(j)}, \theta_1^{(j)})$ is the probability distribution of a binary random variable we have that $\theta_1^{(j)} = 1 - \theta_0^{(j)}$ so this model actually only has 2 free parameters.

Now since X_1 and X_2 are independent, their joint distribution is given by

$$p_{i_1 i_2} = P(X_1 = i_1, X_2 = i_2) = P(X_1 = i_1)P(X_2 = i_2) = \theta_{i_1}^{(1)}\theta_{i_2}^{(2)}. \tag{1.1}$$

We can view the above equation as a polynomial map which parameterizes the model M in the following way. We will consider joint distributions as matrices of the form

$$p = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix}.$$

Let $\Delta_{n-1} = \{p \in \mathbb{R}^n \ : \ \sum_{i=1}^n p_i = 1, \ p_i \geq 0 \text{ for all } i\}$ be the standard simplex. Then the formula for the joint distribution shown in Equation 1.1 can then be viewed as the following polynomial map

$$\begin{aligned} \phi : \ & \Delta_1 \times \Delta_1 \to \Delta_3 \\ & (\theta_0^{(1)}, \theta_1^{(1)}, \theta_0^{(2)}, \theta_1^{(2)}) \mapsto \begin{pmatrix} \theta_0^{(1)} \\ \theta_1^{(1)} \end{pmatrix} \begin{pmatrix} \theta_0^{(2)}, \theta_1^{(2)} \end{pmatrix}. \end{aligned}$$

In this instance we are viewing Δ_3 as a subset of $\mathbb{R}^{2\times 2}$. More precisely, it is the set $\Delta_3 = \{p \in \mathbb{R}^{2\times 2} \mid p_{i_1 i_2} \geq 0, \ \sum_{i_1, i_2} p_{i_1 i_2} = 1\}$. The image of the polynomial map ϕ is exactly the model M which is the set of all possible joint distributions of X_1 and X_2. In

this case it is not hard to see that

$$M = \{p \in \mathbb{R}^{2\times 2} \mid p_{i_1 i_2} \geq 0,\ \sum_{i_1,i_2} p_{i_1 i_2} = 1,\ p_{00}p_{11} - p_{01}p_{10}\}.$$

This is because every $p \in M$ is an outer product of two column vectors and thus has rank at most 1.

We can also study M in a purely algebraic context. This means that we extend ϕ to a complex polynomial map

$$\begin{aligned} \phi :\ & \mathbb{C} \times \mathbb{C} \to \mathbb{C}^{2\times 2} \\ & (\theta_0^{(1)}, \theta_0^{(2)}) \mapsto \begin{pmatrix} \theta_0^{(1)} \\ 1 - \theta_0^{(1)} \end{pmatrix} \left(\theta_0^{(2)}, 1 - \theta_0^{(2)}\right). \end{aligned}$$

The vanishing ideal of the image of this map can be computed using elimination though in this case it is clearly $J = \mathcal{I}(\mathrm{im}(\phi)) = \langle p_{00} + p_{01} + p_{10} + p_{11} - 1, p_{00}p_{11} - p_{01}p_{10}\rangle$. Observe that this ideal retains much of the algebraic structure of the original model. In fact, if we consider M as a subset of $\mathbb{C}^{2\times 2}$ then the Zariski closure of M is exactly $\mathcal{V}(J)$. So studying the ideal J or its corresponding variety $\mathcal{V}(J)$ can still help us understand the underlying statistical model M.

1.2 Polytopes

In this section we introduce some basic concepts related to polytopes. The background discussed here will be used in Chapter 2. For additional information on polytopes we refer the reader to [48].

Definition 1.2.1. A set $S \subseteq \mathbb{R}^n$ is *convex* if for all $x, y \in S$ and all $\lambda \in [0,1]$ $\lambda x + (1-\lambda)y \in S$.

Given an arbitrary set $S \subseteq \mathbb{R}^n$, the smallest convex set containing S is called the *convex hull* of S and can be formulated as

$$\mathrm{conv}(S) := \{\lambda_1 x^{(1)} + \ldots \lambda_k x^{(k)} \mid x^{(k)} \in S,\ \lambda_i \geq 0,\ \sum_i \lambda_i = 1\}.$$

Convex hulls will be the main objects of study in Chapter 2 of this thesis. In particular, we will study the following types of convex sets.

Definition 1.2.2. A *polytope* is the convex hull of finitely many points in $\mathbb{R}^n$.

Example 1.2.3. Let $e_1, \ldots, e_n$ be the standard basis vectors of $\mathbb{R}^n$. Their convex hull is the standard $n-1$-dimensional simplex denoted Δ_{n-1} which is

$$\Delta_{n-1} = \{x \in \mathbb{R}^n \mid x_i \geq 0, \sum_i x_i = 1\}$$

Definition 1.2.4. Let $P \subset \mathbb{R}^n$ be a polytope. A linear inequality $c \cdot x \leq c_0$ is *valid* for P if every point in P satisfies it. A *face* of P is a set F of the form

$$F = P \cap \{x \in \mathbb{R}^n \mid c \cdot x = c_0\}$$

where $c \cdot x \leq c_0$ is a valid inequality for P.

Zero-dimensional faces are called *vertices*, one-dimensional faces are called *edges*, and codimension one faces are called facets. We denote the set of vertices of P by $\text{vert}(P)$.

Proposition 1.2.5. *Let $P \subseteq \mathbb{R}^n$ be a polytope. Then*

1. *Every polytope is the convex hull of its vertices.*
2. *If P can be written as $P = \text{conv}(S)$ for some S then $\text{vert}(P) \subseteq S$.*

This proposition tells us that if we take a polytope $P = \text{conv}(S)$, then we know the vertices are contained in S but it may often be the case that S contains extra points which are not vertices. It is then natural to ask if we can determine exactly which of the points in S are vertices and which are not. The following example illustrates this.

Example 1.2.6. Let $S = \{(0,0), (1,0), (0,1), (\frac{1}{2}, \frac{1}{2}), (\frac{1}{2}, \frac{1}{4})\}$ and let $P = \text{conv}(S)$. Then the vertices of this polytope are the points $(0,0), (1,0), (0,1)$. To see this we examine the following valid inequalities.

First note that the inequality $-x_1 - x_2 \leq 0$ for all points $x \in P$ and furthermore,

$$P \cap \{x \in \mathbb{R}^n \mid -x_1 - x_2 = 0\} = (0,0).$$

It is not hard to see that the inequalities $x_1 \leq 1$ and $x_2 \leq 1$ are also valid inequalities for P and the intersection of the corresponding equalities with P yields $(1,0)$ and $(0,1)$ respectively.

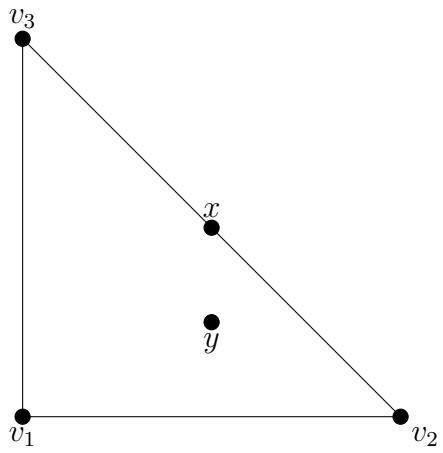


Figure 1.1: A polytope which is discussed in Example 1.2.6. The vertices of this polytope are the points labelled v_1, v_2, v_3. The other two points labelled x and y are not vertices.

On the other hand the points $(\frac{1}{2}, \frac{1}{2})$ and $(\frac{1}{2}, \frac{1}{4})$ are not vertices of P.
Denote $(0,0), (1,0), (0,1)$ with v_1, v_2, v_3 and observe that

$$\left(\frac{1}{2}, \frac{1}{2}\right) = \frac{1}{2}v_2 + \frac{1}{2}v_3$$
$$\left(\frac{1}{2}, \frac{1}{4}\right) = \frac{1}{4}v_1 + \frac{1}{2}v_2 + \frac{1}{4}v_3.$$

This implies that $P = \text{conv}((0,0), (1,0), (0,1))$ and so the other points are not vertices. This polytope is pictured in Figure 1.1.

1.3 Matroids

In this section we introduce some basic concepts from matroid theory and discuss algebraic matroids in particular. The results collected here will be the main tools that we utilize to prove identifiability results in Chapter 3. For further information on matroids we refer the reader to [39].

Definition 1.3.1. A *matroid* $\mathcal{M} = (E, \mathfrak{I})$ is a pair where E is a finite set and $\mathfrak{I} \subseteq 2^E$ satisfies

1. $\emptyset \in \mathfrak{I}$.
2. If $I' \subseteq I \in \mathfrak{I}$, then $I' \in \mathfrak{I}$.

3. If $I_1, I_2 \in \mathfrak{I}$ and $|I_2| > |I_1|$, then there exists $e \in I_2 \setminus I_1$ such that $I_1 \cup \{e\} \in \mathfrak{I}$.

The set E is called the *ground set* of $\mathcal{M}$ and the elements of $\mathfrak{I}$ are the *independent sets* of $\mathcal{M}$. There are many equivalent formulations of the axioms of a matroid but Definition 1.3.1 will be sufficient for the purpose of this thesis. We will focus on two specific types of matroids which are *linear matroids* and *algebraic matroids*.

Definition 1.3.2. Let $A \in \mathbb{K}^{d\times n}$ be a matrix with entries in a field $\mathbb{K}$ and $a_1, \dots a_n$ be the columns of A. Then letting $E = [n]$ and taking $\mathfrak{I}$ to be the subsets of E such that the corresponding columns of A are linearly independent over $\mathbb{K}$, defines a matroid. A matroid defined in this way is called a *linear matroid* over the field $\mathbb{K}$.

Example 1.3.3. Let

$$A = \begin{bmatrix} 1 & 1 & -1 & -2 \\ 3 & 1 & 2 & 4 \\ 0 & -1 & 1 & 2 \end{bmatrix}$$

and for any $S \subseteq [4]$ let A_S denote the submatrix of A obtained by taking only the columns indexed by S. A set S is an independent set in the matroid $\mathcal{M}(A)$ defined by A if and only if $\operatorname{rank}(A_S) = |S|$. In this case the independent sets of $\mathcal{M}(A)$ are

$$\emptyset, \{1\}, \{2\}, \{3\}, \{4\}, \{1,2\}, \{1,3\}, \{1,4\}, \{2,3\}, \{2,4\}, \{1,2,3\}, \{1,2,4\}.$$

Linear matroids are one of the key examples of matroids, and the name matroid itself is supposed to indicate that matroids form a generalization of this linear algebraic independence structure arising from a matrix.

Definition 1.3.4. A set $B \in \mathfrak{I}$ is called a *basis* of $\mathcal{M}$ if it is a maximal independent set with respect to inclusion.

Proposition 1.3.5. *Let B_1 and B_2 be bases of a matroid $\mathcal{M}$. Then B_1 and B_2 have the same cardinality which is called the* rank *of $\mathcal{M}$.*

Example 1.3.6. Consider again the linear matroid $\mathcal{M}(A)$ defined by the matrix

$$A = \begin{bmatrix} 1 & 1 & -1 & -2 \\ 3 & 1 & 2 & 4 \\ 0 & -1 & 1 & 2 \end{bmatrix}.$$

The bases of A are $\{1,2,3\}$ and $\{1,2,4\}$ and the rank of $\mathcal{M}(A)$ is equal to the rank of A which is 3.

There is also a way to naturally associate a matroid to an algebraic variety. All these matroids are examples of *algebraic matroids* though for practical purposes it can be more useful to think of the following geometric characterization.

Definition 1.3.7. Let $V \subset \mathbb{K}^n$ be an irreducible variety over the field $\mathbb{K}$ and for $S \subseteq [n]$ let $\pi_S : \mathbb{K}^n \to \mathbb{K}^{|S|}$ be the projection onto the coordinates in S. Let $\overline{\pi_S(V)}$ denote the Zariski closure of the projection of V. Then the pair $([n], \mathfrak{I}_V)$ defines a matroid where

$$\mathfrak{I}_V = \{S \subseteq [n] \ : \ \overline{\pi_S(V)} = \mathbb{K}^{|S|}\}$$

which is called the *coordinate projection matroid* of V and denoted by $\mathcal{M}(V)$.

The geometric perspective on algebraic matroids can also be phrased in an algebraic language.

Proposition 1.3.8. *Let $V \subset \mathbb{K}^n$ be an irreducible variety. Let $P \subseteq \mathbb{K}[p_1, \ldots, p_n]$ be the vanishing ideal of V. A set S is an independent set of the coordinate projection matroid $\mathcal{M}(V)$ if and only if*

$$P \cap \mathbb{K}[p_i : i \in S] = \langle 0 \rangle.$$

Proof. This follows directly from the fact that $P \cap \mathbb{K}[p_i : i \in S]$ is the vanishing ideal of the coordinate projection $\pi_S(V)$ and the fact that the vanishing ideal of a set is $\langle 0 \rangle$ if and only if its Zariski closure is all of space. □

Recall the more familiar definition of an algebraic matroid.

Definition 1.3.9. Let $\mathbb{L}/\mathbb{K}$ be a field extension and let $E = \{\alpha_1, \ldots, \alpha_n\} \subseteq L$. The *algebraic matroid* $(E, \mathcal{I})$ consists of all sets $S \subseteq E$ that are algebraically independent over $\mathbb{K}$.

Note that Proposition 1.3.8 shows that the coordinate projection matroid is an algebraic matroid where the field extension is $\text{Frac}(\mathbb{K}[p_1, \ldots, p_n]/P)/\mathbb{K}$ and $E = \{p_1, \ldots, p_n\}$, the images of the variables in the fraction field $\text{Frac}(\mathbb{K}[p_1, \ldots, p_n]/P)$.

When the variety V is parameterized we are able to construct the matroid $\mathcal{M}(V)$ using the Jacobian matrix of the parameterization (see [40]).

Proposition 1.3.10. *Suppose that $\phi(\theta_1, \ldots, \theta_d) = (\phi_1(\theta), \ldots, \phi_n(\theta))$ parameterizes V (that is, $V = \overline{\phi(\mathbb{K}^d)}$). Let*

$$J(\phi) := \left(\frac{\partial \phi_j}{\partial \theta_i}\right), 1 \leq i \leq d, \ 1 \leq j \leq n \tag{1.2}$$

be the transpose of the Jacobian matrix of ϕ. Then the matroid defined by the matrix $J(\phi)$ using linear independence over the fraction field Frac$(\mathbb{K}[\theta]) = \mathbb{K}(\theta)$ gives the same matroid as $\mathcal{M}(\overline{\phi(\mathbb{K}^d)})$.

Thus we have multiple different ways that we can view the same matroid which will be convenient to use at different times. For any of the above constructions that produce a matroid $\mathcal{M}$, we use the notation $\mathfrak{I}(\mathcal{M})$ to denote the set of independent sets of $\mathcal{M}$. We end this section with an example that illustrates these different versions of the same matroid.

Example 1.3.11. Let $M \subset \mathbb{P}^2$ be the model for a binomial random variable with 2 trials in projective space. This model is parameterized by the homogeneous map $\phi : \mathbb{P}^1 \to \mathbb{P}^2$ defined by $\phi_i(t, \theta) = t\binom{2}{i}\theta^i(1-\theta)^{2-i}$ for $i = 0, 1, 2$. The variable t is used to homogenize the map so that the resulting vanishing ideal is homogeneous. The transposed Jacobian is

$$J(\phi) = \begin{bmatrix} (1-\theta)^2 & 2\theta(1-\theta) & \theta^2 \\ -2t(1-\theta) & 2t(1-2\theta) & 2t\theta \end{bmatrix}.$$

Let $\mathcal{M}_\phi$ denote the corresponding matroid which has ground set $\{0, 1, 2\}$ corresponding to the columns of $J(\phi)$. The independent sets are sets $S \subseteq \{0, 1, 2\}$ such that columns in S are linearly independent over the fraction field $\mathbb{C}(t, \theta)$. One can verify through direct computation that the independent sets are exactly $S \subseteq \{0, 1, 2\}$ such that $\#S < 3$.

On the other hand, the homogeneous vanishing ideal of M is $\mathcal{I}(M) = \langle 4p_0p_2 - p_1^2 \rangle$. Its corresponding matroid, which we denote by $\mathcal{M}_{\mathcal{I}(M)}$, also has ground set $\{0, 1, 2\}$ and a set $S \subseteq \{0, 1, 2\}$ is an independent set in $\mathcal{M}_{\mathcal{I}(M)}$ if $\mathcal{I}(M) \cap \mathbb{C}[S] = \langle 0 \rangle$ where $\mathbb{C}[S] = \mathbb{C}[p_i : i \in S]$. In this case it is straightforward to see that the independent sets are again the sets S such that $\#S < 3$ and so $\mathcal{M}_\phi = \mathcal{M}_{\mathcal{I}(M)}$.

Note that, as we have done in Example 1.3.11, we will usually work with homogeneous vanishing ideals of algebraic statistical models. This has the advantage of simplifying some computations, but does not affect the underlying theory.

1.4 Trees, Networks, and Phylogenetic Models

In this section, we provide some background on phylogenetic trees, networks, and Markov models on them. The models and results discussed here will be used in chapters 3 and 4.

1.4.1 Phylogenetic Trees

In this section we review the basics of phylogenetic trees and define some structures that will be used throughout the rest of this thesis. Our terminology is adapted from [45]. For additional information on phylogenetic trees we refer the reader to [41].

Definition 1.4.1. A *tree* $T = (V, E)$ is a connected graph with no cycles. A *leaf* of T is a degree 1 vertex. T is *rooted* if it has a distinguished node, typically denoted ρ, called the *root*. A tree is *binary* if every non-leaf vertex has degree 3. A rooted tree is binary if every non-root, non-leaf vertex has degree 3, and the root has degree 2.

In phylogenetics we are typically concerned with reconstructing a tree that best represents the evolutionary history of a group of extant species which are associated with the leaves. This type of evolutionary history is typically represented by a binary tree whose leaves are labelled but internal vertices are not labelled.

Definition 1.4.2. Let X be a set of labels, $T = (V, E)$ be a (binary) tree, and $\phi : X \to V$ be an injective map whose image is exactly the set of leaves of $\mathcal{T}$. Then the pair (T, ϕ) is called a *(binary) phylogenetic X-tree*.

In this thesis we will always take the label set X to be $[n] = \{1, 2, \ldots n\}$ and focus on binary phylogenetic $[n]$-trees which we will abbreviate as $[n]$-trees.

Definition 1.4.3. A *split* of $[n]$ is a set partition $A|B$ of the set $[n]$. A split $A|B$ is valid for an unrooted binary $[n]$-tree T if it can be obtained as the leaf sets of the two connected components of $T \setminus e$ for some edge e of T. The set of all valid splits of T is denoted by $\Sigma(T)$. The *trivial splits* of T are those obtained by deleting the edges e which include a leaf of T.

When describing a tree by its splits we often omit the trivial splits associated to the leaves of T. We will also typically suppress the parentheses when writing splits so we write $123|456$ to represent the split $\{1, 2, 3\}|\{4, 5, 6\}$. The following example illustrates how the splits of a tree are determined from the tree.

Example 1.4.4. Let T be the binary phylogenetic tree pictured in Figure 1.2. T has the trivial splits $i \mid [6] \setminus i$ for all $i \in [6]$. These trivial splits are obtained by deleting the edges that contain the leaves. The nontrivial splits of T are $12|3456, 123|456$, and $1234|56$ which are induced by deleting the edges a, b, and c respectively.

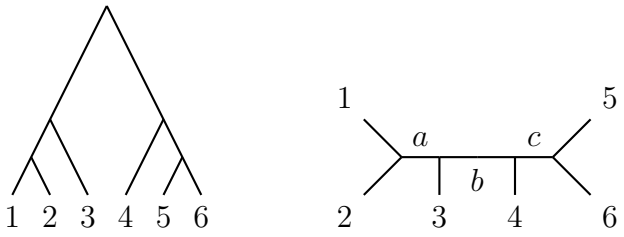


Figure 1.2: A rooted binary phylogenetic tree with 6 leaves labelled by {1, 2, 3, 4, 5, 6} is pictured on the left and an unrooted version of the same tree is pictured on the right.

Definition 1.4.5. A pair of splits $A_1|B_1$ and $A_2|B_2$ is *compatible* if at least one of the intersections

$$A_1 \cap A_2,\ A_1 \cap B_2,\ B_1 \cap A_2,\ B_1 \cap B_2$$

is empty. A set of splits Σ is *pairwise compatible* if every pair of splits is compatible.

Proposition 1.4.6. *If T is an $[n]$-tree then the set of splits $\Sigma(T)$ is pairwise compatible.*

Example 1.4.7. Again let T be the tree pictured in Figure 1.2. Then we see that

- $12|3456$ and $123|456$ are compatible since $12 \cap 456 = \emptyset$,
- $12|3456$ and $1234|56$ are compatible since $12 \cap 56 = \emptyset$,
- $123|456$ and $1234|56$ are compatible since $123 \cap 56 = \emptyset$.

Since every pair of splits is compatible, the set of splits $\Sigma(T)$ is pairwise compatible.

Theorem 1.4.8 (Splits Equivalence Theorem)**.** *Let Σ be a pairwise compatible set of splits of $[n]$. Then there exists a unique $[n]$-tree such that $\Sigma = \Sigma(T)$.*

This theorem combined with Proposition 1.4.6 tells us that $[n]$-trees and pairwise compatible sets of splits are equivalent so we will frequently move between these two ways of representing a $[n]$-tree.

1.4.2 Phylogenetic Networks

In this section we review the basics of phylogenetic networks and define some network structures that we will use throughout this thesis. Our notation and terminology is adapted from [21, 22]. For additional information on the combinatorial properties of networks and definitions associated to them we refer the reader to [22, 41].

Definition 1.4.9. A phylogenetic network N on leaf set $[n] = \{1, 2, \dots n\}$ is a rooted acyclic digraph with no multiple edges which satisfies the following properties:

1. the root has out-degree two;
2. a vertex with out-degree zero has in-degree one, and the set of vertices with out-degree zero is $[n]$;
3. all other vertices have either in-degree one and out-degree two, or in-degree two and out-degree one.

Vertices with in-degree one and out-degree two are called *tree vertices* while vertices with in-degree two and out-degree one are called *reticulation vertices*. Edges directed into a reticulation vertex are called *reticulation edges* and all other edges are called *tree edges*. In this thesis we will focus on group-based phylogenetic models which are *time-reversible*. This means that it is impossible to identify the location of the root under these models so we are only interested in the underlying *semi-directed* network structure of the phylogenetic network. The underlying semi-directed network of a phylogenetic network is obtained by suppressing the root and undirecting all tree edges in the network. The reticulation edges remain directed into the reticulation vertex though. Note that since the reticulation edges are implicitly directed into the reticulation vertex, we typically omit the arrows when drawing semi-directed networks. This is illustrated in Figure 1.3.

As the number of reticulation vertices in the network increases, the parameterization of the model becomes increasingly complicated. A common restriction is to limit the number of reticulation vertices in each biconnected component of the network. A network is called level-k if there is a maximum of k reticulation vertices in each biconnected component of the network. In this thesis we will focus on level-1 networks and a special subclass of these networks called *sunlet networks* which were first studied in [21].

Definition 1.4.10. A *n-sunlet network* is a semi-directed network with one reticulation vertex and whose underlying graph is obtained by adding a leaf to every vertex of a n-cycle. We denote with $\mathcal{S}_n$ the n-sunlet network with reticulation vertex adjacent to the leaf 1 and the other leaves labelled clockwise from 1 in increasing order.

Note that any level-1 network can be constructed by gluing sunlets of possibly different sizes along trees. It was noted in [21] that this corresponds to a toric fiber product of their ideals. We develop this further in Section 4.1. The following example corresponds to the 4-sunlet, $\mathcal{S}_4$, which we will use throughout Chapter 4

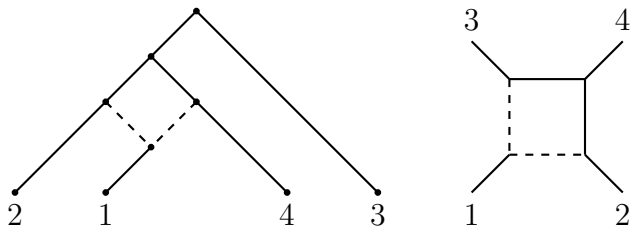


Figure 1.3: A four leaf, level-1 network pictured on the left with all edges directed away from the root. On the right is the associated semidirected network obtained by suppressing the root and undirecting all tree edges. The edges are implicitly assumed to be directed into the vertex adjacent to the leaf 1.

Example 1.4.11. Consider the network pictured on the left in Figure 1.3. This is a 4 leaf, level-1 network. The reticulation edges are dashed and the reticulation vertex is the vertex adjacent to the leaf labelled 1. Its underlying semi-directed network is pictured on the right. This semi-directed network is a 4-sunlet with reticulation vertex 1. Observe that deleting either of the reticulation edges in the sunlet network yields an unrooted binary tree with 4 leaves but that these two trees are not the same.

Another type of network which are closely related to sunlet networks are *cycle networks* which were first introduced in [21].

Definition 1.4.12. A cycle-network is a semi-directed network with one reticulation vertex. A k-cycle network is a cycle-network with cycle size k. Every k-cycle network can be built by attaching a binary tree with at least one leaf to every vertex of a k-cycle and specifying a single vertex of the cycle as the reticulation vertex.

1.4.3 Preliminaries on Phylogenetic Models

A κ-state phylogenetic Markov model on a n-leaf, leaf-labelled rooted binary tree T gives us a joint distribution on the states of the leaves of T. This joint distribution is determined by associating a κ-state random variable X_v to each internal vertex v of T and a $\kappa \times \kappa$ transition matrix M^e to each directed edge $e = (u, v)$ of T such that $M^e_{i,j} = P(X_v = j | X_u = i)$. A root distribution π for the root ρ of T is also needed. The transition matrices $\{M^e\}_{e \in E(T)}$ and the root distribution π are called the continuous parameters of the model.

We let $[\kappa]$ be the state space of these random variables and $Int(T)$ be the set of internal vertices of T. Also let X_i be the random variable associated to the leaf labelled i

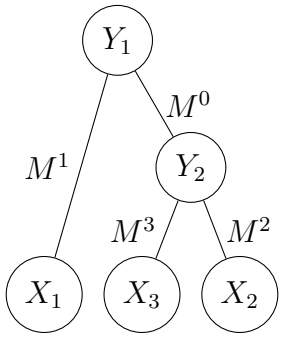


Figure 1.4: A three leaf tree with a random variable associated to each node of the tree. The matrices M^i are the transition matrices encoding the probabilities of the random variables changing states.

for $i \in [n]$. Then the probability of observing a configuration $(x_1, \dots x_n) \in [\kappa]^n$ of states at the leaves is

$$P(X_1 = x_1, \dots, X_n = x_n) \ = \sum_{j \in [\kappa]^{Int(T)}} \pi_{j_\rho} \prod_{(u,v) \in E(T)} M^{(u,v)}_{j_u, j_v}.$$

Example 1.4.13. Let T be the three leaf tree pictured in Figure 1.4. The random variables Y_1 and Y_2, which correspond to internal nodes, are hidden random variables of the model whereas the random variables X_1, X_2, X_3, which correspond to leaves, are observed.

We let M^i be transition matrices associated to each edge as pictured in Figure 1.4. The transition matrix M^i gives the probability of the random variables changing states along the corresponding edge. For instance, if we let $i, j \in [\kappa]$, then $P(X_1 = j | Y_1 = i) = M^1_{i,j}$. Lastly we choose a root distribution π to be the distribution of the random variable Y_1. Then the probability of observing $(x_1, x_2, x_3) \in [\kappa]^3$ at the leaves is

$$P(X_1 = x_1, X_2 = x_2, X_3 = x_3) \ = \sum_{(y_1, y_2) \in [\kappa]^2} \pi_{y_1} M^0_{y_1, y_2} M^1_{y_1, x_1} M^2_{y_2, x_2} M^3_{y_2, x_3}.$$

The first coordinate of $(y_1, y_2) \in [\kappa]^2$ corresponds to the root which has associated random variable Y_1. The second coordinate corresponds to the other internal vertex which has associated random variable Y_2.

We can see that the joint distribution of $(X_1, \dots X_n)$ is given by polynomials in the entries of π and the M^e. In other words, the model can be thought of as a polynomial

map

$$\psi_T : \Theta_T \to \Delta_{\kappa^n - 1}$$

where Θ_T is the *stochastic parameter space* of the model and Δ_{κ^n-1} is the probability simplex. We can also consider the variety V_T we get by taking the Zariski closure of the image of ψ_T. Polynomials in the vanishing ideal $\mathcal{I}(V_T)$ are called *phylogenetic invariants* and were first studied in [9, 30]. For more information on these models we refer the reader to [41].

With such a model, the 2-tree mixture model for trees T_1 and T_2 leaf-labelled by $[n]$ is obtained by taking the image of the map

$$\psi_{T_1,T_2} : \Theta_{T_1} \times \Theta_{T_2} \times [0,1] \to \Delta_{\kappa^n - 1}$$

defined by

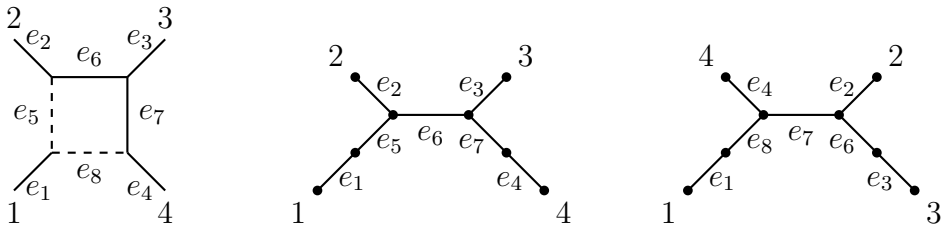
$$\psi_{T_1,T_2}(\theta_1, \theta_2, \lambda) = \lambda \psi_{T_1}(\theta_1) + (1-\lambda)\psi_{T_2}(\theta_2).$$

The 2-tree mixture model is the image of the map ψ_{T_1,T_2} but the main object of interest in this thesis is the variety naturally obtained by taking the Zariski closure of the image. Denote this variety by $V_{T_1} * V_{T_2}$ which is the *join variety* of the varieties V_{T_1} and V_{T_2}. For additional information of join varieties we refer the reader to [24].

Phylogenetic Markov models can also be extended to networks in the following way. Let N be a network with reticulation vertices $v_1, \ldots v_m$ and let e_i^0 and e_i^1 be the reticulation edges adjacent to v_i. Associate a transition matrix to each edge of N. Independently at random we delete e_i^0 with probability λ_i and otherwise delete e_i^1 and record which edge is deleted with a vector $\sigma \in \{0,1\}^m$ where $\sigma_i = 0$ indicates that edge e_i^0 was deleted. Each σ corresponds to a different tree T_σ. Then the parameterization ψ_N is given by

$$\psi_N = \sum_{\sigma \in \{0,1\}^m} \left(\prod_{i=1}^{m} \lambda_i^{1-\sigma_i} (1-\lambda_i)^{\sigma_i} \right) \psi_{T_\sigma} \tag{1.3}$$

where ψ_{T_σ} is the parameterization corresponding to the tree T_σ with transition matrices inherited from the original network N. Note that this is similar to a mixture model but with many additional relations among the parameters. The parameterization ψ_N is still a polynomial map though which means we can still consider the Zariski closure of the image ψ_N and the corresponding ideal of phylogenetic invariants, I_N. As mentioned previously, if the phylogenetic model is time-reversible then we get the same model by considering the Markov process on the underlying semi-directed network. We end this section with



(a) $\mathcal{S}_4$ (b) T_0 (c) T_1

Figure 1.5: A 4 leaf 4-cycle network N and the two trees T_0 and T_1 that are obtained by deleting the reticulation edges e_8 and e_5 respectively.

our running example.

Example 1.4.14. Consider the 4-sunlet $\mathcal{S}_4$ pictured in Figure 1.5 with reticulation vertex adjacent to the leaf 1 and reticulation edges e_5 and e_8. The trees T_0 and T_1 are obtained by deleting edges e_8 and e_5 respectively. Since there is only one reticulation vertex in $\mathcal{S}_4$, the sum in Equation 1.3 simplifies to

$$\psi_{\mathcal{S}_4} = \lambda\psi_{T_0} + (1-\lambda)\psi_{T_1}.$$

The transition matrices used in the parameterization maps ψ_{T_i} are inherited from the original network. For instance the edge e_6 in the original network has a transition matrix M^{e_6} associated to it and thus the edge e_6 that appears in T_0 and the edge e_6 that appears in T_1 both use the same transition matrix M^{e_6}.

1.4.4 Group-Based Phylogenetic Models in Fourier Coordinates

Group-based models are a family of phylogenetic Markov models where the random variables associated to each vertex take values in a finite abelian group. This allows for a linear change of coordinates in which the models are given by monomial maps.

Definition 1.4.15. Let G be a finite abelian group of order κ and T a rooted binary tree. Then a group-based model on T is a phylogenetic Markov model on T such that for each transition matrix M^e, there exists a function $f_e : G \to \mathbb{R}$ such that $M^e_{g,h} = f(g-h)$.

As mentioned above, we think of the random variables X_v as taking values in the group G, and the transition matrices as being indexed by the elements of the group. We will focus on the Cavendar-Farris-Neyman (CFN), Jukes-Cantor (JC), Kimura 2-Parameter

$$\begin{bmatrix} \alpha & \beta \\ \beta & \alpha \end{bmatrix} \qquad \begin{bmatrix} \alpha & \beta & \gamma & \delta \\ \beta & \alpha & \delta & \gamma \\ \gamma & \delta & \alpha & \beta \\ \delta & \gamma & \beta & \alpha \end{bmatrix}$$

CFN K3P

Figure 1.6: Transition matrices in the CFN and K3P models have the above forms

(K2P), and Kimura 3-Parameter (K3P) models. The CFN model is associated to the group $\mathbb{Z}_2$ while the other three models are associated to the group $\mathbb{Z}_2 \times \mathbb{Z}_2$. The form of the transition matrices for the CFN and K3P models are pictured in Figure 1.6.

Group-based models allow for a linear change of coordinates that makes ψ_T a monomial map, thus the variety V_T is a *toric variety* [42]. This change of coordinates is called the discrete Fourier transform and was first applied to phylogenetic models in [16, 27]. The new image coordinates, commonly called the Fourier coordinates, are denoted with $q_{g_1,\ldots,g_n}$ for $g_1, \ldots, g_n \in G$. This map is defined even more simply in the case that G is $\mathbb{Z}_2$ or $\mathbb{Z}_2 \times \mathbb{Z}_2$ which we will restrict to. In this case, the map can be described in terms of the *splits* of the tree which we briefly describe first.

Now for each split $A|B \in \Sigma(T)$ and each group element $g \in G$ we have a parameter $a_g^{A|B}$. The parameterization of the model ψ_T in the Fourier coordinates is given by

$$q_{g_1,\ldots g_n} = \begin{cases} \prod_{A|B \in \Sigma(T)} a^{A|B}_{\sum_{i \in A} g_i} & \text{if } \sum_{i \in [n]} g_i = 0 \\ 0 & \text{otherwise} \end{cases} \tag{1.4}$$

In the JC and the K2P models, further conditions are imposed on the parameters $a_g^{A|B}$ but in the *generic group based models*, which are the CFN and K3P models, there are no other restrictions on the parameters.

Example 1.4.16. Let T_1 be the tree pictured in Figure 3.1. The nontrivial splits of T_1 are $\{12|3456, 123|456, 1234|56\}$. Since each split is a set partition of [6] into two parts, we can just use one of the parts of the set partition to denote the parameter corresponding to that split. So the parameterization ψ_{T_1} in the Fourier coordinates will be

$$q_{g_1,\ldots g_6} = \begin{cases} a^1_{g_1} a^2_{g_2} a^3_{g_3} a^4_{g_4} a^5_{g_5} a^6_{g_6} a^{12}_{g_1+g_2} a^{123}_{g_1+g_2+g_3} a^{56}_{g_5+g_6} & \text{if } \sum_{i \in [6]} g_i = 0 \\ 0 & \text{otherwise} \end{cases}$$

The linearity of the Fourier transform allows us to also apply this change of coordinates to 2-tree mixture models and network models as well [2] which makes the map ψ_{T_1,T_2} into a binomial map. The following example illustrates this for network models.

Example 1.4.17. Let $\mathcal{S}_n$ be the 4-sunlet pictured in Figure 1.5. As we saw in the previous example, the trees T_0 and T_1 that are also pictured in Figure 1.5 are obtained from $\mathcal{S}_n$ by deleting the reticulation edges e_8 and e_5 respectively. We denote the Fourier parameter corresponding to the edge e_i and group element g_j by $a^i_{g_j}$. The parameterization $\psi_{\mathcal{S}_n}$ in the Fourier coordinates is

$$q_{g_1,g_2,g_3,g_4} = \begin{cases} a^1_{g_1}a^2_{g_2}a^3_{g_3}a^4_{g_4}a^5_{g_1}a^6_{g_1+g_2}a^7_{g_4} + a^1_{g_1}a^2_{g_2}a^3_{g_3}a^4_{g_4}a^6_{g_3}a^7_{g_1+g_4}a^8_{g_1} & \text{if } \sum_{i\in[4]} g_i = 0 \\ 0 & \text{otherwise} \end{cases}$$

The first term in the above parameterization comes from the parameterization ψ_{T_0} in the Fourier coordinates and the second term comes from ψ_{T_1}.

This new parameterization for network models is easier to work with than the previous parameterization but we can see that the ideal of a sunlet network is still not a toric ideal in the new coordinates. This means the techniques used to analyze the ideal I_T can not be directly used to analyze $I_{\mathcal{S}_n}$.

1.5 Outline of the Thesis

We will now outline the remaining chapters of this thesis.

1.5.1 Exchangeable and Sampling Consistent Distributions on Rooted Binary Trees

Chapter 2 focuses on characterizing the set of *exchangeable* and *sampling consistent* probability distributions on phylogenetic $[n]$-trees. The contents of this chapter are joint work with Seth Sullivant and come from a paper which was published in *Journal of Applied Probability* [29].

Exchangeability and sampling consistency are two properties a distribution on trees can satisfy and are desirable for biological reasons. In Chapter 1 we introduce a finite notion of sampling consistency for distributions on $[n]$-trees and show that the set of such distributions is a polytope. We then study the vertices of this polytope for small n.

Next we introduce a new family of exchangeable and sampling consistent distributions which come from doing a multinomial sample on the edges of a fixed tree. We then show that any exchangeable and sampling consistent distribution on $[n]$-trees is either a convex combinations of limits of multinomial distributions or a limit point of points in that set.

1.5.2 Identifiability in Phylogenetics using Algebraic Matroids

In Chapter 3 we develop a new method for proving that discrete parameters in parametric algebraic statistical models are identifiable which uses algebraic matroids associated to the models. The contents of this chapter are also joint work with Seth Sullivant and come from a paper that was published in *Journal of Symbolic Computation* [28].

Identifiability is a crucial property for a statistical model since it ensures that distributions in the model uniquely determine the parameters that produce them. In phylogenetics, the identifiability of the tree parameter is of particular interest since it means that phylogenetic models can be used to infer evolutionary histories from data. In this chapter we introduce a new computational strategy for proving the identifiability of discrete parameters in algebraic statistical models that uses algebraic matroids naturally associated to the models. The main idea of this algorithm is to compute independent sets in the algebraic matroid defined by the vanishing ideal of the statistical model without actually computing the ideal. This allows us to avoid time consuming Gröbner basis computations and prove identifiability results that were computationally infeasible beforehand.

In particular, we use this algorithm to prove that the tree parameters are generically identifiable for 2-tree CFN and K3P mixtures. We also show that the k-cycle phylogenetic network parameter is identifiable under the K2P and K3P models.

While the results discussed in this chapter primarily focus on phylogenetic models and are developed within the context of models for discrete random variables, our techniques work in a broader setting. Our main algorithm can actually be applied to other continuous models with finite dimensional natural parameter spaces.

1.5.3 Invariants for level-1 phylogenetic networks under the Cavendar-Farris-Neyman Model

In Chapter 4 we study the vanishing ideals of CFN level-1 phylogenetic network models. The contents of this chapter are joint work with Joseph Cummings and Chris Manon [12].

Phylogenetic networks can model more complicated evolutionary phenomena that

trees fail to capture such as horizontal gene transfer and hybridization. The same Markov models that are used to model evolution on trees can also be extended to networks and similar questions, such as the identifiability of the network parameter or the invariants of the model, can be asked. In this Chapter we focus on finding the invariants of the Cavendar-Farris-Neyman (CFN) model on level-1 phylogenetic networks. We do this by reducing the problem to finding invariants of sunlet networks, which are level-1 networks consisting of a single cycle with leaves at each vertex. We then determine all quadratic invariants in the sunlet network ideal which we conjecture generate the full ideal.

We determine the quadratic invariants by first showing that the ideal I_n associated to the n-sunlet network $\mathcal{S}_n$ is homogeneous with respect to a particualr multigrading. We then break up the ideal into graded pieces which we call *gloves* and give an explicit description of the quadratics that are in each glove.

CHAPTER

2

EXCHANGEABLE AND SAMPLING CONSISTENT DISTRIBUTIONS ON ROOTED BINARY TREES

2.1 Introduction

Leaf-labelled binary trees, which are commonly called phylogenetic trees, are frequently used to represent the evolutionary relationships between species. In this section we will restrict our attention to rooted binary trees and our label set for a tree with n leaves will always be $[n] = \{1, 2, \ldots n\}$. We call these trees $[n]$-trees and denote the set of $[n]$-trees with $\mathrm{RB}_{\mathrm{L}}(n)$.

Processes for generating random $[n]$-trees play an important role in phylogenetics. Two common examples are the uniform distribution (where a tree is chosen uniformly at random from among all trees in $\mathrm{RB}_{\mathrm{L}}(n)$) and the Yule-Harding distribution (a simple Markov branching process). Some other examples of random tree models include Aldous' β-splitting model [1], the α-splitting model [17], and the coalescent process (which generates

trees with edge lengths) [47]. Two features common to all these random tree processes and desirable for any such tree process is that they are exchangeable and sampling consistent.

Let p_n denote a probability distribution on $\mathrm{RB}_{\mathrm{L}}(n)$. *Exchangeability* refers to the fact that relabeling the leaves of the tree does not change its probability. That is, for all $T \in \mathrm{RB}_{\mathrm{L}}(n)$ and $\sigma \in S_n$, $p_n(T) = p_n(\sigma T)$. Exchangeability is a natural condition since it does not allow the names of the species to play any special role in the probability distribution. A family of distributions, $\{p_n\}_{n=2}^{\infty}$, on trees has *sampling consistency* if for each n, the distribution p_n, which is on $[n]$-trees, can be realized as the marginalization of distributions p_m, which is on $[m]$-trees, for $m > n$. That is the probability of a $[n]$-tree, T, under p_n can be written as

$$\pi_n(p_m)(T) = p_n^m(T) = \sum_{\{S \in \mathrm{RB}_{\mathrm{L}}(m) | T = S|_{[n]}\}} p_m(S).$$

Sampling consistency is a natural condition for a random tree model because it means that randomly missing species do not affect the underlying distribution on the species that were observed.

Our motivation for this study is two-fold. First of all, there has been significant work on understanding the set of exchangeable, sampling consistent distributions on other discrete objects, including rooted trees. A classic result in this theory is de Finetti's Theorem for infinitely exchangeable sequences of binary random variables which shows that every subsequence of the infinite sequence can be expressed as a mixture of independent and identically distributed sequences. This does not hold for finitely exchangeable sequences but Diaconis later developed a finite form of de Finetti's theorem. He showed that if a finite exchangeable sequence of binary random variables, $\{X_i\}_{i=1}^n$, can be extended to an exchangeable sequence, $\{X_i\}_{i=1}^m$ where $m > n$, then the original sequence can be approximated with a mixture of independent and identically distributed sequences with error $O(\frac{1}{m})$ [14]. A substantial amount of work has been done on exchangeable arrays (see [15] for example) as well, which has been used to prove de Finetti theorems for other discrete objects. For instance, Lauritzen, Rinaldo, and Sadeghi recently developed a de Finetti Theorem for exchangeable random networks [31].

There has also been considerable work characterizing exchangeable and sampling consistent distributions on trees using weighted real trees as limit objects in [18, 19, 23]. In [23] a characterization of the exchangeable and sampling consistent Markov branching models we discuss in Section 2.3.1 is obtained. A true de Finetti theorem for trees

is conjectured in [19] and proven in Theorem 3 of [18]. The approach taken in these papers is to characterize all infinitely sampling consistent distributions on trees using a limiting object called a weighted real tree. In this paper, we instead take a geometric and combinatorial approach to the study of exchangeable and finitely sampling consistent distributions on binary trees and examine what happens as we take the limit.

A second motivation comes from the combinatorial phylogenetics problem of studying properties of the distribution of the maximum agreement subtree of pairs of random trees. Let $T \in \mathrm{RB_L}(n)$ and $S \subseteq [n]$. The restriction tree $T|_S$ is the rooted binary tree with leaf label set S obtained by removing all leaves of T not in S and suppressing all vertices of degree 2 except the root. Two trees, $T_1, T_2 \in \mathrm{RB_L}(n)$, agree on a set $S \subseteq [n]$ if $T_1|_S = T_2|_S$. A maximum agreement set is an agreement set of the largest size for T_1 and T_2. The size of a maximum agreement subtree of these two trees is the cardinality of the largest subset S that T_1 and T_2 agree on and is denoted $\mathrm{MAST}(T_1, T_2)$. If S is an agreement set with $|S| = \mathrm{MAST}(T_1, T_2)$ then the resulting tree $T_1|_S = T_2|_S$ is a *maximum agreement subtree* of T_1 and T_2.

Understanding the distribution of $\mathrm{MAST}(T_1, T_2)$ for random tree distributions would help in conducting hypothesis tests that the similarity between the trees is no greater than the similarity between random trees. For example, it was suggested in [13] that $\mathrm{MAST}(T_1, T_2)$ could be used to test the hypothesis that no cospeciation occurred between a family of host species and a family of parasite species that prey on them. The study of the distribution of $\mathrm{MAST}(T_1, T_2)$ for random trees T_1, T_2 is primarily conducted with the assumption that T_1 and T_2 are drawn from an exchangeable, sampling consistent distribution on rooted binary trees. Bryant, Mackenzie, and Steel began the study of the distribution of $\mathrm{MAST}(T_1, T_2)$ and obtained some first bounds on $\mathbb{E}(\mathrm{MAST}(T_1, T_2))$ for random trees T_1 and T_2 drawn from the Uniform or Yule-Harding distributions [7]. Later work on the distribution obtained an upper bound on the order of $O(\sqrt{n})$ for $\mathbb{E}(\mathrm{MAST}(T_1, T_2))$ when T_1 and T_2 are drawn from any exchangeable, sampling consistent distribution [6]. A lower bound on the order of $\Omega(\sqrt{n})$ has been conjectured for all exchangeable, sampling consistent distributions as well but this remains an open problem. Our hope in pursuing this project is that developing a better understanding of the set of all exchangeable sampling consistent distributions might shed light on this conjecture.

In this chapter we study the structure of exchangeable, sampling consistent distributions on leaf labelled, rooted binary trees. We introduce a notion of a polytope of exchangeable and finitely sampling consistent distributions. We use it to study the set of exchangeable and sampling consistent distributions on trees and get some characterizations

for trees with a small number of leaves. We show that set of all exchangeable and sampling consistent distributions on four leaf trees come from the β-splitting model that was first introduced by Aldous in [1]. We have not been able to find a similar characterization for exchangeable and sampling consistent distributions on five leaf trees but we describe some of the vertices of the polytope of exchangeable and finitely sampling consistent distributions. Lastly, we introduce a new exchangeable and sampling consistent model on trees, called the multinomial model, and show that every sampling consistent and exchangeable distribution can be realized as a convex combination of limits of sequences of multinomial distributions.

2.2 Exchangeability and Finite Sampling Consistency

In this section we describe how the set of exchangeable distributions relates to the set of all distributions on leaf labelled, rooted binary trees. We then introduce a notion of finite sampling consistency and discuss how it relates to traditional sampling consistency.

Recall that $\mathrm{RB}_{\mathrm{L}}(n)$ denotes the set of all leaf labelled, rooted binary trees with label set $[n]$, which we call $[n]$-trees, and that $|\mathrm{RB}_{\mathrm{L}}(n)| = (2n-3)!!$. The set of all distributions on $\mathrm{RB}_{\mathrm{L}}(n)$ is the probability simplex $\Delta_{(2n-3)!!-1} \subseteq \mathbb{R}^{(2n-3)!!}$ where the coordinates are indexed by $[n]$-trees. The symmetric group S_n denotes the group of permutations of $[n]$. For each $\sigma \in S_n$ and $T \in \mathrm{RB}_{\mathrm{L}}(n)$ let σT denote the tree obtained by applying σ to the leaf labels.

Definition 2.2.1. A distribution p on $\mathrm{RB}_{\mathrm{L}}(n)$ is *exchangeable* if for all permutations $\sigma \in S_n$ and $[n]$-trees $T \in \mathrm{RB}_{\mathrm{L}}(n)$, $p(T) = p(\sigma T)$. The set of all exchangeable distributions on $\mathrm{RB}_{\mathrm{L}}(n)$ is denoted $\mathcal{E}_n$.

As previously mentioned, exchangeability requires that the probability of a $[n]$-tree under a particular distribution depend only on the shape of the tree. Thus we only need to consider distributions on the set of tree shapes. Let $\mathrm{RB}_{\mathrm{U}}(n)$ denote the set of unlabelled rooted binary trees, which we may also call trees or tree shapes. This idea is summarized in the next lemma which is the $[n]$-tree analogue of Lemma 2 in [31].

Lemma 2.2.2. *The set of exchangeable distributions on* $\mathrm{RB}_{\mathrm{L}}(n)$, $\mathcal{E}_n$, *is a simplex of dimension* $|\mathrm{RB}_{\mathrm{U}}(n)| - 1$ *with coordinates indexed by tree shapes.*

Proof. First we define a distribution $p_T \in \mathcal{E}_n$ for each tree shape $T \in \mathrm{RB}_{\mathrm{U}}(n)$. To do so, we let $O(T)$ be the set of trees $T' \in \mathrm{RB}_{\mathrm{L}}(n)$ such that $\mathrm{shape}(T') = T$. For any tree

$S \in \mathrm{RB}_{\mathrm{L}}(n)$ we set

$$p_T(S) = \begin{cases} \frac{1}{|O(T)|} & \text{shape}(S) = T \\ 0 & \text{shape}(S) \neq T. \end{cases}$$

Then $p_T \in \mathcal{E}_n$ since it is a probability distribution on trees and all trees of the same shape have the same probability. We claim that $\mathcal{E}_n = \text{conv}\left(\{p_T : T \in \mathrm{RB}_{\mathrm{U}}(n)\}\right)$, where $\text{conv}(A)$ denotes the convex hull of the set A. Since $p_T \in \mathcal{E}_n$ for all $T \in \mathrm{RB}_{\mathrm{U}}(n)$, it is enough to show that any distribution $p \in \mathcal{E}_n$ can be written as a convex combination of the p_T. If $p \in \mathcal{E}_n$, then the probability of any tree $T' \in \mathrm{RB}_{\mathrm{L}}(n)$ depends only on the shape of T' not the leaf labelling so we can write

$$p = \sum_{T \in \mathrm{RB}_{\mathrm{U}}(n)} p(T) p_T$$

where $p(T)$ is $|O(T)|$ times the probability of any $[n]$-tree in $\mathrm{RB}_{\mathrm{L}}(n)$ with shape T. Since the original p is a probability distribution on all leaf labelled trees the weights in the linear combination are nonnegative and sum to 1. Lastly we note that the vectors p_T are affinely independent since there is no overlap of coordinate indices where the entries in p_T are nonzero. So $\mathcal{E}_n = \text{conv}\left(\{p_T : T \in \mathrm{RB}_{\mathrm{U}}(n)\}\right)$ is a simplex and has coordinates indexed by $\mathrm{RB}_{\mathrm{U}}(n)$. □

Lemma 2.2.2 allows us to move from studying exchangeable distributions on leaf labelled $[n]$-trees to all distributions on unlabelled trees. We will primarily focus on understanding the set of sampling consistent distributions within $\mathcal{E}_n$ now. First recall that for $p_m \in \mathcal{E}_m$ the marginalization or projection map π_n gives a new distribution p_n^m on $\mathrm{RB}_{\mathrm{L}}(n)$ for $n < m$, defined for all $T \in \mathrm{RB}_{\mathrm{L}}(n)$ by

$$\pi_n(p_m)(T) = \sum_{\{S \in \mathrm{RB}_{\mathrm{L}}(m) | T = S|_{[n]}\}} p_m(S)$$

We will use this marginalization map to define a notion of finite sampling consistency.

Definition 2.2.3. A family of distributions $\{p_k\}_{k=n}^m$ is *finitely sampling consistent* or *m-sampling consistent*, if for each $n \leq k < m$, $p_k = \pi_k(p_m)$. We denote the set of all distributions in $\mathcal{E}_n$ that are m-sampling consistent by

$$\mathcal{E}_n^m = \pi_n(\mathcal{E}_m).$$

It is immediate that if a distribution in $\mathcal{E}_n$ is m-sampling consistent, then for any k, such that $n < k < m$, the distribution is also k-sampling consistent. This leads to the following:

Lemma 2.2.4. *For all $m > k > n$,*

$$\mathcal{E}_n^m \subseteq \mathcal{E}_n^k.$$

A distribution in $\mathcal{E}_n$ is sampling consistent if it is part of a m-sampling consistent family of distributions for all $m > n$. In other words, a distribution is sampling consistent if it is in $\mathcal{E}_n^m$ for all $m > n$. Thus we can define the following notation for the set of exchangeable distributions on $\mathrm{RB_L}(n)$ that are sampling consistent:

$$\mathcal{E}_n^\infty := \cap_{m=n}^\infty \mathcal{E}_n^m.$$

Lemma 2.2.5. *Let $p_T \in \mathcal{E}_m$ be defined as it is in Lemma 2.2.2, then*

$$\mathcal{E}_n^m = \mathrm{conv}\left(\{\pi_n(p_T) : T \in \mathrm{RB_U}(m)\}\right).$$

Proof. Clearly it holds that $\mathrm{conv}(\{\pi_n(p_T) : T \in \mathrm{RB_U}(m)\}) \subseteq \mathcal{E}_n^m$ since $\pi_n(p_T) \in \mathcal{E}_n^m$ for all $T \in \mathrm{RB_U}(m)$. It is enough to show that if we have a distribution $p_n^m \in \mathcal{E}_n^m$, then it can be written as a convex combination of the $\pi_n(p_T)$. If $p_n^m \in \mathcal{E}_n^m$, then there exists $p_m \in \mathcal{E}_m$ such that $\pi_n(p_m) = p_n^m$. Since $p_m \in \mathcal{E}_m$, we know from Lemma 2.2 that we can write $p_m = \sum_{T \in \mathrm{RB_U}(n)} p_m(T) \cdot p_T$. Then evaluating $\pi_n(p_m)$ at a $[n]$-tree $S \in \mathrm{RB_L}(n)$ gives

$$\pi_n(p_m)(S) = \sum_{\{Q \in \mathrm{RB_L}(m) | S = Q|_{[n]}\}} \sum_{T \in \mathrm{RB_U}(m)} p_m(T) p_T(Q)$$

Changing the order of summation we have

$$\pi_n(p_m)(S) = \sum_{T \in \mathrm{RB_U}(m)} p_m(T) \sum_{\{Q \in \mathrm{RB_L}(m) | S = Q|_{[n]}\}} p_T(Q)$$

but $\sum_{\{Q \in \mathrm{RB_L}(m) | S = Q|_{[n]}\}} p_T(Q) = \pi_n(p_T)(S)$ so we get that

$$\pi_n(p_m)(S) = \sum_{T \in \mathrm{RB_U}(m)} p_m(T)(\pi_n(p_T)(S))$$

which shows that $p_n^m = \pi_n(p_m)$ can be written as a convex combination of the $\pi_n(p_T)$. □

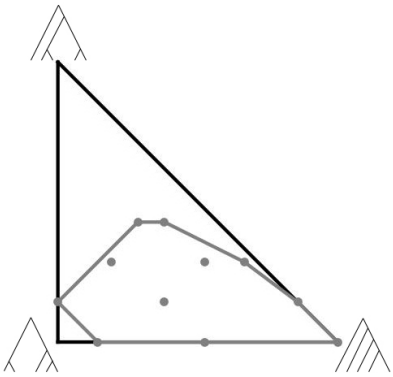


Figure 2.1: The projection of $\mathcal{E}_5^7$ onto the first two coordinates of the simplex $\mathcal{E}_5$. The gray points correspond to the points $\pi_n(p_T)$ for $T \in \mathrm{RB_U}(7)$. The vertices of the simplex are labelled with the corresponding unrooted tree. Note that the balanced tree is at the origin since we've projected onto the coordinates corresponding to the other two trees.

Example 2.2.6. While it is the case that $\mathcal{E}_n^m = \mathrm{conv}(\{\pi_n(p_T) : T \in \mathrm{RB_U}(m)\})$, not every $\pi_n(p_T)$ will be a vertex of $\mathcal{E}_n^m$. Figure 2.1 illustrates this.

Lemma 2.2.5 implies that understanding how the marginalization map acts on the vertices of $\mathcal{E}_m$ will allow us to compute all of $\mathcal{E}_n^m$. The following lemma and corollary will give us a method for calculating the vertices of $\mathcal{E}_n^m$ by computing subtree densities.

Lemma 2.2.7. *Let $S \in \mathrm{RB_L}(n)$ and $T \in \mathrm{RB_U}(m)$. Also let $c_T(S) = |\{Q \in \mathrm{RB_L}(m) | S = Q|_{[n]}, \mathrm{shape}(Q) = T\}|$. Then $\pi_n(p_T)(S) = \frac{c_T(S)}{|O(T)|}$.*

Proof. By definition of the map π_n

$$\pi_n(p_T)(S) = \sum_{\{Q \in \mathrm{RB_L}(m) | S = Q|_{[n]}\}} p_T(Q)$$

but $p_T(Q)$ is nonzero if and only if $\mathrm{shape}(Q) = T$, in which case it is $\frac{1}{|O(T)|}$. So the above sum becomes

$$\pi_n(p_T)(S) = \sum_{\{Q \in \mathrm{RB_L}(m) | S = Q|_{[n]}, \mathrm{shape}(Q) = T\}} \frac{1}{|O(T)|} = \frac{c_T(S)}{|O(T)|}.$$

□

Corollary 2.2.8. *Let $S' \in \mathrm{RB_U}(n)$ and $T \in \mathrm{RB_U}(m)$. Then $\pi_n(p_T)(S')$, which is used to denote the sum of $\pi_n(p_T)(S)$ over all $S \in O(S')$, is the induced subtree density of S' in T.*

That is, for any fixed $Q \in O(T)$

$$\pi_n(p_T)(S') = \frac{|\{I \subseteq [m] : |I| = n \text{ and } \text{shape}(Q|_I) = S'\}|}{\binom{m}{n}}.$$

Proof. From the previous lemma, we know that for any $S \in O(S')$, $\pi_n(p_T)(S) = \frac{c_T(S)}{|O(T)|}$ where $c_T(S) = |\{Q \in \text{RB}_\text{L}(m) | S = Q|_{[n]}, \text{shape}(Q) = T\}|$. Then we have

$$\pi_n(p_T)(S') = \sum_{S \in O(S')} \frac{c_T(S)}{|O(T)|}$$

So for each labelling S of S', we are counting which fraction of labellings of T yield S when restricted to $[n]$. As we sum over all labellings of S, this gives us the total fraction of times that the shape S' appears as a restriction tree of the shape T when $(n-m)$ of its leaves are marginalized out which is exactly

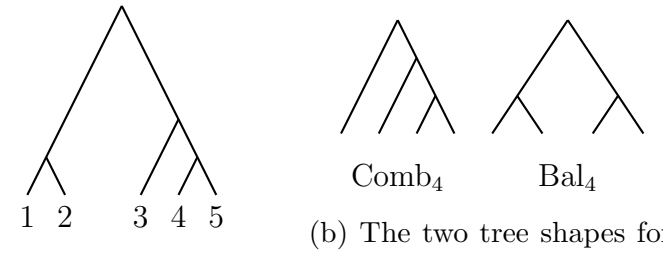
$$\frac{|\{I \subseteq [m] : |I| = n \text{ and } \text{shape}(Q|_I) = S'\}|}{\binom{m}{n}}.$$

□

The following examples elucidates what is meant by induced subtree density and shows how we can explicitly calculate this quantity.

Example 2.2.9. We show how to find the projection of one vertex of $\mathcal{E}_5$ down to $\mathcal{E}_4$. $\mathcal{E}_4^5$ is the convex hull of the projection of all of the vertices of $\mathcal{E}_5$. Begin with the tree shape T pictured in Figure 2.2a. We label the leaves of T for the sake of the calculation but it should be thought of as an unlabelled tree. We then find the shape of the restriction $\{1,2,3,4\}, \{1,2,3,5\}, \{1,2,4,5\}$, gives the shape Bal_4 and the restriction to the sets $\{1,3,4,5\}, \{2,3,4,5\}$ gives the shape Comb_4, pictured in Figure 2.2b. We let the first coordinate of $\mathcal{E}_4$ be the probability of obtaining Comb_4 and the second be the probability of obtaining Bal_4. As mentioned above, these probabilities will simply be the number of times each shape appears as a restriction tree over the total number of restriction trees. Thus this vertex of $\mathcal{E}_5$ will give us the distribution $(2/5, 3/5)$ in $\mathcal{E}_4$.

We have now seen how to compute the vertices of $\mathcal{E}_n^m$ explicitly but not every distribution $\pi_n(p_T)$ is a vertex of $\mathcal{E}_n^m$. However, the comb tree always yields a vertex of $\mathcal{E}_n^m$.



(a) A five leaf tree

(b) The two tree shapes for four leaf trees

Figure 2.2

Lemma 2.2.10. *For all $m \geq n$, let $\mathrm{Comb}_m \in \mathrm{RB}_{\mathrm{U}}(m)$ be the m-leaf comb tree, then $\pi_n(p_{\mathrm{Comb}_m})$ is a vertex in $\mathcal{E}_n^m$.*

Proof. The comb tree has only smaller comb trees as restriction trees, so the image of the comb distribution on m leaves under the marginalization map will be the comb distribution on n leaves. Since p_{Comb_n} is a vertex of $\mathcal{E}_n$ and $\mathcal{E}_n^m$ is a subset of $\mathcal{E}_n$, then p_{Comb_n} is also a vertex of $\mathcal{E}_n^m$. □

2.3 Examples of Exchangeable and Sampling Consistent distributions

In this section we discuss some of the well-known exchangeable and sampling consistent families of distributions particularly, the Markov branching models. We also introduce a new family of exchangeable sampling consistent tree distributions, namely the multinomial family.

2.3.1 Markov Branching Models

An important example of sampling consistent and exchangeable distributions are the families of Markov branching models which can be constructed in the following way as first introduced in [1] by Aldous.

Suppose that for every integer $n \geq 2$, we have a probability distribution on $\{1, 2, \ldots, n-1\}$, $q_n = (q_n(i) : i = 1, 2, \ldots n-1)$ which satisfies $q_n(i) = q_n(n-i)$. Using this family of distributions we can define a probability distribution on $\mathrm{RB}_{\mathrm{U}}(n)$ by taking the probability that i leaves fall on the left of the root-split and $n-i$ leaves fall on the right of the root-split to be $q_n(i)$ with each choice of i labels to fall on the left having

the same probability. Repeating recursively in each branch will yield the probability of a rooted binary tree. Aldous called these models Markov branching models.

Haas et al. classified the sampling consistent Markov branching models on rooted binary trees in [23]. They show that every sampling consistent Markov branching model, defined by the splitting rules q_n, $n \geq 2$, has an integral representation of the form

$$q_n(i) = a_n^{-1}\left(\binom{n}{i}\int_0^1 x^i(1-x)^{n-i}\nu(dx) + nc1_{i=1}\right) \tag{2.1}$$

where $c \geq 0$, ν is a symmetric measure on $(0,1)$ such that $\int_0^1 x(1-x)\nu(dx) < \infty$, and a_n is a normalization constant. $c1_{i=1}$ accounts for the comb distribution. A subclass of these models are those where the measure ν in equation (2.1) has the form $\nu(dx) = f(x)dx$ for a probability density function f on $(0,1)$ that is symmetric on the interval (i.e. $f(x) = f(1-x)$) and where $c = 0$. These Markov branching models can be thought of as uniformly choosing n points in the interval $(0,1)$ at random and then splitting the interval with respect to the density f. Repeating the splitting process recursively in each subinterval until each of the original n points is contained in its own subinterval gives a tree shape. This process is pictured in Figure 6 in [1].

One particularly important family of Markov branching distributions is the beta-splitting model. It is a Markov branching model that belongs to the subclass mentioned above where the function f in the above description has the form

$$f(x) = \frac{\Gamma(2\beta+2)}{\Gamma^2(\beta+1)}x^\beta(1-x)^\beta$$

for $-1 < \beta < \infty$. For the beta-splitting model we can calculate the values $q_n(i)$ explicitly in terms of β. By plugging in the beta-splitting density function f into (2.1) for $q_n(i)$ we get the following formulas:

$$q_n(i) = a_n^{-1}\binom{n}{i}\frac{\Gamma(\beta+i+1)\Gamma(\beta+n-i+1)\Gamma(2\beta+2)}{\Gamma(2\beta+n+2)\Gamma^2(\beta+1)} \tag{2.2}$$

for $-1 < \beta < \infty$. Note that Equation (2.2) can be analytically continued on $-2 < \beta \leq -1$ and so it is natural to extend the beta-splitting model to those values of β. As β approaches -2 the beta-splitting model approaches the distribution which puts all probability on the comb tree, so we also include $\beta = -2$ in the beta splitting model as the comb distribution.

An important note here is that for the beta-splitting model each $q_n(i)$ is actually a

rational function in β. Using properties of the gamma function one can see that the above formula simplifies to

$$q_n(i) = \frac{\binom{n}{i}(i+\beta)_i(n-i+\beta)_{n-i}}{(n+2\beta+1)_n - 2(n+\beta)_n}$$

Since each $q_n(i)$ is a rational function in β, we can see that the probability of obtaining a certain tree shape is a rational function in β as well because the probability of obtaining that tree shape under the beta-splitting model is simply the product of the probability of all of the splits in the tree.

Example 2.3.1. Let Comb_4 and Bal_4 be the trees pictured in Figure 2.2b. Then the probabilities of obtaining them under the beta-splitting model are

$$p(\text{Comb}_4) = 2q_4(1) = \frac{12+4\beta}{18+7\beta}$$
$$p(\text{Bal}_4) = q_4(2) = \frac{6+3\beta}{18+7\beta}$$

This model also has a nice characterization among all of the sampling consistent Markov branching models. In [38], McCullagh, Pitman, and Winkel show that the beta-splitting models are the only sampling consistent Markov branching models whose splitting rules admit a particular factorization.

We are interested in examining how the sampling consistent Markov branching models and in particular the beta-splitting model fits inside inside of $\mathcal{E}_n$ as a whole. These distributions are infinitely sampling consistent and so lie in $\mathcal{E}_n^\infty$ as well. A priori, it might seem that to determine the probability of a tree shape with n leaves under a Markov branching model that one would need to have not only the distribution q_n but also distributions q_k where $2 \leq k \leq n-1$. This is actually not the case for any sampling consistent Markov branching model though. Ford showed in Proposition 41 of [17] that if $(q_k | 2 \leq k \leq n)$ are the splitting rules for a distribution in $\mathcal{E}_n^\infty$, then in fact it must be that

$$q_{n-1}(i) = \frac{(n-i)q_n(i) + (i+1)q_n(i+1)}{n - 2q_n(1)} \tag{2.3}$$

This implies that all that is needed to define a distribution in $\mathcal{E}_n^\infty$ is the first splitting rule q_n which gives the following corollary.

Corollary 2.3.2. *The dimension of the set of all sampling consistent Markov branching models in $\mathcal{E}_n$ is at most $\lceil \frac{n-1}{2} \rceil - 1$*

Proof. As explained above, a Markov branching model is completely determined by the distribution $q_n = (q_n(i) : i = 1, 2, \dots n-1)$ which determines all of the distributions $q_k = (q_k(i) : i = 1, 2, \dots k-1)$ where $2 \leq k \leq n-2$. Since q_n must be symmetric we immediately get that the values $q_n(1), q_n(2), \dots, q_n(\lceil \frac{n-1}{2} \rceil)$ determine all of q_n. Also since q_n must be a distribution we lose one of these as a free parameter, thus the dimension of the set of sampling consistent Markov branching models is bounded above by $(\lceil \frac{n-1}{2} \rceil - 1)$. □

Note that when $n = 4$, the space of sampling consistent Markov branching models has dimension 1. We will see in Section 2.4 that the set of beta-splitting models is equal to the set of sampling consistent Markov branching models in this case.

2.3.2 Multinomial model

The multinomial model associates to each tree shape $T \in \mathrm{RB}_{\mathrm{U}}(m)$ for any $m \geq 2$ a family of probability distributions on $\mathrm{RB}_{\mathrm{L}}(n)$ for each n. We first add an extra leaf to the root of T to obtain a new tree which we denote by $\tilde{T}$. We then associate to every edge, e, in $\tilde{T}$ a parameter $t_e \geq 0$. This gives us a vector of parameters $t = (t_e | e \in E(\tilde{T}))$ of length $2m-1$, and we assume that $\sum_e t_e = 1$, so that these parameters give a probability distribution on the edges of $\tilde{T}$. We will now use this probability distribution to define a set of distributions on $\mathrm{RB}_{\mathrm{U}}(n)$ for any $n \geq 2$. Note that n and m do not have to be related to each other.

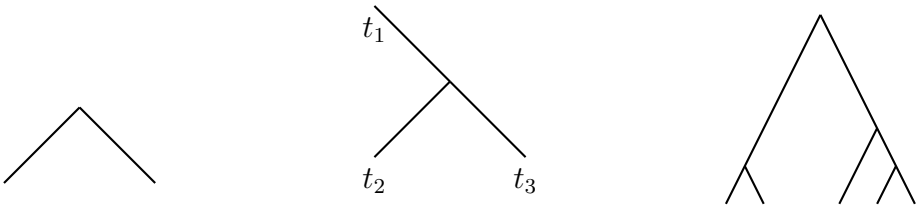
Using the distribution t, we draw a multiset A of edges from the tree $\tilde{T}$, where edge e occurs with probability t_e. There is a natural way to take the tree $\tilde{T}$ and a multiset A of size n on the set of parameters and construct a new tree which we will call $\tilde{T}_A \in \mathrm{RB}_{\mathrm{U}}(n)$. Each time that an edge e appears in A, we add a new leaf to the edge e, which will give us a new tree with $m + n$ leaves. We then simply take $\tilde{T}_A$ to be the induced subtree on only the leaves that come from A. Hence, the multinomial model on the tree T gives a way to produce random trees with an underlying skeleton that is the tree T. For large n, the resulting random trees look like T with many extra leaves added.

The multinomial probability of observing a particular multiset of edges A is the monomial

$$p_A = \binom{n}{m_A} \prod_{e \in \tilde{T}} t_e^{m_A(e)}$$

where $m_A(e)$ denotes the number of times that e appears in the multiset A, and m_A is the resulting vector.

Letting $M_n^{\tilde{T}}$ be the set of all n element multisets of edges of $\tilde{T}$, we can calculate the



(a) T (b) $\tilde{T}$ (c) Bal_5

Figure 2.3

probability of observing any particular tree shape S by

$$p_{\tilde{T},t}(S) = \sum_{\substack{A \in M_n^{\tilde{T}} \\ \tilde{T}_A = S}} p_A.$$

Example 2.3.3. Consider the tree $\tilde{T}$ from Figure 2.3b with edge parameters (t_1, t_2, t_3). To calculate the probability of the tree, Bal_5, in Figure 2.3c we use the formula

$$p_{\tilde{T},t}(\mathrm{Bal}_5) = \sum_{\substack{A \in M_5^{\tilde{T}} \\ \tilde{T}_A = \mathrm{Bal}_5}} p_A.$$

The only multisets that satisfy this condition are the sets $A_1 = \{2, 2, 2, 3, 3\}$ and $A_2 = \{2, 2, 3, 3, 3\}$. This is because if 1 appears in a multiset A any positive number of times, the tree $\tilde{T}_A$ will have a single leaf on one side of the root and four leaves on the other side, regardless of what other parameters appear in the set. So A_1 and A_2 are the only elements of $M_5^{\tilde{T}}$ that we sum over so

$$p_{\tilde{T},t}(\mathrm{Bal}_5) = \binom{5}{3,2} t_2^3 t_3^2 + \binom{5}{2,3} t_2^2 t_3^3$$

The multinomial model gives a family of distributions as we let the parameter vector t range over the entire simplex. Equivalently, the model can be described as the image of the simplex under the polynomial map

$$p_{\tilde{T}} : \Delta_{|E(\tilde{T})|-1} \to \mathcal{E}_n^\infty$$

where the coordinate corresponding to $S \in \mathrm{RB}_{\mathrm{U}}(n)$ has value $p_{\tilde{T},t}(S)$ for $t \in \Delta_{2m-2}$. Since Δ_{2m-2} is a semialgebraic set and $p_{\tilde{T}}$ is a polynomial map, the multinomial model is also a

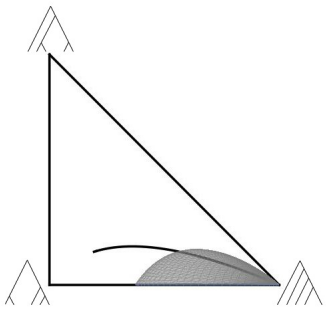


Figure 2.4: This is a projection onto the first two coordinates of the simplex $\mathcal{E}_5$. The beta-splitting model on $\mathrm{RB}_{\mathrm{U}}(5)$ is pictured in black and the multinomial model on the two leaf tree is pictured in gray.

semialgebraic set.

It also holds that if we take any tree $T \in \mathrm{RB}_{\mathrm{U}}(m)$, and any subtree $T' \in \mathrm{RB}_{\mathrm{U}}(m')$ of T, then we have that $\mathrm{Im}(p_{\tilde{T}'}) \subseteq \mathrm{Im}(p_{\tilde{T}})$. This is because if the parameters corresponding to edges that appear in T but not in T' are set to 0 in p_T, the map will simply become $p_{T'}$. Setting these parameters to 0 just corresponds to restricting p_T to a subset of the simplex and thus we get the image containment.

A last interesting note is that this model is perhaps similar in spirit to the W-random graphs when W is a graphon obtained from a finite graph G as described in [34]. The construction begins with a finite graph G and uses it to define a distribution on graphs with k vertices similarly to how we begin with a tree T and define a distribution on trees with k leaves.

We end this section with Figure 2.4, which shows both the beta-splitting model and the multinomial model inside $\mathcal{E}_5$. In the next section we will discuss the exchangeable and sampling consistent distributions on four leaf trees and how they relate to the models discussed in this section.

2.4 Distributions in $\mathcal{E}_4^\infty$

In this section we classify all of the distributions in $\mathcal{E}_4^\infty$. In particular, we show that $\mathcal{E}_4^\infty$ is equal to the beta-splitting model.

First we note that since there are only two distinct tree shapes with four leaves (see Figure 2.2b), the set of exchangeable distributions is just a 1-dimensional simplex Δ_1 in $\mathbb{R}^2$. We take coordinates (p_1, p_2) on $\mathbb{R}^2$ and let the first coordinate correspond to Comb_4

and the second coordinate to Bal_4. The subset of distributions that are also sampling consistent must be some line segment within the simplex. We know from Lemma 2.2.10 that the comb distribution, which is $(1,0)$ in these coordinates, is a vertex in $\mathcal{E}_4^\infty$. If we can bound the probability of obtaining Bal_4 then we will have a complete characterization of all distributions in $\mathcal{E}_4^\infty$. Theorem 2 in [11] will be the main tool to achieve this.

Theorem 2.4.1. *[11, Thm 2] The most balanced tree in* $\text{RB}_\text{U}(n)$ *has the complete symmetric tree on four leaves appear more frequently as a subtree than any other tree in* $\text{RB}_\text{U}(n)$.

By the *most balanced tree* in $\text{RB}_\text{U}(n)$, we mean the unique tree shape in $\text{RB}_\text{U}(n)$ that has the property that for any internal vertex of the tree, the number of leaves on the left and right subtrees below that differ by at most one.

Theorem 2.4.2. *The four leaf beta-splitting model equals the set of all exchangeable and sampling consistent distributions on* $\text{RB}_\text{U}(4)$.

Proof. Note that $\mathcal{E}_4^n$ only has two vertices since it is a line segment. The comb distribution $(1,0)$ is always a vertex in $\mathcal{E}_4^n$, by Lemma 2.2.10. The other vertex will be the projection of the vertex of $\mathcal{E}_n$ that places the most mass on Bal_4. The projection of a vertex $p_T \in \mathcal{E}_n$, is $(p_1, p_2) = \frac{1}{\binom{n}{4}}(m_1, m_2)$ where m_1 is the number of 4 element subsets $S \subset [n]$ such that $T|_S = \text{Comb}_4$ and m_2 is the number of 4 element subsets $S \subset [n]$ such that $T|_S = \text{Bal}_4$. By Theorem 2.4.1 we can restrict to the most balanced tree in $\text{RB}_\text{U}(n)$. We will use $m_{2,n}$ to denote this highest value of m_2 that we get from the most balanced tree in $\text{RB}_\text{U}(n)$.

The beta-splitting model on $\text{RB}_\text{U}(4)$, on the other hand, is the line segment from $(1,0)$ to $(\frac{4}{7}, \frac{3}{7})$. Indeed, under the beta splitting model, the probability of Bal_4 is just

$$\begin{aligned} q_4(2) &= \frac{\binom{4}{2}(\beta+2)_2^2}{(2\beta+5)_4 - 2(\beta+4)_4} \\ &= \frac{6\beta^4 + O(\beta^3)}{14\beta^4 + O(\beta^3)}. \end{aligned}$$

As $\beta \to \infty$, this converges to $\frac{3}{7}$. So if we can show that $\lim_{n\to} \frac{m_{2,n}}{\binom{n}{4}} = \frac{3}{7}$ then we will be done.

To prove that $\lim_{n\to} \frac{m_{2,n}}{\binom{n}{4}} = \frac{3}{7}$, we can restrict to the subsequence of values $n = 2^k$, since Lemma 2.2.4 implies that $\frac{m_{2,n}}{\binom{n}{4}}$ is a monotone decreasing sequence. This subsequence is easier to deal with since $m_{2,2^n}$ counts the number of 4-subsets, $S \subset [2^n]$ of the leaves of

the complete symmetric tree T_{2^n} in $\mathrm{RB}_{\mathrm{U}}(2^n)$ such that $T_{2^n}|_S = \mathrm{Bal}_4$. Using the recursive structure of T_{2^n}, we see that $m_{2,2^n} = 2m_{2,2^{n-1}} + \binom{2^{n-1}}{2}^2$. The only ways we can choose a subset S such that $T_{2^n}|_S = \mathrm{Bal}_4$ are that the leaves in S fall either entirely within the left or right subtrees or that S has two leaves from both the left and right subtrees. The number of ways to choose a subset S that falls entirely on the left or right side is $m_{2,2^{n-1}}$ by definition. The number of ways to choose two leaves from each side is $\binom{2^{n-1}}{2}^2$. This recurrence can be solved to find an explicit formula for $m_{2,2^n}$ which is

$$m_{2,2^n} = \sum_{i=1}^{n-1} 2^{n-i-1} \binom{2^i}{2}^2$$

Now we can simplify $\frac{m_{2,2^n}}{\binom{2^n}{4}}$ to get

$$\frac{m_{2,2^n}}{\binom{2^n}{4}} = \frac{3(2^n) - 5}{7(2^n) - 21}$$

which converges to $\frac{3}{7}$ as n tends to infinity. □

Note that Theorem 2.4.2 does not generalize to higher dimensions as the set of beta splitting distributions is of strictly smaller dimension than the set of exchangeable sampling consistent distributions. We explore the discrepancy between these sets in more detail in the next sections.

2.5 Distributions on $\mathcal{E}_5^\infty$

There are three distinct tree shapes with five leaves so $\mathcal{E}_5$ is a 2-dimensional simplex in $\mathbb{R}^3$. For the rest of this section we will use Comb_5, Gir_5, and Bal_5 to represent the trees pictured in Figure 2.5. Specifically, let Comb_5 denote the comb tree on five leaves, Bal_5 denote the balanced tree on five leaves and Gir_5 denote the giraffe tree on five leaves. We take coordinates (p_1, p_2, p_3) on $\mathbb{R}^3$ where p_1, p_2, p_3 represent the probability of obtaining Comb_5, Gir_5, and Bal_5, respectively.

While we have not been able to give a complete description of the vertices of $\mathcal{E}_5^n$ for all n, we are able to define some tree structures in $\mathrm{RB}_{\mathrm{U}}(n)$ that do yield vertices of $\mathcal{E}_5^n$. We have already seen that the comb tree Comb_m always yields a vertex of $\mathcal{E}_n^m$ for all m and n. Here we provide some other examples.

Definition 2.5.1. For a tree $T \in \mathrm{RB}_{\mathrm{U}}(m)$ let $\mathrm{comb}(T, n)$ be the tree that is obtained

Figure 2.5: Tree shapes on five leaves

by creating a comb tree with n leaves and replacing one of the two leaves at the deepest level with the tree T.

Generally, if $T \in \mathrm{RB_U}(m)$ then $\mathrm{comb}(T, n)$ has $m + n - 1$ vertices. For example, $\mathrm{Gir}_5 = \mathrm{comb}(\mathrm{Bal}_4, 2)$. Note that does not matter which of the leaves is replaced with T since our trees are unlabelled.

Proposition 2.5.2. *Let $T_n = \mathrm{comb}(\mathrm{Gir}_5, n-4)$. Then $\pi_5(p_{T_n})$ is a vertex in $\mathcal{E}_5^n$.*

Proof. First note that T_n and Comb_n are the only trees with n leaves that do not have Bal_5 as a subtree. This means T_n and Comb_n are the only tree shapes $T \in RB_U(n)$ such that $\pi_5(p_T)$ fall on the line $p_3 = 0$ so the segment $[\pi_5(p_{T_n}), \pi_5(p_{\mathrm{Comb}_n})]$ is a face of $\mathcal{E}_5^n$. □

We now introduce another tree structure that will yield a vertex in $\mathcal{E}_5^n$.

Definition 2.5.3. For two positive integers m and n let $\mathrm{bicomb}(m, n)$ denote the tree made by joining a comb tree of size m and a comb tree of size n together at a new root. We call such trees *bicomb trees*.

For example, $\mathrm{Bal}_5 = \mathrm{bicomb}(2, 3)$.

Lemma 2.5.4. *Let $T_n = \mathrm{bicomb}(\lfloor \frac{n}{2} \rfloor, \lceil \frac{n}{2} \rceil)$. Then $\pi_5(p_{T_n})$ is a vertex of $\mathcal{E}_5^n$.*

Proof. First note that for $n \geq 5$, the only trees in $\mathrm{RB_U}(n)$ that never contain Gir_5 as a restriction tree are the comb tree and the bicomb trees. This means that in $\mathcal{E}_5^n$, they are the only trees that fall on the edge $p_2 = 0$. To show that $\pi_5(p_{T_n})$ is a vertex of $\mathcal{E}_5^n$ it remains to to show that $\pi_5(p_{T_n})$ is extremal on this edge. We know that the comb tree is one of the extremal points on this edge and so the other extremal point will correspond to the bicomb tree with the highest density of Bal_5 as a restriction tree. Let $T' = \mathrm{bicomb}(i, n-i)$ be a bicomb tree for some $1 \leq i \leq n-1$. We let $b_5(T')$ denote

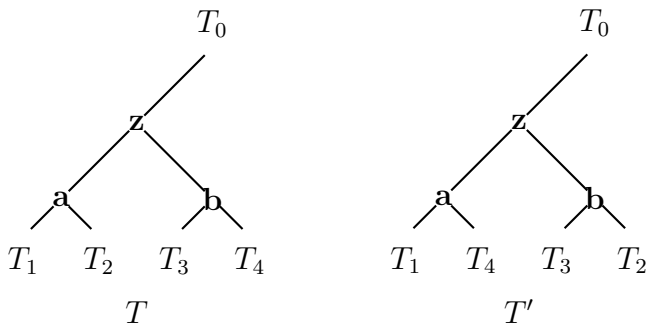


Figure 2.6: The two trees from the Proof of Lemma 2.5.5. Note that T_0 denotes all of the part of the tree that lies above the vertex z.

the number of times that Bal_5 occurs as a restriction tree of T'. From the structure of a bicomb tree we have

$$b_5(T') = \binom{i}{2}\binom{n-i}{3} + \binom{i}{3}\binom{n-i}{2}.$$

This function is maximized when $i = \lfloor \frac{n}{2} \rfloor$. □

Now we will show that the projection of the most balanced tree in $\mathrm{RB}_{\mathrm{U}}(n)$ is a vertex of $\mathcal{E}_5^n$. To do this, we prove a few lemmas about the number of Comb_5 trees that can appear as subtrees of a tree. These results follow the basic outline of Lemma 9 in [11], and are in some sense an extension of those results to 5 leaf trees.

For a tree $T \in \mathrm{RB}_{\mathrm{U}}(n)$ let $c_5(T)$ count the number of 5-subsets, S, of the leaves of T such that $T|_S = \mathrm{Comb}_5$. Let $c_4(T)$ and $b_4(T)$ be defined similarly but for Comb_4 and Bal_4 respectively.

Lemma 2.5.5. *Let T be as it is pictured in Figure 2.6 and T' obtained from T by swapping the positions of T_2 and T_4. For $i = 0, 1, 2, 3, 4$, let n_i be the number of leaves of T_i so $n = \sum_{i=0}^{4} n_i$. Without loss of generality choose $n_1 \geq n_2$ and $n_3 \geq n_4$. If $n_1 > n_3$ and $n_2 > n_4$ then $c_5(T) \geq c_5(T')$. Furthermore, if $n \geq 7$, then $c_5(T) > c_5(T')$.*

Proof. Without loss of generality assume that $n_1 \geq n_2$ and $n_3 \geq n_4$ and let Σ_z denote the set of leaves of T below the vertex z. Note that by construction, this is the same as the set of leaves below the vertex z in T'. If we take a 5-subset, S, of the leaves of T and T' then it is only possible for $T|_S \neq T'|_S$ if $|S \cap \Sigma_z| \geq 4$. It is straightforward to see that if $S \cap \Sigma_z$ has zero, one, two, or three elements, $T|_S = T'|_S$.

This means

$$c_5(T) - c_5(T') = (c_5(T_z) - c_5(T'_z)) + n_0(c_4(T_z) - c_4(T'_z))$$

where T_z and T'_z denote the subtrees of T and T' below z. Note that for any tree $S \in \mathrm{RB}_{\mathrm{U}}(n)$, it holds that

$$\binom{n}{4} = c_4(S) + b_4(S)$$

which gives

$$n_0(c_4(T_z) - c_4(T'_z)) = n_0(b_4(T'_z) - b_4(T_z))$$

and $(b_4(T'_z) - b_4(T_z))$ is guaranteed to be positive by Lemma 9 of [11] so the term $n_0(b_4(T'_z) - b_4(T_z))$ is nonnegative. It remains to show that $(c_5(T_z) - c_5(T'_z))$ is nonnegative. We can explicitly enumerate these quantities in the following way:

$$\begin{aligned} c_5(T_z) &= \sum_{i=1}^{4} c_5(T_i) + \sum_{i=1}^{4} c_4(T_i) \sum_{j=1, j\neq i}^{4} n_j + \binom{n_1}{3} n_2(n_3 + n_4) \\ &\quad + \binom{n_2}{3} n_1(n_3 + n_4) + \binom{n_3}{3} n_4(n_1 + n_2) + \binom{n_4}{3} n_3(n_1 + n_2) \end{aligned}$$

$$\begin{aligned} c_5(T'_z) &= \sum_{i=1}^{4} c_5(T_i) + \sum_{i=1}^{4} c_4(T_i) \sum_{j=1, j\neq i}^{4} n_i + \binom{n_1}{3} n_4(n_2 + n_3) \\ &\quad + \binom{n_4}{3} n_1(n_2 + n_3) + \binom{n_2}{3} n_3(n_1 + n_4) + \binom{n_3}{3} n_2(n_1 + n_4) \end{aligned}$$

We can simplify this to get that

$$c_5(T_z) - c_5(T'_z) = \frac{1}{6}(n_1 - n_3)(n_2 - n_4)(n_1 n_3(-3 + n_1 + n_3) + n_2 n_4(-3 + n_2 + n_4)).$$

Note that this quantity is nonnegative since $n_1 > n_3$ and $n_2 > n_4$ by assumption and $n_i \geq 1$ for $i = 1, 2, 3, 4$. Note that if $n \geq 7$, then we either have that $n_0 \geq 1$, or $\sum_{i=1}^{4} n_i \geq 7$ which both guarantee that $c_5(T) - c_5(T') > 0$. □

This lemma essentially tells us that if the tree has an internal node that is unbalanced, we can find a tree that has Comb_5 appear less frequently as a restriction tree. We now have another lemma following in the style of [11].

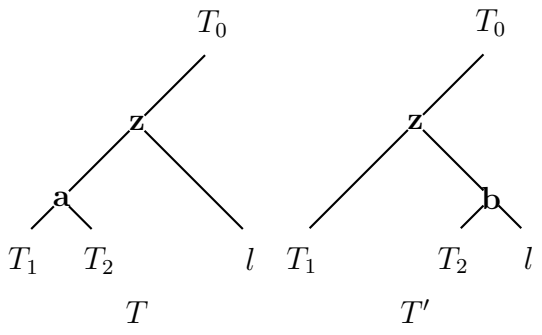


Figure 2.7

Lemma 2.5.6. *Let T be as it is pictured in Figure 2.7 and for $i = 0, 1, 2$, let n_i be the number of leaves of T_i and assume $n_1 \geq n_2$. We also assume that $n_1 + n_2 \geq 3$. Then $c_5(T) \geq c_5(T')$. Furthermore, if $n \geq 7$, then $c_5(T) > c_5(T')$.*

Proof. By the same reasoning as that given in the last lemma we know that

$$c_5(T) - c_5(T') = c_5(T_Z) - c_5(T'_z) + n_0(c_4(T_z) - c_4(T'_z))$$

and the nonnegativity of the second term follows by Lemma 10 in [11]. Now we can easily see that

$$c_5(T_z) = c_5(T_1) + c_5(T_2) + (n_2 + 1)c_4(T_1) + (n_1 + 1)c_4(T_2) + \binom{n_1}{3}n_2 + \binom{n_2}{3}n_1$$

$$c_5(T'_z) = c_5(T_1) + c_5(T_2) + (n_2 + 1)c_4(T_1) + (n_1 + 1)c_4(T_2) + \binom{n_2}{3}n_1$$

and so

$$c_5(T_Z) - c_5(T'_z) = \binom{n_1}{3}n_2$$

It is clear that the right hand side is always nonnegative. Note that if $n \geq 7$, then either $n_0 \geq 1$ or $n_1 \geq 3$. In both cases this guarantees that $c_5(T) - c_5(T') > 0$.

□

Combining these two lemmas together we get the following theorem. This theorem will immediately allow us to show that the projection of the most balanced tree in $\mathrm{RB_U}(n)$ will always be a vertex in $\mathcal{E}_5^n$.

Theorem 2.5.7. *For $n \geq 7$, the maximally balanced tree is the unique minimizer of $c_5(T)$ among all trees $T \in \mathrm{RB}_{\mathrm{U}}(n)$.*

Proof. This proof also follows the strategy of [11]. We assume that c_5 obtains it minimum value in $\mathrm{RB}_{\mathrm{U}}(n)$ at T but that T is not maximally balanced. We will try to find a contradiction. We let z be a non-balanced internal node with balanced children a and b. We let n_a and n_b be the number of leaves of the trees rooted at a and b respectively. Then since z is not balanced we have, without loss of generality, that $n_a \geq n_b + 2$. If b is a leaf then by Lemma 2.5.6 we immediately have that $c_5(T)$ is not minimum since $n \geq 7$. So we have that $n_b \geq 2$ and thus both a and b are balanced and must be internal nodes.

We now let v_1, v_2 be the children of a and v_3, v_4 be the children of b and take $n_i = \#L(T_{v_i})$ for $i = 1, 2, 3, 4$ and once again without loss of generality assume that $n_1 \geq n_2$ and $n_3 \geq n_4$. Since both a and b are balanced it must be that $n_1 = n_2$ or $n_1 = n_2 + 1$ and $n_3 = n_4$ or $n_3 = n_4 + 1$. Then the assumption that $n_a \geq n_b + 2$ immediately gives us that

$$n_1 + n_2 = n_a \geq n_b + 2 = n_3 + n_4 + 2$$

Then by previous assumptions we get that $n_1 > n_3$. Now since c_5 is minimum at T and $n \geq 7$, we can apply Lemma 2.5.5 to get that $n_4 \geq n_2$. Stringing together these inequalities we get that

$$n_1 > n_3 \geq n_4 \geq n_2$$

But since $n_1 = n_2$ or $n_1 = n_2 + 1$, the only possibility we have is that

$$n_1 - 1 = n_2 = n_3 = n_4$$

But then we get that $n_1 + n_2 = 2n_1 - 1$ and $n_3 + n_4 = 2n_1 - 2$ which contradicts the inequality $n_1 + n_2 \geq n_3 + n_4 + 2$. This tells us that any tree with at least 7 leaves must be maximally balanced around every internal node if it obtains the minimum value of c_5 on $\mathrm{RB}_{\mathrm{U}}(n)$. Since there is only one tree that is maximally balanced at every internal node, there is a unique minimizer of $c_5(T)$ in $\mathrm{RB}_{\mathrm{U}}(n)$ for $n \geq 7$ which is the maximally balanced tree. □

Corollary 2.5.8. *Let T_n be the maximally balanced tree in $\mathrm{RB}_{\mathrm{U}}(n)$. Then $\pi_5(p_{T_n})$ is a vertex of $\mathcal{E}_5^n$.*

Proof. The Corollary can be verified computationally for $n = 6$. For $n \geq 7$ Theorem 2.5.7

shows that T_n is the unique tree that attains the minimum value of c_5 among all trees in $\mathrm{RB}_{\mathrm{U}}(n)$. So it holds that $\{(c_5(T_n)/\binom{n}{5}, p_2, p_3) \in \mathcal{E}_5\} \cap \mathcal{E}_5^n = \{\pi_5(p_{T_n})\}$, thus $\pi_5(p_{T_n})$ is a vertex of $\mathcal{E}_5^n$. □

We have another Corollary that relates the exchangeable and sampling consistent distributions to the β-splitting model.

Corollary 2.5.9. *The projection of the most balanced tree in $\mathcal{E}_5^n$ approaches the $\beta = \infty$ point on the beta-splitting model as $n \to \infty$.*

Proof. It is enough to show that the complete symmetric tree $T_{2^n} \in \mathrm{RB}_{\mathrm{U}}(2^n)$ satisfies this property. We can just count the number of times that Gir_5 and Bal_5 occur as restriction trees when we restrict to a 5-subset of the leaves. We use the structure of T_{2^n} to write down a simple recurrence for $g_5(T_{2^n})$ and $b_5(T_{2^n})$ and then solve the recurrence. Since we can choose our subset to be on the right side of the root of T_{2^n}, the left side of the root of T_{2^n}, or to have 3 leaves from one side and 2 leaves from the other we have that

$$b_5(T_{2^n}) = 2b_5(T_{2^{n-1}}) + 2\binom{2^{n-1}}{3}\binom{2^{n-1}}{2}.$$

As for g_5, we can once again choose our subset to be on either the right or left side of the root of T_{2^n} or we can choose to have 1 leaf on a side of the tree and a 4 leaf symmetric tree on the other. This can be done in just $2^{n-1}m_{2,2^{n-1}}$ ways which implies

$$g_5(T_{2^n}) = 2g_5(T_{2^{n-1}}) + 2(2^{n-1}m_{2,2^{n-1}}) = 2g_5(T_{2^{n-1}}) + 2^n m_{2,2^{n-1}}.$$

Both of these recurrences can be solved explicitly using a computer algebra system. We get that

$$b_5(T_{2^n}) = \frac{1}{315}2^{n-2}(2^n-4)(2^n-2)(2^n-1)(7\cdot 2^n - 11)$$

$$g_5(T_{2^n}) = \frac{1}{105}2^{n-3}(2^n-4)(2^n-3)(2^n-2)(2^n-1)$$

We can then find the probabilities p_2 and p_3 of Gir_5 and Bal_5 by simply dividing out by $\binom{2^n}{5}$. This yields

$$p_3 = \frac{b_5(T_{2^n})}{\binom{2^n}{5}} = \frac{2}{3} + \frac{20}{21(2^n-3)}$$

$$p_2 = \frac{g_5(T_{2^n})}{\binom{2^n}{5}} = \frac{1}{7}$$

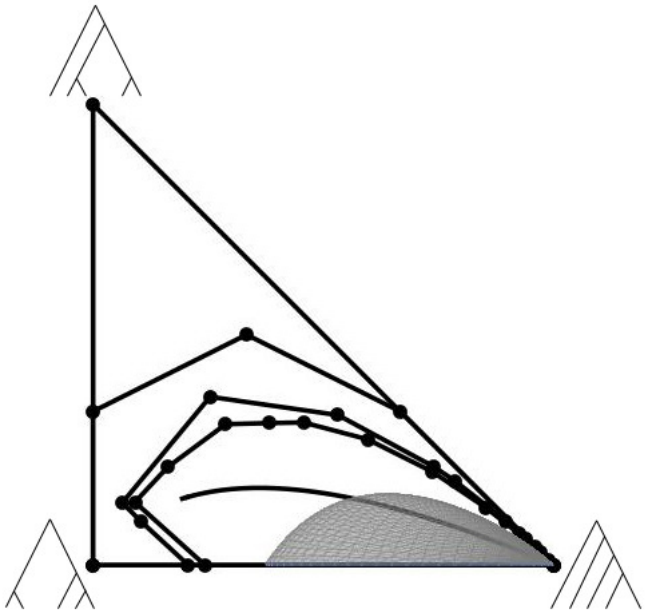


Figure 2.8: The multinomial model on the two leaf tree is in grey and the β-splitting model is the thick black curve. The thinner black lines are the boundary of $\mathcal{E}_5^n$ for n=5,6,9,12.

Clearly as $n \to \infty$ we have $p_3 \to \frac{2}{3}$ and $p_2 \to \frac{1}{7}$.

On the other hand, we recall that the probability of obtaining a tree under the beta-splitting model is just a rational function in β that can be explicitly calculated. We can then find the limit of these rational functions to get that the beta-splitting curve approaches the point

$$(p_1, p_2, p_3) = (\frac{4}{21}, \frac{1}{7}, \frac{2}{3})$$

as $\beta \to \infty$ as well and so the projection of T_{2^n} in $\mathcal{E}_5^{2^n}$ is approaching the $\beta = \infty$ point on the curve. □

These are all of the tree structures in $\mathrm{RB}_{\mathrm{U}}(n)$ we have been able to find that always appear as vertices in $\mathcal{E}_5^n$. We end this section with Figure 2.8, which pictures all of the families of exchangeable and sampling consistent distributions that we have discussed and the vertices of $\mathcal{E}_n^m$ for some small values of m.

2.6 Distributions on $\mathcal{E}_n^\infty$

While we are not able to get a description of the vertices of $\mathcal{E}_n^m$ for general m and n, it is possible to to describe $\mathcal{E}_n^\infty$ using the multinomial model that was introduced in Section 2.3.2. In particular, this shows that multinomial models converge as an inner limit to $\mathcal{E}_n^\infty$.

Theorem 2.6.1. *Let $\{T_m\}_{m=n}^\infty$ be a sequence of tree shapes and $p^{(m)} = \pi_n(T_m)$ be the corresponding sequence of distributions. If $p^{(m)}$ converges to some $p \in \mathcal{E}_n^\infty$ as m goes to infinity, then there exists a sequence of multinomial distributions $\{d^{(m)}\}_{m=n}^\infty$ that also converges to p as m goes to infinity.*

Proof. Define $d^{(m)}$ to be the multinomial distribution on the tree T_m with the edge parameter vector $(t_e | e \in E(T_m))$ such that $t_e = \frac{1}{m}$ if one of the vertices in e is one of the original m leaves of T_m and $t_e = 0$ otherwise. Note that these nonzero edge parameters are bijectively associated to the leaves of T_m and we may call the set of nonzero edge parameters $L(T_m)$ meaning the leaf set of T_m. To show that $d^{(m)}$ also converges to p, it is enough to show that for every tree $T \in \mathrm{RB}_{\mathrm{U}}(n)$, $\lim_{m\to\infty} d^{(m)}(T) = \lim_{m\to\infty} p^{(m)}(T)$. Fix a labelling of T_m and let $c_{T_m}(T)$ be the number of sets $S \subseteq [m]$ such that $\mathrm{shape}(T_m|_S) = T$. By Corollary 2.2.8, $p^{(m)}(T)$ is the induced subtree density of T in T_m, so $p^{(m)}(T) = \frac{c_{T_m}(T)}{\binom{m}{n}}$. So

$$\lim_{m\to\infty} p^{(m)}(T) = \lim_{m\to\infty} \frac{c_{T_m}(T)}{\binom{m}{n}} = \lim_{m\to\infty} \frac{n!}{m^n} c_{T_m}(T)$$

On the other hand, let $M^{(m)} = \{A \in M_n^{T_m} | (T_m)_A = T, p_A \neq 0\}$, then

$$d^{(m)}(T) = \sum_{A \in M^{(m)}} p_A$$

by definition and we note by requiring that multisets $A \in M^{(m)}$ have that $p_A \neq 0$, $M^{(m)}$ only includes multisets whose support is contained in $L(T_m)$. Also note that p_A is either 0 or $\binom{n}{m_A(t_{e_1}), m_A(t_{e_2}), \ldots m_A(t_{e_{2m-1}})} \frac{1}{m^n}$ since all the edge parameters are 0 or $\frac{1}{m}$. So to understand the quantity $d^{(m)}(T)$ it is enough to know the coefficient of $\frac{1}{m^n}$. Note that any multiset A has a naturally associated integer partition of n to it, formed by taking the multiplicities of each unique element that appears in it. Call this integer partition the weight of A, denoted $wt(A)$, and let $M_\lambda^{(m)}$ be the set of multisets in $M^{(m)}$ with weight λ. Now observe that for $A, B \in M_\lambda^{(m)}$, $p_A = p_B$ since the value of the multinomial coefficient is totally determined by the weight and the product of the edge parameters is always $\frac{1}{m^n}$. If we let $\binom{n}{\lambda}$ be the value of the multinomial coefficient then the formula for $d^{(m)}(T)$ can be

rewritten as

$$d^{(m)}(T) = \frac{1}{m^n} \sum_{\lambda \vdash n} \binom{n}{\lambda} |M_\lambda^{(m)}|$$

but we can bound the quantity $|M_\lambda^{(m)}|$. We note that the quantity $|M_\lambda^{(m)}|$, is at most $l(\lambda)!\binom{m}{l(\lambda)}$ where $l(\lambda)$ is the length of the partition λ. This is because there are $\binom{m}{l(\lambda)}$ choices for which elements to use in the multiset and at most $l(\lambda)!$ unique multisets for each choice of elements. Recall that there are only $\binom{m}{l(\lambda)}$ choices to use in a multiset since any $A \in M_\lambda^{(m)}$ must have $p_A \neq 0$ which means A must be a multiset on the leaves of T_m. Since $l(\lambda)!\binom{m}{l(\lambda)}$ is a polynomial in m of degree $l(\lambda)$ though, we have that

$$\lim_{m\to\infty} \frac{1}{m^n} \sum_{\lambda \vdash n} \binom{n}{\lambda} |M_\lambda^{(m)}| = \lim_{m\to\infty} \frac{n!}{m^n} |M_{(1,1,\ldots,1)}^{(m)}|$$

since the partition $\lambda = (1, 1, \ldots 1)$ is the only partition where $|M_{(1,1,\ldots,1)}^{(m)}|$ is of the order m^n, and so is the only term that contributes to the limit. Now we note that the multisets $A \in M_{(1,1,\ldots,1)}^{(m)}$ correspond exactly to choosing subsets of the leaves of T_m that yield T upon restriction since the only edges that can be in A are those corresponding to leaves, every leaf can be chosen at most once, and $\text{shape}((T_m)_A) = T$. So $|M_{(1,1,\ldots,1)}^{(m)}| = c_{T_m}(T)$, and so

$$\lim_{m\to\infty} d^{(m)} = \lim_{m\to\infty} \frac{n!}{m^n} c_{T_m}(T) = \lim_{m\to\infty} p^{(m)}$$

and since $p^{(m)}$ converges, to p, it must be that $d^{(m)}$ also does. □

Theorem 2.6.2. *For all $n \geq 1$, there exists a constant $C > 0$ such that for all $m > n$ and $p \in \mathcal{E}_n^m$ there exists $d \in \mathcal{E}_n^\infty$ such that*

$$\max_{S \in \mathrm{RB}_{\mathrm{U}}(n)} |p(S) - d(S)| \leq \frac{C}{m}.$$

Proof. Note that if $p \in \mathcal{E}_n^m$, then we have for every $S \in \mathrm{RB}_{\mathrm{U}}(n)$,

$$p(S) = \sum_{T \in \mathrm{RB}_{\mathrm{U}}(m)} \lambda_T \pi_n(p_T)(S)$$

where the above combination is convex by Lemma 2.2.5. Then let d^T be defined as the multinomial distribution d^T on T just as $d^{(m)}$ is defined for T_m in the previous theorem.

Then recall from the proof of the previous theorem that

$$d^T(S) = \frac{1}{m^n} \sum_{\lambda \vdash n} \binom{n}{\lambda} |M_\lambda^T|$$

where $M_\lambda^T = \{A \in M_n^T | T_A = S,\ p_A \neq 0,\ wt(A) = \lambda\}$. Also recall from the proof of the previous theorem that $|M_{(1,1,\ldots,1)}^T| = c_T(S)$. Combining these facts with the definition of $\pi_n(p_T)$ and the triangle inequality gives

$$|\pi_n(p_T)(S) - d^T(S)| \leq \left| \frac{c_T(S)}{\binom{m}{n}} - \frac{n! c_T(S)}{m^n} \right| + \left| \frac{1}{m^n} \sum_{\substack{\lambda \vdash n \\ \lambda \neq (1,1,\ldots,1)}} \binom{n}{\lambda} |M_\lambda^T| \right| \tag{2.4}$$

and we now bound each term on the right hand side of this inequality.

To bound the first term in equation (2.4), note that $c_T(S)$ is a nonnegative quantity and is bounded above by $\binom{m}{n}$. This gives the inequality

$$\left| \frac{c_T(S)}{\binom{m}{n}} - \frac{n! c_T(S)}{m^n} \right| \leq \left| 1 - \frac{\frac{m!}{(m-n)!}}{m^n} \right| \leq \left| \frac{m^n - (m-n)^n}{m^n} \right| \leq \frac{n^2}{m}. \tag{2.5}$$

Note that this bound does not depend on the trees T and S.

To bound the second term we again recall from the proof of the previous theorem that $|M_\lambda^T| \leq l(\lambda)! \binom{m}{l(\lambda)}$ for each partition λ of n. Then we have that

$$\left| \frac{1}{m^n} \sum_{\substack{\lambda \vdash n \\ \lambda \neq (1,1,\ldots,1)}} \binom{n}{\lambda} |M_\lambda^T| \right| \leq \sum_{\substack{\lambda \vdash n \\ \lambda \neq (1,1,\ldots,1)}} \binom{n}{\lambda} \frac{l(\lambda)! \binom{m}{l(\lambda)}}{m^n} \tag{2.6}$$

but since $\lambda \neq (1, 1, \ldots, 1)$, it must be that $l(\lambda) \leq n - 1$ so $l(\lambda)! \binom{m}{l(\lambda)} \leq m^{n-1}$ for all the remaining partitions λ. Applying this fact to the right hand side of equation (2.6) gives the bound

$$\left| \frac{1}{m^n} \sum_{\substack{\lambda \vdash n \\ \lambda \neq (1,1,\ldots,1)}} \binom{n}{\lambda} |M_\lambda^T| \right| \leq \frac{1}{m} \sum_{\substack{\lambda \vdash n \\ \lambda \neq (1,1,\ldots,1)}} \binom{n}{\lambda} \leq \frac{\tilde{C}}{m} \tag{2.7}$$

where $\tilde{C} \in \mathbb{R}$ is a constant that also does not depend on the trees T and S but only on n. Applying the bounds for each term to equation (2.4) and setting $C = \tilde{C} + n^2$ gives

$$|\pi_n(p_T)(S) - d^T(S)| \leq \frac{C}{m} \tag{2.8}$$

and again we note that this bound is independent of the trees T and S. We are now ready to construct a distribution $d \in \mathcal{E}_n^\infty$ that gives the desired result. From the discussion of the multinomial model, we have that each distribution $d^T \in \mathcal{E}_n^\infty$ and so from the convexity of $\mathcal{E}_n^\infty$ we get

$$d = \sum_{T \in \mathrm{RB_U}(m)} \lambda_T d^T \in \mathcal{E}_n^\infty.$$

We can now use the expression for p we began with and the bound obtained in equation (2.8) to get that

$$|p(S) - d(S)| \leq \sum_{T \in \mathrm{RB_U}(m)} \lambda_T |\pi_n(p_T)(S) - d^T(S)| \leq \frac{C}{m}.$$

□

Theorem 2.6.1 gives that the limit of any convergent sequence $(v_m)_{m \geq 1}$ where $v_m \in V(\mathcal{E}_n^m)$ can also be realized as the limit of points coming from multinomial models. Theorem 2.6.2 shows that if we have a distribution in $\mathcal{E}_n$ that can be extended to part of a finitely sampling consistent family, then it can be approximated with an infinitely sampling consistent distribution. With Theorem 2.6.1 and the following proposition, we will show that $\mathcal{E}_n^\infty$ is actually the convex hull of all limits of convergent sequences of vertices, and thus the convex hull of limits of distributions drawn from the multinomial model. To do this we need a basic proposition from convex analysis which the proof of is included for completeness.

Proposition 2.6.3. *Let $(P_m)_{m \geq 1}$ be a sequence of polytopes in $\mathbb{R}^n$ such that for all $m \geq 1$, $P_{m+1} \subseteq P_m$. Let*

$$P = \overline{\mathrm{conv}(\{ \lim_{m \to \infty} v_{i_m}^{(m)} | v_{i_m}^{(m)} \in V(P_m) \text{ and } (v_{i_m}^{(m)})_{m \geq 1} \text{ converges } \})}$$

where the bar denotes the closure in the Euclidean topology. Then $P = \cap_{m=1}^\infty P_m$.

Proof. It is straightforward to see that $P \subseteq \cap_{m=1}^\infty P_m$. To show that the sets are equal suppose that there is $p \in (\cap_{m=1}^\infty P_m) \setminus P$. Then the Basic Separation Theorem of convex analysis implies there must exist an affine functional ℓ with $\ell(p) \leq 0$ and $\ell(w) > 0$ for all $w \in P$. We also have that since $p \in \cap_{m=1}^\infty P_m$, for each $m \geq 1$, p can be written as

$$p = \sum_{j=1}^{k_m} \lambda_j v_j^{(m)}$$

where the $v_j^{(m)}$ are the vertices of P_m. Then because $\ell(p) \leq 0$ it must be that for each m, there exists at least one vertex $v_{i_m}^{(m)}$ of P_m such that $\ell(v_{i_m}^{(m)}) \leq 0$. Since all the points $v_j^{(m)}$ lie in P_1 which is a compact set, there exists a convergent subsequence $(v_{i_{m_k}}^{(m_k)})_{k\geq 1}$ with limit $v \in P$, thus $\ell(v) > 0$. But it also holds that

$$\ell(v) = \lim_{k\to\infty} \ell(v_{i_{m_k}}^{(m_k)}) \leq 0$$

which is a contradiction. □

Corollary 2.6.4. *Let $d_{T_m}^{(m)}$ denote the specific multinomial model construction on the tree $T_m \in \mathrm{RB}_{\mathrm{U}}(m)$ described in Theorem 2.6.1. Then*

$$\mathcal{E}_n^\infty = \overline{\mathrm{conv}(\{\lim_{m\to\infty} d_{T_m}^{(m)} | \pi_n(T_m) \in V(\mathcal{E}_n^m) \text{ and } (d_{T_m}^{(m)})_{m\geq n} \text{ converges } \})}.$$

Proof. Recall that $\mathcal{E}_n^\infty = \cap_{m=n}^\infty \mathcal{E}_n^m$, thus by Proposition 2.6.3,

$$\mathcal{E}_n^\infty = \overline{\mathrm{conv}(\{\lim_{m\to\infty} \pi_n(p_{T_m}) | T_m \in \mathrm{RB}_{\mathrm{U}}(m) \text{ and } (\pi_n(p_{T_m}))_{m\geq 1} \text{ converges } \})}$$

since the vertices of $\mathcal{E}_n^m$ correspond to a subset of the points $\pi_n(T_m)$. Applying Theorem 2.6.1 to the sequence $(\pi_n(T_m))_{m\geq 1}$ gives the result. □

Corollary 2.6.4 shows that every exchangeable and infinitely sampling consistent distribution is either a convex combinations of limits of multinomial distributions or a limit point of points in that set. Understanding the structure of the multinomial models may shed greater light on the structure of $\mathcal{E}_n^\infty$ as a whole. We view Theorem 2.6.2 and Corollary 2.6.4 as the rooted binary tree analogue to Theorems 3 and 4 in [14], in essence they are finite forms of a de Finetti-type theorem for rooted binary trees. As previously mentioned, the work done in [19] and [18] establishes a more typical de Finetti theorem in the sense that it shows every infinitely sampling consistent sequence of distributions can be obtained by sampling from a limit object using techniques from Probability theory.

We also note that the requirement that the induced subtree densities converge is quite similar to the idea of graph convergence that appears in [34] and that many of the ideas in the theory of graph limits may also be applied to trees. The very well developed theory of graph limits contains many equivalent versions of the limiting object (see Theorem 11.52 in [34]). The work done in [19] and [18] makes the connection between the limiting object, a random real tree, and an infinitely sampling consistent model. It is still unknown if this can be connected to ideas such as tree parameters (the induced subtree density for

instance) and to metrics on finite trees as has been done in the theory of graph limits. It seems that many of these equivalences hold but differences in techniques will be required.

CHAPTER

3

IDENTIFIABILITY IN PHYLOGENETICS USING ALGEBRAIC MATROIDS

3.1 Introduction

A statistical model is identifiable if the map parameterizing the model is injective. This means that the parameters producing a probability distribution in the model can be uniquely determined from the distribution itself which is a critical property for meaningful data analysis. In phylogenetic models, the identifiability of the tree parameter is especially important since this allows for evolutionary histories to be inferred from observed genetic data.

The identifiability of the tree parameter in basic models has already been established [10] and a natural next step is to investigate the identifiability of the tree parameters in phylogenetic mixture models. Mixture models can be used to represent more complicated evolutionary events such as horizontal gene transfer. [37] showed that the tree parameters are not identifiable for 2-tree mixtures on four leaf trees under the Cavendar-Farris-Neyman (CFN) model. On the other hand, positive results for the identifiability of tree

parameters in other group-based models were obtained in both [2] and [32]. In [2], the authors constructed linear invariants for 2-tree Jukes-Cantor (JC) and Kimura 2-Parameter (K2P) mixtures to show that the tree parameters were identifiable and [32] used direct computation to construct invariants for 3-tree JC mixtures to obtain identifiability results. These computations often involve time consuming Gröbner basis computations, which are not possible to do for larger models. Similar calculations were also done in [21] to establish the identifiability of the network parameters in Jukes-Cantor network models.

Our goal in this chapter is to introduce a new algorithm that can be used to show that parameters of an algebraic statistical model are identifiable by computing independent sets in a naturally associated algebraic matroid. This allows us to avoid dealing with the vanishing ideals that are typically used and thus avoid Gröbner basis calculations. We begin with a short background on generic identifiability and algebraic matroids in Section 3.2. We then introduce the main algorithm we employ to prove identifiability results in Section 3.3. We provide both an exact verification based on symbolic computation and a randomized algorithm with probabilistic guarantees based on the Schwartz-Zippel Lemma. In Section 3.4 we use the algorithm and the Six-To-Infinity Theorem [36] to show that the tree parameters are generically identifiable in 2-tree CFN and K3P mixture models. We end by showing how our algorithm can be used to extend the results in [21] for JC phylogenetic networks to K2P and K3P networks.

3.2 Preliminaries

In this section we provide some background on identifiability and describe some common tools used to prove identifiability results.

Our main objects of focus in this chapter will be parametric algebraic statistical models for discrete random variables. This means we have a rational map

$$\phi : \Theta \to \Delta_{n-1} = \left\{ p \in \mathbb{R}^n \ : \ \sum_{i=1}^{n} p_i = 1, \ p_i \geq 0 \text{ for all } i \right\}$$

whose image, which we denote by M, is the model of interest. This is a broad setting that includes many classic statistical models such as distributions of discrete random variables and the phylogenetic models that we will discuss in the later sections. The definitions and techniques presented in this chapter could also be adapted for Gaussian random variable and other continuous models with finite dimensional natural parameter spaces.

If we have a family of these models $\{M_s\}_{s=1}^k$ that all sit inside Δ_{n-1} and are indexed by a discrete parameter s, then we say that the discrete parameter s is *globally identifiable* if $M_{s_1} \cap M_{s_2} = \emptyset$ for every distinct pair $\{s_1, s_2\}$ of values of s. Most models are not globally identifiable but may still satisfy a slightly weaker notion of identifiability instead.

Definition 3.2.1. Let $\{M_s\}_{s=1}^k$ be a collection of algebraic models that sit inside the probability simplex Δ_{n-1}, then the parameter s is *generically identifiable* if for each 2-subset $\{s_1, s_2\} \subset [k]$,

$$\dim(M_{s_1} \cap M_{s_2}) < \min(\dim(M_{s_1}), \dim(M_{s_2}))$$

Another way to think about generic identifiability is that the overlap of any two models in the family is a Lebesgue measure zero subset of both of the overlapping models. A typical tool for proving generic identifiability of algebraic models is the following proposition that uses the *vanishing ideal* $\mathcal{I}(M) = \{f \in \mathbb{C}[p] \ : \ f(p) = 0 \text{ for all } p \in M\}$ of the model M.

Proposition 3.2.2. *[45, Proposition 16.1.12] Let M_1 and M_2 be two algebraic models which sit inside the probability simplex Δ_{n-1} and have irreducible Zariski closures. If there exists polynomials f_1 and f_2 such that*

$$f_1 \in \mathcal{I}(M_1) \setminus \mathcal{I}(M_2) \text{ and } f_2 \in \mathcal{I}(M_2) \setminus \mathcal{I}(M_1)$$

then $\dim(M_1 \cap M_2) < \min(\dim(M_1), \dim(M_2))$.

If the models M_1 and M_2 have the same dimension, then to ensure their intersection is lower dimensional, it suffices to show that $\mathcal{I}(M_1) \neq \mathcal{I}(M_2)$. This means it is enough to find either $f \in \mathcal{I}(M_1) \setminus \mathcal{I}(M_2)$ or $f \in \mathcal{I}(M_2) \setminus \mathcal{I}(M_1)$.

We note here that the vanishing ideal of M also completely defines the *Zariski closure* of the model which is the algebraic variety $\overline{M} = \{p \in \mathbb{C}^n \ : \ f(p) = 0 \text{ for all } f \in \mathcal{I}(M)\}$. Essentially, $\mathcal{I}(M)$ gives part of the *implicit* description of the model M. A full implicit description of M also requires finding polynomial inequalities that define M as a semialgebraic set. Computing the ideal $\mathcal{I}(M)$ typically requires Gröbner basis computations which can be difficult, especially as the number of variables involved increases.

3.3 Certifying Generic Identifiability With Algebraic Matroids

In this section we make a few basic observations that will lead to a new algorithm for certifying the generic identifiability of a family of models using their associated algebraic matroids. Our starting point for proving identifiability using algebraic methods is Proposition 3.2.2. However, it often difficult to find the polynomials required by Proposition 3.2.2 to certify identifiability. The following proposition is the driver of our algebraic matroid based procedure for verifying identifiability.

Proposition 3.3.1. *Let M_1 and M_2 be two algebraic models which sit inside the probability simplex Δ_{n-1} and have irreducible Zariski closures. Without loss of generality assume* $\dim(M_1) \geq \dim(M_2)$. *If there exists a subset S of the coordinates such that*

$$S \in \mathfrak{I}(\mathcal{M}(M_2)) \setminus \mathfrak{I}(\mathcal{M}(M_1)) \tag{3.1}$$

then $\dim(M_1 \cap M_2) < \min(\dim(M_1), \dim(M_2))$.

Note that we abuse notation and write $\mathcal{M}(M)$ to denote the matroid $\mathcal{M}(\overline{M})$.

Proof. Since M_1 and M_2 have irreducible Zariski closures their vanishing ideals $\mathcal{I}(M_1)$ and $\mathcal{I}(M_2)$ are prime and so define the same matroid as M_1 and M_2 respectively. First suppose that $\dim(M_1) > \dim(M_2)$. This dimension inequality implies that there is a polynomial $f_2 \in \mathcal{I}(M_2) \setminus \mathcal{I}(M_1)$. Then since $S \in \mathfrak{I}(\mathcal{M}(M_2)) \setminus \mathfrak{I}(\mathcal{M}(M_1))$, it holds that $\mathcal{I}(M_1) \cap k[S] \neq \langle 0 \rangle$ but $\mathcal{I}(M_2) \cap k[S] = \langle 0 \rangle$ which implies that there exists $f_1 \in \mathcal{I}(M_1) \setminus \mathcal{I}(M_2)$ and so the result follows by Proposition 3.2.2.

Now suppose that $\dim(M_1) = \dim(M_2)$. The existence of $S \in \mathfrak{I}(\mathcal{M}(M_2)) \setminus \mathfrak{I}(\mathcal{M}(M_1))$ implies that $\mathcal{I}(M_1) \neq \mathcal{I}(M_2)$. Since these ideals are prime and of the same dimension we must have the mutual noncontainments $\mathcal{I}(M_1) \not\subseteq \mathcal{I}(M_2)$ and $\mathcal{I}(M_2) \not\subseteq \mathcal{I}(M_1)$. This immediately implies the existence of $f_1 \in \mathcal{I}(M_1) \setminus \mathcal{I}(M_2)$ and $f_2 \in \mathcal{I}(M_2) \setminus \mathcal{I}(M_1)$ and so again the result follows by Proposition 3.2.2. □

In essence, Proposition 3.3.1 can certify the existence of the desired polynomials for applying Proposition 3.2.2, without necessarily finding them, only proving they exist. Note that Proposition 3.3.1 is weaker then Proposition 3.2.2. This is because there can be models with different ideals but that have the same matroid. This is due to the fact that the matroid only keeps track of which sets of coordinates have polynomial relations in

the ideals of the models but not the nature of the polynomial relations themselves. This is illustrated in Example 3.5.8.

Proposition 3.3.2. *[40, Proposition 2.5] Let $\mathbb{K}$ be a field of characteristic zero and $V \subset \mathbb{K}^n$ be a variety parameterized by ϕ with Jacobian $J(\phi)$ defined as in Equation (1.2). Then the matrix obtained by plugging in generic parameter values into $J(\phi)$ gives a linear matroid over $\mathbb{K}$ which is the same as that defined by $J(\phi)$ with symbolic parameters over $\mathbb{K}(\theta)$ and thus the same as $\mathcal{M}(V)$.*

We use $\mathcal{M}(J(\phi), \mathbb{K})$ to denote the linear matroid we get by plugging in random parameter values for θ and $\mathcal{M}(J(\phi), \mathbb{K}(\theta))$ to denote the symbolic matroid. With these two propositions we are ready to define the main algorithm that we use to prove identifiability results.

Algorithm 3: matroidSeparate

Input : Two maps ϕ_1, ϕ_2 parameterizing models M_1 and M_2 in $\mathbb{K}^n$ with $\dim(M_1) \geq \dim(M_2)$, a number of trials t.

Output : A certificate S satisfying Equation (3.1) in Proposition 3.3.1

```
for i = 0 to t do
    Randomly select T ⊆ [n] such that |T| ≤ dim(M_2);
    if T ∈ I(M(J(φ_2), K)) \ I(M(J(φ_1), K)) then
        if T ∈ I(M(J(φ_2), K(θ))) \ I(M(J(φ_1), K(θ))) then
            S = T;
            Break;
return S or report that no certificate was found.
```

In summary, the algorithm works by randomly plugging in a numerical value for θ and testing random subsets until it finds an example of a set S where the submatrices of the Jacobians have different rank. Random rational numbers are used so that the rank computations are exactly calculated symbolically (rather than using a numerical rank test with floating point numbers). Once a candidate set is found, then an exact symbolic computation over $\mathbb{K}(\theta)$ is performed to verify the result exactly.

In cases where it is too time consuming to compute over $\mathbb{K}(\theta)$ we can use the Schwartz-Zippel Lemma from polynomial identity testing to produce a certificate that satisfies equation (3.1) with probability $1 - \epsilon$.

Lemma 3.3.3. *(Schwartz-Zippel) Let $f \in \mathbb{K}[p_1, \ldots p_d]$ be a non-zero polynomial of total*

degree α. Let E be a finite subset of k and $r_1, \ldots r_d$ be selected at random independently and uniformly from E. Then

$$P(f(r_1, \ldots, r_d) = 0) \leq \frac{\alpha}{|E|}.$$

Determining if $S \in \mathfrak{I}(\mathcal{M}(J(\phi_2), \mathbb{K}(\theta))) \setminus \mathfrak{I}(\mathcal{M}(J(\phi_1), \mathbb{K}(\theta)))$ can be done by evaluating minors of the submatrices of the Jacobian matrices corresponding to S. Since minors are polynomials in the entries of the matrices, we can use the this lemma to bound the probability that $S \in \mathfrak{I}(\mathcal{M}(J(\phi_2), \mathbb{K}(\theta))) \setminus \mathfrak{I}(\mathcal{M}(J(\phi_1), \mathbb{K}(\theta)))$ without ever computing over $\mathbb{K}(\theta)$.

Corollary 3.3.4. *Let $S \in \mathfrak{I}(\mathcal{M}(J(\phi), \mathbb{K}(\theta)))$ and let $E \subseteq \mathbb{K}$ be a finite set such that $|E| > \alpha$ where α is the degree of an $|S| \times |S|$ minor of $J(\phi)_S$ that is not identically zero. Let $r_1, \ldots r_d$ be selected independently and uniformly at random from E and let $\mathcal{M}(J(\phi), \mathbb{K})$ be the linear matroid obtained by plugging in $r_1, \ldots r_d$ for the parameters. Then*

$$P(S \notin \mathfrak{I}(\mathcal{M}(J(\phi), \mathbb{K}))) \leq \frac{\alpha}{|E|}.$$

Proof. First note that the $|S| \times |S|$ minors of the matrix $J(\phi)_S$ are polynomials in θ which we denote by $f_i(\theta) \in \mathbb{K}[\theta]$ for $1 \leq i \leq \binom{d}{|S|}$. Since $S \in \mathfrak{I}(\mathcal{M}(J(\phi), \mathbb{K}(\theta)))$ there exists at least one f_j that is not identically zero. On the other hand, $S \notin \mathfrak{I}(\mathcal{M}(J(\phi), \mathbb{K}))$ if and only if $f_i(r) = 0$ for all $1 \leq i \leq \binom{d}{|S|}$ so

$$P(S \notin \mathfrak{I}(\mathcal{M}(J(\phi), \mathbb{K}))) = P\left(f_i(r) = 0, 1 \leq i \leq \binom{d}{|S|}\right) \leq P(f_j(r) = 0)$$

Letting $\deg(f_j) = \alpha$ and applying the Schwartz-Zippel Lemma gives

$$P(f_j(r) = 0) \leq \frac{\alpha}{|E|}$$

which gives us the desired result. □

Corollary 3.3.4 suggests a new strategy for deciding if $S \in \mathfrak{I}(\mathcal{M}(J(\phi), \mathbb{K}(\theta)))$ without performing any symbolic computation. We let E be as described in Corollary 3.3.4 and repeatedly sample $r_1, \ldots r_d$ independently and uniformly at random, plug these values into $J(\phi)$, and calculate the rank of $J(\phi)_S$. If $\text{rank}(J(\phi)_S) = |S|$, then we immediately know $S \in \mathfrak{I}(\mathcal{M}(J(\phi), \mathbb{K}(\theta)))$. If after l trials, we find for each trial that $\text{rank}(J(\phi)_S) < |S|$, then

we say that $S \notin \mathfrak{I}(\mathcal{M}(J(\phi), \mathbb{K}(\theta)))$. Corollary 3.3.4 and our amplification by independent trials guarantee that the probability that we made an error is less than $\left(\frac{\alpha}{|E|}\right)^l$. This procedure is utilized in Algorithm 4 to avoid symbolic computation completely.

Algorithm 4: matroidSeparateSZ

Input : Two maps ϕ_1, ϕ_2 parameterizing models M_1 and M_2 in $\mathbb{K}^n$ with $\dim(M_1) \geq \dim(M_2)$, a number of trials t, a tolerance ϵ.

Output : A certificate S satisfying Equation (3.1) in Proposition 3.3.1 with probability at least $1 - \epsilon$.

1 Choose a finite subset $|E| \subseteq \mathbb{K}$ such that $|E| > \alpha$ where α is the maximum degree of any $dim(M_2) \times dim(M_2)$ minor of $J(\phi_1)$;
2 **for** $i = 0$ **to** t **do**
3 Randomly select $T \subseteq [n]$ such that $|T| \leq \dim(M_2)$;
4 Sample points $r_1, \ldots r_d$ independently and uniformly at random from E and plug in to each $J(\phi_i)$;
5 **if** $T \in \mathfrak{I}(\mathcal{M}(J(\phi_2), \mathbb{K})) \setminus \mathfrak{I}(\mathcal{M}(J(\phi_1), \mathbb{K}))$ **then**
6 Choose l such that $\left(\frac{\alpha}{|E|}\right)^l \leq \epsilon$;
7 **for** $j = 0$ **to** l **do**
8 Sample points $r'_1, \ldots r'_d$ independently and uniformly at random from E and plug in to $J(\phi_1)$;
9 **if** $T \in \mathfrak{I}(\mathcal{M}(J(\phi_1), \mathbb{K}))$ **then**
10 Break and return to line 3;
11 S = T;
12 Break;
13 **return** *S or report that no certificate was found.*

Both Algorithm 3 and Algorithm 4 can be modified in the case that $\dim(M_1) = \dim(M_2)$. In that case, we can also accept T as a certificate if it is an independent set for M_1 but not M_2 in both the numerical step and the symbolic step. This is because we just need to certify that the models are not equal in the case where they are the same dimension. This is a very general version of the algorithm and it can be fine tuned in many ways depending on the specifics of the models. One such modification in the case that the models are the same dimension would be to only check for sets T such that $|T| = \dim(M_2)$. This is equivalent to searching for a basis for the matroid of one model that is not a basis for the other. If the matroids are not the same, then such a basis must exist since a matroid is uniquely determined by its bases. However, from a practical

standpoint, it is faster to perform the symbolic rank calculations on smaller matrices, so hunting for small sized sets that verify that the matroids are different can speed up computations.

Using Algorithm 3 and Proposition 3.3.1 to certify identifiability has several advantages over approaches that rely on Proposition 3.2.2. Algorithm 3 does not require an implicit description of the models M_1 and M_2 so time-consuming elimination computations are avoided. Symbolic computation is also only done to verify that a test set T is in fact a certificate but not to find the candidate set. Proposition 3.3.2 guarantees that if there is a certificate S that can be found symbolically, then it can be found numerically with probability 1 so we can minimize the amount of symbolic computation necessary. Lastly, we are frequently able to find certificates by just randomly searching for them which avoids the combinatorial complexity of computing the whole matroid. The downside of this is that the failure of the algorithm does not imply that the matroids are the same or that a discrete parameter is not identifiable. This type of failure is illustrated by Example 3.5.8.

3.4 Identifiability of 2-tree Mixtures for Generic Group-Based Models

In this section we demonstrate how Algorithm 3 can be used to certify the identifiability of the tree parameters in group-based phylogenetic models. In Section 3.5 we apply the method to the identifiability of phylogenetic network models.

Many of our proofs in the following sections use supplementary files. We will reference relevant supplementary files or methods as needed. All of these files are located at the website:

https://github.com/bkholler/MatroidIdentifiability

Since 2-tree mixture models are a family of algebraic models indexed by 2-multisets of n-leaf trees we have the following definition of generic identifiability for this family.

Definition 3.4.1. The tree parameters of a 2-tree mixture model are *generically identifiable* if for every pair of distinct multisets of n-leaf trees $\{T_1, T_2\}$ and $\{S_1, S_2\}$,

$$\dim((V_{T_1} * V_{T_2}) \cap (V_{S_1} * V_{S_2})) < \min(\dim(V_{T_1} * V_{T_2}),\ \dim(V_{S_1} * V_{S_2})).$$

We now discuss how Algorithm 3 can be specialized for separating 2-tree CFN mixtures of 6-leaf trees and show how it can be used to prove generic identifiability of the tree parameters CFN model when combined with the following theorem of Matsen, Mossel, and Steel [36]. All of the computations involved are available in the supplementary materials.

Theorem 3.4.2. *(Six-To-Infinity Theorem) [36, Theorem 23] Suppose that the tree parameters T_1, T_2 are identifiable for a 2-tree mixture model for trees with six leaves. Then the tree parameters are identifiable for trees with n leaves for all $n \geq 6$.*

Theorem 3.4.3. *The tree parameters of the 2-tree CFN mixture model are generically identifiable for trees with at least 6 leaves.*

Proof. If we can show that the tree parameters are identifiable for 2-tree CFN mixtures of six leaf trees then we are done by the Six-To-Infinity Theorem. Our proof of this is computational and simply an application of Algorithm 3 with some simplifications.

First, we note that instead of comparing every possible pair of 2-multisets of six leaf trees of which there $15,481,830$ it is enough to check up to the symmetry induced by the permutation action of S_6 on leaf labels. Consideration of symmetry reduces the problem to checking only $22,773$ distinct cases.

Next we note that 2-tree CFN mixtures of six leaf trees have the *expected dimension* [3, Proposition 5.5]. This means that for every pair of six leaf trees $\{T_1, T_2\}$, the variety $V_{T_1} * V_{T_2}$ satisfies

$$\dim(V_{T_1} * V_{T_2}) = \dim(V_{T_1}) + \dim(V_{T_2}) + 1 = 19$$

which implies that join varieties of this form have the same dimension regardless of the tree parameters. This means we can use the specialized version of the algorithm for models of the same dimension. Furthermore, as a result of the Fourier transform, the parameters that correspond to the identity element in $\mathbb{Z}_2$ are actually identically 1. By removing them we are able to greatly reduce the number of variables which significantly speeds up the symbolic computation step required for verification of certificates. Our algorithm is able to produce a certificate for all but one of the 22,773 cases. These certificates are stored in the file *certsCFN* and code to verify that they are certificates can be found in the Mathematica file *CFN_6Leaf_Mixtures.nb*. The certificates were originally found using the function *matroidSeparate* in the Mathematica package *PhylogeneticMatroids.m* which is our implementation of Algorithm 3.

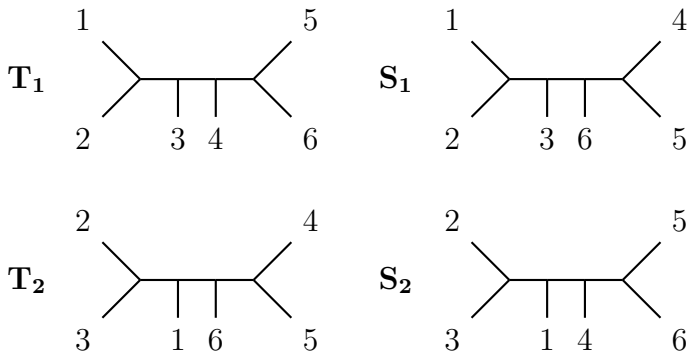


Figure 3.1: The two pairs of trees described by Equation (3.2) which have the same sets of splits when combined.

The case that the algorithm fails to find a certificate has tree parameters $\{T_1, T_2\}$ and $\{S_1, S_2\}$ of the following form up to symmetry

$$\begin{aligned} T_1 &= \{12|3456, 123|456, 1234|56\} \\ T_2 &= \{23|1456, 123|456, 1236|45\} \\ S_1 &= \{12|3456, 123|456, 1236|45\} \\ S_2 &= \{23|1456, 123|456, 1234|56\}. \end{aligned} \tag{3.2}$$

In this case, we were able to find invariants that separate the join varieties by computing a degree-bounded Gröbner basis for the varieties $V_{T_1} * V_{T_2}$ and $V_{S_1} * V_{S_2}$ up to degree 4. These computations can be found in *CFN_last_pair.m2*. This separates all pairs up to symmetry and so the tree parameters are identifiable for six leaf trees. □

A natural question to ask is why the more typical Gröbner basis algorithm that was employed to deal with the last case could not simply be used to deal with every case. This is because even the degree bounded Gröbner basis calculation can take a significant amount of time compared to our algorithm. For instance if we take

$$\begin{aligned} T_1 &= \{12|3456, 125|346, 1256|34\} \\ T_2 &= \{13|2456, 134|256, 1346|25\} \\ S_1 &= \{12|3456, 126|345, 1246|35\} \\ S_2 &= \{15|2346, 156|234, 1356|24\}. \end{aligned}$$

then computing a Gröbner basis up to degree four took slightly over eight minutes whereas

our algorithm took slightly under four minutes in this case. This computational difference is quite significant given the large number of cases that need to be dealt with.

For the main computation we ran Algorithm 3 on each case in batches of about 1000 cases over a month. We do not have a precise time estimate for how long this computation took but in the 22,772 cases where the algorithm worked, it seems to find potential certificates quickly and most of the computation time came from computing matrix rank over the fraction field $k(\theta)$. On the other hand, using Algorithm 4 with tolerance $\epsilon = 10^{-10}$, we can find a list of certificates in slightly over 19 minutes running the algorithm in parallel on a laptop with four processors.

The identifiability of the tree parameters for 2-tree K3P mixtures actually follows almost immediately from the CFN case but we are also able to use our method along with some results from [2] to get identifiability results for smaller trees in the K3P case.

Theorem 3.4.4. *The tree parameters of the 2-tree K3P mixture model are generically identifiable for trees with at least four leaves.*

Proof. The generic identifiability of the tree parameters of 2-tree K3P mixtures for trees with at least six leaves follows immediately from Theorem 3.4.3. This is because the CFN model can be obtained from the K3P model via a coordinate projection. More explicitly, let $\{T_1, T_2\}$ and $\{S_1, S_2\}$ be two distinct multisets of six leaf trees and suppose $V_{T_1}^{K3P} * V_{T_2}^{K3P}$ and $V_{S_1}^{K3P} * V_{S_2}^{K3P}$ are the join varieties associated to the K3P model. Theorem 3.4.3 guarantees that the same varieties associated the CFN model satisfy

$$V_{T_1}^{CFN} * V_{T_2}^{CFN} \neq V_{S_1}^{CFN} * V_{S_2}^{CFN}.$$

Let G be a subgroup of $\mathbb{Z}_2 \times \mathbb{Z}_2$ isomorphic to $\mathbb{Z}_2$, and $\pi : \mathbb{C}^{4^n} \to \mathbb{C}^{2^n}$ be the linear map obtained by projecting onto the coordinates of $\mathbb{C}^{4^n}$ indexed only by the elements of G. Then for any tree T, $\pi(V_T^{K3P}) = V_T^{CFN}$. For example, let $G = \langle (1,0) \rangle \subseteq \mathbb{Z}_2 \times \mathbb{Z}_2$ and let $\pi(V_T^{K3P})$ be the projection onto these coordinates. Then $\pi(V_T^{K3P}) \subseteq \mathbb{C}^{2^n}$ is parameterized by the map

$$q_{g_1,\dots g_n} = \begin{cases} \prod_{A|B \in \Sigma(T)} a_{\sum_{i \in A} g_i}^{A|B} & \text{, if } \sum_{i \in [n]} g_i = 0 \\ 0 & \text{, otherwise} \end{cases}$$

where the g_i are in G. Since $G \cong \mathbb{Z}_2$ we could simply replace every occurrence of $(1,0) \in \mathbb{Z}_2 \times \mathbb{Z}_2$ in this map with $1 \in \mathbb{Z}_2$ without changing the map at all. The resulting parameterization would be exactly the parameterization of V_T^{CFN} and so we see that the two varieties are parameterized by the same map. Since linear maps commute with taking

joins of varieties, it holds that

$$\pi\left(V_{T_1}^{K3P} * V_{T_2}^{K3P}\right) = V_{T_1}^{CFN} * V_{T_2}^{CFN}.$$

This together with the inequality of the 2-tree CFN join varieties implies the inequality of the 2-tree K3P join varieties and so the generic identifiability of the tree parameters for trees with at least six leaves follows.

For trees with four leaves we once again apply Algorithm 3 to all distinct 2-multisets of four leaf trees up to symmetry. In all four cases the algorithm quickly finds a certificate numerically but in the last case it seems to only find certificates that are very large. When the potential certificate sets are large, the verification step can still be time consuming so in this case we instead constructed a smaller certificate that we verified symbolically. The five leaf case then follows from Proposition 7 of [2] which guarantees that if $\{T_1, T_2\}$ and $\{S_1, S_2\}$ are distinct multisets of five leaf trees, then there exists a 4-subset $K \subseteq [5]$ of the leaves such that restricting each tree to K gives two distinct multisets of four leaf trees or in symbols $\{T_1|_K, T_2|_K\} \neq \{S_1|_K, S_2|_K\}$. The result then follows from Lemma 3 of [2] which shows that if $V_{T_1|_K} * V_{T_2|_K} \not\subseteq V_{S_1|_K} * V_{S_2|_K}$ then $V_{T_1} * V_{T_2} \not\subseteq V_{S_1} * V_{S_2}$. □

In the proof of Theorem 3.4.3, there was a single pair of trees up to symmetry that our matroid-based algorithm failed to find a certificate for. We also attempted to run the algorithm for the same pair of trees under the K3P model and also did not find a certificate but in both cases we know the corresponding ideals of phylogenetic invariants are not equal. As mentioned in section 3.3, it is possible for two different ideals to define the same matroid. We conjecture this to be the case in this instance.

Conjecture 3.4.5. *Let $\{T_1, T_2\}$ and $\{S_1, S_2\}$ be the pairs of trees defined in Equation (3.2) and let $V_{T_1} * V_{T_2}$ and $V_{S_1} * V_{S_2}$ be the associated CFN join varieties. Then*

$$\mathcal{M}(V_{T_1} * V_{T_2}) = \mathcal{M}(V_{S_1} * V_{S_2}).$$

3.5 Identifiability for Phylogenetic Networks

Recently phylogenetic network models have emerged as a tool to account for events in the evolutionary history of organisms that trees cannot represent. Non-treelike evolutionary processes include horizontal gene transfer and hybridization [35, 46]. Similar to the case of trees, an important question to address is the identifiability of the network parameter in

network-based phylogenetic models. [21] showed that the network parameter is identifiable in *large-cycle* JC network models by explicitly computing the associated ideals. We will show how Algorithm 3 can be used to extend their results to large-cycle K2P and K3P network models.

Definition 3.5.1. The set of large-cycle networks is the collection of all k-cycle networks with $k \geq 4$.

Definition 3.5.2. The large-cycle network parameter of a phylogenetic network model is generically identifiable if for every pair of n-leaf large-cycle networks N_1 and N_2,

$$\dim(V_{N_1} \cap V_{N_2}) < min(\dim(V_{N_1}), \dim(V_{N_2}))$$

We now describe the proof strategy that Gross and Long used to prove the generic identifiability of the network parameter for large-cycle JC networks. As they remarked in [21], the combinatorial arguments they make to prove the final result still apply but the necessary computational results are more difficult since K2P and K3P are higher dimensional models with more parameters.

Let M be a phylogenetic model for which the tree parameter is generically identifiable. Gross and Long showed in [21, Section 4.2] that if Lemma 3.5.3, Lemma 3.5.4, and Lemma 3.5.5 also hold for M, then the large-cycle network parameter is identifiable for M. They prove this by finding subsets of the leaves of the networks that when restricted to, yield a situation that can be addressed with one of the lemmas or the generic identifiability of the tree parameter. In [21], they proved that the three following lemmas hold when M is the JC model by computing a degree-bounded Gröbner basis for I_N and then verifying that the degree-bounded basis generates a prime ideal of the correct dimension, thus it must be a Gröbner basis for the prime ideal I_N. This computation becomes more difficult though as the number of parameters in the model increases. We instead use Algorithm 3 to prove these three lemmas also hold for the K2P and K3P models as well. For the remainder of this chapter we let M be either K2P or K3P and denote the variety associated to the network N under the model M with V_N^M.

Lemma 3.5.3. *Let N_1 be a k_1-cycle network and N_2 be a k_2 cycle network. If $2 \leq k_1 < k_2 \leq 4$, then $V_{N_2}^M \not\subseteq V_{N_1}^M$.*

Proof. We prove this by explicitly computing dimensions of the associated varieties. If V_N^M, is a network variety parameterized by ψ_N^M, then the dimension of V_N^M can be computed

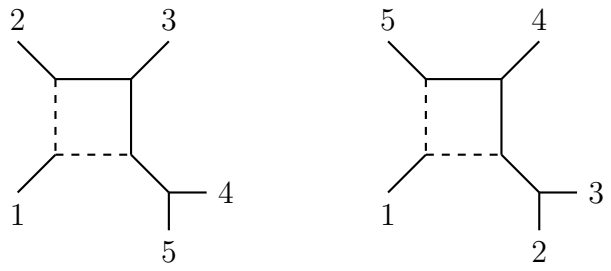


Figure 3.2: The two possibilities for N_1 in Lemma 3.5.5.

by calculating the rank of the Jacobian of ψ_N^M over the fraction field $k(\theta)$. In each case, we find that $\dim(V_{N_2}) > \dim(V_{N_1})$ which implies $V_{N_2}^M \not\subseteq V_{N_1}^M$. These computations can be found in the Mathematica files *K2P_Networks.nb* and *K3P_Networks.nb*. □

Lemma 3.5.4. *Let N_1 and N_2 be distinct 4-leaf 4-cycle networks. Then $V_{N_2}^M \not\subseteq V_{N_1}^M$ and $V_{N_1}^M \not\subseteq V_{N_2}^M$.*

Proof. In this case V_{N_1} and V_{N_2} both have the same dimension so we can run the specialized version of Algorithm 3. For both models we ran *matroidSeparate* and were once again able to find a certificate separating each pair of 4-leaf 4-cycle networks. These computations can also be found in the Mathematica files *K2P_Networks.nb* and *K3P_Networks.nb*. □

Lemma 3.5.5. *Let N_1 be either of the two 5-leaf 4-cycle networks pictured in Figure 3.2 and let N_2 be the 5-leaf 5-cycle network with reticulation edges directed toward the leaf-labelled by 1. Then $V_{N_1} \not\subseteq V_{N_2}$.*

Proof. V_{N_1} and V_{N_2} once again have the same dimension in this case so we can again run the specialized version of Algorithm 3 to show $V_{N_1} \not\subseteq V_{N_2}$ for both possible choices of N_1. As before, we ran *matroidSeparate* to find certificates that show $V_{N_1} \not\subseteq V_{N_2}$. These computations can also be found in the Mathematica files *K2P_Networks.nb* and *K3P_Networks.nb*. □

Corollary 3.5.6. *The semi-directed network parameter of large-cycle K2P and K3P network models is generically identifiable.*

Proof. Since Lemmas 3.5.3, 3.5.4, 3.5.5 hold for K2P and K3P cycle-networks, Lemmas 4.11, 4.12, and 4.13 of [21] hold for K2P and K3P networks as well. This means for any

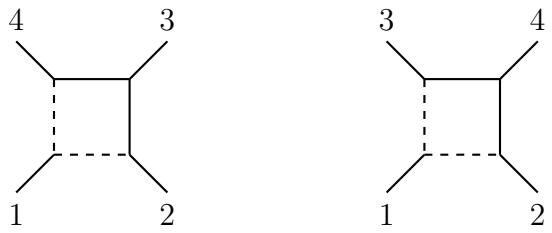


Figure 3.3: 4-cycle networks with four leaves. Under the CFN model these two networks have different ideals but the same matroid.

two large-cycle networks N_1 and N_2, $V_{N_1} \not\subseteq V_{N_2}$ and $V_{N_2} \not\subseteq V_{N_1}$. Since these varieties are irreducible, this mutual non-containment implies

$$\dim(V_{N_1}^M \cap V_{N_2}^M) < \min(\dim(V_{N_1}^M), \dim(V_{N_2}^M))$$

and so the semi-directed network parameter of large-cycle K2P and K3P network models is generically identifiable. □

Remark 3.5.7. *In our original computations we were also able to separate the 3-cycle networks from the 4-cycle networks for both the K2P and K3P models. It may be possible to extend these identifiability results to cycle networks with cycle size at least 3. As previously mentioned though, it will always be impossible for trees to be generically identifiable from cycle networks.*

This serves as another example of how Algorithm 3 can be used to obtain identifiability results for discrete parameters in algebraic models. While this algorithm has nice computational advantages over computing vanishing ideals, there can be times when it fails to separate varieties whose intersection is actually lower dimensional. It is important to remember that when this algorithm fails to separate two models, it does not imply that the discrete parameter is not identifiable. The example below shows that even if we compute the entire matroid of both models, we still may not be able to separate models whose intersection is actually lower dimensional.

Example 3.5.8. Let N_1 and N_2 be the networks pictured on the left and right in Figure 3.3 respectively. We can directly compute the vanishing ideals I_{N_1} and I_{N_2} of the CFN

network models on N_1 and N_2 via elimination and get

$$I_{N_1} = \langle q_{0110}q_{1001} - q_{0101}q_{1010} + q_{0011}q_{1100} - q_{0000}q_{1111} \rangle$$
$$I_{N_2} = \langle -q_{0110}q_{1001} + q_{0101}q_{1010} + q_{0011}q_{1100} - q_{0000}q_{1111} \rangle.$$

These ideals are of the same dimension and not equal so the intersection of their corresponding varieties is lower dimensional. Despite that, we can compute their entire matroid explicitly and see that they are equal. This stems from the fact that the polynomials that generate I_{N_1} and I_{N_2} involve the same variables.

CHAPTER

4

INVARIANTS FOR LEVEL-1 PHYLOGENETIC NETWORKS UNDER THE CAVENDAR-FARRIS-NEYMAN MODEL

In this Chapter we focus on finding the invariants of the Cavendar-Farris-Neyman (CFN) model on level-1 phylogenetic networks. This means that we are trying to determine a generating set for the vanishing ideal I_N of the CFN network model on the level-1 phylogenetic network N. Recall that the discrete Fourier transform, which is used to simplify the parameterization of group-based models, such as the CFN model, can also be applied to network models as well [21]. After applying this transform, CFN tree models become *toric varieties* but the same is not true for CFN network models which makes analyzing their algebraic structure more difficult.

As observed in [21], the toric fiber product of [44] can still be applied to group-based network models. Our approach leverages this toric fiber product structure to reduce the

problem to that of finding the invariants for *sunlet networks* which consist of only a single cycle. While sunlet network varieties are still not toric, they do have a lower-dimensional torus action on them meaning they are *T-varieties* [26]. We use this torus action to break up the ideal of invariants of a n-leaf sunlet network into homogeneous graded pieces we call *gloves*.

We then explicitly produce all quadratic generators of the sunlet network ideal that lie in a given graded piece which gives a complete set of quadratic generators of the sunlet network ideal under the CFN model. We conjecture that the sunlet network ideal is generated by quadratics which would imply our set of quadratic generators actually generate the entire ideal. An implementation of our algorithm to find quadratic generators and computational evidence for our conjectures can be found at:

`https://github.com/bkholler/CFN_Networks`.

This Chapter is organized as follows. In Section 2 we utilize the toric fiber product to reduce the problem of finding a generating set for I_N to the problem of finding a generating set for $I_{\mathcal{S}_n}$ where $\mathcal{S}_n$ is the n-leaf sunlet network. In Section 3, we give a complete description of the quadratic invariants for any sunlet network. In Section 4, we discuss some open problems and conjectures concerning network ideals and give some possible directions for approaching them. In particular, we conjecture that the CFN sunlet network ideal is generated by quadratics and is dimension $2n$ when the network has n leaves.

4.1 Reduction to Sunlet Networks

In this section, we show that gluing level-1 networks together along a leaf corresponds to a toric fiber product of their corresponding ideals. This was pointed out in [21] but the authors do not prove it. We include a more detailed discussion and the proof here for completeness. This means that the ideal of invariants for any network can be constructed by taking toric fiber products of sunlet networks and trees.

Let N be a level-1 network and observe that we can either find an edge e such that when e is cut, N is split into two new networks N_- and N_+ where N_- and N_+ are level-1 networks with fewer leaves or that no such e exists in which case N is a sunlet network or 3-leaf tree. We can of course recover the network N by gluing N_- and N_+ along the edge e which is a leaf of both new networks. We denote the operation of gluing these networks along a leaf edge as $N = N_- * N_+$. This operation is pictured in Figure 4.1.

We now assume N does admit a decomposition $N = N_- * N_+$ and denote the ambient polynomial rings of these networks with $\mathbb{C}[q]$, $\mathbb{C}[q]_-$, $\mathbb{C}[q]_+$. Note that their corresponding ideals I_N, I_{N_-}, I_{N_+} are all homogeneous in the grading determined by $\deg(q_{\mathbf{g}}) = e_{g_e}$ where e_{g_e} is the corresponding standard basis vector.

Example 4.1.1. Let N_- be the corresponding network pictured in Figure 4.1 then

$$\mathbb{C}[q]_- = \mathbb{C}[q_{\mathbf{g}} \mid \mathbf{g} = (g_1, g_2, g_3, g_e) \in \mathbb{Z}_2 \text{ and } g_1 + g_2 + g_3 + g_e = 0]$$

and one can compute explicitly that

$$I_{N_-} = \langle q_{0000}q_{1111} - q_{0011}q_{1100} + q_{0101}q_{1010} - q_{0110}q_{1001} \rangle \subseteq \mathbb{C}[q]_-.$$

We can clearly see that this polynomial is homogeneous of degree $e_0 + e_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ by simply examining the last entry of the label sequence of each monomial.

Proposition 4.1.2. *Assume N is not a sunlet network or 3-leaf tree and let $N = N_- * N_+$ be a decomposition of N into two smaller level-1 networks. Let each variable $q_{\mathbf{g}}$ in $\mathbb{C}[q]$, $\mathbb{C}[q]_-$, $\mathbb{C}[q]_+$ have degree e_{g_e}. Then I_N is the toric fiber product:*

$$I_N = I_{N_-} \times_{\mathcal{A}} I_{N_+}$$

with $\mathcal{A} = \{e_0, e_1\}$ linearly independent.

Proof. We prove this by slightly modifying the parameterization ψ_N and then factoring it which is a standard technique introduced in [44]. Recall that for a tree T, I_T can be thought of as the kernel of the map

$$\psi_T : \mathbb{C}[q] \to \mathbb{C}[a_g^i \mid g \in \mathbb{Z}_2, \quad i \in E(N)]$$

given by Equation 1.4 and I_N is then the kernel of the map

$$\psi_N = \sum_{\sigma \in \{0,1\}^m} \left(\prod_{i=1}^{m} \lambda_i^{1-\sigma_i} (1 - \lambda_i)^{\sigma_i} \right) \psi_{T_\sigma}.$$

Note that squaring the variables associated to the edge e, which are $a_{g_e}^e$, everywhere they appear does not change the parameterization. Furthermore, the edge e which we have

glued along is an edge in every tree T_σ and so we can also split each T_σ along this edge to get two new trees T_σ^+ and T_σ^-. Then we have from [44, Theorem 3.10] that

$$\psi_{T_\sigma}(q_{\mathbf{g}}) = \psi_{T_\sigma^-}(q_{\mathbf{g}})\psi_{T_\sigma^+}(q_{\mathbf{g}}). \tag{4.1}$$

That is the parameterization for the tree T_σ factors as a product of the parameterizations for the trees T_σ^+ and T_σ^-.

Without loss of generality let $v_1, \ldots, v_\ell$ be the reticulation vertices of N that lie in N_- and $v_{l+1}, \ldots, v_m$ be those that lie in N_+. Then we can substitute Equation 4.1 into ψ_N and regroup to get

$$\psi_N(q_{\mathbf{g}}) = \sum_{\sigma\in\{0,1\}^m} \left[\left(\prod_{i=1}^{\ell} \lambda_i^{1-\sigma_i}(1-\lambda_i)^{\sigma_i}\right)\psi_{T_\sigma^-}(q_{\mathbf{g}})\right]\left[\left(\prod_{i=\ell+1}^{m} \lambda_i^{1-\sigma_i}(1-\lambda_i)^{\sigma_i}\right)\psi_{T_\sigma^+}(q_{\mathbf{g}})\right] =$$
$$\left(\sum_{\sigma\in\{0,1\}^\ell}\left(\prod_{i=1}^{\ell} \lambda_i^{1-\sigma_i}(1-\lambda_i)^{\sigma_i}\right)\psi_{T_\sigma^-}(q_{\mathbf{g}})\right)\left(\sum_{\sigma\in\{0,1\}^{m-\ell}}\left(\prod_{i=\ell+1}^{m} \lambda_i^{1-\sigma_i}(1-\lambda_i)^{\sigma_i}\right)\psi_{T_\sigma^+}(q_{\mathbf{g}})\right)$$
$$= \psi_{N_-}(q_{\mathbf{g}})\psi_{N_+}(q_{\mathbf{g}})$$

since trees T_σ^- and T_σ^+ are exactly the trees that appear in the parameterization of ψ_{N_-} and ψ_{N_+} respectively.

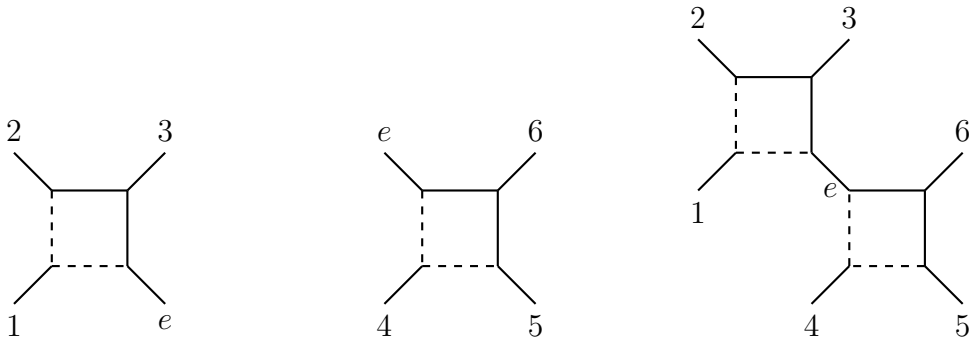
This implies that ψ_N factors through the map

$$\phi : \mathbb{C}[q] \to \mathbb{C}[q]_- \otimes \mathbb{C}[q]_+$$
$$q_{\mathbf{g}} \mapsto q_{\mathbf{g}_-} \otimes q_{\mathbf{g}_+}$$

and thus I_N is the desired toric fiber product. □

Remark 4.1.3. *The exact same proof can be used to extend the above proposition to all group-based models on level-1 phylogenetic networks. We present it in terms of the CFN model here since that is the main focus of this chapter.*

The above proposition gives an immediate algorithm for constructing the ideal I_N if the ideals for all sunlet networks and trees are known. The original network N is recursively decomposed into sunlet networks and trees. One then builds the ideal back up by taking toric fiber products of the sunlet network ideals and tree ideals. Since the ideals corresponding to trees are completely known, the problem of finding the ideal I_N now amounts to understanding the sunlet network ideals $I_{\mathcal{S}_n}$. This is our main focus for



(a) N_- (b) N_+ (c) N

Figure 4.1: We can glue two four leaf networks along identified leaves to get a six leaf network. This corresponds to taking a toric fiber product of the corresponding ideals.

the remainder of this chapter.

4.2 Quadratic Invariants of Sunlet Networks

The goal of this section is to describe the quadratic phylogenetic invariants for sunlet networks. This is done by leveraging a $\mathbb{Z}^{n+1}$-grading on their corresponding phylogenetic ideals. Throughout this section, we will consider the n-sunlet network $\mathcal{S}_n$ whose edges are labelled as follows: the leaves are labelled $e_1, \ldots, e_n$ going clockwise starting with the leaf adjacent to the reticulation vertex; all cycle edges are labelled $e_{n+1}, \ldots, e_{2n}$ oriented clockwise starting and ending at the reticulation vertex. The phylogenetic variety $V_{\mathcal{S}_n}$ is the variety associated to the kernel of the ring homomorphism $\psi_n : R_n \to S_n$ where

$$R_n = \mathbb{C}[q_{g_1,\ldots,g_n} \mid (g_1, \ldots, g_n) \in \mathbb{Z}_2^n \text{ and } \sum_{i=1}^{n} g_i = 0],$$
$$S_n = \mathbb{C}[a_g^i \mid g \in \mathbb{Z}_2 \text{ and } 1 \leq i \leq 2n],$$

and ψ_n is defined by $q_{g_1,\ldots,g_n} \mapsto \prod_{j=1}^{n} a_{g_j}^j \left(\prod_{j=1}^{n-1} a^{n+j}_{\sum_{\ell=1}^{j} g_\ell} + \prod_{j=2}^{n} a^{n+j}_{\sum_{\ell=2}^{j} g_\ell} \right)$. The kernel of ψ_n will be denoted by I_n throughout the rest of the Chapter.

Proposition 4.2.1. *The homomorphism ψ_n is graded by $\mathbb{Z}^{n+1}$.*

Proof. R_n is graded by $\mathbb{Z}^{n+1}$ as follows: $\deg(q_{g_1 \ldots g_n} = (1, g_1, \ldots, g_n)$ where the g_i are

considered as elements of $\mathbb{Z}$. S_n can also be given a $\mathbb{Z}^{n+1}$-grading as follows:

$$\deg(a^j_{g_j}) = \begin{cases} \mathbf{0} & \text{if } j > n \\ (1, g_1, 0, \ldots, 0) & \text{if } j = 1 \\ (0, 0, \ldots, g_j, \ldots, 0) & \text{if } 2 \leq j \leq n \end{cases}$$

where again the g_i's are considered as elements of $\mathbb{Z}$ instead of $\mathbb{Z}_2$. One checks that the $\deg(q_{\mathbf{g}}) = \deg(\psi_n(q_{\mathbf{g}}))$, and the claim follows. □

Remark 4.2.2. *We have shown that the coordinate ring of the n-sunlet variety is graded by $\mathbb{Z}^{n+1}$. In particular, this makes the variety into a T-variety: there is a $\mathcal{T} \cong (\mathbb{C}^\times)^{n+1}$-action on the variety. We note that $\mathcal{S}_n$ does not yield a toric variety since in general $\dim(\mathcal{T}) < \dim \mathcal{S}_n$.*

4.2.1 Quadratic phylogenetic invariants for sunlet networks

In this subsection, we will leverage the grading from Proposition 4.2.1 to find all quadratic invariants of $\mathcal{S}_n$. While this approach differs from the standard description of phylogenetic invariants of trees in terms of splits, we shall see that our approach not only produces quadratic invariants for n-sunlets, but can also be used to describe the invariants for trees as well (Section 4.2.2).

Throughout this section, let $\psi_n : R_n \to S_n$ be the parameterization of the network variety $\mathcal{S}_n$ as defined in Section 4.2.1, and let $I_n = \ker \psi_n$. We begin with a definition.

Definition 4.2.3. Fix $\mathcal{F} \subseteq [n]$ and $\mathbf{a} \in \mathbb{Z}_2^{\mathcal{F}}$. The *glove*, $\mathcal{G}(n, \mathcal{F}, \mathbf{a})$, is the $\mathbb{C}$-vector space spanned by all quadratic monomials $q_{\mathbf{g}}q_{\mathbf{h}}$ in R_n so that $\mathbf{g}|_{\mathcal{F}} = \mathbf{h}|_{\mathcal{F}} = \mathbf{a}$ and $\mathbf{g}|_{\mathcal{F}^c} + \mathbf{h}|_{\mathcal{F}^c} = \mathbf{1}$ where $\mathbf{1}$ is the all ones vector in $\mathbb{Z}_2^{\mathcal{F}^c}$. If $\mathcal{F} = \emptyset$, then we simply write $\mathcal{G}(n, \emptyset)$.

Remark 4.2.4. *It is not efficient to consider all possible gloves since for some choices of $\mathcal{F}$ and $\mathbf{a}$, the corresponding glove intersects I_n trivially. In fact, given a glove, $\mathcal{G}(n, \mathcal{F}, \mathbf{a}) \subseteq R_n$, if $\mathcal{G}(n, \mathcal{F}, \mathbf{a}) \cap I_n \neq \{0\}$, then $|[n] \setminus \mathcal{F}| \geq 4$ and is even. In order to prove the claim, we first show that if $|[n] \setminus \mathcal{F}|$ is odd, then $\mathcal{G}(n, \mathcal{F}, \mathbf{a}) = \{0\}$. Indeed, if one considers a monomial $q_{\mathbf{g}}q_{\mathbf{h}} \in \mathcal{G}(n, \mathcal{F}, \mathbf{a})$, then it is not possible for $\sum_{i=1}^n g_i$ and $\sum_{i=1}^n h_i$ to both be 0; hence, no such monomial exists. Now, suppose that $|[n] \setminus \mathcal{F}| = 0$ or 2. In either case, $\dim_{\mathbb{C}}(\mathcal{G}(n, \mathcal{F}, \mathbf{a})) = 1$. Then as $\psi_n(q_{\mathbf{g}}) \neq 0$ for any $\mathbf{g}$ and as S_n is an integral domain, all non-trivial polynomials from $\mathcal{G}(n, \mathcal{F}, \mathbf{a})$ lie outside the kernel of ψ_n.*

Remark 4.2.5. *Note that when $n \geq 4$ and is even, $\dim_{\mathbb{C}}(\mathcal{G}(n,\emptyset)) = 2^{n-2}$. One way to see this is to note that the indices $\{\mathbf{g},\mathbf{h}\}$ that appear in the chosen basis for $\mathcal{G}(n,\emptyset)$ are exactly the cosets of $\langle \mathbf{1} \rangle \leq \{\mathbf{g} \in \mathbb{Z}_2^n \mid \sum_{i=1}^n g_i = 0\}$.*

With respect to the $\mathbb{Z}^{n+1}$ grading from Section 4.2.1, each glove $\mathcal{G}(n,\mathcal{F},\mathbf{a})$ is $(R_n)_{\mathbf{c}}$ where $c_1 = 2$ and $c_{i+1} = 1$ if $i \notin \mathcal{F}$ and $c_{i+1} = 2a_i$ when $i \in \mathcal{F}$. Moreover, this encompasses all graded components whose total degree is 2. Therefore, in order to describe all quadratic phylogenetic invariants of $\mathcal{S}_n$, it is enough to find a basis for $\mathcal{G}(n,\mathcal{F},\mathbf{a}) \cap I_n$ for each choice of $\mathcal{F}$ and $\mathbf{a}$ where $|[n] \setminus \mathcal{F}| = 2k$ for all k in $\{1,\ldots,\lfloor \frac{n}{2} \rfloor\}$.

In order to state the main result of this section, we need to define two linear maps obtained out of a glove $\mathcal{G}(n,\mathcal{F},\mathbf{a})$. Consider two following subsets of $[n]$:

$$\mathbb{E}(n,\mathcal{F}) = \{i \mid |[i] \setminus \mathcal{F}| \text{ is even and } 2 \leq i \leq n-1\}$$
$$\mathbb{O}(n,\mathcal{F}) = \{i \mid |[i] \setminus \mathcal{F}| \text{ is odd and } 2 \leq i \leq n-1\}.$$

When n and $\mathcal{F}$ are clear from context, we will just write $\mathbb{E}$ and $\mathbb{O}$, respectively. Using these subsets of $\{2,\ldots,n-1\}$, we color the monomials lying in $\mathcal{G}(n,\mathcal{F},\mathbf{a})$ in two ways. If we have a monomial lying in $\mathcal{G}(n,\mathcal{F},\mathbf{a})$, and we know that one of the factors is $q_{\mathbf{g}}$, then the other factor is determined by $\mathbf{g}$. Thus, it is convenient for us to only record "half" of each term, so we set

$$L(n,\mathcal{F},\mathbf{a}) = \{\mathbf{g} \mid q_{\mathbf{g}}q_{\mathbf{h}} \in \mathcal{G}(n,\mathcal{F},\mathbf{a}) \text{ and } \mathbf{g} <_{\text{lex}} \mathbf{h}\}.$$

If $q_{\mathbf{g}}q_{\mathbf{h}} \in \mathcal{G}(n,\mathcal{F},\mathbf{a})$ and $\mathbf{g} \in L(n,\mathcal{F},\mathbf{a})$, we define our two colorings as follows.

$$c_{\mathbb{E}}(q_{\mathbf{g}}q_{\mathbf{h}}) = \left(\sum_{i=1}^{j} g_i\right)_{j \in \mathbb{E}} \in \mathbb{Z}_2^{\mathbb{E}}$$
$$c_{\mathbb{O}}(q_{\mathbf{g}}q_{\mathbf{h}}) = \left(\sum_{i=1}^{j} g_i\right)_{j \in \mathbb{O}} \in \mathbb{Z}_2^{\mathbb{O}}$$

Now we can define our two maps $M_{\mathbb{E}}^{n,\mathcal{F},\mathbf{a}} : \mathcal{G}(n,\mathcal{F},\mathbf{a}) \to \mathbb{C}^{\mathbb{Z}_2^{\mathbb{E}}}$ and $M_{\mathbb{O}}^{n,\mathcal{F},\mathbf{a}} : \mathcal{G}(n,\mathcal{F},\mathbf{a}) \to \mathbb{C}^{\mathbb{Z}_2^{\mathbb{O}}}$. With respect to the bases $\{q_{\mathbf{g}}q_{\mathbf{h}} \mid \mathbf{g} \in L(n,\mathcal{F},\mathbf{a}\}$, $\{e_{\mathbf{c}} \mid \mathbf{c} \in \mathbb{Z}_2^{\mathbb{E}}\}$, and $\{e_{\mathbf{c}} \mid \mathbf{c} \in \mathbb{Z}_2^{\mathbb{O}}\}$, these maps have the following matrix representations.

$$(M_{\mathbb{E}}^{n,\mathcal{F},\mathbf{a}})_{(\mathbf{c},q_{\mathbf{g}}q_{\mathbf{h}})} = \begin{cases} 1 & \text{if } \mathbf{c} = c_{\mathbb{E}}(q_{\mathbf{g}}q_{\mathbf{h}}) \\ 0 & \text{otherwise} \end{cases}$$

and

$$(M_{\mathbb{O}}^{n,\mathcal{F},\mathbf{a}})_{(\mathbf{c},q_{\mathbf{g}}q_{\mathbf{h}})} = \begin{cases} 1 & \text{if } \mathbf{c} = c_{\mathbb{O}}(q_{\mathbf{g}}q_{\mathbf{h}}) \\ 0 & \text{otherwise} \end{cases}.$$

At this point, we are fully equipped to state the main theorem of this section; however, we will delay the proof until Section 4.2.3.

Theorem 4.2.6. *Let $\mathcal{G}(n,\mathcal{F},\mathbf{a})$ be a glove so that either 1 is not in $\mathcal{F}$ or 1 is in $\mathcal{F}$ but $a_1 = 1$. Then*

$$I_n \cap \mathcal{G}(n,\mathcal{F},\mathbf{a}) = \ker M_{\mathbb{E}}^{n,\mathcal{F},\mathbf{a}} \cap \ker M_{\mathbb{O}}^{n,\mathcal{F},\mathbf{a}}.$$

On the other hand, if 1 is in $\mathcal{F}$ and $a_1 = 0$, then

$$I_n \cap \mathcal{G}(n,\mathcal{F},\mathbf{a}) = \ker M_{\mathbb{E}}^{n,\mathcal{F},\mathbf{a}}.$$

Remark 4.2.7. *As we shall see in Section 4.2.2, Theorem 4.2.6 can be reformulated as follows: $f \in I_n \cap \mathcal{G}(n,\mathcal{F},\mathbf{a})$ if and only if f is a phylogenetic invariant for both underlying trees. If we let I_{T_0} and I_{T_1} be the defining ideals for the two underlying trees, then it is always true that I_n is contained in the intersection of I_{T_0} and I_{T_1}; however, in general, I_n is* not *the intersection of these two toric ideals as can be seen even when $n = 4$. Indeed, the ideals for the two underlying trees are given by*

$$I_{T_0} = \langle q_{0011}q_{1100} - q_{0000}q_{1111}, q_{0110}q_{1001} - q_{0101}q_{1010} \rangle$$

$$I_{T_1} = \langle q_{0101}q_{1010} - q_{0011}q_{1100}, q_{0110}q_{1001} - q_{0000}q_{1111} \rangle.$$

However, $I_{T_0} \cap I_{T_1}$ is generated by one quadratic and one quartic, while J_4 is generated by just the quadratic.

Proposition 4.2.8. *If n is at least 4 and is even, then $\dim_{\mathbb{C}}(I_n \cap \mathcal{G}(n,\emptyset)) = (2^{n/2-1}-1)^2$. Moreover, as long as $1 \notin \mathcal{F}$ or $1 \in \mathcal{F}$ but $a_1 = 1$, $I_n \cap \mathcal{G}(n,\mathcal{F},\mathbf{a}) \cong J_{n-|\mathcal{F}|} \cap \mathcal{G}(n,\emptyset)$.*

Proof. For the first claim, note that n must be even; otherwise, $\mathcal{G}(n,\emptyset)$ is trivial. By Theorem 4.2.6, $I_n \cap \mathcal{G}(n,\emptyset)$ is the intersection of $\ker M_{\mathbb{E}}^{n,\emptyset}$ and $\ker M_{\mathbb{O}}^{n,\emptyset}$. Let $M^{n,\emptyset}$ be the map $M_{\mathbb{E}}^{n,\emptyset} \oplus M_{\mathbb{O}}^{n,\emptyset} : \mathcal{G}(n,\emptyset) \to \mathbb{C}^{\mathbb{Z}_2^{\mathbb{E}}} \oplus \mathbb{C}^{\mathbb{Z}_2^{\mathbb{O}}}$, so $I_n \cap \mathcal{G}(n,\emptyset) = \ker M^{n,\emptyset}$.

We will demonstrate that $\dim_{\mathbb{C}}(I_n \cap \mathcal{G}(n,\emptyset)) = (2^{n/2-1}-1)^2$ by showing that the rank of $M^{n,\emptyset}$ is $2^{n/2}-1$. Then, as $\dim_{\mathbb{C}}(\mathcal{G}(n,\emptyset)) = 2^{n-2}$, we will see by rank-nullity that $\dim_{\mathbb{C}}(I_n \cap \mathcal{G}(n,\emptyset)) = 2^{n-2} - 2^{n/2} + 1 = (2^{n/2-1}-1)^2$.

Note that $|\mathbb{E}| = |\mathbb{O}| = \frac{n}{2} - 1$. If we think of $M^{n,\emptyset}$ as a matrix, its columns are indexed by monomials $q_{\mathbf{g}}q_{\mathbf{h}} \in \mathcal{G}(n, \emptyset)$, and its first $2^{n/2-1}$ rows are indexed by the elements of $\mathbb{Z}_2^{\mathbb{E}}$ and the last $2^{n/2-1}$ rows are indexed by $\mathbb{Z}_2^{\mathbb{O}}$. We claim that the matrix for $M^{n,\emptyset}$ takes the following form: (1) every column is of the form $e_{\mathbf{c}_1} + e_{\mathbf{c}_2}$ where $\mathbf{c}_1 \in \mathbb{Z}_2^{\mathbb{E}}$ and $\mathbf{c}_2 \in \mathbb{Z}_2^{\mathbb{O}}$, (2) each column is distinct, and (3) every possible combination of $e_{\mathbf{c}_1} + e_{\mathbf{c}_2}$ occurs.

The first point is clear by the definition of the maps $M_{\mathbb{E}}^{n,\emptyset}$ and $M_{\mathbb{O}}^{n,\emptyset}$. For the second and third points, we will show that for any $\mathbf{c}_1 \in \mathbb{Z}_2^{\mathbb{E}}$ and $\mathbf{c}_2 \in \mathbb{Z}_2^{\mathbb{O}}$ there is a unique $q_{\mathbf{g}}q_{\mathbf{h}} \in \mathcal{G}(n, \emptyset)$ so that $c_{\mathbb{E}}(q_{\mathbf{g}}q_{\mathbf{h}}) = \mathbf{c}_1$ and $c_{\mathbb{O}}(q_{\mathbf{g}}q_{\mathbf{h}}) = \mathbf{c}_2$. Note that uniqueness will follow immediately since if there were two monomials whose colors are $\mathbf{c}_1$ and $\mathbf{c}_2$, then they must be the same since $\mathbf{c}_1$ and $\mathbf{c}_2$ record all the partial sums of each of their corresponding group elements. We will build up $\mathbf{g} \in L(n, \mathcal{F}, \mathbf{a})$ whose partial sums are given by $\mathbf{c}_1$ and $\mathbf{c}_2$. Let $\mathbf{c} \in \mathbb{Z}_2^{\{2,\dots,n-1\}}$ be the unique vector with $\mathbf{c}|_{\mathbb{E}} = \mathbf{c}_1$ and $\mathbf{c}|_{\mathbb{O}} = \mathbf{c}_2$. If we let $\widetilde{\mathbf{c}} = (0, \mathbf{c}, 0) \in \mathbb{Z}_2^n$, then we set $g_i = \widetilde{c}_i + \widetilde{c}_{i-1}$ for $i \geq 2$ and $g_1 = 0$. One can see that $\sum_{i=1}^{j} g_j = c_j$ for any $2 \leq j \leq n-1$. In order to get a monomial in the glove, we consider $q_{\mathbf{g}}q_{\mathbf{1}+\mathbf{g}} \in \mathcal{G}(n, \emptyset)$. By construction, $c_{\mathbb{E}}(q_{\mathbf{g}}q_{\mathbf{1}+\mathbf{g}}) = \mathbf{c}_1$ and $c_{\mathbb{O}}(q_{\mathbf{g}}q_{\mathbf{1}+\mathbf{g}}) = \mathbf{c}_2$.

Now, we can show that the row rank of $M^{n,\emptyset}$ is one less than the number of rows. Up to scaling there is only one linear relation among the rows which is given by adding up the first $2^{n/2-1}$ rows and subtracting off the last $2^{n/2-1}$ rows. Points (2) and (3) above guarantee that this is the only relation among the rows. Since the rank of $M^{n,\emptyset}$ is $2^{n/2} - 1$ and $\dim_{\mathbb{C}} \mathcal{G}(n, \emptyset) = 2^{n-2}$, we have that $\dim_{\mathbb{C}}(I_n \cap \mathcal{G}(n, \emptyset)) = (2^{n/2-1} - 1)^2$.

For the second statement fix a glove $\mathcal{G}(n, \mathcal{F}, \mathbf{a})$. First, suppose that $\sum_{i \in \mathcal{F}} a_i = 0$. Then for any $\mathbf{g} \in \mathbb{Z}_2^{n-|\mathcal{F}|}$, define $\mathbf{g}(\mathcal{F}, \mathbf{a}) \in \mathbb{Z}_2^n$ as $\mathbf{g}(\mathcal{F}, \mathbf{a})|_{\mathcal{F}} = \mathbf{a}$ and $\mathbf{g}(\mathcal{F}, \mathbf{a})|_{\mathcal{F}^c} = \mathbf{g}$. Then define a linear map $T : \mathcal{G}(n, \emptyset) \to \mathcal{G}(n, \mathcal{F}, \mathbf{a})$ defined by $T(q_{\mathbf{g}}q_{\mathbf{h}}) = q_{\mathbf{g}(\mathcal{F},\mathbf{a})}q_{\mathbf{1}+\mathbf{g}(\mathcal{F},\mathbf{a})}$. T is an isomorphism, and it is not hard to see that there is a map which makes the diagram commute and is an isomorphism when restricted to the images of the horizontal maps.

$$\begin{array}{ccc} \mathcal{G}(n - |\mathcal{F}|, \emptyset) & \xrightarrow{M^{n-|\mathcal{F}|,\emptyset}} & \mathbb{C}^{\mathbb{E}(n-|\mathcal{F}|,\emptyset)} \oplus \mathbb{C}^{\mathbb{O}(n-|\mathcal{F}|,\emptyset)} \\ \downarrow T & & \downarrow \\ \mathcal{G}(n, \mathcal{F}, \mathbf{a}) & \xrightarrow{M^{n,\mathcal{F},\mathbf{a}}} & \mathbb{C}^{\mathbb{E}(n,\mathcal{F})} \oplus \mathbb{C}^{\mathbb{O}(n,\mathcal{F})} \end{array}$$

It then follows that $I_n \cap \mathcal{G}(n, \mathcal{F}, \mathbf{a}) \cong J_{n-|\mathcal{F}|} \cap \mathcal{G}(n, \emptyset)$ in this case. The other case, when $\sum_{i \in \mathcal{F}} a_i = 1$, is exactly the same except $\mathbf{g}(\mathcal{F}, \mathbf{a})$ is defined as $\mathbf{g}(\mathcal{F}, \mathbf{a})|_{\mathcal{F}} = \mathbf{a}$ and $\mathbf{g}(\mathcal{F}, \mathbf{a})|_{\mathcal{F}^c} = \mathbf{g} + e_{n-|\mathcal{F}|}$. □

By the propsoition, in order to find a basis for $I_n \cap \mathcal{G}(n, \mathcal{F}, \mathbf{a})$ when 1 is not in $\mathcal{F}$ or 1 is in $\mathcal{F}$ but $a_1 = 1$, it is enough to find a basis for $J_{n-|\mathcal{F}|} \cap \mathcal{G}(n - |\mathcal{F}|, \emptyset)$ and then apply

the map T. In the next proposition, we provide an explicit basis for $I_n \cap \mathcal{G}(n, \emptyset)$ for any even n greater than or equal to 4.

Theorem 4.2.9. *Fix an even integer $n \in \mathbb{Z}_{\geq 4}$, and a group element $\mathbf{c} \in \mathbb{Z}_2^{\{2,\ldots,n-1\}}$ so that $\mathbf{c}|_{\mathbb{E}(n,\emptyset)} \neq \mathbf{0}$ and $\mathbf{c}|_{\mathbb{O}(n,\emptyset)} \neq \mathbf{0}$. Then we define the polynomial*

$$f_{\mathbf{c}} = q_{\mathbf{g}(\mathbf{0},\mathbf{0})} q_{\mathbf{h}(\mathbf{0},\mathbf{0})} - q_{\mathbf{g}(\mathbf{c}|_{\mathbb{E}},\mathbf{0})} q_{\mathbf{h}(\mathbf{c}|_{\mathbb{E}},\mathbf{0})} + q_{\mathbf{g}(\mathbf{c}|_{\mathbb{E}},\mathbf{c}|_{\mathbb{O}})} q_{\mathbf{h}(\mathbf{c}|_{\mathbb{E}},\mathbf{c}|_{\mathbb{O}})} - q_{\mathbf{g}(\mathbf{0},\mathbf{c}|_{\mathbb{O}})} q_{\mathbf{h}(\mathbf{0},\mathbf{c}|_{\mathbb{O}})}$$

in $I_n \cap \mathcal{G}(n, \emptyset)$. Here $\mathbf{g}(\mathbf{c}', \mathbf{c}'')$ is defined by setting $g_1 = 0$ and for $i \geq 2$ we have that $g_i = c_{i-1} + c_i$ where $\mathbf{c} \in \mathbb{Z}_2^n$ has $c_1 = c_n = 0$ and $\mathbf{c}|_{\mathbb{E}} = \mathbf{c}'$ and $\mathbf{c}|_{\mathbb{O}} = \mathbf{c}''$, and $\mathbf{h}(\mathbf{c}_1, \mathbf{c}_2) = \mathbf{1} + \mathbf{g}(\mathbf{c}_1, \mathbf{c}_2)$. Then

$$\mathcal{B}_n = \{f_{\mathbf{c}} \mid \mathbf{c} \in \mathbb{Z}_2^{\{2,\ldots,n-1\}} \text{ and } \mathbf{c}|_{\mathbb{E}} \neq \mathbf{0}, \mathbf{c}|_{\mathbb{O}} \neq \mathbf{0}\}$$

is a basis for $I_n \cap \mathcal{G}(n, \emptyset)$.

Proof. Note that by definition $f_{\mathbf{c}} \in \mathcal{G}(n, \mathcal{F}, \mathbf{a})$. To see that $f_{\mathbf{c}} \in I_n$, note that

$$M^{n,\emptyset}(f_{\mathbf{c}}) = e_{\mathbf{0}|_{\mathbb{E}}} + e_{\mathbf{0}|_{\mathbb{O}}} - e_{\mathbf{c}|_{\mathbb{E}}} - e_{\mathbf{0}|_{\mathbb{O}}} + e_{\mathbf{c}|_{\mathbb{E}}} + e_{\mathbf{c}|_{\mathbb{O}}} - e_{\mathbf{0}|_{\mathbb{E}}} - e_{\mathbf{c}|_{\mathbb{O}}} = 0.$$

By Theorem 4.2.6, $f_{\mathbf{c}} \in I_n$.

Since $|\mathcal{B}_n|$ is $(2^{n/2-1} - 1)^2$, it is enough show that $\mathcal{B}_n$ is independent. Consider any linear combination of the elements of $\mathcal{B}_n$

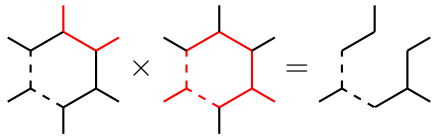
$$0 = \sum_{\mathbf{c}} a_{\mathbf{c}} f_{\mathbf{c}}.$$

Projecting $\sum_{\mathbf{c}} a_{\mathbf{c}} f_{\mathbf{c}}$ onto $\text{span}_{\mathbb{C}}\{q_{\mathbf{g}(\mathbf{c}|_{\mathbb{E}},\mathbf{c}|_{\mathbb{O}})} q_{\mathbf{h}(\mathbf{c}|_{\mathbb{E}},\mathbf{c}|_{\mathbb{O}})}\}$, yields $a_{\mathbf{c}} q_{\mathbf{g}(\mathbf{c}|_{\mathbb{E}},\mathbf{c}|_{\mathbb{O}})} q_{\mathbf{h}(\mathbf{c}|_{\mathbb{E}},\mathbf{c}|_{\mathbb{O}})}$ from which it follows that $a_{\mathbf{c}} = 0$ for all such $\mathbf{c}$. □

Remark 4.2.10. *Let J_n be the ideal generated by all quadratics in I_n. Then Propositions 4.2.8 and 4.2.9 give a recipe for obtaining generators of $I_n \cap \mathcal{G}(n, \mathcal{F}, \mathbf{a})$ where either $1 \notin F$ or $1 \in \mathcal{F}$ but $a_1 = 0$. The case when $1 \in \mathcal{F}$ and $a_1 = 0$ is easily taken care of using previously known techniques. In this case, the parameterization restricts to a monomial map. These phylogenetic invariants are obtained from the underlying tree T in $\mathcal{S}_n$ where all the edges containing the reticulation vertex are deleted. Of course, this tree only has $n-1$ leaves, so we lift these phylogenetic invariants to the network via the map defined by $q_{\mathbf{g}} \mapsto q_{(0,\mathbf{g})}$. These facts along with Propositions 4.7 and 4.8 allow us to find all quadratic*

generators of the sunlet network ideal very quickly. Our implementation of this can be found in the Macaulay2 file `sunletQuadGens.m2`.

In [8], the authors produce a combinatorial interpretation of the phylogenetic invariants of a tree T in terms of systems of paths on T. In a similar vein, each variable $q_{\mathbf{g}}$ can be thought of as a system of paths on the network. The paths connecting the vertices ℓ such that $g_\ell = 1$ though are not necessarily unique. Indeed, there is a unique system of paths connecting all such vertices for each tree. For a monomial, $q_{\mathbf{g}}$, we consider all the edges in the network which are supported in either of these two path systems. Now, we fix a glove $\mathcal{G}(n, \mathcal{F}, \mathbf{a})$ so that $1 \notin \mathcal{F}$. For any monomial $q_{\mathbf{g}}q_{\mathbf{h}} \in \mathcal{G}(n, \mathcal{F}, \mathbf{a})$, we take the symmetric difference of the collection of edges obtained from each monomial. Below is an example with $q_{001100}q_{100010} \in \mathcal{G}(\{2, 6\}, (0, 0)) \subset R_6$.



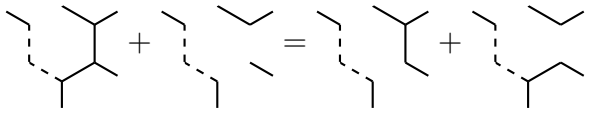
Note that in this example, the leaves which are omitted correspond to $\mathcal{F} = \{2, 6\}$. Note $\mathbb{E} = \{3, 5\}$, $\mathbb{O} = \{2, 4\}$, $c_{\mathbb{E}}(q_{001100}q_{100010}) = (1, 0) \in \mathbb{Z}_2^{\{3,5\}}$, and $c_{\mathbb{O}}(q_{001100}q_{100010}) = (0, 0) \in \mathbb{Z}_2^{\{2,4\}}$. Putting these two colorings together gives us $(0, 1, 0, 0) \in \mathbb{Z}_2^{\{2,3,4,5\}}$. We see that the 1 in the coloring indicates that e_{6+3} should be removed while the zeros in positions $2, 4$, and 5 indicate that the edges e_{6+2}, e_{6+4}, and e_{6+5} should remain in the resulting diagram. In fact, these observations hold true as long as $1 \notin \mathcal{F}$. Therefore, we define the diagram for $q_{\mathbf{g}}q_{\mathbf{h}} \in \mathcal{G}(n, \mathcal{F}, \mathbf{a})$ (for any $\mathcal{F}$) by omitting any leaves which are in $\mathcal{F}$ and any edge e_{n+k} when the coloring of the monomial in position k is 1. These diagrams gives us a visual interpretation of the colorings $c_{\mathbb{E}}$ and $c_{\mathbb{O}}$.

Example 4.2.11. These diagrams give us an easy way to tell if an element $f \in \mathcal{G}(n, \mathcal{F}, \mathbf{a})$ is in an invariant. For example, take $f = q_{101111}q_{111000} - q_{101011}q_{111100} + q_{101101}q_{111010} - q_{101110}q_{111001} \in \mathcal{G}(\{1, 3\}, (1, 1)) \subset R_6$. Here $\mathbb{E} = \{4\}$ and $\mathbb{O} = \{2, 3, 5\}$. Then $c_{\mathbb{E}}$ and $c_{\mathbb{O}}$ on each monomial is as follows.

$$
\begin{aligned}
c_{\mathbb{E}}(q_{101000}q_{111111}) &= 0 & \qquad c_{\mathbb{O}}(q_{101000}q_{111111}) &= (1, 0, 0)\\
c_{\mathbb{E}}(q_{101011}q_{111100}) &= 0 & \qquad c_{\mathbb{O}}(q_{101011}q_{111100}) &= (1, 0, 1)\\
c_{\mathbb{E}}(q_{101101}q_{111010}) &= 1 & \qquad c_{\mathbb{O}}(q_{101101}q_{111010}) &= (1, 0, 1)\\
c_{\mathbb{E}}(q_{101110}q_{111001}) &= 1 & \qquad c_{\mathbb{O}}(q_{101110}q_{111001}) &= (1, 0, 0)
\end{aligned}
$$

Pictorially, this is as follows.

$$\psi_6(q_{101111}q_{111000} + q_{101101}q_{111010}) = \psi_6(q_{101011}q_{111100} + q_{101110}q_{111001})$$



We can tell that $f \in J_6 \cap \mathcal{G}(\{1,3\},(1,1))$ by noting that the odd colors, $(1,0,0)$ and $(1,0,1)$, and the even colors, 0 and 1, appear once on each side of the equation, i.e. $M_{\mathbb{E}}^{\{1,3\},(1,1)}(f)$ and $M_{\mathbb{O}}^{\{1,3\},(1,1)}(f)$ are both 0.

On the other hand, one can also see that $J_6 \cap \mathcal{G}(\{1,3\},(1,1))$ contains no binomials of the form

$$z_1 q_{\mathbf{g}_1} q_{\mathbf{h}_1} - z_2 q_{\mathbf{g}_2} q_{\mathbf{h}_2}$$

for any suitable group elements and complex numbers $z_i \in \mathbb{C} \setminus \{0\}$. The reason being that if this were to vanish under ψ_6 that would mean that $z_1 = z_2$ and the colorings of each $q_{\mathbf{g}_i} q_{\mathbf{h}_i}$ would need to be identical, but this would imply that $\mathbf{g}_1 = \mathbf{g}_2$ and $\mathbf{h}_1 = \mathbf{h}_2$.

Example 4.2.12. Let $n = 6$ and $\mathcal{F} = \emptyset$. In this case, $\mathcal{G}(6,\emptyset)$ is spanned by the following 16 monomials.

$$q_{000000}q_{111111}, q_{000011}q_{111100}, q_{000101}q_{111010}, q_{000110}q_{111001},$$

$$q_{001001}q_{110110}, q_{001010}q_{110101}, q_{001100}q_{110011}, q_{001111}q_{110000},$$

$$q_{010001}q_{101110}, q_{010010}q_{101101}, q_{010100}q_{101011}, q_{010111}q_{101000},$$

$$q_{011000}q_{100111}, q_{011011}q_{100100}, q_{011101}q_{100010}, q_{011110}q_{100001}$$

We have the following matrices where the columns are indexed by the monomials above and the rows are indexed by elements of $\mathbb{Z}_2^2$ lexicographically.

$$M_{\mathbb{E}}^{6,\emptyset} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

$$M_{\mathbb{O}}^{6,\emptyset} = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Then $J_6 \cap \mathcal{G}(6,\emptyset)$ is a 9 dimensional $\mathbb{C}$-vector space spanned by the following polynomials.

$$q_{000000}q_{111111} - q_{000110}q_{111001} + q_{000101}q_{111010} - q_{000011}q_{111100}$$

$$q_{000000}q_{111111} - q_{000110}q_{111001} + q_{001010}q_{110101} - q_{001100}q_{110011}$$

$$q_{000000}q_{111111} - q_{000110}q_{111001} + q_{001001}q_{110110} - q_{001111}q_{110000}$$

$$q_{000000}q_{111111} - q_{011000}q_{100111} + q_{011011}q_{100100} - q_{000011}q_{111100}$$

$$q_{000000}q_{111111} - q_{011000}q_{100111} + q_{010100}q_{101011} - q_{001100}q_{110011}$$

$$q_{000000}q_{111111} - q_{011000}q_{100111} + q_{010111}q_{101000} - q_{001111}q_{110000}$$

$$q_{000000}q_{111111} - q_{011110}q_{100001} + q_{011101}q_{100010} - q_{000011}q_{111100}$$

$$q_{000000}q_{111111} - q_{011110}q_{100001} + q_{010010}q_{101101} - q_{001100}q_{110011}$$

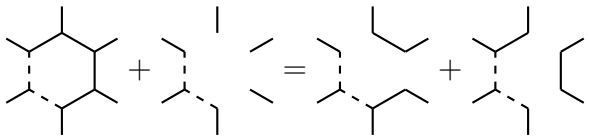
$$q_{000000}q_{111111} - q_{011110}q_{100001} + q_{010001}q_{101110} - q_{001111}q_{110000}$$

Let us consider the colorings of the monomials in the last polynomial. Note $\mathbb{E} = \{2,4\}$ and $\mathbb{O} = \{3,5\}$

$$c_{\mathbb{E}}(q_{000000}q_{111111}) = (0,0) \qquad c_{\mathbb{O}}(q_{000000}q_{111111}) = (0,0)$$
$$c_{\mathbb{E}}(q_{011110}q_{100001}) = (1,1) \qquad c_{\mathbb{O}}(q_{011110}q_{100001}) = (0,0)$$
$$c_{\mathbb{E}}(q_{010001}q_{101110}) = (1,1) \qquad c_{\mathbb{O}}(q_{010001}q_{101110}) = (1,1)$$
$$c_{\mathbb{E}}(q_{001111}q_{110000}) = (0,0) \qquad c_{\mathbb{O}}(q_{001111}q_{110000}) = (1,1)$$

This relation can be viewed pictorially as

$$\psi_6(q_{000000}q_{111111} + q_{010001}q_{101110}) = \psi_6(q_{011110}q_{100001} + q_{001111}q_{110000})$$



We also note that the dimension of $\mathcal{S}_6$ is 12, its codimension is 20, and J_6 is minimally generated by 79 polynomials; thus, contrary to say the 4-leaf case, $\mathcal{S}_6$ is not a complete intersection even set-theoretically.

4.2.2 Quadratic phylogenetic invariants of trees

Let T be a binary tree with leaf set $[n]$, and let I_T be the defining ideal for the corresponding variety under the CFN model. It was shown in [43] that the phylogenetic invariants for this model are given purely in terms of 2×2 minors of certain matrices. In this section, we give a separate description for the generating set which is in line with the approach from Section 4.2.1. Using the same reasoning as in Section 4.2.1, we can see that I_T is also graded by $\mathbb{Z}^{n+1}$; hence, the quadratic generators can be described by the $\mathbb{C}$-vector spaces $I_T \cap \mathcal{G}(n, \mathcal{F}, \mathbf{a})$ where we can again restrict to when $[n] \setminus \mathcal{F}$ is even has cardinality at least 4. Recall that for any edge $e \in \Sigma(T)$, the edge induces a split of the tree $A_e|B_e$. In this section, given a glove $\mathcal{G}(n, \mathcal{F}, \mathbf{a})$, we define

$$\mathbb{E}_T(\mathcal{F}) = \{e \in E(T) \mid |A_e \setminus \mathcal{F}| \text{ is even}\}.$$

When it is clear from context, we will simply write $\mathbb{E}_T$. Similarly, we let the linear map $M_{\mathbb{E}_T}^{\mathcal{F},\mathbf{a}} : \mathcal{G}(n, \mathcal{F}, \mathbf{a}) \to \mathbb{C}^{\mathbb{Z}_2^{\mathbb{E}_T}}$ be defined by the following matrix as in the previous subsection.

$$(M_{\mathbb{E}_T}^{n,\mathcal{F},\mathbf{a}})_{\mathbf{c},q_\mathbf{g}q_\mathbf{h}} = \begin{cases} 1 & \text{if for all } e \in \mathbb{E}_T,\ c_e = \sum_{i \in A_e} g_i \\ 0 & \text{otherwise} \end{cases}$$

where $\mathbf{g} <_{\text{lex}} \mathbf{h}$. Then we have the following theorem which is analogous to Theorem 4.2.6, but for trees.

Theorem 4.2.13. *Given a glove $\mathcal{G}(n, \mathcal{F}, \mathbf{a})$ and a phylogenetic tree T, the $\mathbb{Z}^{n+1}$-graded piece $I_T \cap \mathcal{G}(n, \mathcal{F}, \mathbf{a})$ is the kernel of $M_{\mathbb{E}_T}^{n,\mathcal{F},\mathbf{a}}$.*

Proof. Let $S_T = \mathbb{C}[a_g^e \mid g \in \mathbb{Z}_2 \text{ and } e \in E(T)]$. Recall that I_T is the kernel of $\psi_T : R_n \to S_T$ defined by

$$q_\mathbf{g} \mapsto \prod_{A_e|B_e \in \Sigma(T)} a^e_{\sum_{i \in A_e} g_i}$$

Now, fix a glove $\mathcal{G}(n, \mathcal{F}, \mathbf{a})$, and note that if $q_\mathbf{g}q_\mathbf{h} \in \mathcal{G}(n, \mathcal{F}, \mathbf{a})$, then $\sum_{i \in A_e} g_i = \sum_{i \in A_e} h_i$ if and only if $e \in \mathbb{E}_T$. Consider any polynomial $f = \sum_{\mathbf{g} \in L(n,\mathcal{F},\mathbf{a})} c_\mathbf{g} q_\mathbf{g} q_\mathbf{h} \in \mathcal{G}(n, \mathcal{F}, \mathbf{a})$. If we

apply ψ_T, we get the following.

$$\begin{aligned}\psi_T(f) &= \sum_{\mathbf{g}\in L(n,\mathcal{F},\mathbf{a})} c_{\mathbf{g}} \left(\prod_{e\in E(T)} a^e_{\sum_{i\in A_e} g_i}\right)\left(\prod_{e\in E(T)} a^e_{\sum_{i\in A_e} h_i}\right)\\ &= \sum_{\mathbf{g}\in L(n,\mathcal{F},\mathbf{a})} c_{\mathbf{g}} \left(\prod_{e\notin \mathbb{E}} a^e_0 a^e_1\right)\left(\prod_{e\in \mathbb{E}_T} (a^e_{\sum_{i\in A_e} g_i})^2\right)\\ &= \left(\prod_{e\notin \mathbb{E}_T} a^e_0 a^e_1\right) \sum_{\mathbf{g}\in L(n,\mathcal{F},\mathbf{a})} c_{\mathbf{g}} \left(\prod_{e\in \mathbb{E}} (a^e_{\sum_{i\in A_e} g_i})^2\right)\end{aligned}$$

The monomials, $\prod_{e\in\mathbb{E}_T}(a^e_{\sum_{i\in A_e} g_i})^2$, can be identified as standard basis vectors in $\mathbb{C}^{\mathbb{Z}_2^{\mathbb{E}_T}}$. After making this identification, it becomes evident that $\psi_T(f) = 0$ if and only if $M^{\mathcal{F},\mathbf{a}}_{\mathbb{E}_T}(f) = 0$. □

Consider $\mathcal{S}_n$ and its two underlying trees T_0 and T_1, and fix any glove $\mathcal{G}(n,\mathcal{F},\mathbf{a})$ where either 1 is not in $\mathcal{F}$ or 1 is in $\mathcal{F}$ but $a_1 = 1$. Recall that T_0 is obtained by deleting the reticulation edge that lies between the leaves e_1 and e_2, and T_1 is obtained by deleting the reticulation edge that lies between the leaves e_1 and e_n. The defining ideals for T_0 and T_1 are generated by quadratic binomials. Here we will show that the polynomials $f_{\mathbf{c}}$ from Proposition 4.2.9 are either sums or differences of binomials coming from I_{T_0} and I_{T_1}. In the following proposition, we only consider the case when n is even, at least 4, and $\mathcal{F} = \emptyset$ since any other glove of the form stated can be obtained from this case.

Proposition 4.2.14. *Let $n \in \mathbb{Z}_{\geq 4}$ be even, and consider a polynomial of the following form*

$$f_{\mathbf{c}} = q_{\mathbf{g}(\mathbf{0},\mathbf{0})}q_{\mathbf{h}(\mathbf{0},\mathbf{0})} - q_{\mathbf{g}(\mathbf{c}|_{\mathbb{E}},\mathbf{0})}q_{\mathbf{h}(\mathbf{c}|_{\mathbb{E}},\mathbf{0})} + q_{\mathbf{g}(\mathbf{c}|_{\mathbb{E}},\mathbf{c}|_{\mathbb{O}})}q_{\mathbf{h}(\mathbf{c}|_{\mathbb{E}},\mathbf{c}|_{\mathbb{O}})} - q_{\mathbf{g}(\mathbf{0},\mathbf{c}|_{\mathbb{O}})}q_{\mathbf{h}(\mathbf{0},\mathbf{c}|_{\mathbb{O}})}$$

in $I_n \cap \mathcal{G}(n,\emptyset)$ from Proposition 4.2.9. Then the binomials $q_{\mathbf{g}(\mathbf{c}|_{\mathbb{E}},\mathbf{0})}q_{\mathbf{h}(\mathbf{c}|_{\mathbb{E}},\mathbf{0})} - q_{\mathbf{g}(\mathbf{c}|_{\mathbb{E}},\mathbf{c}|_{\mathbb{O}})}q_{\mathbf{h}(\mathbf{c}|_{\mathbb{E}},\mathbf{c}|_{\mathbb{O}})}$ and $q_{\mathbf{g}(\mathbf{0},\mathbf{0})}q_{\mathbf{h}(\mathbf{0},\mathbf{0})} - q_{\mathbf{g}(\mathbf{0}\mathbf{c}|_{\mathbb{O}})}q_{\mathbf{h}(\mathbf{0},\mathbf{c}|_{\mathbb{O}})}$ are elements of I_{T_0}, and the binomials $q_{\mathbf{g}(\mathbf{0},\mathbf{0})}q_{\mathbf{h}(\mathbf{0},\mathbf{0})} - q_{\mathbf{g}(\mathbf{c}|_{\mathbb{E}},\mathbf{0})}q_{\mathbf{h}(\mathbf{c}|_{\mathbb{E}},\mathbf{0})}$ and $q_{\mathbf{g}(\mathbf{c}|_{\mathbb{E}},\mathbf{c}|_{\mathbb{O}})}q_{\mathbf{h}(\mathbf{c}|_{\mathbb{E}},\mathbf{c}|_{\mathbb{O}})} - q_{\mathbf{g}(\mathbf{0},\mathbf{c}|_{\mathbb{O}})}q_{\mathbf{h}(\mathbf{0},\mathbf{c}|_{\mathbb{O}})}$ are elements of I_{T_1}.

Proof. Note that $\mathbb{E} = \mathbb{E}_{T_0}$ and $\mathbb{O} = \mathbb{E}_{T_1}$. Then the claim follows by Theorem 4.2.13. □

4.2.3 Proof of Theorem 4.2.6

The goal of this section is to prove Theorem 4.2.6. This is done in three cases:

1. when $1 \in \mathcal{F}$ and $a_1 = 0$.

2. when $1 \in \mathcal{F}$ and $a_1 = 1$,

3. and when $1 \notin \mathcal{F}$.

Case (1) follows from Theorem 4.2.13. Indeed, consider a monomial $q_{\mathbf{g}}$ so that $g_1 = 0$. Let ψ_T be the parameterization for the tree obtained from $\mathcal{S}_n$ by deleting all edges adjacent to the reticulation vertex. Then the parameterization is as follows.

$$\begin{aligned}\psi_n(q_{\mathbf{g}}) &= \prod_{i=1}^{n} a_{g_i}^i \left(\prod_{i=1}^{n-1} a_{\sum_{\ell=1}^{i} g_i}^{n+i} + \prod_{i=2}^{n} a_{\sum_{\ell=2}^{i} g_i}^{n+i} \right) \\ &= a_0^1(a_0^{n+1} + a_0^{2n}) \prod_{i=2}^{n} a_{g_i}^i \prod_{i=2}^{n-1} a_{\sum_{\ell=2}^{i} g_i}^{n+i} \\ &= a_0^1(a_0^{n+1} + a_0^{2n}) \psi_T(q_{\mathbf{g}})\end{aligned}$$

Note that the term, $a_0^1(a_0^{n+1} + a_0^{2n})$, does not affect the kernel, so any glove where $1 \in \mathcal{F}$ and $a_1 = 0$ has the desired form.

Consider the second case when $1 \in \mathcal{F}$ and $a_1 = 1$. For each monomial, $q_{\mathbf{g}} q_{\mathbf{h}}$ in $\mathcal{G}(n, \mathcal{F}, \mathbf{a})$, both g_1 and h_1 are 1. We will always assume that $\mathbf{g} <_{\text{lex}} \mathbf{h}$, so $\mathbf{g} \in L(n, \mathcal{F}, \mathbf{a})$. As $\mathbf{h}$ is completely determined by $\mathbf{g}, n, \mathcal{F}$, and $\mathbf{a}$, any polynomial f in $\mathcal{G}(n, \mathcal{F}, \mathbf{a})$ can be written as follows: $f = \sum_{\mathbf{g} \in L(n,\mathcal{F},\mathbf{a})} c_{\mathbf{g}} q_{\mathbf{g}} q_{\mathbf{h}} \in \mathcal{G}(n, \mathcal{F}, \mathbf{a})$. Computing $\psi_n(f)$ yields that $\psi_n(f) = \sum_{\mathbf{g} \in L(n,\mathcal{F},\mathbf{a})} c_{\mathbf{g}} \psi_n(q_{\mathbf{g}} q_{\mathbf{h}}) =$

$$\sum_{\mathbf{g} \in L(n,\mathcal{F},\mathbf{a})} c_{\mathbf{g}} \left(\prod_{j=1}^{n} a_{g_j}^j \right) \left(\prod_{j=1}^{n-1} a_{\sum_{\ell=1}^{j} g_\ell}^{n+j} + \prod_{j=2}^{n} a_{\sum_{\ell=2}^{j} g_\ell}^{n+j} \right) \times \left(\prod_{j=1}^{n} a_{h_j}^j \right) \left(\prod_{j=1}^{n-1} a_{\sum_{\ell=1}^{j} h_\ell}^{n+j} + \prod_{j=2}^{n} a_{\sum_{\ell=2}^{j} h_\ell}^{n+j} \right)$$

Which we can rewrite as

$$\sum_{\mathbf{g} \in L(n,\mathcal{F},\mathbf{a})} c_{\mathbf{g}} \left(\prod_{j \in \mathcal{F}} (a_{g_j}^j)^2 \right) \left(\prod_{j \notin \mathcal{F}} a_0^j a_1^j \right) \left(\prod_{j=1}^{n-1} a_{\sum_{\ell=1}^{j} g_\ell}^{n+j} + \prod_{j=2}^{n} a_{\sum_{\ell=2}^{j} g_\ell}^{n+j} \right) \times \left(\prod_{j=1}^{n-1} a_{\sum_{\ell=1}^{j} h_\ell}^{n+j} + \prod_{j=2}^{n} a_{\sum_{\ell=2}^{j} h_\ell}^{n+j} \right)$$

The monomial $\left(\prod_{j\in\mathcal{F}}(a^j_{g_j})^2\right)\left(\prod_{j\notin\mathcal{F}} a^j_0 a^j_1\right)$ depends only on $\mathcal{F}$ and $\mathbf{a}$, so it can be factored out of the sum, and it will be denoted as $m_{\mathcal{F},\mathbf{a}}$. So we get

$$\psi_n(f) = m_{\mathcal{F},\mathbf{a}} \sum_{\mathbf{g}\in L(n,\mathcal{F},\mathbf{a})} c_{\mathbf{g}} \left(\prod_{j=1}^{n-1} a^{n+j}_{\sum_{\ell=1}^{j} g_\ell} + \prod_{j=2}^{n} a^{n+j}_{\sum_{\ell=2}^{j} g_\ell}\right)\left(\prod_{j=1}^{n-1} a^{n+j}_{\sum_{\ell=1}^{j} h_\ell} + \prod_{j=2}^{n} a^{n+j}_{\sum_{\ell=2}^{j} h_\ell}\right).$$

Which can be rewritten as

$$m_{\mathcal{F},\mathbf{a}} \sum_{\mathbf{g}\in L(n,\mathcal{F},\mathbf{a})} c_{\mathbf{g}} \left(a^{n+1}_{g_1}\prod_{j=2}^{n-1} a^{n+j}_{\sum_{\ell=1}^{j} g_\ell} + a^{2n}_{g_1}\prod_{j=2}^{n-1} a^{n+j}_{\sum_{\ell=2}^{j} g_\ell}\right) \times \left(a^{n+1}_{h_1}\prod_{j=2}^{n-1} a^{n+j}_{\sum_{\ell=1}^{j} h_\ell} + a^{2n}_{h_1}\prod_{j=2}^{n-1} a^{n+j}_{\sum_{\ell=2}^{j} h_\ell}\right).$$

This then gives us that

$$m_{\mathcal{F},\mathbf{a}} \sum_{\mathbf{g}\in L(n,\mathcal{F},\mathbf{a})} c_{\mathbf{g}} \left(a^{n+1}_{1}\prod_{j=2}^{n-1} a^{n+j}_{\sum_{\ell=1}^{j} g_\ell} + a^{2n}_{1}\prod_{j=2}^{n-1} a^{n+j}_{\sum_{\ell=2}^{j} g_\ell}\right) \times \left(a^{n+1}_{1}\prod_{j=2}^{n-1} a^{n+j}_{\sum_{\ell=1}^{j} h_\ell} + a^{2n}_{1}\prod_{j=2}^{n-1} a^{n+j}_{\sum_{\ell=2}^{j} h_\ell}\right).$$

We proceed by expanding and regrouping terms using the following observations about the various sums in the subscripts.

- $\sum_{\ell=1}^{j} g_\ell = 1 + \sum_{\ell=2}^{j} g_\ell$ since $g_1 = 1$
- $\sum_{\ell=1}^{j} g_\ell = \sum_{\ell=1}^{j} h_\ell \iff [j]\setminus\mathcal{F}$ has even cardinality.
- $\sum_{\ell=2}^{j} g_\ell = \sum_{\ell=2}^{j} h_\ell \iff [j]\setminus\mathcal{F}$ has even cardinality.
- $\sum_{\ell=1}^{j} g_\ell = \sum_{\ell=2}^{j} h_\ell \iff [j]\setminus\mathcal{F}$ has odd cardinality.
- $\sum_{\ell=2}^{j} g_\ell = \sum_{\ell=1}^{j} h_\ell \iff [j]\setminus\mathcal{F}$ has odd cardinality.

Using the observations above, all subscripts in $\psi_n(f)$ can be written in terms of $\mathbf{g} \in$

$L(n, \mathcal{F}, \mathbf{a})$.

$$\begin{aligned}\psi_n(f) = m_{\mathcal{F},\mathbf{a}} \sum_{\mathbf{g}\in L(n,\mathcal{F},\mathbf{a})} c_{\mathbf{g}} \Bigg(& a_1^{n+1}a_1^{n+1}\prod_{j\in\mathbb{E}}(a^{n+j}_{\sum_{\ell=1}^{j} g_\ell})^2 \prod_{j\in\mathbb{O}} a_0^{n+j}a_1^{n+j} \\ & + a_1^{n+1}a_1^{2n}\prod_{j\in\mathbb{E}} a_0^{n+j}a_1^{n+j}\prod_{j\in\mathbb{O}}(a^{n+j}_{\sum_{\ell=1}^{j} g_\ell})^2 \\ & + a_1^{n+1}a_1^{2n}\prod_{j\in\mathbb{E}} a_0^{n+j}a_1^{n+j}\prod_{j\in\mathbb{O}}(a^{n+j}_{\sum_{\ell=2}^{j} g_\ell})^2 \\ & + a_1^{2n}a_1^{2n}\prod_{j\in\mathbb{E}}(a^{n+j}_{\sum_{\ell=2}^{j} g_\ell})^2 \prod_{j\in\mathbb{O}} a_0^{n+j}a_1^{n+j}\Bigg)\end{aligned}$$

The products, $\prod_{j\in\mathbb{E}} a_0^{n+j}a_1^{n+j}$ and $\prod_{j\in\mathbb{O}} a_0^{n+j}a_1^{n+j}$, depend only on $\mathcal{F}$, and will be denoted $m_{\mathcal{F},\mathbb{E}}$ and $m_{\mathcal{F},\mathbb{O}}$, respectively.

$$\begin{aligned}\psi_n(f) &= m_{\mathcal{F},\mathbf{a}} \sum_{\mathbf{g}\in L(n,\mathcal{F},\mathbf{a})} c_{\mathbf{g}} \Bigg(a_1^{n+1}a_1^{n+1} m_{\mathcal{F},\mathbb{O}} \prod_{j\in\mathbb{E}}(a^{n+j}_{\sum_{\ell=1}^{j} g_\ell})^2 \\ &\qquad + a_1^{n+1}a_1^{2n} m_{\mathcal{F},\mathbb{E}} \prod_{j\in\mathbb{O}}(a^{n+j}_{\sum_{\ell=1}^{j} g_\ell})^2 + a_1^{n+1}a_1^{2n} m_{\mathcal{F},\mathbb{E}} \prod_{j\in\mathbb{O}}(a^{n+j}_{\sum_{\ell=2}^{j} g_\ell})^2 \\ &\qquad + a_1^{2n}a_1^{2n} m_{\mathcal{F},\mathbb{O}} \prod_{j\in\mathbb{E}}(a^{n+j}_{\sum_{\ell=2}^{j} g_\ell})^2 \Bigg) \\ &= m_{\mathcal{F},\mathbf{a}} m_{\mathcal{F},\mathbb{O}} (a_1^{n+1})^2 \sum_{\mathbf{g}\in L(n,\mathcal{F},\mathbf{a})} c_{\mathbf{g}} \prod_{j\in\mathbb{E}}(a^{n+j}_{\sum_{\ell=1}^{j} g_\ell})^2 \\ &\qquad + m_{\mathcal{F},\mathbf{a}} m_{\mathcal{F},\mathbb{O}} (a_1^{2n})^2 \sum_{\mathbf{g}\in L(n,\mathcal{F},\mathbf{a})} c_{\mathbf{g}} \prod_{j\in\mathbb{E}}(a^{n+j}_{\sum_{\ell=2}^{j} g_\ell})^2 \\ &\qquad + m_{\mathcal{F},\mathbf{a}} m_{\mathcal{F},\mathbb{E}} a_1^{n+1}a_1^{2n} \sum_{\mathbf{g}\in L(n,\mathcal{F},\mathbf{a})} c_{\mathbf{g}} \prod_{j\in\mathbb{O}}(a^{n+j}_{\sum_{\ell=1}^{j} g_\ell})^2 \\ &\qquad + m_{\mathcal{F},\mathbf{a}} m_{\mathcal{F},\mathbb{E}} a_1^{n+1}a_1^{2n} \sum_{\mathbf{g}\in L(n,\mathcal{F},\mathbf{a})} c_{\mathbf{g}} \prod_{j\in\mathbb{O}}(a^{n+j}_{\sum_{\ell=2}^{j} g_\ell})^2\end{aligned}$$

In the final expression of the equation above, there are four sums. The superscripts appearing in the first two sums are identical, and similarly, the superscripts appearing in the last two sums are the same. Moreover, they are completely disjoint, so there is no cancellation between the first two sums and the second two sums. It follows that

$\psi_n(f) = 0$ if and only if the following equations hold.

$$0 = (a_1^{n+1})^2 \sum_{\mathbf{g}\in L(n,\mathcal{F},\mathbf{a})} c_{\mathbf{g}} \prod_{j\in\mathbb{E}} (a^{n+j}_{\sum_{\ell=1}^{j} g_\ell})^2 + (a_1^{2n})^2 \sum_{\mathbf{g}\in L(n,\mathcal{F},\mathbf{a})} c_{\mathbf{g}} \prod_{j\in\mathbb{E}} (a^{n+j}_{\sum_{\ell=2}^{j} g_\ell})^2 \tag{4.2}$$

$$0 = \sum_{\mathbf{g}\in L(n,\mathcal{F},\mathbf{a})} c_{\mathbf{g}} \prod_{j\in\mathbb{O}} (a^{n+j}_{\sum_{\ell=1}^{j} g_\ell})^2 + \sum_{\mathbf{g}\in L(n,\mathcal{F},\mathbf{a})} c_{\mathbf{g}} \prod_{j\in\mathbb{O}} (a^{n+j}_{\sum_{\ell=2}^{j} g_\ell})^2 \tag{4.3}$$

In equation (5), there can be no cancellation among these two sums because of the coefficients $(a_1^{n+1})^2$ and $(a_1^{2n})^2$ in front. The subscripts in each of these sums are all off by exactly 1; therefore, the first term is 0 if and only if the second term is 0. In equation (6), the subscripts in each sum are also again off by exactly 1. In order to show there is no cancellation among these sums, we argue that the set of monomials appearing in each sum are disjoint.

Lemma 4.2.15. *There are no distinct $\mathbf{g}, \mathbf{g}' \in L(n, \mathcal{F}, \mathbf{a})$ so that $\sum_{\ell=1}^{j} g_\ell = \sum_{\ell=2}^{j} g'_\ell$ for all $2 \le j \le n-1$ such that $[j] \setminus \mathcal{F}$ has odd cardinality. In other words, in (5), the two sets of monomials appearing in the two sums are disjoint.*

Proof. Let $\{i_1, \ldots, i_m\} = [n-1]\setminus\mathcal{F}$. Suppose $\mathbf{g}, \mathbf{g}' \in L(n, \mathcal{F}, \mathbf{a})$ and $\sum_{\ell=1}^{j} g_\ell = \sum_{\ell=2}^{j} g'_\ell$ for all $j \in \{i_1, \ldots, i_m\}$. Since $g'_1 = 1$, we have $\sum_{\ell=1}^{j} g_\ell = 1 + \sum_{\ell=1}^{j} g'_\ell$ for all $j \in \{i_1, \ldots, i_m\}$. Since $\mathbf{g}|_{\mathbf{a}} = \mathbf{g}'|_{\mathbf{a}}$, we see that $g_{i_1} = 1 + g'_{i_1}$. However, this contradicts that $\mathbf{g}' \in L(n, \mathcal{F}, \mathbf{a})$. Since $L(n, \mathcal{F}, \mathbf{a}) = \{\mathbf{g} \mid q_{\mathbf{g}} q_{\mathbf{h}} \in \mathcal{G}(n, \mathcal{F}, \mathbf{a}) \text{ and } \mathbf{g} <_{lex} \mathbf{h}\}$, there is some $\mathbf{h}'$ so that $q_{\mathbf{g}'} q_{\mathbf{h}'} \in \mathcal{G}(n, \mathcal{F}, \mathbf{a})$, and since $i_1 \notin \mathcal{F}$, $h'_{i_1} = 0$ which implies $\mathbf{h}' <_{\text{lex}} \mathbf{g}'$ and $\mathbf{g}' \notin L(n, \mathcal{F}, \mathbf{a})$. □

From the arguments above, we see that equations (5) and (6) reduce to (7) and (8) below. It then follows that $\psi_n(f) = 0$ if and only if (7) and (8) hold.

$$0 = \sum_{\mathbf{g}\in L(n,\mathcal{F},\mathbf{a})} c_{\mathbf{g}} \prod_{j\in\mathbb{E}} (a^{n+j}_{\sum_{\ell=1}^{j} g_\ell})^2 \tag{4.4}$$

$$0 = \sum_{\mathbf{g}\in L(n,\mathcal{F},\mathbf{a})} c_{\mathbf{g}} \prod_{j\in\mathbb{O}} (a^{n+j}_{\sum_{\ell=1}^{j} g_\ell})^2 \tag{4.5}$$

Equations (7) and (8) imply that f lies in $\ker M_{\mathbb{E}}^{n,\mathcal{F},\mathbf{a}}$ and $\ker M_{\mathbb{O}}^{n,\mathcal{F},\mathbf{a}}$, respectively; thus, $f \in I_n \cap \mathcal{G}(n, \mathcal{F}, \mathbf{a})$ if and only if it lies in the intersection of $\ker M_{\mathbb{E}}^{n,\mathcal{F},\mathbf{a}}$ and $\ker M_{\mathbb{O}}^{n,\mathcal{F},\mathbf{a}}$.

Finally, we can consider the third case, when $1 \notin \mathcal{F}$. Here we will omit the proof, since it is nearly identical to the proof of the second case.

4.3 Open Problems

In this section, we discuss some conjectures for which we have computational evidence and suggest some possible techniques for solving them. We also provide some interesting open problems surrounding sunlet network ideals.

One of the main drawbacks to the techniques used in Section 4.2 is that it only yields quadratic generators for I_n. For n-sunlet networks with $4 \leq n \leq 7$, we have verified that their ideals are quadratically generated. This was done in Macaulay2 by showing that over $\mathbb{Q}$, $\ker \psi_n = J_n$ for $n = 4, 5, 6$, and 7. Since we had equality over $\mathbb{Q}$, the ideals must still be equal after extending to the complex numbers. While we have verified that I_n is generated by quadratics for $4 \leq n \leq 7$, it remains open as to whether these generate I_n for $n \geq 8$. For the CFN model, the ideals for trees are always generated by quadratics, and as we have seen the quadratic invariants obtained for the sunlet ideals are built from invariants from the underlying trees; hence, we suspect that I_n is always quadratically generated.

Conjecture 4.3.1. *Let J_n be the ideal generated by all quadratic invariants in I_n. Then $J_n = I_n$ for all $n \geq 4$.*

In order to prove Conjecture 4.3.1, it would be enough to show that J_n is prime and of the correct dimension. To this end, we have the following conjecture which would prove Conjecture 4.3.1.

Conjecture 4.3.2. *For $n \geq 5$, $\dim I_n = 2n = \dim J_n$ and J_n is prime.*

A possible approach to proving that J_n is prime is that taken in [33]. The main workhorse of their technique is the following lemma which was originally stated in [20, Proposition 23].

Lemma 4.3.3. *[33, Lemma 2.5] Let k be a field and $J \subset k[x_1, \ldots x_n]$ be an ideal containing a polynomial $f = gx_1 + h$ with g, h not involving x_1 and g a non-zero divisor modulo J. Let $J_1 = J \cap k[x_2, \ldots x_n]$ be the elimination ideal. Then J is prime if and only if J_1 is prime.*

This lemma can be used to create a descending chain of ideals each one involving one less variable. As long as a polynomial f of the required form can be found, then one can prove that the original ideal is prime by verifying that the last ideal in the chain is prime. For $4 \leq n \leq 7$ we have done this with J_n by repeatedly eliminating variables in reverse

lexicographic order until we are left with an ideal in only the variables $q_{\mathbf{g}}$ such that $g_1 = 0$. That is we build a chain

$$J_n \supset J_n^{(1)} \supset \cdots \supset J_n^{(2^{n-2})}$$

where $J_n^{(j)}$ is obtained by eliminating the jth variable in reverse lexicographic order from $J_n^{(j-1)}$ and at each step we ensure that a polynomial f of the form described in Lemma 4.3.3 exists. Typically one would then need to verify that $J_n^{(2^{n-2})}$ is prime but the following lemma shows there is no need for this. Our implementation of this can be found in the macaulay2 file `primeDescent.m2`.

Lemma 4.3.4. *Let $J_n^{(2^{n-2})} = J_n \cap \mathbb{C}[q_{\mathbf{g}} : g_1 = 0]$. Then $J_n^{(2^{n-2})} \cong I_T$ where T is the tree obtained by deleting the reticulation vertex of $\mathcal{S}_n$ and all adjacent edges.*

This lemma implies that if one can always find a polynomial f of the desired form in each of the intermediate elimination ideal $J_n^{(j)}$ then J_n is prime since the last ideal $J_n^{(2^{n-2})}$ is isomorphic to a tree ideal; thus, it must be prime.

For the question of the dimension of I_n, we have the following bound.

Proposition 4.3.5. *For $n \geq 4$ it holds that $2n - 1 \leq \dim(I_n) \leq 2n + 1$.*

Proof. First we note that I_n is properly contained in the ideals I_{T_0} and I_{T_1} for the trees T_0 and T_1 that are obtained from $\mathcal{S}_n$ by deleting reticulation edges. It is well known that each of these ideals has $\dim(I_{T_i}) = 2n - 2$ (see for example [4]). Since we have that I_n is a prime ideal properly contained in these two prime ideals which are not equal, we get the lower bound $2n - 1 \leq \dim(I_n)$. For the other bound recall that $V_{\mathcal{S}_n}$ can also be thought of as a projective variety the map $\psi_{\mathcal{S}_n}$ parameterizing I_n can be thought of as a map

$$\psi_{\mathcal{S}_n} : \prod_{e \in E(\mathcal{S}_n)} \mathbb{P}^1 \to \mathbb{P}^{2^{n-1}-1}$$

where each copy of $\mathbb{P}^1$ in the domain corresponds to an edge of $\mathcal{S}_n$. This immediately implies that the projective variety corresponding to $\mathcal{S}_n$ has dimension at most $\#E(\mathcal{S}_n) = 2n$ and so $\dim(I_n) \leq 2n + 1$. □

We also have that $\dim I_n \leq \dim J_n$ as $J_n \subseteq I_n$. Moreover, using the rank of Jacobian of $\psi_{\mathcal{S}_n}$, we have shown for $5 \leq n \leq 8$ that the dimension of I_n is $2n$. We've also computed the rank of the Jacobian with random values substituted in for the parameters for n up to 17. In each case we've found that the rank is also $2n$ which means that $\dim(I_n) = 2n$ with probability 1 for $9 \leq n \leq 17$. These computations can be found in the file `sunletDim.m2`.

Through computational experiments we have found very well-behaved *toric degenerations* of the ideals corresponding to $\mathcal{S}_4$ and $\mathcal{S}_5$. In the case when $n = 5$, there are 116 cones in the tropical variety which yield normal toric varieties; however, most of them are somewhat less well-behaved. For example, we were able to find a weight vector w such that the quadratic invariants produced in Section 4.2 actually form a Gröbner basis with respect to w but this does *not* happen for most of the weights in the tropical variety. Moreover, the initial forms of these quadratic invariants are always invariants for at least one of the underlying trees T_0 or T_1. To this end, we ask the following.

Question 4.3.6. *For $n \geq 5$, is there a weight vector w on R_n for which $in_w(I_n)$ is a prime binomial ideal? If so, can it be shown that there is a combinatorial rule for finding such a w where a Gröbner basis of I_n with respect to w can be deduced combinatorially?*

REFERENCES

[1] David Aldous. Probability distributions on cladograms. In *Random discrete structures (Minneapolis, MN, 1993)*, volume 76 of *IMA Vol. Math. Appl.*, pages 1–18. Springer, New York, 1996.

[2] Elizabeth S Allman, Sonia Petrovic, John A Rhodes, and Seth Sullivant. Identifiability of two-tree mixtures for group-based models. *IEEE/ACM transactions on computational biology and bioinformatics*, 8(3):710–722, 2010.

[3] Hector Baños, Nathaniel Bushek, Ruth Davidson, Elizabeth Gross, Pamela E. Harris, Robert Krone, Colby Long, Allen Stewart, and Robert Walker. Dimensions of group-based phylogenetic mixtures. *Bull. Math. Biol.*, 81(2):316–336, 2019.

[4] Hector Baños, Nathaniel Bushek, Ruth Davidson, Elizabeth Gross, Pamela E. Harris, Robert Krone, Colby Long, Allen Stewart, and Robert Walker. Dimensions of group-based phylogenetic mixtures. *Bull. Math. Biol.*, 81(2):316–336, 2019.

[5] Magali Bardet, Jean-Charles Faugère, and Bruno Salvy. On the complexity of the f5 gröbner basis algorithm. *Journal of Symbolic Computation*, 70:49–70, 2015.

[6] Daniel Irving Bernstein, Lam Si Tung Ho, Colby Long, Mike Steel, Katherine St. John, and Seth Sullivant. Bounds on the expected size of the maximum agreement subtree. *SIAM J. Discrete Math.*, 29(4):2065–2074, 2015.

[7] David Bryant, Andy McKenzie, and Mike Steel. The size of a maximum agreement subtree for random binary trees. In *Bioconsensus (Piscataway, NJ, 2000/2001)*, volume 61 of *DIMACS Ser. Discrete Math. Theoret. Comput. Sci.*, pages 55–65. Amer. Math. Soc., Providence, RI, 2003.

[8] Weronika Buczyńska and Jarosław A. Wiśniewski. On geometry of binary symmetric models of phylogenetic trees. *J. Eur. Math. Soc. (JEMS)*, 9(3):609–635, 2007.

[9] James A Cavender and Joseph Felsenstein. Invariants of phylogenies in a simple case with discrete states. *Journal of classification*, 4(1):57–71, 1987.

[10] Joseph T Chang. Full reconstruction of markov models on evolutionary trees: identifiability and consistency. *Mathematical biosciences*, 137(1):51–73, 1996.

[11] Tomás M. Coronado, Arnau Mir, Francesc Rosselló, and Gabriel Valiente. A balance index for phylogenetic trees based on rooted quartets. *J. Math. Biol.*, 79(3):1105–1148, 2019.

[12] Joseph Cummings, Benjamin Hollering, and Christopher Manon. Invariants for level-1 phylogenetic networks under the cavendar-farris-neyman model. *arXiv preprint arXiv:2102.03431*, 2021.

[13] D. M. de Vienne, T. Giraud, and O.C. Martin. A congruence index for testing topological similarity between trees. *Bioinformatics*, 23:3119–3124, 2007.

[14] Persi Diaconis. Finite forms of de finetti's theorem on exchangeability. *Synthese*, 36(2):271–281, Oct 1977.

[15] Persi Diaconis and Svante Janson. Graph limits and exchangeable random graphs. *Rend. Mat. Appl. (7)*, 28(1):33–61, 2008.

[16] Steven N. Evans and T. P. Speed. Invariants of some probability models used in phylogenetic inference. *Ann. Statist.*, 21(1):355–377, 1993.

[17] Daniel J. Ford. *Probabilities on cladograms: Introduction to the alpha model*. ProQuest LLC, Ann Arbor, MI, 2006. Thesis (Ph.D.)–Stanford University.

[18] Noah Forman. Mass-structure of weighted real trees. *arXiv e-prints*, page arXiv:1801.02700, Jan 2018.

[19] Noah Forman, Chris Haulk, and Jim Pitman. A representation of exchangeable hierarchies by sampling from random real trees. *Probab. Theory Related Fields*, 172(1-2):1–29, 2018.

[20] Luis David Garcia, Michael Stillman, and Bernd Sturmfels. Algebraic geometry of Bayesian networks. *J. Symbolic Comput.*, 39(3-4):331–355, 2005.

[21] Elizabeth Gross and Colby Long. Distinguishing phylogenetic networks. *SIAM Journal on Applied Algebra and Geometry*, 2(1):72–93, 2018.

[22] Elizabeth Gross, Colby Long, and Joseph Rusinko. Phylogenetic Networks. *arXiv e-prints*, page arXiv:1906.01586, Jun 2019.

[23] Bénédicte Haas, Grégory Miermont, Jim Pitman, and Matthias Winkel. Continuum tree asymptotics of discrete fragmentations and applications to phylogenetic models. *Ann. Probab.*, 36(5):1790–1837, 2008.

[24] Joe Harris. *Algebraic geometry*, volume 133 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1992. A first course.

[25] Brendan Hassett. *Introduction to algebraic geometry*. Cambridge University Press, 2007.

[26] Jürgen Hausen, Christoff Hische, and Milena Wrobel. On torus actions of higher complexity. *Forum Math. Sigma*, 7:Paper No. e38, 81, 2019.

[27] Michael D Hendy and David Penny. Complete families of linear invariants for some stochastic models of sequence evolution, with and without the molecular clock assumption. *Journal of Computational Biology*, 3(1):19–31, 1996.

[28] Benjamin Hollering and Seth Sullivant. Identifiability in phylogenetics using algebraic matroids. *J. Symbolic Comput.*, 104:142–158, 2021.

[29] Benjamin Hollering and Seth Sullivant. Exchangeable and sampling-consistent distributions on rooted binary trees. *Journal of Applied Probability*, page 1–21, 2022.

[30] James A Lake. A rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony. *Molecular biology and evolution*, 4(2):167–191, 1987.

[31] S. L. Lauritzen, A. Rinaldo, and K. Sadeghi. On Exchangeability in Network Models. *ArXiv e-prints*, September 2017.

[32] Colby Long and Seth Sullivant. Identifiability of 3-class Jukes-Cantor mixtures. *Adv. in Appl. Math.*, 64:89–110, 2015.

[33] Colby Long and Seth Sullivant. Tying up loose strands: defining equations of the strand symmetric model. *J. Algebr. Stat.*, 6(1):17–23, 2015.

[34] László Lovász. *Large Networks and Graph Limits.*, volume 60 of *Colloquium Publications*. American Mathematical Society, 2012.

[35] Wayne P Maddison. Gene trees in species trees. *Systematic biology*, 46(3):523–536, 1997.

[36] Frederick A. Matsen, Elchanan Mossel, and Mike Steel. Mixed-up trees: the structure of phylogenetic mixtures. *Bull. Math. Biol.*, 70(4):1115–1139, 2008.

[37] Frederick A Matsen and Mike Steel. Phylogenetic mixtures on a single tree can mimic a tree of another topology. *Systematic Biology*, 56(5):767–775, 2007.

[38] Peter McCullagh, Jim Pitman, and Matthias Winkel. Gibbs fragmentation trees. *Bernoulli*, 14(4):988–1002, 2008.

[39] James G. Oxley. *Matroid theory*. Oxford Science Publications. The Clarendon Press, Oxford University Press, New York, 1992.

[40] Zvi Rosen. Computing algebraic matroids. *arXiv preprint arXiv:1403.8148*, 2014.

[41] Mike Steel. *Phylogeny: discrete and random processes in evolution*. SIAM, 2016.

[42] Bernd Sturmfels and Seth Sullivant. Toric ideals of phylogenetic invariants. *Journal of Computational Biology*, 12(2):204–228, 2005.

[43] Bernd Sturmfels and Seth Sullivant. Toric ideals of phylogenetic invariants. *Journal of Computational Biology*, 12(2):204–228, 2005.

[44] Seth Sullivant. Toric fiber products. *J. Algebra*, 316(2):560–577, 2007.

[45] Seth Sullivant. *Algebraic statistics*, volume 194 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2018.

[46] Michael Syvanen. Horizontal gene transfer: evidence and possible consequences. *Annual review of genetics*, 28(1):237–261, 1994.

[47] J Wakeley. *Coalescent Theory: An Introduction*. W.H. Freeman, 2008.

[48] Günter M. Ziegler. *Lectures on polytopes*, volume 152 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1995.