ABSTRACT

LONG, COLBY EDWARD. Algebraic Geometry of Phylogenetic Models. (Under the direction of Seth Sullivant.)

A phylogenetic model is a statistical model of the evolutionary relationships between a group of species. The Zariski closure of the set of probability distributions associated to a phylogenetic model is an algebraic variety and the ideal of this variety is the ideal of phylogenetic invariants for the model. These invariants have proven useful both for phylogenetic reconstruction and for determining properties of the models, such as whether or not the parameters are identifiable. In this thesis, we use the tools of algebraic statistics and algebraic geometry to prove several results for phylogenetic models.

First, we prove that the tree parameters of the 3-class Jukes-Cantor mixture model are identifiable. The proof uses ideas from algebraic statistics, in particular: finding phylogenetic invariants that separate the varieties associated to different triples of trees; computing dimensions of the resulting phylogenetic varieties; and using the disentangling number to reduce to trees with a small number of leaves. Symbolic computation also plays a key role in handling the many different cases and finding relevant phylogenetic invariants.

Next, we determine defining equations for the ideal of the strand symmetric model. The strand symmetric model is a phylogenetic model designed to reflect the symmetry inherent in the double-stranded structure of DNA. We show that the set of known phylogenetic invariants for the general strand symmetric model of the 3-leaf claw tree entirely defines the ideal. This knowledge allows one to determine the vanishing ideal of the general strand symmetric model on any binary tree. Our proof of the main result is computational. We use the fact that the Zariski closure of the strand symmetric model is the secant variety of a toric variety to compute its dimension. We then show that the known equations generate a prime ideal of the correct dimension using elimination theory.

Finally, we study 2-class mixtures of the binary Jukes-Cantor model. The ideal of phylogenetic invariants of a phylogenetic mixture model on trees with the same topology is a secant ideal. It has been shown that the Hilbert series of the ideal of the binary Jukes-Cantor model on an n-leaf tree is independent of the tree topology. We show that for trees with six or fewer leaves the same result holds for secants of these ideals and

conjecture that this is true for all n. We also resolve a conjecture about a class of binomial initial ideals of $I_{2,n}$, the ideal of the Grassmannian, $\mathrm{Gr}(2,\mathbb{C}^n)$, which are associated to phylogenetic trees. For a weight vector ω in the tropical Grassmannian, $in_\omega(I_{2,n}) = J_{\mathcal{T}}$ is the ideal associated to the tree $\mathcal{T}$. The ideal generated by the $2r \times 2r$ subpfaffians of a generic $n \times n$ skew-symmetric matrix is precisely $I_{2,n}^{\{r-1\}}$, the $(r-1)$-secant of $I_{2,n}$. We prove necessary and sufficient conditions on the topology of $\mathcal{T}$ in order for $in_\omega(I_{2,n})^{\{2\}} = J_{\mathcal{T}}^{\{2\}}$. We also give a new class of prime initial ideals of the Pfaffian ideals.

Algebraic Geometry of Phylogenetic Models

by
Colby Edward Long

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

2016

APPROVED BY:

Eric Stone	Agnes Szanto
Cynthia Vinzant	Seth Sullivant Chair of Advisory Committee

DEDICATION

For Mom, Dad, and Keeri.

BIOGRAPHY

Colby spent most of his early life in Maryland and graduated from Kent County High School in 2002. After brief stints in the Navy and on the Appalachian Trail, he attended St. Mary's College of Maryland where he divided his time between mathematics, sailing, and running. After graduating in 2008, he taught English in Seoul, South Korea before starting graduate studies in mathematics at North Carolina State University.

In the Fall of 2016, he will start a postdoctoral research fellowship at the Mathematical Biosciences Institute at Ohio State University. In his spare time, he plays and watches entirely too much soccer and hangs out with his future wife Keeri and their two cats, Charleston and Furmat (named after Pierre de Fermat, but spelled with a "u," because she's a cat, and furry).

ACKNOWLEDGEMENTS

I would like to thank the many people that made it possible for me to complete my degree. First among them, my advisor Seth Sullivant, who guided me through all the stages of starting a mathematical career and never failed to be available over the last several years. He introduced me to a way of doing mathematics that I thoroughly enjoy and enabled me to accomplish much more than I thought possible.

I would also like to thank my committee, Ruth Davidson, Elizabeth Gross, and everyone who has been a part of my research group for their advice and support. I owe a debt to Mark, Jason, Daniel, and the many officemates that have made our windowless office a place for enriching mathematical discussion and idle chatter. I would especially like to thank Mandi, without whom this would literally not have been possible because she was my ride to campus.

Finally, of course, I would like to thank my parents for their constant encouragement, and Keeri, who pushed me to work, enforced breaks, listened to talks, and worked every bit as hard for this as I did.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

Introduction

Phylogenetics is the field which seeks to untangle the evolutionary relationships between species. Researchers construct phylogenetic models designed to model the process of molecular sequence evolution and compare the results to existing biological data. The goal is to infer species trees which have applications in evolutionary biology, epidemiology, and species conservation [SS03]. While essential to our understanding of evolution, reconstructing trees is far from straigthtforward. Biological phenomena such as hybridization, horizontal gene transfer, and incomplete lineage sorting all complicate the evolutionary picture. Moreover, a weak phylogenetic signal or a paucity of data can make the task even more difficult. Thus, there are a number of theoretical issues that must be resolved in order to interpret the data and ensure inference is consistent. This is done with a variety of mathematical tools from several fields, including discrete mathematics, probability and statistics, and algebraic geometry.

Our primary tool for studying several different phylogenetic models of DNA evolution will be algebraic geometry. In this chapter, we begin with a description of the basic components and construction of phylogenetic models. We then reexamine them from the perspective of algebraic geometry and establish the terminology, notation, and tools that we require for the later chapters.

1.1 Phylogenetic Models

A central problem in phylogenetics is to describe the evolutionary history of n species from their aligned DNA sequences. In this case, statistical models are constructed on a

tree T to mimic the process of molecular evolution at a single locus.

Definition 1.1.1. A *tree* $T = (V, E)$ is a connected acyclic graph with vertex set V and edge set E.

Definition 1.1.2. A vertex of T of degree at most one is called a *leaf*. The vertices that are not leaves are called *interior vertices*.

In our models, the leaves of T correspond to the species under consideration. Thus, most often we will be dealing with *phylogenetic X-trees*. A phylogenetic X-tree is an ordered pair $\mathcal{T} = (T, \phi)$ where T is a tree and ϕ is a bijective function from a label set X to the set of leaf vertices of T. Typically, our label set will be $X = [n] := \{1, \ldots, n\}$. Additionally, we assume that the tree parameter is *binary*.

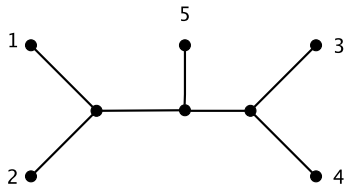
Definition 1.1.3. A tree is *binary* if every interior vertex has degree three.

The binary assumption on the tree parameter reflects our assumptions about the process of evolution. We think of each interior vertex of the tree as corresponding to an ancestral species for which no data is available. At each interior vertex, a speciation event occurs wherein the ancestral species diverges into two different species.

Removing an edge from a binary phylogenetic X-tree $\mathcal{T}$ creates two connected components. The labels in each component form a partition $\{B, B'\}$ of the leaf label set which we refer to as an *X-split* or more simply a *split* when the set X is understood. We use the notation $B|B'$ to denote splits and $\Sigma(\mathcal{T})$ to denote the set of all splits of $\mathcal{T}$. A split in which either $|B|$ or $|B'| = 1$ is called a *trivial split* since it is a split of every binary phylogenetic X-tree.

Example 1.1.4. For the 5-leaf binary phylogenetic X-tree $\mathcal{T}$ pictured below, $X = [5]$ and

$$\Sigma(\mathcal{T}) = \{1|2345, 2|1345, 3|1245, 4|1235, 5|1234, 12|345, 34|125\}.$$



The set of trivial splits is $\{1|2345, 2|1345, 3|1245, 4|1235, 5|1234\}$.

A phylogenetic X-tree is uniquely determined by its set of splits. This is the result of the well-known Splits-Equivalence Theorem [SS03, p. 44]. This result also gives necessary and sufficient conditions on an arbitrary set of splits in order for there to exist an X-tree displaying that set of splits. For this reason, we are justified in distinguishing trees by their sets of splits and indexing edges by the splits they induce.

From the evolutionary standpoint, it makes sense to distinguish one interior vertex as the common ancestor of all of our species and to think of the process of evolution as directional away from this vertex.

Definition 1.1.5. A *rooted binary phylogenetic X-tree* is a phylogenetic X-tree where each interior vertex has degree three except one interior vertex which has degree two. We call the distinguished vertex of degree two the *root* and denote it ρ.

Thus, the underlying *tree parameter* of the phylogenetic model is a rooted binary phylogenetic X-tree $\mathcal{T}$ with $X = [n]$ (as pictured in Example 1.1.7). To each vertex v we associate a random variable X_v with state space $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ corresponding to the four DNA bases. For the unique edge e between nodes v and w, the 4×4 *transition matrix* A^e is the matrix with $A^e_{ij} = P(X_v = i | X_w = j)$. The entries of the transition matrices are called the *numerical parameters*.

Definition 1.1.6. The *n-dimensional probability simplex* is the set

$$\Delta^n = \left\{ (x_1, \ldots, x_{n+1}) \in \mathbb{R}^{n+1} : x_i \geq 0 \text{ and } \sum_{i=1}^{n+1} x_i = 1 \right\}.$$

The *root distribution* of a phylogenetic model is the vector $\pi = (\pi_A, \pi_C, \pi_G, \pi_T) \in \Delta^3$ where $P(X_\rho = i) = \pi_i$. For any assignment of states to the vertices of $\mathcal{T}$, we can use the transition matrices and root distribution to calculate the probability of observing that particular state. The following explanation for how this is done is adapted from [SS03, Ch. 8] where the formal framework is fully developed. It is intuitive if we think of the process of evolution at a particular site unfolding along a tree. It may be helpful to refer to Example 1.1.7 which demonstrates the calculation for a simplified model of DNA evolution.

We view the edges of $\mathcal{T}$ as directed away from the root and define a total order on the vertex set V. This order must satisfy that for each edge vw of the tree, $w < v$ if w is the vertex closer to the root. Let $\chi : V \to \sigma$ be an assignment of states to the vertices. We then insist that the random variables satisfy the *Markov property*, that is,

$$P\left(X_v = \chi(v) | \bigcap_{u<v} X_u = \chi(u)\right) = P(X_v = \chi(v) | X_w = \chi(w)).$$

This implies that the probability of observing a mutation between species depends only on the state of the immediate ancestor. Combining the product rule of conditional probability with the Markov property gives us

$$P\left(\bigcap_{u\leq v} X_u = \chi(u)\right) = P\left(X_v = \chi(v) | \bigcap_{u<v} X_u = \chi(u)\right) = P(X_v = \chi(v) | X_w = \chi(w)).$$

Note that the last term of this expression is an entry of the transition matrix associated to the edge vw. Iteratively applying the product rule and using the entries of the transition matrices for the conditional probabilities gives us

$$P\left(\bigcap_{u\in V} X_u = \chi(u)\right) = \pi_{\chi(\rho)} \prod_{e=uv\in E(\mathcal{T})} A^e_{\chi(v),\chi(w)} \tag{1.1}$$

Our primary interest is in determining the probability of observing a particular n-tuple of states at the leaves. Thus, for a fixed n-tuple we marginalize over all possible states of the interior vertices to obtain the probability of observing that n-tuple at the leaves. Once we have fixed $\mathcal{T}$, each choice of numerical parameters gives us a probability distribution on the 4^n distinct n-tuples of DNA bases in Δ^{4^n-1}.

While the 4-state model of DNA evolution serves as our motivating example we may also wish to consider other phylogenetic models with different state spaces. For example, the binary Jukes-Cantor model or Cavender-Farris-Neyman (CFN) model is a two-state phylogenetic model useful for modeling the evolution of purine (`A,G`) and pyrimidine (`C,T`) sequences. We may also look at sequences at the level of *codons* which are sequences of three DNA bases that code for the 20 different amino acids. While there are 64 possible three nucleotide sequences, only 61 of these that actually code for amino acids. Thus, the codon model is a 61-state model of codon evolution. Or, we may group together codons that code for the same amino acid, and model the evolution of amino acid sequences via a 20-state model.

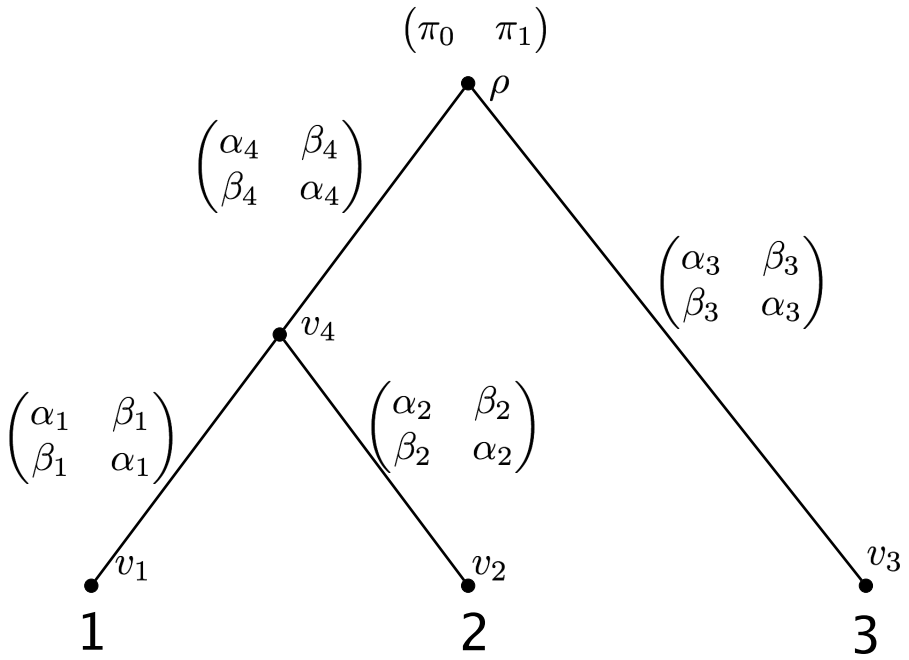
In general, if we let σ be the state space of the k-state random variables associated

to each node of the tree then there are k^n possible states at the leaves and each choice of numerical parameters yields a probability distribution on the elements of σ^n. For a fixed tree, this gives a map from the space of numerical parameters into the probability simplex, $\psi_{\mathcal{T}} : \Theta_{\mathcal{T}} \to \Delta^{k^n-1} \subseteq \mathbb{R}^{k^n}$.

Example 1.1.7. Consider the CFN model ($k = 2$) for three species. We let $\sigma = \{0, 1\}$ be the state space where 0 and 1 represent purines (`A,G`) and pyrimidines (`C,T`) respectively.

The tree parameter $\mathcal{T}$ is a 3-leaf binary phylogenetic X-tree with label set $X = [3]$. The leaves correspond to three distinct species. There are two interior vertices of $\mathcal{T}$, the root, ρ, and v_4. The edges are labeled by their 2×2 transition matrices with rows and columns indexed by $\{0, 1\}$.

The set of numerical parameters for this model is $\{\pi_0, \pi_1, \alpha_1, \beta_1, \alpha_2, \beta_2, \alpha_3, \beta_3, \alpha_4, \beta_4\}$. Although there are ten parameters, the parameter space is only 4-dimensional as we must have $\alpha_i + \beta_i = 1$ for $1 \le i \le 4$ and because we insist that for the CFN model the root distribution is uniform, that is, $\pi = (\frac{1}{2}, \frac{1}{2})$.



For what follows let $p_{i_1 i_2 i_3} := P(X_{v_1} = i_1, X_{v_2} = i_2, X_{v_3} = i_3)$ and $p_{i_1 i_2 i_3 i_4 i_5} := P(X_{v_1} = i_1, X_{v_2} = i_2, X_{v_3} = i_3, X_{v_4} = i_4, X_\rho = i_5)$. As an example, we will compute p_{011}, the probability that at a particular DNA locus, there is a purine in the sequence of the first species

and a pyrimidine in the sequences of the other two species. We marginalize over all of the unobserved states of the internal vertices, so

$$p_{011} = p_{01100} + p_{01110} + p_{01101} + p_{01111}.$$

Based on the preceding discussion and (1.1) in particular, we see that this is a polynomial with four terms,

$$p_{011} = \pi_0\alpha_1\beta_2\beta_3\alpha_4 + \pi_0\beta_1\alpha_2\beta_3\beta_4 + \pi_1\alpha_1\beta_2\alpha_3\beta_4 + \pi_1\beta_1\alpha_2\alpha_3\alpha_4.$$

Each point in the image of $\psi_{\mathcal{T}}$ is a probability distribution in $\mathbb{R}^8$. The coordinates of $\mathbb{R}^8$ are indexed by the eight possible states at the leaves so that a generic point is of the form

$$(p_{000}, p_{001}, p_{010}, p_{100}, p_{011}, p_{101}, p_{110}, p_{111}) \in \mathbb{R}^8.$$

Notice that each of these coordinates is parameterized by a degree 5 polynomial in the numerical parameters as was shown for p_{011}.

For model-based reconstruction, we assume that the DNA sites are independent and that the distribution of n-tuples of DNA bases is the same at each site. The goal is then to find a choice of parameters that yields a distribution close to that observed in the aligned DNA sequences. Suppose, for example, that we observe the n-tuple `ACTTG` at .01% of the sites of the aligned DNA sequences of five species. Then we would like to find a model and choice of parameters where $p_{ACTTG} \approx .0001$, and of course, we would also like to match as closely as possible the site pattern frequencies of the other n-tuples. If we are able to find such a model then it reasonable to assume that it is a good approximation of the process that produced our sequences. In particular, we infer that the tree parameter of the model is the phylogeny of the species.

One approach to actually finding a model that matches our distribution is maximum likelihood estimation [Fel81, SS03]. This is the process of finding the parameter pair $(\mathcal{T}, \theta)$, where $\mathcal{T}$ is a binary phylogenetic X-tree and θ is the associated set of numerical parameters, that maximizes the probability of producing the observed data. Even for a fixed tree, computing the maximum likelihood estimation of the numerical parameters is a difficult problem. Moreover, because there are $(2n-3)!! = 1 \cdot 3 \cdot 5 \cdot \ldots \cdot (2n-5) \cdot (2n-3)$ n-leaf rooted binary phylogenetic X-trees, computing this estimation over the set of all such trees becomes infeasible as n increases. In practice, heuristic methods work well

enough that maximum likelihood estimation is widely used in phylogenetics. One can also take a Bayesian approach to phylogenetic reconstruction, though as noted in [SS03], this is less frequently used because of the difficulty in specifying a prior distribution on the pairs $(\mathcal{T}, \theta)$.

This model-based approach raises a serious concern about the *identifiability* of the parameters of the model. A model parameter is identifiable if a distribution coming from the model uniquely determines the parameter that produced it. Of particular importance for phylogenetic inference is the identifiability of the tree parameter. Indeed, there is no way to infer the true phylogeny if multiple models with different tree parameters explain our data equally well. One important application of algebraic geometry to phylogenetics has been proving that different phylogenetic models are identifiable (See e.g., [ARS12, MMS08, MS07, RS12].)

We obtain different phylogenetic models by placing various restrictions on the entries of the transition matrices. For example, the Jukes-Cantor (JC) model of DNA evolution is a 4-state model where we assume $A^e_{ij} = \alpha$ if $i = j$ and $A^e_{ij} = \beta$ otherwise. In other words, we assume that along each edge there is some probability of mutation and if a mutation occurs it does so with equal likelihood to any of the other DNA bases. Thus, the transition matrices for the Jukes-Cantor model have the structure depicted in Figure 1.1c.

At the other extreme for DNA models, the 4-state general Markov model imposes only the stochastic condition on the rows of the transition matrices. That is, each row of A^e belongs to Δ^3. While this model allows greater flexibility, it is more difficult to determine its properties, such as whether or not it is identifiable. Existing somewhere in between these two extremes are the Kimura 2-parameter (K2P) and Kimura 3-parameter models (K3P) which are designed to better reflect the realities of DNA substitution. The transition matrices for several of these models are shown in Figure 1.1.

The set of all possible distributions we obtain by varying the numerical parameters is what we call the *model*. We denote the model $\mathcal{M}_\mathcal{T}$ and observe that $\mathcal{M}_\mathcal{T} = \mathrm{Im}(\psi_\mathcal{T})$. Though our notation involves only $\mathcal{T}$, as per the discussion above, it is possible to have different models on the same tree. For example, the set of probability distributions associated to the CFN model on $\mathcal{T}$ and the K2P model on $\mathcal{T}$ will not be the same set and are not even contained in the same space. However, in all of our discussions the particular restrictions on the transition matrices will be clear from context and no further notation

$$\begin{pmatrix} \alpha & \beta & \gamma & \gamma \\ \beta & \alpha & \gamma & \gamma \\ \gamma & \gamma & \alpha & \beta \\ \gamma & \gamma & \beta & \alpha \end{pmatrix}$$

(a) K2P

$$\begin{pmatrix} \alpha & \beta \\ \beta & \alpha \end{pmatrix}$$

(b) CFN

$$\begin{pmatrix} \alpha & \beta & \beta & \beta \\ \beta & \alpha & \beta & \beta \\ \beta & \beta & \alpha & \beta \\ \beta & \beta & \beta & \alpha \end{pmatrix}$$

(c) JC

$$\begin{pmatrix} \alpha & \beta & \gamma & \delta \\ \beta & \alpha & \delta & \gamma \\ \gamma & \delta & \alpha & \beta \\ \delta & \gamma & \beta & \alpha \end{pmatrix}$$

(d) K3P

Figure 1.1: Transition matrices for phylogenetic models.

is needed.

For all of the models discussed in this thesis the location of the root vertex is unidentifiable [AR07]. This means that if $\mathcal{T}$ and $\mathcal{T}'$ differ only by the location of their root vertex then $\mathcal{M}_{\mathcal{T}} = \mathcal{M}_{\mathcal{T}'}$. For this reason, despite the intuitive appeal of rooted trees for modeling evolution, all of the tree parameters that we consider will be unrooted. Given an assignment of states to the vertices, we compute the probability of observing that state by regarding any vertex as the root and proceeding as in Example 1.1.7. Proposition 1.1.8 implies that for a k-state model each coordinate function of $\psi_{\mathcal{T}}$ is a degree $2n-3$ polynomial in the numerical parameters with k^{n-2} terms.

Proposition 1.1.8. *[SS03] Let $\mathcal{T}$ be a binary phylogenetic X-tree (unrooted) and let $n = |X|$. Then, for all $n \geq 2$, $\mathcal{T}$ has $2n-3$ edges and $n-2$ interior vertices.*

One other type of phylogenetic model that we will consider in Chapter 2 is a *mixture model*. Mixture models account for the fact that different portions of DNA may evolve differently due to issues such as incomplete lineage sorting, horizontal gene transfer, and different rates of mutation across sites [DR09]. The idea is to weight the distributions from multiple models according to the proportion of data that evolved according to each to produce a single probability distribution. Therefore, if s_i is a choice of numerical parameters for the k-state model associated to the tree $\mathcal{T}_i$, the r-th mixture model is the closure of the image of the map

$$\psi_{\mathcal{T}_1,\ldots,\mathcal{T}_r} : \Theta_{\mathcal{T}_1} \times \ldots \times \Theta_{\mathcal{T}_r} \times \Delta^{r-1} \to \Delta^{k^n-1},$$

where $(s_1, \ldots, s_r, \pi) \mapsto \pi_1 \psi_{\mathcal{T}_1}(s_1) + \ldots + \pi_r \psi_{\mathcal{T}_r}(s_r)$. We use the notation $\mathcal{M}_{\mathcal{T}_1} * \ldots * \mathcal{M}_{\mathcal{T}_r}$ for the set of distributions of the mixture model. Just as with the general Markov model, the additional parameters in phylogenetic mixture models make them both more flexible and more difficult to analyze.

1.2 The Algebraic Perspective

We may study phylogenetic models by finding algebraic relationships between coordinates satisfied by all points in $\mathcal{M}_{\mathcal{T}}$. To do so, we associate algebraic objects to the models and analyze them using tools from algebraic geometry. The rest of this thesis relies heavily on this algebraic perspective. In this section we demonstrate the connections between algebraic geometry and phylogenetic models. In Section 1.3, we discuss the theory that underlies many of the tools of algebraic geometry that we deploy in the subsequent chapters. Most of the standard notation, definitions, and results are adapted from [Eis04, Has07].

Let $\mathbb{K}$ be an algebraically closed field of characteristic zero and let $\mathbb{K}[\mathbf{x}] = \mathbb{K}[x_1, \ldots, x_n]$ denote the algebra of polynomials in the variables $x_1, \ldots, x_n$ over $\mathbb{K}$.

Definition 1.2.1. Given $S \subseteq \mathbb{K}^n$, the *vanishing ideal of* S is

$$I(S) := \{f \in \mathbb{K}[x_1, \ldots, x_n] : f(s) = 0 \text{ for each } s \in S\}.$$

For a phylogenetic model, $\mathcal{I}_{\mathcal{T}} := I(\mathcal{M}_{\mathcal{T}}) \subseteq \mathbb{C}[p_i : i \in \sigma^n]$ is called the *ideal of phylogenetic invariants* and the elements of $\mathcal{I}_{\mathcal{T}}$ are called *phylogenetic invariants*. Recall that the number of leaves of the tree parameter is n and that σ^n is the set of n-tuples of the elements of the state space σ. The variable p_i is the probability of observing the n-tuple i at the leaves. The study of phylogenetic invariants was originally proposed as a method for reconstructing phylogenetic trees [CF87, Lak87], but as mentioned above, they have also been useful for proving identifiability results.

Theorem 1.2.2 (Hilbert Basis Theorem)**.** *Every polynomial ideal in* $\mathbb{K}[\mathbf{x}]$ *is finitely generated.*

In light of the Hilbert Basis Theorem, often we will be interested in finding a finite set of polynomials that generate $\mathcal{I}_{\mathcal{T}}$.

Example 1.2.3. For the phylogenetic model from Example 1.1.7,

$$\mathcal{I}_{\mathcal{T}} \subseteq \mathbb{C}[p_{000}, p_{001}, p_{010}, p_{100}, p_{011}, p_{101}, p_{110}, p_{111}].$$

Since all of the transition matrices are symmetric, it is not surprising that, for example, $p_{001} = p_{110}$. In this case, the ideal of phylogenetic invariants is entirely generated by such linear relations along with the trivial invariant forced by the stochastic condition on our parameters.

$$\begin{aligned}\mathcal{I}_{\mathcal{T}} = \langle & p_{000} - p_{111}, p_{100} - p_{011}, p_{010} - p_{101}, p_{001} - p_{110}, \\ & p_{000} + p_{001} + p_{010} + p_{100} + p_{011} + p_{101} + p_{110} + p_{111} - 1 \rangle.\end{aligned}$$

Example 1.2.4. Let $\mathcal{T}$ be the 4-leaf tree with nontrivial split 12|34. Modulo the linear invariants,

$$\mathcal{I}_{\mathcal{T}} \subseteq \mathbb{C}[p_{0000}, p_{0001}, p_{0010}, p_{0011}, p_{0100}, p_{0101}, p_{0110}, p_{0111}]$$

Here, we will ignore the stochastic assumption on the parameters so that the trivial invariant is not contained in the ideal of phylogenetic invariants. This makes the ideal of phylogenetic invariants a *homogeneous ideal*. The advantages of doing this will become evident in Chapter 2 and are related to the discussion in Section 1.3 surrounding homogeneous ideals and *projective varieties*.

In this case, the ideal of phylogenetic invariants is generated by two quadratic equations,

$$\begin{aligned}\mathcal{I}_{\mathcal{T}} = \langle & p_{0010}p_{0100} - p_{0011}p_{0101} - p_{0000}p_{0110} + p_{0001}p_{0111}, \\ & p_{0001}p_{0100} - p_{0000}p_{0101} - p_{0011}p_{0110} + p_{0010}p_{0111} \rangle.\end{aligned}$$

Definition 1.2.5. An *affine variety* is the locus where a collection of polynomials is satisfied. Given $F \subseteq \mathbb{K}[x_1, \ldots, x_n]$,

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

The set F is said to *define* the variety. The *Zariski topology* on $\mathbb{K}^n$ is the topology in

which a set is closed if and only if it is an affine variety. The closure of $S \subseteq \mathbb{K}^n$ in the Zariski topology is $\overline{S} := V(I(S))$. For a phylogenetic model, the *variety of the model* is $V_{\mathcal{T}} := \overline{\mathcal{M}_{\mathcal{T}}}$.

The algebraic perspective also applies to mixture models. In the language of algebraic geometry, $\overline{\mathcal{M}_{\mathcal{T}_1} * \ldots * \mathcal{M}_{\mathcal{T}_r}} = V_{\mathcal{T}_1} * \ldots * V_{\mathcal{T}_r}$, the *join* of the varieties $V_{\mathcal{T}_1}, \ldots, V_{\mathcal{T}_r}$.

Definition 1.2.6. The *join* of the algebraic varieties $V_1, \ldots, V_r$ is the set

$$V_1 * \ldots * V_r = \overline{\{\pi_1 v_1 + \ldots + \pi_r v_r : v_i \in V_i \text{ and } \pi_1 + \ldots + \pi_r = 1\}}.$$

Put another way, the join is the closure of the union of all linear spaces spanned by one point from each variety in the join. In the case where $V_1 = \ldots = V_r$ this is called the *r-secant variety* and we use the notation $V^{\{r\}}$ for the r-secant variety of V. Letting $I_j := I(V_j) \subseteq \mathbb{K}[\mathbf{x}]$, we define $I_1 * \ldots * I_r := I(V_1 * \ldots * V_r) \subseteq \mathbb{K}[\mathbf{x}]$ to be the *join* (or *r-secant*) ideal. As for varieties, $I^{\{r\}}$ is the r-secant ideal of I.

In order to compute the join ideal we introduce $r(n+1)$ new unknowns grouped into r vectors of the form $\mathbf{y}_j := (y_{j1}, \ldots, y_{jn})$ and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_r)$. We then construct the ring of $rn + r + n$ variables $\mathbb{K}[\mathbf{x}, \mathbf{y}, \boldsymbol{\pi}]$. We define the ideal $I_j(\mathbf{y}_j)$ to be the image of the ideal I_j under the map $x_i \mapsto y_{ji}$.

Proposition 1.2.7. *[SS06] The join ideal $I_1 * \ldots * I_r$ is equal to*

$$(I_1(\boldsymbol{y}_1) + \ldots + I_r(\boldsymbol{y}_r) + \langle \pi_1 y_{1i} + \ldots + \pi_r y_{ri} - x_i : 1 \leq i \leq n \rangle + \langle \pi_1 + \ldots + \pi_r - 1 \rangle) \cap \mathbb{K}[\boldsymbol{x}].$$

This proposition gives us a way to express the ideal of phylogenetic invariants for a mixture model $\mathcal{I}_{\mathcal{T}_1} * \ldots * \mathcal{I}_{\mathcal{T}_r}$ in terms of the ideals of the constituent models. However, to actually compute a generating set for $\mathcal{I}_{\mathcal{T}_1} * \ldots * \mathcal{I}_{\mathcal{T}_r}$ requires the theory of Gröbner bases and elimination.

1.3 Gröbner Bases and Applications

In this section, we discuss Gröbner bases which form the theoretical foundation for much of the computational algebraic geometry presented in this thesis.

Definition 1.3.1. A *monomial* of $\mathbb{K}[\mathbf{x}]$ is an element of the form $\mathbf{x}^\alpha = x_1^{\alpha_1} \ldots x_n^{\alpha_n}$ for $\alpha \in \mathbb{N}^n$.

Definition 1.3.2. A *monomial order* is a total order $\prec$ on the set of monomials of $\mathbb{K}[\mathbf{x}]$ that satisfies

(i) Multiplicative Property: if $\mathbf{x}^\alpha \prec \mathbf{x}^\beta$ then $\mathbf{x}^\gamma \mathbf{x}^\alpha \prec \mathbf{x}^\gamma \mathbf{x}^\beta$.

(ii) Well Ordering: An arbitrary set of monomials $\{\mathbf{x}^\alpha\}_{\alpha \in \mathcal{A}}$ has a least element.

Example 1.3.3. The *standard lexicographic order* defined by $\mathbf{x}^\alpha \prec \mathbf{x}^\beta$ if and only if the leftmost nonzero entry of $\alpha - \beta$ is less than zero is a monomial order. Let $\prec$ be the standard lexicographic order in the ring $\mathbb{C}[x_1, x_2, x_3]$. Then,

$$x_1^3 x_2^4 x_3^5 \prec x_1^3 x_2^6 x_3$$

since the leftmost nonzero entry of $(3,4,5) - (3,6,1) = (0,-2,4)$ is -2.

Definition 1.3.4. Fix a monomial order $\prec$ on $\mathbb{K}[\mathbf{x}]$ and consider a polynomial

$$f = \sum_{\alpha \in \mathcal{A}} c_\alpha \mathbf{x}^\alpha \in \mathbb{K}[\mathbf{x}],$$

where each $c_\alpha \neq 0$. Let $\mathbf{x}^\alpha$ be the largest monomial of f with respect to $\prec$. Then the *initial term* of f is $in_\prec(f) := c_\alpha \mathbf{x}^\alpha$.

Definition 1.3.5. For an ideal $I \subseteq \mathbb{K}[\mathbf{x}]$, the *initial ideal of I with respect to* $\prec$ is the ideal

$$in_\prec(I) := \langle in_\prec(f) : f \in I \rangle.$$

Definition 1.3.6. Let $I \subseteq \mathbb{K}[\mathbf{x}]$ be an ideal. A finite set of polynomials $\{f_1, \ldots, f_k\} \subseteq I$ forms a *Gröbner basis* for I if $in_\prec(I) = \langle in_\prec(f_1), \ldots, in_\prec(f_k) \rangle$.

A priori, it is not clear that every ideal has a Gröbner basis. In fact, not only is this the case, but *Buchberger's algorithm* is an algorithm that takes as input any generating set for an ideal and returns a Gröbner basis for that ideal with respect to a given term order. A short proof also shows that a Gröbner basis for I with respect to any term order is also a generating set [Has07].

Example 1.3.7. Let $\prec$ be the standard lexicographic order on the ring $\mathbb{C}[x, y, z]$. Let $f_1 = x^3 + y^2 z$, $f_2 = xy - z^2$, and

$$I = \langle f_1, f_2 \rangle \subseteq \mathbb{C}[x, y, z]$$

Then $in_\prec(f_1) = x^3$ and $in_\prec(f_2) = xy$. Since I is an ideal, it also contains

$$f_3 = yf_1 - x^2 f_2 = x^2z^2 + y^3z.$$

Then $in_\prec(f_3) = x^2z^2 \notin \langle x^3, xy\rangle$. Therefore, the set $\{f_1, f_2\}$ is not a Gröbner basis for I with respect to the standard lexicographic term order. A Gröbner basis for I with respect to this term order is given by

$$\{y^5z + z^6, xz^4 + y^4z, xy - z^2, x^2z^2 + y^3z, x^3 + y^2z\}.$$

and

$$in_\prec(I) = \langle y^5z, xz^4, xy, x^2z^2, x^3\rangle.$$

In Chapter 4 we will also wish to consider initial ideals with respect to partial monomial orders.

Definition 1.3.8. Let $\omega \in \mathbb{N}^n$. Define the *ω-weight* of the monomial $\mathbf{x}^\alpha$ to be $\omega(\mathbf{x}^\alpha) := \omega \cdot \alpha$. Then $\prec_\omega$ is the partial monomial order given by $\mathbf{x}^\alpha \prec_\omega \mathbf{x}^\beta \iff \omega(\mathbf{x}^\alpha) < \omega(\mathbf{x}^\beta)$.

Definition 1.3.9. Let $f = \sum_{\alpha\in\mathcal{A}} c_\alpha \mathbf{x}^\alpha \in \mathbb{K}[\mathbf{x}]$, with each $c_\alpha \neq 0$. The *initial form of f with respect to $\prec_\omega$* is

$$in_\omega(f) := \sum_{\gamma\in\Gamma} c_\gamma \mathbf{x}^\gamma$$

where $\Gamma = \{\gamma \in \mathcal{A} : \omega(\mathbf{x}^\gamma) = \max_{\alpha\in\mathcal{A}} \omega(\mathbf{x}^\alpha)\}$.

Likewise, for an ideal $I \subseteq \mathbb{K}[\mathbf{x}]$ we define $in_\omega(I) := \langle in_\omega(f) : f \in I\rangle$.

Example 1.3.10. Let $\omega = (3, 1, 2)$, then for the polynomials f_1 and f_2 from Example 1.3.7, $in_\omega(f_1) = x^3$ since $\omega(x^3) = (3, 1, 2) \cdot (3, 0, 0) = 9$ and $\omega(y^2z) = (3, 1, 2) \cdot (0, 2, 1) = 4$. Both monomials of f_2 have weight four, so $in_\omega(f_2) = xy - z^2$.

Example 1.3.10 illustrates that $\prec_\omega$ is not necessarily a monomial order. As a consequence, $in_\omega(I)$ is not necessarily a monomial ideal. Still, in some sense, every monomial order can be realized in this way.

Proposition 1.3.11. *[Stu96, Proposition 1.11] For any monomial order $\prec$ and any ideal $I \subseteq \mathbb{K}[\mathbf{x}]$, there exists a nonnegative integer weight vector $\omega \in \mathbb{N}^n$ such that $in_\omega(I) = in_\prec(I)$.*

There is strong motivation for computing Gröbner bases because of their numerous applications. For example, given an ideal I, testing whether or not a polynomial belongs to I is straightforward if one knows a Gröbner basis with respect to any term order. Below, we discuss two other applications, computing elimination ideals and determining the Hilbert series of an ideal.

In Proposition 1.2.7 we saw that to compute a join ideal we could compute the intersection of an ideal with a ring generated by a subset of the variables. Given an ideal $I \subseteq \mathbb{K}[\mathbf{x}, \mathbf{y}]$ the process of finding generators for $I \cap \mathbb{K}[\mathbf{y}]$ is called *elimination*. The resulting ideal is called an *elimination ideal*. The variety $V(I \cap \mathbb{K}[\mathbf{y}])$ is the closure of the image of $V(I)$ under the map that projects away the $\mathbf{x}$ coordinates.

Definition 1.3.12. A monomial order $\prec$ on $\mathbb{K}[\mathbf{x}, \mathbf{y}]$ is an *elimination order* for $x_1, \ldots, x_n$ if for each polynomial $g \in \mathbb{K}[\mathbf{x}, \mathbf{y}]$, $in_{\prec}(g) \in \mathbb{K}[\mathbf{y}] \Rightarrow g \in \mathbb{K}[\mathbf{y}]$.

Theorem 1.3.13 (Elimination Theorem)**.** *Let $I \subseteq \mathbb{K}[\mathbf{x}, \mathbf{y}]$ be an ideal and $\prec$ an elimination order for $x_1, \ldots, x_n$. If $\mathcal{G}$ is a Gröbner basis for I then $\mathcal{G} \cap \mathbb{K}[\boldsymbol{y}]$ is a Gröbner basis for $I \cap \mathbb{K}[\mathbf{y}]$ and hence generates $I \cap \mathbb{K}[\mathbf{y}]$.*

As shown in Proposition 1.2.7, this gives us an algorithm for finding generating sets of join ideals. Another application is the general implicitization problem. Given an algebraic variety $V \subseteq \mathbb{K}^n$ and a morphism $\phi : V \to \mathbb{K}^m$, this is the problem of finding generators for $I(\phi(V))$.

Definition 1.3.14. A *morphism*, $\phi : \mathbb{K}^n \to \mathbb{K}^m$ is a map given by a polynomial rule

$$\mathbf{x} \mapsto (\phi_1(\mathbf{x}), \ldots, \phi_m(\mathbf{x}))$$

with each $\phi_i(\mathbf{x}) \in \mathbb{K}[\mathbf{x}]$.

Example 1.3.15. Consider the morphism $\phi : \mathbb{C}^2 \to \mathbb{C}^3$ given by (s^2, st, t^2). Labeling the coordinates of the image space by y_1, y_2, and y_3, we define

$$J = \langle y_1 - s^2, y_2 - st, y_3 - t^2 \rangle \subset \mathbb{C}[y_1, y_2, y_3, s, t].$$

Then $I(\mathrm{Im}(\phi)) = J \cap \mathbb{R}[y_1, y_2, y_3]$. The lexicographic term order with $y_1 < y_2 < y_3 < s < t$ is an elimination order for s, t. Using the computer algebra package Macaulay2, we find that

$$\mathcal{G} = \{y_1 y_3 - y_2^2, t^2 - y_3, s y_3 - t y_2, s y_2 - t y_1, st - y_2, s^2 - y_1\}$$

is a Gröbner basis for J with respect to this term order.

By Theorem 1.3.13,

$$I(\mathrm{Im}(\phi)) = \langle y_2^2 - y_1 y_3 \rangle.$$

Theoretically, elimination should allow us to compute $\mathcal{I}_\mathcal{T}$, the ideal of phylogenetic invariants for a model, from the morphism $\psi_\mathcal{T}$. Computing the requisite Gröbner basis turns out to be difficult in practice even with the aid of computers.

Another application of Gröbner bases is the computation of the Hilbert series of an ideal. Before we begin our discussion of Hilbert series, we present a few basic definitions from projective algebraic geometry.

Definition 1.3.16. *Projective n-space* $\mathbb{P}^n(\mathbb{K})$ is the set of all lines in $\mathbb{K}^{n+1}$ containing the origin.

Each such line can be defined as $\mathrm{span}(\{(a_0, \ldots, a_n)\})$ for some $(a_0, \ldots, a_n) \in \mathbb{K}^{n+1} - \{0\}$. This representation is of course not unique since $(a_0, \ldots, a_n)$ and $\lambda(a_0, \ldots, a_n)$ define the same line for any $\lambda \in \mathbb{K}^*$. Thus, there is an equivalence relation on $\mathbb{K}^{n+1} - \{0\}$ where two points are equivalent if they define the same line. We can identify $\mathbb{P}^n(\mathbb{K})$ with the set of equivalence classes in $\mathbb{K}^{n+1} - \{0\}$ modulo this relation. We denote the equivalence class of $(a_0, \ldots, a_n)$ by $[a_0 : \ldots : a_n]$ and simply use $\mathbb{P}^n$ for projective n-space when the field is understood.

Definition 1.3.17. An ideal $I \subseteq \mathbb{K}[\mathbf{x}]$ is *homogeneous* if it has a homogeneous system of generators.

Definition 1.3.18. Let $J \subset \mathbb{K}[x_0, \ldots, x_n]$ be a homogeneous ideal. The set

$$X(J) := \{[a_0 : \ldots : a_n] : F(a_0, \ldots, a_n) = 0 \text{ for each homogeneous } F \in J\}$$

is the *projective variety defined by* J.

A *projective variety* is a set $X \subseteq \mathbb{P}^n$ that is equal to $X(J)$ for some homogeneous ideal J. While this definition is different than that given in [Has07], Proposition 9.14 of the same reference ensures that these definitions are equivalent.

Example 1.3.19. Consider the ideal

$$J = \langle x - y, x^2 - yz \rangle \subseteq \mathbb{C}[x, y, z].$$

This is a homogeneous ideal with homogeneous generators of degree one and two. The projective variety it defines is $X(J) = \{[0:0:1],[1:1:1]\} \subseteq \mathbb{P}^2$.

In order to define the Hilbert series, we need to introduce the notion of a graded ring and a graded R-module. The definition of an R-module can be found for example in [Eis04, p.15].

Definition 1.3.20. A *graded ring* is a ring together with a direct sum decomposition

$$R = \bigoplus_{i=0}^{\infty} R_i$$

as abelian groups such that $R_i R_j \subseteq R_{i+j}$ for $i, j \geq 0$.

Definition 1.3.21. If $R = R_0 \oplus R_1 \oplus \ldots$ is a graded ring, then a *graded module* over R is a module M with a decomposition

$$M = \bigoplus_{i=0}^{\infty} M_i$$

as abelian groups, such that $R_i M_j \subseteq M_{i+j}$ for all i and j.

The situation we are most interested in is when R is the ring of polynomials $\mathbb{K}[\mathbf{x}]$ graded by degree. If $J \subseteq \mathbb{K}[\mathbf{x}]$ is a homogeneous ideal, then $\mathbb{K}[\mathbf{x}]/J$ is a graded $\mathbb{K}[\mathbf{x}]$-module where the grading is again by degree.

Definition 1.3.22. Let M be a finitely generated graded module over $\mathbb{K}[\mathbf{x}]$ graded by degree. Then the *Hilbert series of* M is

$$HS(M,t) := \sum_{d=0}^{\infty} \dim_{\mathbb{K}}(M_d) t^d.$$

As noted in [Eis04, Section 1.9], the Hilbert series of $\mathbb{K}[\mathbf{x}]/J$ encodes invariants of the projective variety $X(J)$ such as the dimension (which we define in Section 1.4.2) and degree.

Our earlier claim was that determining the Hilbert series is an application of Gröbner bases. To see how the two are related, let J be a homogeneous ideal and choose a monomial order $\prec$. By calculating a Gröbner basis, we can determine generators for $in_{\prec}(J)$. The monomials which do not lie in J are called the *standard monomials of J with respect to*

$\prec$. If we let $M = \mathbb{K}[\mathbf{x}]/J$ then the standard monomials of degree d form a basis for M_d as a $\mathbb{K}$-vector space [Stu96, Proposition 1.11].

Example 1.3.23. Let $J = \langle x^4+xy^3-y^4, x^3-xy^2+2y^3\rangle \subset \mathbb{C}[x,y]$ and let $\prec$ be the standard lexicographic term order.

J is a homogeneous ideal with Gröbner basis

$$\mathcal{G} = \{y^6, xy^4 + 3y^5, x^2y^2 - xy^3 - y^4, x^3 - xy^2 + 2y^3\}$$

and

$$in_\prec(J) = \langle y^6, xy^4, x^2y^2, x^3\rangle.$$

We can visualize the monomials of $in_\prec(J)$ via a *staircase diagram* in the plane. The lattice vector $(a,b) \in \mathbb{N}^2$ corresponds to the monomial $x^a y^b$. If a monomial appears in the initial ideal, then so must every monomial to the right and above it, since these can be obtained by repeatedly multiplying this monomial by x and y.

The aptness of the term "staircase diagram" is illustrated in Figure 1.2. The black lattice points correspond to generators of $in_\prec(J)$, grey to the other monomials in $in_\prec(J)$, and white to the standard monomials. The diagonal dotted lines connect monomials of the same degree. The standard monomials are those that lie underneath the staircase.

The degree d standard monomials in the diagram form a basis for $(\mathbb{C}[x,y]/J)_d$ as a $\mathbb{C}$-vector space. As an illustration, consider the degree four homogeneous polynomial $g = x^4+x^2y^2 \notin J$. Adding or subtracting elements of J gives us an equivalent representative of g in the ring $\mathbb{C}[x,y]/J$. Therefore, we can write

$$\begin{aligned} g &\sim (x^4 + x^2y^2) - x(x^3 - xy^2 + 2y^3) \\ &\sim 2x^2y^2 - 2xy^3 \\ &\sim (2x^2y^2 - 2xy^3) - 2(x^2y^2 - xy^3 - y^4) \\ &\sim 2y^4. \end{aligned}$$

And $2y^4 \in \operatorname{span}_{\mathbb{C}}(\{xy^3, y^4\})$. To compute the Hilbert series, we observe,

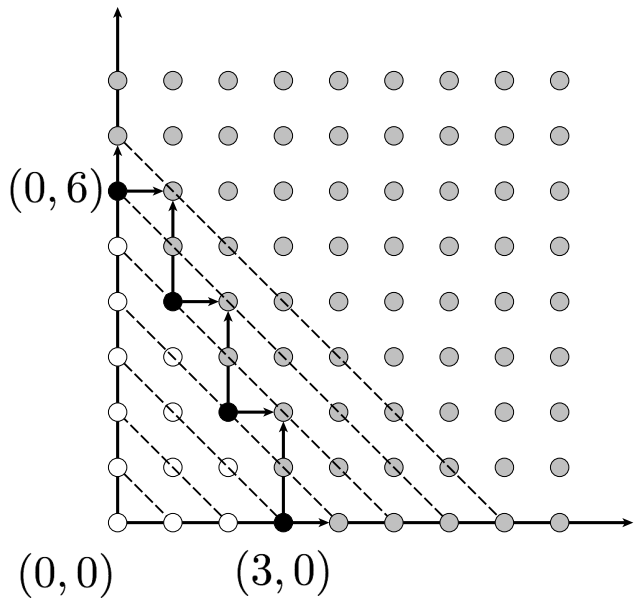


Figure 1.2: The staircase diagram for $in_{\prec}(J)$ from Example 1.3.23.

$$
\begin{aligned}
(\mathbb{C}[x,y]/J)_0 &= \mathrm{span}_{\mathbb{C}}(\{1\}),\\
(\mathbb{C}[x,y]/J)_1 &= \mathrm{span}_{\mathbb{C}}(\{x,y\}),\\
(\mathbb{C}[x,y]/J)_2 &= \mathrm{span}_{\mathbb{C}}(\{x^2,xy,y^2\})\\
(\mathbb{C}[x,y]/J)_3 &= \mathrm{span}_{\mathbb{C}}(\{x^2y,xy^2,y^3\}),\\
(\mathbb{C}[x,y]/J)_4 &= \mathrm{span}_{\mathbb{C}}(\{xy^3,y^4\}),\\
(\mathbb{C}[x,y]/J)_5 &= \mathrm{span}_{\mathbb{C}}(\{y^5\}),\\
(\mathbb{C}[x,y]/J)_i &= \mathrm{span}_{\mathbb{C}}(\{0\}) \text{ for } i \geq 6.
\end{aligned}
$$

Therefore, $HS(\mathbb{C}[x,y]/J,t) = 1 + 2t + 3t^2 + 3t^3 + 2t^4 + t^5$.

As a corollary to [Stu96, Proposition 1.11], a homogeneous ideal and all of its initial ideals have the same Hilbert series. This also implies that any two ideals that share a common initial ideal must have the same Hilbert series. We will make extensive use of these properties in Chapter 4 when we examine the initial ideals of some secant ideals

arising in phylogenetics.

1.4 Algebraic Tools for Phylogenetics

By viewing phylogenetic models as algebraic objects we can leverage the methods of algebra to explore their properties. For example, in the following chapters, we will see how some questions about these models can be reduced to questions about the *dimension* and *primality* of certain ideals. To address these questions we will tailor existing algebraic methods to our purposes. Three tools in particular that we will use repeatedly are the Fourier-Hadamard transformation, the prime-dimension approach, and the tropical secant dimension approach.

1.4.1 The Fourier-Hadamard Transformation

In this thesis, we will work primarily with *group-based phylogenetic models*. These are models in which the probability of a particular state change along an edge is dependent only on the difference between the group elements associated to the states at the endpoints. Formally,

Definition 1.4.1. A phylogenetic model is *group-based* if there exists a group G, a map $L : \sigma \to G$, and functions $f_e : G \to \mathbb{R}$ associated to the edges of $\mathcal{T}$, such that if v and w are the vertices of e, then $P(X_v = i | X_w = j) = f_e(L(i) - L(j))$.

The JC, K2P, and K3P models discussed above are all group-based models where $G = \mathbb{Z}_2 \times \mathbb{Z}_2$. The CFN model is also group-based for $G = \mathbb{Z}_2$.

Example 1.4.2. For the Jukes-Cantor model, the state space is $\sigma = \{A, C, G, T\}$. Let $G = \mathbb{Z}_2 \times \mathbb{Z}_2$ and set $L(A) = (0,0)$, $L(C) = (1,0)$, $L(G) = (0,1)$, and $L(T) = (1,1)$. Then choosing $\beta \in [0,1]$ and defining $f_e((1,0)) = f_e((0,1)) = f_e((1,1)) = \beta$ and $f_e((0,0)) = 1 - 3\beta$ realizes the JC model as a group-based model.

For group-based models the Fourier-Hadamard coordinate transformation is a linear change of coordinates that makes each coordinate function of the parameterization a monomial. We will present a practical outline demonstrating how to recover the monomials; a thorough explanation of the transform can be found in [ES93], [SS05], and [SESP93].

Let $p_{g_1,\ldots,g_n}$ be the probability of observing the state $(g_1,\ldots,g_n)$ at the leaves of $\mathcal{T}$ and let $q_{g_1,\ldots,g_n}$ be the image of this coordinate after the Fourier-Hadamard transformation. The image of the parameter $f_e(g)$ after transformation is $a_g^{B|B'}$ where $B|B'$ is the split induced by removing e. Thus, the set of new parameters is $\{a_g^{B|B'} : g \in G, B|B' \in \Sigma(\mathcal{T})\}$. The stochastic assumption under the Fourier transform forces $a_g^{B|B'} = 1$ when g is the identity element [SS05]. If every element of G is its own inverse, then after transformation,

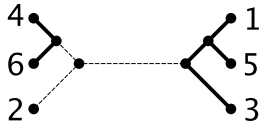
$$q_{g_1,\ldots,g_n} = \begin{cases} \prod_{B|B' \in \Sigma(T)} a^{B|B'}_{\sum_{i\in B} g_i} & \text{if } \sum_{i=1}^{n} g_i = 0 \\ 0 & \text{otherwise} \end{cases}$$

The assumption that every element of the group is its own inverse means that if the leaf elements sum to the identity, then for every partition $B|B'$, $\sum_{i\in B} g_i = \sum_{i\in B'} g_i$. Therefore, the monomial above does not depend on our labeling of the splits. All of the group-based models discussed in this thesis will be based on $\mathbb{Z}_2$ or $\mathbb{Z}_2 \times \mathbb{Z}_2$, both of which have this property.

Example 1.4.3. Consider the Jukes-Cantor model on the 6-leaf tree with nontrivial splits given by $\{15|2346\}$, $\{135|246\}$, and $\{1235|46\}$. We have already noted that since $L(A) = (0,0)$, $a_A^{B|B'} = 1$. In Example 1.4.2 we saw that $f_e((1,0)) = f_e((0,1)) = f_e((1,1))$ which forces $a_C^{B|B'} = a_G^{B|B'} = a_T^{B|B'} \in (0,1]$. Therefore, we will use only the label $a_C^{B|B'}$ for each of these parameters. Then for each coordinate and for each split, either $a_A^{B|B'}$ or $a_C^{B|B'}$ appears in the monomial parameterization of the coordinate. We encode the resulting monomials in tree diagrams as follows. Redraw the tree $\mathcal{T}$, but make each edge solid if $a_C^{B|B'}$ appears and dotted if $a_A^{B|B'}$ appears. The solid edges of the diagram form a subforest of $\mathcal{T}$ and the number of distinct nontrivial Fourier coordinates are in bijection with the subforests of $\mathcal{T}$ [SF95].

The parameterization of a particular coordinate as well as the subforest induced by this coordinate are shown. The superscript i means that the variable corresponds to the trivial split $i|([n] \setminus \{i\})$

$$q_{CAGGTG} = a_C^1 a_A^2 a_C^3 a_C^4 a_C^5 a_C^6 a_C^{15|2346} a_A^{135|246} a_A^{1235|46}$$



In the transformed coordinates, the ideals $\mathcal{I}_\mathcal{T}$ for the group-based models are seen to be *toric ideals*. That is, they are generated by differences of monomials. For the rest of

this thesis we will work in the transformed coordinates in order to take advantage of the rich combinatorial structure of toric ideals. For an excellent introduction to toric ideals see [Stu96, Chapter 4].

1.4.2 The Prime-Dimension Approach

One technique that we will apply in both Chapters 3 and 4 is the prime-dimension approach. In each of those chapters, we will have a morphism and we will want to determine a generating set for the vanishing ideal of the image of the morphism. As in Example 1.3.15, this can be done theoretically using elimination. However, in these instances, while we are able to find some elements of the elimination ideal, the Gröbner basis computation required to verify that these polynomials generate the entire elimination ideal is too intensive. Instead, we have a set of polynomials generating an ideal which we know to be contained in the ideal we actually wish to compute. The prime-dimension approach will allow us to prove that in fact, the two ideals are equal. The basic definitions of dimension theory presented in this section are adapted from [Eis04, Chapter 9].

Definition 1.4.4. An ideal $I \subseteq R$ is *prime* if $ab \in I$ implies that either $a \in I$ or $b \in I$.

Example 1.4.5. Let $\psi : \mathbb{K}^n \to \mathbb{K}^m$ be a morphism where $\psi(\mathbf{a}) = (f_1(\mathbf{a}), \ldots, f_m(\mathbf{a}))$. Then define $\phi : \mathbb{K}[y_1, \ldots, y_m] \to \mathbb{K}[x_1, \ldots, x_n]$ to be the $\mathbb{K}$-algebra homomorphism that sends $y_i \mapsto f_i(\mathbf{x})$. We claim that $\ker(\phi)$ is a prime ideal. Indeed, if $ab \in \ker(\phi)$, then $\phi(ab) = \phi(a)\phi(b) = 0$. Since $\mathbb{K}[x_1, \ldots, x_n]$ is an integral domain, this implies that either $\phi(a) = 0$ or $\phi(b) = 0$, which implies that either a or b is in $\ker(\phi)$. As a corollary, the ideals $\mathcal{I}_\mathcal{T}$ are always prime.

Definition 1.4.6. The *Krull dimension* of a commutative ring R is the length of the longest chain of prime ideals

$$\mathfrak{p}_0 \subset \ldots \subset \mathfrak{p}_d$$

contained in R. The length of the chain above is d and we write $\dim(R) = d$.

Definition 1.4.7. The dimension of an ideal $I \subseteq R$ is the Krull dimension of R/I.

Example 1.4.8. Let $I = \langle x^2 - y \rangle \subseteq \mathbb{C}[x, y, z]$. Then $\mathbb{C}[x, y, z]/I \cong \mathbb{C}[x, z]$. The sequence

$$\langle 0 \rangle \subseteq \langle x \rangle \subseteq \langle x, z \rangle$$

is a sequence of prime ideals in $\mathbb{C}[x, z]$. It can easily be seen that this chain is maximal, and so $\dim(I) = 2$.

If we let I be the ideal in Example 1.4.8, then $V(I) = \{(s, s^2, t) \in \mathbb{C}^3 : (s, t) \in \mathbb{C}^2\}$. Our intuition based on the number of parameters and the geometry of this set suggests that this variety also has dimension two. In fact, the dimension of the ideal $I(V)$ defines the dimension of V.

Definition 1.4.9. The dimension of an affine variety $V \subset \mathbb{K}^n$ is equal to the dimension of $I(V) \subset \mathbb{K}[\mathbf{x}]$. The dimension of the projective variety $X(J) \subseteq \mathbb{P}^n$ is one less than the dimension of $J \subseteq \mathbb{K}[x_0, x_1, \ldots, x_n]$.

Proposition 1.4.10. *Suppose that I, J are ideals in a commutative ring R with $J \subseteq I$. If J is prime and $\dim(I) = \dim(J)$ then $I = J$.*

Proof. We will use contradiction. Suppose I, J are as above with $\dim(I) = \dim(J) = n$ but that $J \subsetneq I$. Therefore, there exists a maximal chain of prime ideals

$$\mathfrak{p}_0 \subset \mathfrak{p}_1 \subset \ldots \subset \mathfrak{p}_n \subset R/I$$

Given an ideal $\mathfrak{p}$ in R/I, $\tilde{\mathfrak{p}} = \{x \in R : [x] \in R/I\}$ is an ideal in R containing I. Moreover, this construction gives a one-to-one correspondence between prime ideals in R/I and prime ideals in R containing I.

Therefore, since we assumed $J \subsetneq I$,

$$J \subset I \subseteq \tilde{\mathfrak{p}}_0 \subset \tilde{\mathfrak{p}}_1 \subset \ldots \subset \tilde{\mathfrak{p}}_n$$

is a chain of ideals in R and so

$$J \subset \tilde{\mathfrak{p}}_0 \subset \tilde{\mathfrak{p}}_1 \subset \ldots \subset \tilde{\mathfrak{p}}_n$$

is a chain of prime ideals in R of length $n + 1$. This implies the existence of a chain of prime ideals in R/J of length $n + 1$, contradicting that $\dim(J) = n$. □

Proposition 1.4.10 is the basis of the prime-dimension approach. Given two ideals $J \subseteq I$, the approach is to show that $I = J$ by proving that J is prime and that $\dim(I) = \dim(J)$. In the problem described in the introductory paragraph to this section, I is the

vanishing ideal of the image of the morphism and $J \subseteq I$ is the ideal generated by the polynomials known to be in the elimination ideal.

The questions still remain as to how to determine the dimension of an ideal and how to show that an ideal is prime. Section 1.4.3 will address the former question. As for primeness, we already know that the phylogenetic ideals $\mathcal{I}_\mathcal{T}$ are prime by the comments in Example 1.4.5. However, in Chapters 3 and 4 we will need to determine if certain ideals are prime only from a set of generators. There are algorithms for doing so implemented in many computer algebra systems. Unfortunately, as is a recurring theme in our discussion of computational algebraic geometry, these algorithms are too computationally intensive for the ideals in this thesis.

Instead, we use the following result from [GSS05] which in certain cases allows one to determine the primality of an ideal by determining the primality of an ideal in fewer variables.

Lemma 1.4.11. *[GSS05, Proposition 23] Let $\mathbb{K}$ be a field and $J \subseteq \mathbb{K}[\mathbf{x}]$ be an ideal containing a polynomial $f = gx_1 + h$ with g, h not involving x_1 and g not a zero divisor modulo J. Let $J_1 = J \cap \mathbb{K}[x_2, \ldots, x_n]$ be the elimination ideal. Then J is prime if and only if J_1 is prime.*

Proposition 23 of [GSS05] is a known result that was stated without proof. We include a proof here for completeness.

Proof. ($\Rightarrow$) It is true in general that the elimination ideal of a prime ideal is prime. Suppose J is prime and let $a, b \in \mathbb{K}[\mathbf{x}] \setminus J_1$ such that $ab \in J_1$. Since $J_1 \subset J$, it must be that either a or b is in $J \setminus J_1$, otherwise it would contradict that J is prime. Therefore, either a or b is in $\mathbb{K}[\mathbf{x}] \setminus \mathbb{K}[x_2, \ldots, x_n]$ and so ab must have some term that involves x_1, which implies $ab \notin J_1$, a contradiction.

($\Leftarrow$) Suppose J_1 is prime but that J is not. Then there must exist $a, b \in \mathbb{K}[\mathbf{x}] \setminus J$ with $ab \in J \setminus J_1$. Choose a and b so that ab has minimal x_1-degree among all such pairs. Let d be the x_1-degree of a and d' the x_1-degree of b. Since $ab \in J \setminus J_1$, $d + d' \geq 1$, and so without loss of generality we can assume $d \geq 1$. Write

$$a = h_0 + h_1 x_1 + h_2 x_1^2 + \ldots + h_d x_1^d,$$

where each $h_i \in \mathbb{K}[x_2, \ldots, x_n]$ and $h_d \neq 0$. Then since $f \in J$ and g is not a zero divisor modulo J, $a' := (ga - h_d x_1^{d-1} f)$ is not in J and has x_1-degree strictly less than d. It follows

that $a'b$ has x_1-degree strictly less than that of ab. Finally, since ab and f are in J, $a'b = gab - h_d x_1^{d-1} fb$ is in J, contradicting the minimality of the x_1-degree of ab. □

1.4.3 Tropical Secant Dimensions

In order to apply the prime-dimension approach we must be able to determine the dimension of an ideal. Again, given a set of generators, this can be done using computer algebra systems. However, in this thesis, we will also need to determine the dimension of an ideal for which generators of the ideal are not yet known. For example, in the previous section we described how we will want to use the prime-dimension approach to determine generators for an ideal I, the vanishing ideal of the image of a morphism. This requires that we know the dimension of I, but of course, we will not be able to use a generating set to determine this. Instead, we will need a way to determine $\dim(I)$ only from the morphism. In such a case, the tropical secant dimension approach of [Dra08] allows us to establish lower bounds on $\dim(I)$.

When taking the join of r varieties contained in $\mathbb{K}^n$, we introduce $(r-1)$ parameters, giving us the following bound

$$\dim(V_1 * \ldots * V_r) \leq \min\left\{\sum_{i=1}^{k} \dim(V_i) + (r-1), n\right\},$$

This upper bound is called the *expected dimension* and any join variety realizing this bound is called *nondefective*. The expected dimension also gives us an upper bound on the dimension of the ideal $I(V)$ where V is a join variety. Another way to establish an upper bound for $I(V)$ is to find an ideal J of known dimension with $J \subseteq I(V)$, which implies $\dim(I(V)) \leq \dim(J)$. If we can construct equal upper and lower bounds using these methods and the tropical secant dimension approach, we will be able to determine $\dim(I(V))$ exactly.

The remaining definitions and terminology in this section are adapted from the more general presentation in [Dra08]. We refer the reader there and to [MS15] for a background on tropical geometry. The tropical secant dimension approach actually applies to joins of affine cones. That is, affine varieties closed under scalar multiplication. Since we may regard a projective variety as an affine cone we will be able to apply this theorem to projective varieties in Chapters 2, 3 and 4.

Let $C_1, \ldots, C_r$ be affine cones. Suppose further that $C_i = \overline{\mathrm{Im}(f_i)}$ where $f_i : \mathbb{C}^{m_i} \to \mathbb{C}^{|B|}$

is a morphism. For $1 \leq i \leq r$, we write f_i as a list $(f_{i,b})_{b \in B}$. For our purposes, we may assume that each $f_{i,b}$ is a monomial, so that $f_{i,b} = x^{\alpha_{i,b}}$. For affine cones the mixing parameters introduced when constructing the join variety are superfluous. Thus, we can write the *join* of the affine cones $C_1, \ldots, C_r$ as

$$C_1 + \ldots + C_r := \overline{\{c_1 + \ldots + c_r : c_i \in C_i, 1 \leq i \leq r\}}.$$

Definition 1.4.12. For $v = (v_1, \ldots, v_r) \in \bigoplus_{i=1}^{r} \mathbb{R}^{m_i}$, let

$$D_i(v) := \{\alpha_{i,b} : \langle v_i, \alpha_{i,b} \rangle > \langle v_j, \alpha_{j,b} \rangle \text{ for all } j \neq i\}.$$

If $\alpha_{i,b} \in D_i(v)$ then we say that *i wins b at v* and call $D_i(v)$ the set of *winning directions of i at v*.

Finally, we have all the requisite definitions to state the primary result we will need.

Lemma 1.4.13. *[Dra08] The affine dimension of $C_1 + \ldots + C_r$ is at least the maximum, taken over all $v = (v_1, \ldots, v_r) \in \bigoplus_{i=1}^{r} \mathbb{R}^{m_i}$, of the sum*

$$\sum_{i=1}^{r} \dim_{\mathbb{R}} \langle D_i(v) \rangle_{\mathbb{R}}.$$

In the following example, we show how to apply this lemma in the simple case of the second secant variety of the Veronese surface. In Example 2.3.1, we relate the notation and terminology above to join varieties associated to 3-class Jukes-Cantor mixture models.

Example 1.4.14. The Veronese surface is the projective variety V defined by the mapping $v : \mathbb{P}^2 \to \mathbb{P}^5$ where

$$[a : b : c] \mapsto [a^2 : b^2 : c^2 : bc : ac : ab]$$

To apply Lemma 1.4.13, let us consider this variety as an affine cone and write the secant as $C_1 + C_2$.

Now, $C_1 = \overline{\text{Im}(f_1)}$ where $f_1 : \mathbb{C}^3 \to \mathbb{C}^6$ is the morphism

$$f_1(a, b, c) = (a^2, b^2, c^2, bc, ac, ab),$$

and $C_2 = \overline{\mathrm{Im}(f_2)}$ where $f_2 : \mathbb{C}^3 \to \mathbb{C}^6$ is the morphism

$$f_2(d, e, f) = (d^2, e^2, f^2, ef, df, de).$$

In this example we will simply index the coordinates by $B = [6]$. Then the exponent vectors are

$$\begin{aligned}
\alpha_{1,1} &= (2,0,0) \quad & \alpha_{2,1} &= (2,0,0) \\
\alpha_{1,2} &= (0,2,0) \quad & \alpha_{2,2} &= (0,2,0) \\
\alpha_{1,3} &= (0,0,2) \quad & \alpha_{2,3} &= (0,0,2) \\
\alpha_{1,4} &= (0,1,1) \quad & \alpha_{2,4} &= (0,1,1) \\
\alpha_{1,5} &= (1,0,1) \quad & \alpha_{2,5} &= (1,0,1) \\
\alpha_{1,6} &= (1,1,0) \quad & \alpha_{2,6} &= (1,1,0).
\end{aligned}$$

To obtain a lower bound on $\dim(C_1 + C_2)$, we choose $(v_1, v_2) \in \mathbb{R}^3 \oplus \mathbb{R}^3$. For this example we will choose $v_1 = (1,3,5)$ and $v_2 = (2,6,3)$. The vector $v = (v_1, v_2)$ determines the winning directions at b. For example, 2 wins 1 at v because

$$v_2 \cdot \alpha_{2,1} = 4 > 2 = v_1 \cdot \alpha_{1,1}.$$

Continuing, we determine that 1 wins 3 and 5 at v and 2 wins $1, 2, 4$, and 6 at v. Thus,

$$D_1(v) = \left\{ \begin{pmatrix} 0 \\ 0 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \right\} \text{ and } D_2(v) = \left\{ \begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \right\}.$$

Lemma 1.4.13 tells us that $\dim(C_1 + C_2) \geq \dim_{\mathbb{R}}\langle D_1(v)\rangle + \dim_{\mathbb{R}}\langle D_2(v)\rangle = 2 + 3 = 5$. Since the expected dimension is 6, this tells us only that the dimension is either 5 or 6. We can compute the actual dimension in Macaulay2 to get $\dim(C_1 + C_2) = 5$. Thus, as a projective variety, $\dim(V * V) = 4$.

Our choice of v happened to give us a lower bound equal to the actual dimension, but of course this need not be the case. Choosing $v = (1,1,1,2,2,2)$ for example would have only told us that $\dim(C_1 + C_2) \geq 3$. Obtaining a tight bound is not simply a matter of choosing the right vector, as [Dra08] contains several examples where no choice of vector gives a tight lower bound.

1.5 Outline

In this thesis, we will use the tools described above to examine the properties and structure of different phylogenetic models. We begin in Chapter 2 by examining the identifiability of the 3-class Jukes-Cantor mixture model. The importance of identifiability for phylogenetic inference was discussed previously where we noted that the identifiability of the tree parameters is of particular importance. The analogous question for a phylogenetic mixture model is whether or not the multiset of tree parameters is identifiable.

The identifiability of both the tree parameter and the numerical parameters has already been established for the basic models of character evolution [Cha96] as well as for some of the more complex phylogenetic models [AAR08, AR08a, AR06]. A number of papers have examined the identifiability of mixture models with various restrictions on the topologies of the trees in the mixture [ARS12, MMS08, MS07, RS12]. Recent work has established the identifiability of the tree parameters for 2-class mixtures of both the Jukes-Cantor and Kimura 2-parameter models with no restrictions on the tree topologies [APRS11]. Our goal in this chapter is to extend the ideas from [APRS11] to larger class mixture models, in particular, to the 3-class Jukes-Cantor mixture model. We do not address the identifiability of the numerical parameters, but focus instead on the tree parameters of the model. Our main result is the following:

Theorem 2.0.1. *The tree parameters of the* 3*-class Jukes-Cantor mixture model are generically identifiable for trees with* ≥ 6 *leaves.*

The proof of this main result uses tools from algebraic geometry and combinatorics as well as some heavy symbolic computation. First, we use a combinatorial argument to show that to establish identifiability for models on n-leaf trees it is enough to establish identifiability for models on trees with six or fewer leaves. We then establish this result by comparing the ideals of phylogenetic invariants for all 3-class Jukes-Cantor mixture models on trees with fewer than six leaves. Along the way, we will use various arguments to simplify and reduce the total number of computations we need to perform.

The strand symmetric model (SSM) is a phylogenetic model designed to reflect the symmetry inherent in the double-stranded structure of DNA. In Chapter 3, we will determine the ideal of phylogenetic invariants for the SSM on the claw tree $K_{1,3}$. Results in [DK09] imply that generators of the ideal of phylogenetic invariants for the SSM on any binary tree can be determined from knowledge of the ideal on $K_{1,3}$. Thus, this is an

important first step in understanding the theoretical properties of the model for eventual application. In [CS05], the authors were able to determine equations in the ideal of phylogenetic invariants for the SSM on $K_{1,3}$. However, whether or not these equations generate the entire ideal was heretofore unknown. In this chapter, we will apply the prime-dimension approach of Section 1.4.2 to show that in fact these 50 equations generate the ideal of the SSM on $K_{1,3}$.

Finally, in Chapter 4, we study secant ideals associated to the CFN model. It is known that for any two n-leaf binary phylogenetic X-trees, the associated ideals of phylogenetic invariants for the CFN model have the same Hilbert series [BW07]. Moreover, there exists a single ideal, $\mathcal{I}^n$, of which the ideal associated to the CFN model of any n-leaf binary phylogenetic X-tree can be realized as an initial ideal [SX10]. In light of these results, we conjecture the following.

Conjecture 4.1.1. *Let $\mathcal{T}$ be an n-leaf binary phylogenetic X-tree and let ω be a weight vector such that $in_\omega(\mathcal{I}^n) = \mathcal{I}_\mathcal{T}$. Then $in_\omega(\mathcal{I}^n * \mathcal{I}^n) = \mathcal{I}_\mathcal{T} * \mathcal{I}_\mathcal{T}$.*

To investigate this conjecture, we first gather what evidence we can by determining the CFN secant ideals for both 6-leaf tree topologies. We then explore a related conjecture for a class of ideals associated to binary trees that can be constructed as initial ideals of $I_{2,n}$, the Plücker ideal. This last chapter will also reveal some connections between phylogenetic models and the Pfaffian ideals, a well-studied class of ideals in algebraic geometry.

Chapter 2

Identifiability of 3-Class Jukes-Cantor Mixtures

The goal of phylogenetic inference is to find a tree that captures the evolutionary relationships between species. However, as referenced in Section 1.1, various biological phenomena confound this effort. Individual genes may actually conform to different phylogenetic trees, telling conflicting stories about the species in which they reside. The result is that a model-based approach on a single tree may be doomed to fail. Suppose for example that one has aligned DNA sequences and that a certain portion of the sequences evolved according to a model on one tree and a different portion independently according to a model on another. Then the observed distribution on the n-tuples of DNA bases is unlikely to belong to either model. Instead, the observed distribution would be a weighted sum of two distributions, one from each model, where the weighting is according to the proportion of DNA that evolved according to each. Geometrically, the observed distribution would lie on a line between two probability distributions, one from each model.

Mixture models are designed to model the situation above by weighting distributions from multiple models to produce a single probability distribution. Of course, there is no reason that this process should be limited to two models and we call a mixture of r different models an *r-class* mixture. Since different portions of DNA may evolve according to the same tree but at different rates, there is also good reason to consider mixtures of models with the same tree parameter. Thus, instead of a single tree parameter, an r-class mixture model has a multiset of r tree parameters. Just as with a phylogenetic

model on a single tree, if the mixture model is to be informative, the multiset of tree parameters must be identifiable. This is perhaps even more of a concern with mixture models as we introduce new parameters and enlarge the set of distributions. Such issues of overparameterization were observed in [MS07] where it was shown that for the CFN model, a mixture of two models on the same tree could mimic a model on a single, entirely different tree.

In this chapter, we prove the following result.

Theorem 2.0.1. *The tree parameters of the* 3*-class Jukes-Cantor mixture model are generically identifiable for trees with* ≥ 6 *leaves.*

The proof of this main result will occupy the whole of the present chapter. In Section 2.1, we will demonstrate why algebraic geometry is the appropriate tool for studying these models by associating to each set of tree parameters an irreducible algebraic variety containing the possible distributions arising from the 3-class Jukes-Cantor mixture model on those trees. We will then show how the question of identifiability can be reduced to showing that for any two sets of tree parameters, the associated varieties are not contained in one another. To show the varieties are not contained in one another, it is enough to show that their vanishing ideals are not contained in one another. Isolating phylogenetic invariants for the mixture models will be a key part of this proof.

In Section 2.2 we will investigate the combinatorial properties of binary phylogenetic X-trees to show that it is not actually necessary to compare *arbitrary* sets of tree parameters. Instead, we will be able to obtain identifiability results for n-leaf trees by comparing mixtures on trees with six or fewer leaves. Thus, we will have a finite list of pairs of mixtures for which we must show the mutual noncontainment of their varieties.

Finally, in Section 2.4, we will combine the results from the previous sections to construct a finite list of specific pairs of mixtures that we must consider. We will then outline a method for finding phylogenetic invariants that distinguish these mixtures from one another and provide access to computations proving that they exist. Many, but not all pairs of triplets of trees are separated by linear invariants. For the triplets not separated by linear invariants, we will use the linear invariants in a novel way to construct separating invariants of higher degree.

2.1 Preliminaries

The tree parameter of a phylogenetic model is a binary phylogenetic X-tree with $X = [n]$. In this chapter, we will specifically consider the Jukes-Cantor model. There is a transition matrix associated to each edge of $\mathcal{T}$ with structure depicted in Figure 1.1c. Recall that the entries of the transition matrices are called the numerical parameters and that we have a polynomial map from the set of numerical parameters $\Theta_{\mathcal{T}}$ into the probability simplex Δ^{4^n-1},

$$\psi_{\mathcal{T}} : \Theta_{\mathcal{T}} \to \Delta^{4^n-1} \subseteq \mathbb{R}^{4^n}.$$

The image of this map is a set of distributions that we call the model, and we write $\mathcal{M}_{\mathcal{T}} = \mathrm{Im}(\psi_{\mathcal{T}})$.

Because the entries of each row of the transition matrices must sum to one, we essentially have one parameter along each edge in the Jukes-Cantor model, though we will often ignore this in order to homogenize the parameterization. Likewise, every numerical parameter must be a real number between zero and one. If we similarly ignore this restriction and simply regard $\psi_{\mathcal{T}}$ as a complex polynomial map, $\overline{\mathrm{Im}(\psi_{\mathcal{T}})} = V_{\mathcal{T}}$ is a complex algebraic variety.

An r-class mixture model enlarges the space of possible distributions by taking r models and introducing $r-1$ *mixing parameters*. The mixing parameters weight the distribution from each of the models according to the proportion of data arising from each. Note that the underlying tree parameters of the models need not be distinct. This allows us to account for different portions of DNA being explained by the same tree but with different choices of numerical parameters. Just as before, with fixed tree parameters, we have a map that takes a choice of numerical parameters for each model and a choice of mixing parameters and maps them to a probability distribution. Our primary object of interest will be the map for the 3-class Jukes-Cantor mixture, so we have

$$\psi_{\mathcal{T}_1,\mathcal{T}_2,\mathcal{T}_3} : \Theta_{\mathcal{T}_1} \times \Theta_{\mathcal{T}_2} \times \Theta_{\mathcal{T}_3} \times \Delta^2 \to \Delta^{4^n-1}$$

where

$$(s_1, s_2, s_3, \pi) \mapsto \pi_1 \psi_{\mathcal{T}_1}(s_1) + \pi_2 \psi_{\mathcal{T}_2}(s_2) + \pi_3 \psi_{\mathcal{T}_3}(s_3).$$

Here, $\pi = (\pi_1, \pi_2, \pi_3) \in \Delta^2$ is the vector of mixing parameters. Again, regarded as a complex polynomial map, $\overline{\mathrm{Im}(\psi_{\mathcal{T}_1,\mathcal{T}_2,\mathcal{T}_3})}$ is an algebraic variety.

In fact,

$$\overline{\mathrm{Im}(\psi_{\mathcal{T}_1,\mathcal{T}_2,\mathcal{T}_3})} = V_{\mathcal{T}_1} * V_{\mathcal{T}_2} * V_{\mathcal{T}_3},$$

where $V_{\mathcal{T}_1} * V_{\mathcal{T}_2} * V_{\mathcal{T}_3}$ is the join variety of $V_{\mathcal{T}_1}, V_{\mathcal{T}_2}$, and $V_{\mathcal{T}_3}$.

Before we formally define the concept of generic identifiability for r-class mixtures, we will introduce some convenient notation. Let $\mathcal{T}_X$ be the set of binary phylogenetic X-trees and let $\mathcal{T}_{X,r}$ be the set of r element multisets of elements of $\mathcal{T}_X$. Note that as in our mixture models, for $\mathcal{T} = \{\mathcal{T}_1, \ldots, \mathcal{T}_r\}$, the trees in $\mathcal{T}$ are not necessarily distinct. We will now write $\psi_{\mathcal{T}} := \psi_{\mathcal{T}_1,\ldots,\mathcal{T}_r}$.

Definition 2.1.1. [APRS11] The tree parameters of an r-tree mixture model are *generically identifiable* for n-leaf trees if for all $\mathcal{S}, \mathcal{T} \in \mathcal{T}_{[n],r}$ generic $(s_1, \ldots, s_r, \pi) \in \Theta_{\mathcal{S}_1} \times \cdots \times \Theta_{\mathcal{S}_r} \times \Delta^{r-1}$, and any $(t_1, \ldots, t_r, \pi') \in \Theta_{\mathcal{T}_1} \times \cdots \times \Theta_{\mathcal{T}_r} \times \Delta^{r-1}$, the equality

$$\psi_{\mathcal{S}}(s_1, \ldots, s_r, \pi) = \psi_{\mathcal{T}}(t_1, \ldots, t_r, \pi')$$

implies $\mathcal{S} = \mathcal{T}$.

In this chapter we will not need such generality, as we will specifically consider the 3-class Jukes-Cantor mixture model. In order to prove Theorem 2.0.1, we will translate this statement about identifiability into one about algebraic varieties.

Lemma 2.1.2. *[APRS11] Suppose $\mathcal{S}, \mathcal{T} \in \mathcal{T}_{[n],3}$. Then for the 3-class Jukes-Cantor mixture model, $V_{\mathcal{T}} \nsubseteq V_{\mathcal{S}}$ and $V_{\mathcal{S}} \nsubseteq V_{\mathcal{T}}$ implies that the set of numerical parameters mapping into $V_{\mathcal{S}} \cap V_{\mathcal{T}}$ is a set of Lebesgue measure 0.*

Notice that this algebraic characterization means that we are able to obtain results about the models $\mathcal{M}_{\mathcal{S}}$ and $\mathcal{M}_{\mathcal{T}}$ by working with the complex varieties $V_{\mathcal{S}}$ and $V_{\mathcal{T}}$. One strategy for proving generic identifiability of the 3-class Jukes-Cantor mixture model for n-leaf trees is then clear. We can simply list all elements of $\mathcal{T}_{[n],3}$ (which we will call *n-leaf triplets*) and generate the corresponding varieties. By Lemma 2.1.2, if we can show that any two of these varieties are mutually noncontained, then we will have established identifiability for n-leaf trees. As alluded to earlier, we will actually want to look at the ideals of phylogenetic invariants. We will often refer to the elements of $\mathcal{I}_{\mathcal{T}} = I(V_{\mathcal{T}_1} * V_{\mathcal{T}_2} * V_{\mathcal{T}_3})$ as phylogenetic invariants of $V_{\mathcal{T}} = V_{\mathcal{T}_1} * V_{\mathcal{T}_2} * V_{\mathcal{T}_3}$, (or occasionally just phylogenetic invariants of $\mathcal{T} = \{\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3\}$, or of the mixture model). At the level of ideals, our strategy translates into showing that for each $(\mathcal{S}, \mathcal{T}) \in \mathcal{T}_{[n],r} \times \mathcal{T}_{[n],r}$ with $\mathcal{S} \neq \mathcal{T}$,

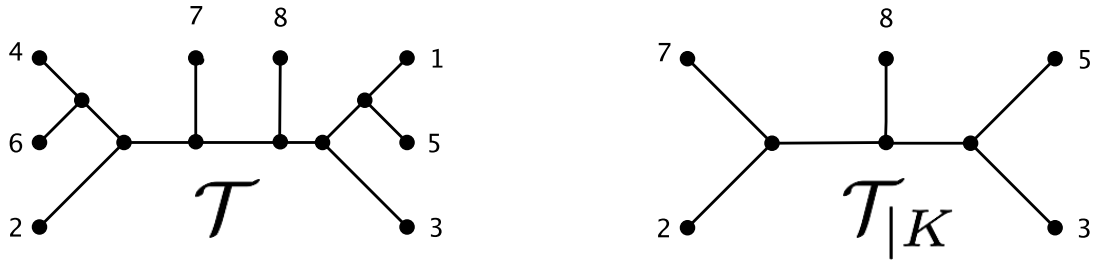
$\mathcal{I}_\mathcal{S} \not\subseteq \mathcal{I}_\mathcal{T}$. But to do this, we need not compute generators for the the ideals involved, but instead we only need to find an invariant of $\mathcal{I}_\mathcal{T}$ that is not an invariant of $\mathcal{I}_\mathcal{S}$, and vice versa. Once we have done this for a specific pair, we will say that we have *separated* $\mathcal{S}$ and $\mathcal{T}$.

This gives us a clear procedure for determining identifiability, but with some obvious drawbacks. First, the number of binary phylogenetic X-trees with n leaves is $(2n-5)!! = 1 \cdot 3 \cdot 5 \cdots (2n-5)$, which makes generating all possible 3-class mixtures computationally prohibitive even for relatively small n. Secondly, on the face of it, this brute force approach does not seem to offer any way of establishing generic identifiability of the tree parameters for arbitrary n. However, as we will see in the next section, it is possible to establish generic identifiability of the 3-class Jukes-Cantor mixture model for all n by separating only a finite number of mixtures.

2.2 Disentangling Trees

In this section we explain how to use trees with few leaves to establish identifiability for trees with an arbitrary number of leaves. The size of the trees we need to consider is bounded by the disentangling number for phylogenetic mixtures. For $\mathcal{T} \in \mathcal{T}_X$ and $K \subset X$, let $\mathcal{T}_{|K}$ be the tree obtained by suppressing all degree two vertices in the subtree of $\mathcal{T}$ induced by the leaves labeled by K. For $\mathcal{T} = \{\mathcal{T}_1, \ldots, \mathcal{T}_r\} \in \mathcal{T}_{X,r}$, $\mathcal{T}_{|K} = \{\mathcal{T}_{1|K}, \ldots, \mathcal{T}_{r|K}\}$.

Example 2.2.1. Consider $\mathcal{T} \in \mathcal{T}_{[8]}$ pictured below and $K = \{2, 3, 5, 7, 8\}$.



Definition 2.2.2. Let $\mathcal{S}, \mathcal{T} \in \mathcal{T}_{X,r}$ with $\mathcal{S} \neq \mathcal{T}$. A subset $K \subseteq X$ is said to disentangle $\mathcal{S}$ and $\mathcal{T}$ if $\mathcal{S}_{|K} \neq \mathcal{T}_{|K}$. Let $d(\mathcal{S}, \mathcal{T})$ be the cardinality of the minimum disentangling set of $\mathcal{S}$

and $\mathcal{T}$. The disentangling number $D(r)$ is

$$D(r) = \max_{n \in \mathbb{N}} \max_{\mathcal{S} \neq \mathcal{T} \in \mathcal{T}_{[n],r}} d(\mathcal{S}, \mathcal{T})$$

The following lemma [APRS11] motivates our interest in the disentangling number.

Lemma 2.2.3. *Let $\mathcal{S}, \mathcal{T} \in \mathcal{T}_{[n],3}$ and $K \subseteq [n]$. If $V_{\mathcal{S}_{|K}} \notin V_{\mathcal{T}_{|K}}$ then $V_{\mathcal{S}} \notin V_{\mathcal{T}}$.*

Now suppose we are able to show identifiability for trees with $D(3)$ leaves. Then given any $(\mathcal{S}, \mathcal{T}) \in \mathcal{T}_{[n],3} \times \mathcal{T}_{[n],3}$ with $\mathcal{S} \neq \mathcal{T}$ and $n > D(3)$, we can find some $K \subset [n]$ with $|K| = D(3)$ such that $\mathcal{S}_{|K} \neq \mathcal{T}_{|K}$, $V_{\mathcal{S}_{|K}} \notin V_{\mathcal{T}_{|K}}$, and $V_{\mathcal{T}_{|K}} \notin V_{\mathcal{S}_{|K}}$. By Lemma 2.2.3, in doing so we will have separated $\mathcal{S}$ and $\mathcal{T}$. Consequently, we would have generic identifiability of the tree parameters of the 3-class Jukes-Cantor mixture model for trees with $n \geq D(3)$ leaves. Thus, as promised, we will have an upper bound on the number of possible varieties we need to consider. In this section we provide some general background on the disentangling number and prove that $D(3) = 6$.

The *rooted* disentangling number, $RD(r)$, is defined analogously for rooted trees. We will omit the short proof of this lemma from [Sul12] that relates $RD(r)$ and $D(r)$.

Lemma 2.2.4. *The disentangling and rooted disentangling numbers satisfy: $D(r) \leq RD(r) + 1$.*

The main result of [Sul12] is the following theorem from which we obtain an upper bound on $D(r)$ as an immediate corollary.

Theorem 2.2.5. $RD(r) = 3(\lfloor log_2(r) \rfloor + 1)$.

Corollary 2.2.6. *For $r \in \mathbb{N}$, $D(r) \leq 3(\lfloor log_2(r) \rfloor + 1) + 1$.*

The original proof Theorem 2.2.5 is obtained by encoding multisets of trees as high-dimensional contingency tables and applying results about marginal maps. We provide an alternative, and hopefully more direct proof by examining the tree topologies directly.

Proof of Theorem 2.2.5. Let $\mathcal{R}_{X,r}$ be the set of r-element multisets of rooted binary phylogenetic X-trees. A construction in [Hum08] shows that $3(\lfloor \log_2(r) \rfloor + 1) \leq RD(r)$, so we need only show that for every pair $\mathcal{S}, \mathcal{T} \in \mathcal{R}_{X,r}$ with $\mathcal{S} \neq \mathcal{T}$, there is a disentangling set of cardinality less than or equal to $3(\lfloor \log_2(r) \rfloor + 1)$. We will proceed by induction on r. Because a rooted tree is determined by its rooted triples ([SS03, Theorem 6.4.1]),

the base case $RD(1) = 3$ is established. Assume this is true for all integers less than r and let $\mathcal{S}, \mathcal{T} \in \mathcal{R}_{X,r}$ with $\mathcal{S} \neq \mathcal{T}$. There must exists some i and j such that $\mathcal{S}_i \neq \mathcal{T}_j$. By our inductive assumption, we can permute the leaf labels so that for $K = \{1, 2, 3\}$, $\mathcal{S}_{i|K} \neq \mathcal{T}_{j|K}$. There are only three topologically distinct 3-leaf rooted binary phylogenetic X-trees, which we will label t_1, t_2, and t_3.

If $\mathcal{S}_{|K} \neq \mathcal{T}_{|K}$, then $\mathcal{S}$ and $\mathcal{T}$ are disentangled and we are done. Otherwise, $\mathcal{S}_{|K} = \mathcal{T}_{|K}$ is an unordered list of the trees t_1, t_2, and t_3 occurring with multiplicity. Partition $\mathcal{S}$ into three multisets,

$$L^l_{\mathcal{S}} := \{\mathcal{S}_j \in \mathcal{S} : \mathcal{S}_{j|K} = t_l\}$$

for $1 \leq l \leq 3$, and likewise for $\mathcal{T}$. Since K was chosen to disentangle an element of $\mathcal{S}$ from an element of $\mathcal{T}$, it must be the case that $\mathcal{S}_{|K}$ and $\mathcal{T}_{|K}$ contain at least two distinct 3-leaf trees. Therefore, we can choose l so that $L^l_{\mathcal{S}}$ is nonempty and $|L^l_{\mathcal{S}}| = r' \leq \frac{r}{2}$. Since $\mathcal{S}_{|K} = \mathcal{T}_{|K}$, $|L^l_{\mathcal{S}}| = |L^l_{\mathcal{T}}|$ and we can consider $L^l_{\mathcal{T}}$ and $L^l_{\mathcal{S}}$ as elements of $\mathcal{R}_{X,r'}$. By our inductive assumption, there exists a disentangling set K' of $L^l_{\mathcal{T}}$ and $L^l_{\mathcal{S}}$ such that

$$\begin{aligned}
|K'| &\leq 3(\lfloor \log_2(r') \rfloor + 1) \\
&\leq 3\left(\left\lfloor \log_2\left(\frac{r}{2}\right)\right\rfloor + 1\right) \\
&= 3((\lfloor \log_2(r) \rfloor - 1) + 1) \\
&= 3(\lfloor \log_2(r) \rfloor).
\end{aligned}$$

Therefore, $|K \cup K'| \leq 3(\lfloor \log_2(r) \rfloor + 1)$. We claim that this set disentangles $\mathcal{S}$ and $\mathcal{T}$. Since K' disentangles $L^l_{\mathcal{S}}$ from $L^l_{\mathcal{T}}$, and $K' \subseteq K \cup K'$, $(L^l_{\mathcal{S}})_{|K\cup K'} \neq (L^l_{\mathcal{T}})_{|K\cup K'}$. If $\mathcal{S}$ and $\mathcal{T}$ are still entangled, then there must be some tree in $(L^l_{\mathcal{S}})_{|K\cup K'}$ equal to some tree in $(L^m_{\mathcal{T}})_{|K\cup K'}$ with $l \neq m$. But since $K \subseteq K \cup K'$, this is impossible, so $K \cup K'$ disentangles $\mathcal{S}$ and $\mathcal{T}$. □

While this assures us that $D(3) \leq 7$, we can actually reduce this bound slightly, vastly reducing the number of triplet pairs we need to consider. For the theorem and proof that follow, we will make use of the following definition.

Definition 2.2.7. For $\mathcal{T} \in \mathcal{T}_{[n]}$ and K a three element subset of $[n]$, if $K|([n] \setminus K)$ is a split of $\mathcal{T}$ then the 3-leaf rooted tree induced by K is a *cluster on K*.

Definition 2.2.8. Let $\mathcal{S}, \mathcal{T} \in \mathcal{T}_{X,r}$ and K a subset of X that does not disentangle $\mathcal{S}$ and $\mathcal{T}$. Label the trees of $\mathcal{S}$ and $\mathcal{T}$ so that $\mathcal{S} = \{\mathcal{S}_1, \ldots, \mathcal{S}_r\}$ and $\mathcal{T} = \{\mathcal{T}_1, \ldots, \mathcal{T}_r\}$. For $1 \leq m \leq r$,

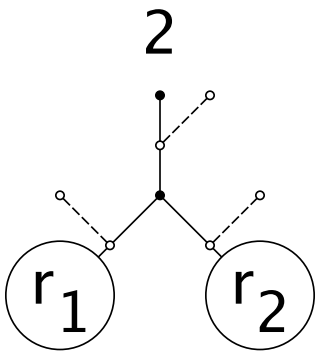


Figure 2.1: Possible locations for e_1 from the proof of Theorem 2.2.9.

let

$$m_i = \min(\{m \in [r] \setminus \{m_1, \ldots, m_{i-1}\} : \mathcal{S}_{m|K} = \mathcal{T}_{i|K}\})$$

Then with respect to the chosen labeling, we say that $\mathcal{S}_{m_i}$ and $\mathcal{T}_i$ are *partners* at K.

Notice that each tree of $\mathcal{T}$ has exactly one partner at K, and that the partnered trees at K are exactly the same if we swap the roles of $\mathcal{S}$ and $\mathcal{T}$ in the definition.

Theorem 2.2.9. $D(3) \leq 6$.

Proof. We will use contradiction. Suppose $D(r) = 7$ and let $K_i = [7] \setminus \{i\}$. Then there must exist $\mathcal{S}, \mathcal{T} \in \mathcal{T}_{[7],3}$ such that $\mathcal{S}_{|K_i} = \mathcal{T}_{|K_i}$ for $1 \leq i \leq 7$. For everything that follows, fix some labeling of the trees of $\mathcal{S}$ and $\mathcal{T}$ so that for each i, every tree of $\mathcal{S}$ and $\mathcal{T}$ has a partner at K_i.

We will collect a few key observations about trees that are partnered together at multiple K_i. If a tree of $\mathcal{S}$ and a tree of $\mathcal{T}$ are partnered together at K_i for exactly j distinct values of i, then we will call them j-partners. Suppose $\mathcal{S}_l \neq \mathcal{T}_m$ are 2-partners and permute the leaf labels so that they are partnered at K_1 and at K_2. Let v_i be the leaf vertex labeled i, and let e_i be the edge adjacent to this vertex. Since $\mathcal{S}_{l|K_1} = \mathcal{T}_{m|K_1}$, there must be a unique edge on this tree where e_1 is attached to form $\mathcal{S}_l$, and a different unique edge where e_1 is attached to form $\mathcal{T}_m$. But in order for our trees to still be equal when restricted to K_2, the two distinct edges where we attached e_1 must collapse to the same edge when we remove e_2. Therefore, $\mathcal{S}_l$ and $\mathcal{T}_m$ must have the structure of the tree in Figure 2.1 where r_1 and r_2 are rooted trees and where e_1 is one of the dashed edges.

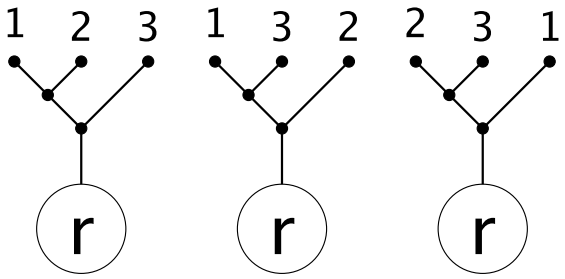


Figure 2.2: Possible structures for 3-partners at $\{K_1, K_2, K_3\}$.

Now suppose that $\mathcal{S}_l$ and $\mathcal{T}_m$ are 3-partners, partnered at K_1, K_2 and K_3. From above, observe that regardless of which edge is e_1, that the length of the path from v_1 to v_2 in both trees must be less than or equal to three. Therefore, the length of the path between each pair of vertices, (v_1, v_3), (v_1, v_2), and (v_2, v_3), must be less than three in both $\mathcal{S}_l$ and $\mathcal{T}_m$. Consequently, $\mathcal{S}_l$ and $\mathcal{T}_m$ must both have a cluster on $\{1, 2, 3\}$, and they must be the same tree apart from these clusters (i.e., $\mathcal{S}_l$ and $\mathcal{T}_m$ must be two different trees from the list in Figure 2.2, where r is some rooted tree). From the figure we also see if $\mathcal{S}_{l|K_i} = \mathcal{T}_{m|K_i}$ for $i \notin \{1, 2, 3\}$, then $\mathcal{S}_l = \mathcal{T}_m$. This implies that for $j > 3$ any two trees that are j-partners must be the same tree.

If any tree of $\mathcal{S}$ is equal to any tree of $\mathcal{T}$, then we can remove these trees to form the multisets $\mathcal{S}', \mathcal{T}' \in \mathcal{T}_{[7],2}$, and any set K that disentangles $\mathcal{S}'$ and $\mathcal{T}'$ will disentangle $\mathcal{S}$ and $\mathcal{T}$. Since $D(2) = 6$ ([MMS08]) this would imply $(\mathcal{S}, \mathcal{T}) \leq 6$ contradicting our assumption that $d(\mathcal{S}, \mathcal{T}) = 7$. Therefore, we can assume that no two trees are j-partners for $j > 3$. Since each tree of $\mathcal{S}$ and each tree of $\mathcal{T}$ must be partnered at all seven K_i, the only possibility for a single tree is that it has one 3-partner and two 2-partners or two 3-partners and one 1-partner. The particular partnering relationships impose restrictions on the possible structures of the trees in $\mathcal{S}$ and $\mathcal{T}$. We will now consider both cases and use these restrictions to arrive at a contradiction.

Case 1: There exists a tree in $\mathcal{S}$ or $\mathcal{T}$ with one 3-partner and two 2-partners.

We will leave all partners fixed according to the original labeling. However, for convenience, we will relabel the multisets, trees, and leaves so that $\mathcal{T}_1$ is the tree from the assumption that is partnered with $\mathcal{S}_1$ at $\{K_1, K_2, K_3\}$, with $\mathcal{S}_2$ at $\{K_4, K_5\}$, and with $\mathcal{S}_3$ at $\{K_6, K_7\}$.

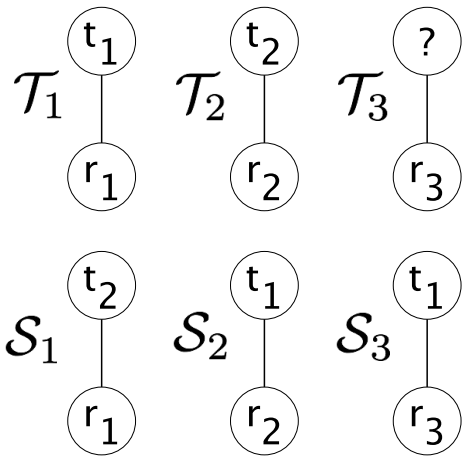


Figure 2.3: Structure of trees satisfying Case 1 of Theorem 2.2.9.

So far then, we know $\mathcal{T}_1$ and $\mathcal{S}_1$ are as in Figure 2.3, where t_1 and t_2 are distinct clusters on $\{1,2,3\}$ and r_1 is a rooted tree with leaf label set $\{4,5,6,7\}$. We also know that $\mathcal{T}_{1|K_4} = \mathcal{S}_{2|K_4}$, so $\mathcal{S}_{2|K_4}$ contains the cluster t_1. If e_4 is connected to an edge of $\mathcal{S}_2$ somewhere in the cluster t_1 of $\mathcal{S}_{2|K_4}$, then it is impossible for $\mathcal{S}_{2|K_5} = \mathcal{T}_1$. Therefore, we see that even without restricting to K_4, $\mathcal{S}_2$ and similarly $\mathcal{S}_3$ must contain the cluster t_1. Thus, $\mathcal{T}_1, \mathcal{S}_1, \mathcal{S}_2$, and $\mathcal{S}_3$ are all as depicted in Figure 2.3.

Every tree must have a 3-partner, so let $\mathcal{T}_2$ be a 3-partner of $\mathcal{S}_2$. From our observations above, $\mathcal{S}_2$ and $\mathcal{T}_2$ differ only by a cluster on some three element set K'. We know that K' cannot contain 4 or 5 since $\mathcal{S}_2$ is partnered with $\mathcal{T}_1$ at K_4 and K_5. Therefore, K' contains at least one element of $\{1,2,3\}$. But to preserve the cluster t_1, it must be that $K' = \{1,2,3\}$. If $\mathcal{T}_{2|K'} = t_3$, then K' disentangles $\mathcal{S}$ and $\mathcal{T}$, and if $\mathcal{T}_{2|K'} = t_1$ then $\mathcal{T}_2 = \mathcal{S}_2$, it follows that $\mathcal{T}_{2|K'} = t_2$.

Finally, $\mathcal{S}_1$ and $\mathcal{T}_1$ as well as $\mathcal{S}_2$ and $\mathcal{T}_2$ are partnered at K_1, K_2, and K_3, which forces $\mathcal{S}_3$ and $\mathcal{T}_3$ to be partnered at K_1, K_2, and K_3. As a result, $\mathcal{S}_3$ and $\mathcal{T}_3$ differ only by a cluster on $\{1,2,3\}$ (Figure 2.3). If $\mathcal{T}_{3|K'} = t_1$ then $\mathcal{T}_3 = \mathcal{S}_3$ and if $\mathcal{T}_3 = t_2$ or $\mathcal{T}_3 = t_3$ then K' is a disentangling set. In any case we have a contradiction.

Case 2: Every tree in $\mathcal{S}$ and $\mathcal{T}$ has two 3-partners and one 1-partner.

As before, leave the partnering relationships fixed and relabel the multisets, trees, and

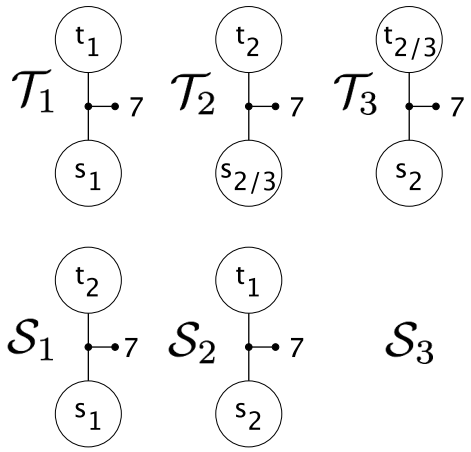


Figure 2.4: Structure of trees satisfying Case 2 of Theorem 2.2.9.

leaves so that $\mathcal{T}_1$ is partnered with $\mathcal{S}_1$ at $\{K_1, K_2, K_3\}$, with $\mathcal{S}_2$ at $\{K_4, K_5, K_6\}$, and with $\mathcal{S}_3$ at $\{K_7\}$. As we've seen, $\mathcal{T}_1$ must have a cluster on $\{1, 2, 3\}$ and a cluster on $\{4, 5, 6\}$, so $\mathcal{T}_1$ is as pictured in Figure 2.4. As 3-partners of $\mathcal{T}_1$, both $\mathcal{S}_1$ and $\mathcal{S}_2$ must have clusters on $\{1, 2, 3\}$ and $\{4, 5, 6\}$ as well and are also as depicted in Figure 2.4.

$\mathcal{S}_1$ must have another 3-partner which we will label $\mathcal{T}_2$. $\mathcal{S}_1$ and $\mathcal{T}_2$ must differ only at a cluster on some three element set $K' \subset [7] \setminus \{1, 2, 3\}$ which must contain elements of $\{4, 5, 6\}$. Then as we argued above, to preserve the $\{4, 5, 6\}$ cluster on $\mathcal{S}_1$ it must be that $K' = \{4, 5, 6\}$. $\mathcal{S}_2$ must also have a second 3-partner. This tree can not be $\mathcal{T}_2$, since then both $\mathcal{T}_1$ and $\mathcal{T}_2$ would have two 3-partners, leaving $\mathcal{T}_3$ as the sole 3-partner for $\mathcal{S}_3$ and putting us back in Case 1. Therefore, $\mathcal{S}_2$ and $\mathcal{T}_3$ must be 3-partners, and the same logic shows that they differ only at a cluster on $\{1, 2, 3\}$. The possible structures of $\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3, \mathcal{S}_1$, and $\mathcal{S}_2$ are all displayed in Figure 2.4 ($t_{i/j}$ indicates that a cluster can be only either t_i or t_j and likewise for $s_{i/j}$). Since the 3-partners of $\mathcal{S}_1$ are $\mathcal{T}_1$ and $\mathcal{T}_2$, the 1-partner of $\mathcal{S}_1$ must be $\mathcal{T}_3$, and $\mathcal{S}_1$ and $\mathcal{T}_3$ must be partnered at K_7. From the diagram, it is clear $\mathcal{S}_{1|K_7} \neq \mathcal{T}_{3|K_7}$, which is a contradiction. □

2.3 Dimension

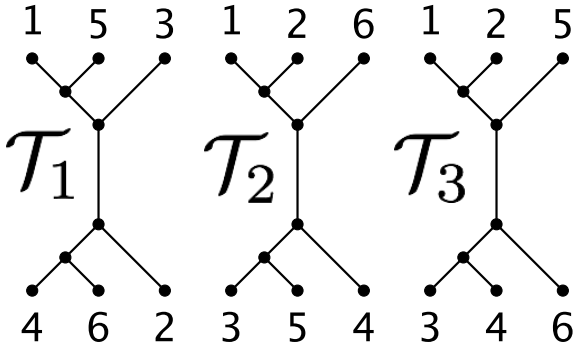
As a corollary to Lemma 2.1.2, we have a strategy for proving the identifiability of the 3-class Jukes-Cantor mixture model for n-leaf trees. We simply look at all pairs $(\mathcal{S}, \mathcal{T}) \in \mathcal{T}_{[n],3} \times \mathcal{T}_{[n],3}$ with $\mathcal{S} \neq \mathcal{T}$ and show that $\mathcal{I}_{\mathcal{S}} \not\subseteq \mathcal{I}_{\mathcal{T}}$ and that $\mathcal{I}_{\mathcal{T}} \not\subseteq \mathcal{I}_{\mathcal{S}}$. Our goal in this section is to prove that for every $(\mathcal{S}, \mathcal{T}) \in \mathcal{T}_{[n],3} \times \mathcal{T}_{[n],3}$, $\dim(\mathcal{I}_{\mathcal{S}}) = \dim(\mathcal{I}_{\mathcal{T}})$. We have already seen in Example 1.4.5 that both ideals are prime. If they are also of the same dimension, then Proposition 1.4.10 tells us that they are either mutually not contained or equal. Thus, proving this dimension result will allow us to establish mutual noncontainment by simply showing that $\mathcal{I}_{\mathcal{S}} \neq \mathcal{I}_{\mathcal{T}}$. Since eventually our proof will necessitate separating many triplet pairs, this will greatly reduce the number of invariants we have to find. Moreover, it will make the task of separating pairs much easier in cases where finding an invariant to establish noncontainment in one direction is more difficult than doing so in the other.

To deduce the dimension of these ideals we will use the parameterization of the underlying varieties and apply the tropical secant dimension technique that we introduced in Section 1.4.3. In particular, we will see that it is enough to show that each of these ideals is nondefective. For all of the work that follows, we will work in the Fourier coordinates that we introduced in Section 1.4.1. Importantly for our purposes, the linearity of the transform means that it commutes with taking mixtures. To see how we will use nondefectiveness, let $\mathcal{T} = \{\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3\}$ be a 6-leaf triplet. We know that each tree in $\mathcal{T}$ has 9 edges, and so from the Fourier parameterization of the Jukes-Cantor model, it is obvious that $\dim(V_{\mathcal{T}_i}) \leq 9$. This comes from a simple count of the parameters, remembering that each of the a_A^e is actually equal to one in the Fourier coordinates. Therefore, if we can show that $\dim(V_{\mathcal{T}_1} * V_{\mathcal{T}_2} * V_{\mathcal{T}_3}) = 29$, we will have shown that this variety is nondefective. Now suppose we could show that the join variety of every 6-leaf triplet is nondefective, then as a corollary, the join varieties of any two 6-leaf triplets are the same dimension, and so too are the ideals.

Our approach is inspired by the proof of the nondefectiveness of the second secant variety associated to a 4-leaf tree as shown in ([APRS11]). Let $\mathcal{T} \in \mathcal{T}_{[n],r}$, we will temporarily regard the a_A^e as variables which homogenizes the parameterization of $V_{\mathcal{T}_i}$ for $1 \leq i \leq r$. Now we have projective varieties which we will regard as affine cones, $C_1, \ldots, C_r$. We can apply Lemma 1.4.13 because each of these cones is the image of a polynomial map, $f_i : \mathbb{C}^{m_i} \to \mathbb{C}^{|B|}$. We illustrate how the components of the lemma correspond to the

3-class Jukes-Cantor mixture model in the example below.

Example 2.3.1. Consider the 3-class Jukes-Cantor mixture model for the trees pictured.



In this example, $r = 3$. After the Fourier transformation there are two parameters associated to each of the nine edges of the tree (a_A^e and a_C^e). The dimension of the parameter space for each cone is the same and so $m_1 = m_2 = m_3 = 18$. The coordinates of the image space are indexed by 6-tuples of the DNA bases, thus $B = \{A, C, G, T\}^6$ and for $1 \leq r \leq 3$, f_i is a map from $\mathbb{C}^{18}$ into $\mathbb{C}^{4^6}$.

As outlined in Section 1.4.1, after the Fourier transformation each coordinate in the image space is parameterized by a monomial where every exponent is either 0 or 1. To illustrate one particular coordinate,

$$
\begin{aligned}
f_{1,CACCGT} &= a_C^1 a_A^2 a_C^3 a_C^4 a_C^5 a_C^6 a_C^{15|2346} a_C^{46|1235} a_C^{135|246} \\
f_{2,CACCGT} &= b_C^1 b_A^2 b_C^3 b_C^4 b_C^5 b_C^6 b_C^{12|3456} b_C^{15|2346} b_C^{126|345} \\
f_{3,CACCGT} &= c_C^1 c_A^2 c_C^3 c_C^4 c_C^5 c_C^6 c_C^{12|3456} c_A^{34|1256} c_C^{125|346}.
\end{aligned}
$$

Defining a different parameter vector for each of the trees in the mixture, we can write each $f_{i,b} = x^{\alpha_{i,b}}$ for some $\alpha_{i,b} \in \{0,1\}^{18}$. Using a colon to separate the a_C^e and a_A^e coordinates, the exponent vectors are

$$\alpha_{1,CACCGT} = (1,0,1,1,1,1,1,1,1:0,1,0,0,0,0,0,0,0)$$
$$\alpha_{2,CACCGT} = (1,0,1,1,1,1,1,1,1:0,1,0,0,0,0,0,0,0)$$
$$\alpha_{3,CACCGT} = (1,0,1,1,1,1,1,0,1:0,1,0,0,0,0,0,1,0).$$

To apply Lemma 1.4.13, we choose a vector $v = (v_1, v_2, v_3) \in \mathbb{R}^{18} \oplus \mathbb{R}^{18} \oplus \mathbb{R}^{18}$, which will partition the exponent vectors above into the "winning directions" of Definition 1.4.12.

Lemma 1.4.13 gives us a way to compute lower bounds on the dimensions of join varieties. Now to show that a join variety is nondefective, we just need to generate a lower bound that is equal to the expected dimension.

Theorem 2.3.2. *Let $\mathcal{T} \in \mathcal{T}_{[n],r}$. For $n \geq 4$ and $r \leq \lceil \frac{n}{2} \rceil$, the join variety $V_{\mathcal{T}_1} * \ldots * V_{\mathcal{T}_r}$ associated to the r-class Jukes-Cantor mixture model is nondefective.*

Proof. Let $\mathcal{T} \in \mathcal{T}_{[n],r}$, by Lemma 1.4.13, to show nondefectiveness it will be enough to find a vector $v = (v_1, \ldots, v_r)$ so that for $1 \leq i \leq r$, $\dim_{\mathbb{R}} \langle D_i(v) \rangle_{\mathbb{R}} = (2n-3)+1$. Thus, $C_1 + \ldots + C_r$ will have affine dimension $r(2n-3)+r$, and after we set each $a^e_A = 1$, the projective dimension of $V_{\mathcal{T}_1} * \ldots * V_{\mathcal{T}_r}$ will be $r(2n-3)+(r-1)$ as desired.

The set of winning directions of i at v, $D_i(v)$, is a set of 0/1 vectors in $\mathbb{R}^{4n-6}$. Our goal will be to construct the vector v in such a way that the vectors in each $D_i(v)$ span a space of dimension $2n-2$. Recall that for a tree $\mathcal{T}$, the distinct Fourier coordinates are in bijection with the subforests of $\mathcal{T}$. Therefore, each $b \in B$ induces a subforest on the trees $\mathcal{T}_1, \ldots, \mathcal{T}_r$, the number of leaf edges of which, t_b, depends only on the number of entries that are not A in b. For example, if $b = AACGT$ then $t_b = 3$ since in *any* 5-leaf tree the subforest induced by b contains three leaf edges.

Case 1: r is even

Construct the vector $v = (v_1, \ldots, v_r)$ as follows. Each v_i has $2n$ entries corresponding to leaf edges, half of which correspond to the variables a^e_C and half to the homogenizing variables a^e_A. Let the entries of v_i corresponding to leaf edges a^e_C be equal to γ_i, to leaf edges a^e_A equal to δ_i, and set all other entries equal to zero.

Then for $1 \leq i \leq r$, $b \in B$,

$$\langle v_i, \alpha_{i,b} \rangle = \gamma_i t_b + (n - t_b)\delta_i = (\gamma_i - \delta_i)t_b + n\delta_i.$$

Notice that this value depends only on the number of leaf edges in the subforest of $\mathcal{T}_i$ induced by b. Let $\mu_i : \mathbb{R} \to \mathbb{R}$ be given by

$$t \mapsto (\gamma_i - \delta_i)t + n\delta_i.$$

The two parameters γ_i and δ_i allow us to make μ_i whatever line we wish in $\mathbb{R}^2$. Now we have $\langle v_i, \alpha_{i,b} \rangle = \mu_i(t_b)$, and as explained, we do not need to know anything about the topology of the tree $\mathcal{T}_i$ to compute this value. Thus, if $\mu_i(t) < \mu_j(t)$ for all $j \neq i$, then for any b with $t_b = t$, i wins b at v.

Choose γ_i and δ_i so that $\min_j(\mu_j(t))$ is a continuous piecewise linear function with

$$\min_j(\mu_j(t)) = \begin{cases} \mu_1(t) & \text{if } t \in [0, \frac{5}{2}] \\ \mu_i(t) & \text{if } t \in [2i - \frac{3}{2}, 2i + \frac{1}{2}] \text{ for } 1 < i \leq r \end{cases}$$

Then 1 wins b at v if $t_b = 0, 2$, and for $1 < i \leq r$, i wins b at v if $t_b = 2i$ or $2i - 1$.

Form the matrices $M_i(v)$ with columns equal to the vectors in $D_i(v)$. Now we just need to show that for $1 \leq i \leq r$, $\operatorname{rank}(M_i(v)) = (2n - 3) + 1$. In order to do so, we will reinterpret our matrices in order to utilize previous results about reconstructing trees from subtree weights. Let $\mathcal{T}'$ be a tree, and assign to each edge a positive weight $w(e)$. Define the weight of a subforest to be the sum of the weights of the edges contained in the subforest. Let $M_i'(v)$ be the matrix consisting of the first $2n - 3$ rows of $M_i(v)$ and

$$w = \begin{pmatrix} w(e_1) \\ \vdots \\ w(e_{2n-3}) \end{pmatrix},$$

then $M_i'(v)^T w$ is a column vector with j-th entry equal to the weight of the subforest corresponding to the j-th column of $M_i'(v)$.

$M_1'(v)$ contains column vectors corresponding to the empty subforest as well as the subforests with two leaf edges. A subforest with exactly two leaf edges with degree one vertices u and v is just the path between u and v with weight $d(u, v)$. Therefore, for fixed w, the entries of the column vector $M_1'(v)w$ determine a tree metric $\delta_{\mathcal{T}'}$ which has

a graph realization $\mathcal{T}'$. By the *Tree-Metric theorem* ([PS04, SS03]) $\mathcal{T}'$ is the unique tree metric representation of $\delta_{\mathcal{T}'}$, and w is the unique solution to

$$M_1'(v)^T x = M_1'(v)^T w.$$

Therefore, we know $\operatorname{rank}(M_1'(v)) = 2n-3$. In fact, we have that the column rank of just the columns corresponding to subforests with two leaf edges is equal to $2n-3$. Let $(x_1, \ldots, x_{2n-3}, y_1, \ldots, y_{2n-3})^T$ be an arbitrary vector in $\mathbb{R}^{4n-6}$ where $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ are the coordinates corresponding to leaf parameters. Then each of the columns is contained in the subspace defined by

$$x_1 + \ldots + x_n = \frac{2}{n-2}(y_1 + \ldots + y_n).$$

The column corresponding to the empty subforest is clearly not contained in this subspace, so its addition increases the column rank by one, which implies $\operatorname{rank}(M_1(v)) = 2n-2$.

For $i \geq 2$, in order to show that $\operatorname{rank}(M_i(v)) = 2n-2$, we will first show that we can recover every edge weight if we know the weight of the subforests on $2i$ and $2i-1$ leaves. To determine the weight of a leaf edge uv with degree three vertex u, choose a subforest with $2i$ leaves that includes all three edges incident to u. Choosing such a subforest is always possible since $2i \geq 3$. Removing uv results in a subforest with $2i-1$ leaves with corresponding vector also in $D_i(v)$. The difference of the weights of these two subforests determines the weight of the leaf edge.

For an internal edge uv, we construct a subforest that includes the edge uv and the other four edges incident to either u or v. The fact that $2i \geq 4$ ensures that such a subforest exists. Again, omitting uv from this subforest gives us a different subforest on $2i$ leaves, and subtracting, we obtain the weight of uv.

For each edge, we found two subforests that differed by exactly that edge. Subtracting these vectors we obtain every column of the matrix $\begin{pmatrix} I \\ -I \end{pmatrix}$, where I is the $(2n-3)\times(2n-3)$ identity matrix. Anything in the column span of these vectors possesses the property that the entry for a_C^e is just the negative of the entry for a_A^e. Therefore, adding any vector without this property to the set, that is, adding any of the other subforest vectors, increases the rank by one. Thus, $\operatorname{rank}(M_i(v)) = 2n-2$.

Case 2: r is odd

We construct the vector $v = (v_1, \ldots, v_r)$ as in the first case so that $\min_j(\mu_j(t))$ is a continuous piecewise linear function, but with

$$\min_j(\mu_j(t)) = \begin{cases} \mu_1(t) & \text{if } t \in [0, \frac{5}{2}] \\ \mu_2(t) & \text{if } t \in [\frac{5}{2}, 4] \\ \mu_3(t) & \text{if } t \in [4, \frac{11}{2}] \\ \mu_i(t) & \text{if } t \in [2i - \frac{5}{2}, 2i - \frac{1}{2}] \text{ for } 3 < i \leq r. \end{cases}$$

Notice that no i wins v at b if $t_b = 4$, but that if $t_b \neq 4$, then

- 1 wins b at v if $t_b = 0$ or 2
- 2 wins b at v if $t_b = 3$
- 3 wins b at v if $t_b = 5$
- for $3 < i \leq r$, i wins b at v if $t_b = 2i - 2$ or $t_b = 2i - 1$.

Let s_b be the number of internal edges of the subforest induced by b on $\mathcal{T}_3$. We will perturb the entries of v_3 so that none of the above winning directions is affected but so that 2 wins b at v if $t_b = 4$ and $s_b = 0$ and 3 wins b at v if $t_b = 4$ and $s_b > 0$.

First we will show that for $t_b = 4$ and $s_b = 0$, $\langle v_3, \alpha_{3,b} \rangle > \langle v_2, \alpha_{2,b} \rangle$. Set the entries of v_3 corresponding to internal edges a^e_C equal to ζ and to internal edges a^e_A equal to η. Let $\tilde{\mu}_3(s,t) = \mu_3(t) + \zeta s + (n-3-s)\eta$. Now $\langle v_3, \alpha_{3,b} \rangle = \tilde{\mu}_3(s_b, t_b)$ and for $i \neq 3$, $\langle v_i, \alpha_{i,b} \rangle = \mu_i(t_b)$. Let $\eta = \epsilon > 0$ and $\zeta = -\epsilon(n - 3.9)$. If $t_b = 4$ and $s_b = 0$, then $\langle v_3, \alpha_{3,b} \rangle = \tilde{\mu}_3(0, 4) = \mu_3(4) + \epsilon(n-3) = \mu_2(4) + \epsilon(n-3) = \langle v_2, \alpha_{2,b} \rangle + \epsilon(n-3) > \langle v_2, \alpha_{2,b} \rangle$.

By our choice of ζ and η, as s increases $\tilde{\mu}_3(s,t)$ decreases. Therefore, to show that $\langle v_3, \alpha_{3,b} \rangle < \langle v_2, \alpha_{2,b} \rangle$ when $t_b = 4$ and $s_b > 0$, it is enough to show that $\langle v_3, \alpha_{3,b} \rangle < \langle v_2, \alpha_{2,b} \rangle$ when $t_b = 4$ and $s_b = 1$. In that case, we have $\langle v_3, \alpha_{3,b} \rangle = \tilde{\mu}_3(1, 4) = \mu_3(4) - \epsilon(n - 3.9) + \epsilon(n-4) < \mu_3(4) = \mu_2(4) = \langle v_2, \alpha_{2,b} \rangle$. Now choose $\epsilon > 0$ small enough so that we do not affect any of the other winning directions. Then as desired, 2 wins b at v if $t_b = 4$ and $s_b = 0$ and 3 wins b at v if $t_b = 4$ and $s_b > 0$.

Now we would like to show that for $1 \leq i \leq r$, $\dim_{\mathbb{R}} \langle D_i(v) \rangle_{\mathbb{R}} = 2n - 2$. When $i = 1$, the proof is the same as in the even case. Likewise, for $i > 3$, $D_i(v)$ contains vectors

corresponding to subforests with $2i-2$ and $2i-1$ leaves. Essentially the same arguments from the even case where $D_i(v)$ contained vectors corresponding to subforests with $2i-1$ and $2i$ leaves show that $\dim_{\mathbb{R}}\langle D_i(v)\rangle_{\mathbb{R}} = 2n-2$. Therefore, we just need to establish the rank of $M_2(v)$ and $M_3(v)$.

$M_2'(v)$ contains vectors corresponding to every 3-leaf subtree of $\mathcal{T}_2$. Just as before, given the weight vector w, $M_2'(v)^T w$ encodes the weight of every 3-leaf subtree. The main theorem in [PS04] implies that these weights uniquely determine w, so $\operatorname{rank}(M_2'(v)) = 2n-3$. Every one of these vectors is contained in the hyperplane defined by

$$x_1 + \ldots + x_n = \frac{3}{n-3}(y_1 + \ldots + y_n).$$

Since any binary tree with four or more leaves contains at least two cherries, $\mathcal{T}_3$ has a 4-leaf subforest with no internal edges, and 2 wins v at the coordinate corresponding to this subforest. The vector corresponding to this subforest is not contained in the hyperplane above, so $\operatorname{rank}(M_2(v)) = 2n-2$.

Again as before, for each edge e of $\mathcal{T}_3$ we use two subforests from the set of 4 and 5-leaf subforests that differ only by e to determine $w(e)$. To carry out this procedure, we never require a subforest with no internal edges. This would only be the case if when isolating an internal edge e with endpoints u and v, every other edge adjacent to u and v was a leaf edge. However, that would imply that $n = 4$, which is contradiction. Again, the columns of the matrix $\begin{pmatrix} I \\ -I \end{pmatrix}$ are in the column span and adding any of the original subforest vectors gives us $\operatorname{rank}(M_3(v)) = 2n-2$. □

2.4 Phylogenetic Invariants

The fact that $D(3) = 6$ means that to show the generic identifiability of the tree parameters of the 3-class Jukes-Cantor mixture model for 6-leaf trees it is enough to separate all 6-leaf triplets. While this list of triplets is at least finite, there are still $19,698,048,370$ pairs of 6-leaf triplets, making it infeasible to try to list all pairs and separate them directly. For some pairs $\mathcal{S}, \mathcal{T} \in \mathcal{T}_{[6],3}$, there is a disentangling set K such that $|K| < 6$, so our strategy is to first generate all pairs of 5-leaf triplets and wherever possible show the mutual noncontainment of their corresponding varieties. The results in Section 2.3 imply that we actually only need to show the two varieties are not identical, and to do so we

will need to find a phylogenetic invariant that holds for one mixture and not the other. Once complete, we will have a short list of 5-leaf triplet pairs for which we are unable to show that their varieties are not identical. From this list, we arrive at a much smaller list of pairs of 6-leaf triplets which we need to separate using invariants. For these 6-leaf triplets, we use linear invariants and higher degree invariants to separate all of the pairs.

This final step in the proof of Theorem 2.0.1 is highly computational. The steps are all completely contained in the three Maple [Map] worksheets

`LinearInvariants_5Leaf.mw`, `LinearInvariants_6Leaf.mw`,
`Higher_Degree_Invariants.mw`

which are located at the website:

`http://www4.ncsu.edu/~smsulli2/Pubs/ThreeTreesWebsite/threetrees.html`

We outline the methods for finding separating linear invariants and higher degree invariants in the next two subsections.

2.4.1 Linear Invariants

Our first step will be to separate all 5-leaf triplet pairs by finding linear invariants that hold for one triplet but not the other. A few observations will help us reduce the dimension of the ambient space of the varieties. For our purposes, it is unnecessary to calculate invariants that hold for all mixtures. For the Jukes-Cantor model, every model will have linear invariants arising from permutations on the set $\{C, G, T\}$. For example, any 5-leaf mixture will have the same parameterization on the coordinates in the set

$$\{q_{CCCGT}, q_{GGGCT}, q_{TTTGC}, q_{CCCTG}, q_{GGGTC}, q_{TTTCG}\}.$$

Therefore, the difference of any two of these elements is an invariant for every 5-leaf 3-class Jukes-Cantor mixture model. We will only consider the lexicographically first element as a representative of each such set. By doing this, and removing coordinates that are always zero, we can perform our calculations in $\mathbb{C}^{51}$ instead of $\mathbb{C}^{1024}$. (In the provided Maple file, we also exclude the coordinate q_{AAAAA}, which is not involved in any linear invariant).

Applying an element of S_5 to the leaf labels of all of the trees in a mixture model merely permutes the coordinates of our parameterization. As a result, once we have

determined that the models corresponding to a 5-leaf triplet pair do not in fact have the same variety, then by applying the elements of S_5 to all six trees in the pair, we generate new 5-leaf triplet pairs that do not have the same variety. To illustrate, although there are 680 different identical 5-leaf triplet pairs ($(\mathcal{S}, \mathcal{S}) \in \mathcal{T}_{[5],3} \times \mathcal{T}_{[5],3}$), all of these pairs can be generated by applying an element of S_5 to one of only 28 such pairs.

The following lemma allows us to compute the linear invariant space of a 3-class mixture from the linear invariant space of the constituent models.

Lemma 2.4.1. *A linear polynomial f is an invariant of $V_{\mathcal{T}_1} * V_{\mathcal{T}_2} * V_{\mathcal{T}_3}$ $\iff$ f is an invariant of $V_{\mathcal{T}_1}, V_{\mathcal{T}_2}$, and $V_{\mathcal{T}_3}$.*

Proof. ($\Rightarrow$) Suppose f is a linear invariant of $V_{\mathcal{T}_1} * V_{\mathcal{T}_2} * V_{\mathcal{T}_3}$. By definition, $f(v) = 0$ for all $v \in V_{\mathcal{T}_1} * V_{\mathcal{T}_2} * V_{\mathcal{T}_3}$. But each $V_{\mathcal{T}_i} \subset V_{\mathcal{T}_1} * V_{\mathcal{T}_2} * V_{\mathcal{T}_3}$, so f must evaluate to zero on each $V_{\mathcal{T}_i}$. Therefore, f is an invariant for $V_{\mathcal{T}_1}, V_{\mathcal{T}_2}$, and $V_{\mathcal{T}_3}$.

($\Leftarrow$) Suppose f is a linear invariant for $V_{\mathcal{T}_1}, V_{\mathcal{T}_2}$, and $V_{\mathcal{T}_3}$. Recall that any element of $v \in V_{\mathcal{T}_1} * V_{\mathcal{T}_2} * V_{\mathcal{T}_3}$ can be written as $\pi_1 v_1 + \pi_2 v_2 + \pi_3 v_3$ for some $(v_1, v_2, v_3, \pi) \in V_{\mathcal{T}_1} \times V_{\mathcal{T}_2} \times V_{\mathcal{T}_3} \times \Delta^2$. Then by linearity, $f(v) = f(\pi_1 v_1 + \pi_2 v_2 + \pi_3 v_3) = \pi_1 f(v_1) + \pi_2 f(v_2) + \pi_3 f(v_3) = 0$ since each $f(v_i) = 0$. □

In the provided code, we compare all 5-leaf triplet pairs, and up to the action of S_5, there are 36 pairs with the same linear invariant space. As mentioned, this list consists of the 28 pairs of identical triplet pairs ($(S, S) \in \mathcal{T}_{[5],3} \times \mathcal{T}_{[5],3}$) as well as 8 additional pairs where the two triplets are distinct. It should be noted that these eight additional pairs are what prevent us from generalizing Theorem 2.0.1 to 5-leaf trees. It may be that these eight pairs are separable, but if that is the case, separating them would require identifying higher degree invariants. Below, we will discuss a method for finding higher degree invariants for 6-leaf trees, but the same method does not yield separating invariants for these 5-leaf triplet pairs.

If there exists $\mathcal{S}, \mathcal{T} \in \mathcal{T}_{[6],3}$ such that $V_{\mathcal{T}_1} * V_{\mathcal{T}_2} * V_{\mathcal{T}_3} = V_{\mathcal{S}_1} * V_{\mathcal{S}_2} * V_{\mathcal{S}_3}$, then it must be the case that for any five element subset $K \subset [n]$, $(\mathcal{S}_{|K}, \mathcal{T}_{|K})$ (or some permutation of the leaves thereof) is one of the 5-leaf triplet pairs in our list. Therefore, the only 6-leaf triplet pairs that are candidates for inseparability are those generated by attaching an additional edge to each of the six trees in an inseparable 5-leaf triplet pair. Since there are 36 pairs, and each 5-leaf tree has 7 edges, the number of 6-leaf triplets we must

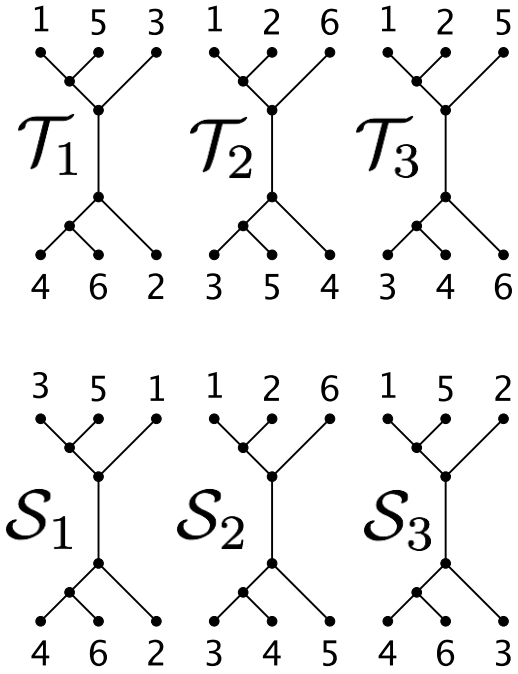


Figure 2.5: A 6-leaf triplet pair that is not separated by linear invariants.

consider is less than $(36)(7^6) = 4,235,364$. This is far fewer than would be expected had we proceeded directly to the 6-leaf case.

Just as in the 5-leaf case, nondefectiveness of all of the involved varieties is ensured by Theorem 2.3.2, so it suffices to show that $V_{\mathcal{T}_1} * V_{\mathcal{T}_2} * V_{\mathcal{T}_3} \not\subseteq V_{\mathcal{S}_1} * V_{\mathcal{S}_2} * V_{\mathcal{S}_3}$. This fact is reflected in our computation of the set `AllSixLeafPairs`, which contains the eighty-five 6-leaf triplet pairs (up to the action of S_6) that are not separated by any linear invariant. For these eighty-five pairs of 6-leaf triplets we must find higher degree invariants that separate them.

2.4.2 Invariants of Higher Degree

If we wish to determine identifiability, we must broaden our search to invariants of higher degree. Unfortunately, we glean little from studying $V_{\mathcal{T}_1}, V_{\mathcal{T}_2}$, and $V_{\mathcal{T}_3}$ individually, as there is no reason to expect that a nonlinear invariant that holds for all three holds for their join.

After removing trivial coordinates and linear invariants that hold for all 6-leaf trees we

can perform our calculations for 6-leaf trees in a 186-dimensional space. When searching for linear invariants, we will also disregard any coordinate in which the character A appears in the index so that the bulk of our computations are done in a 31-dimensional space. If we let $\mathcal{T}$ be the 6-leaf triplet described in Figure 2.5 (in the worksheet, these trees are labeled $\mathcal{T}_{16}, \mathcal{T}_{19}$ and $\mathcal{T}_{63}$), there are 61 linearly independent elements of $I(V_{\mathcal{T}_1} * V_{\mathcal{T}_2} * V_{\mathcal{T}_3})_1$ which define a 125-dimensional linear subspace containing $V_{\mathcal{T}_1} * V_{\mathcal{T}_2} * V_{\mathcal{T}_3}$. Since $\dim(V_{\mathcal{T}_1} * V_{\mathcal{T}_2} * V_{\mathcal{T}_3}) = 29$, at least for this particular triplet, higher degree invariants must exist. In order to find these, we will use the explicit parameterization of the variety in Fourier coordinates.

When writing out the explicit parameterization of a coordinate we dehomogenize by setting all of the $a_A^e = 1$. This allows us to drop the subscripts on the parameters entirely. For arbitrary n-leaf trees, we will further simplify notation by extending the numbering of the trivial splits to a numbering of all of the splits so that every split is labeled by some element of the set $[2n-3]$. Thus, in this more compact notation the coordinate in Example 1.4.3 can be written $q_{CAGGTG} = a_1a_3a_4a_5a_6a_7$.

Example 2.4.2. For the multiset of trees from Figure 2.5, the parameterization of three different coordinates of $V_{\mathcal{T}}$ is listed below. We let $a_C^{i|[6]\setminus\{i\}} = a_i$, $b_C^{i|[6]\setminus\{i\}} = b_i$, $c_C^{i|[6]\setminus\{i\}} = c_i$, $a_C^{15|2346} = a_7$, $a_C^{46|1235} = a_8$, $a_C^{135|246} = a_9$, $b_C^{12|3456} = b_7$, $b_C^{15|2346} = b_8$, $b_C^{126|345} = b_9$, $c_C^{12|3456} = c_7$, $c_C^{34|1256} = c_8$, and $c_C^{125|346} = c_9$.

$$q_{AAAACC} = \pi_1 a_5a_6a_7a_8a_9 + \pi_2 b_5b_6b_8b_9 + \pi_3 c_5c_6c_9$$
$$q_{AAACAC} = \pi_1 a_4a_6 + \pi_2 b_4b_6b_9 + \pi_3 c_4c_6c_8$$
$$q_{AAACCA} = \pi_1 a_4a_5a_7a_8a_9 + \pi_2 b_4b_5b_8 + \pi_3 c_4c_5c_8c_9$$

Theoretically, we should be able use the equations like those in Example 2.4.2 to construct an ideal in

$$\mathbb{C}[y_1, \ldots, y_{186}, a_1, \ldots, a_9, b_1, \ldots, b_9, c_1, \ldots, c_9, \pi_1, \pi_2],$$

and eliminate to obtain a Gröbner basis for $I(V_{\mathcal{T}_1} * V_{\mathcal{T}_2} * V_{\mathcal{T}_3})$ that includes nonlinear invariants. The number of variables and equations involved in the computation make this infeasible.

Instead, we will apply heuristic methods to reduce the number of variables in the ideal before we attempt elimination. Our strategy will be to find linear invariants that

hold for $V_{\mathcal{T}_2}$ and $V_{\mathcal{T}_3}$, but not for $V_{\mathcal{T}_1}$. The resulting equations will not evaluate to zero on all of $V_{\mathcal{T}}$, but will not involve the parameters from $\mathcal{T}_2$ or $\mathcal{T}_3$ at all. To illustrate, we revisit the trees in Figure 2.5, and look at a few more coordinates.

$$\begin{aligned}
q_{CCCAAC} &= \pi_1 a_1 a_2 a_3 a_6 a_7 a_8 + \pi_2 b_1 b_2 b_3 b_6 b_8 b_9 + \pi_3 c_1 c_2 c_3 c_6 c_8 \\
q_{CCGAAG} &= \pi_1 a_1 a_2 a_3 a_6 a_7 a_8 a_9 + \pi_2 b_1 b_2 b_3 b_6 b_8 b_9 + \pi_3 c_1 c_2 c_3 c_6 c_8 \\
q_{CACCGT} &= \pi_1 a_1 a_3 a_4 a_5 a_6 a_7 a_8 a_9 + \pi_2 b_1 b_3 b_4 b_5 b_6 b_7 b_8 b_9 + \pi_3 c_1 c_3 c_3 c_4 c_5 c_6 c_7 c_9 \\
q_{CAGGTG} &= \pi_1 a_1 a_3 a_4 a_5 a_6 a_7 + \pi_2 b_1 b_3 b_4 b_5 b_6 b_7 b_8 b_9 + \pi_3 c_1 c_3 c_3 c_4 c_5 c_6 c_7 c_9,
\end{aligned}$$

Now it is easy to spot linear invariants for $V_{\mathcal{T}_2}$ and $V_{\mathcal{T}_3}$, and subtracting we obtain

$$\begin{aligned}
q_{CCCAAC} - q_{CCGAAG} &= \pi_1 a_1 a_2 a_3 a_6 a_7 a_8 - \pi_1 a_1 a_2 a_3 a_6 a_7 a_8 a_9 \\
q_{CACCGT} - q_{CAGGTG} &= \pi_1 a_1 a_3 a_4 a_5 a_6 a_7 a_8 a_9 - \pi_1 a_1 a_3 a_4 a_5 a_6 a_7.
\end{aligned}$$

In this particular case, $\dim(I(V_{T_2} * V_{T_3})_1) - \dim(I(V_{\mathcal{T}})_1) = 20$, so there are twenty linearly independent relations only involving the parameters from $\mathcal{T}_1$. We introduce new variables for the image space and use these relations to construct the ideal,

$$\begin{aligned}
J = \langle & y_1 - (\pi_1 a_1 a_2 a_3 a_6 a_7 a_8 - \pi_1 a_1 a_2 a_3 a_6 a_7 a_8 a_9), \ldots, \\
& y_{20} - (\pi_1 a_1 a_3 a_4 a_5 a_6 a_7 a_8 a_9 - \pi_1 a_1 a_3 a_4 a_5 a_6 a_7)\rangle,
\end{aligned}$$

where $J \subseteq \mathbb{C}[y_1, \ldots, y_{20}, a_1, \ldots, a_9, \pi_1]$. With fewer parameters, we can now compute elements of the Gröbner basis for $J \cap \mathbb{C}[y_1, \ldots, y_{20}]$ using elimination in Macaulay2 [GS02]. This gives us relations in the y_i variables, which we translate back into our original coordinates. For this particular triplet, we find

$$\begin{aligned}
&(q_{CCCAAC} - q_{CCGAAG})(q_{CACCGT} - q_{CAGGTG}) = \\
&(q_{CCCAGT} - q_{CCGATC})(q_{CACCAC} - q_{CAGGAC}).
\end{aligned}$$

Finally, to separate the triplet pair from Figure 2.5, we substitute the parameterization of $V_{\mathcal{S}}$ into this relation and confirm that it does not evaluate to zero.

This technique allows us to find an invariant for one mixture that does not hold for the

other for all of the triplet pairs contained in `AllSixLeafPairs`. As outlined, the existence of these invariants is sufficient to establish the generic identifiability of the tree parameters of the 3-class Jukes-Cantor mixture model. Among the supplementary materials is the worksheet `Higher_Degree_Invariants.mw` which lists an invariant separating each triplet pair and provides code to quickly generate the coordinate functions for verification.

Chapter 3

The Defining Equations of the Strand Symmetric Model

3.1 Introduction

The strand symmetric model is a phylogenetic model designed to reflect the symmetry inherent in the double-stranded structure of DNA. This symmetry naturally imposes restrictions on the transition probabilities assigned to each edge and imposing only these restrictions gives the general strand symmetric model (SSM). The phylogenetic invariants of a model are algebraic relationships that must be satisfied by the probability distributions arising from the model. As we have seen in Chapter 2, phylogenetic invariants can be useful for proving identifiability results. Results in [DK09] imply the ideal of phylogenetic invariants for the SSM on any binary tree can be determined from the ideal of phylogenetic invariants on the claw tree, $K_{1,3}$.

Though the general strand symmetric model itself is not group-based, Casanellas and Sullivant [CS05] showed that it is still amenable to the Fourier-Hadamard transform technique of [ES93, SESP93]. In the Fourier coordinates, it becomes evident that the parameterization of the model on $K_{1,3}$ is a coordinate projection of the secant variety of the Segre embedding of $\mathbb{P}^3 \times \mathbb{P}^3 \times \mathbb{P}^3$. From this observation, the same authors were able to find 32 degree three and 18 degree four invariants of the homogeneous ideal for $K_{1,3}$ and to show that these invariants generate the ideal up to degree four. Whether or not these polynomials generate the entire ideal was heretofore unknown.

In this chapter, we show that these 50 equations in fact generate the ideal of the SSM

on $K_{1,3}$. The method that we use is the prime-dimension approach introduced in Section 1.4.2. First, we apply the tropical secant dimension technique to the parameterization of the model after the matrix-valued Fourier transform to determine the dimension of the variety of probability distributions associated to the model. Then, using Macaulay2 [GS02], we show that the ideal generated by these fifty equations defines a variety of the same dimension. Finally, with the aid of symbolic computation we generate a decreasing sequence of elimination ideals demonstrating that the ideal in question is prime. Thus, the variety defined by these equations is irreducible, contains the variety of the model, and is of the same dimension as the variety of the model, from which the result follows.

3.2 Phylogenetic Invariants of the SSM

3.2.1 Preliminaries

The general strand symmetric model on an n-leaf rooted tree $\mathcal{T}$ is a phylogenetic model of 4-state character change. Since the SSM is specifically intended to model DNA evolution, we associate to each node v of $\mathcal{T}$ a random variable X_v with state space corresponding to the DNA bases {A,C,G,T}. Associated to each edge is a 4×4 transition matrix with rows and columns indexed by the bases. The entry θ_{ij} encodes the probability of changing from character i to j along that edge. In the double helix structure of DNA it is always the case that the bases A and T are paired together and likewise for C and G. So that our model reflects this strand symmetry, we let $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$ be the distribution of the bases at the root, and set $\pi_A = \pi_T$ and $\pi_C = \pi_G$. Additionally, since a character transition in one strand will induce a corresponding transition in the other, we insist

$$\theta_{AA} = \theta_{TT}, \theta_{AC} = \theta_{TG}, \theta_{AT} = \theta_{TA}, \theta_{CA} = \theta_{GT}, \theta_{CC} = \theta_{GG}, \theta_{CG} = \theta_{GC}, \theta_{CT} = \theta_{GA}.$$

The key observation from [CS05] is that the SSM is a matrix-valued group-based model. Identify the character states of the random variables of a phylogenetic model with elements of $G\times\{0,\dots,l\}$ where G is a finite abelian group and $l \in \mathbb{N}$. Then each character state is indexed by an element $\binom{j}{i}$ where $j \in G$ and $i \in \{0,\dots,l\}$. In these indices, the entries of the transition matrix along edge E are written $E^{j_1 j_2}_{i_1 i_2}$ and the probability that the root is in state $\binom{j}{i}$ is equal to R^j_i.

Definition 3.2.1. A phylogenetic model is a *matrix-valued group-based model* if for each

edge, the matrix transition probabilities satisfy

$$E^{j_1 j_2}_{i_1 i_2} = E^{k_1 k_2}_{i_1 i_2}$$

whenever $j_1 - j_2 = k_1 - k_2$ and the root distribution probabilities satisfy $R^j_i = R^k_i$.

Let $G = \mathbb{Z}_2$ and $l = 1$, then the following identifications make manifest the matrix-valued group-based structure of the SSM: $A = \left(\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right)$, $G = \left(\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right)$, $T = \left(\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right)$, $C = \left(\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right)$.

The tree parameter of an algebraic model determines a polynomial map sending each choice of numerical parameters into the probability space indexed by n-tuples of the characters. For the SSM on a tree $\mathcal{T}$, if we let $\Theta_{\mathcal{T}}$ be the space of numerical parameters we have the following map,

$$\phi_{\mathcal{T}} : \Theta_{\mathcal{T}} \to \Delta^{4^n - 1}.$$

If we do not impose the stochastic conditions on the parameters then $\overline{\mathrm{Im}(\phi_{\mathcal{T}})}$, where the closure is taken in the Zariski topology, is a complex algebraic variety. In Section 16.1 of [CS05], the authors detail the group-valued Fourier transform and show how it can be used to obtain a simple parameterization for the closure of the cone over the SSM on $\mathcal{T} = K_{1,3}$, denoted $CV(\mathcal{T})$. Letting q^{mno}_{ijk} be the transformed coordinates of the image space, we have

$$\psi : q^{mno}_{ijk} = d^{mm}_{0i} e^{nn}_{0j} f^{oo}_{0k} + d^{mm}_{1i} e^{nn}_{1j} f^{oo}_{1k}$$

if $m+n+o \equiv 0$ in $\mathbb{Z}_2$, and $q^{mno}_{ijk} = 0$ otherwise. Now to determine the defining equations for the SSM on $K_{1,3}$, it is enough to determine the defining equations for $\overline{\mathrm{Im}(\psi_{\mathcal{T}})} = CV(\mathcal{T})$. The rest of the paper will be concerned with proving the following theorem.

Theorem 3.2.2. *The vanishing ideal of the strand symmetric model for the graph $K_{1,3}$ is minimally generated by* 32 *cubics and* 18 *quartics. The ideal has dimension* 20*, degree* 9024*, and Hilbert series*

$$\frac{1 + 12t + 78t^2 + 332t^3 + 984t^4 + 1908t^5 + 2394t^6 + 1908t^7 + 984t^8 + 332t^9 + 78t^{10} + 12t^{11} + t^{12}}{(1-t)^{20}}.$$

Note that the Hilbert series suggests that the ideal is Gorenstein though we have not been able to prove this.

3.2.2 Dimension

A toric variety is a variety that is parametrized by monomials. Let $C \subset CV(\mathcal{T})$ be the toric variety parameterized in each coordinate only by the monomial containing variables with zero in the first entry of the subscript. Thus, $CV(\mathcal{T}) = C^{\{2\}}$, the second secant variety of C, and we can compute its dimension using the techniques of Section 1.4.3.

To apply the tropical secant dimension approach we will actually use Theorem 15 from [APRS11] which is a reformulated version of Lemma 1.4.13 more convenient for our purposes in this chapter. We associate to each monomial $x_1^{u_1}x_2^{u_2}\dots x_n^{u_n}$ in the parameterization of a toric variety an integer vector u and let A be the set of these integer vectors. Let $H = \{x \in \mathbb{R}^d : c^T x = e\}$ be a hyperplane in $\mathbb{R}^d$ that splits $\mathbb{R}^d$ into two components which we will label $H^+ = \{x \in \mathbb{R}^d : c^T x > e\}$ and $H^- = \{x \in \mathbb{R}^d : c^T x < e\}$.

In our case, the matrix A is a 12×32 matrix of rank 10, with each column containing exactly threes 1's and nine 0's. If we let $\{e_0^0, e_1^0, e_0^1, e_1^1\}$ denote the standard basis in $\mathbb{R}^{2\times 2}$ then the thirty-two columns of A are

$$\{e_i^m \oplus e_j^n \oplus e_k^o \in \mathbb{R}^{12} : m + n + o \equiv 0 \text{ in } \mathbb{Z}_2\}.$$

For example, the column of A corresponding to the coordinate q_{101}^{110} is given by

$$\begin{pmatrix}0\\0\\0\\1\end{pmatrix} \oplus \begin{pmatrix}0\\0\\1\\0\end{pmatrix} \oplus \begin{pmatrix}0\\1\\0\\0\end{pmatrix}$$

which we write as $(0,0,0,1,0,0,1,0,0,1,0,0)^T$.

Theorem 3.2.3. *[APRS11, Theorem 15] Let V_A be a projective toric variety with corresponding set of exponent vectors $A \subset \mathbb{N}^d$. Let H be a hyperplane not intersecting A. Let $A^+ = A \cap H^+$ and $A^- = A \cap H^-$. Then* $\dim(V_A^{\{2\}}) \geq rank(A^+) + rank(A^-) - 1$.

Let I_F be the ideal generated by the fifty equations found in [CS05].

Lemma 3.2.4. $\dim(CV(\mathcal{T})) = \dim(V(I_F)) = 20$.

Proof. Regard C as a projective variety so that $C = V_A$ from Theorem 3.2.3. The hyperplane defined by the vector $c = (1,0,1,0,1,0,1,0,1,0,1,0)$ and $e = \frac{3}{2}$ gives $|A^+| = |A^-| = 16$

and $\operatorname{rank}(A^+) = \operatorname{rank}(A^-) = 10$. Therefore, by Theorem 3.2.3, as a projective variety $\dim(C^{\{2\}}) \geq 19$ and as an affine cone $\dim(CV(\mathcal{T})) \geq 20$. Using Macaulay2 we determine that $\dim(V(I_F)) = 20$, and since $CV(\mathcal{T}) \subseteq V(I_F)$, we must have $\dim(CV(\mathcal{T})) = 20$. □

3.2.3 Primality

In this section we outline our computations for proving that the ideal I_F is prime. In Section 1.4.2, we introduced Lemma 1.4.11 which allows us to prove that an ideal is prime by showing that a certain elimination ideal is prime. In order to apply the lemma, we will need a way to show that an equation is not a zero divisor modulo an ideal. We will do this in Macaulay2 by computing ideal quotients.

Definition 3.2.5. Let I and J be ideals in a commutative ring R, the *ideal quotient of I by J* is

$$(I : J) := \{r \in R \mid rJ \subset I\}.$$

Lemma 3.2.6. *Let $R = \mathbb{K}[\mathbf{x}], J \subset R$ an ideal, and $g \in R$. Then g is not a zero divisor of $R/J \iff (J : \langle g \rangle) = J$.*

Proof. ($\Rightarrow$) Suppose g is not a zero divisor of R/J and let $f \in (J : \langle g \rangle)$. By the definition of the ideal quotient, $f\langle g \rangle \subset J \Rightarrow fg = 0$ in R/J. Since g is not a zero divisor, $f = 0$ in R/J and so $f \in J$. Therefore, $(J : \langle g \rangle) \subset J$, and it is obvious that $J \subset (J : \langle g \rangle)$.
($\Leftarrow$) Suppose $J = (J : \langle g \rangle)$ and $fg = 0$ in R/J. This implies that $f\langle g \rangle \subset J$ and so $f \in (J : \langle g \rangle) = J$. Therefore, g is not a zero divisor of R/J. □

Lemma 3.2.7. *The ideal I_F generated by the 32 cubics and 18 quartics of the general strand symmetric model on $K_{1,3}$ is prime.*

Proof. The proof is obtained by repeated application of Lemma 1.4.11. The computations we describe can be found at

`http://www4.ncsu.edu/~smsulli2/Pubs/LooseStrandsWebsite/SSM_-Supplement.html`

in the Macaulay2 file `SSM_Supplement` where the symbols 0,1,2, and 3 are substituted for $\binom{1}{1}$, $\binom{1}{0}$, $\binom{0}{1}$, and $\binom{0}{0}$.

First, we let $I_0 = I_F$. Beginning with $k = 1$, we find a polynomial $f_k = g_k x_k + h_k \in I_{k-1}$, verify that g_k is not a zero divisor modulo I_{k-1}, and then eliminate x_k to obtain the ideal I_k. In this way we generate a decreasing chain of elimination ideals

$$I_F = I_0 \supset I_1 \supset I_2 \ldots \supset I_{10}.$$

Using the `isPrime` function in Macaulay2, we show that I_{10}, and hence every ideal in the sequence, is prime.

□

While this is the general outline of our approach, it is actually computationally easier to show that none of the g_k that we encounter is a zero divisor modulo the respective elimination ideal first. Identify the new indices $0, 1, 2$, and 3 with the set of standard basis vectors $\{e_1, e_2, e_3, e_4\}$ and define a multigrading where the weight of q_{ijk} is $e_{i+1} \oplus e_{j+1} \oplus e_{k+1}$. Let $q_\alpha q_\beta - q_\gamma q_\delta$ be a nontrivial binomial that is homogeneous with respect to this grading. For this particular sequence of ideals we are always able to choose $f_k = g_k x_k + h_k$ so that g_k is either such a binomial or a product of such binomials. There are two elementary observations that will be useful:

(i) $g = l_1 l_2$ is a zero divisor modulo J if and only if at least one of l_1 and l_2 is.

(ii) g is not a zero divisor modulo any elimination ideal of J if it is not a zero divisor modulo J.

Thus, to show that none of the g_k is a zero divisor modulo I_{k-1} it is enough to show that none of the homogeneous binomials is a zero divisor modulo I_F. As noted above, we do this by verifying that $(I_F : g_k) = I_F$.

The symmetry of I_F enables us to establish this by considering only a small subset of the homogeneous binomials. There is a group action of $S_4 \times S_4 \times S_4 \rtimes S_3$ on $(Seg(\mathbb{P}^3 \times \mathbb{P}^3 \times \mathbb{P}^3))^{\{2\}}$, that comes from performing the rank-preserving column and transposition operations. Hence, the same group acts on $I((Seg(\mathbb{P}^3 \times \mathbb{P}^3 \times \mathbb{P}^3))^{\{2\}})$, where column operations correspond to changing the indices of the variables and transposition operations correspond to permuting the indices of each variable. Let G be the subgroup of elements of $S_4 \times S_4 \times S_4 \rtimes S_3$ satisfying $g \cdot q_{ijk}^{mno} = q_{i'j'k'}^{m'n'o'}$ with $m + n + o \equiv m' + n' + o'$ in $\mathbb{Z}_2$ for each of the 64 variables. Since

$$I(CV(\mathcal{T})) = I((Seg(\mathbb{P}^3 \times \mathbb{P}^3 \times \mathbb{P}^3))^{\{2\}}) \cap \mathbb{C}[q_{ijk}^{mno} : m + n + o = 0],$$

G acts on $I(CV(\mathcal{T}))$, and since the generators of I_F generate $I(CV(\mathcal{T}))$ up to degree four, G acts on I_F as well. Let H be the subgroup of G generated by elements that correspond to changing the indices. For example, $h = ((01), (01)(23), (01)) \in H$ interchanges $0 \leftrightarrow 1$ in the first index, $0 \leftrightarrow 1$ and $2 \leftrightarrow 3$ in the second, and $0 \leftrightarrow 1$ in the third so that $h \cdot (q_{021}q_{113} - q_{013}q_{121}) = (q_{130}q_{003} - q_{103}q_{030})$. Then

$$H = \langle \quad ((01), id, id), (id, (01), id), (id, id, (01)), ((23), id, id), (id, (23), id), \\ (id, id, (23)), ((0213), (0213), id), ((0213), id, (0213))\rangle$$

is a 256-element normal subgroup and $G \cong H \rtimes S_3$. One can check that the set of homogeneous binomials partitions into three orbits under the action of G with representatives $q_{002}q_{013} - q_{003}q_{012}$, $q_{002}q_{113} - q_{003}q_{112}$, and $q_{002}q_{120} - q_{020}q_{102}$. In the file `SSM_Supplement` we show that none of the homogeneous binomials is a zero divisor by showing that none of these three binomials is a zero divisor modulo I_F.

Having shown that I_F is prime, we are able to give a short proof of Theorem 3.2.2.

Proof of Theorem 3.2.2. The containment $I_F \subset I(CV(\mathcal{T}))$ implies that $CV(\mathcal{T}) \subset V(I)$. By Lemma 3.2.7, I_F is prime and so $V(I_F)$ is an irreducible variety. By Lemma 3.2.4, $CV(\mathcal{T})$ is an irreducible variety contained in an irreducible variety of the same dimension, so $CV(\mathcal{T}) = V(I_F)$ and $I_F = I(CV(\mathcal{T}))$. Knowing explicit generators of the vanishing ideal of the strand symmetric model on the graph $K_{1,3}$, the claims about the rank, degree, and Hilbert series of the ideal are easily verified by the Macaulay2 code in `SSM_Supplement`. □

Chapter 4

Initial Ideals of Phylogenetic Secant Ideals

4.1 Introduction

In Chapter 2, we saw that the ideal of phylogenetic invariants for a mixture model on trees of the same topology is a secant ideal. Secant ideals also appear in the study of the general k-state Markov model on the claw tree $K_{1,3}$. The variety of the k-state model is the k-secant variety of a Segre product, specifically, $(Seg(\mathbb{P}^{k-1} \times \mathbb{P}^{k-1} \times \mathbb{P}^{k-1}))^{\{k\}}$ [GSS05]. As with the strand symmetric model, determining the ideal of phylogenetic invariants for this tree allows one to determine the ideal of phylogenetic invariants for the general Markov model on any binary tree [AR08b]. The set of distributions associated to other statistical models outside of phylogenetics exhibit the structure of secant varieties as well (see e.g., [DSS07] and [Sul08]).

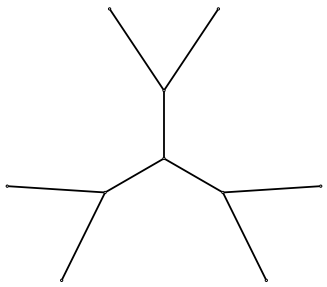
One approach to studying ideals is via initial ideals. We observed at the end of Section 1.3 that an ideal and its initial ideals have the same Hilbert series, and consequently, share many of the same properties. In [SS06], the authors study the relationship between the secant of an initial ideal and the initial ideal of a secant ideal. In particular, they explore under what conditions these operations commute. In this chapter, we investigate the relationship between secant ideals of initial ideals and initial ideals of secant ideals for two classes of ideals connected to binary trees.

The binary Jukes-Cantor model or Cavender-Farris-Neyman (CFN) model is a two-state group-based phylogenetic model. As a group-based model, it is subject to the Fourier

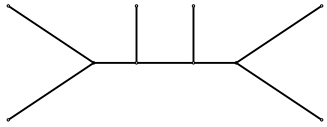
transformation, and we will denote by $\mathcal{I}_{\mathcal{T}}$ the ideal of phylogenetic invariants for the CFN model on the tree $\mathcal{T}$ in the Fourier coordinates. Following the convention for secant ideals, we denote the ideal of the 2-class CFN mixture model on $\mathcal{T}$ by $\mathcal{I}_{\mathcal{T}} * \mathcal{I}_{\mathcal{T}}$ and call this the *CFN secant ideal for* $\mathcal{T}$.

In Section 4.2, we compute the CFN secant ideals for both 6-leaf tree topologies. More precisely, we choose a particular labeling of the leaves for each 6-leaf tree topology and compute the ideals for the labeled trees. Of course, if $\mathcal{T}$ and $\mathcal{T}'$ are binary phylogenetic $[n]$-trees with the same underlying topology, then $\mathcal{I}_{\mathcal{T}} * \mathcal{I}_{\mathcal{T}}$ and $\mathcal{I}_{\mathcal{T}'} * \mathcal{I}_{\mathcal{T}'}$ are isomorphic.

First, we show that in the Fourier coordinates the CFN secant ideal for the 6-leaf snowflake tree (Figure 4.1a), $\mathcal{I}_{\mathcal{S}} * \mathcal{I}_{\mathcal{S}}$, is isomorphic to that of the known ideal of the strand symmetric model on the claw tree. Next, we determine the CFN secant ideal for the 6-leaf caterpillar tree (Figure 4.1b), $\mathcal{I}_{\mathcal{C}} * \mathcal{I}_{\mathcal{C}}$, using the prime-dimension approach. Our computations reveal that $HS(\mathcal{I}_{\mathcal{C}} * \mathcal{I}_{\mathcal{C}}, t) = HS(\mathcal{I}_{\mathcal{S}} * \mathcal{I}_{\mathcal{S}}, t)$.



(a) Snowflake



(b) Caterpillar

Figure 4.1: The two 6-leaf binary tree topologies.

It is known that for any two n-leaf binary phylogenetic X-trees, the ideals of the CFN model have the same Hilbert series [BW07]. Moreover, there exists a single ideal, $\mathcal{I}^n$, of which the ideal associated to the CFN model of any n-leaf binary phylogenetic X-tree can be realized as an initial ideal [SX10]. In light of these results, we conjecture the following.

Conjecture 4.1.1. *Let $\mathcal{T}$ be an n-leaf binary phylogenetic X-tree and let ω be a weight vector such that $in_\omega(\mathcal{I}^n) = \mathcal{I}_{\mathcal{T}}$. Then $in_\omega(\mathcal{I}^n * \mathcal{I}^n) = \mathcal{I}_{\mathcal{T}} * \mathcal{I}_{\mathcal{T}}$.*

In other words, we conjecture that for $\mathcal{I}^n$ and certain weight vectors, the secant of

the initial ideal is equal to the initial ideal of the secant ideal. If true, this would account for our observation that $HS(\mathcal{I}_{\mathcal{C}} * \mathcal{I}_{\mathcal{C}}, t) = HS(\mathcal{I}_{\mathcal{S}} * \mathcal{I}_{\mathcal{S}}, t)$. This is perhaps somewhat unexpected as it was shown in [SS06] that the operations of taking initial ideals and secant ideals do not in general commute even when the initial ideals are monomial.

In Section 4.3, we study a second class of ideals in bijection with the set of binary phylogenetic X-trees which we call the *Plücker tree ideals*. They are so named because they can be constructed as initial ideals of the Plücker ideal, $I_{2,n}$. The Plücker ideal is the vanishing ideal of the Grassmannian, $Gr(2, \mathbb{C}^n)$, in the Plücker coordinates. The secant ideals of the Plücker tree ideals are then initial ideals of the well-known Pfaffian ideals. We let $J_{\mathcal{T}}$ denote the Plücker tree ideal associated to $\mathcal{T}$. These ideals are discussed in [SS04] where the following theorem is proven.

Theorem 4.1.2. *[SS04] Let $\mathcal{T}$ be an n-leaf binary phylogenetic X-tree. There exists a weight vector $\omega \in \mathbb{R}^n$ and a sign vector $\tau \in \{\pm 1\}^{\binom{n}{2}}$ such that $J_{\mathcal{T}} = \tau \cdot in_\omega(I_{2,n})$, where the sign vector multiplies coordinate p_{ij} by τ_{ij}.*

They also appear in [Sul08] which discusses how these ideals and their secants are connected to Gaussian graphical models and concludes with the following conjecture.

Conjecture 4.1.3. *[Sul08, Conjecture 7.10] Let $\mathcal{T}$ be an n-leaf binary phylogenetic X-tree, $\omega \in \mathbb{R}^n$ a weight vector, and $\tau \in \{\pm 1\}^{\binom{n}{2}}$ a sign vector such that $J_{\mathcal{T}} = \tau \cdot in_\omega(I_{2,n})$, then $\tau \cdot in_\omega(I_{2,n}^{\{r\}}) = J_{\mathcal{T}}^{\{r\}}$.*

We show that this conjecture is not true for any r. In the case where $r = 2$, we also prove the following theorem giving necessary and sufficient conditions on the topology of $\mathcal{T}$ for the theorem to hold. In the course of doing so, we also furnish a new class of prime initial ideals of the Pfaffian ideals.

Theorem 4.1.4. *Let $\mathcal{T}$ be an n-leaf binary phylogenetic X-tree with $n \geq 4$ and $\omega \in \mathbb{R}^{\binom{n}{2}}$ be a weight vector such that $in_\omega(I_{2,n}) = J_{\mathcal{T}}$, then $in_\omega(I_{2,n}^{\{2\}}) = J_{\mathcal{T}}^{\{2\}}$ if and only if $\mathcal{T}$ has fewer than five cherries.*

The similarity of Conjectures 4.1.3 and 4.1.1 is obvious. There is also a close relationship between the ideals involved as $J_{\mathcal{T}}$ can be viewed as the intersection of $\mathcal{I}_{\mathcal{T}}$ with a coordinate subring. Thus, our hope is to better understand the structure of CFN secant ideals by studying secants of the Plücker tree ideals. Moreover, we hope to gain some insight into the general structure of initial ideals of secant ideals by studying initial ideals of the Pfaffian ideals.

4.2 Second Secants of the CFN Model

Recall that the CFN model is a two-state group-based phylogenetic model where the two states are identified with $\mathbb{Z}_2$. We dissected the various components of the CFN model on a 3-leaf tree earlier in Example 1.1.7.

Let $\mathcal{T}$ be an n-leaf binary phylogenetic X-tree. After the Fourier transformation, $\mathcal{I}_\mathcal{T} \subseteq \mathbb{C}[q_g : g \in (\mathbb{Z}_2)^n]$. Recall that we actually construct $\mathcal{I}_\mathcal{T}$ as $\mathcal{I}_\mathcal{T} = \mathcal{I}(V_\mathcal{T})$, where $V_\mathcal{T}$ is the variety parameterized as follows.

$$q_{g_1,\ldots,g_n} = \begin{cases} \displaystyle\prod_{B|B' \in \Sigma(T)} a^{B|B'}_{\sum_{i\in B} g_i} & \text{if } \displaystyle\sum_{i=1}^{n} g_i = 0 \\ 0 & \text{otherwise} \end{cases}$$

Thus, we can consider $\mathcal{I}_\mathcal{T} \subseteq \mathbb{C}[q] := \mathbb{C}[q_g : g \in (\mathbb{Z}_2)^n, \sum_{i=1}^n g_i = 0]$, a ring of 2^{n-1} variables. Each variable corresponds to a set of disjoint paths in $\mathcal{T}$ involving an even number of leaves. Notice also that the paths in $\mathcal{T}$ are in bijection with coordinates with exactly two nonzero indices. We will now show that $\mathcal{I}_\mathcal{S} * \mathcal{I}_\mathcal{S}$, the secant ideal for the CFN model on the 6-leaf snowflake tree, is an ideal with which we are already familiar. If we let $\mathcal{I}_{SSM}$ denote the ideal of the SSM on the claw tree in the Fourier coordinates then we have the following theorem.

Theorem 4.2.1. $\mathcal{I}_\mathcal{S} * \mathcal{I}_\mathcal{S} \cong \mathcal{I}_{SSM}$.

Proof. In Chapter 3, we encountered the cone C such that $\mathcal{I}_{SSM} = \mathcal{I}(C * C) = \mathcal{I}(C) * \mathcal{I}(C)$. We will now show $\mathcal{I}_\mathcal{S} \cong \mathcal{I}(C)$ by relabeling parameters and coordinates, from which the theorem follows.

Let $Z_1 = \mathbb{C}[q_g : g \in (\mathbb{Z}_2)^6 : \sum_{i=1}^6 g_i \equiv 0]$ and $Z_2 = \mathbb{C}[q^{g_1g_3g_5}_{g_2g_4g_6} : g \in (\mathbb{Z}_2)^6 : g_1 + g_3 + g_5 \equiv 0]$, so that $\mathcal{I}_\mathcal{S} \subseteq Z_1$ and $\mathcal{I}(C) \subseteq Z_2$. Technically, the variables of Z_2 are of the form q^{mno}_{ijk}, where only the top indices are group elements. However, for the purposes of constructing a map, we now treat the bottom indices as belonging to $\mathbb{Z}_2$ as well. Then the $\mathbb{C}$-algebra homomorphism $\sigma : Z_1 \to Z_2$ defined by

$$q_{g_1g_2g_3g_4g_5g_6} \mapsto q^{(g_1+g_2)(g_3+g_4)(g_5+g_6)}_{g_2g_4g_6}$$

with inverse map

$$q^{g_1g_3g_5}_{g_2g_4g_6} \mapsto q_{(g_1-g_2)g_2(g_3-g_4)g_4(g_5-g_6)g_6}$$

is an isomorphism.

Now, label the cherries of the snowflake tree by $\{1,2\},\{3,4\}$, and $\{5,6\}$. Label the three interior edges of the snowflake connected to these cherries e_7, e_8, and e_9 respectively. Let

$$P_1 = \mathbb{C}[a_0^1, a_1^1, a_0^2, a_1^2, a_0^3, a_1^3, a_0^4, a_1^4, a_0^5, a_1^5, a_0^6, a_1^6, a_0^7, a_1^7, a_0^8, a_1^8, a_0^9, a_1^9].$$

and

$$P_2 = \mathbb{C}[d_0^0, d_1^0, d_0^1, d_1^1, e_0^0, e_1^0, e_0^1, e_1^1, f_0^0, f_1^0, f_0^1, f_1^1].$$

Let $\phi_1 : Z_1 \to P_1$ and $\phi_2 : Z_2 \to P_2$ be the parameterization maps such that $\mathcal{I}_{\mathcal{S}} = \ker(\phi_1)$ and $\mathcal{I}(C) = \ker(\phi_2)$. Note that

$$\phi_1(q_{g_1g_2g_3g_4g_5g_6}) = (a_{g_1}^1 a_{g_2}^2 a_{g_1+g_2}^7)(a_{g_3}^3 a_{g_4}^4 a_{g_3+g_4}^8)(a_{g_5}^5 a_{g_6}^6 a_{g_5+g_6}^9).$$

Therefore, we can construct the 12-dimensional parameter space

$$P_1' = \mathbb{C}[a_{h_1}^1 a_{h_2}^2 a_{h_1+h_2}^7, a_{h_1}^3 a_{h_2}^4 a_{h_1+h_2}^8, a_{h_1}^5 a_{h_2}^6 a_{h_1+h_2}^9 : (h_1,h_2) \in \mathbb{Z}_2^2],$$

and regard $\mathcal{I}_C$ as the kernel of a map $\phi_1' : Z_1 \to P_1'$.

Then the $\mathbb{C}$-algebra homomorphism $\psi : P_1' \to P_2$ defined by

$$\begin{aligned}(a_{g_1}^1 a_{g_2}^2 a_{g_1+g_2}^7) &\mapsto d_{g_2}^{g_1+g_2}\\ (a_{g_3}^3 a_{g_4}^4 a_{g_3+g_4}^8) &\mapsto e_{g_4}^{g_3+g_4}\\ (a_{g_5}^5 a_{g_6}^6 a_{g_5+g_6}^9) &\mapsto f_{g_6}^{g_5+g_6}.\end{aligned}$$

with inverse map

$$\begin{aligned}d_{h_2}^{h_1} &\mapsto (a_{h_1-h_2}^1 a_{h_2}^2 a_{h_1}^7)\\ e_{h_2}^{h_1} &\mapsto (a_{h_1-h_2}^2 a_{h_2}^3 a_{h_1}^8)\\ f_{h_2}^{h_1} &\mapsto (a_{h_1-h_2}^3 a_{h_2}^4 a_{h_1}^9)\end{aligned}$$

is an isomorphism.

Finally, since σ is an isomorphism, we just need to show that σ maps $\ker(\phi_1') = \mathcal{I}_{\mathcal{S}}$ onto $\ker(\phi_2) = \mathcal{I}(C)$. One can verify that the following diagram commutes.

$$\begin{array}{ccc} Z_1 & \overset{\sigma}{=\!=} & Z_2 \\ \downarrow{\scriptstyle \phi_1'} & & \downarrow{\scriptstyle \phi_2} \\ P_1' & \overset{\psi}{=\!=} & P_2 \end{array}$$

By the diagram, we have $\psi^{-1} \circ \phi_2 \circ \sigma = \phi_1'$ and $\psi \circ \phi_1' \circ \sigma^{-1} = \phi_2$. If $z_1 \in \ker(\phi_1')$, $(\psi^{-1} \circ \phi_2 \circ \sigma)(z_1) = 0$, which implies $\phi_2(\sigma(z_1)) = 0$, since ψ^{-1} is an isomorphism. Therefore, $\sigma(\ker(\phi_1')) \subseteq \ker(\phi_2)$. Likewise, if $z_2 \in \ker(\phi_2)$, $\phi_1'(\sigma^{-1}(z_2)) = 0$. Thus, σ maps $\ker(\phi_1')$ onto $\ker(\phi_2)$. □

Next, we compute $\mathcal{I}_\mathcal{C} * \mathcal{I}_\mathcal{C}$, the secant ideal for the CFN model on the 6-leaf caterpillar. We fix a labeling so that the nontrivial splits of the caterpillar tree are $12|3456, 123|456$, and $1234|56$. Since we do not recognize $\mathcal{I}_\mathcal{C} * \mathcal{I}_\mathcal{C}$ as an ideal we have encountered before we must compute it directly. In the Fourier coordinates, the CFN ideals are generated by the 2×2 minors of certain matrices [SX10, Section 7]. Using the known generators for $\mathcal{I}_\mathcal{C}$ we are able to compute $\mathcal{I}_\mathcal{C} * \mathcal{I}_\mathcal{C}$ up to degree 5 in Macaulay2. This computation and all other Macaulay2 and Maple computations referenced in this chapter are contained in the worksheets

`CFN_6leaf_Caterpillar_Secant.m2` and `SecantDimension.mw`

located at the website:

`http://www4.ncsu.edu/~celong2/Thesis%20Supplement.html`.

The generators of $\mathcal{I}_\mathcal{C} * \mathcal{I}_\mathcal{C}$ of degree five or less generate an ideal J_5, and we know that $J_5 \subseteq \mathcal{I}_\mathcal{C} * \mathcal{I}_\mathcal{C}$. However, we would like to show that $J_5 = \mathcal{I}_\mathcal{C} * \mathcal{I}_\mathcal{C}$. This is the same situation we saw in Chapter 3 and again we will apply the prime-dimension approach. Explicitly, to show these ideals are equal we need to show

(i) $\dim(J_5) = \dim(\mathcal{I}_\mathcal{C} * \mathcal{I}_\mathcal{C})$.

(ii) J_5 is prime.

To verify (i), we first find that $\dim(J_5) = 20$ using Macaulay2. Since $J_5 = \mathcal{I}_\mathcal{C} * \mathcal{I}_\mathcal{C}$, we know that $\dim(\mathcal{I}_\mathcal{C} * \mathcal{I}_\mathcal{C}) \leq 20$. Next, we apply the tropical secant dimension approach to get a lower bound on $\dim(\mathcal{I}_\mathcal{C} * \mathcal{I}_\mathcal{C})$. Again, we use the formulation in Theorem 3.2.3 which is convenient when working with the second secant of a toric variety. Using random search, we are able to find a vector as in that theorem which proves $\dim(\mathcal{I}_\mathcal{C} * \mathcal{I}_\mathcal{C}) \geq 20$. This

vector and the computations to verify the lower bound are contained in the supplementary Maple worksheet `SecantDimension.mw`.

To confirm (ii), we use Macaulay2 and repeatedly apply Lemma 1.4.11 just as we did in the proof of Theorem 3.2.2. This gives us a generating set for $\mathcal{I}_\mathcal{C} * \mathcal{I}_\mathcal{C}$ and we can calculate its Hilbert series. All of these computations are contained in the supplementary Macaulay2 file `CFN_6leaf_Caterpillar_Secant.m2` Combining Theorems 3.2.2 and 4.2.1, we know the Hilbert series of $\mathcal{I}_\mathcal{S} * \mathcal{I}_\mathcal{S}$ and we observe the following.

Theorem 4.2.2. $HS(\mathcal{I}_\mathcal{C} * \mathcal{I}_\mathcal{C}, t) = HS(\mathcal{I}_\mathcal{S} * \mathcal{I}_\mathcal{S}, t)$.

4.3 Plücker Tree Ideals

For what follows, it will be useful to have a standard planar embedding our trees. If $\mathcal{T}$ is an n-leaf binary phylogenetic X-tree, then inscribe a regular n-gon on the unit circle in $\mathbb{R}^2$ and choose a planar representation of $\mathcal{T}$ so that the leaves are located at the vertices of the n-gon. Label the leaves of $\mathcal{T}$ in increasing order clockwise around the circle. Recall that the induced 4-leaf subtrees of a tree are called quartets and that a tree is uniquely determined by its quartets [SS03]. With a circular embedding of $\mathcal{T}$ as described, every induced quartet on the leaves $1 \leq i < j < k < l \leq n$ is either $ij|kl$ or $il|jk$. For trees with such a circular embedding the vector τ from Conjecture 4.1.3 is equal to the all ones vector. Thus, for the rest of this chapter we will consider only trees embedded in this manner so that we can ignore the sign vector entirely.

Let $Z^n = \mathbb{C}[p_{ij} : 1 \leq i < j \leq n]$ and

$$I_{2,n} = \langle p_{ij}p_{kl} - p_{ik}p_{jl} + p_{il}p_{jk} : 1 \leq i < j < k < l \leq n \rangle \subseteq Z^n$$

be the the ideal of quadratic Plücker relations. Let $\mathcal{T}$ be a binary phylogenetic $[n]$-tree and assign positive lengths to the edges of $\mathcal{T}$. The choice of edge lengths naturally induces a metric d on the leaves of $\mathcal{T}$ where $d(i,j)$ is the length of the unique path between i and j. Let $\omega \in \mathbb{R}^{\binom{n}{2}}$ be the vector with $\omega_{ij} = d(i,j)$ for $i < j$. Then

$$in_\omega(I_{2,n}) = \langle p_{ik}p_{jl} - p_{il}p_{jk} : ij|kl \text{ is a quartet of } \mathcal{T} \rangle$$

[SS04, Corollary 4.4]. We call $J_\mathcal{T} = in_\omega(I_{2,n})$ the *Plücker tree ideal of* $\mathcal{T}$. Note that any choice of positive edge lengths for $\mathcal{T}$ yields the same initial ideal.

Corollary 4.4 from [SS04] also gives us a way to realize $J_{\mathcal{T}}$ as the kernel of a homomorphism. Let $\mathbb{C}[y] = \mathbb{C}[y_e : e \text{ is an edge of } \mathcal{T}]$ and $\phi_{\mathcal{T}} : Z^n \to \mathbb{C}[y]$ be the homomorphism that sends p_{ij} to the product of all of the parameters y_e corresponding to edges on the unique path from i to j. Then $J_{\mathcal{T}}$ is the toric ideal $\ker(\phi_{\mathcal{T}})$. Notice the close connection between $J_{\mathcal{T}}$ and $\mathcal{I}_{\mathcal{T}}$ alluded to above. If we relabel the coordinates of Z^n by $p_{ij} = p_{e_i \oplus e_j}$, then $J_{\mathcal{T}} = (\mathcal{I}_{\mathcal{T}} \cap Z^n)$.

4.3.1 Pfaffian Initial Ideals

The determinant of a $2r \times 2r$ skew-symmetric matrix is the square of a polynomial called the Pfaffian of the matrix. Let $I_{2,n,r}$ be the ideal generated by the $2r \times 2r$ subpfaffians of a generic $n \times n$ skew-symmetric matrix $P = (p_{ij})$. Each $2r \times 2r$ Pfaffian equation corresponds to a $2r$-element set $K \subseteq [n]$. The terms appearing in each Pfaffian are then in bijection with perfect matchings on the set K. The Pfaffian ideals arise as secants of the Plücker ideal, that is $I_{2,n,r} = I_{2,n}^{\{r-1\}}$, and for the rest of this paper we will use the latter notation. This result as well as background and examples for the Pfaffian ideals can be found in [KL80, PS05]. In this section, we will collect a number of facts about the Pfaffian ideals which will be useful for proving the results that follow.

Definition 4.3.1. Let p be the $2r \times 2r$ Pfaffian equation corresponding to perfect matchings on the set $\{i_1, \ldots, i_{2r}\}$ with $i_1 < \ldots < i_{2r}$. The *crossing monomial* of p is the monomial $p_{i_1,i_{r+1}} p_{i_2,i_{r+2}} \cdots p_{i_r,i_{2r}}$.

Theorem 4.3.2. *[JW07, Theorem 2.1] There exists a term order $\prec_{circ}$ on Z^n that selects the crossing monomial as the lead term of the Pfaffian equations. Furthermore, the $2r \times 2r$ Pfaffians form a Gröbner basis for $I_{2,n}^{\{r-1\}}$ with respect to this term order and*

$$in_{\prec_{circ}}(I_{2,n}^{\{r-1\}}) = \langle p_{i_1,i_{r+1}} p_{i_2,i_{r+2}} \cdots p_{i_r,i_{2r}} : 1 \leq i_1 < i_2 < \ldots < i_{2r} \leq n \rangle.$$

We also have the following corollary.

Corollary 4.3.3. *Let $\mathcal{T}$ be a binary phylogenetic X-tree and ω a term order for Z^n derived from $\mathcal{T}$ as above. Then the initial forms of the $2r \times 2r$ Pfaffians with respect to ω form a Gröbner basis for $in_\omega(I_{2,n}^{\{r-1\}})$ with respect to $\prec_{circ}$ and hence generate $in_\omega(I_{2,n}^{\{r-1\}})$.*

Proof. Since all of our trees are circularly embedded, for $1 \leq i < j < k < l \leq n$, the ω-weight of $p_{ik}p_{jl}$ is greater than or equal to that of both $p_{ij}p_{kl}$ and $p_{il}p_{jk}$. Therefore, if we

let p be the $2r \times 2r$ Pfaffian equation with monomials corresponding to perfect matchings of the set $\{i_1, \ldots, i_{2r}\}$ with $1 \leq i_1 < i_2 < \ldots < i_{2r} \leq n$, then $in_\omega(p)$ contains the term $p_{i_1,i_{r+2}} p_{i_2,i_{r+3}} \cdots p_{i_{r+1},i_{2r}}$. Thus, the term order $\prec_{circ}$ refines the weight vector ω [SS04]. The result follows from [Stu96, Corollary 1.9]. □

One component of our proof of Theorem 4.1.4 will involve showing that for a weight vector constructed from a circularly embedded binary phylogenetic $[n]$-tree, $in_\omega(I_{2,n}^{\{2\}})$ is prime. In fact, we obtain the much stronger result below giving an entire class of prime initial ideals for the Pfaffian ideals.

Theorem 4.3.4. *Let ω be a weight vector constructed from a circular embedding of a binary phylogenetic $[n]$-tree $\mathcal{T}$. Then for all $r, n \in \mathbb{N}$, $in_\omega(I_{2,n}^{\{r\}})$ is a prime ideal.*

We will prove Theorem 4.3.4 using the prime-dimension approach and induction. We restate here the lemma from Section 1.4.2 which formed the basis for the prime-dimension approach.

Lemma 1.4.11. *Let $\mathbb{K}$ be a field and $J \subset \mathbb{K}[x_1, \ldots, x_n]$ be an ideal containing a polynomial $f = gx_1 + h$ with g, h not involving x_1 and g not a zero divisor modulo J. Let $J_1 = J \cap \mathbb{K}[x_2, \ldots, x_n]$ be the elimination ideal. Then J is prime if and only if J_1 is prime.*

In order to apply Lemma 1.4.11 in our proof of Theorem 4.1.4, we will need the following two lemmas.

Lemma 4.3.5. *Let $\mathcal{T}$ be an $(n+1)$-leaf binary phylogenetic X-tree and ω a weight vector constructed from $\mathcal{T}$. If $n \geq 2r + 1$, then for $2r < j \leq n$, there exists a polynomial in $in_\omega(I_{2,n+1}^{\{r\}})$ in which $p_{j,n+1}$ occurs linearly.*

Proof. Proving this lemma requires choosing a particular circular planar embedding of the tree $\mathcal{T}$ which we now describe. If there exists a split $A|B$ in $\mathcal{T}$ such that $\#A = r + 1$, then circularly label the leaves in A clockwise by the labels $\{n+1, 1, 2, \ldots, r\}$ and then complete the circular labeling of $\mathcal{T}$. Now consider the $(2r + 2) \times (2r + 2)$ Pfaffian equation $p \in I_{2,n+1}^{\{r\}}$ that is the sum of monomials corresponding to perfect matchings on the set $\{1, 2, \ldots, 2r, j, n + 1\}$ with $2r < j \leq n$. As in Corollary 4.3.3, the monomial $p_{1,r+2} p_{2,r+3} \cdots p_{r-1,2r} p_{r,j} p_{r+1,n+1}$ appears in $in_\omega(p)$. Since $\mathcal{T}_{|\{r,r+1,j,n+1\}}$ contains the split $r(n + 1)|(r + 1)j$, $\omega(p_{r,j} p_{r+1,n+1}) = \omega(p_{r,r+1} p_{j,n+1})$. Therefore, $in_\omega(p)$ also contains the monomial

$$p_{1,r+2} p_{2,r+3} \cdots p_{r-1,2r} p_{r,r+1} p_{j,n+1}$$

of equal ω weight, and so $p_{j,n+1}$ occurs linearly in $I^{\{r\}}_{2,n+1}$.

If there does not exists a split $A|B$ in $\mathcal{T}$ with $\#A = r+1$, then since $\mathcal{T}$ is binary, there exists a split $A|B$ such that $r+1 < \#A \leq 2r$. Choose such a split with $\#A$ as small as possible.

Consider $\mathcal{T}_{|A}$ as a rooted tree and starting on the side of the root with the greater number of leaves (if one exists), circularly label the leaves by $\{n+1, 1, 2, \ldots, r, r+1, \ldots, \#A-1\}$. Complete the labeling to a circular labeling of $\mathcal{T}$. Choosing either of the edges adjacent to the root in $\mathcal{T}_{|A}$ induces the split $(n+1)123\ldots k|(k+1)\ldots(\#A-1)$ in $\mathcal{T}_{|A}$. Notice also that $k < r$. Otherwise, either the set $\{(n+1), 1, 2, 3, \ldots, k\}$ labels a split of $\mathcal{T}$ with exactly $r+1$ leaves, which we assumed was not true, or it labels a split with between $r+1$ and $2r$ leaves, contradicting that A was chosen so that $\#A$ was as small as possible.

As before, for $2r < j \leq n$, consider the Pfaffian generator that is the sum of monomials corresponding to perfect matchings on the set $\{1, \ldots, 2r, j, n+1\}$. Then $in_\omega(p)$ contains the monomial $\mathbf{m} = p_{1,r+2}p_{2,r+3}\ldots p_{r-1,2r}p_{r,j}p_{r+1,n+1}$. The monomial $p_{k,(k+r+1)}p_{(r+1),(n+1)}$ divides $\mathbf{m}$ and we know that $\omega(p_{k,(r+1)}p_{(k+r+1),(n+1)}) = \omega(p_{k,(k+r+1)}p_{(r+1),(n+1)})$ since removing the edge of $\mathcal{T}_{|A}$ adjacent to the root on the side labeled by leaves $\{(n+1), 1, 2, \ldots, k\}$ induces the quartet $(n+1)k|(r+1)(k+r+1)$. Therefore, we can replace $p_{k,(k+r+1)}p_{(r+1),(n+1)}$ in $\mathbf{m}$ by the equal weight term $p_{k,(r+1)}p_{(k+r+1),(n+1)}$ to produce a monomial $\mathbf{m}'$ of $in_\omega(p)$.

Notice that now $p_{r,j}p_{(k+r+1),(n+1)}|\mathbf{m}'$. Since $k \geq (\#A)/2-1$ and $r \geq (\#A)/2$, it must be that $k+r+1 \geq \#A$. Therefore, the edge that splits $A|B$ in $\mathcal{T}$ also splits $r(n+1)|j(k+r+1)$, since the leaves in A are labeled by $\{n+1, 1, 2, \ldots, r, r+1, \ldots, \#A-1\}$. So we can replace $p_{r,j}p_{(k+r+1),(n+1)}$ in $\mathbf{m}'$ with $p_{r,(k+r+1)}p_{j,(n+1)}$ to produce another monomial of $in_\omega(p)$. Thus, $p_{j,n+1}$ occurs linearly in $in_\omega(p)$. □

Lemma 4.3.6. *If $n \geq 2r+1$ then for $2r < j \leq n$ let $in_\omega(p)$ be the polynomial found in Lemma 4.3.5 in which $p_{j,n+1}$ occurs linearly. Then $in_\omega(p) = g \cdot p_{j,n+1} + h$ with g, h not involving $p_{j,n+1}$ and g not a zero divisor modulo $in_\omega(I^{\{r\}}_{2,n+1})$.*

Proof. We write $in_\omega(p) = g \cdot p_{j,n+1} + h$ and observe that the polynomial g is the sum of monomials corresponding to perfect matchings on the set $\{1, \ldots, 2r\}$ with equal ω-weight. In other words, $g \in in_{\omega'}(I^{\{r\}}_{2,n})$, where ω' is the subvector of ω without coordinates containing $(n+1)$ in the index. So we just need to show that g is not a zero divisor modulo $in_\omega(I^{\{r\}}_{2,n+1})$.

Recall the term order $\prec_{circ}$ from Theorem 4.3.2 with respect to which the Pfaffian

equations form a Gröbner basis for $in_\omega(I^{\{r\}}_{2,n+1})$. Then

$$in_{\prec_{circ}}(g) = p_{1,r+1}p_{2,r+2}\dots p_{r-2,2r-1}p_{r,2r}.$$

Suppose that there exists $g' \notin in_\omega(I^{\{r\}}_{2,n+1})$ such that $gg' \in in_\omega(I^{\{r\}}_{2,n+1})$. Then choose such a g' with standard leading term with respect to the Gröbner basis given by $\prec_{circ}$. Then

$$in_{\prec circ}(gg') = (p_{1,r+1}p_{2,r+2}\dots p_{r-2,2r-1}p_{r,2r})in_{\prec circ}(g'),$$

and $in_{\prec circ}(gg')$ must be in $in_{\prec_{circ}}(I^{\{r\}}_{2,n+1})$. Therefore, $in_{\prec circ}(gg')$ must be divisible by one of the crossing monomials which are the lead terms of the $(2r+2)\times(2r+2)$ Pfaffian equations. But if p_{ij} appears in the crossing monomial of a $(2r+2)\times(2r+2)$ Pfaffian equation, then $j-i \geq r+1$. This implies that $in_{\prec circ}(g)$ is relatively prime to every crossing monomial. Therefore, $in_{\prec circ}(g')$ must be in the leading term ideal of $in_{\prec_{circ}}(I^{\{r\}}_{2,n+1})$ with respect to $\prec_{circ}$, which is a contradiction since we assumed it was standard. □

Proof of Theorem 4.3.4. We will proceed by induction. Fix $r \in \mathbb{N}$. For $n < 2r+1$, $in_\omega(I^{\{r\}}_{2,n+1}) = \langle 0\rangle$ which is prime.

Now suppose $in_\omega(I^{\{r\}}_{2,n}) \subseteq Z^n$ is prime and consider the ideal $I^{\{r\}}_{2,n+1} \subseteq Z^{n+1}$. First, we show that $(in_\omega(I^{\{r\}}_{2,n+1}) \cap Z^n) = in_{\omega'}(I^{\{r\}}_{2,n})$, where again ω' is the subvector of ω that does not include coordinates with $(n+1)$ in the index. Define a grading on Z^{n+1} where $\deg(p_{ij}) = 1$ if $j = (n+1)$ and $\deg(p_{ij}) = 0$ otherwise. Then $Z^{n+1} = \bigoplus_{i=0}^{\infty} Z^{n+1}_i$ and $I^{\{r\}}_{2,n+1}$ is homogeneous with respect to this grading. It is true in general that for a homogeneous ideal I contained in a graded ring $R = \bigoplus_{i=0}^{\infty} R_i$ and a weight vector ω, that $I = \bigoplus_{i=0}^{\infty} I \cap R_i$ and

$$\begin{aligned} in_\omega(I) &= \bigoplus_{i=0}^{\infty} in_\omega(I \cap R_i) \\ &= \bigoplus_{i=0}^{\infty} in_\omega(I) \cap R_i. \end{aligned}$$

In our case, we have $(in_\omega(I^{\{r\}}_{2,n+1}) \cap Z^{n+1}_0) = in_\omega(I^{\{r\}}_{2,n+1} \cap Z^{n+1}_0)$. Since $(I^{\{r\}}_{2,n+1} \cap Z^{n+1}_0) = I^{\{r\}}_{2,n}$ and Z^n is precisely Z^{n+1}_0, the degree zero piece of Z^{n+1}, $(in_\omega(I^{\{r\}}_{2,n+1}) \cap Z^n) = in_{\omega'}(I^{\{r\}}_{2,n})$.

So now assume the statement is true for all integers less than or equal to $n \geq 2r+2$. We note by Lemma 4.3.5 that each $p_{j,n+1}$ appears in some equation of $in_\omega(I^{\{r\}}_{2,n+1})$. Lemma

4.3.6 tells us that the coefficient of $p_{j,n+1}$ is not a zero divisor modulo $in_\omega(I_{2,n+1}^{\{r\}})$, but this also implies that each coefficient is not a zero divisor modulo any elimination ideal of $in_\omega(I_{2,n+1}^{\{r\}})$. So now beginning with $j = n$, we eliminate $p_{j,n+1}$ for $2r < j \leq n$ from $in_\omega(I_{2,n+1}^{\{r\}})$. Importantly, the equation in which $p_{j,n+1}$ occurs linearly found in Lemma 4.3.5 does not contain any variables of the form $p_{k,n+1}$ for $k > j$ and so is still contained in the elimination ideal after we have eliminated all of these variables. Therefore, at each step, we meet the conditions of Lemma 1.4.11, which implies that each successive elimination ideal is prime if and only if $in_\omega(I_{2,n+1}^{\{r\}})$ is prime.

After eliminating, we have the ideal $in_\omega(I_{2,n+1}^{\{r\}}) \cap Z^n[p_{1,n+1}, \ldots, p_{2r,n+1}]$ which we will now show is equal to $in_\omega(I_{2,n+1}^{\{r\}}) \cap Z^n = in_\omega(I_{2,n}^{\{r\}})$. In other words, we will show that after eliminating $\{p_{2r+1,n+1}, \ldots, p_{n,n+1}\}$, there are no equations involving any variable with $n+1$ in the index in the elimination ideal. Then by induction, the proof will be complete.

The dimension of $I_{2,n}^{\{r\}}$, and hence the dimension of all of its initial ideals, is $2rn-2r^2-r$ [KL80]. Since $I_{2,n+1}^{\{r\}}$ is prime, every irreducible component of $in_\omega(I_{2,n+1}^{\{r\}})$ has dimension $2r(n+1) - 2r^2 - r$ [KS95]. The birational projection of Lemma 1.4.11 preserves the dimension of each component, which implies

$$\dim(in_\omega(I_{2,n+1}^{\{r\}}) \cap Z^n[p_{1,n+1}, \ldots, p_{2r,n+1}]) = 2r(n+1) - 2r^2 - r = \dim(in_\omega(I_{2,n}^{\{r\}})) + 2r.$$

Therefore, eliminating the remaining $2r$ variables must decrease the dimension of each component by $2r$, which implies that the variables in $\{p_{1,n+1}, \ldots, p_{2r,n+1}\}$ are free in each component of $in_\omega(I_{2,n+1}^{\{r\}}) \cap Z^n[p_{1,n+1}, \ldots, p_{2r,n+1}]$. We conclude that $in_\omega(I_{2,n+1}^{\{r\}}) \cap Z^n[p_{1,n+1}, \ldots, p_{2r,n+1}] = in_{\omega'}(I_{2,n}^{\{r\}})$. □

4.3.2 Second Secants of the Plücker Tree Ideals

To address Conjecture 4.1.3 we first construct a simple bound on $\dim(J_{\mathcal{T}}^{\{r\}})$.

Lemma 4.3.7. *Let $\mathcal{T}$ be a tree with c cherries, then* $\dim(J_{\mathcal{T}}^{\{r\}}) \leq 2rn - 3r - (r-1)c$.

Proof. The variables corresponding to cherries do not appear in any of the binomials generating $J_{\mathcal{T}}$. Thus, we can write $V(J_{\mathcal{T}}) = V \times \mathbb{C}^c$ and $V(J_{\mathcal{T}})^{\{r\}} = V^{\{r\}} \times \mathbb{C}^c$, since $\mathbb{C}^c$ is a linear space. The expected dimension of $V^{\{r\}}$ is $r\dim(V) + (r-1)$. However, $J_{\mathcal{T}}$ being homogeneous implies that V is a cone and that $\dim(V^{\{r\}}) \leq r\dim(V)$. Since they share the same Hilbert series, the dimension of $J_{\mathcal{T}}$ is equal to that of $I_{2,n}$ which is $2n-3$ [KL80].

Thus, we have

$$\begin{aligned}\dim(V(J_{\mathcal{T}})^{\{r\}}) &\leq r\dim(V)+c\\ &= r(2n-3-c)+c.\\ &= 2rn-3r-(r-1)c\end{aligned}$$

□

Corollary 4.3.8. *Conjecture 4.1.3 does not hold for any r.*

Proof. Every initial ideal of $I_{2,n}^{\{r\}}$ has dimension $2rn-2r^2-r$ [KL80]. Therefore, it is impossible for $J_{\mathcal{T}}^{\{r\}} = I_{2,n}^{\{r\}}$ if

$$\begin{aligned}2rn-3r-(r-1)c &< 2rn-2r^2-r\\ -(r-1)c &< -2r^2+2r\\ c &> 2r.\end{aligned}$$

Thus, for any r, trees with more than $2r$ cherries serve as a counterexample. □

Theorem 4.1.4 claims that when $r=2$, trees with strictly more than 4 cherries are the only obstructions. Before we begin the proof of Theorem 4.1.4 we will discuss the specific structure of $I_{2,n}^{\{2\}}$ and the initial ideals $in_\omega(I_{2,n}^{\{2\}})$. The ideal $I_{2,n}^{\{2\}}$ is the vanishing ideal of the set of $n\times n$ rank four skew-symmetric matrices and is generated by the 6×6 Pfaffian equations. There are $\binom{n}{6}$ of these degree 3 equations each with 15 terms corresponding to the perfect matchings on the 6-element subset of $[n]$ to which the equation corresponds. Theorem 4.3.2 tells us that the initial forms of these equations with respect to ω form a Gröbner basis for $in_\omega(I_{2,n}^{\{2\}})$. Without loss of generality, let p be the 6×6 Pfaffian equation for the set $K=\{1,2,3,4,5,6\}\subseteq[n]$ and let $\mathcal{T}_{|K}$ be the restriction of $\mathcal{T}$ to the leaves of K. Up to relabeling of the leaves, there are only two 6-leaf tree topologies and the structure of $in_\omega(p)$ is completely determined by the topology of $\mathcal{T}_{|K}$.

If $\mathcal{T}$ is the 6-leaf caterpillar tree with nontrivial splits 12|3456, 123|456, and 1234|56, then

$$in_\omega(p) = p_{14}p_{25}p_{36}-p_{14}p_{26}p_{35}-p_{15}p_{24}p_{36}+p_{15}p_{26}p_{34}+p_{16}p_{24}p_{35}-p_{16}p_{25}p_{34}.$$

If $\mathcal{T}$ is the 6-leaf snowflake tree with nontrivial splits 12|3456, 34|1256, and 56|1234, then

$$\begin{aligned} in_\omega(p) =& p_{14}p_{25}p_{36} - p_{14}p_{26}p_{35} - p_{15}p_{24}p_{36} + p_{13}p_{25}p_{46} + \\ & p_{16}p_{24}p_{35} - p_{13}p_{26}p_{45} + p_{15}p_{23}p_{46} - p_{16}p_{23}p_{45}. \end{aligned}$$

Thus, $in_\omega(I_{2,n}^{\{2\}})$ has a Gröbner basis consisting of $\binom{n}{6}$ equations each with either six or eight terms. We call the n-leaf binary phylogenetic X-tree with exactly two cherries the *n-leaf caterpillar*. Although the following theorem for caterpillar trees does not generalize to a proof of Theorem 4.1.4, we include it because it is rather straightforward and establishes one of the base cases for our inductive argument.

Theorem 4.3.9. *Let $\mathcal{C}$ be an n-leaf caterpillar tree and $\omega \in \mathbb{R}^{\binom{n}{2}}$ be a weight vector such that $in_\omega(I_{2,n}) = J_\mathcal{C}$, then $in_\omega(I_{2,n}^{\{2\}}) = J_\mathcal{C}^{\{2\}}$.*

Proof. For a given term order, the initial ideal of a secant ideal is contained inside the secant of the initial ideal [SS06], so we have the inclusion,

$$in_\omega(I_{2,n}^{\{2\}}) \subseteq (in_\omega(I_{2,n}))^{\{2\}} = J_\mathcal{C}^{\{2\}}. \tag{4.1}$$

Let P be the poset on the variables of Z^n given by $p_{ij} \leq p_{kl}$ if $i \leq k$ and $j \leq l$ and $J(P)$ the monomial ideal generated by incomparable pairs $p_{ij}p_{kl}$ in P. There exists a term order ω' for which $in_{\omega'}(J_C) = J(P)$ [MS05, Theorem 14.16]. Taking initial ideals with respect to ω' in (1), we have

$$in_{\omega'}(in_\omega(I_{2,n}^{\{2\}})) \subseteq in_{\omega'}(J_\mathcal{C}^{\{2\}}) \subseteq (in_{\omega'}(J_\mathcal{C}))^{\{2\}} = J(P)^{\{2\}}. \tag{4.2}$$

In fact, there exists $\omega'' = \omega + \epsilon\omega'$ such that $in_{\omega'}(in_\omega(I_{2,n}^{\{2\}})) = in_{\omega''}(I_{2,n}^{\{2\}})$ [Stu96, Proposition 1.13]. It is also shown in [SS06] Example 4.13, that we can choose a term order $\prec$ for which $in_\prec(I_{2,n}^{\{2\}}) = J(P)^{\{2\}}$. This implies

$$HS(J(P)^{\{2\}}, t) = HS(in_\prec(I_{2,n}^{\{2\}}), t) = HS(in_{\omega''}(I_{2,n}^{\{2\}})), t),$$

which gives equality all across (4.2). This further implies that

$$HS(J_C^{\{2\}}, t) = HS(in_{\omega''}(I_{2,n}^{\{2\}}), t) = HS(in_\omega(I_{2,n}^{\{2\}}), t),$$

giving equality in (4.1) and completing the proof. □

As mentioned in the proof of Theorem 4.3.9, given any binary phylogenetic X-tree $\mathcal{T}$ (not necessarily a caterpillar) and a weight vector ω such that $in_\omega(I_{2,n}) = J_\mathcal{T}$, we have the inclusion $in_\omega(I_{2,n}^{\{2\}}) \subseteq J_\mathcal{T}^{\{2\}}$. To complete the proof of Theorem 4.1.4, we will again use the prime-dimension approach to show that this containment of ideals is actually equality for trees with exactly 3 or 4 cherries. Lemma 4.3.4 establishes that $in_\omega(I_{2,n}^{\{2\}})$ is prime, so it will suffice to show that for such trees $\dim(in_\omega(I_{2,n}^{\{2\}})) = \dim(J_\mathcal{T}^{\{2\}})$.

To prove the dimension result we will first obtain a lower bound on the dimension of $J_\mathcal{T}^{\{2\}}$. Just as in Chapters 2 and 3, we will use the tropical secant dimension approach. To show the correspondence, note that the variety $V(J_\mathcal{T})$ is an affine cone. Therefore,

$$V(J_\mathcal{T})^{\{2\}} = V(J_\mathcal{T}) + V(J_\mathcal{T}) = \{v_1 + v_2 : v_1, v_2 \in V(J_\mathcal{T})\}.$$

As described in Section 4.3, $J_\mathcal{T}$ is the Zariski closure of the monomial map $\phi_\mathcal{T} : Z^n \to \mathbb{C}[y]$. Then $\phi_\mathcal{T}(p_{ij})$ is the square-free monomial parametrizing p_{ij}. Let $\alpha_{ij}^\mathcal{T} \in \mathbb{R}^{2n-3}$ be the 0/1 coefficient vector of $\phi_\mathcal{T}(p_{ij})$. Following the setup of Section 1.4.3 and Definition 1.4.12 we have a simplified version of Lemma 1.4.13.

Lemma 4.3.10. *The dimension of $V(J_\mathcal{T}) + V(J_\mathcal{T})$ is at least the maximum, taken over all $v = (v_1, v_2) \in \mathbb{R}^{2n-3} \oplus \mathbb{R}^{2n-3}$, of the sum*

$$\dim_\mathbb{R}\langle D_1^\mathcal{T}(v)\rangle_\mathbb{R} + \dim_\mathbb{R}\langle D_2^\mathcal{T}(v)\rangle_\mathbb{R}.$$

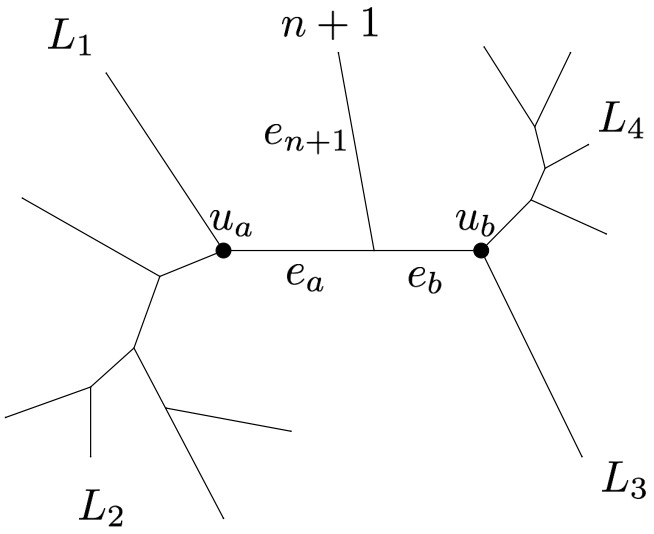
Lemma 4.3.11. *Let $\mathcal{T}$ be an n-leaf binary phylogenetic X-tree with exactly 3 or 4 cherries, then $\dim(J_\mathcal{T}^{\{2\}}) \geq 4n - 10$.*

Proof. We will prove by induction on n that there exists a vector $v = (v_1, v_2) \in \mathbb{R}^{2n-3} \oplus \mathbb{R}^{2n-3}$ such that $\dim_\mathbb{R}\langle D_1^\mathcal{T}(v)\rangle_\mathbb{R} = \dim_\mathbb{R}\langle D_2^\mathcal{T}(v)\rangle_\mathbb{R} = 2n-5$. First, note that every tree with exactly 3 cherries can be constructed by successively attaching leaves to the snowflake tree so that the new leaf is not involved in a cherry. Every tree with exactly 4 cherries can be constructed in the same manner from the unique 8-leaf tree with 4 cherries. By random search, we can find vectors that give us the lower bound for these two trees establishing our base cases. These vectors and the computations to verify the lower bounds can be found in the Maple worksheet `SecantDimension.mw` at the website listed in Section 4.2.

Assume the statement is true for all n-leaf binary phylogenetic X-trees and let $\mathcal{T}$ be an $(n+1)$-leaf binary phylogenetic X-tree with exactly 3 or 4 cherries. Label $\mathcal{T}$ so that the leaf labeled by $(n+1)$ is not part of a cherry. Let $\mathcal{R} = \mathcal{T}_{|[n]}$, by our inductive assumption, there exists $v = (v_1, v_2) \in \mathbb{R}^{2n-3} \oplus \mathbb{R}^{2n-3}$ such that $\dim_{\mathbb{R}}\langle D_1^{\mathcal{R}}(v)\rangle_{\mathbb{R}} = \dim_{\mathbb{R}}\langle D_2^{\mathcal{R}}(v)\rangle_{\mathbb{R}} = 2n-5$. Our goal will be to construct a new vector $w = (w_1, w_2) \in \mathbb{R}^{2(n+1)-3} \oplus \mathbb{R}^{2(n+1)-3}$ so that $\dim_{\mathbb{R}}\langle D_1^{\mathcal{T}}(w)\rangle_{\mathbb{R}} = \dim_{\mathbb{R}}\langle D_2^{\mathcal{T}}(w)\rangle_{\mathbb{R}} = 2(n+1) - 5$.

When adding the $(n+1)$ leaf to $\mathcal{R}$, we introduce "new" edges e_a, e_b, and e_{n+1} and eliminate the edge e. Let u_a be the vertex of e_a not shared with e_b and likewise let u_b be the vertex of e_b not shared with e_a. Arbitrarily choose two leaves L_1 and L_2 such that the path from these leaves to $(n+1)$ passes through e_a. Also choose leaves L_3 and L_4 such that the path from these leaves to $(n+1)$ passes through e_b. Such leaves exist since $(n+1)$ is not contained in a cherry. Figure 4.2 depicts the situation.

Figure 4.2: An example of the labeling scheme described in Lemma 4.3.11.



Delete the entry of v_1 and that of v_2 corresponding to the parameter y_e to form $v_1', v_2' \in \mathbb{R}^{2n-4}$. Define $w_1 = (v_1', w_1^a, w_1^b, w_1^{n+1})$ and $w_2 = (v_2', w_2^a, w_2^b, w_2^{n+1})$ where the entries of w correspond to the edges of $\mathcal{T}$ in the obvious way.

Our goal will be to choose the six new vector entries so that s wins $\alpha_{ij}^{\mathcal{T}}$ at w if and only if s wins $\alpha_{ij}^{\mathcal{R}}$ at v for $1 \leq i < j \leq n$. Moreover, we will want both 1 and 2 to win one

of $\{\alpha^{\mathcal{T}}_{L_1,n+1}, \alpha^{\mathcal{T}}_{L_2,n+1}\}$ and $\{\alpha^{\mathcal{T}}_{L_3,n+1}, \alpha^{\mathcal{T}}_{L_4,n+1}\}$. First, we will see why this will guarantee that $\dim_{\mathbb{R}}\langle D^T_1(w)\rangle_{\mathbb{R}} = \dim_{\mathbb{R}}\langle D^T_2(w)\rangle_{\mathbb{R}} = 2(n+1)-5$.

Form the matrix $A(\mathcal{T})$ with rows equal to *all* the vectors $\alpha^{\mathcal{T}}_{ij}$. Let

$$\omega = \begin{pmatrix} \omega(e_1) \\ \vdots \\ \omega(e_{2(n+1)-3}) \end{pmatrix},$$

be a vector of edge lengths for $\mathcal{T}$. Since $\alpha^{\mathcal{T}}_{ij} \cdot \omega$ gives us the distance between leaves i and j in $\mathcal{T}$, $A(\mathcal{T})\omega$ determines a metric on the leaves of $\mathcal{T}$. By the *Tree-Metric theorem* ([PS04, SS03]) ω is the unique solution to $A(\mathcal{T})x = A(\mathcal{T})\omega$. Therefore, the rank of $A(\mathcal{T})$ is $2n-3$. Thus, if we can uniquely recover all of the edge lengths assigned to $\mathcal{T}$ from a matrix, the matrix has rank at least $2(n+1)-3$.

Let ω' be a vector of edge lengths for $\mathcal{R}$ where the lengths of edges shared between $\mathcal{R}$ and $\mathcal{T}$ are the same and $\omega'(e) = \omega(e_a) + \omega(e_b)$. Form the matrix $M^{\mathcal{R}}_s(v)$ with rows equal to the vectors in $D^{\mathcal{R}}_s(v)$. By induction, this matrix has rank equal to $2n-5$. Let $M^{\mathcal{R}}_s(v)'$ be the matrix $M^{\mathcal{R}}_s(v)$ augmented with two additional columns from $A(\mathcal{R})$ so that $\operatorname{rank}(M^{\mathcal{R}}_s(v)') = 2n-3$. Since this matrix is full rank, there is again a unique solution to

$$M^{\mathcal{R}}_s(v)'x = M^{\mathcal{R}}_s(v)'\omega'.$$

This implies that we can uniquely determine the lengths of all $2n-3$ edges in $\mathcal{R}$. As a corollary, we can recover the lengths of all edges in $\mathcal{T}$ that are also in $\mathcal{R}$ and $\omega(e_a)+\omega(e_b)$, the sum of the lengths of edges e_a and e_b in $\mathcal{T}$.

Without loss of generality, suppose we have constructed $w = (w_1, w_2)$ so that $M^{\mathcal{T}}_1(v)$ contains all of the columns from $M^{\mathcal{R}}_1(v)$ and columns corresponding to $\alpha^{\mathcal{T}}_{L_1,n+1}$ and $\alpha^{\mathcal{T}}_{L_3,n+1}$. Then let $M^{\mathcal{T}}_1(v)'$ be the matrix that contains all of the columns from $M^{\mathcal{R}}_1(v)'$ and columns corresponding to $\alpha^{\mathcal{T}}_{L_1,n+1}$ and $\alpha^{\mathcal{T}}_{L_3,n+1}$. These columns enable us to recover the lengths of the paths from L_1 to $(n+1)$ and from L_3 to $(n+1)$ in $\mathcal{T}$. We will now show how this will enable us to determine the lengths of the remaining edges, e_{n+1}, e_a, and e_b uniquely. As explained, being able to determine all of the edge lengths of $\mathcal{T}$ from $M^{\mathcal{T}}_1(v)'$ shows that $M^{\mathcal{T}}_1(v)'$ has rank $2(n+1)-3$.

Since we know the length of the path from L_1 to $(n+1)$ and the length of every edge between L_1 and $(n+1)$ except e_{n+1} and e_a, we can determine $\omega(e_{n+1}) + \omega(e_a)$. Likewise, we know the length of the path from L_3 to $(n+1)$ and the length of every edge between

L_3 and $(n+1)$ except e_{n+1} and e_b, so we can recover $\omega(e_{n+1})+\omega(e_b)$. Combined with our knowledge of $\omega(e_a)+\omega(e_b)$ we can determine the lengths of e_{n+1}, e_a, and e_b. Uniqueness implies that the augmented matrix $M_1^{\mathcal{T}}(w)'$ has rank $2(n+1)-3$ and so $M_1^{\mathcal{T}}(w)$ has rank $2(n+1)-5$ as desired. If we have also chosen $w=(w_1,w_2)$ so that $M_2^{\mathcal{T}}(v)$ contains all of the columns from $M_2^{\mathcal{R}}(v)$ and columns corresponding to $\alpha^{\mathcal{T}}_{L_2,n+1}$ and $\alpha^{\mathcal{T}}_{L_4,n+1}$, then the same is true for $M_2^{\mathcal{T}}(w)$, and the theorem is complete.

It remains to show that we can actually choose the six vector entries w_1^a, w_1^b, w_1^{n+1}, w_2^a, w_2^b, and w_2^{n+1} in the manner specified. First, note that every edge in $\mathcal{T}$ along the path from u_a to L_1 or L_2 and u_b to L_3 or L_4 is contained in $\mathcal{R}$. Therefore, we let a_i^s be the v_s-weight of the path from u_a to L_i with $i=1,2$ and we have:

$$\begin{aligned}
w_1\cdot\alpha^{\mathcal{T}}_{L_1,n+1} &= a_1^1+w_1^a+w_1^{n+1},\\
w_2\cdot\alpha^{\mathcal{T}}_{L_1,n+1} &= a_1^2+w_2^a+w_2^{n+1},\\
w_1\cdot\alpha^{\mathcal{T}}_{L_2,n+1} &= a_2^1+w_1^a+w_1^{n+1},\\
w_2\cdot\alpha^{\mathcal{T}}_{L_2,n+1} &= a_2^2+w_2^a+w_2^{n+1}.
\end{aligned}$$

Recall that our goal is for both 1 and 2 to win one of $\{\alpha^{\mathcal{T}}_{L_1,n+1},\alpha^{\mathcal{T}}_{L_2,n+1}\}$ and $\{\alpha^{\mathcal{T}}_{L_3,n+1},\alpha^{\mathcal{T}}_{L_4,n+1}\}$. Rearranging, we would like to have

$$\begin{aligned}
a_1^1+w_1^a+w_1^{n+1} &< a_1^2+w_2^a+w_2^{n+1},\\
a_2^1+w_1^a+w_1^{n+1} &> a_2^2+w_2^a+w_2^{n+1}\\
\Rightarrow (w_1^a+w_1^{n+1})-(w_2^a+w_2^{n+1}) &< a_1^2-a_1^1\\
(w_1^a+w_1^{n+1})-(w_2^a+w_2^{n+1}) &> a_2^2-a_2^1
\end{aligned}$$

.

If we let $(w_1^a+w_1^{n+1})=(a_1^2-a_1^1)/2$ and $(w_2^a+w_2^{n+1})=-(a_2^2-a_2^1)/2$ then $(w_1^a+w_1^{n+1})-(w_2^a+w_2^{n+1})$ is just the average of $(a_1^2-a_1^1)$ and $(a_2^2-a_2^1)$. For w chosen sufficiently generic, the inequalities above may both be switched, but regardless, we will have sent the vectors $\{\alpha^T_{L_1,n+1},\alpha^T_{L_2,n+1}\}$ into different matrices. By symmetry, we let b_i^s be the v_s-weight of the path from u_b to L_i for $i=3,4$. Then we will be done if the following system has a solution:

$$
\begin{aligned}
w_1^a + w_1^{n+1} &= (a_1^2 - a_1^1)/2 \\
w_2^a + w_2^{n+1} &= -(a_2^2 - a_2^1)/2 \\
w_1^b + w_1^{n+1} &= (b_1^2 - b_1^1)/2 \\
w_2^b + w_2^{n+1} &= -(b_2^2 - b_2^1)/2 \\
w_1^a + w_1^b &= v_1^e \\
w_2^a + w_2^b &= v_2^e.
\end{aligned}
$$

The last two equations are necessary so that s wins $\alpha_{ij}^{\mathcal{T}}$ at w if and only if s wins $\alpha_{ij}^{\mathcal{R}}$ at v. The resulting matrix is full rank. □

Finally, we are able to complete our proof.

Proof of Theorem 4.1.4. Corollary 4.3.8 shows that if $\mathcal{T}$ has five or more cherries then $in_\omega(I_{2,n}^{\{2\}}) \neq J_{\mathcal{T}}^{\{2\}}$.

Let $\mathcal{T}$ be a binary phylogenetic X-tree with fewer than 5 cherries. By Lemma 4.3.11, $\dim(J_{\mathcal{T}}^{\{2\}}) \geq 4n - 10$, and since $in_\omega(I_{2,n}^{\{2\}}) \subseteq J_{\mathcal{T}}^{\{2\}}$, and $\dim(in_\omega(I_{2,n}^{\{2\}})) = 4n - 10$, $\dim(in_\omega(I_{2,n}^{\{2\}})) = \dim(J_{\mathcal{T}}^{\{2\}})$. By Theorem 4.3.4, $in_\omega(I_{2,n}^{\{2\}})$ is prime and of the same dimension as $J_{\mathcal{T}}^{\{2\}}$, which implies $in_\omega(I_{2,n}^{\{2\}}) = J_{\mathcal{T}}^{\{2\}}$. □

4.3.3 Beyond the Second Secant

Based on the proof of Theorem 4.1.4 and the result of Lemma 4.3.4, we have the following corollary which is a modification of the statement of Conjecture 4.1.3.

Corollary 4.3.12. *Let $\mathcal{T}$ be an n-leaf binary phylogenetic X-tree, $\omega \in \mathbb{R}^n$ a weight vector, and $\tau \in \{\pm 1\}^{\binom{n}{2}}$ a sign vector such that $J_{\mathcal{T}} = \tau \cdot in_\omega(I_{2,n})$. Then $\tau \cdot in_\omega(I_{2,n}^{\{r\}}) = J_{\mathcal{T}}^{\{r\}}$ if and only if $\dim(J_{\mathcal{T}}^{\{r\}}) = 2rn - 2r^2 - r$.*

We have already seen that Conjecture 4.1.3 is not true for trees with more than $2r$ cherries. However, as r increases, the number of cherries is not the only obstruction. The presence of other tree structures factors into a bound on the possible dimension of $J_{\mathcal{T}}^{\{r\}}$.

Removing an edge from a binary phylogenetic X-tree creates two connected components each of which is a rooted binary phylogenetic K-tree for some $K \subseteq X$. If one of these

rooted trees is a k-leaf rooted caterpillar then we call this rooted subtree a *k-cluster* of $\mathcal{T}$. Cherries, then, may alternatively be referred to as 2-clusters. We let c_k be the number of k-clusters in a tree. If leaves i and j are contained in an s-cluster, then we let k be the smallest such s and call the variable p_{ij} a *k-cluster variable for $\mathcal{T}$*.

Example 4.3.13. Let $\mathcal{T}$ be the tree in Figure 4.3. Then $\mathcal{T}$ has three 3-clusters on the leaves $\{1,2,3\}$, $\{4,5,6\}$, and $\{11,12,13\}$. The set of 2-cluster variables is

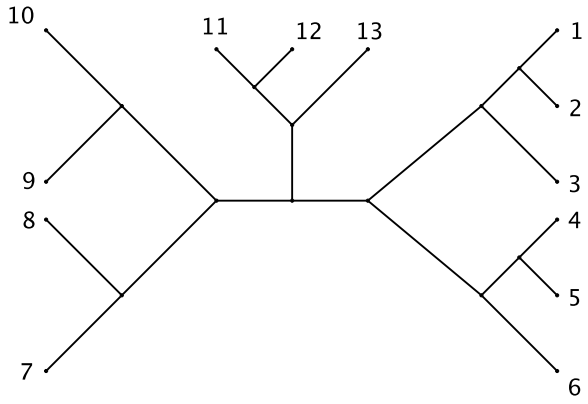
$$\{p_{1,2}, p_{4,5}, p_{7,8}, p_{9,10}, p_{11,12}\}$$

and the set of 3-cluster variables is

$$\{p_{1,3}, p_{2,3}, p_{4,6}, p_{5,6}, p_{11,13}, p_{12,13}\}.$$

Notice that the way clusters are nested, the number of k-cluster variables in a tree will be $(k-1)c_k$.

Figure 4.3: A 13-leaf tree with five 2-clusters and three 3-clusters.



Lemma 4.3.14. *Let c_k be the number of k-clusters in $\mathcal{T}$, then*

$$\dim(J_{\mathcal{T}}^{\{r\}}) \leq 2rn - 3r - \sum_{k=2}^{r}(r-k+1)c_k.$$

Proof. Let ω be a weight vector such that $J_{\mathcal{T}} = in_{\omega}(I_{2,n})$. Let $J_{\mathcal{T}}^{E}$ be the ideal constructed

by eliminating all k-cluster variables from $J_{\mathcal{T}}$ for $1 \leq k \leq r-1$, and embedding this ideal in Z^n. Define $V(J_{\mathcal{T}}^E) = W$ and note that $V(J_{\mathcal{T}}) \subseteq W$ and $\dim(J_{\mathcal{T}}^{\{r\}}) \leq \dim(\mathcal{I}(W)^{\{r\}})$. There are no restrictions on the $\sum_{k=2}^{r}(k-1)c_i$ eliminated variables in W, so we may write $W = W' \times \mathbb{C}^{\sum_{k=2}^{r}(k-1)c_i}$. Since $\mathbb{C}^{\sum_{k=2}^{r}(k-1)c_k}$ is a linear space, $W^{\{r\}} = W'^{\{r\}} \times \mathbb{C}^{\sum_{k=2}^{r}(k-1)c_k}$. We also observe that W' is a cone since it is a coordinate projection of a cone. Thus, $\dim(W^{\{r\}}) \leq r\dim(W'^{\{r\}}) + \sum_{k=2}^{r}(k-1)c_k$.

Now we seek a bound for $\dim(W'^{\{r\}})$. Choose a specific k-cluster in $\mathcal{T}$, define a grading with every k-cluster variable in that k-cluster having weight one and every other variable having weight zero. Observe that each binomial generator of $J_{\mathcal{T}}$ is homogeneous with respect to this grading. Thus, the equations in any reduced Gröbner basis for $J_{\mathcal{T}}$ with respect to any monomial order contain at least two distinct k-cluster variables from the designated k-cluster if they contain any at all. If not, by homogeneity, there exists an equation in the reduced Gröbner basis in which p_{ij}, a k-cluster variable from the designated k-cluster, can be factored. Since $J_{\mathcal{T}}$ is prime, that implies that p_{ij} is zero, which is evidently not true from the parameterization. Therefore, choosing an elimination order and eliminating any $(k-2)$ of the designated k-cluster variables from $J_{\mathcal{T}}$ eliminates all of the $(k-1)$ designated k-cluster variables. Thus, projecting away all of the k-cluster variables from a given k-cluster in $V(J_{\mathcal{T}})$ yields a variety of at least one dimension less. Applying the same argument to each k-cluster implies $\dim(W') \leq 2n - 3 - \sum_{k=2}^{r} c_k$, and the result follows. □

In the case where $r = 2$, this is just a restatement of Lemma 4.3.7. When $r = 3$, we have $\dim(I_{2,n}^{\{3\}}) = 6n-21$, so this tells us that it is impossible for $J_{\mathcal{T}}^{\{3\}} = I_{2,n}^{\{3\}}$ when $2c_2 + c_3 > 12$.

Example 4.3.15. Let $\mathcal{T}$ be the 13-leaf tree pictured in Figure 4.3. Then $c_2 = 5$, $c_3 = 3$, and $2c_2 + c_3 = 13$. Lemma 4.3.14 tells us that $\dim(J_{\mathcal{T}}^{\{3\}}) \leq 66 < 67 = \dim(I_{2,n}^{\{3\}})$ so that $J_{\mathcal{T}}^{\{3\}} \neq I_{2,n}^{\{3\}}$. Evaluating the Jacobian matrix at a point we find that $\dim(J_{\mathcal{T}}^{\{3\}}) = 66$.

Finally, one might wonder if we can modify Conjecture 4.1.3 as follows.

Conjecture 4.3.16. *Let $\mathcal{T}$ be an n-leaf binary phylogenetic X-tree, $\omega \in \mathbb{R}^n$ a weight vector, and $\tau \in \{\pm 1\}^{\binom{n}{2}}$ a sign vector such that $J_{\mathcal{T}} = \tau \cdot in_\omega(I_{2,n})$. Then $\tau \cdot in_\omega(I_{2,n}^{\{r\}}) = J_{\mathcal{T}}^{\{r\}}$ if and only if*

$$\sum_{k=2}^{r}(r-k+1)c_k < 2r^2 - 2r.$$

We have investigated $\dim(J_{\mathcal{T}}^{\{r\}})$ for $r = 3$ and $r = 4$ and several trees up to 18 leaves. By evaluating the Jacobian matrix at random points, we have found in each case that the conjecture holds. While this evidence supports Conjecture 4.3.14, to prove it rigorously, one might like to utilize induction as we did in Lemma 4.3.11. For example, when $r = 3$, there exists a finite set of trees with $2c_2 + c_3 \leq 12$ from which every tree with $2c_2 + c_3 \leq 12$ can be constructed by attaching leaves outside of the clusters to one of the trees in this set. We could verify the conjecture for all of the trees in this set and then proceed by induction.

However, there are two difficulties. First, the number and size of the trees in the base case becomes unwieldy quickly. Consider that for $r = 3$, we would now have to establish the dimension of the 12-leaf trees constructed by attaching a cherry to each of the vertices in the snowflake and the 6-leaf caterpillar, among others. Secondly, we would have to show that attaching a new leaf increases the rank of each of the $2r$ vector partitions by two. In the proof of Lemma 4.3.11, for $1 \leq s \leq 2$, we constructed w_s so that s won a direction corresponding to a path between the $(n+1)$ leaf and a leaf from each side of the split induced by removing the edge where the $(n+1)$ leaf was attached. For the r-secant, we would need to do the same thing for $1 \leq s \leq r$, which would require that there are at least r leaves on both sides of the induced split. There would be many trees that we would have to consider where this is not the case. Even when it is the case, before we were able to construct w from v to dictate the new winning directions. For $r = 3$, we have found vectors that induce partitions of the correct rank for a certain tree that can not be similarly modified to induce partitions of the correct rank when a new leaf is attached.

REFERENCES

[AAR08] Elizabeth S. Allman, C. Ané, and John A. Rhodes, *Identifiability of a Markovian model of a molecular evolution with gamma-distributed rates*, Adv. Appl. Prob. **40** (2008), 229–249.

[APRS11] Elizabeth S. Allman, Sonja Petrović, John A. Rhodes, and Seth Sullivant, *Identifiability of 2-tree mixtures for group-based models*, IEEE/ACM Trans. Comp. Biol. Bioinformatics **8** (2011), no. 3, 710–722.

[AR06] Elizabeth S. Allman and John A. Rhodes, *The identifiability of tree topology for phylogenetic models, including covarion and mixture models*, J. Comp. Biol. **13** (2006), no. 5, 1101–1113.

[AR07] ———, *Reconstructing evolution: New mathematical and computational advances*, ch. Phylogenetic Invariants, Oxford University Press, UK, June 2007.

[AR08a] ———, *Identifying evolutionary trees and substitution parameters for the general Markov model with invariable sites*, Math. Biosci. **211** (2008), no. 1, 18–33.

[AR08b] Elizabeth S. Allman and John A. Rhodes, *Phylogenetic ideals and varieties for the general Markov model*, Adv. in Appl. Math. **40** (2008), no. 2, 127–148.

[ARS12] Elizabeth S. Allman, John A. Rhodes, and Seth Sullivant, *When do phylogenetic mixture models mimic other phylogenetic models?*, Syst. Biol. **61** (2012), no. 6, 1049–1059.

[BW07] Weronika Buczynska and Jaroslaw Wisniewski, *On the geometry of binary symmetric models of phylogenetic trees*, Journal of the European Mathematical Society **9** (2007), no. 3, 609–635.

[CF87] J.A. Cavender and Joseph Felsenstein, *Invariants of phylogenies in a simple case with discrete states*, J. of Class. **4** (1987), 57–71.

[Cha96] J.T. Chang, *Full reconstruction of Markov models on evolutionary trees: identifiability and consistency.*, Math. Biosci. **137** (1996), no. 1, 51–73.

[CS05] Marta Casanellas and Seth Sullivant, *Algebraic statistics for computational biology*, ch. 16, Cambridge University Press, Cambridge, United Kingdom, 2005.

[DK09] Jan Draisma and Jochen Kuttler, *On the ideals of equivariant tree models*, Math. Ann. **344** (2009), no. 3, 619–644. MR 2501304 (2010e:62245)

[DR09] James H. Degnan and Noah A. Rosenberg, *Gene tree discordance, phylogenetic inference and the multispecies coalescent*, Trends in Ecology & Evolution **24** (2009), no. 6, 332–340.

[Dra08] Jan Draisma, *A tropical approach to secant dimensions*, J. Pure Appl. Algebra **212** (2008), no. 2, 349–363.

[DSS07] Mathias Drton, Bernd Sturmfels, and Seth Sullivant, *Algebraic factor analysis: Tetrads, pentads and beyond*, Probability Theorey and Related Fields **138** (2007), no. 3–4, 463–493.

[Eis04] David Eisenbud, *Commutative algebra with a view toward algebraic geometry*, Graduate Texts in Mathematics, no. 150, Springer-Verlag, 2004.

[ES93] S.N. Evans and T.P. Speed, *Invariants of some probability models used in phylogenetic inference*, Ann. Statist. **21** (1993), no. 1, 355–377.

[Fel81] Joseph Felsenstein, *Evolutionary trees from DNA sequences: A maximum likelihood approach*, J. Mol. Evol. **17** (1981), 368–376.

[GS02] D.R. Grayson and M.E. Stillman, *Macaulay2, a software system for research in algebraic geoemetry*, Available at http://www.math.uiuc.edu/Macaulay2/, 2002.

[GSS05] Luis David Garcia, Michael Stillman, and Bernd Sturmfels, *Algebraic geometry of Bayesian networks*, Journal of Symbolic Computation **39** (2005), no. 3-4, 331–355.

[Has07] Brendan Hassett, *Introduction to algebraic geometry*, Cambridge University Press, New York, 2007.

[Hum08] P.J. Humphries, *Combinatorial aspects of leaf-labelled trees*, Ph.D. thesis, University of Canterbury, 2008.

[JW07] Jakob Jonsson and Volkmar Welker, *A spherical initial ideal for Pfaffians*, Illinois J. Math **51** (2007), no. 4, 1397–1407.

[KL80] H. Kleepe and D. Laksov, *The algebraic structure and deformation of Pfaffian schemes*, J. Algebra **64** (1980), 167–189.

[KS95] Michael Kalkbrener and Bernd Sturmfels, *Initial complexes of prime ideals*, Advances in Mathematics **116** (1995), no. 2, 365–376.

[Lak87] J. A. Lake, *A rate-independent technique for analysis of nucleaic acid sequences: Evolutionary parsimony*, Molecular Biology and Evolution **4** (1987), 167–191.

[Map] *Maple 14*, Maplesoft, a division of Waterloo Maple Inc. Waterloo, Ontario.

[MMS08] Frederick A. Matsen, Elchanan Mossel, and Mike Steel, *Mixed-up trees: the structure of phylogenetic mixtures*, Bull. Math Biol. **70** (2008), no. 4, 1115–1139.

[MS05] Ezra Miller and Bernd Sturmfels, *Combinatorial commutative algebra*, Springer, 2005.

[MS07] Frederick A. Matsen and Mike Steel, *Phylogenetic mixtures on a single tree can mimic a tree of another topology*, Syst. Biol. **56** (2007), no. 5, 767–775.

[MS15] Diane Maclagan and Bernd Sturmfels, *Introduction to tropical geometry*, Graduate Studies in Mathematics, vol. 161, American Mathematical Society, 2015.

[PS04] Lior Pachter and David Speyer, *Reconstructing trees from subtree weights*, Applied Mathematical Letters **17** (2004), 615–621.

[PS05] Lior Pachter and Bernd Sturmfels (eds.), *Algebraic statistics for computational biology*, p. 101, Cambridge University Press, Cambridge, United Kingdom, 2005.

[RS12] John A. Rhodes and Seth Sullivant, *Identifiability of large phylogenetic mixtures*, Bull. Math Biol. **74** (2012), no. 1, 212–231.

[SESP93] L. Székely, P.L. Erdös, M.A. Steel, and D. Penny, *A Fourier inversion formula for evolutionary trees*, Applied Mathematics Letters **6** (1993), no. 2, 13–17.

[SF95] Mike A. Steel and Y.X. Fu, *Classifying and counting linear phylogenetic invariants for the Jukes-Cantor mixture model*, J. Comp. Biol. **2** (1995), no. 1, 39–47.

[SS03] Charles Semple and Mike Steel, *Phylogenetics*, Oxford University Press, Oxford, 2003.

[SS04] David Speyer and Bernd Sturmfels, *The tropical Grassmannian*, Adv. Geom. **4** (2004), 389–411.

[SS05] Bernd Sturmfels and Seth Sullivant, *Toric ideals of phylogenetic invariants*, J. Comp. Biol. **12** (2005), no. 2, 204–228.

[SS06] ———, *Combinatorial secant varieties*, Quarterly Journal of Pure and Applied Mathematics **2** (2006), 285–309.

[Stu96] Bernd Sturmfels, *Gröbner bases and convex polytopes*, vol. 8, American Mathematical Soc., 1996.

[Sul08] Seth Sullivant, *Algebraic geometry of Gaussian Bayesian networks*, Advances in Applied Mathematics **40** (2008), no. 4, 482–513.

[Sul12] ———, *The disentangling number for phylogenetic mixtures*, SIAM J. Discrete Math **26** (2012), no. 2, 856–859.

[SX10] Bernd Sturmfels and Zhiqiang Xu, *Sagbi bases of Cox-Nagata rings*, Journal of the European Mathematical Society **12** (2010), no. 2, 429–459.