PHYLOGENETIC IDEALS AND VARIETIES FOR THE GENERAL MARKOV MODEL

ELIZABETH S. ALLMAN AND JOHN A. RHODES

Abstract. The general Markov model of the evolution of biological sequences along a tree leads to a parameterization of an algebraic variety. Understanding this variety and the polynomials, called phylogenetic invariants, which vanish on it, is a problem within the broader area of Algebraic Statistics.

For an arbitrary trivalent tree, we determine the full ideal of invariants for the 2-state model, establishing a conjecture of Pachter-Sturmfels. For the κ-state model, we reduce the problem of determining a defining set of polynomials to that of determining a defining set for a 3-leaved tree.

Along the way, we prove several new cases of a conjecture of Garcia-Stillman-Sturmfels on certain statistical models on star trees, and reduce their conjecture to a family of subcases.

1. Introduction

An important problem arising in modern biology is that of sequence-based *phylogenetic inference.* Suppose we obtain a collection of biological sequences, such as genomic DNA, from currently extant species, or *taxa*. Assuming these sequences evolved from a common ancestral one, how can we infer a tree that describes their evolutionary descent? The use of algebraic methods for this problem was first proposed in 1987 in independent works by Lake [Lak87], and Cavender and Felsenstein [CF87]. Recently, Garcia, Stillman, and Sturmfels [GSS03] initiated a more general algebraic study of statistical models, of which phylogenetic models are a particularly interesting example. In this new field, Algebraic Statistics, the viewpoints of algebraic geometry are central to investigations of probabilistic models arising in applied contexts.

In model-based phylogenetics, evolution is usually assumed to proceed along a bifurcating (*i.e.*, trivalent) tree from an ancestral sequence at the root of the tree, to sequences found in the taxa, which label the leaves of the tree. The $\kappa = 4$ bases A, C, G, T of which DNA is composed are viewed as states of random variables. Each site in the

Date: October 20, 2004.

sequence might be assumed to evolve i.i.d., so that different sites can be viewed as trials of the same process. Probabilities of the various base substitutions along an edge of the tree can then be given by a Markov transition matrix along that edge. Additional biologically reasonable, or mathematically convenient, assumptions as to the form of these transition matrices are often imposed. The basic problem is to assume some model along these lines and use it to infer, from observations of DNA sequences only at the leaves, a tree topology that might describe their evolutionary descent. An excellent overview of the field of phylogenetics is provided by the recent volume of Felsenstein [Fel04].

In the phylogenetics literature, a *phylogenetic invariant* for a particular model and tree is a polynomial that vanishes on all joint distributions of bases at the leaves that arise from the model, regardless of the values of the model parameters. In the terminology of algebraic geometry, the model and tree imply a parameterization of a dense subset of a variety, and phylogenetic invariants are the elements of the ideal defining that variety. For applications, one might hope that the near-vanishing of phylogenetic invariants on observed frequencies of bases in DNA data could be used as a test of model-fit and/or tree topology. See Chapter 22 of [Fel04] for a guide to all but the most recent literature on phylogenetic invariants.

In this paper we investigate the phylogenetic variety for the general Markov model of base substitution for an arbitrary bifurcating tree. (A detailed specification of this model will be given in the next section.)

One main result is the proof of Conjecture 13 of Pachter and Sturmfels [PS04] on the ideal of phylogenetic invariants for the general Markov model in the case of $\kappa = 2$ states: the invariants arising from all 3×3 minors of '2-dimensional flattenings' of an array along the edges of a n-taxon tree T generate the full ideal. This is Theorem 2, which is stated more fully in Section 4 and proved in Section 8.

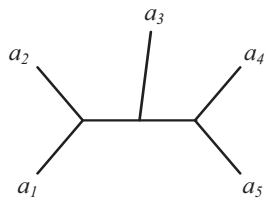

FIGURE 1. A 5-taxon tree

For an explicit example of this theorem, consider the 5-taxon tree of Figure 1. Then for the 2-state model, denote the states by 0 and

1. A $2 \times 2 \times 2 \times 2 \times 2$ tensor P encodes the probabilities of various states at the leaves, where $P(i_1, i_2, i_3, i_4, i_5) = p_{i_1 i_2 i_3 i_4 i_5}$ is the joint probability of observing state i_j in the sequence at leaf a_j, $j = 1, \ldots, 5$. Now P has two natural flattenings according to the partitions of leaves produced by deleting an internal edge of the tree. The partitions, or *splits*, are $\{\{a_1, a_2\}, \{a_3, a_4, a_5\}\}$, and $\{\{a_1, a_2, a_3\}, \{a_4, a_5\}\}$, and the corresponding flattenings

$$\begin{pmatrix} p_{00000} & p_{00001} & p_{00010} & p_{00011} & p_{00100} & p_{00101} & p_{00110} & p_{00111} \\ p_{01000} & p_{01001} & p_{01010} & p_{01011} & p_{01100} & p_{01101} & p_{01110} & p_{01111} \\ p_{10000} & p_{10001} & p_{10010} & p_{10011} & p_{10100} & p_{10101} & p_{10110} & p_{10111} \\ p_{11000} & p_{11001} & p_{11010} & p_{11011} & p_{11100} & p_{11101} & p_{11110} & p_{11111} \end{pmatrix}$$

and

$$\begin{pmatrix} p_{00000} & p_{00001} & p_{00010} & p_{00011} \\ p_{00100} & p_{00101} & p_{00110} & p_{00111} \\ p_{01000} & p_{01001} & p_{01010} & p_{01011} \\ p_{01100} & p_{01101} & p_{01110} & p_{01111} \\ p_{10000} & p_{10001} & p_{10010} & p_{10011} \\ p_{10100} & p_{10101} & p_{10110} & p_{10111} \\ p_{11000} & p_{11001} & p_{11010} & p_{11011} \\ p_{11100} & p_{11101} & p_{11110} & p_{11111} \end{pmatrix}.$$

The theorem states that the 3×3 minors of these two matrices generate the ideal of all phylogenetic invariants for the 2-state general Markov model on this tree. In particular, the ideal of invariants has a natural set of generators that correspond to the splits, and therefore to specific topological features of the tree.

We note that this theorem provides the first determination of all phylogenetic invariants for an arbitrary tree for any non-group-based model. Sturmfels and Sullivant [SS04] solved the similar problem for group-based models, using the Hadamard conjugation ([ES93, HP93]) to recognize the varieties as toric. While algebraic models intermediate to the group-based and general ones have been introduced recently [AR04a, ERSS04], our knowledge of them is less complete.

We also investigate the question of the explicit determination of the phylogenetic variety and ideal for larger κ. We show in Theorem 11 that if we have a set of polynomials whose zero set is the variety for the 3-taxon tree, then we can construct a set of polynomials whose zero set is the variety for an n-taxon tree. Similar to the conjecture of [PS04], our constructions involve 'flattenings', though both 2- and 3-dimensional ones are now needed, as might be expected from [AR03]. Thus the only remaining obstruction to our determination of a defining set of polynomials for the phylogenetic variety for any bifurcating tree

and any number of states κ is the determination of a defining set for the 3-taxon tree variety.

In Conjecture 1 we suggest that the same construction yielding set-theoretic defining polynomials for the variety would yield generators of the full ideal vanishing on the variety, provided we begin with generators of the ideal for the 3-taxon tree. This is the analog for arbitrary κ of the Pachter-Sturmfels conjecture.

Theorem 2, Theorem 11, and Conjecture 1, as well as the Sturmfels-Sullivant group-based result, can all be viewed as statements that the phylogenetic varieties and ideals arise from the 'local structure' of the tree. Exploiting this observation to provide better ways of characterizing the statistical support a data set might provide for specific local tree features would be interesting work for the future. In particular, invariants might provide a means of characterizing support for particular splits or tripartitions of the taxa.

Despite our primary focus on phylogenetic models, to prove Theorem 11 we must consider certain other statistical models on star trees. In Section 6, we therefore investigate models with a κ-state hidden variable associated to the internal node, and l_i-state observed variables associated to the n leaves. Such models are of course interesting in applications outside of phylogenetics, as they are examples of rather common 'mixture models' in statistics. Following [GSS03], they are termed *hidden naive Bayes models*.

Our work here focuses on such models in the case that for each i the number of states l_i is at least as large as the number of hidden states κ. Theorems 6 and 7 describe how set-theoretic and ideal-theoretic defining sets of the associated varieties can be deduced from set-theoretic and ideal-theoretic defining sets of the variety of the related model which has κ-state variables on each leaf.

As a consequence of this work on star tree models, in Corollary 9 we prove several cases of Conjecture 21 of [GSS03], on ideal generators for the hidden naive Bayes model with $\kappa = 2$. While one of these cases, for the 3-leaf tree, has been recently proved in [LM04], even for that case our argument is different, and perhaps more direct. Moreover, our work indicates that establishing the special cases mentioned in Conjecture 2 of this paper is sufficient to prove the full conjecture of [GSS03].

Before obtaining these results, we begin with several background sections. In Section 2, we define the phylogenetic variety for the general Markov model through the natural parameterization arising from modeling molecular evolution along a tree T by associating Markov matrices to each edge. In Section 3 we then give a more convenient

parameterization of (a dense subset of) the cone over the phylogenetic variety, which associates an arbitrary $\kappa \times \kappa$ matrix to each edge of T, rather than a Markov matrix. Section 4 introduces flattenings of tensors along edges and vertices of trees, while Section 5 develops the relationship of a form of multiplication of tensors to the varieties under investigation. Subsequent sections contain our primary results.

Finally, we note that throughout this paper we make the common assumption that all phylogenetic trees are bifurcating. In fact, higher valency of internal nodes could be allowed for Theorem 11, at the expense of further complicating the statement of the theorem and much of the earlier development of the paper. For clarity, we have chosen to present only the bifurcating case, but the results of Section 6 are sufficiently general to allow straightforward modifications to be made for higher valency.

Our proof of Theorem 2, on the other hand, requires that trees be bifurcating.

2. Affine and Projective Phylogenetic Varieties

Let T denote an n-taxon tree, by which we mean a tree with all internal vertices trivalent and unlabeled, with n leaves labeled by taxa $a_1, \ldots, a_n$. Choosing as a root any vertex r of T, either internal or a leaf, denote the rooted tree by T^r. Parameters for the κ-state *general Markov model* of sequence evolution on T^r consist of a root distribution vector $\boldsymbol{\pi}_r = (\pi_1, \pi_2, \ldots, \pi_\kappa)$ with non-negative entries summing to 1, together with a $\kappa \times \kappa$ Markov matrix M_e, which has non-negative entries with each row summing to 1, for each of the $2n-3$ edges e of T^r directed away from r.

This models the evolution of biological sequences as follows. The κ states $[\kappa] = \{1, 2, \ldots, \kappa\}$ correspond to the alphabet from which sequences are composed. The root r represents the most recent common ancestor of the currently extant taxa, and other internal nodes of the tree represent most recent common ancestors of those taxa separated from the root by that node. The root distribution vector encodes the frequencies π_i with which each state i occurs in an ancestral sequence at r. The (i, j)-entry of a Markov matrix along a particular edge of T directed away from r is the conditional probability of state i changing to state j at any particular site in the sequence during evolution along that edge. Thus each site in a biological sequence is assumed to evolve independently, according to the same process (i.i.d.). Note the biological term 'sequence' as used here implies no mathematical structure other than a correspondence between sites based on ancestry; except

for matching sites by common ancestry, the ordering of the sites within the sequences is irrelevant.

Since T has $2n-3$ edges, for this model of evolution along a rooted n-taxon tree T^r, the parameter space S can thus be identified with a subset of $[0,1]^N$, where $N = (\kappa-1) + (2n-3)\kappa(\kappa-1)$. Furthermore, there is a polynomial map $\phi_r : S \to [0,1]^L$, $L = \kappa^n$, giving the joint distribution of states in sequences at the leaves resulting from any parameter choice. We view points in $\phi_r(S)$ or $\mathbb{C}^L$ as $\kappa \times \cdots \times \kappa$ tensors, with the ith index referring to the state at leaf a_i. Indices thus typically range through $[\kappa]$, and a fixed ordering of the taxa is reflected in the ordering of indices of tensors. Assuming the model adequately reflects real molecular evolution, from biological sequence data we can estimate entries of $\phi_r(s)$, but usually have little or no direct information about s.

The map ϕ_r is explicitly given by $\phi_r(s) = P$, where

$$P(i_1, \ldots, i_n) = \sum_{(b_v) \in H} \left(\boldsymbol{\pi}_r(b_r) \prod_e M_e(b_{s(e)}, b_{f(e)}) \right), \tag{1}$$

where the product is taken over all edges e of T^r directed away from r, edge e has initial vertex $s(e)$ and final vertex $f(e)$ and associated Markov matrix M_e, and the sum is taken over the set

$$H = \{(b_v)_{v \in Vert(T)} \mid b_v \in [\kappa] \text{ if } v \neq a_j,\ b_v = i_j \text{ if } v = a_j\} \subset [\kappa]^{2n-2}.$$

Thus H represents the set of all 'histories' consistent with the specified states at the leaves.

The map ϕ_r can also be defined inductively, using matrix algebra, by viewing the tree T^r as built up from smaller trees by the addition of pairs of terminal edges, as we now explain.

A *cherry* of T is a pair of distinct leaves a_{i_1}, a_{i_2} whose incident edges contain a common (internal) vertex of T. For $n \geq 3$, any n-taxon tree contains at least two cherries, and any rooted n-taxon tree contains at least one cherry in which neither taxon of the cherry is the root of the tree.

For $n \geq 3$ let $T^r_n = T^r$ denote a rooted n-taxon tree labeled by taxa $a_1, \ldots, a_n$. Choose a cherry of T^r_n which does not contain the root r. Let T^r_{n-1} denote the rooted $(n-1)$-taxon tree obtained by deleting the cherry and its two incident edges from T^r_n and labeling as a new taxon, say b, the (formerly internal) common vertex of the incident edges.

Applying this definition recursively, we obtain from T^r a sequence of rooted trees $T^r_n, T^r_{n-1}, \ldots, T^r_2$, which of course may depend on some

arbitrary choices of cherries. We assume such choices have been made and fixed.

The map ϕ_r described above can now be described inductively as follows:

A rooted 2-taxon tree has only one edge e directed away from r, so with parameters $\boldsymbol{\pi}_r$ and M_e,

$$\phi_r(\boldsymbol{\pi}_r; M_e) = \operatorname{diag}(\boldsymbol{\pi}_r) M_e,$$

where $\operatorname{diag}(\mathbf{v})$ denotes the square matrix with $\mathbf{v}$ on its main diagonal and zeros elsewhere.

To define ϕ_r for T_m^r, direct edges e away from r and suppose parameters

$$s = (\boldsymbol{\pi}_r; \{M_e\})$$

are given. Then one obtains parameters $\tilde{s}$ for T_{m-1}^r by simply discarding from s the two Markov matrices associated to the edges of T_m^r not appearing in T_{m-1}^r. Inductively, we may assume $\tilde{\phi}_r : \tilde{S} \to [0,1]^{\kappa^{m-1}}$, the map giving the joint distribution of states at leaves for T_{m-1}^r as a function of parameters on T_{m-1}^r, has been given. For convenience, we also assume that taxa of T_m^r are $a_1, a_2, \ldots, a_m$ and those of T_{m-1}^r are $a_1, a_2, \ldots, a_{m-2}, b$, with the given orderings, and that e_1 and e_2 are the edges of T_m^r containing a_{m-1}, a_m respectively.

Then $\phi_r(s) = P$, where P is an m-dimensional tensor with 2-dimensional slices given by first letting $\tilde{P} = \tilde{\phi}_r(\tilde{s})$, $\mathbf{v} = \tilde{P}(i_1, \ldots, i_{m-2}, \cdot)$ and setting

$$P(i_1, \ldots, i_{m-2}, \cdot, \cdot) = M_{e_1}^T \operatorname{diag}(\mathbf{v}) M_{e_2}.$$

One can check that this definition of ϕ_r agrees with our earlier one, and so is independent of the choice of cherries defining the sequence $T_2^r, T_3^r, \ldots, T_n^r$.

We also denote by ϕ_r the unique extension of this map to a polynomial map $\phi_r : \mathbb{C}^N \to \mathbb{C}^L$. The *affine phylogenetic variety* $V(T)$ for the general Markov model on T is defined as the closure in $\mathbb{C}^L$ of the image of ϕ_r. (Throughout, all topological terminology of course refers to the Zariski topology.) As has been shown elsewhere [SSH94, AR03], this definition is independent of the choice of the root r. $V(T)$ is irreducible, as it is the zero set of a prime ideal, the kernel of the map between polynomial rings associated to ϕ_r.

Now one readily sees the image of ϕ_r lies on the hyperplane defined by the trivial phylogenetic invariant $\sum_{\mathbf{i} \in [\kappa]^n} P(\mathbf{i}) - 1 = 0$. It is therefore natural to pass to the *projective phylogenetic variety* in $\mathbb{P}^{L-1}$ by taking a projective closure. We denote this by $V(T)$ also, making clear by context whether the affine or projective version is meant.

The *phylogenetic ideals* of all polynomials vanishing on the affine phylogenetic variety or vanishing on the projective phylogenetic variety are of course closely related. Generators of the homogeneous ideal $\mathfrak{a}_T$ of the projective variety, together with the trivial invariant, generate the ideal of the affine variety. Conversely, any homogeneous polynomial in the ideal of the affine variety is in the homogeneous ideal of the projective variety. Thus identifying phylogenetic invariants for the general Markov model essentially means identifying those polynomials vanishing on the projective phylogenetic variety.

3. Reparameterization

For any projective variety $V \subseteq \mathbb{P}^m$, let $CV \subseteq \mathbb{C}^{m+1}$ denote the cone over V, that is, the union of the lines represented by points in V. Equivalently, CV is the affine variety defined by the same polynomials as V.

A dense subset of the cone $CV(T)$ admits a parameterization that will be more useful than the parameterization ϕ_r above. This new parameterization simplifies many arguments, since it allows matrices with any row sums to be associated to edges, and no longer requires a root distribution, or even a specification of a root.

Definition. Consider an n-taxon tree T. Let $U = \mathbb{C}^K$ with $K = (2n-3)\kappa^2$. Choose any vertex of T as a root, directing all edges of T^r away from r. View $u \in \mathbb{C}^K$ as a $(2n-3)$-tuple of complex $\kappa \times \kappa$ matrices M_e, one for each edge e of T^r.

Then let $\psi : U \to \mathbb{C}^L$ be given inductively as follows, using $T^r = T^r_n, T^r_{n-1}, \ldots, T^r_2$ as in the discussion in Section 2 :

If $n = 2$, $\psi(u) = \psi(M_e) = M_e$, so ψ is the identity map.

If $n > 2$, let $\tilde{\psi} : \tilde{U} \to \mathbb{C}^{\kappa^{n-1}}$ be the map associated to T^r_{n-1}. Then for $u \in U$, define $\tilde{u} \in \tilde{U}$, by omitting from u the matrices associated to the edges e_1, e_2 of T^r_n not in T^r_{n-1}. Then $\psi(u) = P$, where P is a n-dimensional tensor with 2-dimensional slices given by first letting $\tilde{P} = \tilde{\psi}(\tilde{u})$, $\mathbf{v} = \tilde{P}(i_1, \ldots, i_{n-2}, \cdot)$ and setting

$$P(i_1, \ldots, i_{n-2}, \cdot, \cdot) = M_{e_1}^T \operatorname{diag}(\mathbf{v}) M_{e_2}.$$

As with ϕ_r, one sees that this map is independent of the choice of cherries determining the sequence $T^r_2, T^r_3, \ldots, T^r_n$. Although ψ apparently depends on the choice of r, one can further check that if r is moved from one vertex of an edge e to the other vertex, we need only transpose the matrix M_e associated to that edge and the map is unchanged. Thus the map is independent of the choice of r, though our

conception of how components of $\mathbb{C}^K$ are placed into matrices does depend on r. Indeed, all these observations follow from the observation that ψ can also be defined by a formula like that in equation (1), but with the factor $\boldsymbol{\pi}_r(b_r)$ omitted.

Proposition 1. *The closure of $\psi(U)$ in $\mathbb{C}^L$ is the cone $CV(T)$ over the phylogenetic variety $V(T)$.*

Proof. To see $\phi_r(S) \subseteq \psi(U)$, suppose $s = (\boldsymbol{\pi}_r; \{M_e\}) \in S$. Let e_0 be the one edge of T_2^r, and define $M'_{e_0} = \mathrm{diag}(\boldsymbol{\pi}_r)M_{e_0}$. Then with $u = (M'_{e_0}, \{M_e\}_{e \neq e_0})$, we find that $\phi_r(s) = \psi(u)$. Thus $V(T) \subseteq \overline{\psi(U)}$. Furthermore $\psi(U)$ is a cone, since if $u = (\{M_e\}) \in U$ and $\lambda \in \mathbb{C}$, by picking any particular edge e_0 of T and defining $u' \in U$ to be identical to u but with λM_{e_0} replacing M_{e_0}, then $\psi(u') = \lambda\psi(u)$. Thus $CV(T) \subseteq \overline{\psi(U)}$.

We next show there is a non-empty open, and therefore dense, subset of U whose image under ψ lies in the cone over $\phi_r(S)$, and hence in $CV(T)$. This will imply $\overline{\psi(U)} \subseteq CV(T)$

First, if $n = 2$, then $\phi_r(S)$ certainly contains those 2-dimensional arrays whose entries add to 1 and none of whose row sums are 0. Now the subset of U on which all row sums of $M_e(= u)$ are non-zero and the total sum of the entries of $M_e(= u)$ is non-zero is an open set. The points in the image under ψ of this open set lie in the cone over $\phi_r(S)$.

Proceeding inductively, let e_1,e_2 be the edges of T_m^r which are not in T_{m-1}^r, and e_3 the third edge meeting them. We may also suppose r does not lie at the common vertex of e_1, e_2, e_3. Now there is an open $\mathcal{O}_1 \subset \mathbb{C}^K$ such that for points $u \in \mathcal{O}_1$, M_{e_1} and M_{e_2} have all row sums non-zero. Letting D_i be the invertible diagonal matrix constructed from the row sums of M_{e_i}, we may write

$$M_{e_i} = D_i M'_{e_i}, \quad i = 1, 2,$$

where M'_{e_i} has rows summing to 1. Let $M'_{e_3} = M_{e_3} D_1 D_2$. Then for any $u \in \mathcal{O}_1$, we define a new $u' \in \mathcal{O}_1$ as

$$u' = (\{M_e\}_{e \notin \{e_1,e_2,e_3\}}, M'_{e_1}, M'_{e_2}, M'_{e_3}),$$

so that $\psi(u') = \psi(u)$. Note that $\omega : \mathcal{O}_1 \to \mathcal{O}_1$ mapping $u \mapsto u'$ is given by rational functions.

Let $\tilde{\psi} : \tilde{U} \to \mathbb{C}^{\kappa^{m-1}}$ and $\tilde{\phi}_r : \tilde{S} \to \mathbb{C}^{\kappa^{m-1}}$ be the parameterizations associated to T_{m-1}^r. Then by induction there is a non-empty open $\tilde{\mathcal{O}} \subset \tilde{U}$ such that the image of all points in $\tilde{\mathcal{O}}$ under $\tilde{\psi}$ lie in the cone over $\tilde{\phi}_r(\tilde{S})$. Then $\mathcal{O} = \omega^{-1}(\tilde{\mathcal{O}} \times \mathbb{C}^{2\kappa^2})$ is a non-empty open subset of U, and the image of any point of $\mathcal{O}$ under ψ lies in the cone over $\phi_r(S)$. □

While the definition of ψ has introduced many unnecessary parameters, in the sense that the dimension of the image is much smaller than the dimension of the parameter space, it offers us the advantage of dropping inconvenient requirements — that row sums of vectors and matrices be 1 — that arose from the original probabilistic setting of the general Markov model.

4. Flattenings and phylogenetic invariants

To describe the set of phylogenetic invariants we are concerned with, we require the notion of *flattening* a tensor $P \in \mathbb{C}^{\kappa^n}$ according to an n-taxon tree T.

Let e be an edge of T. Then e induces a split of the taxa according to the connected components of $T \setminus \{e\}$. By reordering the indices in P if necessary, we may assume the split is $\{\{a_1, \ldots, a_k\},\ \{a_{k+1}, \ldots, a_n\}\}$. A *flattening of P on e* is a $\kappa^k \times \kappa^{n-k}$ matrix $F = Flat_e(P)$ defined as follows: Fix any ordering of $J_1 = [\kappa]^k$ and $J_2 = [\kappa]^{n-k}$, and for $u \in J_1$, $v \in J_2$, let $F(u, v) = P(u_1, \ldots, u_k, v_1, \ldots, v_{n-k})$.

If the tensor $P = \phi_r(s)$ gives the joint distribution of states for some parameter choice for the general Markov model on T, then $Flat_e(P)$ can be thought of as a joint distribution for a related graphical model with less complicated structure: With the root r chosen to be at one vertex of e, we imagine at r a κ-state hidden variable. The possible joint states at the taxa $a_1, \ldots, a_k$ are viewed as a single κ^k-state observed variable. Similarly, the joint states at the taxa $a_{k+1}, \ldots, a_n$ are described through a single κ^{n-k}-state variable. We thus have a "coarser" graphical model with one hidden κ-state internal node and two descendent nodes with κ^k and κ^{n-k} states, respectively, as depicted in Figure 2. The flattening of P simply prevents one from examining the finer structure in the joint distribution array that arises from the branching of T on either side of e.

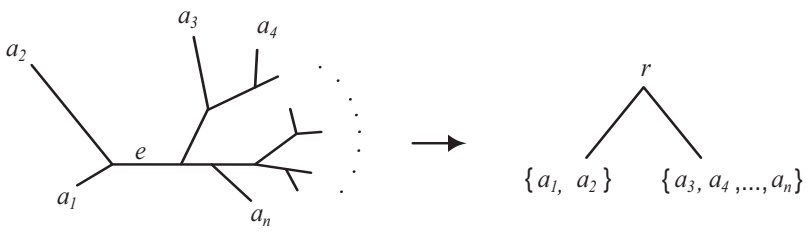

Figure 2. Flattening on an edge e

From this interpretation one readily sees that for any $P \in \phi_r(S)$, $Flat_e(P)$ has rank at most κ. Indeed, for the coarser graphical model, the joint distribution matrix must have the form

$$Flat_e(P) = M_1^T \operatorname{diag}(\pi_r) M_2$$

where M_1 and M_2 are $\kappa \times \kappa^k$ and $\kappa \times \kappa^{n-k}$ Markov matrices.

As a result, all $(\kappa+1) \times (\kappa+1)$ minors of $Flat_e(P)$ must vanish. As is classically known, such minors generate the full ideal of polynomials vanishing on matrices of rank $\leq \kappa$, and thus generate all invariants associated to the coarser model. For the original model on T, these minors therefore give phylogenetic invariants, which we call *edge invariants* associated to the edge e.

We denote by $\mathcal{F}_{edge}(T)$ the set of all $(\kappa+1) \times (\kappa+1)$ minors of all flattenings of a $\kappa \times \cdots \times \kappa$ tensor of κ^n indeterminants on edges of T. Of course the choice of ordering of rows and columns in the flattening introduces factors of ± 1, but as our goal is to determine ideal generators, we may ignore this issue.

In Section 8 we will establish the following, which was conjectured in [PS04].

Theorem 2. *For $\kappa = 2$ and any number of taxa n, the phylogenetic ideal $\mathfrak{a}_T$ for the general Markov model on an n-taxon tree T is generated by $\mathcal{F}_{edge}(T)$, the 3×3 minors of all edge flattenings of a $2 \times \cdots \times 2$ tensor of indeterminants.*

However, for larger κ it is not enough to consider only 2-dimensional edge flattenings (*i.e.*, flattenings to matrices) to obtain generators of the phylogenetic ideal. This can be seen already for the 3-taxon tree. In this case, $\mathcal{F}_{edge}(T)$ is empty, but for any $\kappa > 2$ the phylogenetic ideal contains polynomials of degree $\kappa + 1$ (see [AR03]; for $\kappa = 3$ see also [GSS03]). Thus we need at least to consider flattenings of P at internal nodes of T producing 3-dimensional tensors.

More specifically, let v be an internal node of T, contained in edges e_1, e_2, e_3. Then v induces a tripartition of the taxa according to the connected components of $T \setminus \{v, e_1, e_2, e_3\}$. By reordering the indices in P if necessary, we may assume the tripartition is

$$\{\{a_1, \ldots, a_k\},\ \{a_{k+1}, \ldots, a_{k+l}\},\ \{a_{k+l+1}, \ldots, a_n\}\}.$$

Then a *flattening of P at v* is a $\kappa^k \times \kappa^l \times \kappa^{n-k-l}$ array $F = Flat_v(P)$ defined as follows: Fix an ordering of $J_1 = [\kappa]^k$, $J_2 = [\kappa]^l$, and $J_3 = [\kappa]^{n-k-l}$, and for $u \in J_1$, $v \in J_2$, $w \in J_3$, let $F(u, v, w) = P(u_1, \ldots, u_k, v_1, \ldots, v_l, w_1, \ldots, w_{n-k-l})$.

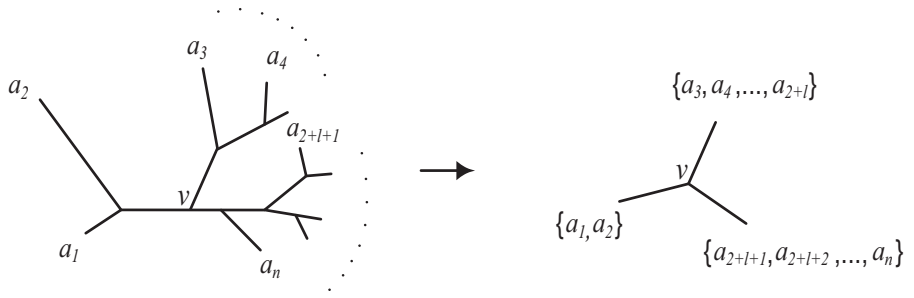

FIGURE 3. Flattening at a vertex v

As illustrated in Figure 3, we think of this flattening as producing a joint distribution array associated to a graphical model with one hidden κ-state internal node and three descendent nodes with κ^k, κ^l, and κ^{n-k-l} states, respectively. Similar to flattenings on edges, a flattening at an internal node ignores the finer structure in the joint distribution array that arises from the branching of T in the three directions leading away from v.

An ideal is associated to such a graphical model (1 hidden κ-state ancestral node, 3 descendent nodes), and so to the flattening at a vertex. While we will investigate such ideals further in Section 6, already we can formulate a natural extension of the conjecture of [PS04].

Conjecture 1. *For any κ and any number of taxa n, the phylogenetic ideal $\mathfrak{a}_T$ for the general Markov model on an n-taxon tree T is the sum of the ideals associated to the flattenings of P at vertices of T.*

That this conjecture is identical to Theorem 2 when $\kappa = 2$ follows from work of Landsberg and Manivel [LM04]. They show that in this special case the ideal associated to a vertex flattening is the sum of those associated to the edge flattenings on the three edges containing the vertex. (The Landsberg-Manivel result is a special case of a conjecture in [GSS03]. We will give a new proof of this case, and several additional cases, as Corollary 9.)

Although we will primarily need to refer to the 2- and 3-dimensional flattenings of a tensor P on an edge or at a vertex of a tree T, the notion naturally extends to flattenings based on any partition of the set of labels (taxa) associated to the indices of P. For instance, an n-dimensional $\kappa \times \cdots \times \kappa$ tensor P with associated labels $a_1, \ldots, a_n$ can be flattened according to the partition $\{\{a_1\}, \ldots, \{a_{n-2}\}, \{a_{n-1}, a_n\}\}$ to give an $(n-1)$-dimensional $\kappa \times \cdots \times \kappa \times \kappa^2$ tensor. We use such

a flattening, where a_{n-1}, a_n are in a cherry, in Section 8. Flattenings according to arbitrary bipartitions also appear in Section 6.

5. The algebra of tensors, trees, and parameters

In this section we define binary operations on trees, model parameters on trees, and tensors. These operations, all denoted by the same symbol '$*$', exhibit relationships that will make them useful in later sections.

Tensors: If Q and R are m- and n-dimensional tensors of 'matching size κ' in the last and first index respectively, then we define an $l = (m+n-2)$-dimensional tensor $Q * R$ by

$$(Q * R)(i_1, \dots i_l) = \sum_{j=1}^{\kappa} Q(i_1, \dots, i_{m-1}, j) R(j, i_m, \dots, i_l).$$

For $m = n = 2$, this is of course just matrix multiplication.

More generally, if the pth index of Q and the qth index of R both run through $[\kappa]$, we may define $Q *_{p,q} R$ by a similar sum. However, to keep our notation less cumbersome, we will generally try to express products using the last and first indices.

Trees: Suppose T' is a tree with taxa $a_1, a_2, \dots, a_m$, and T'' is a tree with taxa $b_1, b_2, \dots, b_n$. Then by $T' * T''$ we mean the $(m+n-2)$-taxon tree with taxa $a_1, \dots, a_{m-1}, b_2, \dots, b_n$ obtained by first identifying the vertices a_m and b_1, and then deleting this vertex, replacing the two edges it lies in with a single conjoined edge, as illustrated in Figure 4.

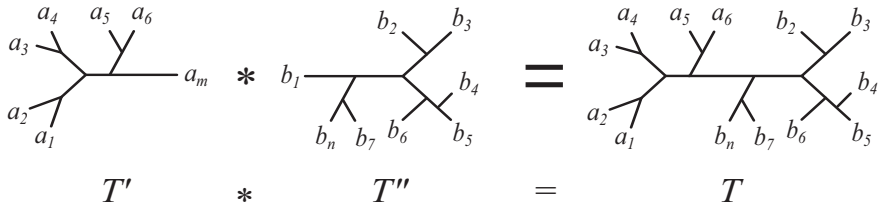

FIGURE 4. The $*$ operation on trees

Parameters: Consider trees T', T'', and $T = T' * T''$ with m, n, and $m+n-2$ taxa. Then from Section 3 we have the parameterizations

$$\psi' : U' \to \mathbb{C}^{\kappa^m},$$
$$\psi'' : U'' \to \mathbb{C}^{\kappa^n},$$
$$\psi : U \to \mathbb{C}^{\kappa^{m+n-2}},$$

of the cones over the associated phylogenetic varieties.

To impose directions on the edges of the trees for notational purposes, root T' and T at a_1, and T'' at b_1. Then for $u' \in U'$, $u'' \in U''$, we define $u' * u'' \in U$ by retaining for each edge of T except the conjoined one the matrix associated to the edge in either u' or u'', and for the conjoined edge using the product of the matrices in u' and u'' associated to its parts.

One readily sees that these three definitions imply the following.

Lemma 3. $\psi(u' * u'') = \psi'(u') * \psi''(u'')$.

Lemma 4. *If $T = T' * T''$, then $CV(T) = \overline{CV(T') * CV(T'')}$.*

Proof. It is clear that

$$U = U' * U'' = \{u' * u'' \mid u' \in U',\ u'' \in U''\}.$$

Thus by Lemma 3,

$$CV(T) = \overline{\psi(U)} = \overline{\psi'(U') * \psi''(U'')} = \overline{CV(T') * CV(T'')}.$$

□

This result will be strengthened in Corollary 14.

In the special case when T'' is a 2-taxon tree, $T' * T''$ is isomorphic to T'. Then $u'' = \psi(u'')$ is simply a $\kappa \times \kappa$ matrix. Informally, one can think of $\psi'(u') * \psi''(u'')$ as the result of 'extending' the edge of T' terminating at a_m and associating to the edge extension the matrix u''.

Considering invertible matrices u'', we get an action of $GL(\kappa, \mathbb{C})$ on both U' and $\psi'(U')$. Thus $GL(\kappa, \mathbb{C})$ acts on the closure, $CV(T')$, as well. Viewing the action described here as operating in 'the last index' of a tensor in $V_{T'}$, we similarly have an action in the other indices. These actions of $GL(\kappa, \mathbb{C})$ are of course just restrictions of the natural actions of that group on the set of all $\kappa \times \cdots \times \kappa$ tensors: For $j = 1, \ldots, n$, the 'jth index' action is defined by $P \mapsto P *_{j,1} A$ for $A \in GL(\kappa, \mathbb{C})$.

6. Models on star trees

In this section, we step back from the phylogenetic tree setting, and consider in more depth the hidden naive Bayes models of [GSS03]. Most of our results will be needed for application to phylogenetic varieties. However, we develop this material in slightly greater generality than we need, and so obtain partial results on a conjecture of [GSS03] as well.

The graphical models of this section are based on a star tree, as in Figure 5, with one internal vertex r, connected by edges to n leaves

$a_1, a_2, \ldots, a_n$. A hidden random variable associated to r has κ possible states, with probability distribution given by a vector $\boldsymbol{\pi}_r$. Each leaf a_i has associated to it a random variable with l_i states, and Markov matrices M_i of size $\kappa \times l_i$ give conditional probabilities of observing the various states at a_i given the state at r.

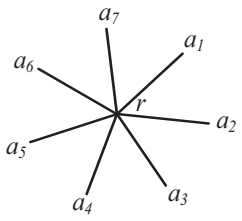

FIGURE 5. Graphical depiction of a hidden naive Bayes model

As in the phylogenetic situation, such a model defines a projective variety, the closure of the set of joint distributions of observations at the leaves arising from this parameterization. We denote this variety by $V(\kappa; l_1, l_2, \ldots, l_n)$, and the homogeneous ideal defining it by $\mathfrak{a}(\kappa; l_1, l_2, \ldots, l_n)$. As pointed out in [GSS03], the variety $V(\kappa; l_1, l_2, \ldots, l_n)$ can be viewed more geometrically as the κ-secant variety of the Segre product $\mathbb{P}^{l_1-1} \times \mathbb{P}^{l_2-1} \times \cdots \times \mathbb{P}^{l_n-1}$.

Note that $V(\kappa; \kappa, \kappa, \kappa) = V(T_3)$, the phylogenetic variety for a κ-state, 3-taxon tree. The varieties $V(\kappa; \kappa^k, \kappa^l, \kappa^{n-k-l})$, with $k, l, n-k-l \geq 1$, are the ones that arose in Section 4, in the discussion of flattenings of tensors at vertices of phylogenetic trees. Moreover, flattenings on edges involve $V(\kappa; \kappa^k, \kappa^{n-k})$, the variety of rank κ matrices of size $\kappa^k \times \kappa^{n-k}$, which is well understood classically.

Our first goals are to show Theorems 6 and 7, that given a set $\mathcal{F}$ of polynomials set-theoretically (respectively, ideal-theoretically) defining $V(\kappa; \kappa, \kappa, \ldots, \kappa)$ for the n-leaf star tree, then for any $l_i \geq \kappa$ we can explicitly construct polynomials set-theoretically (respectively, ideal-theoretically) defining $V(\kappa; l_1, l_2, \ldots, l_n)$.

Previous to these theorems, we know of only one general result concerning defining polynomials of $V(\kappa; l_1, l_2, \ldots, l_n)$: When $\kappa = 2$, for any number of leaves, [LM04] gives a natural set of polynomials defining the variety as a set.

For our application to phylogenetic trees, the assumption that internal nodes are trivalent means only the case $n = 3$ is needed. We therefore summarize known results on $V(\kappa; \kappa, \kappa, \kappa) = V(T_3)$ for small κ.

For $\kappa = 2$, as noted in [GSS03, AR03, LM04], $V(T_3) = \mathbb{P}^7$, and so $\mathcal{F} = \emptyset$ generates the full ideal defining the variety.

For $\kappa = 3$, $\mathcal{F}$ may be taken to be the 27 quartic polynomials in [GSS03], first found in [Str83] but also obtained from the construction in [AR03].

For $\kappa \geq 4$, giving an explicit set $\mathcal{F}$ even set-theoretically defining the variety is still an open problem. However, any polynomial vanishing on the variety must be of degree at least $\kappa + 1$.

When $\kappa = 4$, all degree 5 polynomials vanishing on the variety form an explicitly-known 1728-dimensional vector space. This dimension is computed in [LM04], and an explicit construction for general κ is given in [AR03] that produces a spanning set when $\kappa = 4$. Moreover, off another explicitly-known variety, the vanishing of these polynomials does distinguish points of $V(T_3)$. However, an explicit degree 9 polynomial is known which vanishes on $V(4; 3, 3, 3)$ (see [GSS03] for a statement, or [Str83] for the construction), and from this polynomial one can obtain degree 9 polynomials vanishing on $V(4; 4, 4, 4) = V(T_3)$ which are not in the ideal generated by the degree 5 invariants.

By modifying the approach of Section 3, it is possible to parameterize a dense subset of the cone $CV(\kappa; l_1, l_2, \ldots, l_n)$ using parameters which are arbitrary matrices. We leave the details to the reader, but denote this parameterization by $\psi_{\kappa;l_1,\ldots,l_n}$, where

$$\psi_{\kappa;l_1,\ldots,l_n} : U_{\kappa;l_1,\ldots,l_n} \to \mathbb{C}^L, \quad U_{\kappa;l_1,\ldots,l_n} = \mathbb{C}^{\kappa(l_1+\cdots+l_n)}, \ L = l_1 l_2 \cdots l_n,$$

and if $P = \psi(M_1, M_2, \ldots, M_n)$ then

$$P(i_1, \ldots, i_n) = \sum_{k=1}^{\kappa} \prod_{j=1}^{n} M_j(k, i_j).$$

Here $M_j \in M(\kappa, l_j, \mathbb{C})$, the set of complex $\kappa \times l_j$ matrices.

In order to relate $V(\kappa; l_1, l_2, \ldots, l_n)$ to $V(\kappa; \kappa, \kappa, \ldots, \kappa)$ we need the following lemma. It can be interpreted as describing the effect of extending one edge of the star tree, and associating a (non-square) matrix to that extension.

Lemma 5. *Let $P \in CV(\kappa; l_1, l_2, \ldots, l_n)$ and let $A \in M(l_n, l'_n, \mathbb{C})$. Then A defines a map $CV(\kappa; l_1, l_2, \ldots, l_n) \to CV(\kappa; l_1, l_2, \ldots, l'_n)$ by $P \mapsto P * A$. Furthermore,*

- *(i) If $rank(A) = l'_n$, then $CV(\kappa; l_1, l_2, \ldots, l_n) * A$ is dense in $CV(\kappa; l_1, l_2, \ldots, l'_n)$.*
- *(ii) If $\kappa \leq l_n$ then $CV(\kappa; l_1, l_2, \ldots, l_n) * M(l_n, l'_n, \mathbb{C})$ is dense in $CV(\kappa; l_1, l_2, \ldots, l'_n)$.*

Proof. Suppose first that $P = \psi_{\kappa;l_1,\ldots,l_n}(M_1, M_2, \ldots, M_n)$, with complex $\kappa \times l_i$ matrix parameters M_i, $i = 1, 2, \ldots, n$ associated to the n edges of T directed away from the internal node. Then

$$P * A = \psi_{\kappa;l_1,\ldots,l_{n-1},l'_n}(M_1, M_2, \ldots, M_{n-1}, M_n A),$$

hence $P * A \in CV(\kappa; l_1, l_2, \ldots, l'_n)$. Since $P * A \in CV(\kappa; l_1, l_2, \ldots, l'_n)$ for P in a dense subset of $CV(\kappa; l_1, l_2, \ldots, l_n)$, it follows that $P * A \in CV(\kappa; l_1, l_2, \ldots, l'_n)$ for all $P \in CV(\kappa; l_1, l_2, \ldots, l_n)$.

Now suppose $\operatorname{rank}(A) = l'_n$. Then as M_n ranges through all $\kappa \times l_n$ complex matrices, $M_n A$ ranges through all $\kappa \times l'_n$ complex matrices. Thus

$$\psi_{\kappa;l_1,\ldots,l_n}(U_{\kappa;l_1,\ldots,l_n}) * A = \psi_{\kappa;l_1,\ldots,l_{n-1},l'_n}(U_{\kappa;l_1,\ldots,l'_n}),$$

and so a subset of $CV(\kappa; l_1, \ldots, l_n) * A$ is dense in $CV(\kappa; l_1, \ldots, l'_n)$.

Finally suppose $\kappa \leq l_n$. Then as M_n ranges through all $\kappa \times l_n$ complex matrices and A through all $l_n \times l'_n$ matrices, $M_n A$ ranges through all $\kappa \times l'_n$ complex matrices. Thus

$$\psi_{\kappa;l_1,\ldots,l_n}(U_{\kappa;l_1,\ldots,l_n}) * M(l_n, l'_n, \mathbb{C}) = \psi_{\kappa;l_1,\ldots,l_{n-1},l'_n}(U_{\kappa;l_1,\ldots,l'_n}).$$

Therefore a subset of $CV(\kappa; l_1, \ldots, l_n) * M(l_n, l'_n, \mathbb{C})$ is dense in $CV(\kappa; l_1, \ldots, l'_n)$. □

Remark. For non-zero P and A as in the proof, it is possible that $P * A$ be a zero tensor. Thus while the above lemma could be formulated in terms of a rational map between the underlying projective varieties, it is slightly easier for us to consider a polynomial map on the cones.

By permuting indices Lemma 5 can be applied in any index, not just the last. As shorthand, we will refer to this as letting a $l_k \times l'_k$ matrix 'act in the kth index.' By considering only invertible $l_k \times l_k$ matrices, we have a group action of $GL(l_k, \mathbb{C})$ in the kth index, and so an action of $GL(l_1, \mathbb{C}) \times \cdots \times GL(l_n, \mathbb{C})$ on $V(\kappa; l_1, \ldots, l_n)$. While this group action underlies the dimension computations of [LM04], our work will emphasize the utility of non-square and non-invertible matrices as well.

Theorem 6. *Consider an n-leaf star tree. Suppose $l_1, l_2, \ldots, l_n \geq \kappa$. Let $\mathcal{F}$ be any set of polynomials whose zero set is $V(\kappa; \kappa, \kappa, \ldots, \kappa)$. For $k = 1, 2, \ldots, n$, let $Z_k = (z^k_{ij})$ be $l_k \times \kappa$ matrices of indeterminants. For an $l_1 \times l_2 \times \cdots \times l_n$ tensor P of indeterminants, let $\tilde{P}$ be the $\kappa \times \kappa \times \cdots \times \kappa$ tensor that results from letting each Z_k act formally in the kth index of P. Let $\tilde{\mathcal{F}}$ denote the set of polynomials in the entries of P obtained from those in $\mathcal{F}$ by substituting into them the entries of $\tilde{P}$, expressing the results as polynomials in the z^k_{ij}, and then extracting the coefficients.*

Let $\mathcal{F}_{edge}$ denote the set of $(\kappa+1)\times(\kappa+1)$ minors of the n flattenings of P on edges of the star tree. Finally, let $\mathcal{F}(\kappa; l_1, l_2, \ldots, l_n) = \tilde{\mathcal{F}} \cup \mathcal{F}_{edge}$.

Then $\mathcal{F}(\kappa; l_1, l_2, \ldots, l_n)$ defines $V(\kappa; l_1, l_2, \ldots, l_n)$ set-theoretically.

Proof. We first observe that all polynomials in $\mathcal{F}(\kappa; l_1, l_2, \ldots, l_n)$ vanish on the cone $CV(\kappa; l_1, l_2, \ldots, l_n)$: Polynomials in $\mathcal{F}_{edge}$ must vanish there, since the model has κ states at the internal node, so all 2-dimensional flattenings on edges must have rank $\leq \kappa$ on the parameterized subset of the variety, and hence on the whole variety. Polynomials in $\tilde{\mathcal{F}}$ must vanish there, since for all assignments of values to the z_{ij}^k, if $P \in CV(\kappa, l_1, \ldots, l_n)$ then, by Lemma 5, $\tilde{P} \in CV(\kappa; \kappa, \ldots, \kappa)$.

Now suppose all polynomials in $\mathcal{F}(\kappa; l_1, l_2, \ldots, l_n)$ vanish on a tensor $P_0 \in \mathbb{C}^{l_1 l_2 \cdots l_n}$. Then, flattening P_0 on the edge of the tree leading to a_n gives a matrix of rank $l \leq \kappa$, so we can write

$$P_0 = Q_0' * B_n'$$

where Q_0' is a $l_1 \times l_2 \times \cdots \times l_{n-1} \times l$ tensor and B_n' is a $l \times l_n$ matrix. Construct a $\kappa \times l_n$ matrix B_n of rank κ by augmenting B_n' with additional rows. Similarly augment Q_0' with additional zero entries to obtain a $l_1 \times l_2 \times \cdots \times l_{n-1} \times \kappa$ tensor Q_0 with $P_0 = Q_0 * B_n$. Now there exists a $l_n \times \kappa$ matrix A_n so that $B_n A_n = I$, the identity matrix. Thus $P_0 * A_n * B_n = Q_0 * B_n * A_n * B_n = Q_0 * I * B_n = P_0$.

Proceeding similarly for the other taxa, we obtain matrices A_k, B_k such that

$$(P_0 *_{k,1} A_k) *_{k,1} B_k = P_0 *_{k,1} (A_k * B_k) = P_0.$$

By simultaneously letting each A_k act in the kth index of P_0, we obtain a $\kappa \times \kappa \times \cdots \times \kappa$ tensor $\tilde{P}_0$. Because all polynomials in $\tilde{\mathcal{F}}$ vanish on P_0, all polynomials in $\mathcal{F}$ vanish on $\tilde{P}_0$. Thus by our choice of $\mathcal{F}$, $\tilde{P}_0 \in CV(\kappa; \kappa, \kappa, \ldots, \kappa)$. Since, by repeated applications of Lemma 5, letting each B_k act in the kth index maps $CV(\kappa; \kappa, \kappa, \ldots, \kappa)$ to $CV(\kappa; l_1, l_2, \ldots, l_n)$, and maps $\tilde{P}_0$ to P_0, we see $P_0 \in CV(\kappa; l_1, l_2, \ldots, l_n)$. □

We now state an ideal-theoretic version of this result.

Theorem 7. *Suppose $l_1, l_2, \ldots, l_n \geq \kappa$, and $\mathcal{F}$ is a set of polynomials generating $\mathfrak{a}(\kappa; \kappa, \kappa, \ldots, \kappa)$. Then the set $\mathcal{F}(\kappa; l_1, l_2, \ldots, l_n)$ constructed from $\mathcal{F}$ as in Theorem 6 generates $\mathfrak{a}(\kappa; l_1, l_2, \ldots, l_n)$.*

Since the key argument in the proof will be used again in Section 8, we present it as a lemma.

Lemma 8. *Let V_1 and V_2 be subvarieties of $\mathbb{C}^{nm_1}$ and $\mathbb{C}^{nm_2}$, respectively, with $m_1 \leq m_2$, such that, when points are written as $n \times m_1$ and $n \times m_2$ matrices,*

$$V_1 = V_1 * M(m_1, m_1, \mathbb{C}),$$

and

$$V_2 = \overline{V_1 * M(m_1, m_2, \mathbb{C})}.$$

Let $\mathfrak{a}_i$ denote the ideal of all polynomials vanishing on V_i.

Then $\mathfrak{a}_2$ is generated by the $(m_1+1) \times (m_1+1)$ minors of an $n \times m_2$ matrix P of indeterminants, together with all polynomials of the form $f(P * A)$, where $f \in \mathfrak{a}_1$ and $A \in M(m_2, m_1, \mathbb{C})$.

Proof. Let $\mathfrak{b}$ denote the ideal generated by the $(m_1 + 1) \times (m_1 + 1)$ minors, together with the polynomials $f(P * A)$ described above.

First we show $\mathfrak{a}_2 \supseteq \mathfrak{b}$. It is enough to show the specified generators of $\mathfrak{b}$ vanish on $V_1 * M(m_1, m_2, \mathbb{C})$. Since all points in this set are matrices of rank at most m_1, the specified minors vanish there. To see the $f(P * A)$ vanish there, consider a point $Q_0 * B$ where $Q_0 \in V_1$, $B \in M(m_1, m_2, \mathbb{C})$. Then $Q_0 * B * A \in V_1$ since $B * A \in M(m_1, m_1, \mathbb{C})$. Thus $f(P * A)$ vanishes at $Q_0 * B$.

Our argument that $\mathfrak{a}_2 \subseteq \mathfrak{b}$ is more involved.

Note $GL(m_2, \mathbb{C})$ acts on $V_1 * M(m_1, m_2, \mathbb{C})$, and hence on V_2 as well. Consider the mth degree homogeneous component $\mathfrak{a}_2^{(m)}$ of $\mathfrak{a}_2$. Then the $GL(m_2, \mathbb{C})$-action on V_2 gives a representation of $GL(m_2, \mathbb{C})$ on $\mathfrak{a}_2^{(m)}$, in which $C \in GL(m_2, \mathbb{C})$ maps $g(P) \mapsto g(P * C)$. Since $GL(m_2, \mathbb{C})$ is reductive, this representation decomposes into a sum of irreducible ones. Consider now one of the irreducible subspaces, W. It will be enough to show that $W \subseteq \mathfrak{b}$.

Consider any non-zero polynomial $g(P) \in W$. Let Q denote a $n \times m_1$ matrix of indeterminants. Then for any $B \in M(m_1, m_2, \mathbb{C})$, the polynomial $g_B(Q) = g(Q * B)$ vanishes on V_1, since $Q_0 \mapsto Q_0 * B$ maps V_1 to V_2. Thus $g_B \in \mathfrak{a}_1$.

Suppose first that for all $B \in M(m_1, m_2, \mathbb{C})$ the polynomial $g_B(Q)$ is identically zero. Then g must vanish on all $n \times m_2$ matrices of rank at most m_1, since any such matrix can be written as $Q_0 * B$ for some complex matrices $Q_0 \in M(n, m_1, \mathbb{C})$, $B \in M(m_1, m_2, \mathbb{C})$, and then $g(Q_0 * B) = g_B(Q_0) = 0$. Thus if all g_B are identically zero, then g is in the ideal generated by $(m_1 + 1) \times (m_1 + 1)$ minors of P, and hence $g \in \mathfrak{b}$.

Suppose, then, that for some B the polynomial g_B is not identically zero. Let $D \in M(m_2, m_1, \mathbb{C})$ be chosen so that $h(P) = g_B(P * D)$ is a non-zero polynomial. Such a D must exist since $m_1 \leq m_2$. (For

instance, D may be taken so that its first m_1 rows form an identity and the remaining rows are zero.) Then $h(P) = g(P * DB)$, where DB is a complex $m_2 \times m_2$ matrix that is generally not invertible.

Nonetheless, the irreducibility of W implies that $h(P) \in W$. This is simply because W is closed in $\mathfrak{a}_2^{(m)}$, and so must contain the closure of the orbit of g under $GL(m_2, \mathbb{C})$, and this closure contains $g(P * DB)$.

Now since $g(P) \in W$, $h(P) = g_B(P * D) \in W$, and $g_B \in \mathfrak{a}_1$, the irreducibility of W implies $g(P)$ is in the span of polynomials of the form $f(P * A)$ where $f \in \mathfrak{a}_1$ and $A \in M(m_2, m_1, \mathbb{C})$. Thus in this case as well, $g \in \mathfrak{b}$. □

Proof of Theorem 7. Let $\mathfrak{a} = \mathfrak{a}(\kappa; l_1, \ldots, l_n)$, and let $\mathfrak{b}$ be the ideal generated by $\mathcal{F}(\kappa; l_1, \ldots, l_n)$, the set defined in Theorem 6. Note that $\mathfrak{b}$ is equivalently described as generated by $\mathcal{F}_{edge} \cup \tilde{\tilde{\mathcal{F}}}$, where $\tilde{\tilde{\mathcal{F}}}$ denotes the set of all polynomials of the form $f(\tilde{P})$ where $f \in \mathcal{F}$ and $\tilde{P}$ is obtained from a tensor P of indeterminants by the action of numerical matrices $Z_k \in M(l_k, \kappa, \mathbb{C})$ in each index k.

That $\mathfrak{a} \supseteq \mathfrak{b}$ was shown in the proof of Theorem 6. To establish $\mathfrak{a} \subseteq \mathfrak{b}$. we proceed by induction on the number of $l_k > \kappa$, the base case of zero being trivial.

If at least one such $l_k > \kappa$ exists, assume $l_n > \kappa$. Then let $V_1 = CV(\kappa; l_1, \ldots, l_{n-1}, \kappa)$ and $V_2 = CV(\kappa; l_1, l_2, \ldots, l_n)$. We view points on V_1 and V_2 as $l_1 \cdots l_{n-1} \times \kappa$ and $l_1 \cdots l_{n-1} \times l_n$ matrices, respectively, by flattening on the edge of the star tree leading to the nth leaf. Using Lemma 5 we see that $V_1 * M(\kappa, \kappa, \mathbb{C}) = V_1$ and, since $l_n > \kappa$, that $V_2 = \overline{V_1 * M(\kappa, l_n)}$. Therefore we may apply Lemma 8, and obtain that $\mathfrak{a}$ is generated by the $(\kappa+1) \times (\kappa+1)$ minors of the flattening of P on the edge to the nth leaf, together with all polynomials $f(P * A)$ where $f \in \mathfrak{a}(\kappa; l_1, \ldots, l_{n-1}, \kappa)$ and $A \in M(l_n, \kappa, \mathbb{C})$. We thus need only show such $f(P * A)$ are in $\mathfrak{b}$.

Now by induction, $\mathfrak{a}(\kappa; l_1, \ldots, l_{n-1}, \kappa)$ is generated by $(\kappa+1) \times (\kappa+1)$ minors of edge flattenings of a $l_1 \times \cdots \times l_{n-1} \times \kappa$ tensor Q of indeterminants, together with polynomials of the form $h(\tilde{Q})$, where $h \in \mathfrak{a}(\kappa; \kappa, \ldots, \kappa)$ and $\tilde{Q}$ is a $\kappa \times \cdots \times \kappa$ tensor obtained from Q by letting elements of $M(l_i, \kappa, \mathbb{C})$ (respectively $M(\kappa, \kappa, \mathbb{C})$) act on Q in the ith index for each $i \neq n$ (respectively $i = n$). We may thus assume f itself has one of these forms.

In the first case, where $f \in \mathfrak{a}(\kappa; l_1, \ldots, l_{n-1}, \kappa)$ is a minor of an edge flattening for the model, we see f vanishes on all tensors Q that have rank at most κ when flattened on a certain edge e not leading to the nth leaf. But if P is a $l_1 \times \cdots \times l_n$ tensor with $\text{rank}(Flat_e(P)) \leq \kappa$,

then $\operatorname{rank}(Flat_e(P * A)) \leq \kappa$ as well, for all $A \in M(l_n, \kappa, \mathbb{C})$. Thus $f(P * A)$ vanishes on all tensors such that $\operatorname{rank}(Flat_e(P)) \leq \kappa$, and so $f(P * A)$ is in the ideal generated by $(\kappa+1) \times (\kappa+1)$ minors from edge flattenings of P.

In the second case, where $f = h(\tilde{Q})$, we find $f(P * A) = h(\tilde{P})$ where $\tilde{P}$ is obtained from P by letting elements of $M(l_i, \kappa, \mathbb{C})$ act on P in the ith index for each i, and h vanishes on $V(\kappa; \kappa, \cdots, \kappa)$.

Thus in either case $f(P * A) \in \mathfrak{b}$. □

As a corollary, we prove several cases of Conjecture 21 in [GSS03] on the ideals $\mathfrak{a}(2; l_1, \ldots, l_n)$. We note the $n = 3$ case was first proved in [LM04] by invoking sophisticated methods of [Wey03].

Corollary 9. *For $n \leq 5$, the ideal $\mathfrak{a}(2; l_1, \ldots, l_n)$ associated to the hidden naive Bayes model with a 2-state hidden variable and n observed variables with $l_1, \ldots, l_n$ states, is generated by the 3×3 minors of all 2-dimensional flattenings associated to bipartitions of the observed variables.*

Proof. Since there are no polynomials vanishing on $V(2; 2, 2, 2) = \mathbb{P}^7$, by Theorem 7 the set of polynomials vanishing on $V(2; l_1, l_2, l_3)$ is generated by edge invariants.

By calculations of [GSS03], the statement holds for the two cases $V(2; 2, 2, 2, 2)$ and $V(2; 2, 2, 2, 2, 2)$. The corollary then follows from Lemma 10 below. □

Lemma 10. *Suppose for the n-leaf star tree $\mathfrak{a}(2; 2, \ldots, 2)$ is generated by the 3×3 minors of all 2-dimensional flattenings of $2 \times \cdots \times 2$ tensors according to bipartitions of the observed variables. Then $\mathfrak{a}(2; l_1, \ldots, l_n)$ is generated by the 3×3 minors of all 2-dimensional flattenings of $l_1 \times \cdots \times l_n$ tensors according to bipartitions of the observed variables.*

Proof. By Theorem 7, $\mathfrak{a}(2; l_1, \ldots, l_n)$ is generated by all 3×3 minors of edge flattenings of an $l_1 \times \cdots \times l_n$ tensor of indeterminants P, together with all 3×3 minors of all 2-dimensional flattenings of all $\tilde{P}$, where $\tilde{P}$ denotes a $2 \times \cdots \times 2$ tensor obtained from P by an action in each index i by matrices $A_i \in M(l_i, 2, \mathbb{C})$. One readily sees such flattenings of $\tilde{P}$ can be expressed as $\tilde{F} = B_1 * F * B_2$, where F is the corresponding flattening of P and the B_j are matrices depending on the A_i. But then the 3×3 minors of such a flattening of $\tilde{P}$ will be zero provided F has rank ≤ 2. Thus these polynomials are in the ideal $\mathfrak{b}$ generated by 3×3 minors of flattenings of P. Therefore $\mathfrak{a}(2; l_1, \ldots, l_n) \subseteq \mathfrak{b}$.

That $\mathfrak{a}(2; l_1, \ldots, l_n) \supseteq \mathfrak{b}$ is clear. □

A proof of the full Conjecture 21 of [GSS03] would follow from the following special cases:

Conjecture 2. *(Garcia,Stillman,Sturmfels) The ideal* $\mathfrak{a}(2; 2, 2, \ldots, 2)$*, that is, the ideal associated to the hidden naive Bayes model with a 2-state hidden variable and n 2-state observed variables, is generated by the* 3×3 *minors of all 2-dimensional flattenings arising from bipartitions of the observed variables.*

7. Set-theoretic description of the phylogenetic variety: arbitrary κ.

For the remainder of this paper, we return to the consideration of models on phylogenetic trees. We first establish a set-theoretic result that provides some evidence for Conjecture 1, for arbitrary κ.

Theorem 11. *For a 3-leaf star tree, let* $\mathcal{F}$ *be a set of polynomials defining* $V(\kappa; \kappa, \kappa, \kappa)$ *set-theoretically, and let* $\mathcal{F}(\kappa; l_1, l_2, l_3)$ *be as defined in Theorem 6. For an n-taxon tree* T*, let* $\mathcal{F}(T)$ *be the union of all sets* $\mathcal{F}(\kappa; \kappa^{n_1}, \kappa^{n_2}, \kappa^{n-n_1-n_2})$ *associated to 3-dimensional flattenings at nodes of* T*. Then the zero set of* $\mathcal{F}(T)$ *is the phylogenetic variety* $V(T)$*.*

More informally, from polynomials whose zero set is $V(T_3)$, one can explicitly construct polynomials whose zero set is $V(T)$ for any n-taxon tree T. While one might naively view the case of $V(T_3)$ as the simplest, in fact it is the only remaining barrier to the determination of polynomials defining the n-taxon variety, for any n.

For the remainder of this section, fix a set $\mathcal{F}$ of polynomials whose zero set is $V(\kappa; \kappa, \kappa, \kappa)$, and let $V_{Flat}(T)$ denote the zero set of the resulting $\mathcal{F}(T)$. Our proof of the theorem will follow several lemmas. The first is an analog for $V_{Flat}(T)$ of Lemma 4.

Lemma 12. *Let* T' *and* T'' *be n-taxon and m-taxon trees, with* $T = T' * T''$*. If* $Q \in CV_{Flat}(T')$ *and* $R \in CV_{Flat}(T'')$*, then* $Q * R \in CV_{Flat}(T)$*.*

Proof. Consider any internal node v of T, which we may assume arises from an internal node of T'. Flattening Q at v, the resulting 3-dimensional tensor lies on $CV(\kappa; \kappa^{n_1}, \kappa^{n_2}, \kappa^{n_3})$, with $n_3 = n - n_1 - n_2$, where we assume taxon a_n of T' (where taxon b_1 of T'' is to be joined) is included in the last index of the flattening. Then the flattening of $Q * R$ at v is obtained from the flattening of Q at v by an action in the third index by a matrix R' whose entries are determined by those of R. By Lemma 5 the flattening of $Q * R$ at v lies in $CV(\kappa; \kappa^{n_1}, \kappa^{n_2}, \kappa^{n_3+m-2})$.

Thus $Q * R \in CV_{Flat}(T)$. □

We also need a converse to this lemma.

Lemma 13. *Let T' and T'' be n-taxon and m-taxon trees, with $T = T' * T''$. Then if $P \in CV_{Flat}(T)$, there exist $Q \in CV_{Flat}(T')$ and $R \in CV_{Flat}(T'')$ with $P = Q * R$.*

Proof. Let e be the edge of T formed by conjoining edges of T' and T''. Since any $P \in CV_{Flat}(T)$ satisfies the edge invariants for e, we may flatten it on e to obtain a $\kappa^{n-1} \times \kappa^{m-1}$ matrix of rank $l \leq \kappa$, and write

$$P = Q * R,$$

where Q and R are n- and m-dimensional tensors, respectively, with all indices running through $[\kappa]$. We may further assume the non-zero $Q_k = Q(\cdot, \cdots, \cdot, k)$ are linearly independent, as are the non-zero $R_k = R(k, \cdot, \cdots, \cdot)$, and that Q_k, R_k are non-zero only for $k = 1, \ldots, l \leq \kappa$.

We next show $Q \in CV_{Flat}(T')$. First observe that since the non-zero R_k are independent, if we write them as row vectors, there is a $\kappa^{m-1} \times \kappa$ matrix A so that $R_k A = \mathbf{e}_k$ for all $k \leq l$. Now supposing the taxa of T' and T'' are $a_1, \ldots, a_n$ and $b_1, \ldots, b_m$, respectively, flatten P according to the partition $\{\{a_1\}, \ldots, \{a_{n-1}\}, \{b_2, \ldots, b_m\}\}$ to an n-dimensional $\kappa \times \cdots \times \kappa \times \kappa^{m-1}$ tensor F. Letting R' denote the $\kappa \times \kappa^{m-1}$ flattened form of R with rows R_k, we have $F = Q * R'$. Thus $F * A = Q * R' * A = Q$. (Note that A does not act in a single index of P here, but does act in a single index of the flattening F.) It is now straightforward to see that any 3-dimensional flattening of Q at an internal vertex of T' is obtained from a flattening of P at a vertex of T, followed by an action in one of the three resulting indices of a matrix determined by A. Thus by the definition of $\mathcal{F}(T)$, Q will satisfy all polynomials in $\mathcal{F}(T')$.

Similarly, $R \in CV_{Flat}(T'')$. □

Proof of Theorem 11. We already know that $V_{Flat}(T) \supseteq V(T)$.

The proof that $V_{Flat}(T) = V(T)$ proceeds by induction on the number n of taxa. The cases of $n = 2, 3$ hold by the definition of $\mathcal{F}(T)$.

Consider an n-taxon tree $T = T_n$, $n \geq 4$, and picking a cherry of T, let T_{n-1} and T_3 be such that $T = T_{n-1} * T_3$. Suppose $P \in CV_{Flat}(T)$. By Lemma 13, we have $P = Q * R$, for $Q \in CV_{Flat}(T_{n-1})$ and $R \in CV_{Flat}(T_3)$. This, in combination with Lemma 12, means the map

$$\mu : CV_{Flat}(T_{n-1}) \times CV_{Flat}(T_3) \to CV_{Flat}(T_n)$$

defined by $(Q, R) \mapsto Q * R$ is surjective.

Denote the parameterizations of the cones over the phylogenetic varieties for T_k by $\psi_k : U_k \to \mathbb{C}^{L_k}$. With the map $\alpha : U_{n-1} \times U_3 \to U_n$

defined by $\alpha(u_{n-1}, u_3) = u_{n-1} * u_3$, the diagram

$$\begin{array}{ccc} U_{n-1} \times U_3 & \xrightarrow{\psi_{n-1}\times\psi_3} & CV_{Flat}(T_{n-1}) \times CV_{Flat}(T_3) \\ \alpha\downarrow & & \downarrow\mu \\ U_n & \xrightarrow{\psi_n} & CV_{Flat}(T) \end{array}$$

commutes, by Lemma 3.

Now α and μ are surjective, and the image of $\psi_{n-1} \times \psi_3$ is dense in $CV_{Flat}(T_{n-1}) \times CV_{Flat}(T_3)$ by the inductive hypothesis, so the image of ψ_n is dense in $CV_{Flat}(T)$. Thus $V_{Flat}(T) = V(T)$. □

Theorem 11 and the preceeding lemmas yield the following strengthening of Lemma 4.

Corollary 14. *If $T = T' * T''$, then $CV(T) = CV(T') * CV(T'')$.*

8. The phylogenetic ideal: $\kappa = 2$.

We now prove Theorem 2. Our arguments will use in several ways that for $\kappa = 2$ the variety $V(T_3)$ fills its ambient space: $V(T_3) = \mathbb{P}^{\kappa^3-1}$. Note, however, that for $\kappa > 2$, $V(T_3) \subsetneq \mathbb{P}^{\kappa^3-1}$, and so the approach here cannot be successfully modified in a simple way.

The first use of this special fact is to note that for our chosen κ, $V(2; 2, 2, 2) = V(T_3) = \mathbb{P}^7$ means the set $\mathcal{F}$ defining $V(T_3)$ is empty. Thus the set $\mathcal{F}(T_n)$ of the set-theoretic result Theorem 11 is the set of edge invariants. While our goal is to show $\mathcal{F}(T_n)$ generates the full ideal vanishing on $V(T_n)$, we will not, in fact, appeal to Theorem 11 to do so.

The second use of $V(T_3) = \mathbb{P}^{\kappa^3-1}$ is more subtle. Recall that regardless of κ, there are actions of $GL(\kappa, \mathbb{C})$ on $V(T_n)$ in each index. However, in the case $\kappa = 2$, the special nature of $V(T_3)$ gives us actions of $GL(\kappa^2, \mathbb{C})$ on $V(T_n)$ via the cherries of T_n. This is really the key point in our argument, as it underlies the application of Lemma 8. Nonetheless, this action is in some respect an 'unnatural' consequence of $\kappa = 2$. The following lemma provides a more careful statement of the special structure we use.

Lemma 15. *Let T_n denote an n-taxon tree, labeled so that taxa a_{n-1} and a_n form a cherry. Write $T_n = T_{n-1} * T_3$, where a_{n-1}, a_n are taxa on T_3. Let e denote the edge of T_n formed from conjoining edge $\tilde{e}$ of T_{n-1} and the appropriate edge of T_3. View points in $CV(T_n)$ and $CV(T_{n-1})$ as $2^{n-2} \times 4$ and $2^{n-2} \times 2$ matrices by flattening them on the edges e and*

$\tilde{e}$, respectively. Then

$$CV(T_n) = \overline{CV(T_{n-1}) * M(2,4,\mathbb{C})}$$

and

$$CV(T_{n-1}) = CV(T_{n-1}) * M(2,2,\mathbb{C}).$$

Proof. The first claim is simply Lemma 4 applied to T_{n-1} and T_3, combined with the observation that $CV(T_3) = \mathbb{C}^8$ flattens to give $M(2,4,\mathbb{C})$. (Note that by Corollary 14, we could also remove the closure symbol here.)

For the second claim, apply the same argument to T_{n-1} and T_2, observing that $CV(T_2) = M(2,2,\mathbb{C})$. □

Proof of Theorem 2. We proceed by induction on the number n of taxa for T_n, with the cases of $n = 2, 3$ known.

Let $\mathfrak{a} = \mathfrak{a}_T$ denote the ideal vanishing on $CV(T)$, and $\mathfrak{b}$ the ideal generated by $\mathcal{F}_{edge}(T)$. That $\mathfrak{a} \supseteq \mathfrak{b}$ has been discussed already; we must show the opposite inclusion.

With $T_n = T$, choose a cherry so that $T_n = T_{n-1} * T_3$, with notation as in Lemma 15. By that lemma, we may apply Lemma 8 with $V_1 = CV(T_{n-1})$ and $V_2 = CV(T_n)$. We thus find $\mathfrak{a}$ is generated by the 3×3 minors of the edge flattening $Flat_e(P)$ on the conjoined edge e of an n-dimensional tensor of indeterminants P, together with all polynomials of the form $g(P) = f(Flat_e(P) * B)$ where $f(Q)$ vanishes on $CV(T_{n-1})$, Q is a $(n-1)$-dimensional tensor of indeterminants, and $B \in M(4,2,\mathbb{C})$.

Now, by induction, the ideal of such f is generated by 3×3 minors of $Flat_{e'}(Q)$ as e' ranges through edges of T_{n-1}. Consider one such minor, say f_0, obtained from the flattening on an edge e_0 of T_{n-1}. We may assume $e_0 \neq \tilde{e}$, since otherwise there are no 3×3 minors. It will be enough to show $f_0(Flat_e(P) * B) \in \mathfrak{b}$.

We claim that $f_0(Flat_e(P) * B)$ vanishes on all P that have rank at most 2 when flattened on the edge e_0 in T_n. For such a P, since $Flat_{e_0}(P)$ is $2^m \times 2^{n-m}$, there is an expression $P = P_1 * P_2$, where P_1 is an $(m+1)$-dimensional $2 \times \cdots \times 2$ tensor, and P_2 an $(n-m+1)$-dimensional $2 \times \cdots \times 2$ tensor. Then writing P and P_2 as $2 \times \cdots \times 2 \times 4$ tensors by flattening to combine the taxa a_{n-1}, a_n, we have $P * B = P_1 * (P_2 * B)$. This shows $P * B$ also has rank at most 2 when flattened on e_0, and so f_0 vanishes on it, as claimed.

But since $f_0(Flat_e(P) * B)$ vanishes on all P of rank at most 2 when flattened on e_0, it is contained in the ideal generated by 3×3 minors of flattenings on e_0. Thus it is in $\mathfrak{b}$. □

References

[AR03] Elizabeth S. Allman and John A. Rhodes. Phylogenetic invariants for the general Markov model of sequence mutation. *Math. Biosci.*, 186:113–144, 2003.

[AR04a] Elizabeth S. Allman and John A. Rhodes. Phylogenetic invariants for stationary base composition. *J. Symbolic Comp.*, 2004. to appear, `arXiv:q-bio.PE/0407035`.

[AR04b] Elizabeth S. Allman and John A. Rhodes. Quartets and parameter recovery for the general Markov model of sequence mutation. *App. Math. Res. Express (AMRX)*, 2004:4:107–131, 2004.

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

[Eri04] Nicholas Eriksson. Toric ideals of homogeneous phylogenetic models. 2004. `arXiv:math.CO/0401175`.

[ERSS04] Nicholas Eriksson, Kristian Ranestad, Bernd Sturmfels, and Seth Sullivant. Phylogenetic Algebraic Geometry. 2004. `arXiv:math.AG/0407033`.

[ES93] Steven N. Evans and T. P. Speed. Invariants of some probability models used in phylogenetic inference. *Ann. Statist.*, 21(1):355–377, 1993.

[Fel04] Joseph Felsenstein. *Inferring Phylogenies*. Sinauer Associates, Sunderland, MA, 2004.

[GSS03] Luis David Garcia, Michael Stillman, and Bernd Sturmfels. Algebraic geometry of Bayesian networks, 2003. `arXiv:math.AG/0301255`.

[HP93] Michael D. Hendy and David Penny. Spectral analysis of phylogenetic data. *J. Class.*, 10:1–20, 1993.

[Lak87] J.A. Lake. A rate independent technique for analysis of nucleic acid sequences: Evolutionary parsimony. *Mol. Bio. Evol.*, 4:167–191, 1987.

[LM04] J. M. Landsberg and L. Manivel. On the ideals of secant varieties of Segre varieties. *Found. Comput. Math.*, 2004. to appear.

[PS04] Lior Pachter and Bernd Sturmfels. Tropical geometry of statistical models. *Proc. Natl. Acad. Sci. USA*, 2004. to appear, `http://arxiv.org/abs/q-bio.QM/0311009`.

[SS04] Bernd Sturmfels and Seth Sullivant. Toric ideals of phylogenetic invariants. *J. Comput. Biol.*, 2004. to appear, `arXiv:q-bio.PE/0402015`.

[SSH94] M.A. Steel, L. Székely, and M.D. Hendy. Reconstructing trees from sequences whose sites evolve at variable rates. *J. Comput. Biol.*, 1(2):153–163, 1994.

[Str83] V. Strassen. Rank and optimal computation of generic tensors. *Linear Algebra Appl.*, 52/53:645–685, 1983.

[Wey03] Jerzy Weyman. *Cohomology of vector bundles and syzygies*, volume 149 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 2003.

ESA: Department of Mathematics and Statistics, University of Southern Maine, Portland, Maine 04104

E-mail address: `eallman@maine.edu`

JAR: Department of Mathematics, Bates College, Lewiston, Maine 04240
E-mail address: jrhodes@bates.edu